FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Biomedical Multimodal Explanations – Increasing Diversity and Complementarity in Explainable Artificial Intelligence

Diogo Baptista Martins da Mata

Mestrado em Engenharia Biomédica

Supervisors: Jaime S. Cardoso, Wilson Silva

June 19, 2022

Biomedical Multimodal Explanations – Increasing Diversity and Complementarity in Explainable Artificial Intelligence

Diogo Baptista Martins da Mata

Mestrado em Engenharia Biomédica

Faculdade de Engenharia da Universidade do Porto

June 19, 2022

Resumo

Ao longo dos últimos anos, as abordagens de Deep Learning têm sido o "gold standard" para uma miríade de tarefas e aplicações quer a nível de investigação quer em várias indústrias. Isto deve-se principalmente à obtenção de grandes resultados e desempenhos.

Deep Learning está enraizado no conceito de redes neuronais artificiais - estruturas que pretendem traduzir a conjuntura do cérebro humano para um contexto computacional. Nesse sentido, redes neuronais artificiais reproduzem, de forma simplificada, o complexo sistema cerebral através das suas unidades, que representam os neurónios, e das ligações entre unidades, que implicam a transmissão de informação que existe entre as mesmas. A conformação que se reflecte em tal formulação envolve uma grande quantidade de cálculos que, consequentemente, conduz a um alto grau de complexidade que, por sua vez, se repercute como impraticável para compreensão e análise humana. Assim, torna-se normal percepcionar estes sistemas como "caixas negras" que recebem certos inputs e produzem outputs diferenciados. Tal paradigma revela um contratempo considerável para a sua implementação em determinados contextos. Em áreas altamente regulamentadas, como os cuidados de saúde, existe uma procura de estruturas que sejam capazes de fornecer algum tipo de fundamento ou raciocínio lógico para a sua funcionalidade. Têm de ser explicáveis e dignos de confiança. A "Explainable Artificial Intelligence" é o domínio que visa satisfazer estes requisitos, estudando e concebendo estratégias e métodos que possibilitem a explicação da ação de sistemas de inteligência artificial, e que por conseguinte viabilizem uma compreensão razoável e acessível por parte de agentes humanos.

Actualmente, a maioria dos algoritmos de explicabilidade de última geração focam-se no fornecimento de mapas de saliência que realçam as ocorrências mais importantes para as decisões de modelos de Machine Learning. No entanto, estes mapas de saliência são frequentemente difíceis de ler, ou porque o seu consumidor não é um perito ou porque a sua representação e uma ideia semântica compreensível para o ser humano não estão diretamente ligados. A multimodalidade pode ajudar a solucionar este problema. A incorporação de dados multimodais nas explicações criadas reforça a sua diversidade e complementaridade, aumentando a probabilidade de ter pelo menos uma explicação que o consumidor compreenda. Esta dissertação tem como objectivo fundamental investigar e desenvolver metodologias de geração de explicações multimodais para sistemas de apoio ao diagnóstico médico. Nesse sentido, o trabalho proposto visa a utilização de imagens de raios-X e seus relatórios de radiologia associados para produzir explicações visuais e textuais de qualidade, no âmbito de tarefas de classificação.

O trabalho concretizado é segmentado em três vertentes: experiências independentes de geração de explicações unimodais a partir de dados unimodais; experiências de produção de explicações visuais recorrendo a dados multimodais; experiências de geração de explicações textuais usando dados multimodais. As abordagens exploradas revelaram ter potencial, dado que foi possível produzir explicações viáveis de forma automática sem prejudicar a robustez dos exercícios de diagnóstico. A nível visual, introduziram-se claras melhorias na qualidade das explicações e das próprias tarefas em si, enquanto no âmbito textual os resultados acabaram por ser menos claros.

Abstract

Over the past few years, Deep Learning approaches have been the "gold standard" for a myriad of tasks and applications both in research and in various industries. This is mainly due to the achievement of great results and performances.

Deep Learning is rooted in the concept of artificial neural networks - structures that aim to translate the conjuncture of the human brain into a computational context. In this sense, artificial neural networks reproduce, in a simplified way, the complex brain system through its units, which represent neurons, and through the connections between units, which involve the transmission of information that exists between the latter. The conformation that is reflected in such a formulation involves a large amount of computation which, consequently, leads to a high degree of complexity that, in turn, resonates as impractical for human understanding and analysis. Thus, it becomes normal to perceive these systems as "black boxes" that receive certain inputs and produce differentiated outputs. Such a paradigm reveals a considerable setback for implementation in certain contexts. In highly regulated areas, such as healthcare, there is a demand for structures that are able to provide some kind of rationale or logical reasoning for their functionality. They have to be explainable and trustworthy. Explainable Artificial Intelligence is the domain that aims to satisfy these requirements by studying and devising strategies and methods that make it possible to explain the action of artificial intelligence systems, and therefore enable a reasonable and accessible understanding by human agents.

Currently, most state-of-the-art explainability algorithms focus on providing saliency maps that highlight the most important instances for Machine Learning models' decisions. However, these saliency maps are often difficult to read, either because their consumer is not an expert or because their representation and a human-understandable semantic idea are not directly connected. Multimodality may be one way to address this problem. Incorporating multimodal data into explanations strengthens their diversity and complementarity, increasing the likelihood of having at least one explanation that the consumer understands. This dissertation fundamentally aims to investigate and develop methodologies for generating multimodal explanations in medical diagnostic support systems. Accordingly, the proposed work targets the production of quality visual and textual explanations within classification exercises, by using X-ray images and their associated radiology reports.

The work is segmented into three threads: independent experiments on generating unimodal explanations from unimodal data; experiments on producing visual explanations using multimodal data; experiments on generating textual explanations using multimodal data. The various explored approaches proved to have a lot of potential as it was possible to yield viable explanations in an automated way without undermining the robustness of the diagnostic exercises. At the visual level, clear improvements were made in the quality of the explanations and the tasks themselves, while at the textual level the results turned out to be less satisfactory and clear.

Acknowledgements

Initially, I would like to start by thanking my supervisor professor Jaime S. Cardoso for the opportunity to work on a subject that gave and gives me tremendous pleasure, for the guidance, and advisory. I would also like to thank my co-supervisor Wilson Silva for the support, ideas, availability, and proximity.

I would further like to say a word to the VCMI group, and especially to the interpretability subgroup that welcomed me with open arms and helped me so much with all the suggestions and useful discussions that became so important for the progress of my work.

The dissertation was developed within the scope of the Transparent Artificial Medical Intelligence (TAMI) project at INESC TEC. In this respect, I want to highlight the motivation that such context brought me for its execution.

To my amazing friends, thank you for the adventures, stories, laughter, comprehension, fun, and for motivating me to live my life to the fullest.

I am hugely thankful for my girlfriend Margarida. Thank you for brightening my days, for making me smile, for all the companionship and for being the most caring person I have ever met.

Finally, I want to express the immense gratitude I have for my family - my mother, father, and sister for always being there for me, for putting up with my tantrums, and essentially for all the love and loving environment where I grew up in. It made me who I am today, and I couldn't ask for better.

Diogo Mata

"Os casos excepcionais são todos os que há no mundo"

Agostinho da Silva

Contents

List of Figures

List of Tables

Abbreviations and Symbols

AI	Artificial Intelligence
ANN	Artificial Neural Network
ATH	Attention-based Triplet Hashing
BERT	Bidirectional Encoder Representations from Transformers
CBIR	Content-Based Image Retrieval
CCA	Canonical Correlation Analysis
CNN	Convolutional Neural Networks
CV	Computer Vision
DBM	Deep Boltzmann Machine
DBN	Deep Belief Network
DCCA	Deep Canonical Correlation Analysis
DL	Deep Learning
DNN	Deep Neural Networks
FC	Fully Connected
GAM	Generalised Additive Model
GAN	Generative Adversarial Networks
GLM	Generalised Linear Model
GPT	Generative Pre-Training
GRU	Gated Recurrent Units
KNN	K-Nearest Neighbours
LSTM	Long short-term memory
ML	Machine Learning
MLE	Maximum Likelihood Estimation
NER	Named Entity Recognition
NLG	Natural Language Generation
NLP	Natural Language Processing
NLU	Natural Language Understanding
PGM	Probabilistic Graphical Model
RBM	Restricted Boltzmann Machine
RNN	Recurrent Neural Network
SCDA	Selective Convolutional Descriptor Aggregation
VQA	Visual Question and Answering
XAI	Explainable Artificial Intelligence
mAP	Mean Average Precision
nDCG	Normalized Discounted Cumulative Gain
ViT	Vision Image Transformer
DeiT	Data-Efficient Image Transformer
BEiT	Bidirectional Encoder Image Transformer

KD	Knowledge Distillation
BPE	Byte-Pair-Encoding
MLP	Multilayer Perceptron
BLEU	Bilingual Evaluation Understudy
METEOR	Metric for Evaluation of Translation with Explicit Ordering
ROUGE-L	Recall-Oriented Understudy for Gisting Evaluation
CIDEr	Consensus-based Image Description Evaluation

Chapter 1

Introduction

1.1 Context

The advancements in modern computer science technology and computational resources allow Machine Learning (ML) to be present in our daily lives and routines, applying its functionalities to devices like mobile phones and cars. Besides that, the increasing accuracy and ever-growing performance improvements of this type of system, regarding countless different tasks, has made them a very popular resource in various industries and research areas. It is important to denote that the majority of recent developments within the ML spectrum have arisen from Deep Learning (DL) approaches and techniques. Thus, varied Deep Neural Network (DNN) architectures have emerged in different domains.

In the medical field, there has been considerable progress towards the integration of Artificial Intelligence (AI) and specifically DNNs, over the recent years. Diagnostic support [1] and retrieval systems [2] represent that reality. However, there is still a big gap between the current panorama and what can be the potential of these technologies, which may not be so true in other areas. This is mainly due to two reasons: healthcare is an extremely high regulated environment, and DL models are the majority of times seen as black boxes.

Despite the many different configurations and tasks assigned to DNNs, the general sense that they function as opaque contrivances, where inputs are fed and a distinct result or prediction is produced as an output, prevails. This happens thanks to what differs DL from the traditional ML procedures: the high complexity of the designed structures that lead to millions of non-linear calculations, which in themselves are virtually impossible to follow and make a simple reasoning of, and the automatic feature extraction, which adds another layer of opacity to the models. The entailed hindrances provide for mechanisms to become more and more secluded and unpredictable. Normally, this behavior is not permitted within highly regulated fields like finances [3] or medicine [4]. There exist various constraints related to safety and legal issues that ensue some sort of insight on the function of the mentioned strategies. In other words, there is a need for explainability and interpretability. Taking the example of medicine, it is imperative that a clinician trusts a supportive diagnostic-related device that makes use of artificial intelligence. In that manner, if an

explanation is given in the context of some specific result, if a trace or a relatively logical decision structure is displayed, it is possible to work towards the goal of trusting the computerized output.

There are already studies dedicated to examining ways of making more transparent arrangements, either by causing them to be inherently interpretable [5, 6] or by providing them with tools that allow the provision of explanations [7, 8]. This theme is still an evolving topic and over time, new strategies and mechanisms, that are more effective in executing more perceptible and trustworthy configurations without compromising or residually compromising operability, will appear.

Nevertheless, there is a clear overload in unimodal efforts when compared with multimodal ones, which exposes the lacking diversity within methodologies, in the realm of Explainable AI (XAI). Presently, the majority of state-of-the-art interpretability algorithms provide saliency maps that emphasize the most important elements regarding model decisions. Notwithstanding, these saliency maps are frequently challenging to interpret, either because the explanation target consumer is not an expert or because the image/saliency map and a human-understandable semantic idea are not intuitively connected. The incorporation of multimodal data in explanations increases its complementarity aspect, and consequently improves the likelihood of having at least one explanation that the consumer understands. Such a circumstance becomes significantly indispensable in healthcare services since adopting procedures of this sort can certainly bring advantages, as it facilitates compliance with the industry demands. Furthermore, in this sector, there exists a natural coexistence between data of many kinds (images, text, tabular data), which creates a chance to explore more and more differentiated ways of developing multimodal explanations.

1.2 Motivation

The growth of technology in the medical field, and more specifically, the evolution of the digitalization of health records allowed for a greater possibility of access and to take advantage of large amounts of data and resources. Nowadays, it is increasingly more practical for clinicians to consult all kinds of information with a view to improve their interventions and their diagnostic efforts. Such expanding availability enabled a wider transversality concerning the kinds of data that a professional can have at his hands. The opportunity to benefit from the multiple valences and multiple knowledge sources linked with such dynamic is crucial in seeking the progress and optimization of health-related services. Additionally, the requirement for trustworthy systems that can provide transparency and explanations for their actions/decisions makes it of the utmost relevance to address the issue of interpretability and/or explainability for medical AI apparatus with a multimodal perspective, that is, it is fundamental to rely on data from various modalities in order not only to increase knowledge extraction and but also to enhance the explanatory capacity of such frameworks. This last matter naturally involves expanding the variety of obtainable explanations in a supplementary way, i.e. the insight revealed by a visual map can and should be complemented by a brief written reasoning paragraph.

The main motivation behind the dissertation work is, essentially, to harness this multimodal awareness amongst medical data in order to enhance and inherently render more engaging AI

multimodal explanations. In this fashion, it is possible to help improve diagnostic support systems and their viable implementation in realistic clinical scenarios.

1.3 Goals

The goal of this dissertation is to increase diversity and complementarity in explanations for Artificial Intelligence systems, more specifically in AI-related medical diagnostic support systems in radiology. The actual intent is to develop strategies for generating visual saliency map-type explanations and natural language textual explanations that can complement each other in an AI-related radiology classification diagnostic support context. In this manner, the complete system can offer not only a diagnostic prediction but also two kinds of explanations that help to justify its functioning as a whole and thus increase the likelihood of acceptance and trust by the stakeholders - clinicians and radiologists - regarding such tools.

The overall approaches should use chest X-rays and associated radiology reports in order to optimize the extraction of information and knowledge from the available resources and therefore be able to contribute to the quality improving of explanations.

1.4 Contributions

The dissertation's major contributions are disclosed within the ambit of the distinct experiments. The initial unimodal to unimodal trials mainly enabled the consideration and comparison between several saliency map-producing algorithms.

As a primary contribution, the assignment concerning generation of visual explanations from multimodal inputs introduced a textual-driven regularization mechanism that demonstrated to robust and improve regular image classification tasks by bettering metric performances, and also by ensuring more concise, consistent, less dispersed and more clinically correct saliency maps. Such contribute led to the elaboration of a paper [9] - "*Increased Robustness in Chest X-Ray Classification Through Clinical Report-Driven Regularization*" - which was accepted and presented at the 10th Iberian Conference on Pattern Recognition and Image Analysis (IbPria 2022) in May of the present year.

The progression towards the generation of textual explanations has mostly provided solid methodologies for producing those textual explanations from X-ray images, automatically, as well as it handles novelty regularization adjustments that conduct to apparent slight enhancements on the clinical relevance front.

1.5 Document's Structure

The structure of the document comprises the following formulation: chapters 2, 3 and 4 are concerned to literature review and core subjects central to the theoretical framing of the dissertation;

chapters 5 to 7 entail the experimental work; chapter 8 finalizes the document by disclosing several final comments.

In a more detailed overview, chapter 2 presents some basic Deep Learning principles along with discussing their role in the fields of computer vision and natural language processing. Chapter 3 gives a survey-like overview of some of the current literature on multimodal Deep Learning, focusing on the various existing methodologies and implementations. It ends with detailed case-based demonstrations of real multimodal approaches in healthcare and medicine. Chapter 4 is a literature review on interpretability and explainability in machine learning. It covers taxonomy, and different ways of making systems more interpretable or explicable, as well as discerns current methods for generating explanations.

The experimental work pertaining to the dissertation is segmented into three parts that roughly correspond to a chronological sequence of the project's development as a whole, as illustrated by Figure 1.1

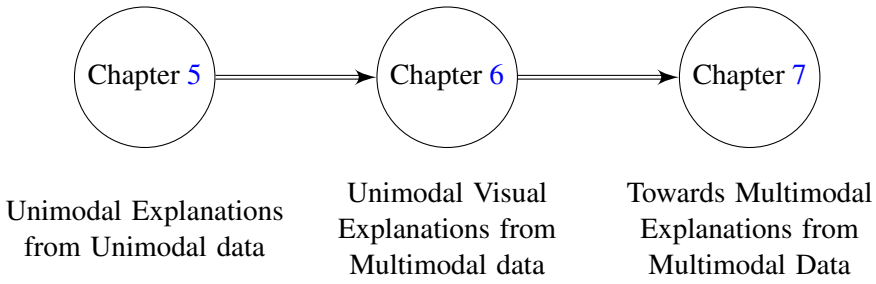


Figure 1.1: Experimental work pipeline.

In the first set of experiments (chapter 5), the purpose was not only to study and explore different post-model explainability techniques in order to produce saliency maps, but also to acquaintance with the particularities of each image and text modalities. On that note, the plan was to probe unimodal explanations for each modality via the execution of separate binary classification tasks, and so reach unimodal explanations from unimodal data.

The accounted work in chapter 6 aimed to leverage a multimodal learning paradigm using chest X-rays and their associated medical reports in a binary classification context, in such a manner as to consider possible improvements of quality for visual explanations. Accordingly, we reach unimodal visual explanations from multimodal data.

Chapter 7 reports on the generation of textual explanations by conditioning on the aforementioned dual modality data. Such a development becomes fundamental for the deployment of multimodal explanations from multimodal data.

Chapter 2

Background: Deep Learning

Since the introduction of the world wide web, the study of data has been widely increasing thanks to its growing availability and quantity, and for this reason, it has fully contributed to the development of information-related technology. Areas linked to the analysis of data such as Data Mining have seen an exponential evolution until modern days and, therefore, are prominent both in terms of research and industry. Work involving the applicability of data processing to real-life solutions is served, to a large extent, by the field of Machine Learning which is embodied by a brand of algorithms that are able to learn patterns in sets of data without the need to be explicitly hardcoded. This fundamental characteristic leads to an adaptive and modeling behavior, which is extremely convenient concerning the objectives set by these methodologies.

A basic procedure across all involved practices is to split the set into two components, one for training the algorithm and one for testing it. The logical sequence of these phases implies that, firstly, it is needed to train the system, offering it data with which it infers the reading of patterns. Then to test, a group of independent and identically distributed data, never provided before to the model, is made available, to evaluate, with defined metrics, how the latter has learned the representations of the patterns in vogue. These metrics vary depending on the objective of the work being promoted. As an example, in the context of a classification problem, where it is intended for a category to be defined for a given input, a coherent metric would be to compare its actual label with that predicted by the system in use.

The current ML panorama is dominated by the term Deep Learning. DL is a subspace within ML that unlike typical machine learning approaches, which need feature extraction to be done independently of the learning process, can pick up what features to obtain given a specific task. This integrated configuration more often translates into superior results and performance across a wide range of domains, hence its recent popularity [10, 11].

The fundamental proposition that lies at the heart of Deep Learning is the concept of Artificial Neural Network (ANN). An ANN is a Machine Learning model that attempts to replicate the basic communication structure of our human brain. It is composed of units that mimic neurons and by connections between them that are conceived to emulate synapses and the intricacies of neural interactions. Combining these building blocks the most basic form of a neural network, the

Perceptron, can be created, as described by Figure 2.1. In it, three inputs (x1, x2, x3) are fed into the neuron N that linearly combines them with its respective weights (w1, w2, w3) while adding a constant value, bias. That end product is then mapped into an output y via an activation function, h.

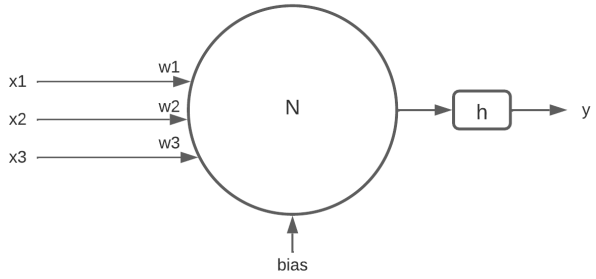


Figure 2.1: Perceptron.

Figure 2.1 can also be expressed mathematically via Equation 2.1, and so there is a base formulation to explore. In an ANN context where the primary structure is replicated and surrounded by higher architectural complexity, such as in biological conditions, the output is passed through the aforementioned activation function, also h in Equation 2.1, which re-scales it conforming to the function's body. This operation primarily exists to assist the network in learning intricate patterns, and so it helps in the training of the concerned models. It is also significant to state that the activation function has the capacity to inject non-linearity into the system, providing the ability for a neural network to convey complexity non-linearity.

$$y = h(\sum_{i=0}^{n} x_i \times w_i + b) \tag{2.1}$$

Examples of regularly used activation functions are as follows [12]:

Sigmoid: This function returns a value ranging from 0 to 1, which can be mapped as a probability. Since its derivative is never greater than one, an issue known as vanishing gradients is frequently caused during network training - the gradients grow so tiny that they are almost zero, preventing the unit from optimizing the values of its weights. Due to that, it is usually only used as the final output activation function.

$$f(x) = \frac{1}{1+e^{-x}} \tag{2.2}$$

Hyperbolic Tangent (Tanh): It resembles the sigmoid function. However, it produces a value between -1 and 1, and it is centered around zero, resulting in larger gradients within the [0, 1] range. The vanishing gradient is also revealed as an existing problem.

$$f(x) = \frac{2}{1+e^{-2x}} - 1 \tag{2.3}$$

Softmax: The function is also used to make a probability distribution out of a set of real values and the assigned output ranges between 0 and 1. It simply differs from the Sigmoid at context level, since it is the latter's application in a multivariate setting.

$$f(x) = \frac{e^x}{\sum_j e^{x_j}} \tag{2.4}$$

Rectified Linear Unit (ReLU): For this function, negative inputs translate to null gradients. In opposition, positive inputs get unitary gradients. The non-activation of units with negative values can cause the known issue, dead ReLU.

$$f(x) = \begin{Bmatrix} x, & if\ x > 0 \\ 0, & if\ x \leq 0 \end{Bmatrix} \tag{2.5}$$

Leaky Rectified Linear Unit (Leaky ReLU): The creation of this function emerged to remedy the dead ReLU issue. In this case, for negative inputs, a linear function with a slight slope is employed, enabling minor negative gradients to modify the unit's weights.

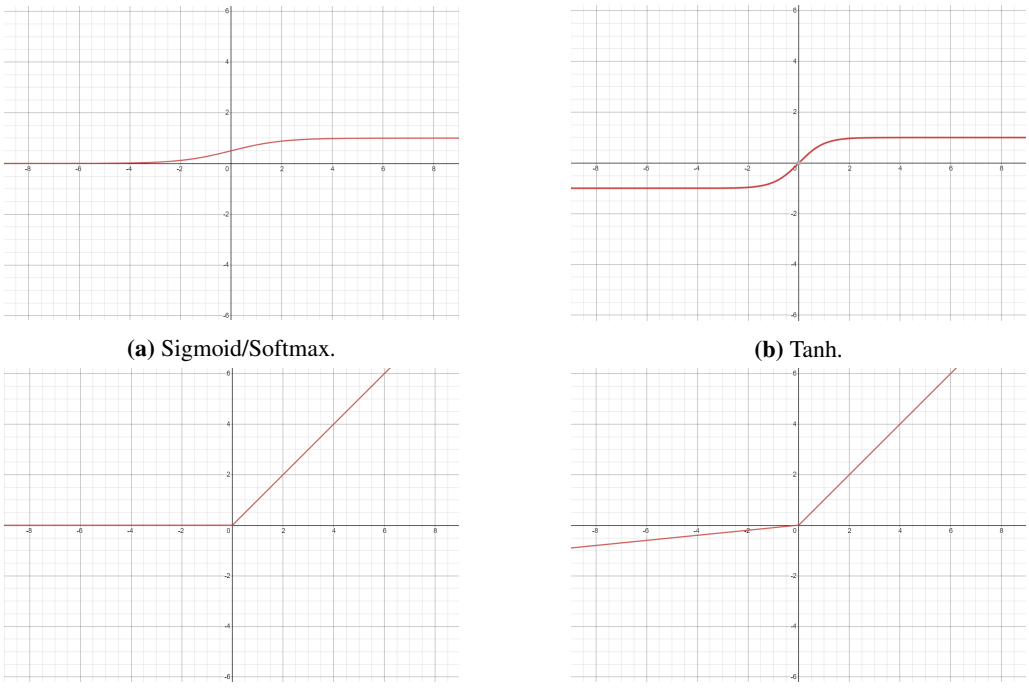
$$f(x) = \begin{Bmatrix} x, & if\ x > 0 \\ ax, & if\ x \leq 0 \end{Bmatrix} \tag{2.6}$$

Graphic representations of aforementioned functions are prompted in Figure 2.2

For what can be called DL purposes, and as hinted before, deep neural networks require that multiple perceptrons organize themselves in a consecutive and transverse manner, in such a fashion that the whole constructed architecture can be defined by various layers. Thus, the handling of layering and network design depends on the problem to be tackled and on the subsequent experimentation.

Reflecting on the adaptive capacity of these models, it is by updating its internal settings, that a feed-forward neural network alters and improves itself. This translates into a change in the values of the weights (w in Figure 2.1) of the network in question. The standard way to guarantee such an update is achievable by what is called the backpropagation algorithm. This technique is divided into two phases: the forward phase, in which the network propagates signals from the input layer to the output layer in a forward direction, and the backward phase, in which the network propagates gradients that will be used to update the weights of the units in a backward fashion.

Regarding the forward pass, each node in an ANN gets its input from the preceding layer, which is a weighted sum of the weights at each of the connections multiplied by the output of the



(a) Sigmoid/Softmax.

(b) Tanh.

(c) ReLU.

(d) Leaky ReLU.

Figure 2.2: Activation Function Graphs.

previous layer. This weighted total is supplied to an activation function, and the result is the output for a specific node, which is subsequently passed as part of the input for the nodes in the following layer. This occurs for each layer in the network until we reach the output.

An essential aspect pertaining to the whole training stage is the procedure of studying the difference between what the model anticipated for a given input and what the provided input is in reality. This analysis is quantified by the computation of the loss value which is no more than a measurement of this disparity. There exist a variety of diverse functions that perform the computation, and so the choice of its use will also vary with the overall scope of the problem at hand. Janocha et al. [13] reflect and contemplate on some of the most familiar loss functions. Said calculation is concluded for every sample in a training set. When the end of the training set is reached, an "epoch" has been finalized and the abovementioned difference measure is attained via the loss function. Then, in the backward phase, the loss is minimized using an optimization algorithm. The gradients from the loss with respect to each of the weights in the model are computed. These gradients permit the optimizer to reach new weights and update them.

The mentioned optimization algorithm is a pre-defined algorithm or formulation that attempts to find a loss function minimum. Although there exist multiple optimization approaches, as explored in [14], the standard procedure from where a lot of the modern optimizers originated is the stochastic gradient descent.

2.1 Computer Vision

Computer vision (CV) is an Artificial Intelligence branch that aims at developing and deploying digital systems that can process, assess, and interpret visual inputs (image and video). There are plenty of tasks and sub-domains that fall under the CV category, from image classification to object detection, video tracking, and others.

Throughout recent years, deep learning and deep neural networks have been essential for the growth of the field. On that matter, the next subsections delve into some of the most prominent types of models in CV.

2.1.1 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are still state-of-the-art DNNs for assignments that involve working with multidimensional data, including images. Thus, they are a widely used tool in computer vision, especially when discussing computer vision through a lens of deep learning strategies. They are rooted in two types of particular operations: convolution and pooling. The alternate use of both operations is combined in various ways, depending on the conformed architecture, but typically, the goal is to identify and extract patterns or features at the layer level. These may be reckoned as simple features, at the beginning - recognizable edges or lines in an image - but as a network broadens, there is an increasing integration and combination of calculations on such elements, which are then decoded to represent more complex patterns and properties within the analyzed inputs. Between the mentioned operations there is always an activation function that alters the convolution end product for pooling. Convolutional Neural Networks employ the conventional and transversal DNN methodologies for training instances, and therefore resort to optimizers for minimizing the associated loss function(s) and to backpropagation for updating its weights.

Over the years, since the relatively recent growth of deep learning, various architectures have been developed with this mindset of improving and enhancing this functionality of information extraction and conservation. According to the current status quo, these architectures are usually introduced to the AI community and so, they can be widely used by the professionals that work in the area. Networks like VGG [15], DenseNet [16], ResNet [17], or U-nets [18] are part of such catalogue and are naturally found in the most distinctive research papers across the CV realm. Moreover, it is relatively simple for any participant in the field to have access to particular model designs, that are pre-trained in several types of visual domains, which makes a convergence in training increasingly attainable, optimal, and robust.

Figure 2.3 serves to show how a simple sequential structure of a CNN would be organized. Even though it details the number of parameters to optimize for such projection and respective magnitudes, the important portion attains the arrangement. In a complementary manner, the following subsubsections serve to dive into more detailed descriptions of the concrete operations and layers connected to CNNs.

```
----------------------------------------------------------------
        Layer (type)               Output Shape         Param #
================================================================
            Conv2d-1         [-1, 32, 222, 222]             896
              ReLU-2         [-1, 32, 222, 222]               0
         MaxPool2d-3         [-1, 32, 111, 111]               0
            Conv2d-4         [-1, 32, 109, 109]           9,248
              ReLU-5         [-1, 32, 109, 109]               0
         MaxPool2d-6           [-1, 32, 54, 54]               0
            Conv2d-7           [-1, 64, 52, 52]          18,496
              ReLU-8           [-1, 64, 52, 52]               0
         MaxPool2d-9           [-1, 64, 26, 26]               0
           Flatten-10                [-1, 43264]               0
            Linear-11                   [-1, 64]       2,768,960
              ReLU-12                   [-1, 64]               0
           Dropout-13                   [-1, 64]               0
            Linear-14                    [-1, 2]             130
================================================================
Total params: 2,797,730
Trainable params: 2,797,730
Non-trainable params: 0
----------------------------------------------------------------
Input size (MB): 0.57
Forward/backward pass size (MB): 36.89
Params size (MB): 10.67
Estimated Total Size (MB): 48.13
----------------------------------------------------------------
```

Figure 2.3: Simple CNN structure.

2.1.1.1 Fully connected layer

Fully connected (FC) layers or dense layers are the most basic or fundamental layer of any artificial neuronal network, and consequently, any DNN. It connects the neurons of one layer with those of the previous one and represents the combination of all perceptrons that materialize at the same level. Figure 2.4 illustrates a simple neural network with two fully connected layers, and so there is a demonstration of the inherent essential relationships/activities.

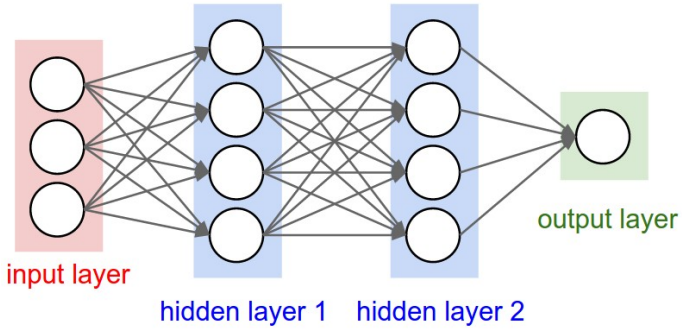
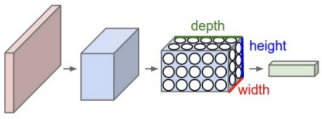


Figure 2.4: Simple ANN with two fully connected layers. Extracted from https://cs231n.github.io/convolutional-networks/ (Accessed on 01-01-2022).

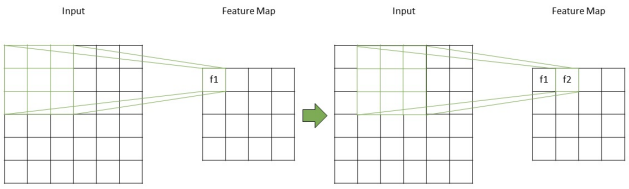
In a CNN setup, these layers are normally applied at the end for the purpose of translating the last defined representations into a decision or classification. Nevertheless, they can also be used in other segments to reduce or increase dimensionality.

2.1.1.2 Convolutional layer

The convolutional layer is deemed as the building block of a CNN. Rather than handling an image at an individual pixel stage, as a regular DNN would, the first convolutional layer implies an area-driven technique, that allows for parameter sharing, therefore avoiding a substantial amount of gradient assessment, and thereupon computations. Such mechanics are reproduced across the many convolutional layers, resulting in a hierarchical structure that concentrates on learning low-level features in early phases and that withdraws higher-level ones as the layers add up. The applied transformation - convolution - resolves in the creation of feature maps that are aggregated on top of each other, throughout the natural sequence of the model and rendered in Figure 2.5a. In these, the main actors are the filters or kernels. Commonly smaller than the input image, they behave as sliding windows that glide over the input object, computing the dot product between its values and the corresponding overlapping image pixels. Such procedure is shown in Figure 2.5b, where also it is possible to notice that the results of each sliding calculation f1 and f2, are annexed in order to form the so-called feature map. Figure 2.5b also provides further information as it implies an iteration where a 3x3 filter moves with a stride of 1 across a 6x6 representation, that can either be the initial treated image, or, in a more advanced step of the CNN, a previous feature map.



(a) Typical CNN architecture. Extracted from https://cs231n.github.io/convolutional-networks/ (Accessed on 01-01-2022).



(b) Convolution operation

Figure 2.5: CNN and convolutional layer.

So far an instance of downsampling has been advocated, but the truth is that there are ways to exert the convolution as an upsampling resolution, or even maintain the intake's size. Resorting to the preceding padding of the latter it is practicable to achieve such effects.

2.1.1.3 Pooling layer

A pooling layer operation happens after the convolution process and is used as a downsampling function since it transmutes the information of the entity on which it exercises to a compressed representation with a smaller dimension. By, once again, manipulating a sliding window dynamically, an operation (often an average or maximum) is applied to the set of values that match the window position. Looking at the example of Figure 2.6, we have a 2x2 window slithering over a 4x4 input, with a step (stride) equal to 2, producing a 2x2 compact matrix.

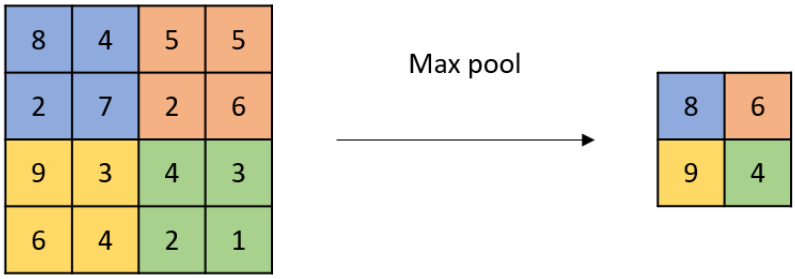


Figure 2.6: Max pooling operation.

2.1.2 Other Neural Networks

Numerous other neural network conformations can be included and employed for computer vision applications. Examples of such are capsule networks [19], vision transformers [20], or generative models [21]. For the sake of being more concise and contained in the description of existing ANNs for computer vision, and denoting that the Transformer is covered in the NLP section, only generative models will be lightly further detailed, since they probably allude to the most research work, in the field, after CNNs.

2.1.2.1 Generative Models

Generative models, as the name implies, generate samples from some determined inputs. This function is, generally, achieved by learning probabilistic data distributions, which according to the goal of these frameworks, can lead to the generation of outputs that, ideally, fall within the spectrum of such distribution. These models use recognized/learned information patterns to produce new instances.

Generative approaches are developed around the concept of maximum likelihood estimation (MLE). Simply put, maximum likelihood estimation is the method of defining model parameters that maximize the training data's likelihood. It can be formulated by virtue of Equation 2.7, where θ is the model's parameters and x is the data itself.

$$\hat{\Theta}_{ML} = argmax \prod_{i=1}^{m} p_x(x_1|\Theta) \tag{2.7}$$

The way that a system learns the data distribution for applying such MLE vision is what differentiates them. Consequently, it is also what helps to build a generative model taxonomy. Goodfellow et al. [21] delineate a pretty robust taxonomy by dividing approaches into explicit and implicit density models. Explicit density models define themselves as systems that can explicitly describe the data distribution and, consequently can maximize the likelihood estimation directly.

On the opposite side, implicit density models are apparatus that can extract samples from a probability distribution without explicitly reporting it. Table 2.1 displays such taxonomy, as it also entails a brief description of how these types of models model probability.

Table 2.1: Overview of Deep Generative Models.

Model	Taxonomy	Description
Generative Adversarial Model (GAN)	Implicit	The probability distribution of the created data converges to the real data distribution as the generator and discriminator elements play an adversarial minimax game.
Variational Autoencoder	Explicit	The encoder-decoder design of the model intends to maximize the likelihood of data. The encoder transforms an initial input, x into an abstract latent space z by modeling p(z|x), and the decoder approximates an output y via maximizing p(y|z), reconstructing that z latent representation.
Normalising Flow	Explicit	Calculates the data's joint distribution by multiplying the conditional distributions of each data dimension space.
Autoregressive Model	Explicit	Converts a straightforward data distribution into a more sophisticated one by means of a sequence of transformations.

The most notable generative systems within the DL spectrum, in recent times, are, most likely, Generative Adversarial Networks (GANs). GANs are a type of deep generative neural network that uses adversarial learning as the main instrument to learn data behaviour [21]. Its generative nature gives an important capacity as an unsupervised learning method. In terms of structure, a GAN is made up of two sub-networks, a generator, and a discriminator, the actors in the adversarial interaction that fuels the learning process. The generator is responsible for creating artificial samples from random noise, and the discriminator compares real data with the outputs of the generator and affects a decision on whether the last belongs to a real data class or is a falsification.

The adversarial aspect reveals itself when analyzing Equation 2.8 and its framing as the standard loss calculation for such implementation.

$$\min_G \max_D V(D,G) = E_{x\sim p_{data}(x)}[\log D(x)] + E_{z\sim p_z(z)}[\log(1-D(G(z)))] \tag{2.8}$$

Considering that G stands for generator, z for noise, D for discriminator, and that p_z and p_x imply the generated samples distribution and the real data distribution respectively, it is perceivable that both entities are opposing each other in a min-max game. The generator attempts to minimize the loss function V and consequently tries to fool the discriminator into taking the input constructed from z as real instances, whereas the discriminator works towards maximizing the disparity between its inputs.

2.2 Natural Language Processing

Natural Language Processing (NLP) is an extremely broad topic that can be approached in a number of ways, taking into account the context and task at hand. Either way, the term can be defined

as a field of research that focuses on investigating how computers can work and manipulate text in such a way that they can replicate human-like capabilities in a variety of applications [22–24].

NLP encompasses two often dissociated terms: Natural Language Understanding (NLU) and Natural Language Generation (NLG) [25]. NLU is generally related to the syntactic and semantic analysis of text and speech so as to establish the meaning of a sentence [26]. On the other side of the spectrum, NLG builds on semantics as a means of generating human-readable and understandable text [27].

Table 2.2 lists and describes a number of common tasks related to the NLP field of study.

Table 2.2: Common NLP tasks.

Task	Description
Text Classification	To label a text sequence input [28].
Language Modeling	The technique of implementing a model to forecast words or simple linguistic components based on prior elements [29].
Summarization	A method for extracting a brief and understandable summary of text from an array of sources [30].
Question and Answering	A task which is to find a concise and precise responses to an user's input via enabling the learning of representations of the interaction between questions and texts [31].
Machine Translation	The exercise of automatically translating content from one language to another [32].
Sentiment Analysis	Refers to systems that extract and analyse emotional states and subjective information from a text object [33].
Named Entity Recognition (NER)	Responsible for recognizing and classifying important information (entities) in text as in excerpting piece of a category from which each word in the sequence belongs to [34].
Image Captioning	Aims at generating a caption for an image input [35].

Arguably, the most commonly adopted systems in NLP studies and experiments are RNNs (Recurrent Neural Networks), LSTMs (Long short-term memory), GRUs (Gated Recurrent Units), and more recently Transformer based models. These solutions are further discussed in the subsequent sections.

2.2.1 Recurrent Neural Networks

Recurrent Neural Networks or RNNs are deep neural networks designed to deal with sequential data, i.e time associated information, video, and natural text. These models must have an abstract concept of sequential memory, which is achievable because an RNN has a looping mechanism that acts as an unfolding that grants information to follow a flow across certain steps or time steps, according to its sequential input configuration [36]. To make sense of this, it is useful to observe Figure 2.7. The textual inputs are consecutively fed into the fundamental units of these systems and converted to hidden states, h_i, at distinct time steps. Given the first one, the RNN encodes the word "*How*" and generates an output O_1. At time step two, the input "*time*" and the hidden state from the previous stage are supplied to h_2, which, in this way, encompasses information on both

terms. The apparatus is replicated throughout the remaining steps. It appears that, by the final one, the RNN has encoded information from all the words.

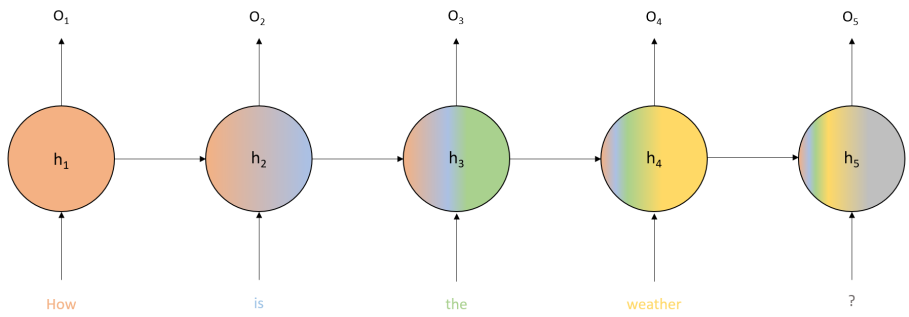


Figure 2.7: RNN basic structure.

For a classification intent, only the final output O_5 would be utilized, since it is the one that results from the series' aggregated knowledge [36].

Figure 2.7 also addresses a key aspect related to such DNNs. The color gradient is meant to express a known issue with RNN's - short-term memory. Short-term memory is reflected in the difficulty of the network to contain information over a relatively long sequence. As is noteworthy, the representation of the word "How" in the last hidden state is already highly reduced. With a few more iterations, it would completely disappear. Such phenomena are caused by the aforementioned vanishing gradients problematic that is linked to the backpropagation phase during training.

On a final look, Figure 2.8 unravels what happens inside an RNN cell. The hidden state coming from the previous cell is concatenated with the input of the current cell and the resultant vector is passed through an activation function (traditionally tanh) [36]. Exiting the cell are already discussed entities.

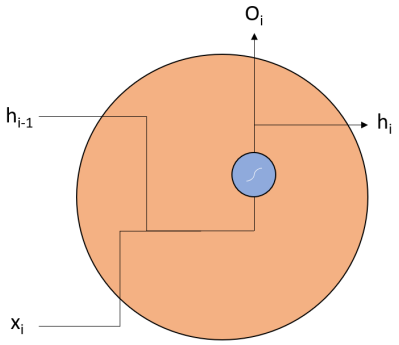
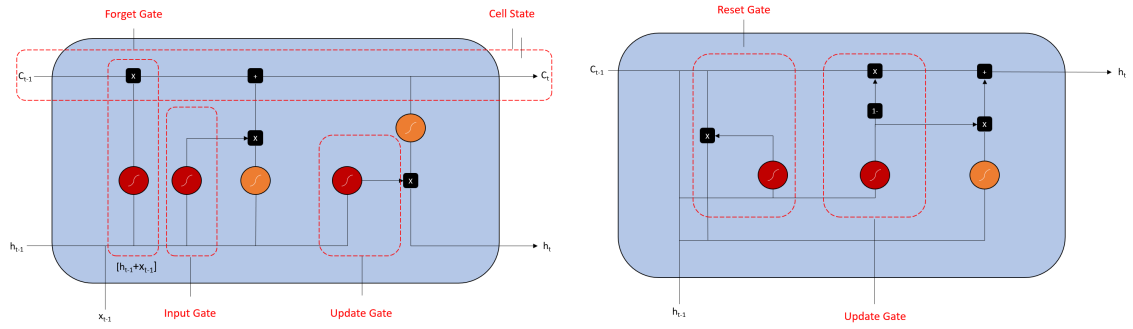


Figure 2.8: RNN cell.

2.2.2 LSTM and GRU

LSTM (Long short-term memory) and GRUs (Gated Recurrent Units) were developed to try to combat short-term memory. They feature inbuilt devices known as gates that may control the flow of data. The control flows of an LSTM and a GRU are comparable to that of an RNN. They process data and transmit information as the signal moves along. The difference resides in the internal structures of LSTM and GRU cells, which can be quite more complex. The internal contrivances function in a gate-like manner.

Figure 2.9 introduces both types of cells in fair detail. The red and orange circles respectively represent sigmoid and tanh activation functions, and the "X" and "+" blocks express pointwise multiplication and addition, correspondingly. The mirrored gates implement the information flow in agreement with the logical reasoning behind these systems. Their high-level rationale works towards retaining only relevant data while discarding the rest [37]. On that note, the gates need to figure out which entries in an input sequence should be kept and which should be dismissed, across the various internal routes.



(a) LSTM unit cell. **(b)** GRU unit cell.

Figure 2.9: LSTM and GRU unit cells.

Focusing on the LSTM framework (Figure 2.9a), there are two main agents: the cell state (c_t) and the gates. The cell state serves as a transit route for meaningful information as it goes deeper into the circuit. While such trail is traversed, information is added or withdrawn from the cell state through gate action [37]. Performing as the forget gate, the sigmoid function regulates the activity of the concatenation of knowledge from the past hidden state h_{t-1} and from the current input h_t. The applied activation can go from 0 to 1, whereby the closer to 0, the bigger the "forgetting" factor, and the closer to 1, the more the network "remembers". The next phase regards the input gate that replicates its adjacent gate's operation as it also forces the same representation to pass through the tanh activation function, so as to multiply both outcomes. This stage is stated to induce some regularization [37]. On the output gate, there's a multiplication of the activated cell state and a sigmoid adjustment. The result follows as the representation of this' cell hidden state (output for the final unit), whereas the cell state is also transferred to the next building block [37].

The GRU arrangement, outlined in Figure 2.9b, was built from the LSTM framing. In this case cell states are abandoned in favor of only using the concealed state to convey the information

stream. Two gates regulate the implied processing: the reset gate, and the update gate [38]. The update gate forges a reproduction of an LSTM's forget and input gates [38]. It determines what information should be discarded and what should be included for circulation [38]. The reset gate intuition centers on how much previous knowledge to select or forget for the latter final step. Here, a new "1-" operation term is established. The definitive differentiate factor for GRUs is the fact that only exist a hidden state as a yielded product.

2.2.3 Transformer Based Models

Since the publication of [39], transformer-based models have dominated the state-of-the-art across nearly all text-related tasks. Their ability to establish relationships between far apart words and retain knowledge about context is unmatched by all other models used until then. It is, thus, much more capable of analyzing larger sequences than, for example, LSTMS or GRUs.

The transformer's framework is detailed by Figure 2.10. Firstly, at a high-level observation, two main components are to be noted: the encoder and the decoder. Both are actually rendering a variable size ("Nx") stack of encoders and decoders, respectively. The encoder blocks are structurally equivalent to each other and are composed of three types of sub-layers: multi-head self-attention layer, feed-forward or fully connected layer, and normalization layer.

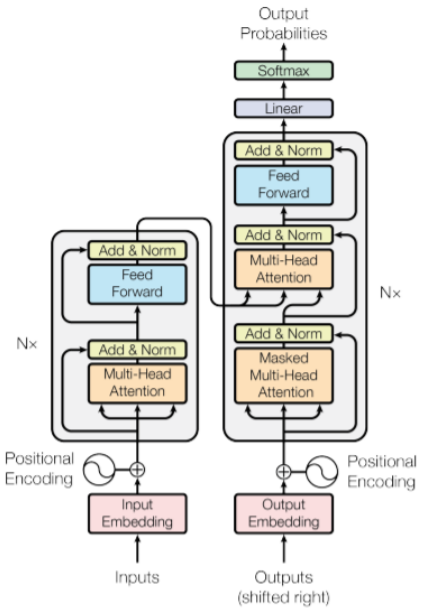


Figure 2.10: The Transformer model architecture from [39].

The self-attention mechanism, in a general and superficial overview, allows the model to glance at various positions in a sequence, while encoding each word, in such a way that is able to pick up clues that help optimize that same encoding. In other words, it is capable of identifying other relevant words to the one that is being treated at said moment. On a deeper perception, such functionality is achieved thanks to the ensuing successive steps:

- Primarily and discerning on Figure 2.11, the encoder's input embeddings are individually mapped into a query vector, a key vector, and a value vector via multiplying them with weight matrices.

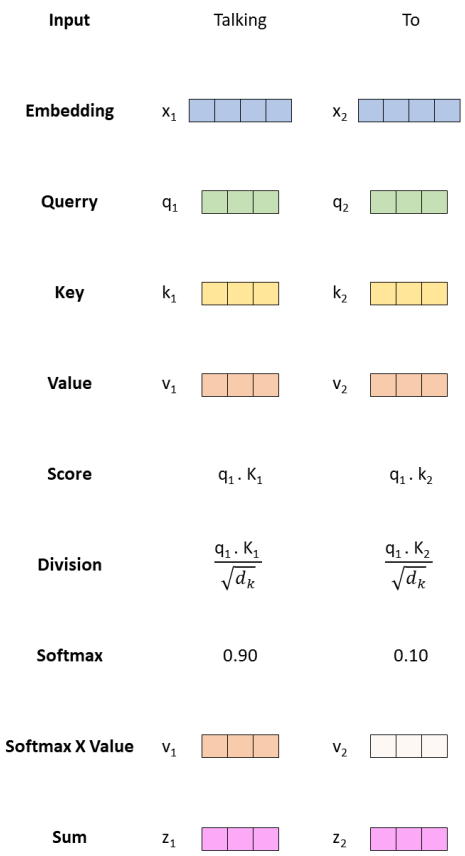


Figure 2.11: Representation of a self attention procedure.

- Secondly, a score is calculated for each instance for the sake of defining a magnitude for the attention allocation. It is computed by exercising the dot product of the query vector with the key vector. Granted that effort, a score metric per word relationship is created.

- The third and fourth stages revolve around dividing the scores by the square root of the key vectors' size, which according to [39] leads to more stable gradients, and normalizing those values with the help of a softmax operation.

- On the fifth step, each value vector is multiplied with the matching "*softmaxed*" score, provided the idea of preserving the values of the word(s) we wish to focus on while drowning out unnecessary ones.

- The sixth and final phase consists of adding up the weighted value vectors. The consequent vector is considered the output of the self-attention regime. As is, it ensues its way to the feed-forward network constituent.

In a multi-headed panorama, this whole process is scaled up to a matrix configuration, where the dimensions are elevated in consonance with the number of heads.

The previous process leads to the example illustrated in Figure 2.12, where it is possible to track the inter-word associations and their significance, for a certain input sentence.

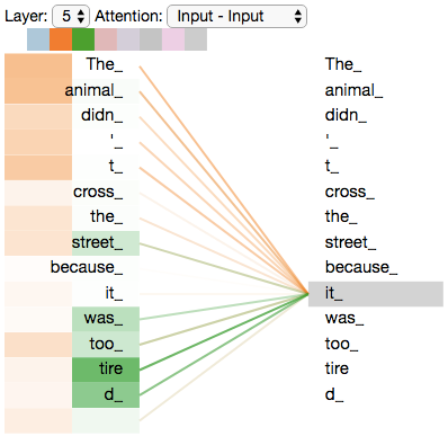


Figure 2.12: Word relationship as computed by a self attention framework. Extracted from https://jalammar.github.io/illustrated-transformer/ (Accessed on 04-01-2022).

Another important transformer contrivance to denote is the positional encoding strategy. The positional encoding is a key instrument for supplying a tactic of accounting for word order in the input sample. It comprises of adding vectors that reflect a learnable particular pattern to each input representation so as to assist in resolving every word relative location in the sequence. Such practice is supposed to deliver meaningful distance concepts between the embeddings whenever they are projected into the query, key, value tensors, and throughout the entirety of the self-attention procedure [39].

The encoder-decoder interaction is produced by the passage of the top encoder's key and value representations to the decoder second attention layer. The vectors are seemingly meant to aid the decoder in converging to the essential position in the output sequence. It is this first communication that generates the first output of the decoder block, which is then used to generate the subsequent outcome. Such process is monitored until the end of the exercise and is cataloged in Figure 2.10 as "*Output (shifted right)*".

The decoder's masked self-attention layer differs slightly from the encoder one. Here, only the outputs already revealed to the system, at the given time step, are employed, whilst the remaining sequence positions are masked. The ensuing multi-head attention arrangement operates just like its encoder counterpart, despite benefiting from the encoder's value and key tensors. In this manner, it only creates a queries matrix.

The final linear and softmax layers are in charge of outlining the decoder stack production into a probability-defined output vector that announces each learned word. Taking a fictitious example, if a trained transformer model is up to learning a 10000 words vocabulary, its final result

will be a 10000 sized unidimensional tensor that describes the vocabulary probabilities of being the predicted word at said time stage.

For a final remark on the transformer analysis, the residual connections that follow each sub-layer and the normalization ones are not to be overlooked, since they are argued to induce greater training stability and regularization [39].

Nowadays, the two most acclaimed and proven families of Transformer-based models used in state-of-the-art experiments are BERT and GPT. BERT stands for bidirectional encoder representations from Transformers and it bases itself on a stacked batch of Transformer encoder structures [40]. BERT models are usually pre-trained with various methodologies with different tasks in mind. Nevertheless, there are two pre-training approaches specific to the base BERT. The bidirectional paradigm stems from a "masked language modeling" pre-training task where 15% of the input tokens are randomly masked, and the model by having access to the remaining tokens, from left to right and from right to left, predicts the masked inputs. Additionally, the pre-training process also applies to what is called a two-sentence task where given two sentences, BERT predicts the likelihood of one of the sentences belonging after the other and vice versa.

On the other hand, GPT (generative pre-training) language models are essentially built on stacked transformer decoder blocks [41] and are generally conditioned on several large-sized corpus and vast amounts of data.

2.3 Conclusion

After a broad overview, it is noticeable that there is a myriad of typologies and models present in both computer vision activities and NLP ventures. Nevertheless, there exists also, inherently, a more selective group of approaches, which due to their greater performances and/or reproducible deployment, are more widely used and make up what is often called the state of the art. For computer vision, CNN-based models continue to be at the forefront of the work being accomplished. For example, in classification tasks, ResNet and DenseNet frameworks are widely used as benchmarks or baselines, whether, for more specific assignments there are other architectures which are adapted to each reality and circumstance, such as R-CNN [42] and YOLO [43] for object detection or U-Net for segmentation exercises. Even for prevailing generative efforts, namely StyleGAN [44] and VQGAN [45], the convolutional paradigm remains at the core of the implementations.

Regarding NLP, the state-of-the-art techniques and systems may vary from task to task. Notwithstanding, nowadays across most, Transformer-like algorithms (in an encoder-only setup, in a decoder-only arrangement, or applying both sub-structures of the Transformer) are the most common instances to address challenges. Considering classification, sentiment analysis, or NER exercises, BERT-like models as, certainly, BERT, DistilBERT, or RoBERTa [46] are very popular, whilst for generative readings or machine translation, GPT-like models predominate. On account of Language modeling, question and answering, and summarization problems often an encoder-decoder guideline is involved.

Chapter 3

Literature Review: Multimodal Learning

The concept of modality invokes a group with certain characteristics. When referring to data, it addresses certain conditions that define the data type in a specific way. Information may be textual, as may be an image, or an audio signal, each representing a different modality. Now, transferring this concept to the ML and/or DL universe, the focus is precisely on the modality of the data to be treated and handled. In this sense, a multimodal approach implies solutions integrating data of different domains.

Multimodal Deep Learning has evolved more and more until the present day and thus contemplates more and more manners and configurations to manage and link data from several modalities. Tasks such as multimodal emotion recognition [47], video and image captioning [35, 48], Audio-visual speech recognition [49], multimedia retrieval [50] are examples of current multimodal Deep Learning advancements and tendencies.

3.1 Multimodal Representation

Data representation is an essential exercise within any ML/DL setup. In fact, it is essential to make the most of it, so that the implemented algorithms can draw the greatest amount of knowledge possible in the most effective way, in contemplation of a proposed challenge. The terms feature and representation are used interchangeably and abundantly, taking into account the relevance of this more abstract perception. In a practical fashion, the attribution is made to a vector or tensor representation of an item, regardless of origin and form. A multimodal representation is one that uses information from different nature and combines them, mathematically, in pursuance of settling on an aggregate and feasible reproduction. This effort might not be always trivial, since there are numerous mishaps across the spectrum in question, such as the diversity of noise perpetuated by various sources, the choice of the aggregation function itself, dealing with the incumbent loss of meaning inherent to the initial inputs, and others [51]. Anyhow, the capacity to describe data in a meaningful way is absolutely critical.

According to Bengio et al. [52] good representations are usually associated with traits such as temporal and spatial coherence, smoothness, sparsity, and natural grouping. No less important, and possibly more generally intuitive is the assertion that similarity in the representation space should reflect the similarity of the corresponding concepts, as depicted by Srivastava et al. [53]. Moreover, such work additionally identifies properties like: the accessibility of obtaining representations even if some modalities are missing, and the possibility to fill in missing modalities given the observed ones.

In [51] there is an important distinction to report, which differentiates multimodal representations into two scopes: joint and coordinated. Joint representations encompass unimodal signals into a single representation space, whereas the coordinated manner analyzes unimodal tokens individually but applies similarity constraints in order to cluster them together in a so-called coordinated space. The work of Guo et al. [54] expands on the categorization and underlines encoder-decoder strategies should also embody a representation mechanism.

3.1.1 Joint Representation

Joint representations are typically utilized in scenarios where multimodal data is present during both the training and inference phases. They attempt to project separate unimodal representations into a shared semantic dimension, where the information is merged, in such a manner that the initial heterogeneity gap is extinct [54]. A simple concatenation of the separate modality features, also assigned as early fusion [55, 56] is the broadest, most direct proposition. As shown in Fig 3.1, each encoded modality is mapped onto a common latent space, where they coalesce into a single vector. Subsequent to the junction, there comes the pass responsible for effecting the blending operation, where the altered modality-specific vectors are combined. This attribution is reported in Equation 3.1, where y is the output, h typifies the activation function, x exemplifies the modality-specific encoding, w represents the respective weights, and the subscript indexes indicate the distinct categories. For demonstration purposes, practicable bias terms are ignored.

$$y = h(x_1 w_1 + ... + x_n w_n) \tag{3.1}$$

In DL, the natural disposition of DNNs enforces a tendency of using the ultimate or penultimate layers to perform such operation, as it is conjectured that each succeeding layer is thought to describe the data in an increasing abstraction fashion [57]. The combined tensor is then routed via n-hidden layers or directly employed on the task [58, 59]. It is also frequent to encounter unsupervised appeals that resort to such fusion after extracting latent individual reproductions, largely thanks to the use of autoencoders [60, 61].

The main benefits of neural network-based joint representations, as stated in [51] come from typically producing greater performances and its suitability in unsupervised scenarios. On another note, [54] points out the regularly simpler implementation when compared to other frameworks,

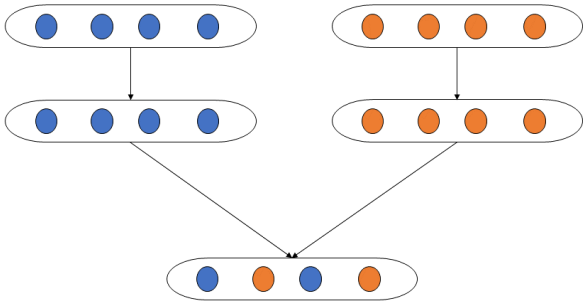


Figure 3.1: Illustration of a joint representation.

since there's no need for categorical coordination. For the authors of [62] the acquired modality-invariant characteristic of such layout conveys itself as an advantage because it implies a facilitated aid knowledge movement. In relation to drawbacks, the overall and comprehensive critique leans on the incapability, demonstrated by this procedure, to infer different representations for each modality.

In terms of employment, it is of notice to remark some requisitions of the discussed mechanism in event detection [63], emotion recognition [64], video classification [65], among others.

3.1.2 Coordinated Representation

A coordinated representation, such as manifested in Fig 3.2, is another form of multimodal learning. Instead of learning representations in a shared subspace, it reflects a coordinated system that learns independent but interconnected representations for each modality while adhering to certain constrains [51]. These restrictions address relational concepts between representations and are often divided into two classes: cross-modal similarity and cross-modal correlation. Cross-modal similarity-based methodologies seek to explore similarities by computing distances between inter-model latent subspaces [66], while cross-modal correlation strategies focus on correlation metrics across the same [67].

Cross-modal similarity algorithms try to maintain the inter-modality and intra-modality similarity formation, which renders the attempt of approximating diversely expressed data if they convey alike semantics. For example, the distance between the word "dog" and an image of a dog should be promoted to be less than the distance between the word "dog" and an image of a clothing item [68]. The inverse logic is clearly justified in cases where the aim is to maximize the distance among concepts. These procedures happen at the time of loss calculation, as it is intended to optimize the supposed relationships and for this end, the construction of a loss function is dependent on the desired constraints and results. Cross-modal ranking and derivative techniques are broadly

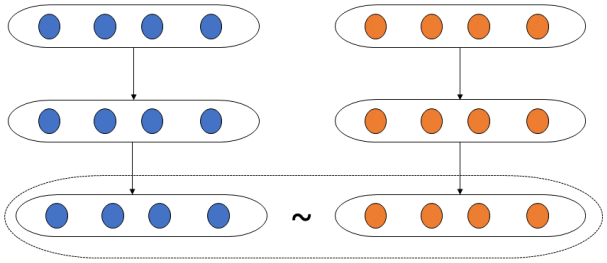


Figure 3.2: Illustration of a coordinated representation.

used [69–71]. The starting point of such methodologies usually points to the margin rank loss, presented in [72] and exemplified in Equation 3.2. In it, v and t ponder different modalities (visual and textual, for the case), α exemplifies a margin, S represents the similarity measurement, and the - subscript intends to identify the unmatched vector embeddings.

$$rankLoss = \sum_{v}\sum_{t^-} max(0, \alpha - S(v,t) + S(v,t^-)) + \sum_{t}\sum_{v^-} max(0, \alpha - S(t,v) + S(t,v^-)) \tag{3.2}$$

Other extensively utilized routines are established from the Euclidian distance idea. In this category, the standard technique relies on simply reducing the distance between paired samples [73, 74]. Granted the effort formulated in [74], it seeks to learn a visual-semantic embedding that could be utilized to produce video descriptions. Thereupon, the projection of both representations onto a low-dimensional space operates toward reducing the distance between them and ensures that the semantics of visual embeddings are consistent with their textual counterparts. Equation 3.3 accounts for such circumstance in a mathematical fashion, where V entails the vector representation and the v and t indexes advert to visual and text feature subspaces. The identification of modalities is purely exemplificative.

$$distance = \sum_{(v,t)} \|V_v - V_t\|^2 \tag{3.3}$$

In light of and building on the distance premise, there are methodologies like cross-modal matching [75] which try to minimize the modality gap of matched data by reducing the difference of hidden representations across all layers while trying to maintain some form of intra-modality structure of similarity.

When compared to alternative frameworks, coordinated mechanisms tend to maintain the unique and valuable modality-specific traits notwithstanding the multimodality purpose, which, in an ample number of circumstances can be seen as an advantage [76]. Moreover, the separate inference process assists transfer learning when required. In contrast, the biggest downside considers the limitations and difficulties connected to learning representations with more than two modalities, as it seems that the larger the number of modalities to encompass, the more difficult it is to guarantee a viable transfer of knowledge [54].

Application-wise, illustrations such as cross-modal retrieval [77–79] and image captioning [80] make use of this coordinated learning mechanism to prove their cases and proposals. Furthermore, there is also a motivating idea that coordinated representations may be engaged for cross-domain transfer learning, normally as a means of reducing the need for labeled data to some extent. For instance, [81] suggests training a couple of networks, each for a certain domain, and coordinating them by reducing the maximum mean discrepancy metric between the addressed datasets to transfer information from one to the other.

3.1.3 Encoder-Decoder

The encoder-decoder paradigm, in a multimodal framework, functions as a mapping performance that transforms a certain modality input into an output with a distinct one [82, 83]. Figure 3.3 depicts such a procedure. The initial representation is encoded into an abstract high-level concept that suggests unraveling or decoding into a product of a dissimilar type. These two oncoming actions are respectively perpetrated by two main incorporated elements: the encoder, and the decoder.

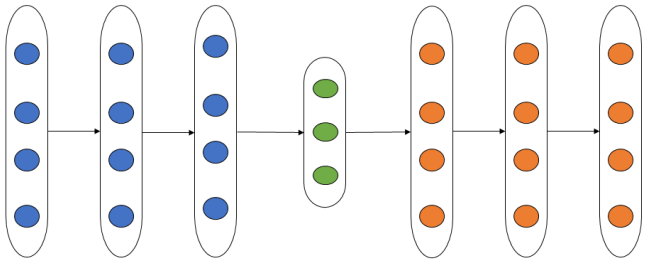


Figure 3.3: Illustration of a encoder-decoder representation.

The majority of encoder-decoder systems apply one overall conversion. Besides that, there exist reports of the usage of configurations that include several encoders or decoders. In [84], the discussed visual question and answering (VQA) model relies on a gate attention system that

encodes two different modalities (text and image) and enforces a replicated decoding phase in which both embeddings are tackled by separate decoders. As another example, the authors of [85] present a model that uses a single encoder to conceal a musical audio signal and multiple decoders to convert it into music across musical instruments, since each decoder has an intended domain.

Resolving into a mathematical standpoint and given an image/video to text transformation, the generalized learning aim of encoder-decoder models can be expressed as follows:

$$\Theta^* = argmax \sum_{(V,T)} \log p(T|V;\Theta) \tag{3.4}$$

The expression maximizes the log-likelihood of the textual target, T, granted the visual intake, V, and the model's parameters θ. Providing such equation and Figure 3.3, it could be argued that the encoder-decoder latent vector appears to be related just to the supplier pattern, but in reality, it is tightly linked to both source and target modalities. This duality is driven by the fact that the encoder is led by the decoder during training since the error correction signal flows from the decoder to its counterpart. As a result, the produced subspace tends to incorporate both modalities' common semantics.

Making use of an image captioning task [86], the image contributions encapsulate an arrangement of information that, somehow, has to surpass the outcome as text. These can include patterns, objects, colors, background, and spatial organization. The encoder is responsible for correctly retaining the features, and the decoder accounts for reasoning high-level semantic content and so creating the caption. A prominent and standard way to seize the shared semantics more effectively is to enforce regularization terms [87, 88].

More distinct approaches as [89] implement a hybrid methodology by reducing the representation discrepancy (often the case in coordinated circumstances), while maximizing the probability described in Equation 3.4. The example of [90] implies a textual-visual retrieval effort that uses a joint encoder-decoder system, in which the extraction of textual and visual information is inserted into a cross-modality embedding, that in turn takes advantage of this joint understanding to generate outputs of the same types.

The landmark recognition of encoder-decoder architectures, as Guo et al. [54] argue, is the fact that they are able to construct unique objects from a determined goal modality. Notwithstanding, the great bulk of the mechanisms in question can only discharge an individual encoding effort. Besides, the present generative attitude still faces a lot of challenges, in various fields and tasks [91, 92].

Image captioning [87, 88, 93], text to image synthesis [94], video description [95], but also retrieval assignments [90] are a considerable share of the exercises involving the discoursed model building technique.

3.2 Common Frameworks

There are several methodologies and frameworks that explore the above-discussed representations in different ways and with diverse motivations. Thus, it is also important to meet some of the most common structures within the multimodal domain.

3.2.1 Probabilistic Graphical Models

Probabilistic Graphical Models (PGMs) are statistical models that use graphical representations to interpret sophisticated joint multivariate probability distributions. In this fashion, PGMs try to capture conditional independence connections between interacting random variables [96].

In Deep Learning, PGM is a general name for various systems with distinct configurations that work with the affected sense described in the previous paragraph. Examples of deep PGMs that have been implemented to tackle a multimodal problem are Deep Belief Networks (DBNs) and Deep Boltzmann Machines (DBMs). Both ideas are established on Restricted Boltzmann machines (RBM) (introduced by [97]) but have dissimilarities. While DBMs feature fully undirected connections in the lower layers, DBNs have bidirectional ones.

In [98] the suggested DBN model creates a probability distribution across the space of multimodal inputs that enables to model case-specific events from conditional distributions for each modality in context. The authors indicate that the solution can learn a fair integrated joint subspace representation, so it could be a useful tool for both image annotation and image retrieval. A rather similar plan is motivated in [99] but instead, it uses a Deep Boltzmann Machine for classification and retrieval scenarios.

3.2.2 Deep Canonical Correlation Analysis

The attribution of deep canonical correlation analysis comes from the implementation of canonical correlation analysis (CCA) in a deep neural network context, being CCA a technique for determining the correlation between paired sets [100]. As a tool to guarantee coordinated representations, this method is normally applied in order to maximize the approximation of the concerned modality projections.

Deep CCA achieves the aforementioned incorporation by implying a new functional convention, as the regular CCA is limited to linear relationships and the use of more complex non-linear transformations is necessary. Although it presents relatively poor scalability (inherent to CCA's nature), as well as a high computational load, this methodology can be highly effective in certain contexts. To such an end, it is also important to control the maximization of the correlation of inter-modal spaces to avoid an excessive overlook of relevant modality-specific information. Ideas conveyed in papers such as [101] and [102] formalize some regularisation in the deployment of a Deep CCA tactic, whether for proposals that predict one modality in relation to another as in the case of [102] or for generative ones [101].

One of the advantages of Deep CCA over other multimodal frameworks is its ease of adapting to unsupervised scenarios. Because of that Deep CCA has been connected to a variety of multimodal learning undertakings, including acoustic characteristics representation [103], emotion recognition [104], classification problems [105], and voice recognition [106]. On the other side, Deep CCA's already entailed drawbacks, indicated by Guo et al. [54], can bring some inconveniences to numerous types of other applications.

3.2.3 Convolutional Neural Networks

Convolutional neural networks are used as the visual associated building block for a lot of multimodal systems in a widespread manner, as manifested in works like [90, 93, 107]. In this way, the three examined types of representations are broadly handled in strategies that employ CNNs as key elements.

Despite the universality of the adverted unilateral intervention, one can also find mediations that apply this kind of neural network as the main mechanism attributed to all modalities. The work of [108] inspects fall detection by computational systems as a way of expediting a support process for the elderly population. The contemplated procedures probe different topologies of a multimodal convolution neural network by treating images and information from accelerometers as the operating data to engage on. In this manner, multimodality is introduced as a means of improving the detection paradigm.

With a distinct view, Madhuranga et al. [109] present CNN models as a tool of video recognition, where knowledge is extracted from visual evidence and from audio signal descriptions (usually intended for recognition of activities of daily living). The identification of poses is completed via the usage of diverse information sources.

Sharing the perspective of [110], multimodal CNNs serve as good cross-modal feature extractors, more often than not in instances that include visual assessments, and can also represent spatial patterns from multimodal streams. A possibility that comes with the cost of benefiting from large data collections, and engraving time-consuming inference training procedures.

3.2.4 Multimodal Autoencoders

Multimodal and unimodal autoencoders go hand in hand with the formation of representations administered in encoder-decoder mechanisms. In fact, in these models, one only finds such conformation for compacting initial motifs into latent vectors [111]. Additionally, what, in some way, defines this pipeline is the reconstruction effort, as an autoencoder hinges on building back a decompressed object that can be as close to the original as feasible [111]. The specificity of the structure in vogue, pictured in Figure 3.4, reflects the development of a loss function tailored to its particular conditions. The reconstruction loss aims at what its name implies, as its minimization promotes a greater restoration, mathematically. The expression 3.5 formulates the reconstruction loss in a dual-modality spectrum. In it, x_i and y_i convey a pair of inputs and $\widehat{x_i}$ and $\widehat{y_i}$ express their reconstructed outputs.

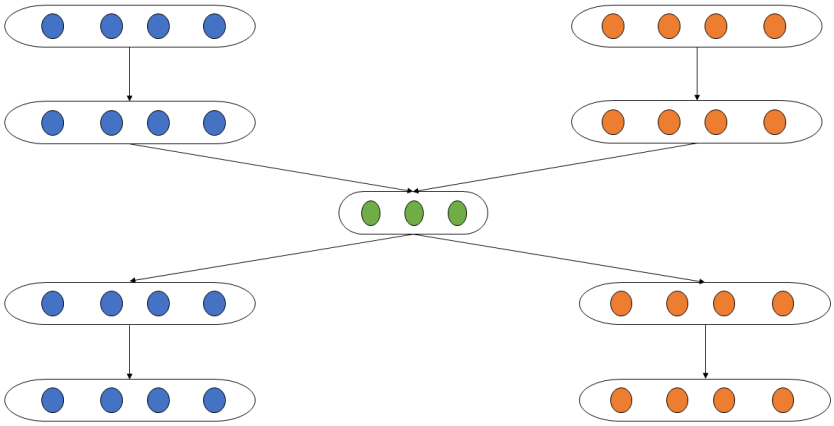


Figure 3.4: Multimodal Autoencoder.

$$Loss = \sum_{i=1}^{M} (\|x_i - \widehat{x}_i\|_2^2 + \|y_i - \widehat{y}_i\|_2^2) \tag{3.5}$$

Works like [112] defend that multimodal autoencoders, by enhancing the interaction across representations, regularize, in a way, the enforced constraints of reconstructing individual inputs. This notion, allied to what the author calls a step reconstruction process optimizes the manner of reconstructing representative features for the desired products and makes a perfect conjuncture for transfer learning tasks. The thesis is successfully tested in cross-modal retrieval, multilingual (cross-language) document categorization, and audio articulation.

Considering another example, Ngiam et al. [113] conserve a utilitarian vision of the opportunity to use data of different sorts, in that they propose a bimodal deep autoencoder that learns, for the case, audio and video projections with the purpose of making it possible to rebuild a deliberate entity without the coexistence of both modalities.

In the biomedical signal processing field, works such as [114] review EMG and EEG pathways as translations into divergent modalities, and engage a multimodal framework to explore the combined restoration process intrinsic to its deep autoencoder setup with the interest of mitigating signal distortion, especially at high compression levels, and improving a sentiment classification problem. Both proposals are achieved and evidenced.

In terms of advantages, multimodal autoencoders stand out as an excellent tool to combat scenarios where there is little or no data, while efficiently benefiting from the possible variety of existing resources. Concerning disadvantages, the workload and time associated with training can drive away potential users of such compositions, as well as the plausible losses of some spatiotemporal information when deepening the encodings.

3.2.5 Generative Adversarial Networks

The effect of GANs on the multimodal field is mainly reflected in cross-modal translation and retrieval issues. For translation purposes, and taking the example of text-to-image synthesis, Figure 3.5 aims at portraying a system whose essential task is to map a myriad of image-related concepts from a text source into a latent representation that, jointly with random noise, needs to be used as input to an image synthesis process. Following that operation, the encoded synthetic visual representation, collectively with the textual one, requires to be deemed compatible with the initial sample, by the discriminator. Regarding the labeling, E refers to an encoder, T and V depict the textual and visual modalities correspondingly, and the remaining identifications follow already defined conventions.

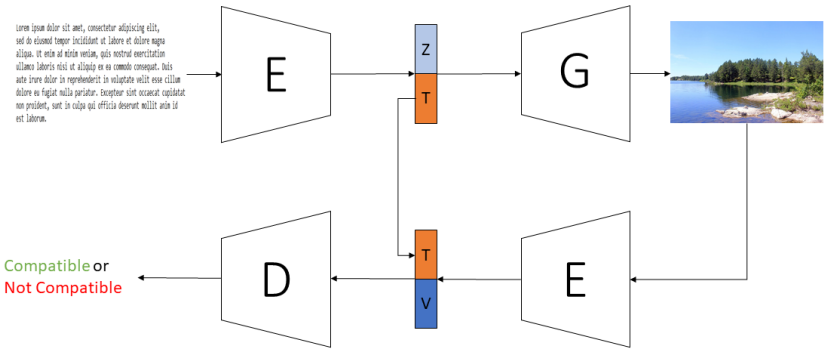


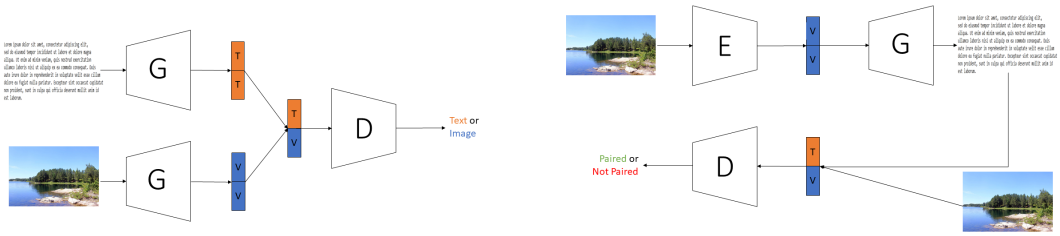
Figure 3.5: Cross-modal translation type GAN.

Reed et al. [89] marked a position, at that moment, with a network that followed precisely the logic presented in Figure 3.5, consolidating a somewhat prevalent approach to the task. Building on this thesis, the authors of [115] revealed a scene-to-graph methodology where they sought to make more of the insight impregnated within the contextual and textual relationships to facilitate the generator's ability to create realistic images. The executed rationale read as, by modeling sentences as graphs, a more developed network that is qualified to translate linguistic relationships into visual connections. Coming from a common starting point and setting a goal of enhancing textural features in the text to synthesized images, [116] introduces a model that integrates a decoder-aided self-supervised discriminator with a so-called feature-aware loss that grants the generator supervision over the multimodal representations from the discriminator. These refinements proved to be crucial in a quest for better picture generation.

The objective of GANs, in cross-modal retrieval, is to map paired inputs into a common representation such that the discriminator is unable to distinguish modality subspaces. The common

topologies of cross-modal adversarial models may be divided into two types based on the discriminator's input subjects.

In a system type depicted in 3.6a the discriminator receives a multimodal mapping that is produced thanks to the encoding done by the generators, that venture on approximating the unimodal latent vectors, for the sake of yielding the most modality-invariant space as possible. In its turn, the job of the discriminator is to dissect its input vector and categorize it into the fed modalities. The lesser the clarity of the distinction, the lesser the distribution split between modalities, albeit in an opposite scenario occurs a maximization of such gap. The exercises put into practice in [117] exemplify the usefulness of this technique, as in addition to their multimodal retrieval base process, the authors enforce this kind of GAN as a modality checking contrivance, to assure that the altered features are indistinguishable, and consequently, they guarantee a quality fusion procedure.



(a) Distinguish input modality. **(b)** Distinguish modality pairs.

Figure 3.6: GAN architectures for cross-modal retrieval.

The instance outlined in 3.6b serves as a modality pair identifier or classifier, where the in-built encoder-decoder structure engenders an object, granted a disparate modality intake, and where the discriminator rules a decision given operated representations of the original network input and the generated one. Tracking the example, the image projection vector is decoded into a textual sample, which is then combined with the initial entity for examination via the discriminator. Provided that the generation method was able to contain key visual information and exert a functional equivalency, the classifier will treat the replicated pair as a real pair. A negative response will be expected if the opposite situation presents itself. Once again there is this sense of game at play, as the generator strives to maintain a modality-invariant knowledge flow across the network, and the discriminator seeks to encounter inconsistencies in cross-modal representation.

The work of Zhang et al. [118] embraced a GAN to ensure consistency on their unsupervised cross-modal hashing problem. The authors contemplated the virtue of conserving manifold structural integrity across multiple modalities, apart from protecting inter-modality and intra-modality interactions in the shared hash space. The generator is trained to choose a modality-specific sample when assigned an example from another one, all within the same manifold. On the other hand, the discriminator will assess whether or not the created pair is part of the same framework. These routines are operated relying heavily on hash codes since the generator picks samples according to them, and the discriminator, in turn, analyzes the connection between modalities trusting on those

hash codes. In this fashion, it is concluded that the adversarial learning strategy is considered relevant to strengthen the cross-modal manifold structure preservation property.

3.2.6 Attention Based Mechanism

Attention-based mechanisms became, in recent years, an increasingly exploited tool, in the most diverse contexts. Initially, it was more revisited in the natural language processing field, due to its major valence when compared to other methods, for sequence data analysis. Moreover, it is now state of the art in XAI applications [119], in computer vision examples [120], and in other areas related with ML [121, 122]. The widespread interest is specifically due to the capacity of these contrivances to signal its system which instances, regions, or steps it should focus on. This actuality reveals an appealing discriminative ability to give more importance to certain patterns or representations or even neglect certain elements. This property has shown tremendous potential, both in terms of insight into models' functionalities, as well as achieving unparalleled results, in many cases, and promising results in other assignments

For multimodality purposes, attention mechanisms may be divided into two types: key-based attention and keyless attention [123]. As the name suggests, Key-based attention uses a key to identify relevant activities. Exploring Figure 3.7 for a relevant multimodal overview of what key-based attention would look like, and taking into account a traditional visual model (i.e CNN) in adjacency with a sequential learning textual one (i.e RNN), we can observe the key-attention dynamics related to the different time steps t.

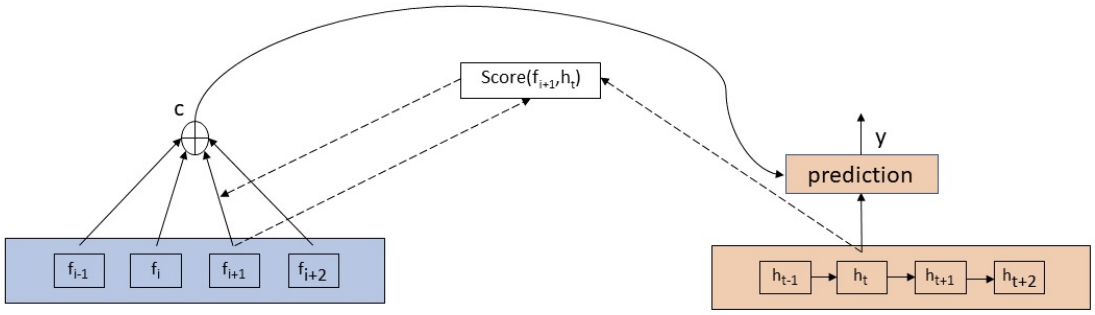


Figure 3.7: Key-based Attention Mechanism.

The visual element highlighted in blue, translates the image features into a set f_i, whereas the orange sequential system decodes an input into step or time-based representations in h_t. In this configuration, the f_i embeddings serve as a source to be sought, while the present state h_t in the decoder acts as a key. The output y, at time step t, is partially affected by the h_t contribution and partially impacted by the contribution of the weighted combination, c, of the various elements of the visual space. The relationship between both modalities is parameterized by the reported scoring mechanics (reflected by the Score block) that embodies the signaling property, as it also defines a clear influence on c.

Image description provides a large source of attention-based multimodal implementations. Namely, in [124] there is an expansion of the typical temporal and/or spatial clarifications attention mechanisms towards a modality established discrimination process, sparingly following modality-specific characteristics like audio features, image patterns, and motion peculiarities. Findings turned out to be very competitively, respecting the state-of-the-art, at the time.

The image question and answering problem exhibited in [125] are confronted with a stacked attention network built on top of the addressed framework that aims at more accurately identifying representative aspects of a particular question by searching for image areas that highly correlate with a possible query reaction. This work imposes itself as another illustration of the utility of key attention mechanisms in multimodal scenarios.

Keyless attention logically follows the same basic principles as key-based attention, but in a spectrum where the inference step is executed in one step. This is circumscribed to non-sequential classification or regression paradigms in which, without the use of a key, attention is focused directly on the localized vector spaces. In this order of thinking, there exists an ingrained considerable suitability for multimodal feature fusion roles that struggle with inter-model concept redundancy, semantic clashes and ambiguities, and even noise from the amalgamation of knowledge. It can be asserted that the produced unified representation with keyless attention intervention can help to retain concise multimodal information.

As practical use cases, [126, 127] describe two different angles of analysis regarding the keyless format. The authors of [126] state that the implication of attention mechanisms in a video classification panorama can be highly fruitful, especially in accessing and understanding interactions between representations of disparate modalities. The developed architecture settles on the application of relevance assignments, subsequently to the modality combination and concatenation of latent vectors. The novelty system suggested in [127] handles video classification with a transformer model (architecture with integrated attention mechanisms) that requires "fusion bottlenecks" so as to blend modalities, across various layers. In this case, the inputs are the sequential frames that make up the video and the corresponding audio signal. Since transformer models are, nowadays, state-of-the-art frameworks for treating sequential data, their employment seems natural and appropriate. Furthermore, unlike classic unimodal paired self-attention, the organization of the system at issue compels the model to aggregate and compress important information in each modality and imposes a limited communication to just what is required. The authors also conclude that imposing inter-modality transmission constraints enhances the merged performance while lowering computing costs.

3.3 Multimodal applications in medical contexts

Multimodal implementations in the medical field and healthcare, intersect a very broad spectrum in terms of purpose, vary in methodology, and also in the way they deal with different types of data. To provide a brief and exploratory overview, the following categorization is centered on two common instances that belong to a diverse number of existing tasks.

3.3.1 Decision Support / Classification

Classification is, probably the most widespread assignment in the ML and DL community. Notwithstanding, such remains a reality in healthcare and medical-related research work, knowing that many of these procedures are revealed in auxiliary diagnostic systems. Multimodal methodologies can be very important in a complementary perspective, where the inter-modality integration of knowledge aims at a more robust and complete decision performance. Thus, works that focus on these particularities may be fundamental in exploring and taking advantage of the resourceful use of medical records, as well as vital to expand the operation of these systems.

The authors of [128] operate on chest X-ray radiography and its associated medical reports for 14 multi-label categorization assignments. Nunes et al. [128] adopt a joint training process, where their multimodal model consists of two main components: a sequential input parsing element, which in this case is a text one, and a visual sensing element. The visual part resides in a CNN and the textual part involves a word embedding constructor, a two-layered block of multiple bidirectional LSTMs (mLSTMs), and a word-level multi-head attention block. The multimodal approach involves concatenating the final unimodal representations that are then passed through a 14-node layer. The latter is mapped by a sigmoid function that reveals the final predictions. Both training and testing procedures are done with the same structure. In this way, the use of images and text inputs is implied in both contexts. For the experimental assessment, 4 layouts were implemented. The first one concerned the evaluation of 4 unimodal visual baselines for grey-scale images and the second one centered on 3 baselines in an RGB setup. The third execution targeted the review of 3 medical report textual baselines' performance. In all three instances, the used models were pre-trained in medical or biomedical data. The fourth and final stage addressed two variations of the proposed model - one with and one without pre-training -, where the chosen CNN was an EfficientNet-B5 [129] and the word embedding system was the BioWordVec [130], while also testing both entities on a pre-trained single-modal setting. In the end, the pre-trained multimodal model proved to be the most capable through all classification tasks, presenting the best results across various metrics such as accuracy, F1 score, recall, precision, label ranking average precision, and classification error. In addition, the authors of [128] slightly confirm the robustness of their model by displaying visual grad-CAM heatmaps for two of the labels, as well as exhibiting the attention outlines for the respective text examples. The heatmaps focus on the expected areas, as well as the textual saliencies reveal higher attention attributions for words that are strongly linked to the classes in question.

The work of Biswal et al. [131] tackles the topic of multimodal data synthesis to aid towards greater availability of training resources in the medical AI field. To this, it adds the scarcity of synthesis studies in the multimodal ambit as a motivation for the proposed system - end-to-end multimodal X-ray generative model (EMIXER). The EMIXER produces synthesized X-ray images and corresponding free-text reports conditioned on diagnosis labeling. Such full functionality is achieved thanks to the model's 6 main segments:

- **X-ray image generator**: generates X-ray images from the combination of a noise vector and the class information that pass through a deconvolution residual block and a self-attention module.
- **X-ray report Generator**: generates a topic-based medical report from the synthesized X-ray images thanks to an encoder-decoder architecture. The sequential process starts by scanning the images through a CNN encoder in order to secure latent feature representations. Then the embeddings are decoded to topic vectors, via an RNN model, for each input sequence. These topic vectors are processed by another RNN decoder to produce the words and consequently sentences of the free radiology text.
- **X-ray image discriminator**: with a ResNet structural design, it ensures that the created instances fit the visual distribution of actual X-ray images.
- **X-ray report discriminator**: allowing for the supervised treatment of the real X-ray image associated report texts, it discriminates whether the newly produced ones match the codified real medical text concepts. Moreover, the framework is rooted in an LSTM model.
- **Joint discriminator for X-ray images and reports**: it forges the true multimodal mechanism of the presented proposal. It is motivated by the authors as a guarantee of further support to the generator network for generating higher quality outputs, that being that they assume a dependence and correlation between radiographs and reports. The joint feature discriminator receives one tensor composed of a concatenation of synthetic image and text embeddings and, subsequently, takes on the role of trying to figure out which of the inputs is constructed and which is false. This block is only constituted of fully connected layers.

The training moment is characterized by the contraction of the min-max dynamics applied to the custom loss function by the different components of the network. While the discriminators try to maximize the loss, the generators try to minimize it. Besides, it is also noteworthy to entail that such a function has some terms that account for fully supervised image rotation values.

For the experimental appraisals, Biswal et al. [131] attempt to answer a set of questions:

1. "*Can EMIXER generate high quality X-ray images?*"
2. "*Can EMIXER generate high quality pairs of X-ray images and reports?*"
3. "*Can EMIXER learn a high quality generative model from limited samples?*"
4. "*Can EMIXER be used to improve COVID X-ray classification?*"

Regarding the first inquiry, an X-ray multi-label image classification arrangement is organized. The performance analysis is executed for five distinct disease-related labels and the evaluation considers an accuracy and an AUC metric. The intention is to analyze the effect that the EXIMER model produces as a data augmentation tool. In this sense, two image classification models were

initially trained with a given data set and then were trained with that initial data set plus a given assemblage of generated data. The last varies depending on its generator model since the EXIMER is compared with three baselines: JointGAN, CoGAN, SMGAN. Performance-wise, an increase of 5.94% was concluded when compared to the initial set classifier and a 3.6 % improvement was established when compared to the best baseline. This demonstrates that EMIXER was able to create synthetic X-ray pictures that supplement an original dataset and enhance classification results.

To respond to the second question, two trials were planned. One of them was analogous to the previous experiment but transposed into a multimodal setting, where now the classification data is both visual and textual. Once again, the EXIMER framework proved to induce better results than the same baselines and than the only real input classifier. For the second instance, the test methodology is maintained, but the evaluation metrics are changed, as it is now a matter of judging the quality of the generated texts. According to several BLEU and CIDEr markers, the model in question improves, in an augmentative perspective, the generation of medical reports.

The third query addressed the system's capacity in a limited resource context. The challenge was faced as a self-supervision problem for classification and image generation exercises and so a partial labeling solution was applied. The EXIMER model guaranteed better outcomes in terms of accuracy and terms of image generation diversity when, once again, benchmarked against the baselines.

For the last case study, there was the prospection of the network as an augmentation instrument to enhance the detection of COVID-19 in radiographs. The detection was done in a multi-labeling scenario where there are 4 highlighted respiratory infections, among which COVID-19. The examination explored three alternative models: trained on a COVID-19 dataset, pre-trained on the *CheXpert* dataset and fine-tuned on a COVID-19 dataset, and pre-trained on real and EXIMER manufactured data and trained on a COVID-19 dataset. The results for AUC, sensitivity, and PPV illustrate how supplementing actual datasets with EMIXER-produced samples enhances overall performance. Moreover, the artificially produced X-ray images and reports were qualitatively inspected by two radiology clinicians. The radiologist's remarks imply that synthetic samples were somewhat identical to genuine ones, besides some X-ray report language incoherence.

3.3.2 Retrieval

Retrieval efforts are probably, along with classification exercises, one of the most popular practices in the ML/DL medical domain, which makes a lot of sense considering the opportunities that exist to improve and optimize the access by professionals to concrete and relevant data. The objective is, in each particular situation, for clinicians to be able to benefit from the best possible conditions to provide the best possible health service. It is in this scope that works like [132] and [133] are inserted.

The aim of Yu et al. [132] is to pursue better and more efficient Content-Based Image Retrieval (CBIR) for radiology purposes. For that, the authors argue that the key resides in bridging semantic gaps and estimating similarities between fine-grained image inputs and retrieved outputs, and so

they present a "*Multimodal Multitask Deep Learning*" technique that scores similarities between query and database instances using X-ray image and report representations in a common subspace.

The training phase reveals several procedures, which in short aim at improving the research and matching process. It is a sort of supervised joint classification setup, where it is intended to approximate the visual and textual representations to the corresponding label matrix via two L2 losses [134]. Before that, both inputs are compressed to individual feature embeddings that manifest in the same latent space, and the link between modalities happens when these representations are combined into a coupling matrix whose columns correspond to different semantic projections. The L2 loss optimization is done for each of these columns. As for the model image branch, there are some unique nuances that are essential to highlight. Initially, a Selective Convolutional Descriptor Aggregation (SCDA) algorithm [135] is applied to the original training images in the interest of localizing the condition-associated regions. Saliency masks are created and then employed to generate image descriptors, which in turn are fed into the proper model. Additionally, complementary to the integrated training, a triplet loss is adjusted for the visual representation production component to improve image discrimination based on minor details. The authors found these imaging modality tweaks to be highly fruitful for the final retrieval outcomes. With such a learning strategy, the authors believe that the system is able to reduce the heterogeneity barrier across distinct modalities and learn common subspaces while simultaneously retaining semantic selectivity and modality invariance.

The test framework, necessarily, has a retrieval nature. In this case, it regards as unimodal, since it only plans to resolve similarity metrics for images present in the database in vogue. Like so, a cosine distance [136] among query images, and each stored sample is quantified by exploiting the training image encoder representation policy. The retrieval output itself is then sorted per distance metric.

The designed model is composed of a VGG16 for the extraction of the image descriptors, a DenseNet121 for the visual encoding, and a Doc2Vec model for the textual representations. All elements are pre-trained on medical data. As for baselines, a DenseNet121 is assigned for unimodal image retrieval, and also a current Attention-based Triplet Hashing (ATH) system [137] is utilized for comparison purposes.

For the experimentation setting, the mean average precision (mAP) score [138] was executed for all testing samples and the average precision [138] was calculated for each particular class prediction on the top 10 retrieved instances. Retrieval is deemed correct if any of the fetched image's positive labels coincide with the search query. The mAP was tracked for the top 1, 5, 20, and k (where k is the number of test cases) retrieved objects.

The proposed multimodal framing induced a 4% improvement in mAP over the DenseNet121 unimodal retrieval baseline, while it also prompted an mAP enhancement concerning the ATH structure. Furthermore, additional ablation studies determined that, for mAP, the multimodal share of knowledge delivered a 3.0% increment over the equivalent unimodal image-only baseline, that the deep descriptor learning supplied a 0.4% growth, and that the ranking loss optimization impelled a 1.1% gain.

To finalize, image saliency maps were generated in order to qualitatively probe the model's attention areas for the executed activity. It was verified that, in general, query and database heatmaps extend a considerable pixel overlap.

The proposal developed in [133] attempts to investigate a multimodal learning approach as a means of verifying X-ray image retrieval quality, in an integrated manner. As such, the system under analysis is capable of separately receiving textual and visual inputs and assembling them in an abstract space. Regarding the proposed architecture there are two main tracks. A text-related section where there is a block to map the embeddings and a layer that scales the latent vector for the implementation of the supervised loss function, and an image portion where a DenseNet121 network receives image inputs and produces a unidimensional condensed tensor, over which PCA is operated to reduce its size.

The modality relationships are modeled in a coordinated way via the progression of the carried out learning exercises, which embody three types of training. The first supervised training process involves minimizing a squared error type loss, thus mitigating the divergence between embeddings of different modalities. The second methodology focuses on Adversarial Domain Adaption (Adv). In this one, a discriminator tries to distinguish the domains of the already mixed representations, and the generator tries to merge and make it increasingly imperceptible to detect discrepancies in the joint representation. A third scenario brought the two previous solutions together to understand how much supervision is needed to make the adversarial study feasible. Across all contexts, some orthogonal regularisation was introduced.

The testing mechanism is consummated as an effort of cross-domain retrieval and, with respect to that two subsets have been defined: one for the image to report query and another for the inverse procedure. At a comparative level, several types of word embeddings were assigned to resolve various test models. Metric wise, cosine similarity measures, Mean reciprocal rank [139], and multiple normalized discounted cumulative gains (nDCG) [140] were studied.

The paper's conclusions begin with the realization that, on a wide scale (top 100 nDCG), the unsupervised algorithms may obtain equivalent performance on disease-related retrieval tasks without the necessity for labeling. Moreover, the experimented bi-gram text embedding model proved to be the most consistent for the in-built retrieval system across the distinct training techniques. On smaller scales, Hsu et al. discovered that even minor supervised interventions (as low as 0.1%) can boost the execution of the retrieval task.

3.4 Conclusion

In the context of the dissertation, multimodality and multimodal learning center on how to develop machine learning strategies that make the best of a set of data of dissimilar kinds in order to accomplish a determined function. The fundamental addressed learning strategies behind most works in multimodal Deep Learning serve as a foundation to achieve such purpose and tackle multimodal challenges. Understanding how they unfold is pertinent in itself, and it becomes essential to understand how the ensuing experiments are framed. Furthermore, it is also significant

to recognize some of the most operated configurations and models in multimodal setups, as well as it is important to illustrate successful examples of how to address multimodal data and Deep Learning within the medical/biomedical spectrum.

Chapter 4

Literature Review: Explainability in Machine Learning

As already briefly scrutinized, interpretability and explainability in machine learning emerge as a necessity of having tools that are capable of directly supporting their decisions or being understood via its mechanism or via its output-input relation [141, 142]. From this perspective, it is comprehensible to stipulate that the two concepts are highly linked since they are both proposed to achieve the same general goal. In spite of that, interpretability and explainability might not necessarily mean exactly the same.

An interpretable machine learning algorithm can be conceived as an algorithm where the connection between the initial inputs/features and the outputs/predictions can be traced, or more importantly, understood by humans - it is inherently interpretable [142, 143]. Often, the operators of interpretable instruments aim to produce logical interpretations for a specific task from which the algorithm is responsible, without sacrificing its efficiency. In this regard, usually, it is an essential part of the work to manage the equilibrium between a simpler model, easier to interpret, and a good accuracy [141]. It can also depend on the end goal and in what scientific area is the model designed [3, 144]. All in all, interpretability relates to understanding what causes the leading of an input to an outcome.

Explainability tries to approach what's behind the process itself. This subject is a lot of times related to *post hoc* methodologies that underline explanations for established complex black box models' forecasts [5]. Hence, it is linked to the idea of explanation as a human-to-human interface that is both a correct proxy for decision-makers and intelligible to humans [145].

Applying another broader distinction, interpretability is centered on a whole strategy of a predictive and interpretable system [5], and explainability can be centralized on methods/algorithms that are able to expose and extrapolate explanations [7].

All differences aside, the two notions are often widely treated, in human terms, as a natural sense of understandability for machine and deep learning procedures [146]. Both are extremely important, especially, in the medical field and for medical/biomedical applications because they give insight on a myriad of essential elements for this type of functions [141]. These elements

include accountability, as in a system may need to be liable and be able to explain its results [147], quality assurance, since we have a thorough look into the model's functionality and performance, ensuring more knowledge and robustness over it [148], and trust since, for example, a clinician needs to be capable of trusting a system that provides him with crucial information about a diagnosis [141, 146]. Allowing for the above premise, both concepts will be invoked in a common manner.

4.1 Taxonomy

Many taxonomies are used to categorize and distinguish XAI techniques. These different criteria handled in literature can be noticed in Table 4.1

Table 4.1: XAI Taxonomies.

Taxonomies	Description
Local	The local criteria catalogues a technique based on whether it is used before, during, or after a model is built, and whether it is applied before, during, or after its development [143, 146, 149].
Scope	It distinguishes between local and global routines, so as the outlook of the exercised procedure [150].
Output	An output type taxonomy explores the divergences based on its own outcomes. [150, 151].
Transparency	The level of transparency and understandability of the internal mechanics defines this criteria. [143].
Agnosticity	The agnosticity convention simply identifies executions per applicability. In consonance with this arrangement there are model-specific proposals - limited to certain types of models - and model-agnostic appeals - may be used on any machine learning model since, by definition, they are *post-hoc* methods (operated after model training). [150].
Methodology	DNN-specific regulation that divides approaches into groups based on how they reach explainability and what they're attempting to achieve, whether it's an explanation of the network's processing or representations, or an explanation-producing system. [152].

Concerning a deeper analysis of the local taxonomy, works such as [153] and [150] scrutinize how it can really rely on a pre-model, in-model, post-model discrimination, but at the same time be alluded to intrinsic or *post-hoc* perceptions, that are more regularly than not associated with in and post-model forays.

Pre-model resolutions occur before beginning any algorithmic endeavor and are independent of it. They are data-centered and advocate the value of studying and comprehending the data regardless of any system considering. Some data characteristics like sparsity, monotonicity, and other intuitive feature patterns can raise this sense of pre-model/data interpretability [6, 146]. Additionally, traditional exploratory data analysis methodologies like Principal Component Analysis (PCA) [154], Clustering methods [155] and Mt-SNE (t-Distributed Stochastic Neighbor Embedding) [155] are heavily enforced tools for accomplishing some insight over the data.

In-model and intrinsic styles are managed by imposing limits on the ML model's complexity. The premise lies in the fact that a simpler constrained structure or one with certain pre-defined conditions leads to an inherent sense of interpretability across its own configuration, providing an answer to the issue of how the mechanism functions [146]. As to the conditions, these can be held as pure organizational and design constraints translating into imposing structural sparsity, monotonicity, causality, or other restrictions and regularisations derived from domain knowledge [5]. On the other hand, they can also be found in rule-based or case-based approaches. Given the example of decision trees, they follow a tree-like representation where each node serves as a data feature, and the corresponding edges are rules that apply to the last. Because a user may mimic the model's behavior while making a choice and ensuing a specific path, it is intrinsically understandable, simply by adhering to its guidelines. This rationale is implemented in [156] in order to generate explanations. Other research endeavors as [157] rely on the use of cluster divisions to create them. Here, each cluster has a prototype and a set of distinguishing characteristics.

Post hoc solutions demand the establishment of a secondary explanatory model and as a result, these tactics do not fully capture the true reasoning behind models' behaviors [158]. These methodologies are requested in three distinct instances: model interpretation, which aims to describe its behavior; model inspection, which enables an algorithmic examination, to grasp some of its attributes; and result explanation, which defines the logic behind a single decision [145].

4.2 Intrinsically Interpretable Models

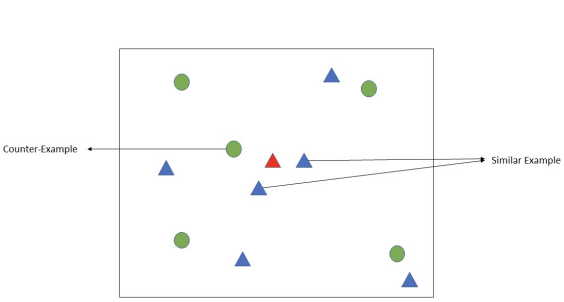
Intrinsically interpretable models are models whose inner workings can be easily understood by a human operator. In a practical and objective view, the required steps or calculations for running the model should be completely and equally inferred by a human actor. Besides the latter should also be conceivably reproducible. The expected characteristics from these systems demand them to be in-model and model-specific by definition [150].

There are several peculiarities within this genre of models, in that there are also different ways of displaying explanations whose logical rationales vary as well. In this case, the understanding and insight may come from example-based explanations, rule-based explanations, relatively accessible mathematical foundations, or simple definitions of probabilities. The accession of explanations by examples is predicated on the assumption that instances that comply with a series of category-specific requisites are eligible to be part of the same set or classification.

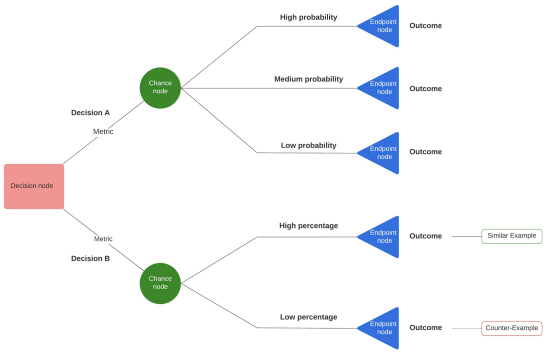
In the ML spectrum, the K-Nearest Neighbours (KNN) algorithm is the most recognizable model that embodies the example-based thesis [159]. To categorize a new observation, the classifier, firstly, tallies a distance between it and all training samples in order to determine which are its nearest neighbors. The label assignment is motivated by the K number training samples that are more identical to the one in analysis, whereby the majority class within those K samples wins. For explanation purposes, this methodology can serve the neighbors' imagery for comparisons. For instance, neighbors that belong to the same class may be exhumed as comparable examples, whereas neighbors of a distinct class can be regarded as counterexamples. The ease of using this

technique as a foundation for explanations is essentially rooted in the ability to "understand" a specific instance in the dataset [150]. Given the instance of a neighborhood of hundreds or thousands of elements, we will hardly have an explainable scenario. However, if we have few comparable features, or there is a way of reducing such features to a salable level, then KNN is stated as a very suitable intrinsically explanation model [150]. Figure 4.1a serves to visually complement the understanding of a KNN model.

Concerning rule-based explainable models, the most immediate and direct illustrations are decision trees. Decision Trees are a non-parametric supervised learning approach that draws a tree-like diagram depicting the various outcomes of a set of connected decisions [160]. It enables to learn and compare data according to certain decision rules and so make predictions ensuing a seemingly logical reasoning choice path. A decision tree usually begins with a single node and branches out into different scenarios. Each of those results leads to new nodes, each of which leads to new possibilities. It takes on a tree-like form as a result of this as seen in Figure 4.1b. Nodes are divided into three categories: chance nodes, decision nodes, and end nodes. The probability of particular outcomes is represented by a chance node, which is represented by a circle. A decision node, depicted by a square, represents a decision that has to be taken, while an end node represents the decision path's final consequence [160]. The explanation process based on decision trees can end up being twofold. On one side, we are only focused on what are the decisions followed along the tree, and on the other, we are focused on the comparison of cases, where similar ones are implicated in the same outcome and where dissimilar ones are taken as counterexamples (Figure 4.1b), in the style of exampled-based systems [161]. This explainable nature of decision trees makes them very appealing for these approaches. Also, the fact that these systems produce divergent sequences makes them easy to visualise [150]. In terms of disadvantages, one must highlight the potential of these models to be quite unstable due to their threshold nature. A few tweaks to the training data can result in an entirely new tree [150].



(a) K-Nearest Neighbours Example



(b) Decision Tree Example

Figure 4.1: KNN and Decision tree examples - Intrinsic explainable models.

A linear regression model estimates an outcome via a weighted sum of the attribute inputs [162]. This methodology is unique to regression tasks and its drawn operation transcribes a linear rela-

tionship between the variables, which renders a context that, in theory, should be interpretable [150]. That notion can be achieved through the exploration of mathematical approach (Equation 4.1), where it is relatively straightforward to assess purely objective correlations, and thus construct a foundation. In visual terms, there are several types of more indirect plots that can induce a simple and concise reasoning, but even a direct expression of a graph, such as the toy example of Figure 4.2 where one maps an output y against an individual input x, there are purely observational inferences one can conclude. Such is the case if we consider the designed line as some sort of a threshold, or even for the direct reading of values that can easily function as explanatory.

$$y = \alpha_0 + \alpha_1 x_1 + ... + \alpha_n x_n \tag{4.1}$$

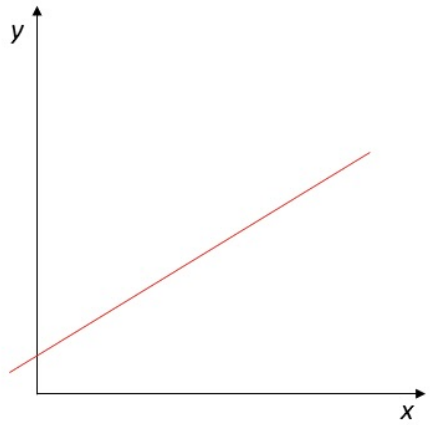


Figure 4.2: Linear Regression graphical toy example.

Advantage-wise, Molnar [150] argues that linear regression models can be generically employed. Besides its mathematical implementation can be considered somewhat trivial. On the other hand, the biggest setback is the fact that this type of method is very limited as to the context in which it can be useful. There are numerous situations, where the scenarios translate into non-linear panoramas, and thus modeling via this procedure becomes unfeasible [150].

Logistic regression models linear regression to a classification task setup [163]. As a result, there's a reformulation of the previous equation into expression 4.2. This restructuration transfers the linear problem to a probabilistic configuration where one wants to acquire the probability of an input being of class 0 or 1.

$$log\left(\frac{p(y=1)}{p(y=0)}\right) = \alpha_0 + \alpha_1 x_1 + ... + \alpha_n x_n \tag{4.2}$$

The logistic view has the great benefit of understanding linear regression for categorization purposes, as well as it provides values that define metrics of certainty or uncertainty [150]. Nonetheless, it suffers from the rest of the drawbacks that typify linear regression.

There are other adaptions of linear models that are worth mentioning such as Generalised Linear Models (GLMs) and Generalised Additive Models (GAMs).

For an interpretable probabilistic view, the Naive Bayes model is most likely the best representation. It utilizes the Naive Bayes theorem to determine the likelihood of a class for each feature based on its value [164]. The naive attitude is concerned with the assumption that all features are deemed independent, and so the estimation of the class probability is carried out granted aforesaid conditional sense. The modelling of a class Ck is as follows:

$$P(C_k|x) = \frac{1}{Z} P(C_k) \prod_{i=1}^{n} P(x_i|C_k) \tag{4.3}$$

The Z parameter accounts for the number of inputs x_i, and so it ensures that the probabilities sum to 1.

The Naive Bayes classifier is an intelligible model thanks to the understandability of a class conditional probability [150]. Like so, it is viable to gauge how much a feature bestows towards a specific categorization.

In a more global overview, there are certain defined characteristics that may indicate or facilitate the explainability of a model on its own [150]. Some of those are: **linearity** - if the linkage between variables and goal outcomes is linear, the model is linear; **monotonicity** - if across a whole feature space, the correlation between an input feature and the output one retains the same trajectory, the concerned system is monotone; **interaction** - the flexibility to incorporate feature interplay can translate to simpler interpretations; **sparsity** - it enhances comprehension since the fewer non-zero valued elements there are, the smoother it is for a person to comprehend them all.

Table 4.2 maps some of those properties to the addressed models in this section.

Table 4.2: Model comparison according to some properties. Adapted from [150].

Model	Monotone	Linear	Interactive
K-Nearest Neighbours	No	No	No
Decision Tree	Some	No	Yes
Linear Regression	Yes	No	No
Logistic Regression	Yes	No	No
Naive Bayes	Yes	No	No

The integration of intrinsic interpretability in DNNs is often thwarted by the setbacks that such imposition may generate at the performance level. However, there are several studies that apply some of the methodologies presented in this chapter to sediment an inherently interpretable approach. Granted the example of [6], Silva et al. presented a deep model that is capable of supplying complimentary style and depth explanations by imposing monotonic restrictions in the system's conformation. Such enforcement led to an increase in the quality of the explanations.

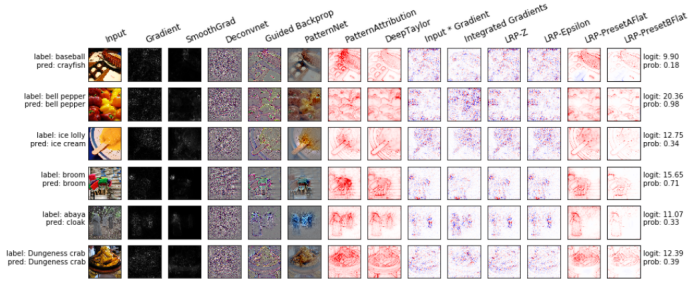
Works like [165] and [166] prospect the inclusion of the KNN algorithm in a deep framework, which by enforcing the interpretable nature of KNN models also enhance the clarity and robustness of their predictive DNNs.

Finally, Yang et al. [167] explore what they call Deep Neural Decision Trees, a hybrid implementation that combines decision trees with deep neural networks, taking advantage of the clear perception and understanding of the model. Moreover, the authors justify that the decision orientation of ruled-base models can be more intuitive and suitable especially for tabular data types, and in that regard, this proposal introduces significant improvements.

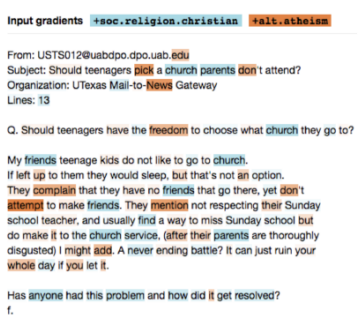
4.3 Explanation Strategies - Saliency Maps

The majority of modern and widely used explanation techniques tend to deliver on a network processing front as understood in explorations of the methodology taxonomy [152]. These approaches are aimed at describing how a neural network processes data, meaning how it consumes data and converts it into outputs. This work is done at the expense of attempting to encapsulate the complex and numerous mathematical operations and conformations, which is often accomplished via producing saliency maps that accentuate key aspects and calculations, or by constructing a proxy model that closely resembles the original one, but it is easier to comprehend and study.

When it comes to explainable DNNs, saliency mapping as a way of tracking the more relevant computations is likely the most popular sort of technique. Its applicability across different contexts and modalities [7, 168, 169], precisely shows transversality and prevalence within the XAI field and community. Figure 4.3 illustrates such plurality since it depicts examples of saliency maps from distinct modality survey works.



(a) Obtained image saliency maps from iNNvestigate library. Extracted from https://github.com/albermax/innvestigate (Accessed on 01-02-2022).



(b) Text saliency map taken from [170].

Figure 4.3: Saliency Maps.

The key premise behind saliency mapping methods is that feature gradients convey the significance of the inputs in relation to the outputs. Scrutinizing such an idea, it is possible to deduce an axiomatic relation, where a relevance score, R_i is established as a metric that expresses the importance that an input x_i has in respect to the decision or task produced by the model under study [171]. The many known techniques present different manners and/or intensities of entailing gradient paths, in fact, some use other additional attributions and mechanisms to obtain the references relevance score. Forging ahead with this idea, Kindermans et al. [172] details a separation

index for saliency approaches. Capitalizing on this account, the next subsubsections go through each of these groupings.

4.3.1 Functionality

The functionality study regards the explanation of a model's function by reporting some sort of description of the complex non-linear transformations that lead to the extraction of an output from a certain input space. Such can reveal to be rather impractical and unreasonable, and so one must resort to leaning on approximations [172]

Sensitivity Analysis is considered to be one of the most basic approaches to examine gradient propagation since its implementation can be rather superficial. It uses the backpropagation mechanism to contemplate local fluctuations at each node, which implies that it only explains the function's local slope [171]. This property induces this strategy to lose a lot of information. There is no valuation sense, as computed nodal outputs are not pondered and so, regularly, the locally gradient courses influence evaluation on an output appears as ineffective. Figure 4.3a labels the sensitivity analysis study as "*Gradient*".

SmoothGrad has a distinctive way to resolve gradient investigation. It artificially creates samples granted added noise to original examples and calculates an average sensitivity map per initial instance [173]. The computation is performed via stochastic approximation since it is unfeasible to directly calculate local means in such high-dimensional space. A SmoothGrad application can be also seen in Figure 4.3a.

4.3.2 Signal

The authors of [172] also claimed the idea that neural networks can identify a signal like an impulse as a gizmo that makes certain internal elements activate, and so define the propagation towards the network finality. Said concept or theory is considered across the following methodologies.

As the name may convey, the **DeConvNet** methodology solely exerts practice in CNNs. It traces events back to its input space, in such a manner that is viable to recognize what input instance triggered a particular activation in a determined layer [174]. The way to achieve this is by enforcing a reversion of the processes that occur in the forward pass. There is, thus, the application of deconvolutions and/or unpoolings at the layer level, in a logical sequence of trailing a network activation reverse path [174]. As a consequence, a generated activation reconstruction stresses the input features that impose weighted contributions at that specific instance. Another interesting fact about this methodology is the fact that each layer extrapolation reflects the network's hierarchical nature, as the initial layers detect simpler patterns like corners and edges, and the deeper ones seize more complex information [174]. The DeConvNet is exemplified in Figure 4.3a.

Guided BackProp was introduced by Springenberg et al. [175], and it builds on the DeConvNet proposal. In fact, the proximity between the two methods is confirmed in Figure 4.3a. The

main difference resides in the ReLU gradient calculation, as in Guided BackProp the ReLU functions are overridden such that exclusively non-negative gradients are backpropagated.

PatternNet is deemed to be an improvement on the previous propositions. The authors of [172] tuned their implementation thanks to the finding that the convolution filters do not obey the direction of a propagated signal across a CNN model, which ultimately revealed to be an impairing factor on the DeConvNet and Guided BackProp visualizations. The input data, in the study, is represented as the combination of a signal component, s, and a distractor element, d. That motivates the idea that the distractor shouldn't quite influence to output, whereas the signal should. As a result, the objective is to estimate the signal as accurately as conceivable. Such an approach leads to way more clear and "clean" saliency maps when compared with its predecessors (check 4.3a).

4.3.3 Attribution

Attribution reflects how much an activation contributes to the produced output across each layer [172]. As a metric definition, this sense of magnitude can be reproduced in a visual map. Multiple solutions for attribution visualization have been addressed.

An **Integrated Gradients** procedure relies on the accumulation of gradients along an exact route, from a baseline pre-defined input to a real one [176]. As it is intended to study an effect on a real input (Figures 4.3a and 4.3b), this path is established as the course integral of the gradients [176]. The incumbent gradient summation process makes the method quite generalizable and, therefore it is widely used to study many different DNN architectures.

DeepLIFT or Deep Learning Important FeaTures exploits backpropagation within DNNs as a means to compute the contributions of all neurons in the network with the respect to the output, decomposing its prediction on a certain input [177]. The score calculation is based on the difference between each neuron's activity and its reference activation. According to the technique's authors, in this way, the saturation problems caused by the existence of null gradients (common in a minimally complex architecture) found in approaches that use the backpropagation mechanism for studying gradients, are no longer a question [177]. Besides, there's a whole extent of discontinuous gradients, which usually deliver misleading relevance scores, that become unaddressed. DeepLIFT also identifies dependencies that are overlooked by other routines like integrated gradients, by considering positive and negative contributions separately. All of this is hailed, in a single backward pass, allowing for scores to be quickly realized [177].

Layer-wise relevance propagation - **LRP** - functions as a backpropagation technique. Its novelty activity explored in [178] concerns a conservation property at neuron level. This means that the share of an activation received by each neuron is evenly reallocated to lower layer neurons, following the backward pass sequence. In this manner, and ensuring the same course, a relevance notion is proportionately backpropagated depending on the lower level nodes' excitatory or inhibitory impact on the activity of the current neuron [178].

Just as the LRP method, **Deep Taylor Decomposition** highlights a repartitioned distribution of neuronal activations in a backward fashion. The relevance mechanics are then also cornerstones

in the Deep Taylor approach. The distinction of this practice mainly resides in the way of substantiating such relevance, since it exerts the mathematical formulations of the Taylor decomposition to the DNN context [179]. This translates to an estimation of every neuron's role as the difference between the neuron's importance metric and its computed root point.

In [172], Kindermans et al. operate upon the concept of Deep Taylor Decomposition to formulate the **PatternAttribution** solution. The uniqueness of the strategy lies in the fact that the root point calculation supports itself on an estimation learned from training data. This and the previous two methods are also previewed in Figure 4.3a.

The **grad-CAM** or class activation maps application is a CNN exclusive mechanism. The intuition behind this method is to follow the track of the gradients with respect, not to the output but the last convolution layer. In this way, a lot of the spatial information lost in the transition to the dense layers can be utilized to reveal where the model's centering is for obtaining certain conclusions [180].

In a completely alternative view to those discussed earlier, **LIME** or Local Interpretable Model-agnostic Explanations is a local model-agnostic method that outputs a Surrogate Intrinsically Interpretable system. As proposed by Ribeiro et al. in [181], the purpose of surrogate/mimic models is to use the input data and conclusions of a black-box model to train a simpler model, generally an inherently interpretable one. LIME exclusively trains local frameworks to explain predictions for particular occurrences, rather than taking a global approach. The LIME procedure makes a perturbation in the neighborhood of each occurrence in the input space and then builds a sparse linear model over that "new" dataset. Using an image as an example, the referenced technique executes these operations to each and every pixel. Finally, the information reflected on each account helps to create a global picture, map, or explanation of the unique input image.

4.4 Conclusion

The XAI field has already proven to be a well-established domain with an extensive amount of published research. In that regard, explainability and interpretability are often handled as a work finality themselves, or as a tool to fulfill major assignments. For instance, several of the demonstrated techniques are broadly used for classification tasks with the purpose of attaining insight into which instances engender higher contributions to influence models' decisions. Kohlbrenner et al. [182] and Han et al. [183] employ saliency maps as a proxy to score or measure instance, pixels and tokens/words respectively, contribution.

For content-based image retrieval (CBIR) regimes, Silva et al. [8] discharge a slightly peculiar take by examining the adoption of Deep Taylor Decomposition saliency maps as a means of directly improving the system, while the work [132] utilizes the grad-CAM methodology to confirm and validate their results.

From a different perspective, Montenegro et al. [184] developed a privacy-preserving generative adversarial network for producing case-based explanations concerning a glaucoma dataset. The authors reiterate the importance of example-driven explanations for clinical contexts.

In the experimentation phase of the dissertation, some of the discussed methodologies are covered via saliency maps. Techniques such as DeepLift, Integrated Gradients, Deep Taylor Decomposition, grad-CAM, LIME, and LRP are applied to yield saliency explanations from X-ray images and radiology reports, in a classification setup. Moreover, sample-based report-like explanations are the foundation and motivation for the last segment of the experimental endeavors.

Chapter 5

Unimodal Explanations from Unimodal data

This section alludes to the findings of the dissertation's first experimental effort. It pertained to investigating unimodal explanations from independent sets of unimodal data. The assays were conducted separately with the intent to gain some insight into the unimodal components to be included in the main work of the dissertation. Later multimodal methodologies were pursued to study how the inherent cross-domain knowledge intertwined in an integrated context. Throughout the experimental procedures, two datasets were explored.

5.1 Data

The main requirement when researching possible datasets was for them to be properly annotated catalogs of X-ray examinations. Ideally, it was important to find at least one multimodal dataset in which the X-ray images were accompanied by medical reports detailing the findings made by clinicians on those instances. In this fashion, it is conceivable to take some advantage of the multimodality nature of existing medical resources. With this configuration in mind, two datasets were selected: *CheXPert* and *MIMIC-CXR-JPG*. *CheXPert* constitutes a vast collection of de-identified and labeled X-ray images whereas *MIMIC-CXR-JPG* provides both types of features since it encompasses a considerable collection of radiographs and their associated free-radiology texts. Thus, *CheXPert* was only used for the initial trials related to visual modality and *MIMIC-CXR-JPG* was used for the text-only experiments and for the multimodal events.

- ***CheXpert***: The *CheXpert* dataset [185] is a public large dataset of chest X-rays released by the Stanford ML Group, which is widely used on studies about X-ray analysis and radiograph interpretation. It is composed of 224,316 unidentified chest radiographs from 65,240 patients. The radiographs were collected at the Stanford Hospital and performed between October 2002 and July 2017. The supplied labels, which are extrapolated from medical reports, cover a range of 14 pathologies/conditions and their occurrence may be positive (label 1), negative (label 0), unmentioned (blank label), or uncertain (label u). The fact that the

data record is substantial, has solid reference standards, and surveys a variety of conditions uncovers its interest.

- ***MIMIC-CXR-JPG***: The *MIMIC-CXR-JPG* database [186] is a publicly accessible collection of JPG-format chest radiographs annexed with structured labels obtained from corresponding free-text radiology reports. The set includes 377,110 JPG files and the associated 227,827 free-text radiology reports. A common reference for data splits is supplied and the dataset is de-identified. Moreover, the labels are also extracted from the semi-structured radiology texts and can also assume positive, negative, uncertain, and blank connotations. The large size, the multimodality setup, and the convenience associated with the prevailing image format were determining factors respecting the preference for the use of *MIMIC-CXR-JPG*.

5.2 Unimodal Experiments

5.2.1 Text related Experiments

The initial textual experiments focused on gaining some insight and perception into the methodologies and models that are able to draw information and knowledge from text. In essence, this segment served as means for conducting an exploratory study over those models, taking into account the targeted medical data in the interest of understanding how to incorporate such strategies in a multimodal context. In addition, it allowed to engage with explainability mechanisms and to understand which instances motivate NLP models to take certain decisions. Altogether, it served as the groundwork for the textual facet of the project.

5.2.1.1 Materials and methods

The data used in this segment were free-text radiology reports from the *MIMIC-CXR-JPG* dataset.

In view of this analysis, the investigation consisted in comparing the performance and interpretable saliency maps produced by three BERT models that differed in their conformation or compliance. The exercise at hand was a binary classification one, where on receiving a medical report as input the model had to decide whether it contained a positive or negative diagnosis of pleural effusion. Pleural effusion is generally referred to as an excess of fluid inside the pleural cavity or within the distinct layers of the pleura [187].

Only clinical reports (example in Figure 6.1b) affiliated to anterior-posterior (AP) radiographs were chosen to provide a strong and coherent assessment, and all uncertain groupings were discarded. At the processing level, every non-alphanumeric component was deleted from the clinical reports before ensuing the direction of the original BERT model [40], in which inputs are tokenized. This means that every instance in a sequence sample (not necessarily a whole word) was assigned to a numeric id (token id), an attention mask, and a token type id. In this case, all the input sentences are adjusted to a 512 token length - maximum size admitted by BERT architecture - being that some input sentences are truncated to this size and others are padded with zeros

until they reach it. The token id transforms the textual representation of each sequence element to a numeric expression that is "readable" by the model. This can happen at the word level as is usually the case, but can also enforce "cutting" the words. This is dependent on the *WordPiece* instituted tokenizer [188] that operates by breaking words into whole forms (i.e., one word becomes one token) or word parts - root formats - (i.e., one word can be broken into many tokens). Taking the example of the word "snowboard", the tokenizer would split it into the tokens "snow" and "##board". The attention masks assigned to each token represent a binary tensor that identifies the position of the padded indices so that the model does not focus on them. The token type ids which are part of the tokenization process do not disclose any relevance, since they are only used for activities involving paired-wise inputs, which does not arise in this instance.

After data partitioning, the training and validation sets were further split in half in an attempt to have more compact sets that would imply less computational burden and would not significantly impair the task at hand. Thereby, the training data collection ended up as a quite unbalanced set of 27697 reports where 72.4% of the labels are positive, and the validation one was composed of 238 samples with a 75.2% unbalance ratio, also favoring positive cases. As for the test set, the tendency was kept as 966 report examples pertained to an 80.4% positive label proportion.

Concerning the methodology itself, a binary cross-entropy loss is utilized. In training, it allows for optimizing and adjusting the network's weights, and in testing, it enables to access quality control of the frameworks. The evaluated models are: a standard BERT pre-trained on plain English language [40], a standard BERT pre-trained on MIMIC notes and other biomedical domain-specific languages (Bio + Clinical BERT) [189], and a smaller dimension BERT model also pre-trained on plain English language [40]. All these share the same configuration scheme, where after the feature extraction operation carried by the fundamental individual BERT structures for the classification task, the vector, progressively, undergoes a linear pre-classifier layer, a dropout layer to add more regularization, and a final linear layer that prefixes the classification. Moreover, all were fine-tuned on the downstream exercise in question over a period of 100 epochs using the Adam optimizer and a batch size of 6.

Several saliency maps were obtained using the Integrated Gradients method [177] (via Pytorch's Captum library [190] to confirm which tokens, and therefore words, the models focus on in order to support their predictions.

5.2.1.2 Results

Concerning results, Table 5.1 presents them taking into account the accuracy metric. It is possible to observe that the models that constitute the total base incorporation of a BERT architecture obtained extremely positive outcomes, being that the one pre-trained on data from the referent database has a slight advantage.

As an illustrative example, an explanation map for each model is shown in Figures 5.1, 5.2 and 5.3.

Table 5.1: Accuracy score for text experiments.

	Accuracy Score
BERT	0.99068
Bio + Clinical BERT	**0.99172**
Smaller BERT	0.74017

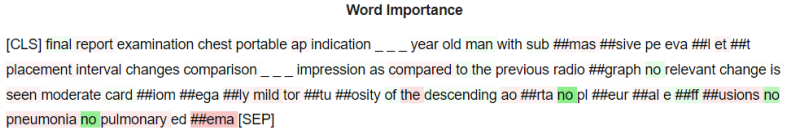


Figure 5.1: BERT integrated gradients saliency map for negative prediction and negative label.

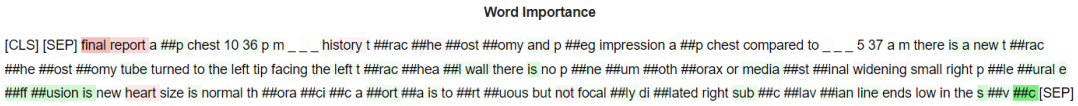


Figure 5.2: Bio + Clinical BERT integrated gradients saliency map for positive prediction and positive label.

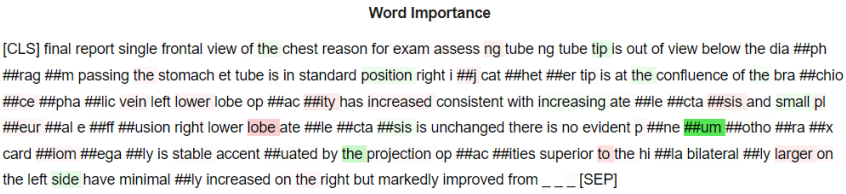


Figure 5.3: Smaller BERT integrated gradients saliency map for positive prediction and positive label.

5.2.1.3 Discussion

The preparatory assessment of the NLP models addressed some vocabulary nuance as well as enabled to experiment with an indicative explainability technique. The inclusion of distinct vocabularies in the methodology process came in as an effort to understand if the more oriented pre-training would facilitate training convergence.

As highlighted in the comparative resolution, the approach with a *MIMIC* specialized vocabulary turns out to have better performance even though by a slight margin when compared to the regular BERT counterpart. In any case, the obtained performance by the tested base BERT models yielded excellent results. This is due to the fact that these frameworks are able to infer knowledge from texts without losing information besides the text's sequential nature. Furthermore, textual resources much more directly and unequivocally convey the essential information to be learned for a decision than, for instance, visual samples, as they represent a human analysis of such data. Thus, it is to be expected that models with a routinized and prolonged training, as is the case here, achieve extremely favorable results. The smaller BERT ended up neglecting some of its predicting

ability despite the attempt to use a simpler and therefore more flexible structure.

The produced explainability maps confirm or build on the classification results since for the best performing classifiers there is a more robust support reasoning. Granted Figure 5.2, the instance reflects a positive prediction and in a supportive manner, the pleural effusion linked token, positively contributes towards the computed outcome. The example in Figure 5.1 reveals a high positive contribution of the negation adverb "no" when it is applied precisely to negate the presence of the condition in question, which makes a reasonable amount of sense given the absence of pleural effusion executed prediction. On the other hand, the map of Figure 5.3 is not very clear, or at least, it is not indicative of any expected contribution in a logical follow-up of the concerned scenario.

5.2.2 Image related Experiments

The experiments on images aimed to use and investigate different algorithms and techniques for generating visual saliency maps. Thus, with a view to complement the classification, it was possible to study which image regions were more and less fundamental for the categorization prediction.

5.2.2.1 Materials and methods

The *CheXpert* dataset described before was selected for the scope of this experiment.

For studying the implementation of various explanation techniques, two simple individual binary classification tasks related to two conditions that are fairly simple to interpret and monitor - positive cases indicate to specific areas on the x-ray images - were prompted. One regarded the identification of the presence or absence of pleural effusion in an X-ray image instance, and the other concerned the determination of a positive or negative cardiomegaly diagnosis, also taking into account chest X-ray images. Cardiomegaly is considered a general term that encompasses a variety of medical conditions that lead to heart enlargement. This enlargement is contemplated or defined as the transverse diameter of the cardiac profile being bigger or equal to half of the transverse diameter of the chest on a posterior-anterior (PA) configuration/projection of a radiograph or computed tomography (CT) scan [191]. With that in mind, and considering all images for the dual functionality of each classifier, two major operations were fashioned including the intention of prevailing coherence throughout the training process and in light of the objective of the exercise at hand. The first decision was held by only providing the posterior-anterior (PA) identified images for analysis, according to the fact that a PA setting allows for a better and clearer recognition of the cardiomegaly characterized physiological transformations. The second compromise concerns the uncertain labels as they were dropped.

The working intake was preprocessed in order to fit the required input format of the exploited ResNet50 network before settling into the different groups. As such, all images were resized to 224x224. Afterward, the training batches suffered augmentation procedures. The completed data augmentation processes render into slight rotations (10º to 20º) and reduced horizontal and vertical translations (to a maximum of 0.3 of the image width and height, respectively) of the

training image data. These operations aimed to avoid or decrease a possible problem of model overfitting to the initial dataset. This way, it is possible to introduce much more variability in the type of inputs and thus increase its versatility. Finally, regarding the dataset split, the cardiomegaly established drill headed an assortment that implied 4999 training images, 1767 validation, and 33 testing ones, while in the pleural effusion investigation, 15895 images were distributed for training, 1767 for validation, and 33 for testing. Both cardiomegaly and pleural effusion-related sets are quite balanced: about 48% positive cases of cardiomegaly across training validation and test and about 45% positive cases of pleural effusion for all subsets.

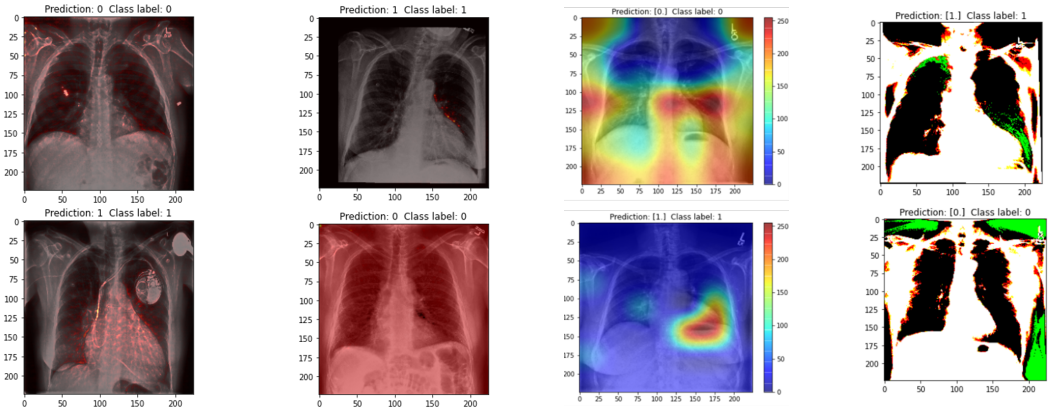
The model in charge of all train and test executions is pre-trained on ImageNet data. The mentioned 50 layer residual convolutional neural network was adapted to suit the binary context, and as such a final classifier layer was added. Concerning hyperparameters, both tasks were carried out during 16 epochs with batch sizes of 32 and resorting to the Adam optimizer.

On the issue of saliency maps, four different techniques were covered: Deep Taylor Decomposition [179], LRP [178], LIME [181], grad-CAM [180]. The Deep Taylor Decomposition and the LRP implementations were tailored from the Keras-based library Innvestigate [7], while the LIME and grad-CAM ones were attired accordingly to their respective original papers.

As to the loss mechanism, a binary cross-entropy conformation was again adhered to, since it perfectly fits the binary classification scenarios.

5.2.2.2 Results

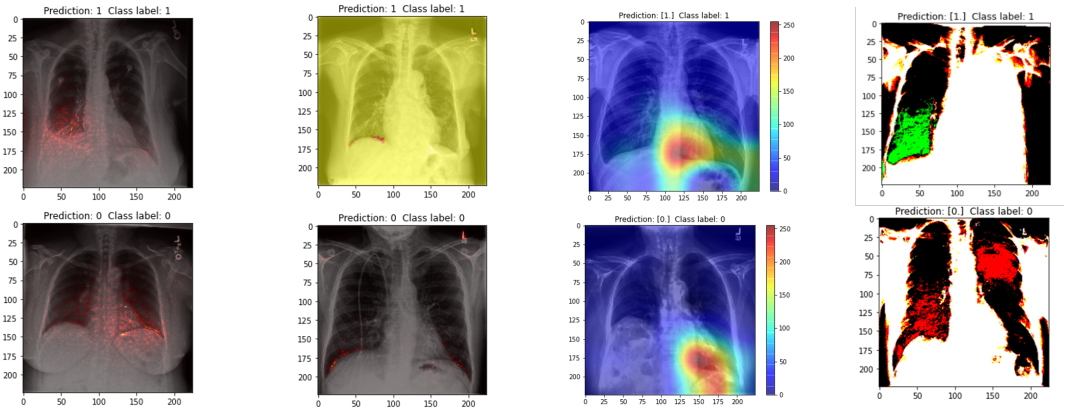
The cardiomegaly classification performance enabled a test accuracy of 96.97% to be achieved and multiple explanatory outlines to be engendered. As a representative sample, Figure 5.4 shows four generated maps for two scenarios where the prediction and actual label are both negative or positive.



(a) Deep Taylor Decomposition. **(b)** LRP. **(c)** grad-CAM. **(d)** LIME.

Figure 5.4: Explanation maps for cardiomegaly image classifier.

Concerning the pleural effusion exercise, the test accuracy was 90.91%. Once again, several saliency maps were obtained, four of which are portrayed in Figure 5.5 with the same conformation as the previous examples.



(a) Deep Taylor Decomposition. **(b)** LRP. **(c)** grad-CAM. **(d)** LIME.

Figure 5.5: Explanation maps for pleural effusion image classifier.

5.2.2.3 Discussion

The observed accuracy results are found to be quite robust considering the process and the envisaged applications.

For the cardiomegaly related explanation maps, it is noticeable that for positive predictions, i.e., presence of cardiomegaly (label 1), the Deep Taylor Decomposition (Figure 5.4a), LRP (Figure 5.4b) and grad-CAM (Figure 5.4c) methods focus mainly on the enlarged heart area. On the other side, the LIME (Figure 5.4d) drafts are not as clear to outline the same, as the green pixels, that favor the final decision, are more scattered and, serving the most optimistic overview, only surround the expected part of the image. When the model's guess points to an absence (label 0) the highlighted areas range is much more dispersed and sparse across all representations, and there is no evident concentration in a specific region. This suggests that the model does not encounter worthy "proof" of a cardiomegaly situation.

The concrete explanations for the scrutiny of pleural effusion usually reveal a spotlight on the lower portion of the lungs where the overflow should prevail, in a positive case, and where a noticeably amount of times, the aforementioned loss of curvature is, according to the literature [187], an indication of an excess of fluid. Unlike the last assignment, the model applies a focus over the mentioned section, independently of class, which may reinforce the idea that the physiological transformation that happens in cases of pleural effusion, fundamental for its diagnosis, is "understood" by the network as a crucial parameter for the outcome of the computerized ruling. As slight drawbacks, the high dispersion of the positive (green) and negative (red) votes within the explanations created by LIME (Figure 5.5d), and the lower concentration in the explanation of the positive prediction constructed by grad-CAM (Figure 5.5c) stand out.

Chapter 6

Unimodal Visual Explanations from Multimodal data

This multimodal predicated assessment was motivated by trying to understand if training a model with a multimodal configuration would bring improvements in performance and robustness for a binary chest X-ray classification exercise. To achieve a conclusion on that matter, several classification methodologies were, naturally, evaluated and saliency maps were generated. Furthermore, being aware of the real-life context on which this task could prove to be of high relevance is also important, as it may forward or suggest a framing for the experience itself. Thus, acknowledging that, when analyzing a new case, a professional usually relies on image-based diagnostic evidence, the carried-out approach follows a coordinated representation setting, where the textual information was used to regularize the embedding learned by a visual encoder, meaning that for inference only one modality is needed (i.e., the chest X-ray images).

6.1 Materials and Methodology

6.1.1 Data

All the available data for this trial comes from the *MIMIC-CXR-JPG* dataset.

In this study, the interest relies on analyzing the interpretability saliency maps generated for the different models. In order to have a clear and consistent comparison, only positive and negative labeled images acquired in an anterior-posterior (AP) orientation and the respective clinical reports were selected. Regarding the images (Figure 6.1a), they were resized to 224x224 and normalized following the same procedure as done for the ImageNet pre-training. To prevent overfitting, data augmentation strategies were adopted, namely, rotations within a 10º to 20º range, and random horizontal translations up to a maximum of 0.5 of the image width. For the clinical reports (Figure 6.1b), all non-alphanumeric elements were removed and a BERT-like [40] tokenization procedure was implemented. Special tokens were created for the beginning and the end of each sentence, and sentences instances were mapped to an id (token id), and an attention mask, which signals the model where to focus. For such a task, the token type ids, that are formally used

in the tokenization process, remain irrelevant since each input sequence is not paired. With respect to the assignment at hand, it centers on a binary classification setting, specifically in detecting the presence of pleural effusion.

(a) Chest X-ray

FINAL REPORT CHEST RADIOGRAPH PERFORMED ON ___ Comparison is made with a CT chest from ___ and a chest radiograph from ___ CLINICAL HISTORY Cough metastatic non small cell lung cancer assess for cause of new cough FINDINGS AP upright portable chest radiograph is obtained Overall there is no significant change from the recent CT performed ___ with innumerable metastatic nodularity involving both lungs and large consolidation occupying the right lower lung with a small to moderate right pleural effusion There is no new area of atelectasis or new area of confluent opacity to suggest a superimposed pneumonia though given the extensive underlying lung disease a subtle acute process would be impossible to exclude Heart size cannot be assessed Mediastinal contour is stable No pneumothorax is seen Bony structures appear stable Known metastatic lesions involving the inferior scapulae are not clearly visualized as well as the recently diagnosed nondisplaced fracture involving the right posterior eighth rib IMPRESSION Overall stable exam with extensive metastatic disease to the lungs with right pleural effusion and right basal consolidation

(b) Clinical Report

Figure 6.1: Test example - Chest X-ray and associated clinical report. Pleural effusion case.

6.1.2 Methods

The proposed methodology (Fig. 6.2) makes use of the available clinical reports to regularize the semantic space previous to classification. Moreover, it is designed in a way that does not require the availability of the clinical reports in inference.

Training: In the training process, the system receives two inputs: Chest X-ray images, and clinical reports. For each modality, a specific branch is built. For the image branch, a Convolutional Neural Network is used to extract the relevant embedding features. On the other hand, for the clinical report branch, a Natural Language Processing network is employed to extract the relevant embedding features. Through an embedding loss function (Equation 6.1), it is possible to promote a similar semantic representation for both branches. Assuming that it is easier to learn a good embedding representation for the natural language branch than for the image branch, since the clinical reports already result from a human description of the image information, the idea involves improving image classification by incorporating this knowledge into the network (via the embedding loss).

$$\mathscr{L}_{emb} = MSE(I_{emb}, T_{emb}) \tag{6.1}$$

Besides this embedding loss function, there is the main classification loss, which is the binary cross-entropy loss (classification task pleural effusion vs. non-pleural Effusion). Thus, the final loss function being used to optimize the architecture's parameters is the one described in Equation 6.2, where $\mathscr{L}_{clf}(y_{true}, y_{pred})$ represents the binary cross-entropy loss term and $\mathscr{L}_{emb}(I_{emb}, T_{emb})$ the embedding loss term, with λ_{clf} and λ_{emb} weighting the importance of each loss term.

$$\mathscr{L}_{final} = \lambda_{clf}\mathscr{L}_{clf}(y_{true}, y_{pred}) + \lambda_{emb}\mathscr{L}_{emb}(I_{emb}, T_{emb}) \tag{6.2}$$

Test: For test, only one type of input is given to the network, the Chest X-ray images. Thus, the inference process is similar to the one of a conventional CNN approach. This is important since at test time, clinical reports are not available.

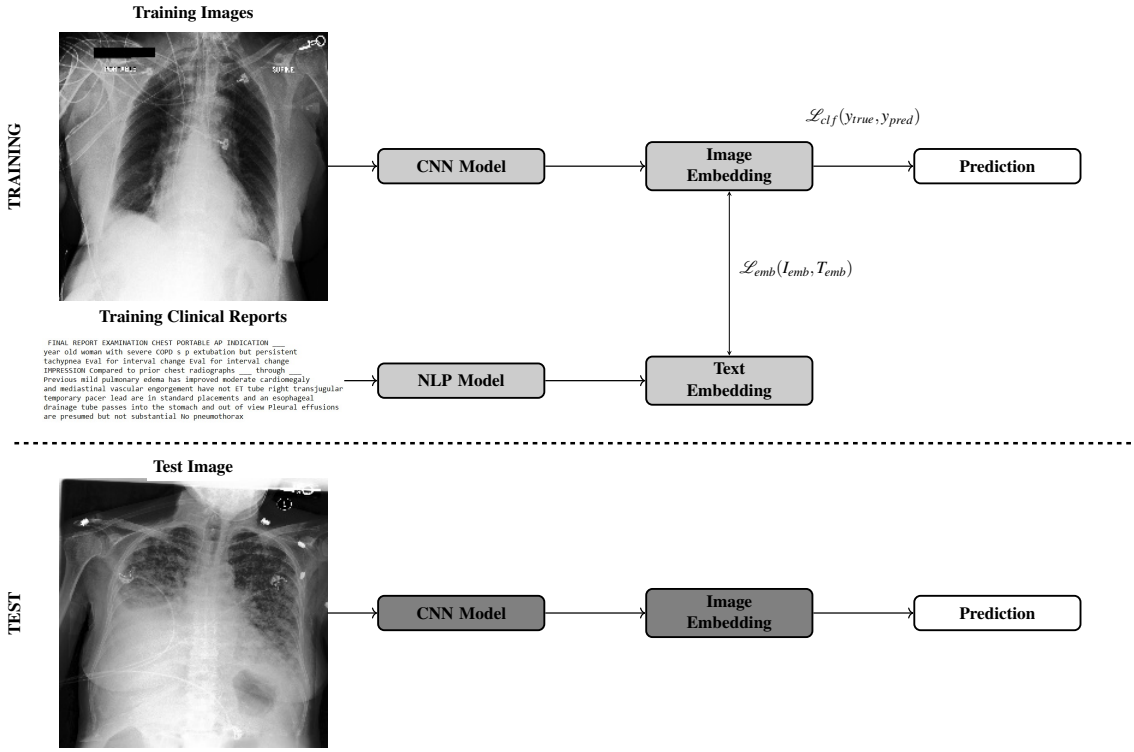


Figure 6.2: Overview of the proposed approach. Blocks in light gray mean deep neural networks are being trained (i.e., weights are being updated), whereas blocks in dark gray represent trained deep neural networks (i.e., weights are fixed). The block in white means there are no weights being learnt. $\mathscr{L}_{emb}(I_{emb}, T_{emb})$ represents the embedding loss, which depends on the Image Embedding (I_{emb}) and Text Embedding (T_{emb}). $\mathscr{L}_{clf}(y_{true}, y_{pred})$ represents the classification loss, which depends on the ground-truth labels (y_{true}) and predicted classes (y_{pred}).

6.1.3 Evaluation

The conducted evaluation process is divided into two main components: the study of classification task performances (accuracy and F1 score), and the analysis of the impact of the regularization process in the interpretability saliency maps generated.

Two different CNN architectures were considered and compared, with and without the clinical reports' regularization (working as baselines/ablation).

6.2 Results

Regarding the experimentation itself, the two deemed different architectures for our CNN model were the well-known DenseNet-121 [16] (pre-trained on ImageNet data) and a simpler CNN model (with three blocks of "Conv-MaxPooling", followed by three fully-connected layers). As the NLP model, we used the state-of-the-art BERT model [40]. Since the CNN and NLP models do not generate embeddings of the same extent, we added a layer in order to have equal size embeddings (for image and text). In the case of having the DenseNet121 as the system's CNN, the resulting image embedding length is 1024 while the textual one is 768, and therefore, before the inter-embedding operation, the image representation is compressed to match that of the other modality. In the matter of employing the simple CNN, the image embedding is smaller, and so the same procedure is applied on the textual side. All models considered in these assays were trained during 20 epochs, using a batch size of 4 and the Adam optimizer [192].

For the sake of observing the impact of the clinical report regularization in the robustness of the models, various interpretability saliency maps were computed. This was accomplished by adopting the DeepLift [193] implementation available in Pytorch's Captum library [190].

The aforementioned pre-processing operations were equally enforced for all listed models: our baselines, DenseNet-121 and the simple CNN, and the equivalent ones with the regularization component given by the textual embeddings calculated by the BERT framework. Accordingly, the data breakdown engaged 55395 training samples, 476 validation samples, and 966 test samples, while the label distribution refers to about 75% of positive cases throughout all sets. It is important to note that in test, only images are used, even for the proposed regularized versions.

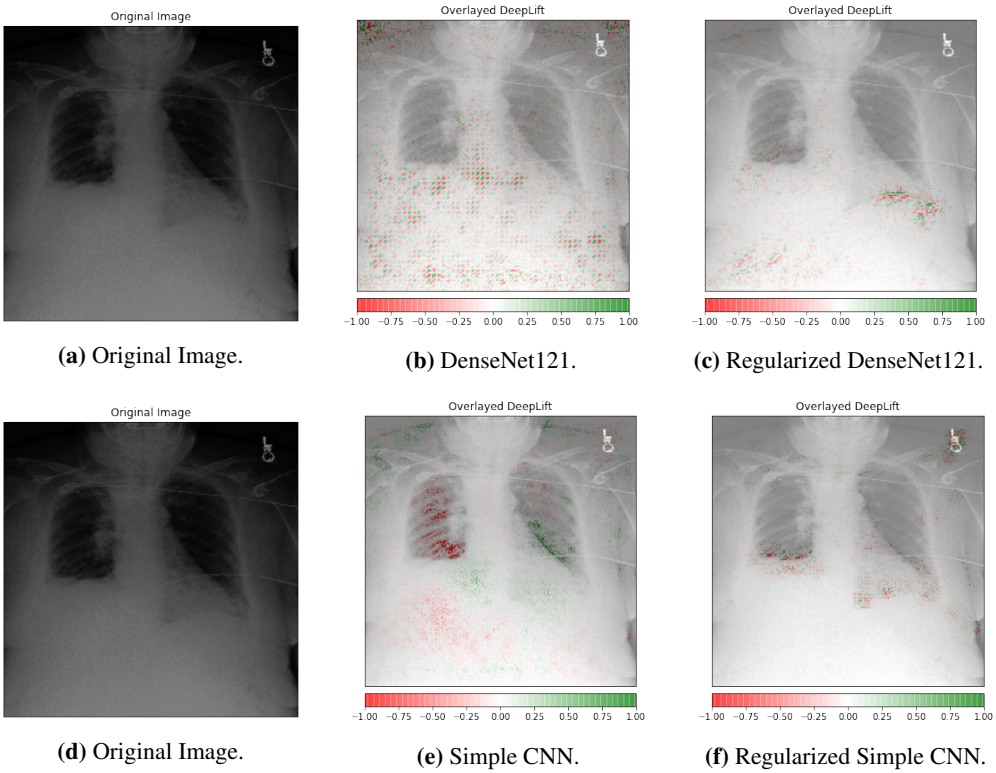
The classification task results are presented in Table 6.1 in accordance with standard accuracy and F1 score metrics. As can be observed, in both CNN architectures, the regularization promoted by the textual embeddings led to a significant improvement in both accuracy and F1 scores.

Table 6.1: Accuracy and F1 test scores.

	Accuracy Score	F1 Score
DenseNet121	0.84886	0.90943
DenseNet121 + BERT Embedding Representation	**0.86749**	**0.92050**
Simple CNN	0.80435	0.86243
Simple CNN + BERT Embedding Representation	0.81677	0.89751

For a more complete analysis, we present the saliency maps of all models for the four possible outcomes: positive prediction and positive label, positive prediction and negative label, negative prediction and positive label, and negative prediction and negative label. Figures 6.3 to 6.6 present those outcomes, with the top row of each figure showing the interpretability saliency maps generated using models based on the DenseNet-121, and the bottom row of each figure showing the interpretability saliency maps generated using models based on the simple CNN. As can be observed, the DeepLift saliency maps produced using the baselines (i.e., without the clinical report regularization) show a dispersed awareness, with several regions of the images being highlighted

as relevant for the decision, which goes against the clinical knowledge of the disease. When the clinical report regularization is introduced, the saliency maps become more concise, producing denser regions of importance, with that happening for both CNN models used.



(a) Original Image. **(b)** DenseNet121. **(c)** Regularized DenseNet121.

(d) Original Image. **(e)** Simple CNN. **(f)** Regularized Simple CNN.

Figure 6.3: DeepLift saliency maps for positive prediction and positive label.

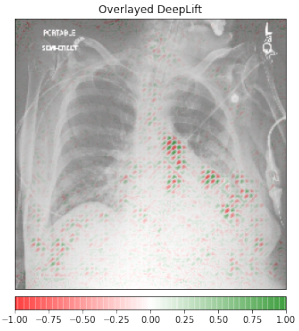
6.3 Discussion

The outlined research centered on the introduction of a multimodal learning setup aiming, not only, the performance improvement of the classification task, but also in the learning of more robust representations, leading to more trustworthy models. The regularization process induced by the clinical reports' data reveals promising outcomes on both counts, as it led to improvements in the classification performance and in the quality of the interpretability saliency maps. These clinical reports are already available, not requiring any additional annotation. Moreover, in inference, the model does not require the text modality to perform the classification and help the diagnosis.

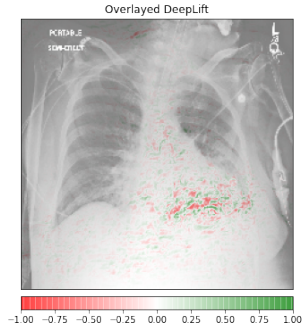
As shown in Table 6.1, the examples trained with BERT-based representations show transversely higher test values for accuracy and F1 scores, when compared to their peer baselines. Additionally and as expected, the DenseNet-121 achieves higher scores than the simple CNN, as it is a more dense and complex network and has already some prior knowledge induced by the ImageNet pre-training.

Across the diverse possible prediction-label combinations, it is possible to make a qualitative assessment and conclude that there is clear diminished dispersion on the interpretability saliency

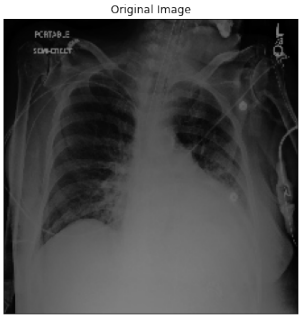
(a) Original Image.



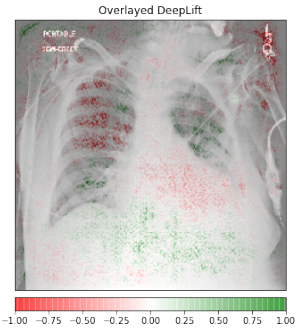
(b) DenseNet121.



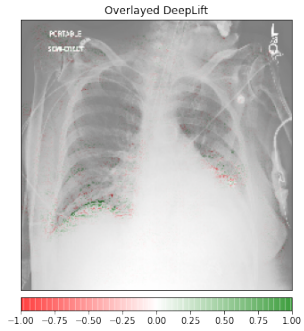
(c) Regularized DenseNet121.



(d) Original Image.

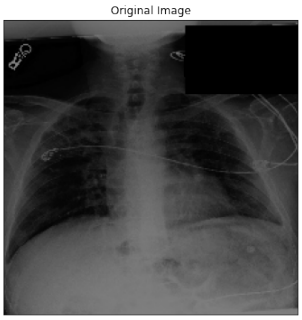


(e) Simple CNN.

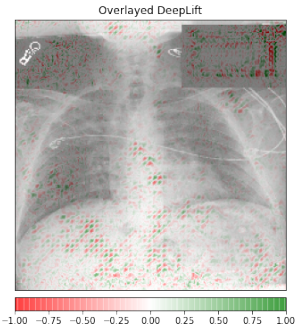


(f) Regularized Simple CNN.

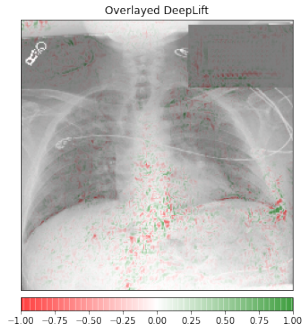
Figure 6.4: DeepLift saliency maps for positive prediction and negative label.



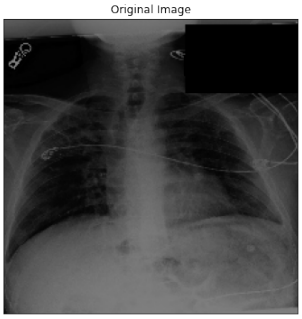
(a) Original Image.



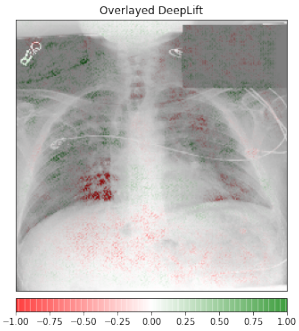
(b) DenseNet121.



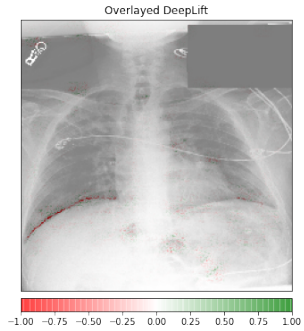
(c) Regularized DenseNet121.



(d) Original Image.

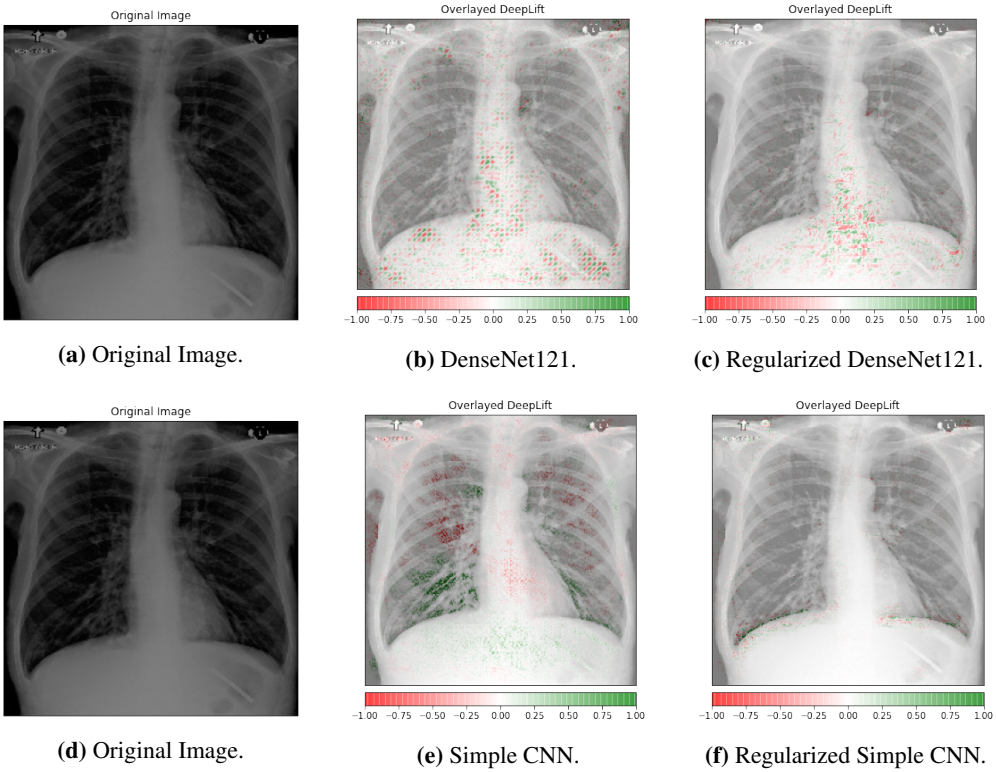


(e) Simple CNN.



(f) Regularized Simple CNN.

Figure 6.5: DeepLift saliency maps for negative prediction and positive label.



(a) Original Image. **(b)** DenseNet121. **(c)** Regularized DenseNet121.

(d) Original Image. **(e)** Simple CNN. **(f)** Regularized Simple CNN.

Figure 6.6: DeepLift saliency maps for negative prediction and negative label.

maps linked to the multimodal-based approaches. Furthermore, it seems that the spotlight is more circumscribed and usually present on the lower portion of the lungs, where a manifestation of the addressed condition should prevail, indicating an excess of fluid. Overall, and in a conclusive fashion, all extrapolated evidence confirms that the proposed approach came to fruition, leading to better classification performance and robustness (verified via interpretability saliency maps).

Concerning forthcoming developments on this exact methodology, there are two main different topics to which this effort could extend. Firstly it would be very interesting to apply the proposal to a Content-Based Image Retrieval (CBIR) context, expanding the work of Silva et al. [8], and secondly scrutinize the feasibility of broadening and generalizing the exercise to a multi-label classification one.

Chapter 7

Towards Multimodal Explanations from Multimodal Data

The creation of a full multimodal quality explanation mechanism from multimodal data, as described and defined in the goals section, involves conjugating the previously investigated and successfully implemented methodologies, in the visual domain, to a textual explanation generation component. In order to handle the problem, the idea and main intuition need to rely on the radiology texts in as a basis for training and optimizing the textual explanations. In this sense, it is clear that the brief rationales and indications present in the dataset are considered to serve as explanations of satisfactory relevance.

The following sections present the conceived encoder-decoder type proposals for producing textual explanations from images, and the experiments that ensued.

7.1 Encoder-Decoder Architecture

In the pursuit of a model capable of generating textual explanations of the desired quality, the architecture found in Figure 7.1 was suggested. The composed configuration is based on the resolution of Liu et al. [194], which developed a full Transformer model for image captioning. Ultimately, the inspiration makes logical sense, since from a practical point of view the produced operation is quite similar - generate text from an image. Following such reasoning and given a taxonomic perspective of the system in question, it can be argued that the network is a vision encoder-decoder model since it leans on its "visual" capabilities (interpreting image inputs) to decode a different modality output, i.e. text.

The implemented methodology is predicated on an encoder-decoder conformation that, as the designation indicates, can be broken down into two main components, the encoder, and the decoder. The employed strategy relied on the use of Transformer-type models either to operate as encoder as well as decoder. The reasons behind such a decision are expounded in the discussion section inherent to the methodology at issue.

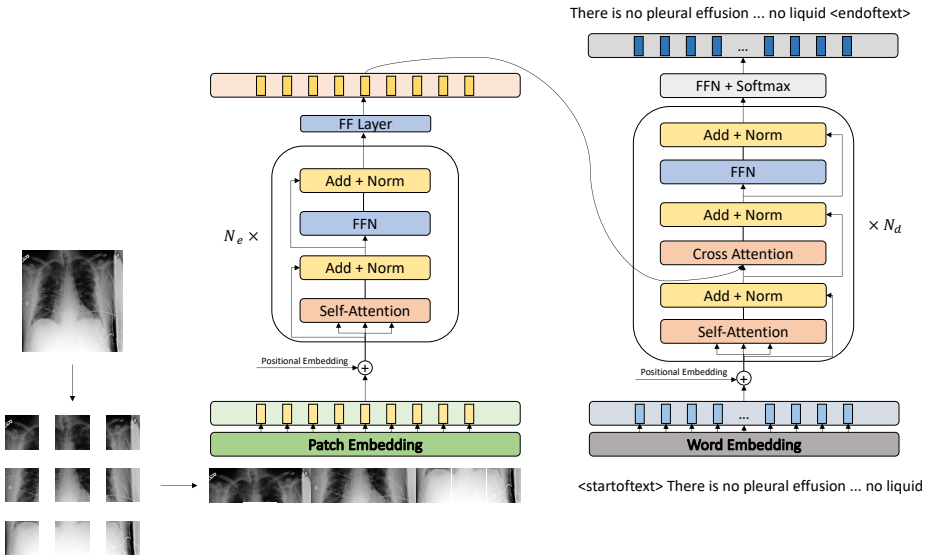


Figure 7.1: Encoder Decoder Model Architecture.

Further inspecting Figure 7.1, it is noticeable that the encoder is constituted by a N_e stack of encoder blocks accompanied by a feed-forward layer, and that each encoder block is composed of a multi-headed self-attention layer and two feed-forward layers (noted as FFN). The decoder component, on the other hand, is a N_d stack of decoder blocks and is also followed by a dense layer (normally mentioned as LM Head). Each decoder block is constructed on a sequence of a self-attention layer, a cross-attention layer, and a two-layer multilayer perceptron (MLP), marked as FFN. Normalization operations and residual skip connections are present in between all these core layers.

The encoder converts an input sequence into a contextualized encoded one. To do so, it needs prior knowledge of the output's length so it can generate, hence, intuitively, it appears to be unable or at least limited in performing sequence-to-sequence tasks. In contrast, decoder models map encoded representations into a target sequence, so they seem to work well for sequence-to-sequence assignments, but they have several restraints that hinder some of their potential, such as not being as efficient as the encoder-based models in conveying context. Thus an encoder-decoder model design, with the desired task in mind, becomes an even more compelling argument. The idea is also to try to take advantage of the strengths of each of the network's modules.

The experiments were conducted by making a few combinations of pre-trained encoder and decoder networks. On the encoder side 3 models were used: the Vision Image Transformer (ViT) [20], the Data-Efficient Image Transformer (DeiT) [195] and the Bidirectional Encoder Image Transformer (BEiT) [196]. On the decoder side 2 configurations were enacted: a Bidirectional BERT [40] model, and a DistilGPT2 [197] (distilled GPT2 version).

For the sake of a more coherent and in-depth analysis, it is indeed interesting to look at the structure of the system, in vogue, in a segmented way to better understand the functionality em-

bedded in the framework as a whole. In this vein, the following subsections are divided in order to describe the processes directly related to the encoder and decoder, separately, and also cover the dissimilarities between the various encoder and decoder models.

7.1.1 Encoder Models

For all pre-trained checkpoints, regardless of individual particularities, there exists a consistent pipeline and certain transversal features. On that note, it is essential to tackle the sequence-to-sequence motif - image to text. The encoder-related preprocessing novelty resides entirely on the sequencing of inputs that are partitioned into a series of image patches. For these experiments, prior to the segmentation, resizing, normalization, and data augmentation operations that have already been mentioned in previous trials, are applied. Considering the resizing of the images to a 224x224 proportion and to maintain their full integrity, the division is done in 16x16 patches and, as portrayed in Figure 7.1, the patches are flattened and reshaped into a 1-dimensional arrangement. Then, the patch embedding - a linear embedding layer - projects each instance into a linear tensor of size $3 \times 16 \times 16 = 768$, being that the calculation is accomplished by multiplying the color channels by the dimensions (height and width). The final transformation is performed by the learnable positional embedding, common to any Transformer. A further matter to underline is that the representation produced at the output of each encoder is passed as input to the next. The last element produces the final image embedding and attention tensors - fed to the decoding mechanism.

One of the advantages of the backbones under discussion is that they have all been pre-trained on ImageNet-21k in a supervised manner, and fine-tuned on ImageNet 2012. This gives some reassurance in a common point of knowledge acquired by all networks at the outset. All also share, similarly to BERT, the specificity of having a classification token at the beginning of each sequence and of each final representation. Accordingly, they are also built on the advent of possibly being fine-tuned for downstream classification exercises.

The DeiT model extends on the Dosovitskiy et al. [20] work as it employs the vision Transformer and is pre-trained and fine-tuned in the same fashion, but reveals a novelty training strategy that produces competitive results and requires considerably fewer resources, hence the prefix Data-Efficient. The supposed optimization of the training process is predicated on the concept of knowledge distillation (KD) [198]. KD is a training paradigm in which the network at the center of the process, coined the student model, uses the output of another model's - teacher - softmax function as a target label. This "soft" labeling instance, instead of trying to achieve the extreme scores (0 or 1), as in a regular supervised classification task ("hard" label), induces a much less aggressive operation. In the DeiT case, the teacher-student relationship implies an even "harder" technique than the illustration KD as the teacher prediction fulfills the same purpose as the true labels. In practice, such is attained via a mechanism imposed by the so-called distillation token. The distillation token is a token that is added at the end of a sequence of patches in the initial embedding. It works in an analogous way to the class token in that it also interacts across all self-attention blocks, propagates along the encoder, and is also transferred to the final representation.

As the classification token is utilized to compute the student model-related loss, the distillation token is operated to get the teacher's one.

The BEiT [196], as the designation suggests intends to translate the masked language modeling from a BERT model to an image setup. It thus applies the concept of masked image modeling to train or pre-train a Vision Transformer. The approach involves tokenizing the image chunks as already discussed and randomly masking some of these input tokens before passing them to the model. In the pre-training phase, the model is set up in a way that tries to predict the masked tokens, in this case, corrupted parts of the image that are then reconstructed. The training is conducted in a bidirectional manner, i.e., the model has access to all the tokens that are not masked, whether they are in a posterior or anterior position in the initial embedding. Furthermore, the BEiT model does not include the linear layer right before the final image embedding, which is depicted in the encoder side of Figure 7.1.

All encoder models executed a 12 stack encoder composition.

7.1.2 Decoder Models

On the decoder side, the differences are more pronounced than on the encoder side. The 2 models used in the undertaken studies configure 2 arrangements that share some nuances (in essence they are all Transformer models) but also entail relevant dissimilarities. Some common structural characteristics are accosted in the latter section, however, a few notes were left undone, more precisely in the last system's framing, where right at the end of the last decoder structure there are two elements through which each token representation propagates before consummating the final model output: a fully connected layer and the softmax activation function. The softmax defines the final probability distribution per token in the vocabulary. The higher probability-related token is the one that is chosen as the predicted word. On that note, it is relevant to highlight the exceptional instance of inference of the first word/token in a sequence. It is required for every input sequence to have a special sentence start token in order to commence the whole decoding procedure. To finish the word prediction another special token is generated - the end of sentence token.

A further fundamental issue, and therefore transversal to any decoder, is the communication between the latter and the encoder. In any circumstance, for the designed context, the encoded image attention tensors are operated in the decoder's cross attention block, as represented in Figure 7.1 and it is at this stage that the representations of both modalities interconnect.

7.1.2.1 BERT as a Decoder

The applied BERT model is pre-trained on a large corpus of English data [40], in a self-supervised fashion, taking into account the regular encoder-based format, and renders a 12-block contrivance. Notwithstanding, that same arrangement needs to fulfill the designated decoder blueprint. In that regard, leveraging a BERT model as a decoder can seem quite counter-intuitive, and that is the reason why, in reality, there is a slight adjustment to the BERT's layout when included in the whole of the implementation. The component remains as portrayed in Figure 7.1 and, consequently, what

occurs is that all the pre-trained parameters are synced to those of the actual decoder. The ones that do not match the parameters of the original BERT are randomly initialized.

Another significant imposed transformation on the nature of the BERT training methodology, allowing it to function as a decoder, is the way it computes its language modeling task. In general circumstances, a BERT network converts an input sequence into a contextualized encoded sequence, by virtue of its bi-directional training mechanism - the modeling takes place both from right to left and left to right on 15% masked sequence tokens. Nonetheless, it is enforced that the activity materializes from the rightmost token to the leftmost instance in the sequence, keeping with the auto-regressive style of a decoder generation process. Accordingly to these alterations, the normally called bi-directional self-attention blocks shall be referred as uni-directional. Nevertheless, at the parameter level, no modification takes place.

The pre-processing of clinical reports that are fed into the decoder can also engender a distinguishing trait. For the BERT case, the tokenization is handled by a WordPiece tokenizer whose operation has already been reported. In addition to the numeric identification detailed by the WordPiece methodology, there are special tokens such as the classification token ([cls]) that sets the start of each sequence, and the separation token ([sep]) which represents space at the textual level and also flags the end of the generated sequence.

7.1.2.2 DistilGPT2 as a Decoder

The GPT2 type models are Transformer decoder-based and therefore fit the provided description about the framework of a decoder in such conditions. They apply a uni-directional modeling technique to establish a mapping from an input sequence to a "next-word" logit tensor, that is after projected onto the probability distribution. In this follow-up, the block described in Figure 7.1 as self-attention, on the decoder side, is often referred to as masked self-attention, since the tokens to the right of the token to be consumed for generation are not available ("masked") for such functionality.

The unloading of weights and biases from pre-trained GPT networks - also preconditioned on a very large corpus of English data [197] in a self-supervised style - to the designed template is seamless since the structure is the same, and there is a perfect correspondence between parameters keys.

Regarding the GPT-specific textual treatment and tokenization procedure, it is based on a Byte-Pair-Encoding or BPE style [199]. The BPE stratagem starts by deeming a pre-tokenizer to divide the word corpus into words. Following the creation of a collection of unique words, the frequency of each one is calculated. Then, a base vocabulary comprising of all symbols found in the unique set of words is developed, and from it, the algorithm learns merge rules to combine two symbols from the base vocabulary in order to construct a new one. It continues to do so until the vocabulary has extended to the required size, being that in this case with the addition of two special tokens that represent the beginning of a sentence and a separation token, it is constituted by 50259 tokens. This strategy leads the tokenizer to address spaces like a token and so thanks to the various frequencies of the various merge operations, a word can be encoded differently depending

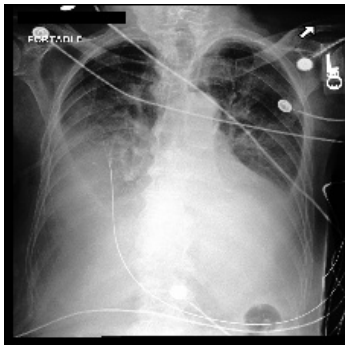
on if it has a space before, after, or neither. It is noteworthy to refer to the fact that, in actuality, the tokenizer uses a byte-level "trick" to force the base vocabulary to be 256 characters long and so guarantee that every base character is present. Hereby it is avoided to consider all Unicode characters as foundational ones, leading to a more compact and less memory-consuming tactic. The 256 bytes basic tokens, the initial special end-of-text token, the symbols learned with 50,000 merges and the 2 new custom tokens make up the 50259 vocabulary size.

The regular GPT2 checkpoint contains 12 decoder blocks, while the DistilGPT2, used in the assessments, configures a distilled version of 6 decoder blocks.

7.2 Materials and Methodology

7.2.1 Data

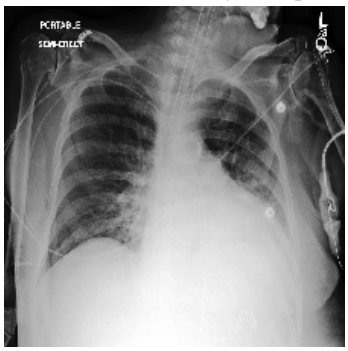
The experiments were conducted over a subset of the *MIMIC-CXR-JPG* dataset that was explored in the *Data* section of chapter 5. This subset consisted of the already leveraged selection of only anterior-posterior (AP) related samples. From it, two work threads were established: one managing the free-radiology texts in their integrity - Set 1 -, and one composed of only clinical report's sections concerning pleural effusion. - Set 2. Data examples for both are depicted in Figure 7.2.



(a) Set 1 - Chest X-ray example.

```
FINAL REPORT EXAMINATION CHEST PORTABLE AP INDICATION ___ year old woman
with heart failure dyspnea Eval for pulmonary edema Eval for pulmonary edema
IMPRESSION Compared to chest radiographs since ___ most recently ___ Large
right and moderate left pleural effusions and severe bibasilar atelectasis
are unchanged Cardiac silhouette is obscured No pneumothorax Pulmonary edema
is mild obscured radiographically by overlying abnormalities
```

(b) Set 1 - Clinical report example.



(c) Set 2 - Chest X-ray example.

```
No interval development of pleural effusion or pneumothorax is seen .
```

(d) Set 2 - Clinical report example.

Figure 7.2: Set 1 example for a Pleural effusion case. Set 2 example for a non Pleural effusion case.

The Set 2 assortment procedure was assembled in an iterative style. With that said, in the first step, every report was segmented into sentences that started with an uppercase letter. For

the second phase, all segments are queried on the presence of the term “pleural effusion”. If the concept is detected, the whole sentence is appointed to a sample. It is important to note that some instances are comprised of more than one sentence as more than one fragment allude to the requisite condition.

For the sake of a more cohesive, compact, and effective experimental setting, especially in terms of resource management, a more specific analysis regarding the number of tokens per report was made. For both sets, making use of the WordPiece algorithm, the demonstration shown in Table 7.1 was produced.

Table 7.1: Token count data analysis for Set 1 and Set 2.

	Set 1 (%)			Set 2 (%)		
	Training	Validation	Test	Training	Validation	Test
Samples with more than 512 tokens	0.12	0.63	0.00	0.00	0.00	0.00
Samples with more than 300 tokens	3.23	2.73	2.17	0.004	0.00	0.00
Samples with more than 100 tokens	71.01	74.16	76.29	3.06	4.12	3.85

The necessity to set a token maximum length for training along with the attempt of avoiding inconvenient and unmanageable deployments resource and time-wise, lead to the decision of settling for a mandatory token length of 300 for Set 1 and 100 for Set 2. The reports that did not reach such size were padded with "ignorable" tokens, and the ones that exceeded the limit were truncated. The judgment did not imply a significant loss of information since only a small percentage of instances render above the stated thresholds.

7.2.2 Method

The proposed methodology regards the exercise of generating explanations from X-ray chest images. For this purpose, at each instance, an image is propagated through the encoder which produces a final image representation. In this process, the embedding attentions are spawned in the encoder self-attention block. Then, the decoder engenders new outputs exploiting the cross attention mechanism that interlinks the attentions of the previous token and the attentions of the visual embedding. Optimization is achieved via calculating the cross-entropy loss between the generated text and the true clinical report example that is intended to be emulated as an explanation. Tactic-wise, teacher forcing dynamics were induced, compelling the replacement of the already produced words with their corresponding real word label. Thus at each mapping, we ensure that the decoder conditions on the correct previous sample.

Aiming to improve the clinical relevance and correctness of the forged explanations, a regularization mechanism was introduced. The intuition is that by utilizing the embeddings of each modality it is possible to extrapolate classification exercises that affect the overall loss. In this respect, three settings are established: explanation generation with text classification regularization, explanation generation with image classification regularization, and explanation generation with image and text classification regularization.

To prompt the textual classification apparatus into the addressed system, some adjustments need to be executed. As depicted in Figure 7.3, the last hidden state of the textual representation (in green), that precedes the probability distribution operation, is pooled and propagated towards binary classification-ready logits.

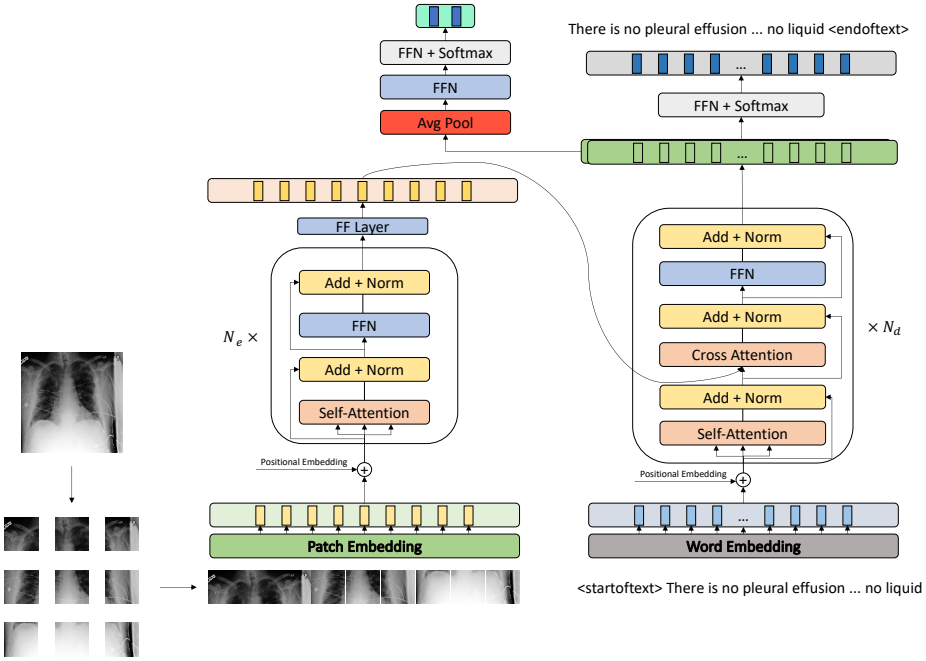


Figure 7.3: Encoder Decoder Model Architecture - Text Classifier regularization.

In such version, the final loss function, represented in Equation 7.1, adds the cross-entropy between the image labels and the predictions produced from the decoder branch to the explanation generation term.

$$\mathscr{L}_{final} = \lambda_{gen}\mathscr{L}_{gen}(y_{true}, y_{gen}) + \lambda_{text}\mathscr{L}_{text}(y_{true}, y_{pred}) \tag{7.1}$$

The image classification regularization stratagem differentiates itself via coupling the classification task to the encoder side. The segment, rendered in Figure 7.4 starts with an average pooling operated on the image embedding last hidden state so that a two-dimension tensor suited label prediction is obtainable.

The final loss for the image-related regularization procedure is represented in Equation 7.2. Once again two clauses arise - the cross-entropy for the text decoding and the cross-entropy image classification assignment.

$$\mathscr{L}_{final} = \lambda_{gen}\mathscr{L}_{gen}(y_{true}, y_{gen}) + \lambda_{img}\mathscr{L}_{img}(y_{true}, y_{pred}) \tag{7.2}$$

Finally, and encompassing both previous proposals, a parallel text and image classification was introduced into the system. In fact, Figure 7.5 describes such combination. Moreover, the

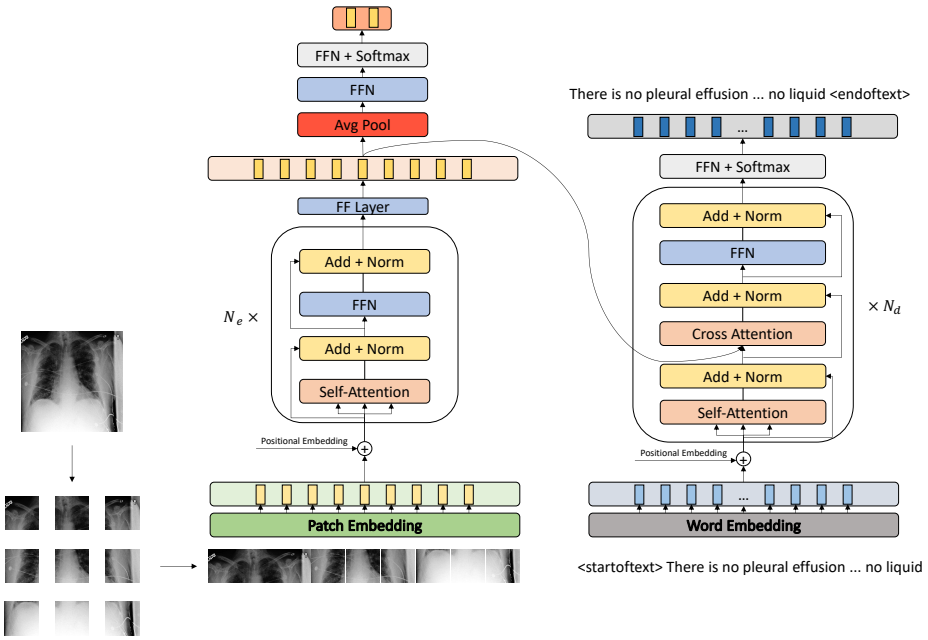


Figure 7.4: Encoder Decoder Model Architecture - Image Classifier regularization.

double regularization-related loss function mirrors such effect with the presence of three terms in Equation 7.3, thus yielding one parcel for the master task and two for the classification functions.

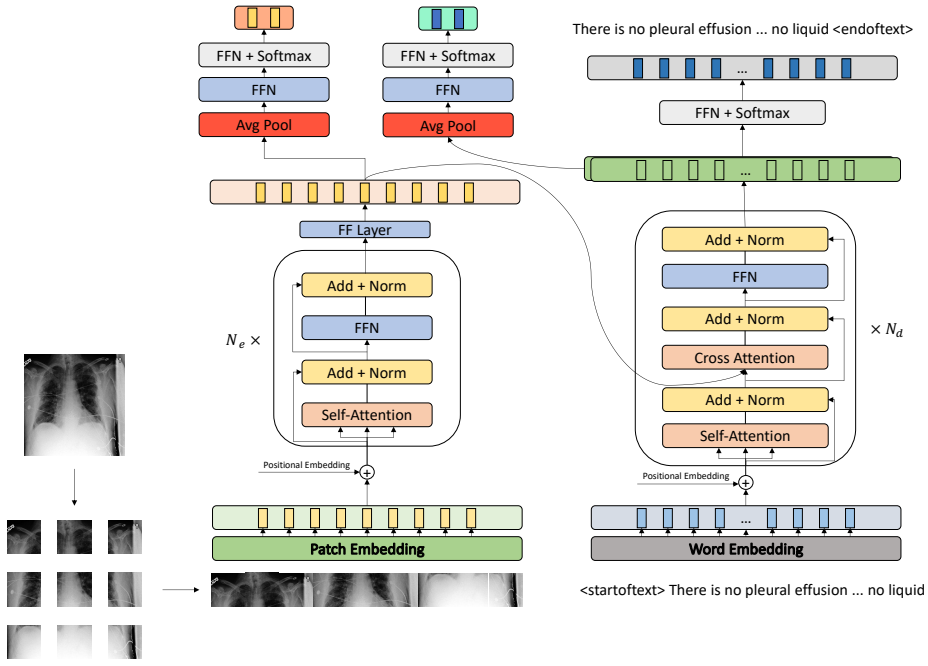


Figure 7.5: Encoder Decoder Model Architecture - Image and Text Classifiers regularization.

$$\mathscr{L}_{final} = \lambda_{gen}\mathscr{L}_{gen}(y_{true}, y_{gen}) + \lambda_{text}\mathscr{L}_{text}(y_{true}, y_{pred}) + \lambda_{img}\mathscr{L}_{img}(y_{true}, y_{pred}) \tag{7.3}$$

7.2.3 Evaluation

The evaluation stage presents itself with a twofold focus: a quantitative component consisting of a set of metrics that are appropriate for text generation tasks such as image captioning and machine translation, and a more subjective dimension based on the interpretation of the clinical coherence and correctness of the generated reports/explanations. The covered metrics are Bilingual Evaluation Understudy (BLEU) [200] in 1-gram, 2-gram, 3-gram and 4-gram style, Metric for Evaluation of Translation with Explicit Ordering (METEOR) [201], Recall-Oriented Understudy for Gisting Evaluation (ROUGE-L) [202], and Consensus-based Image Description Evaluation (CIDEr) [203].

The BLEU metric score is sentence-based. The method involves comparing n-grams in the generated sequence to n-grams in the reference/label one, where each token is considered a 1-gram or unigram. Word order, comprehensibility, or grammatical correctness are not accounted for. METEOR renders a slightly more sophisticated approach. It is averred on the harmonic mean of recall and precision, given that recall is more heavily weighted than precision. It also conveys a broad notion of unigram matching in which the latter can be expressed in surface forms, stemmed forms, and meanings, and applies a fragmentation metric intended to directly portray an order grade. ROUGE-L assesses the longest common subsequence (LCS) between output and label meaning that the longest shared succession of tokens between the two is collected. The premise is that the longer the shared sequence the more similarity is suggested. It addresses recall and precision calculations but for this LCS conjuncture. For CIDEr all words are mapped to their root format and evaluated in a custom n-gram paradigm, as the more common n-grams are lesser weighed. The final score is estimated via the average cosine similarity between the candidate sentence and the label.

Supplementarily, prediction performance, for the configurations with classification segments is also assessed. The accuracy and F1 scores serve as a somewhat of a sanity check, as in an assurance that the network is, at least, fairly concordant with the ground truth label for the addressed condition.

7.3 Results

This time around, Set 1's data partition was settled to 55395 training samples, 476 validation samples, and 966 test samples, while the undertaken handling for the establishment of Set 2 yielded an arrangement of smaller size: 47993 training samples, 412 validation samples, and 856 test samples. This may be because, possibly, some instances whose pleural effusion labels are defined (positive or negative) possess reports without any reference to the aforementioned condition, or that its description has no direct attribution of the expression "pleural effusion". In addition, the label arrangement matches that of chapter 6 for both sets.

The training hyper-parameters are uniform across assays: a 20 epoch training, batch size of 4, a maximum sequence length of 300 tokens for Set 1 and 100 tokens for Set 2, a learning rate of $5x10^{-5}$, and weight decay of $1x10^{-6}$. The chosen optimizer was AdamW [192]. To tackle evaluation, the generation procedure was conducted in batches of 32 with a maximum generation length of 300 (Set 1) or 100 (Set 2) and a minimum generation size of 3.

The results shown in Tables 7.2 to 7.4 reveal a quantitative evaluation of the aforementioned metrics for a number of encoder-decoder configurations, pertaining to Set 1. They also reflect that same evaluation for ablation studies with the text, image, and text and image classification regularization mechanisms. These rely on the DeiT + DistilGPT2 configuration as the baseline, since the latter turned out to show better and more consistent performance in terms of overall metrics.

Regarding Set 2 (Tables 7.5 to 7.7), the quantitative examination is executed only taking into account the ablation studies of the various classification regularization mechanisms. Three versions of the regularization using the textual embedding are assessed: classification segment trained from scratch ("DeiT + DistilGPT2 Text Classification"), classification segment pre-trained in a text classification task ("DeiT + DistilGPT2 Text Classification with pre-trained mlps"), encoder-decoder system pre-trained on the same textual explanation generation task with classification segment trained from scratch ("Pre-trained DeiT + DistilGPT2 Text Classification"). For the image-concerned regularization strategy two renditions are considered: classification head trained from scratch ("DeiT + DistilGPT2 Image Classification"), classification head pre-trained in an image classification task ("DeiT + DistilGPT2 Image Classification with pre-trained mlps"). The dual regularization instance - text and image-related classification - exhibits identical conformations - "DeiT + DistilGPT2 Image and Text Classification" and "DeiT + DistilGPT2 Image and Text Classification with pre-trained mlps ".

A relevant caveat to keep in mind, is that for each metric, the results from Tables 7.2 to 7.7 are averaged on the number of scores obtained per sequence in each set.

For Set 1, the training (Table 7.2), validation (Table 7.3), and test (Table 7.4) outcomes translate the cross-sectional tendency for the ViT + DistilGPT2 and DeiT + DistilGPT2 setups to be more capable of reproducing the intended output. Furthermore, at test time, the image + text classification regularization appears to slightly improve the CIDEr score.

Table 7.2: Set 1 - Explanation Generation Metrics for training set.

	BLEU_1	BLEU_2	BLEU_3	BLEU_4	METEOR	ROUGE-L	CIDEr
ViT + DistilGPT2	**0.273**	**0.174**	0.126	**0.101**	**0.147**	0.262	**0.062**
DeiT + BERT	0.255	0.156	0.107	0.078	0.143	0.239	0.024
DeiT + DistilGPT2	**0.273**	**0.174**	**0.127**	0.1	0.146	**0.265**	0.024
BEiT + DistilGPT2	**0.273**	0.169	0.122	0.097	0.138	0.255	0.049
DeiT + DistilGPT2 Text Classification	0.266	0.167	0.121	0.096	0.142	0.259	0.053
DeiT + DistilGPT2 Image Classification	0.262	0.165	0.119	0.090	0.138	0.260	0.058
DeiT + DistilGPT2 Image and Text Classification	0.269	0.169	0.121	0.096	0.142	0.262	0.054

Table 7.3: Set 1 - Explanation Generation Metrics for validation set.

	BLEU_1	BLEU_2	BLEU_3	BLEU_4	METEOR	ROUGE-L	CIDEr
ViT + DistilGPT2	0.274	0.176	0.128	**0.103**	**0.149**	**0.266**	**0.089**
DeiT + BERT	0.253	0.157	0.107	0.079	0.144	0.241	0.038
DeiT + DistilGPT2	**0,277**	**0,178**	**0.129**	**0.103**	0.148	**0.266**	0.061
BEiT + DistilGPT2	0.265	0.166	0.120	0.096	0.139	0.255	0.039
DeiT + DistilGPT2 Text Classification	0.266	0.168	0.121	0.097	0.145	0.265	0.053
DeiT + DistilGPT2 Image Classification	0.266	0.163	0.115	0.089	0.145	0.260	0.054
DeiT + DistilGPT2 Image and Text Classification	0.265	0.167	0.120	0.095	0.143	0.262	0.055

Table 7.4: Set 1 - Explanation Generation Metrics for test set.

	BLEU_1	BLEU_2	BLEU_3	BLEU_4	METEOR	ROUGE-L	CIDEr
ViT + DistilGPT2	0.247	0.137	0.084	**0.058**	0.131	**0.222**	0.054
DeiT + BERT	0.229	0.126	0.073	0.047	0.126	0.206	0.029
DeiT + DistilGPT2	**0.252**	**0.140**	**0.085**	**0.058**	**0.132**	**0.222**	0.054
BEiT + DistilGPT2	0.248	0.133	0.079	0.054	0.125	0.215	0.046
DeiT + DistilGPT2 Text Classification	0.230	0.126	0.077	0.053	0.127	0.218	0.045
DeiT + DistilGPT2 Image Classification	0.240	0.131	0.079	0.055	0.127	0.219	0.050
DeiT + DistilGPT2 Image and Text Classification	0.242	0.133	0.080	0.055	0.128	**0.222**	**0.057**

In furtherance, for Set 2 there is also a consistent pattern across train (Table 7.5) validation (Table 7.6) and test (Table 7.7), with the highest potential results for BLEU and METEOR being the joint regularization strategy, and for ROUGE-L and CIDEr being the baseline model.

Table 7.5: Set 2 -Explanation Generation Metrics for train set.

	BLEU_1	BLEU_2	BLEU_3	BLEU_4	METEOR	ROUGE-L	CIDEr
DeiT + DistilGPT2	0.088	0.061	0.044	0.030	0.106	**0.304**	**0.664**
DeiT + DistilGPT2 Text Classification	0.090	0.062	0.045	0.030	0.106	0.302	0.652
DeiT + DistilGPT2 Text Classification with pre-trained mlps	0.110	0.074	0.051	0.033	0.108	0.297	0.614
Pre-trained DeiT + DistilGPT2 Text Classification	0.118	0.078	0.052	0.032	0.107	**0.304**	0.656
DeiT + DistilGPT2 Image Classification	0.081	0.056	0.040	0.027	0.103	0.296	0.632
DeiT + DistilGPT2 Image Classification with pre-trained mlps	0.111	0.075	0.050	0.031	**0.113**	0.294	0.610
DeiT + DistilGPT2 Image and Text Classification	**0.130**	**0.087**	**0.061**	**0.039**	**0.113**	0.299	0.592
DeiT + DistilGPT2 Image and Text Classification with pre-trained mlps	0.095	0.066	0.046	0.031	0.107	0.299	0.616

An important contribution towards a genuine consolidation of a proper evaluation, in such a framework, is to precisely review the product explanations. As of today, no effort has been able to engineer metrics that nearly reach or surpass human capacity in encompassing an understanding of what is a good match between two sentences, between two sequences. In that regard, with the infeasibility of including all the test examples, Figure 7.6 serves to depict an example of all

Table 7.6: Set 2 -Explanation Generation Metrics for validation set.

	BLEU_1	BLEU_2	BLEU_3	BLEU_4	METEOR	ROUGE-L	CIDEr
DeiT + DistilGPT2	0.066	0.046	0.032	0.021	0.095	0.282	0.607
DeiT + DistilGPT2 Text Classification	0.069	0.046	0.032	0.020	0.096	0.277	0.624
DeiT + DistilGPT2 Text Classification with pre-trained mlps	0.086	0.059	0.041	0.026	0.100	0.278	0.589
Pre-trained DeiT + DistilGPT2 Text Classification	0.097	0.064	0.042	0.025	0.098	**0.285**	**0.634**
DeiT + DistilGPT2 Image Classification	0.058	0.040	0.029	0.019	0.093	0.273	0.622
DeiT + DistilGPT2 Image Classification with pre-trained mlps	0.080	0.051	0.038	0.022	0.093	0.263	0.555
DeiT + DistilGPT2 Image and Text Classification	**0.107**	**0.070**	**0.047**	**0.029**	**0.103**	0.277	0.604
DeiT + DistilGPT2 Image and Text Classification with pre-trained mlps	0.084	0.057	0.040	0.026	0.101	0.283	0.604

Table 7.7: Set 2 -Explanation Generation Metrics for test set.

	BLEU_1	BLEU_2	BLEU_3	BLEU_4	METEOR	ROUGE-L	CIDEr
DeiT + DistilGPT2	0.098	0.068	0.050	0.034	0.110	0.292	0.542
DeiT + DistilGPT2 Text Classification	0.099	0.070	0.050	0.035	0.111	0.294	0.546
DeiT + DistilGPT2 Text Classification with pre-trained mlps	0.122	0.085	0.061	**0.042**	**0.115**	**0.298**	**0.554**
Pre-trained DeiT + DistilGPT2 Text Classification	**0.137**	**0.091**	0.060	0.037	0.112	0.289	0.476
DeiT + DistilGPT2 Image Classification	0.088	0.062	0.044	0.030	0.106	0.287	0.497
DeiT + DistilGPT2 Image Classification with pre-trained mlps	0.110	0.077	0.054	0.037	0.110	0.290	0.514
DeiT + DistilGPT2 Image and Text Classification	0.132	0.089	**0.062**	0.040	**0.115**	0.289	0.467
DeiT + DistilGPT2 Image and Text Classification with pre-trained mlps	0.102	0.071	0.051	0.034	0.107	0.286	0.486

model's output for Set 1. In it, the allusions to pleural effusion diagnosis are underlined so it is more straightforward to inspect. As beheld in this circumstance, merely the ViT + DistilGPT2 and custom-engineered approaches are clinically concordant with the ground truth.

Figure 7.7 and Figure 7.8 show a group of illustrations for Set 2, where the text classification and text and image classification regularization systems achieve satisfactory deliverables, in contrast to the remaining solutions.

Conversely, Figure 7.9 sustains an example where all four methodologies are concordant with one another but fail to replicate the actual description of the case in hand.

Tables 7.8 and 7.9 illustrate the accuracy and F1 scores for the introduced classification heads, at inference time.

Table 7.8: Set 1 - Accuracy and F1 test scores.

	Accuracy Score	F1 Score
DeiT + DistilGPT2 for Image Classification	0.86542	0.61765
DeiT + DistilGPT2 for Text Classification	0.99362	**0.98832**
DeiT + DistilGPT2 for Image and Text Classification - Image head	0.85921	0.57233
DeiT + DistilGPT2 for Image and Text Classification - Text head	**0.99482**	0.98680

Ground Truth: FINAL REPORT EXAMINATION CHEST PORTABLE AP INDICATION ___ year old man with CHF exacerbation and worsening hypoxia Eval for interval change Eval for interval change IMPRESSION Compared to chest radiographs ___ through ___ Mild pulmonary edema has worsened accompanied by increasing moderate cardiomegaly and increasing small bilateral pleural effusions No pneumothorax

ViT + DistilGPT2: FINAL REPORT EXAMINATION CHEST PORTABLE AP INDICATION ___ year old man with CHF and COPD with worsening hypoxia and tachypnea Eval for worsening pulmonary edema Eval for worsening pulmonary edema IMPRESSION Comparison to ___ The pre existing parenchymal opacities have minimally decreased in extent and severity The lung volumes have increased likely reflecting improved ventilation Moderate cardiomegaly persists left pleural effusion No pneumothorax

DeiT + BERT: final report examination chest portable ap indication _ _ _ year old man with hypoxia and hypoxia eval for interval change eval for interval change impression compared to chest radiographs _ _ _ through _ _ _ mild pulmonary edema has improved substantially since _ _ _ but moderate cardiomegaly is still present and enlarged comparing to last radiographs No pneumothorax

DeiT + DistilGPT2: FINAL REPORT INDICATION ___ year old man with SOB and fever cough PNA TECHNIQUE APsingle view COMPARISON ___ FINDINGS Lines and Tubes Stable right IJ line tip position Lungs Low lung volumes with mild worsening of pulmonary edema Pleura Small left pleural effusion Mediastinum Stable cardiomegaly Bony thorax No change IMPRESSION Mild interval worsening of pulmonary edema with unchanged left pleural effusion and cardiomegaly

BEiT + DistilGPT2: FINAL REPORT EXAMINATION CHEST PORTABLE AP INDICATION ___ year old man with CHF exacerbation and new O2 requirement Please evaluate for pulmonary edema Please evaluate for pulmonary edema IMPRESSION In comparison with the study of ___ there is again enlargement of the cardiac silhouette with pulmonary edema and moderate pleural effusion with compressive atelectasis at the bases In view of the extensive pulmonary changes it would be impossible to exclude superimposed pneumonia in the appropriate clinical setting especially in the absence of a lateral view

DeiT + DistilGPT2 Text Classification: FINAL REPORT EXAMINATION CHEST PORTABLE AP INDICATION ___ year old man with CHF and worsening hypoxia eval for pulm edema eval for pulm edema IMPRESSION In comparison with the study of ___ the cardiac silhouette is more prominent and there is again evidence of elevated pulmonary venous pressure Retrocardiac opacification is consistent with volume loss in the left lower lobe and small pleural effusion

DeiT + DistilGPT2 Image Classification: FINAL REPORT EXAMINATION CHEST PORTABLE AP INDICATION ___ year old man with CHF and COPD exacerbation eval for pulmonary edema eval for pulmonary edema IMPRESSION In comparison with the study of ___ the monitoring and support devices are unchanged Continued enlargement of the cardiac silhouette with pulmonary vascular congestion and bilateral pleural effusions with compressive atelectasis at the bases

DeiT + DistilGPT2 Image and Text Classification: FINAL REPORT EXAMINATION CHEST PORTABLE AP INDICATION ___ year old man with CHF exacerbation and worsening hypoxia eval for worsening pulmonary edema eval for worsening pulmonary edema IMPRESSION In comparison with the study of ___ the cardiac silhouette is more prominent and there is increased indistinctness of pulmonary vessels consistent with elevated pulmonary venous pressure Retrocardiac opacification is consistent with volume loss in the left lower lobe and small pleural effusion

Figure 7.6: Generated explanations example for test set - Set 1.

Table 7.9: Set 2 - Accuracy and F1 test scores.

	Accuracy Score	F1 Score
DeiT + DistilGPT2 for Image Classification	0.85280	0.63372
DeiT + DistilGPT2 for Image Classification with pre-trained mlps	0.87335	0.70422
DeiT + DistilGPT2 for Text Classification	0.98832	0.97175
DeiT + DistilGPT2 for Text Classification with pre-trained mlps	**0.98961**	**0.97287**
Pre-trained DeiT + DistilGPT2 for Text Classification	0.98044	0.96100
DeiT + DistilGPT2 for Image and Text Classification - Image head	0.85514	0.60759
DeiT + DistilGPT2 for Image and Text Classification - Text head	0.98715	0.96906
DeiT + DistilGPT2 for Image and Text Classification with pre-trained mlps - Image head	0.86446	0.62573
DeiT + DistilGPT2 for Image and Text Classification with pre-trained mlps - Text head	0.98882	0.97550

For both sets, text classification clearly displayed better results. This outcome is to be expected on the line of thought that it is more accessible to extract a class from textual data than from image

Ground Truth: Minimal right pleural effusion is noted . Mild to moderate pulmonary edema slightly worse in the interval with trace right pleural effusion and bibasilar atelectasis .

DeiT + DistilGPT2: No pleural effusions.

DeiT + DistilGPT2 Text Classification: Moderate right pleural effusion is larger.

DeiT + DistilGPT2 Image Classification: No pleural effusions.

DeiT + DistilGPT2 Image and Text Classification: There is a moderate right pleural effusion. Moderate right pleural effusion

Figure 7.7: Example 1 of generated explanations for test set - Set 2.

Ground Truth: Possible sml pleural effusions .

DeiT + DistilGPT2: No pleural effusions.

DeiT + DistilGPT2 Text Classification: Small right pleural effusion is unchanged.

DeiT + DistilGPT2 Image Classification: No pleural effusions.

DeiT + DistilGPT2 Image and Text Classification: There is a small right pleural effusion. Small right pleural effusion.

Figure 7.8: Example 2 of generated explanations for test set - Set 2.

data. Such is also proved in both combined regularization settings, where the text-related heads made more effective predictions than the image concerning ones.

7.4 Discussion

In the scope of reproducing quality textual explanations from the chest X-ray images, the experiments were conducted only with Transformer based architectures. Such was mainly due to the combination of three factors: there was limited time for the execution of trial runs, the assembly process of a vision encoder-decoder for full transformer-based models is more convenient granted some tools and frameworks made available in Python libraries such as *transformers* from HuggingFace [204], and some works like [194] report that a full transformer approaches can achieve state-of-the-art performances when it comes to vision encoder-decoder methodologies. Furthermore, as another restriction to be adopted, the ablation study is grounded on a single configuration - DeiT + DistilGPT2 - simply to allow for a more concrete and concise assessment. This encoder and decoder model preference was primarily motivated by the composition's metric appraisal and

Ground Truth: No pleural effusions or pneumothoraces .

DeiT + DistilGPT2: Small right pleural effusion is unchanged.

DeiT + DistilGPT2 Text Classification: Small right pleural effusion is unchanged.

DeiT + DistilGPT2 Image Classification: Small bilateral pleural effusions are present.

DeiT + DistilGPT2 Image and Text Classification: Small right pleural effusion is unchanged.

Figure 7.9: Example 3 of generated explanations for test set - Set 2.

by the fact that the combination requires fewer parameters to optimize, as opposed to other baselines, and therefore is less costly in terms of resources and time.

A precursive thought on which it is worth reflecting before pondering on the scrutiny of the results themselves is the shared resemblance in performance between models that have DistilGPT2 as the decoder. Given that these models use exactly the same decoder and virtually presuppose the same encoder structure - the distinguishing consideration being pre-trained on dissimilar techniques - and provided that fine-tuning is executed on the same downstream task, in equal conditions, it is logical to arrive at comparable outcomes.

In terms of the subjective review of the texts generated at test time, there are several ideas to retain and comparative studies from which some conclusions can be extrapolated. Initially, considering Set 1, and in agreement with what is the reality in the metric investigation, the different variations and combinations of encoder and decoder, that constitute distinct configurations, do not produce very dissimilar outputs, with the exception of the DeiT + BERT conformation, which by having an uncased type decoder and two encoder kind components translates into somewhat divergent end products. Such is verified by the example in Figure 7.6, where, in a singular way, besides being completely under-cased, the explanation does not mention the theme under consideration. Likewise confirming the exclusive lack of effectiveness, the model is also connected to the worst performance across metrics.

In general, these baseline strategies based only on the normal encoder-decoder structure present a fluid discourse but, typically, utterly miss clinical correctness for most of the descriptions. Also, the vast majority of the plotted examples are quite misaligned with information retained in the corresponding images. Additionally and in contrast, oftentimes, the pleural effusion diagnosis is correctly recorded (Table 7.8), however, this is possibly due to the greater number of positive cases in the dataset. Moreover, in this same majority, the concordant indication at the label level does not seem to maintain rigor when identifying characteristics associated with, for example, spatial location or state of the condition in focus. In Figure 7.6 such is described when the ground truth report reveals that there are "small bilateral pleural effusions" and generated texts defend conclusions like "unchanged left pleural effusion" or " moderate pleural effusion". Ultimately, akin

ineffectiveness is aligned with the lack of clinical relevance of observations unrelated to pleural effusion, as well. Without any reinforcement or greater induced weight on the clinical perspective, the high frequency of reference to pleural effusion introduced in training turns out not to be very effective on its own.

A supplementary aspect concerning the exercise on Set 1 is the large repetition of sentence beginnings is evident throughout test examples, which is natural considering the fact that throughout the dataset sequences like "FINAL REPORT EXAMINATION CHEST PORTABLE AP INDICATION ____ year old man/woman" are extremely recurrent. Other very usual terms are also countlessly brought about at generation time, namely "compressive atelectasis at the bases", "worsening hypoxia" and "unchanged".

The submission of the regularization techniques to the system presupposes some relevant aspects to be detailed. Firstly, and for the three forms of regularization there does not appear to exist a comprehensive clinical fidelity improvement on all fronts. When it comes to the totality of the content for each text, the inconsistencies and incongruities remain untouched. Anyhow, with respect to purely-associated pleural effusion subject matter, there are some occurrences, like the ones present in the elated figure, which seem to be vaguely more correct or accurate when compared to the baseline counterparts. This event sounds even more feasible when considering that the operational regularization adjustment is affiliated with a pleural effusion binary classification task. Moreover, across the three mechanisms, there is not a great deal of discrepancy, meaning that in certain examples the image classification regularization entails the better explanation, in others the text classification regularization or the combined one serves as a more viable option. In Figure 7.6, for example, all three introduce some relevant nuance, but they also seem to lead to similar errors in other readings. On that note, none look to be particularly more efficient than the other, as one might expect, since the textual regularization or the joint approach hold information that is already treated or processed by human understanding. In fact, the latter two are almost always in concordance with one another which discloses a stronger influence from the decoder side. It should also be noted that the pertained sequences maintain equivalent repetition tendencies and dispositions.

The results in Set 2 exhibit greater consistency of clinical correctness all across. Thus, it appears that reducing the dataset to only pleural effusion-related references translates into more effective clinical insight, especially in the model's output ability to match the desired label. Nevertheless, there is also a widespread improvement in identifying, for example, the location ("left", "right" and "bilateral") and/or severity ("small", "moderate", "large") when compared to the observations created in Set 1. One pattern that persists is that the training data, by virtue of entailing a somewhat standardized and consequently repetitive writing methodology induces this nature even more visibly, in the test outcomes. Thus, there are pieces of text such as "No pleural effusions.", "pleural effusions are presumed", or "pleural effusion is unchanged" that appear over and over again throughout the various produced examples.

Focusing on the influence exerted by the regularization techniques, for Set 2, there are some notes to be retained. Overall, once again, their weight does not seem to have much impact as

in most cases there seems to be accordance across all methodologies in the identification of the presence or absence and characteristics of pleural effusion. Such is proved by the consistent results exhibited across Table 7.7. Nevertheless, the regularised impression does seem to be greater than that witnessed in the peer set of data. The examples of Figure 7.7 and Figure 7.8 show that the effect of text classification and both text and image classification correct the poor assessment of the baseline approach, even if the description is not completely accurate. Additionally, the text classification and text and image classification regularization approaches' results are even more almost exclusively in agreement, while image classification and the regular model are also typically aligned. This reflects a further stronger impact of the text-driven regularization than the visual counterpart. On another note, Table 7.9 demonstrates, comprehensively, a very strong alignment between the classification outputs and the ground truth labels, and besides that, it also indicates a more consistent text-driven drill.

Metric-wise, the major highlight lies in the fact that there are clear differentiating patterns between the range of metric values from Set 1 to Set 2. If for BLEU and METEOR scores Set 1's experiments lead to better outcomes, such a development follows a reverse course for the more sophisticated metrics - ROUGE-L and CIDEr. Said event, essentially allows for two key readings. Knowing that the BLEU and METEOR metrics mostly contemplate the broad precision and recall of instances per sequence, over contextual insertion or textual significance, Set 1's illustrations end up being favored due to the conjugation of the specific nature and domain in which the dataset is inserted. This means that larger-sized samples - upon a reduced-range vocabulary range and longer sequences - afford a higher probability of yielding words that are contained in the referenced sequence and therefore achieving better performances in such metrics. In the opposite direction, ROUGE-L and CIDEr favor Set 2's samples, because there exists a higher quantity of sequences that are fully concordant with the label sequences, thus, in actuality, promoting similarity in terms of content and sentence order. The nature of shorter sequences in Set 2 plays a substantial role in this distinguishing factor.

A final remarkable trend is the consistency with which the collective regularization delivers better outcomes for "less relevant" metrics in Set 2 and worse results for the remaining ones. Apparently, the effect under review produced a greater affinity between generated sequence and ground truth one only at the word level.

To conclude the discussion it is appropriate to enunciate some proposals that were also thought of in the scope of the experiments, either as complementary ideas, or alternative solutions to overcome the registered limitations. In that regard, ideas such as applying self-critical training [205] in order to directly optimize the metrics at issue could guide toward major improvements, since in theory, the end products are bound to be rather more correspondent to the underlying label clinical reports. In fact, an attempt at emulating the aforesaid technique was executed, but it did not uncover much success due to the very difficult managing of the required resources/time allocation. Another idea that may be useful to consider in the future would be to incorporate a contrastive loss into the system, as explored in [206]. The paradigm of minimizing the distance between similar examples and maximizing it for dissimilar ones may be a valuable addition to the already explored

regularizing dynamics.

7.5 Multimodal Explanation System

The design of an inference-ready fully integrated multimodal explanation system is presented in Figure 7.10.

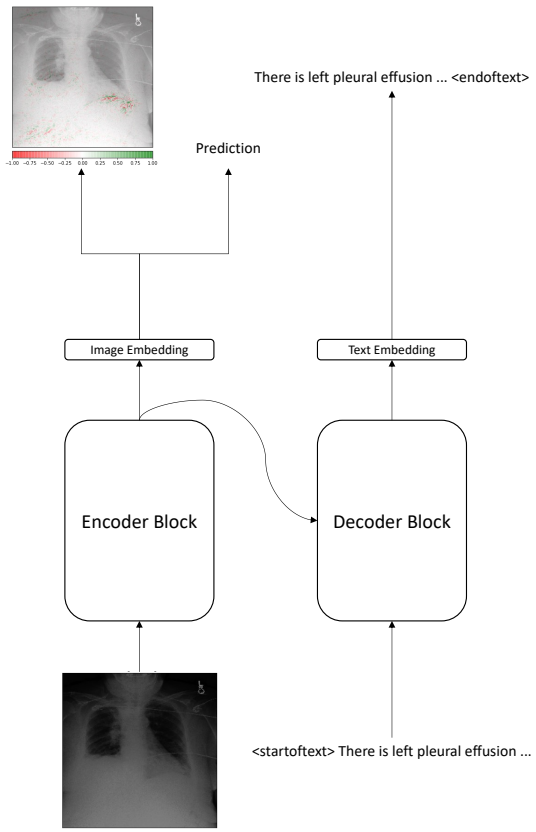


Figure 7.10: Full System for generating multimodal explanations.

Upon receiving a chest x-ray image, on the encoder side, a saliency map is rendered in compliance with certain a prediction. On the decoder side, with the flow of information coming from the encoder, the report-like explanation is generated token by token.

It is substantial to note that the encoder block should be pre-trained according to the methodology presented in chapter 6. In this way, it is possible to transport the visual saliency map quality enhancing characteristics to the desired framework, while maintaining the proposed encoder-decoder structure.

Furthermore, some fine-tuning tactics over the global structure could be tested. For instance, one could witness the effect that fine-tuning, at the same time, the classification and text generation

tasks would have on the visual explanations. On that matter, one could even, for example, freeze the parameters of the entirety of the encoder block and only retrain the remaining components.

Chapter 8

Conclusions

The dissertation proposes to investigate and prospect multimodal solutions (textual and visual) for generating explanations in a medical or clinical setting. In this manner, there is an attempt to promote the complementarity of explanations and increase diversity and, consequently, the likelihood of acceptance by stakeholders. To attain such a proposal, two aspects are required: guarantee the feasibility of unimodal tasks, where both visual and textual explainability techniques can minimally fulfill what is expected of them, being coherent and clearly interpretable; ensure forms of multimodal interaction in which the sharing and/or transfer of knowledge between multimodal approaches is maximized with a perspective of achieving quality compliance for the explanations.

Considering the diagnostic domains or assignments where visual evidence is normally relied upon, the relevance of establishing methodologies for merging visual saliency maps with semantic concepts incorporated in textual explanations is substantial. In this vein, the fundamentally innovative aspect is present in the multiple creations of explanations. That is, granted the example of a chest X-ray, it should be possible to obtain a decision supported by the demonstration of the pixels considered in it, and by a free-text radiology report style description that complements the previous explanation.

The purely unimodal work and the textual regularization experiments translated into a initial clear evolution towards the final and decisive goal of the thesis, given that the carried out experiments explore various algorithms for the accession of saliency maps, and mainly study multimodal learning approaches. The analysis of techniques for generating visual maps serves, precisely, to be better able to establish robust explainability strategies on their own. The multimodal research, on the other hand, motivates the importance of knowledge exchange between representations of different modalities. The covered coordinated approach has proved that the interplay between embeddings can be essential for the improvement of explanation mechanisms, and it is in this sense and perspective of inter-modality linkage that a truly effective diversification in explanations can be implied.

The work in chapter 7 attempts at optimizing the representation inter-operability for a textual generation task, with the intention to deepen and further test the consideration of multimodal learning. It mirrors an effort to fulfill the remaining dissertation objectives.

A more general and encompassing appreciation of what was the work developed in the scope of the dissertation needs to point out that the potential demonstrated by the multimodal regularization approach clearly showed more satisfactory and evident results in the image classification exercise than in the explanation generation experiments. Regardless, one must reflect and realize that the dissertation work constitutes a foundation for more extensive studies and examinations to be had. There are many paths and perspectives to be researched and investigated in order to achieve quality and robust multimodal clinical explanations. Moreover, the employed methodology presented some encouraging signs that result in redoubled motivation to extend and expand the project.

Another discussion to be had with respect to Chapter 7's assay, in particular, concerns the type of targeted explanations and their practical viability. In the defined setting, the conceived texts are extremely dependent on the type of clinical data used in the training process, hence manifesting a report-based quality. The explanation is conditioned on the structure of the ground truth report, on the presence or absence of a template or systematized way of writing, and mainly on the kind of reasoning and conclusions that are demonstrated. In this sense, and in light of the exerted dataset characteristics, it is of accessible understanding that the model's output follows a demonstrative tendency, based on illustrations of condition severity and spatial location instead of cause-effect accounts. In any case, gauging their essence and comparing them with what is established in visual salience maps - considered in Chapter 6 - a correspondence of similar typology regarding the techniques' conveyed information is discernible. The validation of explanations of this kind is reliant on their capacity to highlight instances whose influence or contribution is significant for a model to make a certain decision or perform a certain task. In this work such viewpoint is accomplished for two different modalities, thus there exists framing for combinations of explanations and, consequently expanding the probability of acceptance from stakeholders.

I would further like to point out that it would have been valuable to actually deploy a complete multimodal generation system, not only to validate and formalize all the experimental efforts but also to expand on other exploratory perspectives.

Finalizing this set of concluding remarks, it should be highlighted that the efforts of this thesis aim to make a valuable contribution to the field of XAI, and more specifically to its applicability to healthcare services. This involvement is placed within the scope of the defense of a multiplicity of explanatory formats for systems in a clinical environment. In that fashion, a variety of explanations does not only help to clarify the role of models and neural networks in their tasks but also, allows for greater trust and acceptance. Therefore, complementarity and diversity of explanatory approaches are vital for the real employment of AI solutions in medicine.

References

[1] Talha Khan Burki. The role of AI in diagnosing lung diseases. *The Lancet Respiratory Medicine*, 7(12):1015–1016, December 2019. URL: https://doi.org/10.1016/s2213-2600(19)30331-5, doi:10.1016/s2213-2600(19)30331-5.

[2] K. V. Greeshma, J. Viji Gripsy, Patrick Siarry, M.A. Jabbar, Rajanikanth Aluvalu, Ajith Abraham, and Ana Madureira. *A Review on Classification and Retrieval of Biomedical Images Using Artificial Intelligence*, pages 47–66. Springer International Publishing, Cham, 2021. URL: https://doi.org/10.1007/978-3-030-75220-0_3, doi:10.1007/978-3-030-75220-0_3.

[3] Rong Liu, Feng Mai, Zhe Shan, and Ying Wu. Predicting shareholder litigation on insider trading from financial text: An interpretable deep learning approach. *Information Management*, 57(8):103387, 2020. URL: https://www.sciencedirect.com/science/article/pii/S0378720620303256, doi:https://doi.org/10.1016/j.im.2020.103387.

[4] Erico Tjoa and Cuntai Guan. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–21, 2020. doi:10.1109/TNNLS.2020.3027314.

[5] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

[6] Wilson Silva, Kelwin Fernandes, Maria J Cardoso, and Jaime S Cardoso. Towards complementary explanations using deep neural networks. In *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*, pages 133–140. Springer, 2018.

[7] Maximilian Alber, Sebastian Lapuschkin, Philipp Seegerer, Miriam Hägele, Kristof T. Schütt, Grégoire Montavon, Wojciech Samek, Klaus-Robert Müller, Sven Dähne, and Pieter-Jan Kindermans. innvestigate neural networks!, 2018. arXiv:1808.04260.

[8] Wilson Silva, Alexander Poellinger, Jaime S. Cardoso, and Mauricio Reyes. Interpretability-guided content-based medical image retrieval. In Anne L. Martel, Purang Abolmaesumi, Danail Stoyanov, Diana Mateus, Maria A. Zuluaga, S. Kevin Zhou, Daniel Racoceanu, and Leo Joskowicz, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 305–314, Cham, 2020. Springer International Publishing.

[9] Diogo Mata, Wilson Silva, and Jaime S. Cardoso. Increased robustness in chest x-ray classification through clinical report-driven regularization. In Armando J. Pinho, Petia Georgieva,

Luís F. Teixeira, and Joan Andreu Sánchez, editors, *Pattern Recognition and Image Analysis*, pages 119–128, Cham, 2022. Springer International Publishing.

[10] Kruttika Jain and Shivani Kaushal. A comparative study of machine learning and deep learning techniques for sentiment analysis. In *2018 7th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, pages 483–487, 2018. doi:10.1109/ICRITO.2018.8748793.

[11] Cannannore Nidhi Kamath, Syed Saqib Bukhari, and Andreas Dengel. Comparative study between traditional machine learning and deep learning approaches for text classification. In *Proceedings of the ACM Symposium on Document Engineering 2018*, DocEng '18, New York, NY, USA, 2018. Association for Computing Machinery. URL: https://doi.org/10.1145/3209280.3209526, doi:10.1145/3209280.3209526.

[12] Chigozie Nwankpa, Winifred Ijomah, Anthony Gachagan, and Stephen Marshall. Activation functions: Comparison of trends in practice and research for deep learning, 2018. URL: https://arxiv.org/abs/1811.03378, doi:10.48550/ARXIV.1811.03378.

[13] Katarzyna Janocha and Wojciech Czarnecki. On loss functions for deep neural networks in classification. *Schedae Informaticae*, 25, 02 2017. doi:10.4467/20838476SI.16.004.6185.

[14] Dami Choi, Christopher J. Shallue, Zachary Nado, Jaehoon Lee, Chris J. Maddison, and George E. Dahl. On empirical comparisons of optimizers for deep learning, 2020. arXiv:1910.05446.

[15] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015. arXiv:1409.1556.

[16] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks, 2018. arXiv:1608.06993.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. arXiv:1512.03385.

[18] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015. arXiv:1505.04597.

[19] J Samuel Manoharan. Capsule network algorithm for performance optimization of text classification. *Journal of Soft Computing Paradigm (JSCP)*, 3(01):1–9, 2021.

[20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2020. URL: https://arxiv.org/abs/2010.11929, doi:10.48550/ARXIV.2010.11929.

[21] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. arXiv:1406.2661.

[22] Sethunya R Joseph, Hlomani Hlomani, Keletso Letsholo, Freeson Kaniwa, and Kutlwano Sedimo. Natural language processing: A review. *International Journal of Research in Engineering and Applied Sciences*, 6(3):207–210, 2016.

[23] Gobinda G Chowdhury. Natural language processing. *Annual review of information science and technology*, 37(1):51–89, 2003.

[24] Antoine Ly, Benno Uthayasooriyar, and Tingting Wang. A survey on natural language processing (nlp) and applications in insurance, 2020. arXiv:2010.00462.

[25] Shang-Yu Su, Chao-Wei Huang, and Yun-Nung Chen. Dual supervised learning for natural language understanding and generation, 2020. arXiv:1905.06196.

[26] Madeleine Bates. Models of natural language understanding. *Proceedings of the National Academy of Sciences*, 92(22):9977–9982, 1995.

[27] Girija Godbole. Natural language generation. *Theoretical Issues in Natural Language Processing*, 2018.

[28] Kowsari, Jafari Meimandi, Heidarysafa, Mendu, Barnes, and Brown. Text classification algorithms: A survey. *Information*, 10(4):150, Apr 2019. URL: http://dx.doi.org/10.3390/info10040150, doi:10.3390/info10040150.

[29] Dan Otter, Julian Richard Medina, and Jugal Kumar Kalita. A survey of the usages of deep learning for natural language processing. *IEEE Transactions on Neural Networks and Learning Systems*, 32:604–624, 2021.

[30] Rahul, Surabhi Adhikari, and Monika. Nlp based machine learning approaches for text summarization. In *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, pages 535–538, 2020. doi:10.1109/ICCMC48092.2020.ICCMC-00099.

[31] Zahra Abbasiantaeb and Saeedeh Momtazi. Text-based question answering from information retrieval and deep neural network perspectives: A survey, 2020. arXiv:2002.06612.

[32] Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tie-Yan Liu. Incorporating bert into neural machine translation. *arXiv preprint arXiv:2002.06823*, 2020.

[33] Mickel Hoang, Oskar Alija Bihorac, and Jacobo Rouces. Aspect-based sentiment analysis using bert. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 187–196, 2019.

[34] Zihan Liu, Feijun Jiang, Yuxiang Hu, Chen Shi, and Pascale Fung. Ner-bert: A pre-trained model for low-resource entity tagging, 2021. arXiv:2112.00405.

[35] Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara. From show to tell: A survey on deep learning-based image captioning, 2021. arXiv:2107.06912.

[36] Alex Sherstinsky. Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. *Physica D: Nonlinear Phenomena*, 404:132306, Mar 2020. URL: http://dx.doi.org/10.1016/j.physd.2019.132306, doi:10.1016/j.physd.2019.132306.

[37] Ralf C. Staudemeyer and Eric Rothstein Morris. Understanding lstm – a tutorial into long short-term memory recurrent neural networks, 2019. arXiv:1909.09586.

[38] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation, 2014. arXiv:1406.1078.

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017. URL: http://arxiv.org/abs/1706.03762, arXiv:1706.03762.

[40] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL: http://arxiv.org/abs/1810.04805, arXiv:1810.04805.

[41] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018.

[42] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation, 2013. URL: https://arxiv.org/abs/1311.2524, doi:10.48550/ARXIV.1311.2524.

[43] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection, 2015. URL: https://arxiv.org/abs/1506.02640, doi:10.48550/ARXIV.1506.02640.

[44] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan, 2019. URL: https://arxiv.org/abs/1912.04958, doi:10.48550/ARXIV.1912.04958.

[45] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis, 2020. URL: https://arxiv.org/abs/2012.09841, doi:10.48550/ARXIV.2012.09841.

[46] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019. URL: https://arxiv.org/abs/1907.11692, doi:10.48550/ARXIV.1907.11692.

[47] Jing Chen, Chenhui Wang, Kejun Wang, Chaoqun Yin, Cong Zhao, Tao Xu, Xinyi Zhang, Ziqiang Huang, Meichen Liu, and Tao Yang. Heu emotion: a large-scale database for multimodal emotion recognition in the wild. *Neural Computing and Applications*, Jan 2021. URL: http://dx.doi.org/10.1007/s00521-020-05616-w, doi:10.1007/s00521-020-05616-w.

[48] Mehrdad Hosseinzadeh and Yang Wang. Video captioning of future frames. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 980–989, January 2021.

[49] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Deep audio-visual speech recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[50] Rafael Souza, André Freire, Thiago Teixeira, George Teodoro, and Renato Ferreira. Online multimedia retrieval on cpu–gpu platforms with adaptive work partition. *Journal of Parallel and Distributed Computing*, 148:31–45, 02 2021. doi:10.1016/j.jpdc.2020.10.001.

[51] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

[52] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives, 2014. arXiv:1206.5538.

[53] Nitish Srivastava, Ruslan Salakhutdinov, et al. Multimodal learning with deep boltzmann machines. In *NIPS*, volume 1, page 2. Citeseer, 2012.

[54] Wenzhong Guo, Jianwen Wang, and Shiping Wang. Deep multimodal representation learning: A survey. *IEEE Access*, 7:63373–63394, 2019. doi:10.1109/ACCESS.2019.2916887.

[55] Jing Gao, Peng Li, Zhikui Chen, and Jianing Zhang. A survey on deep learning for multimodal data fusion. *Neural Computation*, 32(5):829–864, 2020.

[56] Kuan Liu, Yanen Li, Ning Xu, and Prem Natarajan. Learn to combine modalities in multimodal deep learning, 2018. arXiv:1805.11730.

[57] Roman Ilin, Thomas Watson, and Robert Kozma. Abstraction hierarchy in deep learning neural networks. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 768–774. IEEE, 2017.

[58] Nicolas Audebert, Catherine Herold, Kuider Slimani, and Cédric Vidal. Multimodal deep networks for text and image-based document classification, 2019. arXiv:1907.06370.

[59] Yue Gu, Xinyu Li, Shuhong Chen, Jianyu Zhang, and Ivan Marsic. Speech intention classification with multimodal deep learning. In *Canadian conference on artificial intelligence*, pages 260–271. Springer, 2017.

[60] Kyle Ross, Paul Hungler, and Ali Etemad. Unsupervised multi-modal representation learning for affective computing with multi-corpus wearable data, 2020. arXiv:2008.10726.

[61] Leanne Nortje and Herman Kamper. Unsupervised vs. transfer learning for multimodal one-shot matching of speech and images, 2020. arXiv:2008.06258.

[62] Yusuf Aytar, Lluis Castrejon, Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Cross-modal scene networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(10):2303–2314, oct 2018. URL: https://doi.org/10.1109/TPAMI.2017.2753232, doi:10.1109/TPAMI.2017.2753232.

[63] Jongmin Yu, Kin Choong Yow, and Moongu Jeon. Joint representation learning of appearance and motion for abnormal event detection. *Machine Vision and Applications*, 29(7):1157–1170, 2018.

[64] James J Deng and Clement HC Leung. Towards learning a joint representation from transformer in multimodal emotion recognition. In *International Conference on Brain Informatics*, pages 179–188. Springer, 2021.

[65] Yanan Liu, Xiaoqing Feng, and Zhiguang Zhou. Multimodal video classification with stacked contractive autoencoders. *Signal Process.*, 120(C):761–766, mar 2016. URL: https://doi.org/10.1016/j.sigpro.2015.01.001, doi:10.1016/j.sigpro.2015.01.001.

[66] Yonghao He, Shiming Xiang, Cuicui Kang, J. J. Wang, and Chunhong Pan. Cross-modal retrieval via deep and bidirectional representation learning. *IEEE Transactions on Multimedia*, 18:1363–1377, 2016.

[67] Yuxin Peng, Jinwei Qi, Xin Huang, and Yuxin Yuan. Ccl: Cross-modal correlation learning with multi-grained fusion by hierarchical network, 2017. arXiv:1704.02116.

[68] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc' Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL: https://proceedings.neurips.cc/paper/2013/file/7cce53cf90577442771720a370c3c723-Paper.pdf.

[69] Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. Hierarchical multimodal lstm for dense visual-semantic embedding. In *Proceedings of the IEEE international conference on computer vision*, pages 1881–1889, 2017.

[70] Spandana Gella, Rico Sennrich, Frank Keller, and Mirella Lapata. Image pivoting for learning multilingual multimodal representations. *arXiv preprint arXiv:1707.07601*, 2017.

[71] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions, 2015. arXiv:1412.2306.

[72] Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models, 2014. arXiv:1411.2539.

[73] Sandeep Singh Sengar, U Hariharan, and K Rajkumar. Multimodal biometric authentication system using deep learning method. In *2020 International Conference on Emerging Smart Computing and Informatics (ESCI)*, pages 309–312. IEEE, 2020.

[74] Yingwei Pan, Tao Mei, Ting Yao, Houqiang Li, and Yong Rui. Jointly modeling embedding and translation to bridge video and language, 2015. arXiv:1505.01861.

[75] Venice Erin Liong, Jiwen Lu, Yap-Peng Tan, and Jie Zhou. Deep coupled metric learning for cross-modal matching. *IEEE Transactions on Multimedia*, 19(6):1234–1244, 2016.

[76] Yuxin Peng, Jinwei Qi, and Yuxin Yuan. Modality-specific cross-modal similarity measurement with recurrent attention network, 2017. arXiv:1708.04776.

[77] Jiamei Fu, Dongyu She, Xingxu Yao, Yuxiang Zhang, and Jufeng Yang. Deep coordinated textual and visual network for sentiment-oriented cross-modal retrieval. In *Pacific Rim International Conference on Artificial Intelligence*, pages 684–696. Springer, 2018.

[78] Samuel G Finlayson, Matthew BA McDermott, Alex V Pickering, Scott L Lipnick, William Yuan, and Isaac S Kohane. Approaching small molecule prioritization as a cross-modal information retrieval task through coordinated representation learning. *arXiv preprint arXiv:1911.10241*, 2019.

[79] Fangxiang Feng, Xiaojie Wang, and Ruifan Li. Cross-modal retrieval with correspondence autoencoder. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 7–16, 2014.

[80] Richard Socher, Andrej Karpathy, Quoc V. Le, Christopher D. Manning, and Andrew Y. Ng. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2:207–218, 2014. URL: https://aclanthology.org/Q14-1017, doi:10.1162/tacl_a_00177.

[81] Xin Huang and Yuxin Peng. Deep cross-media knowledge transfer, 2018. arXiv:1803.03777.

[82] Chongqing Chen, Dezhi Han, and Jun Wang. Multimodal encoder-decoder attention networks for visual question answering. *IEEE Access*, 8:35662–35671, 2020. doi:10.1109/ACCESS.2020.2975093.

[83] Jeff Donahue, Lisa Anne Hendricks, Marcus Rohrbach, Subhashini Venugopalan, Sergio Guadarrama, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description, 2016. arXiv:1411.4389.

[84] Haiyan Li and Dezhi Han. Multimodal encoders and decoders with gate attention for visual question answering. *Computer Science and Information Systems*, (00):32–32, 2021.

[85] Noam Mor, Lior Wolf, Adam Polyak, and Yaniv Taigman. A universal music translation network, 2018. arXiv:1805.07848.

[86] Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. Image captioning: Transforming objects into words. *arXiv preprint arXiv:1906.05963*, 2019.

[87] Xinpeng Chen, Lin Ma, Wenhao Jiang, Jian Yao, and Wei Liu. Regularizing rnns for caption generation by reconstructing the past with the present. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7995–8003, 2018.

[88] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6023–6032, 2019.

[89] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis, 2016. arXiv:1605.05396.

[90] Jiuxiang Gu, Jianfei Cai, Shafiq Joty, Li Niu, and Gang Wang. Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models, 2018. arXiv:1711.06420.

[91] Divya Saxena and Jiannong Cao. Generative adversarial networks (gans): Challenges, solutions, and future directions, 2020. arXiv:2005.00065.

[92] Xiaoran Chen, Nick Pawlowski, Martin Rajchl, Ben Glocker, and Ender Konukoglu. Deep generative models in the real-world: An open challenge from medical imaging. *arXiv preprint arXiv:1806.05452*, 2018.

[93] Harshit Parikh, Harsh Sawant, Bhautik Parmar, Rahul Shah, Santosh Chapaneri, and Deepak Jayaswal. Encoder-decoder architecture for image caption generation. In *2020 3rd International Conference on Communication System, Computing and IT Applications (CSCITA)*, pages 174–179, 2020. doi:10.1109/CSCITA47329.2020.9137802.

[94] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis, 2021. arXiv:2111.14822.

[95] Shaoxiang Chen, Ting Yao, and Yu-Gang Jiang. Deep learning for video captioning: A review. In *IJCAI*, pages 6283–6290, 2019.

[96] Luis Enrique Sucar. Probabilistic graphical models. *Advances in Computer Vision and Pattern Recognition. London: Springer London. doi*, 10(978):1, 2015.

[97] Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800, 2002.

[98] Nitish Srivastava and Ruslan Salakhutdinov. Learning representations for multimodal data with deep belief nets. In *International conference on machine learning workshop*, volume 79, page 3, 2012.

[99] Nitish Srivastava, Ruslan Salakhutdinov, et al. Multimodal learning with deep boltzmann machines. *J. Mach. Learn. Res.*, 15(1):2949–2980, 2014.

[100] Harold Hotelling. Relations between two sets of variates. In *Breakthroughs in statistics*, pages 162–190. Springer, 1992.

[101] Weiran Wang, Raman Arora, Karen Livescu, and Jeff A. Bilmes. Unsupervised learning of acoustic features via deep canonical correlation analysis. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4590–4594, 2015. doi:10.1109/ICASSP.2015.7178840.

[102] Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. On deep multi-view representation learning: Objectives and optimization, 2016. arXiv:1602.01024.

[103] Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. Unsupervised learning of acoustic features via deep canonical correlation analysis. pages 4590–4594, 04 2015. doi:10.1109/ICASSP.2015.7178840.

[104] Wei Liu, Jie-Lin Qiu, Wei-Long Zheng, and Bao-Liang Lu. Multimodal emotion recognition using deep canonical correlation analysis. *arXiv preprint arXiv:1908.05349*, 2019.

[105] Liang Dai, Guodong Du, Jia Zhang, Candong Li, Rong Wei, and Shaozi Li. Joint multilabel classification and feature selection based on deep canonical correlation analysis. *Concurrency and Computation: Practice and Experience*, 32(22):e5864, 2020.

[106] Yuki Takashima, Tetsuya Takiguchi, Yasuo Ariki, and Kiyohiro Omori. Audio-visual speech recognition for a person with severe hearing loss using deep canonical correlation analysis. In *Proc. 1st Int. Workshop Challenges Hearing Assistive Technol.*, pages 77–81, 2017.

[107] Sharmeen M Saleem Abdullah Abdullah, Siddeeq Y Ameen Ameen, Mohammed AM Sadeeq, and Subhi Zeebaree. Multimodal emotion recognition using deep learning. *Journal of Applied Science and Technology Trends*, 2(02):52–58, 2021.

[108] Yves M. Galvão, Janderson Ferreira, Vinícius A. Albuquerque, Pablo Barros, and Bruno J.T. Fernandes. A multimodal approach using deep learning for fall detection. *Expert Systems with Applications*, 168:114226, 2021. URL: https://www.sciencedirect.com/science/article/pii/S0957417420309489, doi:https://doi.org/10.1016/j.eswa.2020.114226.

[109] Danushka Madhuranga, Rivindu Madushan, Chathuranga Siriwardane, and Kutila Gunasekera. Real-time multimodal adl recognition using convolution neural networks. *The Visual Computer*, 37(6):1263–1276, 2021.

[110] Khaled Bayoudh, Raja Knani, Fayçal Hamdaoui, and Mtibaa Abdellatif. A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets. *The Visual Computer*, 06 2021. doi:10.1007/s00371-021-02166-7.

[111] Dor Bank, Noam Koenigstein, and Raja Giryes. Autoencoders, 2021. arXiv:2003.05991.

[112] Gaurav Bhatt, Piyush Jha, and Balasubramanian Raman. Representation learning using step-based deep multi-modal autoencoders. *Pattern Recognition*, 95:12–23, 2019. URL: https://www.sciencedirect.com/science/article/pii/S0031320319302146, doi:https://doi.org/10.1016/j.patcog.2019.05.032.

[113] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *ICML*, 2011.

[114] Ahmed Ben Said, Amr Mohamed, Tarek Elfouly, Khaled Harras, and Z Jane Wang. Multimodal deep learning approach for joint eeg-emg data compression and classification. In *2017 IEEE wireless communications and networking conference (WCNC)*, pages 1–6. IEEE, 2017.

[115] Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image generation from scene graphs, 2018. arXiv:1804.01622.

[116] Eunyeong Jeon, Kunhee Kim, and Daijin Kim. Fa-gan: Feature-aware gan for text to image synthesis, 2021. arXiv:2109.00907.

[117] Xing Xu, Li He, Huimin Lu, Lianli Gao, and Yanli Ji. Deep adversarial metric learning for cross-modal retrieval. *World Wide Web*, 22(2):657–672, 2019.

[118] Jian Zhang, Yuxin Peng, and Mingkuan Yuan. Unsupervised generative adversarial cross-modal hashing, 2017. arXiv:1712.00358.

[119] Xiao Bai, Xiang Wang, Xianglong Liu, Qiang Liu, Jingkuan Song, Nicu Sebe, and Been Kim. Explainable deep learning for efficient and robust pattern recognition: A survey of recent developments. *Pattern Recognition*, page 108102, 2021.

[120] Meng-Hao Guo, Tian-Xing Xu, Jiang-Jiang Liu, Zheng-Ning Liu, Peng-Tao Jiang, Tai-Jiang Mu, Song-Hai Zhang, Ralph R. Martin, Ming-Ming Cheng, and Shi-Min Hu. Attention mechanisms in computer vision: A survey, 2021. arXiv:2111.07624.

[121] Abdelkader Dairi, Fouzi Harrou, Sofiane Khadraoui, and Ying Sun. Integrated multiple directed attention-based deep learning for improved air pollution forecasting. *IEEE Transactions on Instrumentation and Measurement*, 70:1–15, 2021.

[122] Emadeldeen Eldele, Zhenghua Chen, Chengyu Liu, Min Wu, Chee-Keong Kwoh, Xiaoli Li, and Cuntai Guan. An attention-based deep learning approach for sleep stage classification with single-channel eeg. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29:809–818, 2021.

[123] Qipin Chen, Zhenyu Shi, Zhen Zuo, Jinmiao Fu, and Yi Sun. Two-stream hybrid attention network for multimodal classification. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 359–363. IEEE, 2021.

[124] Chiori Hori, Takaaki Hori, Teng-Yok Lee, Kazuhiro Sumi, John R. Hershey, and Tim K. Marks. Attention-based multimodal fusion for video description, 2017. arXiv:1701.03126.

[125] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering, 2016. arXiv:1511.02274.

[126] Xiang Long, Chuang Gan, Gerard De Melo, Xiao Liu, Yandong Li, Fu Li, and Shilei Wen. Multimodal keyless attention fusion for video classification. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[127] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. *Advances in Neural Information Processing Systems*, 34, 2021.

[128] Nelson Nunes, Bruno Martins, Nuno André da Silva, Francisca Leite, and Mário J Silva. A multi-modal deep learning method for classifying chest radiology exams. In *EPIA Conference on Artificial Intelligence*, pages 323–335. Springer, 2019.

[129] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. 2019. URL: https://arxiv.org/abs/1905.11946, doi:10.48550/ARXIV.1905.11946.

[130] Yijia Zhang, Qingyu Chen, Zhihao Yang, Hongfei Lin, and Zhiyong Lu. BioWordVec, improving biomedical word embeddings with subword information and MeSH. *Scientific Data*, 6(1), May 2019. URL: https://doi.org/10.1038/s41597-019-0055-0, doi:10.1038/s41597-019-0055-0.

[131] Siddharth Biswal, Peiye Zhuang, Ayis Pyrros, Nasir Siddiqui, Sanmi Koyejo, and Jimeng Sun. Emixer: End-to-end multimodal x-ray generation via self-supervision, 2021. arXiv:2007.05597.

[132] Yang Yu, Peng Hu, Jie Lin, and Pavitra Krishnaswamy. Multimodal multitask deep learning for x-ray image retrieval. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 603–613. Springer, 2021.

[133] Tzu-Ming Harry Hsu, Wei-Hung Weng, Willie Boag, Matthew McDermott, and Peter Szolovits. Unsupervised multimodal representation learning across medical images and reports, 2018. arXiv:1811.08615.

[134] Ivan Dokmanic, Reza Parhizkar, Juri Ranieri, and Martin Vetterli. Euclidean distance matrices: Essential theory, algorithms, and applications. *IEEE Signal Processing Magazine*, 32(6):12–30, Nov 2015. URL: http://dx.doi.org/10.1109/MSP.2015.2398954, doi:10.1109/msp.2015.2398954.

[135] Xiu-Shen Wei, Jian-Hao Luo, Jianxin Wu, and Zhi-Hua Zhou. Selective convolutional descriptor aggregation for fine-grained image retrieval, 2017. arXiv:1604.04994.

[136] Alfirna Rizqi Lahitani, Adhistya Erna Permanasari, and Noor Akhmad Setiawan. Cosine similarity to determine similarity measure: Study case in online essay assessment. In *2016 4th International Conference on Cyber and IT Service Management*, pages 1–6, 2016. doi:10.1109/CITSM.2016.7577578.

[137] Jiansheng Fang, Huazhu Fu, and Jiang Liu. Deep triplet hashing network for case-based medical image retrieval. *Medical Image Analysis*, 69:101981, Apr 2021. URL: http://dx.doi.org/10.1016/j.media.2021.101981, doi:10.1016/j.media.2021.101981.

[138] Steven M. Beitzel, Eric C. Jensen, and Ophir Frieder. *MAP*, pages 1691–1692. Springer US, Boston, MA, 2009. URL: https://doi.org/10.1007/978-0-387-39940-9_492, doi:10.1007/978-0-387-39940-9_492.

[139] Nick Craswell. *Mean Reciprocal Rank*, pages 1703–1703. Springer US, Boston, MA, 2009. URL: https://doi.org/10.1007/978-0-387-39940-9_488, doi:10.1007/978-0-387-39940-9_488.

[140] Yining Wang, Liwei Wang, Yuanzhi Li, Di He, Tie-Yan Liu, and Wei Chen. A theoretical analysis of ndcg type ranking measures, 2013. arXiv:1304.6480.

[141] Erico Tjoa and Cuntai Guan. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE Transactions on Neural Networks and Learning Systems*, PP, 09 2020. doi:10.1109/TNNLS.2020.3027314.

[142] Ricards Marcinkevics and Julia E. Vogt. Interpretability and explainability: A machine learning zoo mini-tour. *ArXiv*, abs/2012.01805, 2020.

[143] Zachary C. Lipton. The mythos of model interpretability, 2017. arXiv:1606.03490.

[144] Kathryn Schutte, Olivier Moindrot, Paul Hérent, Jean-Baptiste Schiratti, and Simon Jégou. Using stylegan for visual interpretability of deep learning models on medical images, 2021. arXiv:2101.07563.

[145] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Dino Pedreschi, and Fosca Giannotti. A survey of methods for explaining black box models, 2018. arXiv:1802.01933.

[146] Diogo V Carvalho, Eduardo M Pereira, and Jaime S Cardoso. Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8):832, 2019.

[147] Finale Doshi-Velez, Mason Kortz, Ryan Budish, Chris Bavitz, Sam Gershman, David O'Brien, Kate Scott, Stuart Schieber, James Waldo, David Weinberger, Adrian Weller, and Alexandra Wood. Accountability of ai under the law: The role of explanation, 2019. arXiv:1711.01134.

[148] Michael Felderer and Rudolf Ramler. *Quality Assurance for AI-Based Systems: Overview and Challenges (Introduction to Interactive Session)*, pages 33–42. 01 2021. doi:10.1007/978-3-030-65854-0_3.

[149] Gregor Stiglic, Primoz Kocbek, Nino Fijacko, Marinka Zitnik, Katrien Verbert, and Leona Cilar. Interpretability of machine learning-based prediction models in healthcare. *WIREs Data Mining and Knowledge Discovery*, 10(5), Jun 2020. URL: http://dx.doi.org/10.1002/widm.1379, doi:10.1002/widm.1379.

[150] Christoph Molnar, Giuseppe Casalicchio, and Bernd Bischl. Interpretable machine learning – a brief history, state-of-the-art and challenges, 2020. arXiv:2010.09337.

[151] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai, 2019. arXiv:1910.10045.

[152] Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning, 2019. arXiv:1806.00069.

[153] Wilson Silva, Kelwin Fernandes, and Jaime S. Cardoso. How to produce complementary explanations using an ensemble model. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2019. doi:10.1109/IJCNN.2019.8852409.

[154] Ian Jolliffe. *Principal Component Analysis*, pages 1094–1096. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011. URL: https://doi.org/10.1007/978-3-642-04898-2_455, doi:10.1007/978-3-642-04898-2_455.

[155] J. A. Hartigan and M. A. Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979. URL: http://www.jstor.org/stable/2346830.

[156] Yongxin Zhou, Matthieu Boussard, and Agnes Delaborde. Towards an xai-assisted third-party evaluation of ai systems: Illustration on decision trees. In Davide Calvaresi, Amro Najjar, Michael Winikoff, and Kary Främling, editors, *Explainable and Transparent AI and Multi-Agent Systems*, pages 158–172, Cham, 2021. Springer International Publishing.

[157] Been Kim, Cynthia Rudin, and Julie Shah. The bayesian case model: A generative approach for case-based reasoning and prototype classification, 2015. arXiv:1503.01161.

[158] Oscar Li, Hao Liu, Chaofan Chen, and Cynthia Rudin. Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. In *AAAI*, 2018.

[159] Evelyn Fix and J. L. Hodges. Discriminatory analysis. nonparametric discrimination: Consistency properties. *International Statistical Review / Revue Internationale de Statistique*, 57(3):238–247, 1989. URL: http://www.jstor.org/stable/1403797.

[160] Lior Rokach and Oded Maimon. *Decision Trees*, pages 165–192. Springer US, Boston, MA, 2005. URL: https://doi.org/10.1007/0-387-25465-X_9, doi:10.1007/0-387-25465-X_9.

[161] Rich Caruana, Hooshang Kangarloo, John David Dionisio, Usha Sinha, and David Johnson. Case-based explanation of non-case-based learning methods. In *Proceedings of the AMIA Symposium*, page 212. American Medical Informatics Association, 1999.

[162] Khushbu Kumari and Suniti Yadav. Linear regression analysis study. *Journal of the Practice of Cardiovascular Sciences*, 4:33, 01 2018. doi:10.4103/jpcs.jpcs_8_18.

[163] J.S. Cramer. The Origins of Logistic Regression. Tinbergen Institute Discussion Papers 02-119/4, Tinbergen Institute, December 2002. URL: https://ideas.repec.org/p/tin/wpaper/20020119.html.

[164] Vikramkumar, Vijaykumar B, and Trilochan. Bayes and naive bayes classifier, 2014. arXiv:1404.0933.

[165] Eric Wallace, Shi Feng, and Jordan Boyd-Graber. Interpreting neural networks with nearest neighbors, 2018. arXiv:1809.02847.

[166] Nicolas Papernot and Patrick McDaniel. Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning, 2018. arXiv:1803.04765.

[167] Yongxin Yang, Irene Garcia Morillo, and Timothy M. Hospedales. Deep neural decision trees, 2018. arXiv:1806.06988.

[168] Shuoyang Ding and Philipp Koehn. Evaluating saliency methods for neural language models, 2021. arXiv:2104.05824.

[169] Yola Jones, Fani Deligianni, and Jeff Dalton. Improving ecg classification interpretability using saliency maps. *2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE)*, Oct 2020. URL: http://dx.doi.org/10.1109/BIBE50027.2020.00114, doi:10.1109/bibe50027.2020.00114.

[170] Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. A survey of the state of explainable ai for natural language processing, 2020. arXiv:2010.00711.

[171] Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73:1–15, Feb 2018. URL: http://dx.doi.org/10.1016/j.dsp.2017.10.011, doi:10.1016/j.dsp.2017.10.011.

[172] Pieter-Jan Kindermans, Kristof T. Schütt, Maximilian Alber, Klaus-Robert Müller, Dumitru Erhan, Been Kim, and Sven Dähne. Learning how to explain neural networks: Patternnet and patternattribution, 2017. arXiv:1705.05598.

[173] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise, 2017. arXiv:1706.03825.

[174] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks, 2013. arXiv:1311.2901.

[175] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net, 2015. arXiv:1412.6806.

[176] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks, 2017. arXiv:1703.01365.

[177] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences, 2019. arXiv:1704.02685.

[178] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

[179] Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition*, 65:211–222, 2017. URL: https://www.sciencedirect.com/science/article/pii/S0031320316303582, doi:https://doi.org/10.1016/j.patcog.2016.11.008.

[180] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, Oct 2019. URL: http://dx.doi.org/10.1007/s11263-019-01228-7, doi:10.1007/s11263-019-01228-7.

[181] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144, 2016.

[182] Maximilian Kohlbrenner, Alexander Bauer, Shinichi Nakajima, Alexander Binder, Wojciech Samek, and Sebastian Lapuschkin. Towards best practice in explaining neural network decisions with lrp. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2020.

[183] Xiaochuang Han, Byron C Wallace, and Yulia Tsvetkov. Explaining black box predictions and unveiling data artifacts through influence functions. *arXiv preprint arXiv:2005.06676*, 2020.

[184] Helena Montenegro, Wilson Silva, and Jaime Cardoso. Privacy-preserving generative adversarial network for case-based explainability in medical image analysis. *IEEE Access*, PP:1–1, 11 2021. doi:10.1109/ACCESS.2021.3124844.

[185] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison, 2019. arXiv:1901.07031.

[186] Alistair E. W. Johnson, Tom J. Pollard, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih ying Deng, Yifan Peng, Zhiyong Lu, Roger G. Mark, Seth J. Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs, 2019. arXiv:1901.07042.

[187] Yuan Li, Shan Tian, Yajun Huang, and Weiguo Dong. Driverless artificial intelligence framework for the identification of malignant pleural effusion. *Translational Oncology*, 14(1):100896, 2021. URL: https://www.sciencedirect.com/science/article/pii/S1936523320303880, doi:https://doi.org/10.1016/j.tranon.2020.100896.

[188] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's neural machine translation system: Bridging the gap between human and machine translation, 2016. arXiv:1609.08144.

[189] Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B. A. McDermott. Publicly available clinical bert embeddings, 2019. URL: https://arxiv.org/abs/1904.03323, doi:10.48550/ARXIV.1904.03323.

[190] Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. Captum: A unified and generic model interpretability library for pytorch, 2020. arXiv:2009.07896.

[191] Hina Amin and Waqas J Siddiqui. Cardiomegaly. In *StatPearls [internet]*. StatPearls Publishing, 2021.

[192] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[193] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International Conference on Machine Learning*, pages 3145–3153. PMLR, 2017.

[194] Wei Liu, Sihan Chen, Longteng Guo, Xinxin Zhu, and Jing Liu. Cptr: Full transformer network for image captioning, 2021. URL: https://arxiv.org/abs/2101.10804, doi:10.48550/ARXIV.2101.10804.

[195] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers amp; distillation through attention, 2020. URL: https://arxiv.org/abs/2012.12877, doi:10.48550/ARXIV.2012.12877.

[196] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers, 2021. URL: https://arxiv.org/abs/2106.08254, doi:10.48550/ARXIV.2106.08254.

[197] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[198] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. URL: https://arxiv.org/abs/1503.02531, doi:10.48550/ARXIV.1503.02531.

[199] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units, 2015. URL: https://arxiv.org/abs/1508.07909, doi:10.48550/ARXIV.1508.07909.

[200] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. URL: https://aclanthology.org/P02-1040, doi:10.3115/1073083.1073135.

[201] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. URL: https://aclanthology.org/W05-0909.

[202] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL: https://aclanthology.org/W04-1013.

[203] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation, 2014. URL: https://arxiv.org/abs/1411.5726, doi:10.48550/ARXIV.1411.5726.

[204] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface's transformers: State-of-the-art natural language processing, 2019. URL: https://arxiv.org/abs/1910.03771, doi:10.48550/ARXIV.1910.03771.

[205] Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning, 2016. URL: https://arxiv.org/abs/1612.00563, doi:10.48550/ARXIV.1612.00563.

[206] An Yan, Zexue He, Xing Lu, Jiang Du, Eric Chang, Amilcare Gentili, Julian McAuley, and Chun-Nan Hsu. Weakly supervised contrastive learning for chest x-ray report generation, 2021. URL: https://arxiv.org/abs/2109.12242, doi:10.48550/ARXIV.2109.12242.