

FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Generalized Zero-Shot Chest X-ray Classification

David Teixeira Preda



Master in Informatics and Computing Engineering

Supervisor: Prof. Doutor Luís Filipe Pinto de Almeida Teixeira

Co-Supervisor: Isabel Cristina Rio-Torto de Oliveira

July 26, 2024

Generalized Zero-Shot Chest X-ray Classification

David Teixeira Preda

Master in Informatics and Computing Engineering

July 26, 2024

Resumo

Os avanços em *Deep Learning* da última década demonstraram-se capazes de resolver um vasto leque de problemas de diagnóstico médico e a sua performance tem aumentado ano após ano. Assim sendo, o *staff* médico terá cada vez mais tempo para cuidar dos seus pacientes à medida que o apoio em sistemas de inteligência artificial se vai tornando mais ubíquo.

Uma tarefa comum neste domínio é a classificação de imagens raio-X do peito humano (CXR), onde, dada uma imagem do tórax de um paciente, um modelo deve ser capaz de determinar corretamente quais as patologias presentes nessa mesma região do corpo. Embora a classificação de CXRs já tenha sido profundamente explorada na perspetiva de uma só *label* por imagem, os avanços mais recentes encaram o problema numa perspetiva *multi-label* dado que os pacientes frequentemente padecem simultaneamente de várias complicações. Adicionalmente, novas técnicas surgiram para lidar com este problema no contexto de Aprendizagem *Zero-Shot* Generalizada (GZSL), onde, aquando da inferência, um modelo deve prever corretamente as *labels* de uma imagem, independentemente de ter visto instâncias dessa classe ou não durante o seu treino.

Este trabalho visa contribuir para o domínio de classificação *multi-label* de CXRs no contexto GZSL. Métodos desenvolvidos previamente dependem de vetores de atributos para classificarem cada patologia sem ambiguidade. No entanto, definir estes vetores de forma precisa requer um grande investimento de tempo da parte de profissionais médicos altamente especializados. Assim como os desenvolvimentos mais recentes, este trabalho evita o uso de vetores de atributos, favorecendo geração e desambiguação de *features* para descrever cada uma das classes. Mahapatra et al. sugerem uma abordagem similar através de uma *disentanglement network* para criar *features* específicas de cada classe e *embeddings* do modelo BioBERT para as gerar para classes desconhecidas.

Partindo desse trabalho, endereçamos as nossas preocupações ao mesmo tempo que seguimos ideias que pensamos serem fulcrais para a resolução desta tarefa. Propomos um novo método que aproveita a relação entre um conjunto conhecido de *embeddings* para criar um espaço latente a partir do qual podemos gerar novos exemplos de imagens de classes desconhecidas. O nosso trabalho diverge do estado da arte ao igualar o número de *encoders* ao número total de classes (em vez de ter um por classe conhecida) ao mesmo tempo que os faz partilhar um único *decoder* comum. Para além disso, em vez de permitir aos centróides de cada classe que se alterem de forma a cumprir com um conjunto de relações baseadas em similaridade, restringimos as suas posições de forma a moldar o espaço latente com base num conjunto de *embeddings* de origem externa.

Em paralelo, introduzimos a utilização dos *embeddings* do modelo MedCLIP, especializados no domínio alvo e com menor similaridade entre si, um problema que vemos com a extensiva utilização dos *embeddings* do modelo BioBERT na literatura.

Ao testar o método proposto com duas arquiteturas diferentes, descobrimos que, embora a maior dimensionalidade dos *embeddings* do BioBERT providencie uma melhor base para a reconstrução de imagens das classes conhecidas, os *embeddings* do MedCLIP mostram sinais de serem uma alternativa mais robusta para generalizar um método para as classes desconhecidas.

Abstract

Deep learning advances in the last decade have proved capable of solving a wide array of medical diagnosis tasks and their performance keeps improving year over year. Thus, medical staff will have more time to actively attend to their patients as healthcare systems increasingly rely on these new artificial intelligence systems.

A common task in this domain is chest X-ray (CXR) classification, where given an image of the thorax of a patient, a model should correctly label it according to the pathologies present within that body region. While CXR classification has been thoroughly explored from a single-label perspective, recent developments have shifted to multi-label classification, as patients usually exhibit multiple simultaneous complications. Additionally, new techniques have emerged to deal with this problem in the context of Generalized Zero-shot Learning (GZSL), where, at test time, a model is required to correctly predict the labels of an image, regardless of whether it saw instances of its class during training.

This work aims to contribute to the domain of GZSL CXR multi-label classification. Earlier methods relied on class attribute vectors to unambiguously distinguish between each disease. However, accurately defining such vectors requires extensive time from highly specialised medical workers. In line with present-day developments, this work avoids the usage of attribute vectors. Instead, we focus on the generation and disentanglement of features to accurately describe each class. Mahapatra et al. suggested a similar approach using a disentanglement network to create class specific features and BioBERT’s embeddings to generate them for new unseen classes.

We depart from their work to address our concerns while following what we believe to be core ideas for solving this task. We propose a new method that leverages the relationship between a known set of embeddings to create a shared latent space from which we can sample new images for unseen classes. Our work diverges from Mahapatra et al. by matching the number of encoders to total number of classes (instead of having one per seen class) and having them share one common decoder. At the same time, instead of having class centroids change to match a set of similarity-based relationships, we restrict their position to shape the latent space based on a set of externally provided embeddings.

Concurrently, we introduce the use of MedCLIP embeddings, which are specialized in our target domain while bearing less similarity amongst them, an issue we see with the extensive use of BioBERT embeddings in the literature.

By testing our proposed method with two different model architectures, we discover that, while the higher dimensionality of BioBERT’s embeddings provides a better reconstruction ground for the seen classes, MedCLIP embeddings show signs of being a more robust alternative to generalize to unseen classes.

Agradecimentos

Em primeiro lugar, gostaria de agradecer aos meus orientadores por todo o acompanhamento que deram a este trabalho e a mim mesmo, tanto enquanto estudante como ser humano.

Agradeço ao Professor Luís por me ter dado a oportunidade de trabalhar num tópico que me permitiu explorar livremente uma área que me fascina. Agradeço à Isabel pela amizade, pelo apoio constante e por me encorajar a explorar mais e a fazer melhor. Este trabalho não seria possível sem vocês os dois.

Este trabalho é também o fim de uma jornada de cinco anos. Cinco anos de aprendizagem, de descoberta, de frustração, de conquistas e de derrotas. Mas, acima de tudo, cinco anos de crescimento e de memórias que vingarão em mim até ao fim.

Quero agradecer aos meus pais, por me apoiarem durante todo este percurso, independentemente de concordarem ou não com as decisões que tomei pelo caminho. Ao meu irmão Diego, por me manter jovial num mundo mais adulto. Extendo também o meu agradecimento aos meus restantes familiares que acreditaram em mim e me apoiaram a cada passo.

Agradeço também a todos os meus amigos e colegas de curso. Em particular, quero agradecer ao Diogo, ao José, ao Bruno Mendes, ao Pedro, ao Chico e ao Bruno Gomes. Todos vocês contribuíram para que, desde o início da minha vida académica, as dificuldades fossem atenuadas com companheirismo e amizade e a felicidade exponencialmente aumentada. Espero que a vida vos sorria a todos.

Quero também agradecer a todos os meus amigos e colegas do NIAEFEUP. Estou particularmente agradecido ao Luís, ao Tomás Ramos, ao Fernandes, e ao João Pereira. Num ano muito complicado para mim, vocês deram-me um sentido de comunidade e um lugar ao qual pertencer. Continuem o ótimo trabalho, os nossos estudantes precisam de vocês.

Por fim, e de olhos postos no futuro, quero agradecer à minha namorada, noiva e prestes a ser esposa, Mafalda Vieira, que me acompanhou a cada segundo desta longa e árdua caminhada. Por ter celebrado comigo as maiores e menores conquistas e por me ter amparado em todas e quaisquer quedas. Por me ter dado força para não desistir nos tempos difíceis. E acima de tudo, por moldar o homem que sou hoje no fim deste capítulo.

David Preda

Talent is something you make bloom

及川 徹

Contents

1	Introduction	1
1.1	Dissertation Structure	2
2	Generalized Zero-Shot Chest X-ray Classification: Literature Review	3
2.1	Concepts and Definitions	3
2.1.1	Common Datasets	3
2.2	General Domain GZSL	4
2.2.1	Inductive vs Transductive GZSL	5
2.2.2	Taxonomy	5
2.3	GZSL Applied to CXR Classification	10
2.3.1	Taxonomy	10
2.4	Challenges	13
2.5	Summary and Discussion	14
3	Methodology	16
3.1	State-of-the-Art Method Breakdown	16
3.1.1	Feature Disentanglement	16
3.1.2	Feature Generation	17
3.1.3	Inference	18
3.1.4	Analysis	18
3.2	Method Overview	19
3.2.1	Architecture	19
3.2.2	Training Process	20
3.2.3	Inference	21
3.3	On the Importance of Embeddings	21
3.3.1	Multimodal Embeddings	23
3.4	Experimental Model Architectures	25
3.4.1	Pooling Model	25
3.4.2	Upsampling Model	26
4	Results and Discussion	28
4.1	Dataset	28
4.1.1	Training, Validation and Test Splits	29
4.2	First Training Phase	30
4.2.1	Pooling Model	30
4.2.2	Upsampling Model	32
4.3	Second Training Phase	34
4.3.1	Pooling Model	34

4.3.2	Upsampling Model	36
4.4	Test Results	37
5	Conclusions and Future Work	39
	References	42
A	Training hyperparameters	48

List of Figures

2.1	Sample CXR images from NIH Chest X-Ray Dataset	4
2.2	Class Attribute Network-integration	6
2.3	VAE Alignment Approach Architecture	7
2.4	f-VAEGAN-D2 Architecture	8
2.5	ReViSE Architecture	9
2.6	Multi-Attention Zero-Shot Learning Framework	9
2.7	COMe-SEE proposed architecture	10
2.8	Multi-view semantic approach architecture	11
2.9	Semantic embedding-based inductive GZSL for CXR classification architecture .	12
2.10	Class-specific encoder architecture and resulting separation of classes	13
2.11	Two-phased clustering and generative approach architecture	14
3.1	Feature disentanglement step of Mahapatra et al.’s method.	17
3.2	Block diagram of the proposed method’s first training phase	20
3.3	Block diagram of the proposed method’s second training phase	21
3.4	Test time example of our proposed method	22
3.5	t-SNE plot of BioBERT embeddings for general words and CXR classification labels	23
3.6	MedCLIP training process	24
3.7	t-SNE plot of MedCLIP embeddings for general words and CXR classification labels	25
3.8	Encoder architecture for the first proposed model	26
3.9	Decoder architecture for the first proposed model	26
3.10	Encoder architecture for the second proposed model	27
3.11	Decoder architecture for the second proposed model	27
4.1	Class distribution for the NIH dataset, excluding <i>No Finding</i>	29
4.2	Loss evolution against training and validation sets for the pooling model during the first training phase	31
4.3	t-SNE visualizations of the latent encodings of the pooling model during the first training phase	32
4.4	Reconstruction of a random training set sample by the pooling model	32
4.5	Loss evolution against training and validation sets for the upsampling model during the first training phase	33
4.6	t-SNE visualizations of the latent encodings of the upsampling model during the first training phase	33
4.7	Reconstruction of a random training set sample by the upsampling model	34
4.8	Reconstruction of a random sample from the generated set by the pooling model .	35
4.9	Loss evolution against the generated set for the pooling model during the second training phase	35

4.10	Reconstruction of a random sample from the generated set by the upsampling model	36
4.11	Loss evolution against the generated set for the upsampling model during the second training phase	36

List of Tables

3.1	Cosine similarity matrix between BioBERT’s [LYK ⁺ 19] embeddings	22
3.2	Cosine similarity matrix between MedCLIP’s [WWAS22] embeddings	24
4.1	Class distribution for the NIH dataset	28
4.2	Class distribution for the NIH dataset training and validation splits	30
4.3	Class distribution for the NIH dataset test split	30
4.4	Test results, describing average accuracy for seen and unseen classes and their harmonic mean	37
4.5	Comparison against state-of-the-art work by Mahapatra et al.	38
A.1	Training hyperparameters	48

Abreviaturas e Símbolos

a_c	Attribute Vector for Class c
BERT	Bidirectional Encoder Representations from Transformers
BioBERT	Bidirectional Encoder Representations from Transformers for Biomedical Text Mining
C	Set of Classes
CLIP	Contrastive Language-Image Pre-Training
CNN	Convolutional Neural Network
COMe-SEE	Cross-modality Semantic Embedding Ensemble
Conv	Convolution
ConvTransp	Transposed Convolution
CT	Computerized Tomography
CXR	Chest X-ray
D	Decoder
D_S	Set of Labeled Seen Classes' Samples
D_{train}	Set of Training examples
D_U	Set of Unabeled Unseen Classes' Samples
E_y	Encoder associated with Class y
GAN	Generative Adversarial Network
GCN	Graph Convolutional Network
GloVE	Global Vectors
GZSL	Generalized Zero-Shot Learning
MMD	Maximum Mean Discrepancy
MSE	Mean Squared Error
NIH	National Institutes of Health
ReLU	Rectified Linear Unit
S	Set of Seen Classes
SVM	Support Vector Machine
SwAV	Swapping Assignments between multiple Views of the same image
Tanh	Hyperbolic Tangent
t-SNE	t-Distributed Stochastic Neighbor Embedding
U	Set of Unseen Classes
VAE	Variational Autoencoder
x	Sample Image
Y	Sample Label Set
z_y	Latent Encoding from Encoder associated with Class y
z_{emb_y}	Semantic Embedding associated with Class y

Chapter 1

Introduction

Medical services worldwide are increasingly suffering from staff shortages, which adds to the already pressured environment for these professionals. This is especially true for the radiology field. Thus, there is an urgent need to free up the time of these personnel to improve patient care and the overall efficiency and efficacy of healthcare systems.

Concurrently, Deep Learning advances in the last decade have provided the foundations required to have Artificial Intelligence correctly diagnose medical conditions by analysing patient data such as medical reports or X-ray images. These systems can now be leveraged to solve (or at least alleviate) the staff shortages issues.

However, while many image classification problems take a supervised learning approach by utilising large volumes of labelled data, X-ray images, despite being abundant, are usually unlabelled. At the same time, this data is highly unbalanced. While it is fortunate that most images showcase healthy patients with no pathologies associated, this hinders the objective of building a system capable of correctly identifying diseases through these samples. Labelling these images would require the help of medical professionals specialised in the dataset's domain (in our case, the thoracic region), which goes against this work's initial goal of freeing time for them.

This all leads to the conclusion that the supervised approach is not practical. An alternative approach is Generalized Zero-Shot Learning (GZSL), which trains on a **seen** subset of classes and attempts to transpose their learning to both **seen and unseen** classes.

Most GZSL methods rely on attribute vectors. These real-valued vectors allow the usage of unlabeled data by learning the relationship between images and attribute vectors. However, attribute vectors for the medical domain require, once again, in-depth knowledge only available through medical professionals. Consequently, unlike natural domain images, no such vectors are available for X-ray images. This adds an extra layer of complexity to the problem, as there no longer exists a precise definition for each class that can be leveraged for GZSL.

This work aims to produce a GZSL Chest X-ray image classification method that effectively transposes knowledge obtained from seen classes to a superset also containing unseen classes without using attribute vectors. This is an underexplored challenge, as only a handful of works on GZSL for CXR classification have been published.

1.1 Dissertation Structure

Beyond this introduction, this work contains 4 more chapters. **Chapter 2** reviews all the concepts required for understanding this work and the broader scope it is a part of. It starts by providing an overview of fundamental concepts related to Chest X-ray images and the main datasets used in this field. Then, an overview of GZSL methods is presented, categorizing works into three major areas. From there, the focus shifts to works within the CXR field, where we find the state-of-the-art work for this task by Mahapatra et al. [MJYK⁺23].

Chapter 3 focuses on detailing the method proposed in this work. It starts off by analyzing the inner workings of Mahapatra et al.'s [MJYK⁺23] method, from where we take our driving ideas as well as some concerns to address. Then, our approach is broken down into its loss functions, training process and inference. The chapter then ends with a reflection on the importance of the semantic embeddings for our method followed by the introduction of our two experimental model architectures, along with their pros and cons.

Chapter 4 presents and discusses the results of our proposed method after a brief introduction and analysis of the dataset used. At each training step, the behaviour of both models and their variants is explored by examining the evolution of their loss functions, the organization of the latent space and the quality of the reconstructed images. The chapter closes with the test set evaluation and direct comparison with Mahapatra et al. [MJYK⁺23].

Finally, **Chapter 5** presents the conclusions about the developed work and discusses potential future research directions.

Chapter 2

Generalized Zero-Shot Chest X-ray Classification: Literature Review

This chapter delves into Generalized Zero-Shot Learning (GZSL) in Chest X-ray (CXR) classification. It comprehensively covers key concepts, methodologies, and challenges in this domain.

The chapter begins by establishing the foundation with a focus on CXR image classification, including common datasets. Then, in the broader GZSL domain, we distinguish between inductive and transductive approaches. Both general domain and CXR classification-focused methods are systematically categorized according to their strategy, providing a concise taxonomy of GZSL techniques.

The chapter concludes by addressing challenges specific to GZSL, particularly for CXR classification, providing insights into complexities at the intersection of generalized zero-shot learning and medical imaging. A discussion section synthesizes findings, identifies research gaps, and sets the stage for subsequent chapters in the dissertation.

2.1 Concepts and Definitions

Chest X-ray images, commonly abbreviated as CXR images, refer to radiographic representations of the thorax region (as seen in [Figure 2.1](#)) captured through X-ray imaging techniques. These are widely used to detect and analyze pathologies present within that body region, such as pulmonary infections or cardiac abnormalities.

On the other hand, CXR classification refers to the task of correctly predicting one or more labels representing pathologies associated with an input image. This can be formally defined as a function $f : X \rightarrow Y$, where X is the input space representing CXR images (e.g., pixel values), and Y is the set of possible pathologies associated with the images.

2.1.1 Common Datasets

National Institutes of Health (NIH) Chest X-Ray Dataset [[WPL⁺17](#)]. The NIH Chest X-ray Dataset consists of 112,120 X-ray images accompanied by disease labels, representing 30,805

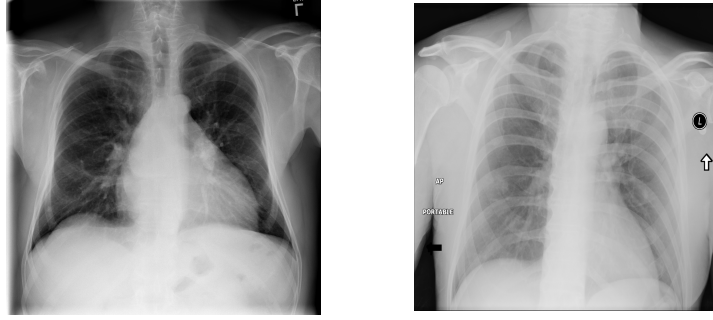


Figure 2.1: Sample CXR images taken from the National Institutes of Health (NIH) Chest X-Ray Dataset ([WPL⁺17]). Left: posteroanterior (PA) view of a patient with cardiomegaly; Right: anteroposterior (AP) view of a patient with lungs revealing mass and node.

unique patients. The authors employed Natural Language Processing techniques to extract disease classifications from corresponding radiological reports. The resulting labels are estimated to be >90% accurate and deemed suitable for weakly supervised learning applications.

CheXpert [IRK⁺19]. The CheXpert Dataset is a comprehensive public dataset for chest radiograph interpretation, comprising 224,316 radiographs from 65,240 patients collected retrospectively at Stanford Hospital between October 2002 and July 2017. Each radiology report associated with the examinations was labeled for the presence of 14 observations, with labels determined based on prevalence and clinical relevance. An automated rule-based labeler, organized into three stages — mention extraction, mention classification, and mention aggregation — was developed to extract structured labels from the free-text radiology reports, providing a detailed and nuanced framework for image interpretation.

MIMIC-CXR [JPB⁺19]. The MIMIC Chest X-ray Database (MIMIC-CXR) comprises 377,110 DICOM-formatted chest radiographs and de-identified radiology reports. The dataset, anonymized in accordance with the Health Insurance Portability and Accountability Act of 1996 (HIPAA) [Act96], originates from Beth Israel Deaconess Medical Center. Radiology reports underwent de-identification through a rule-based and neural network approach.

2.2 General Domain GZSL

Generalized Zero-Shot Learning (GZSL) is a learning paradigm where the model is trained on a set of seen classes and is expected to classify both seen and unseen classes during testing. This means that not only does the model have to be able to deal with classes for which it did not have any examples (i.e., it has to be able to deal with class imbalances), but it also has to accommodate for new, unaccounted-for classes.

More formally, GZSL refers to any method that undergoes training on samples from a specified set of seen classes denoted as $S = \{c_1, c_2, \dots, c_n\}$, where n signifies the total number of seen classes, with the primary objective of achieving accurate predictions within a broader set of classes $C = S \cup U$, where U represents the set of m unseen classes $U = \{c_{n+1}, c_{n+2}, \dots, c_m\}$.

This paradigm shift in learning poses a significant departure from traditional methods, where exhaustive exposure to examples of each class facilitates the establishment of class-specific representations. Unlike conventional learning, where models can precisely tune their parameters based on the examples provided for each class, GZSL necessitates a more generalized approach. The model must rely on alternative sources of information and mechanisms to generate meaningful features for classes that were not part of its training data.

2.2.1 Inductive vs Transductive GZSL

Zero-shot learning and, by definition, its generalized version, can be categorized depending on what data is available during training time.

Let $D_s = \{(x_{c_1}^1, y_{c_1}^1), (x_{c_1}^2, y_{c_1}^2), \dots, (x_{c_n}^d, y_{c_n}^d)\}$ be the set of labeled seen classes' samples, where $x_{c_i}^j$ is the j -th sample of class c_i with label $y_{c_i}^j$ and $D_u = \{x_{c_{n+1}}^1, x_{c_{n+1}}^2, \dots, x_{c_m}^d\}$ the unlabeled set of unseen classes' samples, where $x_{c_i}^j$ is the j -th sample of class c_i .

Inductive GZSL encompasses all methods whose training set is limited to the examples of the seen classes. Therefore, the set of training examples D_{train} can be defined as $D_{train} = D_s$ for this type of method. Another way to formulate this would be:

Definition 1 (Inductive GZSL Assumption) *Only labeled data from seen classes is available at train time.*

On the other hand, transductive methods alter the initial GZSL problem by further assuming that:

Definition 2 (Transductive GZSL Assumption) *Unlabeled data from unseen classes is available at train time.*

Thus, for transductive methods, the training set can be defined as $D_{train} = D_s \cup D_u$.

Several works within the medical domain [HLS21, PSP⁺22] debate the validity of this assumption: whether or not the classes can be really considered *unseen* if their data is being used to train the models and if it is realistic to assume that such data would be available at train time.

As a general rule, whenever not discriminated, a GZSL method is considered to be inductive (i.e., it follows Definition 1).

2.2.2 Taxonomy

Many approaches in the general domain predominantly depend on class attribute vectors. An attribute vector $\mathbf{a}_c$ for a class $c \in \mathcal{C}$ is a real-valued vector defined in an k -dimensional attribute space, where k is the number of attributes. Formally, an attribute vector can be expressed as:

$$\mathbf{a}_c = [a_{c1}, a_{c2}, \dots, a_{ck}]$$

where each a_{ci} represents the value associated with the i -th attribute for class c . The attribute vectors serve as quantitative representations of the intrinsic characteristics or features that define each class.

These vectors serve as a primary source of information for dealing with unseen classes. Notably, acquiring attribute vectors is generally simpler compared to obtaining labels, as a single attribute vector can be applied universally to all images within its corresponding class.

2.2.2.1 Visual to Semantic Augmentation

The following methodologies center around the creation of vectors within the semantic embedding space of the relevant classes. This is achieved by learning to project image features into this space. These generated vectors are then employed either to augment the dataset with samples from previously unseen classes or to classify test examples by comparing them to class attribute vectors.

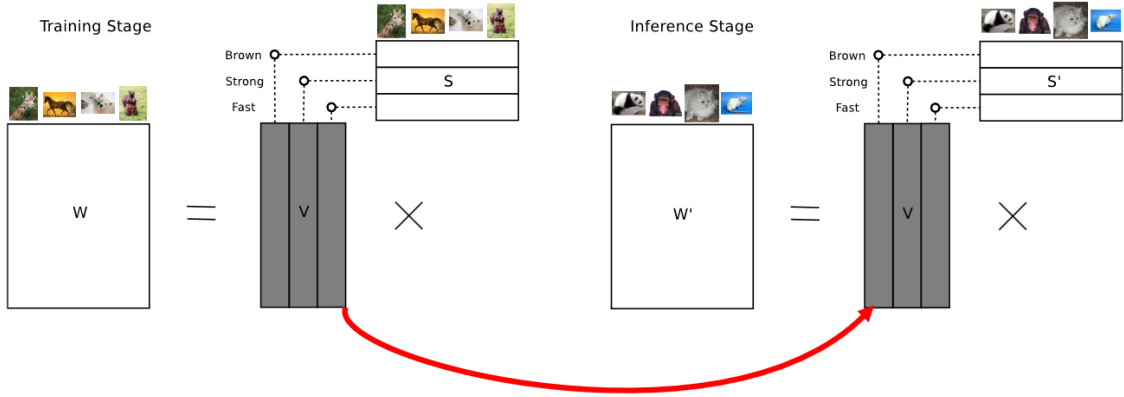


Figure 2.2: Class attribute network-integration, taken from [RPT15]. At train time, the V matrix is learned to convert from feature space to attribute space and obtain the final linear model, defined as W and W' at train and test time, respectively.

Starting with [APHS13], the authors shift the focus from feature extraction to leveraging class attributes for creating label embeddings. They propose a compatibility function F that is trained to assess the compatibility between an image's features and a class label embedding.

In contrast, [RPT15] offers a simpler version of the approach presented in [APHS13]. Here, the authors focus on learning the mapping from the pair of image features and attributes to class (see Figure 2.2). They extend the conventional weight-based approach by integrating class attributes directly into the weight calculation. This method aims to enhance the model's ability to generalize by considering both image features and class attributes in tandem.

Moving on to [SGMN13], the authors present an early approach where images are projected into a semantic word space, derived from word embeddings taken from [HSMN12]. The comparison between the projected images and word vectors is used to determine whether the projection corresponds to one of the seen classes. If this alignment is established, a classifier, previously trained on the seen classes, is employed to assign the correct label. Otherwise, a comparative

analysis is conducted with each word vector. This process identifies the closest word vector and, by extension, associates the projection with its most similar (unseen) class.

Finally, [KXG17] introduces a different strategy by employing a Variational Autoencoder (VAE) [KW14]. In this approach, the VAE is trained to reconstruct an input while approximating it to its class embedding vector within the latent space.

2.2.2.2 Semantic to Visual Augmentation

The following methods focus on generating synthetic samples, particularly exploring the synthesis of images and features conditioned on class attribute vectors. These approaches are, for the vast majority, more recent than visual-to-semantic ones, as they leverage newer Deep Learning architectures such as GANs [GPAM⁺14] and VAEs [KW14].

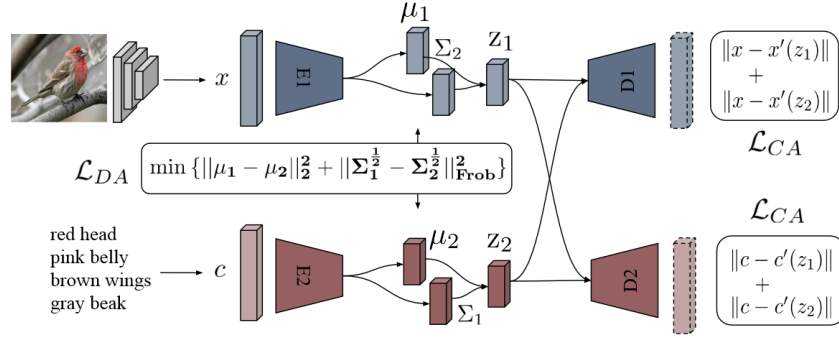


Figure 2.3: VAE alignment approach architecture, taken from [SES⁺19]. Distribution alignment forces the two latent distributions to approach each other, while cross-alignment loss incentivizes cross-modal reconstruction, i.e., latent representations of the same sample should be decoded into the same vector, regardless of which latent distribution they come from.

[VAMR18] propose a method that generates synthetic examples based on image encoding and class attribute vectors, resorting to a regressor network when such vectors are absent.

Concurrently, [XLSA18] leverage a GAN to learn how to generate CNN [LBBH98] features conditioned on class attribute vectors, utilizing pre-trained classifier loss as guidance for improved feature generation. [SES⁺19] present an approach by learning a joint space through training a set of VAEs for each modality, ensuring similarity through Cross-Alignment and Distribution-Alignment Loss (see Figure 2.3).

[MRMM18] employ a VAE to reconstruct input while conditioned on class attribute vectors, utilizing the generated samples to train an SVM on both seen and generated data. Sariyildiz et al. [SC19] introduce a simple GAN-based method for generating synthetic features for both seen and unseen classes, guided by a supervisory signal that matches the gradient of a classifier.

[XSSA19] propose a transductive joint VAE/GAN architecture, where VAE encodes image features, and GAN generates them conditioned on class embeddings. Similarly, [WZZ⁺20] separate seen and unseen classes, employing two GANs to generate synthetic features based on class embeddings and using self-supervised learning for feature alignment via triplet mining [SKP15].

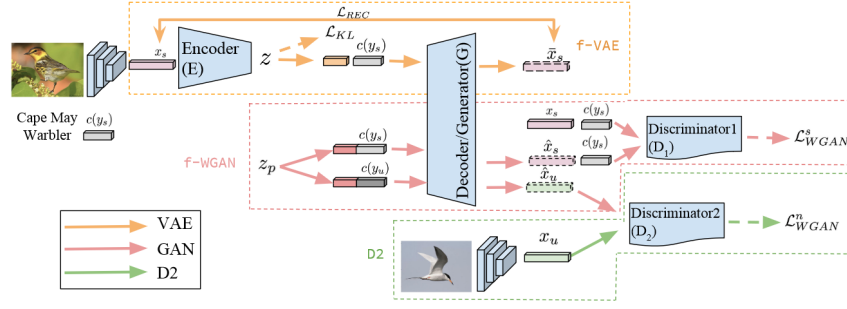


Figure 2.4: f-VAEGAN-D2 architecture, taken from [XSSA19]. Note that the Encoder/Generator of the VAE and GAN, respectively, are shared. The GAN (conditioned on class attribute vectors) learns to generate synthetic features and discriminate them from real ones, while the VAE encourages better feature generation through reconstruction. At the same time, a second discriminator distinguishes real from synthetic features of unseen classes.

Finally, [SLC⁺22] adopt a multi-phased approach utilizing two VAEs to align images and semantic descriptions of seen classes. They synthesize fictitious samples from the semantic descriptions of unseen classes, creating boundaries and deploying two classifiers based on latent representation similarity.

2.2.2.3 Mixed Approaches

Finally, the presented works either aggregate the two previous categories by leveraging their ideas or follow an independent route, including the usage of Graph Convolutional Networks [KW17] or Neural Network Distillation [HVD15].

An early work proposed by [FCS⁺13] combines a visual model, augmented with a projection layer to the embedding space, with an embedding lookup table. [HTHS17] expand upon this idea by introducing an extra step in the fusion of visual features and semantic embeddings: an unsupervised learning phase, aligning distributions through maximum mean discrepancy loss (MMD) [GBR⁺12]. Their architecture is exemplified in Figure 2.5.

[FHGX15] deviate from a linear progression from image to semantic features. Instead, they advocate leveraging multiple semantic spaces and subsequently consolidating them into a multi-dimensional space conducive to label propagation. On a similar note, [ZWS19] identify an information gap between visual features and semantic attributes, addressing it through an intermediate representation known as visually semantic embeddings.

[LLWJ18] introduce a distinctive method, merging image features with class attribute vectors to train a prediction function. This approach, based on [HVD15]’s distillation with temperature calibration, enables confident classification of seen classes as unseen, particularly when features closely match the respective unseen attribute vectors. This facilitates the extraction of shared information between seen and unseen classes.

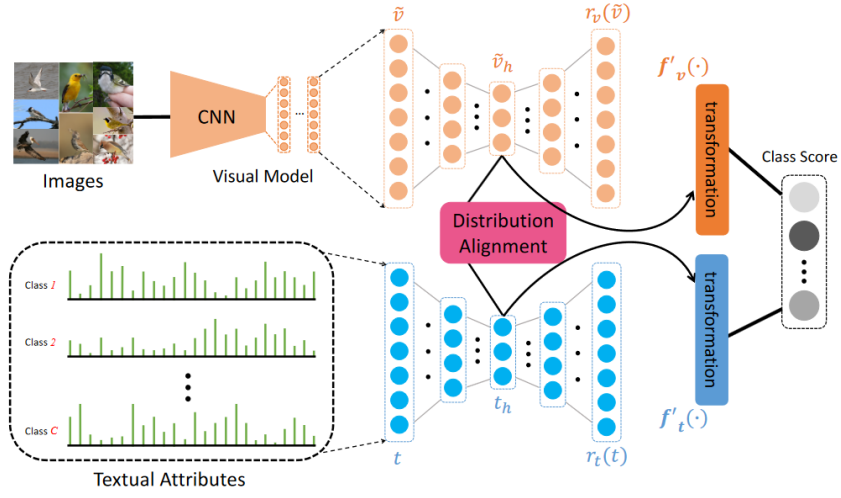


Figure 2.5: ReViSE architecture, as proposed in [HTHS17]. Unlike [FCS⁺13], this architecture aligns the two models by enforcing latent similarity between them before applying transformations.

[?] introduce a logistic regression-based classifier trained on seen examples, coupled with a Graph Convolutional Network [KW17]. This GCN, incorporating classifier weights and embeddings of unseen classes, predicts weights for accurate classification of unseen classes based on their relationships with each other and seen classes.

[HWYW19]’s model combines visual-to-semantic and semantic-to-visual methodologies. Their joint Generator-Regressor-Discriminator architecture, trained with dual adversary loss, allows the generator and regressor to mutually enhance each other, which ultimately results in improved generated samples.

Lastly, [HE20] involves the division of an image into multiple crops, extracting visual features from each. These features are then transformed into attention features [VSP⁺17], projected into a joint visual-semantic embedding space, and compared with semantic class vectors for final prediction. Attribute vectors are derived from the GloVe [PSM14] representation of the label.

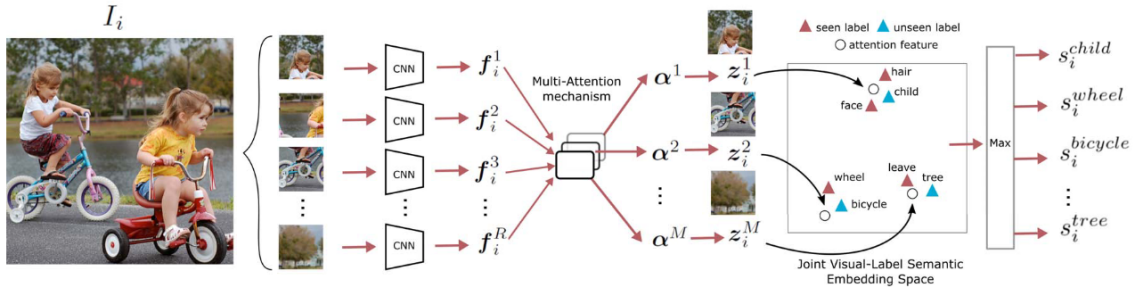


Figure 2.6: Multi-attention zero-shot learning framework, taken from [HE20]. By focusing on different elements, the model is capable of associating the image with several different classes.

2.3 GZSL Applied to CXR Classification

GZSL can be particularly valuable for CXR classification, as it can help identify rare or novel diseases that may not have been included in the initial dataset.

However, in contrast to methods for the general domain, developing attribute vectors for CXR classification proves to be a non-trivial task. The intricate process of crafting these vectors requires a considerable time commitment from highly specialized medical professionals, contradicting the fundamental objective of these systems, which is to enhance the efficiency of medical services. Consequently, in the majority of cases, the unavailability of these attribute vectors introduces an extra layer of complexity to the existing challenges.

2.3.1 Taxonomy

Generalized zero-shot learning has seen very few works applied to CXR classification. Most of the literature falls under the semantic embedding umbrella, where an attribute vector is created to represent its respective class. Most works usually diverge on how to create these class anchors and link samples to them.

2.3.1.1 Semantic Spaces

[PSB⁺20] presented a first effort in this direction by presenting COMe-SEE (Cross-modality Semantic Embedding Ensemble). This two-phased transductive framework comprises a feature extractor and a semantic embedding ensemble, as illustrated in Figure 2.7.

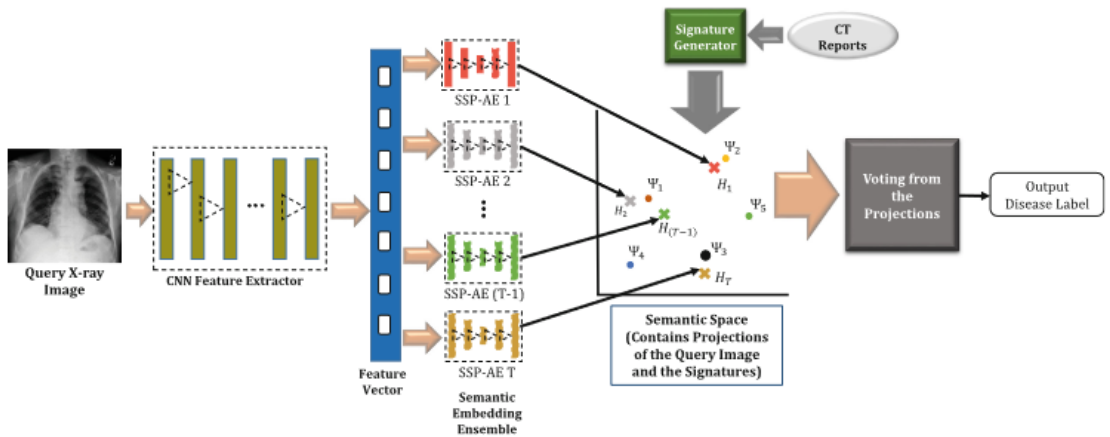


Figure 2.7: COMe-SEE proposed architecture. Each semantic saliency preserving autoencoder (SSP-AE) is trained on a subset of the training data in an attempt to better generalize for unseen classes. Taken from [PSB⁺20].

In the first phase, the authors employed a Word2Vec model specialized in CT reports to construct a set of *signatures* for both seen and unseen target pathologies. Subsequently, each sample underwent encoding using the feature extractor of DenseNet-121 [RIZ⁺17] and was then passed

through a series of autoencoders for projection into the signature space. The projection was matched with its nearest signature, and the decisions were weighted to formulate a final, single label for the image.

A similar approach is presented in [PSL⁺21], where the authors propose the creation of three different semantic spaces. The first two, based on X-ray and CT reports, are, as in [PSB⁺20], created through the usage of Word2Vec models, and the last one is created through visual traits - vectors describing diseases or conditions, manually crafted based on the visual observations of a lesion by radiologists.

The trait semantic space is used as guidance for the other two in order for them to align with expert observation and, thus, better align with their visual counterpart. At inference time, samples are encoded to the X-ray and CT spaces and associated with the closest class signature to produce a singular label.

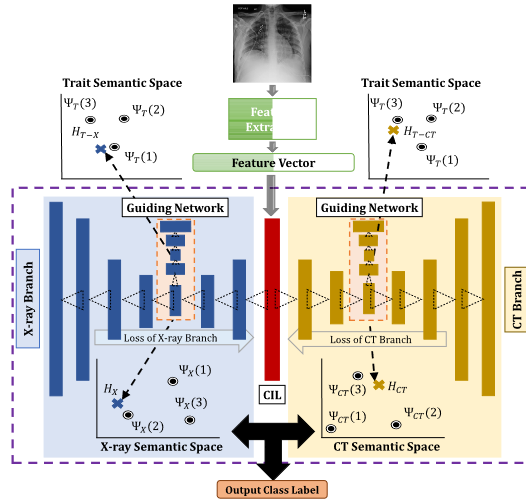


Figure 2.8: Multi-view semantic space approach, taken from [PSL⁺21]. The trait space guides the other two by conditioning their intermediate representation.

Later works [HLS21, PSP⁺22] questioned the value of these approaches due to their transductive nature - the classes are assumed to be unseen, but their examples are used to generate signatures upon which both methods rely (Definition 2), therefore violating the initial assumption.

2.3.1.2 Semantic Embeddings and Similarity

To fix this, [HLS21] propose an inductive method instead. While image features are extracted from each sample with Densenet-121 [RIZ⁺17], the semantic embeddings for each class are extracted from the penultimate layer of BioBERT’s [LYK⁺19] representation of each label. This specific model was chosen due to its contextualized word embeddings learned from biomedical corpora [HLS21].

The benefit of this approach is twofold: it not only eliminates the usage of unseen class data, but it also allows for expanding the model to new classes without having to retrain it.

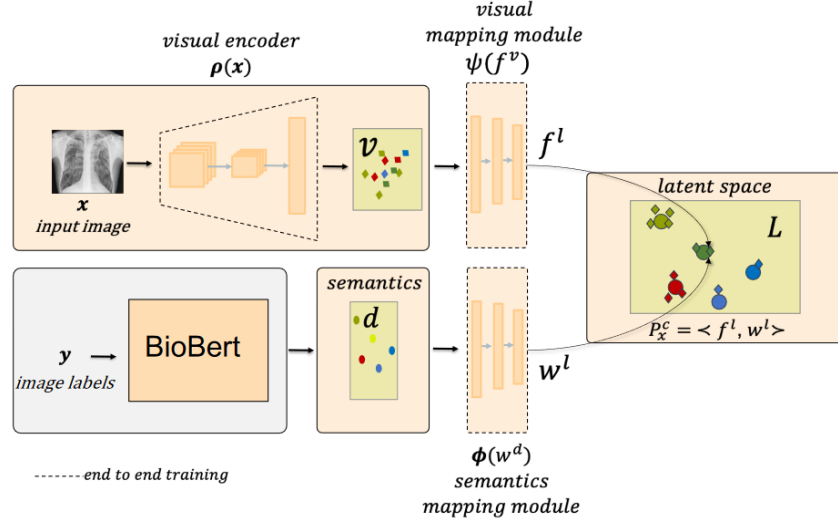


Figure 2.9: Inductive GZSL for CXR classification architecture based on semantic embeddings, taken from [HLS21]. Most subsequent works expand upon this approach.

[PSP⁺22] expanded on the idea of using BioBERT’s embeddings by adding contextual information through knowledge graphs. However, instead of approximating to a semantic embedding, the authors propose a neural network trained on the seen classes. To mitigate the lack of examples from unseen classes, information from the Unified Medical Language System [Bod04] is aggregated into a knowledge graph, from which the semantic embeddings are generated.

In turn, these are used to generate improved classifier head weights for the network through Graph Convolutional Networks [KW17]. A loss term is used in the generation process to enforce alignment between the previously learned weights, trained on the seen classes, and the new, generalized ones. This final classifier is used to predict a single label.

Finally, [MJYK⁺23] propose a novel approach that relies on the utilization of encoders for seen classes to generate class-specific features. These features are then employed to compute class centroids, which serve as semantic embeddings for measuring similarity. Then, for unseen classes, the authors adopt the Mixup technique [ZCDLP18] to generate synthetic features.

However, to ensure the relevance of these synthetic features and prevent divergence from the intended class, the authors introduce a dictionary of similarity based on the embeddings of BioBERT [LYK⁺19] as a guiding mechanism. During the inference phase, the class-specific features of a given sample are compared to the respective class centroids. The decision criterion for considering a sample as a positive example of a particular class involves evaluating the cosine similarity, and a threshold of 0.5 is set for this purpose.

2.3.1.3 Generative Approaches

Concurrently, two noteworthy GZSL generative approaches have been proposed for CXR classification.

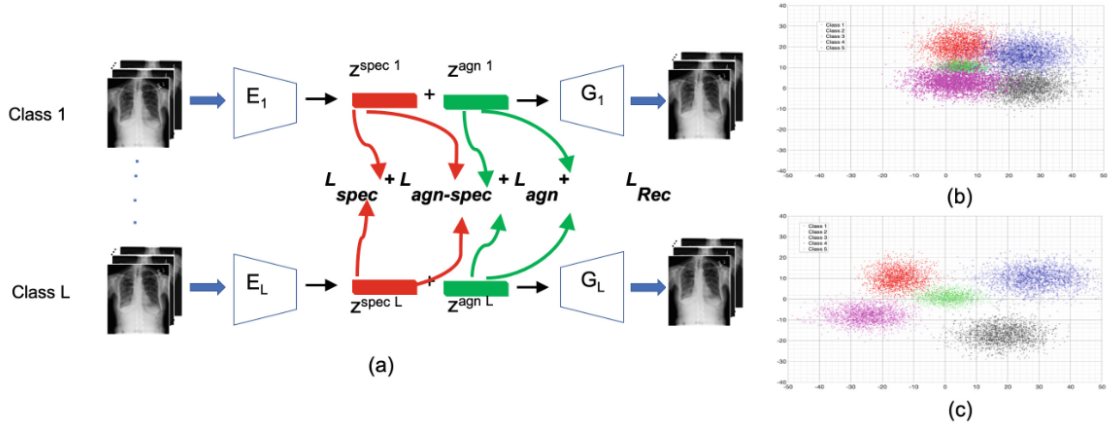


Figure 2.10: Class-specific encoder architecture and resulting separation of classes, taken from [MJYK⁺23]. Class-specific features are extracted by learning them along with class-agnostic features, which encapsulate unwanted common attributes.

[MBG21] present a transductive approach leveraging Self-Supervised Learning (SSL) to address the lack of knowledge on unseen classes. The methodology begins with the computation of centroids for seen classes utilizing SwAV [CMM⁺20].

Importantly, the authors extend this approach to calculate centroids for unseen classes through SSL with specific constraints. These constraints ensure that the centroids of unseen classes maintain a minimum difference from those of seen classes, and samples from seen classes exhibit maximum similarity with unseen class centroids.

A Generative Adversarial Network (GAN) [GPAM⁺14] generates synthetic features, with its discriminator distinguishing between seen and unseen classes based on features created by the generator using random noise and class label vectors (centroids). A pre-trained classifier is also incorporated to validate the correctness of synthetic features, ensuring alignment with real features. The final loss function is a composite of three components: Wasserstein adversarial loss [ACB17], SSL loss to maximize a sample's maximal cosine similarity with its corresponding centroid while minimizing with others, and pre-trained classifier loss.

To construct the training set, the synthesized target features of the unseen class are combined with the real and synthetic features of the seen class. Subsequently, a softmax classifier is trained by minimizing the negative log-likelihood loss.

This was followed up by [MGR22], where additional information is incorporated using saliency maps obtained through GradCAM [SCD⁺19].

2.4 Challenges

Having reviewed the literature around GZSL for the general and CXR domains, it is possible to note some challenges.

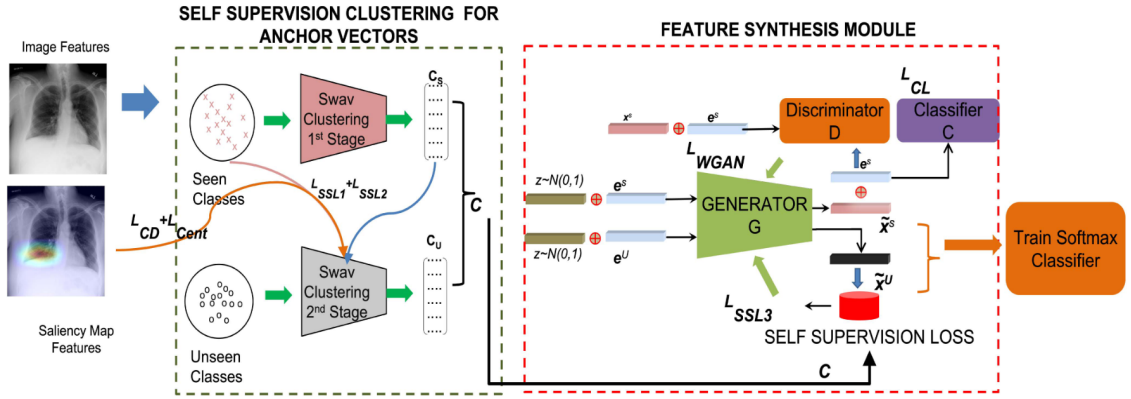


Figure 2.11: Proposed generative approach, taken from [MGR22]. SSL plays a heavy role in this method, being used both at clustering and generation times. [MBG21] follows the same architecture, using only image features.

Inductive GZSL presents a significant challenge in transmitting information from unseen classes using attribute vectors. This difficulty arises from the need to accurately characterize and convey distinctive features of unobserved classes without direct observations, all while maintaining performance on previously seen and trained classes. If addressing the issue through data augmentation, aligning the representation of source and target modalities becomes crucial, demanding meticulous training procedures. Transductive GZSL eases the problem's conditions; however, it falls short of addressing this challenge, primarily because test data is typically unavailable.

This challenge is amplified in the context of CXR classification, where the absence of readily available attribute vectors poses an additional hurdle. The task becomes more intricate as one must either craft attribute vectors for the unseen classes or explore alternative information sources to replace the unattainable attribute vectors. In the absence of a straightforward solution, addressing these challenges necessitates innovative approaches that can effectively bridge the gap between the known and unseen classes.

2.5 Summary and Discussion

In summary, this literature review extensively explored the general landscape of Generalized Zero-Shot Learning, with a particular focus on works applied to CXR classification.

A basic understanding was laid by dissecting key concepts, methodologies, and challenges within the broader domain of GZSL. This included an examination of inductive versus transductive GZSL approaches and a taxonomy categorizing methods into visual-to-semantic, semantic-to-visual, and mixed strategies. GZSL CXR classification methods were categorized as semantic spaces, embeddings, similarity metrics, and generative approaches tailored to medical image analysis.

The challenges section highlighted the complexity of effectively transmitting information from unseen classes using attribute vectors, particularly in inductive GZSL. This difficulty arises from

accurately characterizing distinctive features of unobserved classes without direct observations, emphasizing the importance of aligning source and target modalities during data augmentation, should that route be followed. In the context of CXR classification, the absence of readily available attribute vectors complicates the task, requiring innovative approaches to transport knowledge from seen to unseen classes.

Our approach seeks to address these challenges in a similar fashion to [MJYK⁺23]. By leveraging the inner representation of large pre-trained models, such as MedCLIP [WWAS22], to keep track of the relationships between seen and unseen classes and their common and diverging attributes, one can accommodate for the lack of examples from certain classes. Furthermore, by using a model pre-trained on in-domain data, the synthetic class attribute vectors extracted from it should be more fine-grained and tailored to the task at hand.

In closing, this literature review lays the groundwork for the subsequent chapters, offering a comprehensive understanding of the current state of GZSL in CXR classification and paving the way for further research and advancements in the intersection of generalized zero-shot learning and medical imaging.

Chapter 3

Methodology

This chapter is dedicated to this work’s methodology. We first present and analyze the current state-of-the-art approach as a means to introduce the driving ideas behind our reasoning. We follow by proposing our own method, broken down into loss functions, a two-phased training process and inference. A reflection is then made about the importance of the semantic embeddings for our current method, accompanied by a comparison of multimodal embeddings against the default BioBERT [LYK⁺19] choice prevalent in the literature. Finally, we introduce the model architectures for our experiments and discuss the pros and cons attached to each.

3.1 State-of-the-Art Method Breakdown

Mahapatra et al. [MJYK⁺23] represents the current state-of-the-art architecture for GZSL CXR classification. The following section presents a more detailed overview of this method than the one given in Subsubsection 2.3.1.2 while also providing our own critical analysis of the work. We believe this method can be described in three main steps. Subsection 3.1.1 details how the authors first apply feature disentanglement to deal with the high class-overlap within their data, while Subsection 3.1.2 shows how they generate new samples for seen and unseen classes and aggregate both real and synthetic features leveraging SwAV [CMM⁺20], an unsupervised learning method, and a set of externally provided similarity-based relationships between the target classes. Finally, Subsection 3.1.3 explains how this architecture operates at inference time.

3.1.1 Feature Disentanglement

CXR images rarely have a single label attached to them, as pathologies rarely show themselves alone. This complicates the extraction of features from a CXR image – trying to create feature vectors relating to a single class is made harder due to the constant simultaneous presence of multiple classes within the images.

To mitigate this, [MJYK⁺23] propose that a set of L autoencoders, where L is the number of seen classes, be trained to reconstruct CXR images while simultaneously being capable of

capturing latent features pertaining to each class and separating them from features common to all pathologies being studied.

In order to do this, the authors divide the latent encoding of each image into two sub-vectors, which they dub *class-specific features* and *class-agnostic features*. Then, taking all L encodings for a given batch of images, they promote inter-class similarity of their class-agnostic features while encouraging high intra-class similarity between class-specific features. At the same time, to reduce overlaps, they minimize the similarity between class-specific and class-agnostic features and discourage similarities between class-specific features of different classes.

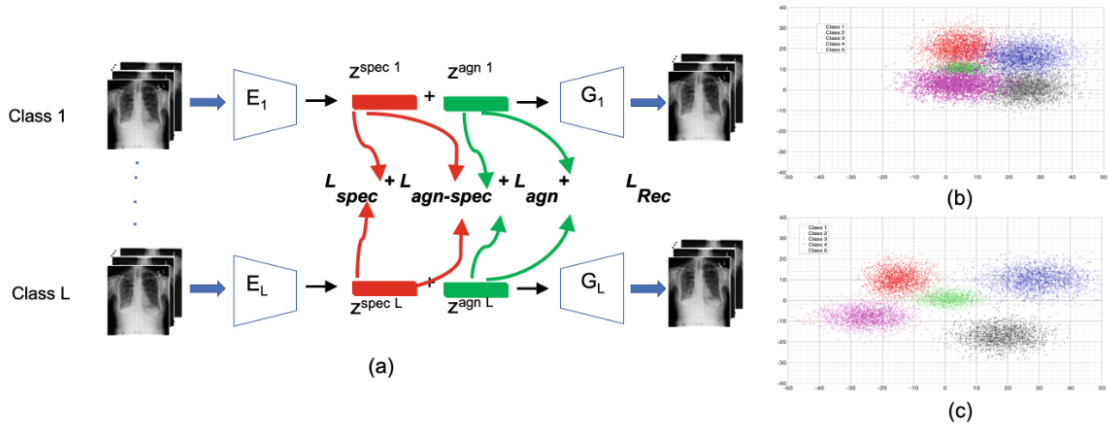


Figure 3.1: Feature disentanglement step of Mahapatra et al.'s method, taken from [MJYK⁺23]. By specializing the encoders in creating features that focus on each class's specificities, the authors are able to mitigate label overlapping problems.

Figure 3.1 (a) shows the architecture suggested by the authors. Here, z_{spec_n} and z_{agn_n} respectively refer to the class-specific and class-agnostic features obtained by E_n , the encoder associated with class n . Figure 3.1 (b) shows a t-SNE [vdMH08] plot of non-disentangled features from the fully connected layer of a DenseNet-121 [RIZ⁺17] model, while Figure 3.1 (c) shows the same experiment for the class-specific features from Mahapatra et al.'s approach, denoting how after this first step it becomes possible to operate at a class-level without being impacted by the data's high class-overlap.

3.1.2 Feature Generation

By encoding all images with the trained encoders, as seen through Figure 3.1 (c), each class can now be associated with a centroid by aggregating the latent vectors. However, these only cover the seen classes, leaving no information on unseen classes.

To create new centroids for the unseen classes, the authors synthesize new class-specific features by aggregating feature vectors sampled from each seen class distribution in a similar fashion to Mixup [ZCDLP18]. The new synthetic vectors are then clustered by use of the SwAV [CMM⁺20] unsupervised learning method.

While this process can create new centroids, it does not ensure that they are in any way related to any unseen classes. To solve this problem, the authors leverage BioBERT’s [LYK⁺19] embeddings to create a dictionary of similarity between each pair of classes. They then introduce a new loss function designed to align the similarity between each pair of centroids with the similarity of their corresponding embeddings in the dictionary.

3.1.3 Inference

Finally, the authors use the L autoencoders at test time to encode each image into L different class-specific vectors. Each one of these is then compared to each of the class centroids via cosine similarity. If any comparison shows a level of similarity superior to 0.5, the image is assigned the label of that class centroid.

3.1.4 Analysis

Mahapatra et al. [MJYK⁺23] present a three-stepped approach to the CXR GZSL problem, which tops every other GZSL method across 3 different datasets (NIH [WPL⁺17], CheXpert [IRK⁺19], and Padchest [BPSdIIV20]). From our interpretation, during the training phase, the authors propose three main driving ideas to guide our work.

The feature disentanglement step suggests disentanglement to mitigate the problems arising from high-class overlap in CXR images. As many pathologies often appear together, overlooking this idea leads to features pertaining to multiple classes, which is prejudicial to any class-level similarity-based approach.

On the other hand, the feature generation step presents a two-part loss, with the first component focused on clustering similar samples and the second on guiding each cluster to form semantically meaningful relationships with the others. Given the previously accomplished disentanglement, the first step aims to assign each class to its own region of latent space, while the second one ensures that each region is positioned in a way that is meaningful in relation to the others – thus allowing us to transpose seen class knowledge to the unseen classes.

We can sum up these ideas into the following definitions:

Definition 3 (Class-Overlap Mitigation Rule) *Given the high class-overlap in CXR images, features representing a single class should be disentangled in order to be considered meaningful.*

Definition 4 (Single-Class Region Rule) *Semantic embeddings should aggregate features via clustering in order to ensure that common patterns stick to common regions of the latent space.*

Definition 5 (Semantic Guidance Rule) *Semantic embeddings should follow the relationship between known sets of embeddings in order to ensure semantic meaningfulness within the latent space.*

These three ideas provide the basis for our work. However, we believe the methodology to implement these rules can be improved.

For one, we believe the number of encoders should match the total number of classes (seen and unseen). As each encoder learns about its assigned class, it has no concept of other unseen classes. Therefore, there is no assurance that the latent output of an encoder will resemble the embedding of any unseen class. This lack of similarity can increase the likelihood of false negatives, as the inference process relies on this comparison.

We also believe that the training process can be improved by directly shaping the latent space based on the known embedding relationship. Instead of disentangling the features to their own regions of space and then shifting those regions to match a given dictionary, we believe the encoders could learn to map directly to target regions in the first place, which reduces training complexity.

3.2 Method Overview

Based on our considerations on Mahapatra et al. [MJYK⁺23], we propose an inductive two-phased approach to the CXR GZSL task which leverages an externally provided set of embeddings. Our goal is to respect Definition 3, Definition 4 and Definition 5 while simultaneously addressing the concerns raised in Subsection 3.1.4. We will follow the notation from Section 2.2 to detail this method and complement it as it is further described.

3.2.1 Architecture

For our method, instead of using an autoencoder per seen class, we opt to expand the number of encoders to match the total number of classes, $|C|$, while making them share one single decoder. Matching the number of encoders with the number of classes aims to ensure that each class, seen or unseen, has an encoder capable of detecting its respective pathology and approximating its latent representation to the correct class embedding. This capacity is promoted by a two-parted loss function:

$$\mathcal{L} = \mathcal{L}_{reconstruction} + \mathcal{L}_{contrastive} \quad (3.1)$$

where $\mathcal{L}_{reconstruction}$ is the sum of the mean squared error between the original image, x , and the reconstructed image, $\hat{x}$:

$$\mathcal{L}_{reconstruction} = \frac{1}{|Y|} \sum_{y \in Y} (\hat{x}_y - x)^2 = (D(E_y(x)) - x)^2 \quad (3.2)$$

and $\mathcal{L}_{contrastive}$ is a cross-entropy loss function applied over the cosine similarity, denoted by $\langle \cdot, \cdot \rangle$, of the encoders' latent encodings, z , and their respective class embeddings, z_{emb} , both normalized:

$$\mathcal{L}_{contrastive} = - \sum_{y \in Y} \log \frac{\exp(\langle z_y, z_{emb_y} \rangle)}{\sum_{c=1}^{|C|} \exp(\langle z_c, z_{emb_c} \rangle)} \quad (3.3)$$

This loss depends on the sample's label set, Y – for any given example, the image is fed through each of the $|Y|$ encoders associated with each of the classes, y , present in Y . This produces a total of $|Y|$ latent encodings and, subsequently, $|Y|$ reconstructed images.

Training each encoder to specialize in a single class is a strategy designed to adhere to [Definition 3](#) and [Definition 4](#). Each encoder focuses on extracting features from its respective class, encoding them into a specific region of the latent space as defined by the corresponding class embedding.

Our choice of using a single decoder is an attempt to align with [Definition 5](#). Unlike the encoders, which are restricted to learning about their designated class, the shared decoder learns the entire latent space and understands the relationships between different classes within it. This provides semantic meaningfulness to the latent space, a quality the individual encoders lack.

3.2.2 Training Process

The first phase of our method focuses on training the seen class encoders, S_n , $n \in S$, and the decoder, as best illustrated in [Figure 3.2](#). The encoders associated with unseen classes are not used during this step, as no samples associated with them are found in our training set, i.e., $D_{train} = D_{set}$.

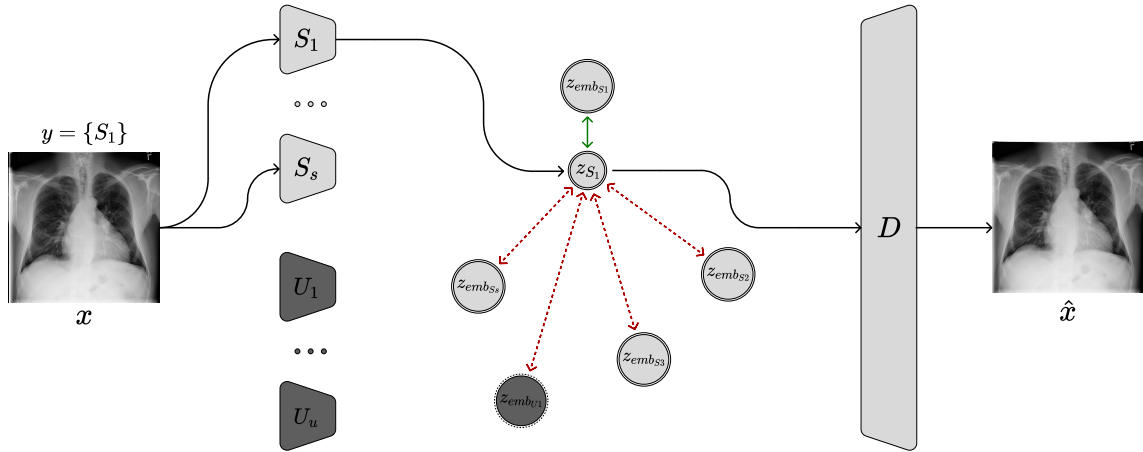


Figure 3.2: Block diagram of the proposed method's first training phase. The encoders associated with the classes in the sample's label set, Y , approximate their encodings of the image with their corresponding class embedding. The decoder reconstructs every encoding produced, thus learning about every class in the sample.

While this approach addresses the seen classes problem, neither the encoders nor the decoder have learned about the unseen classes so far. However, by shaping the latent space to follow the relationships defined by a known set of semantic embeddings, we hypothesize that the latent space will accommodate the unseen classes in a way that respects the same known relationships. This means that the regions of the latent space corresponding to the embeddings of unseen classes should accurately represent and maintain the relational structure established by the first training phase.

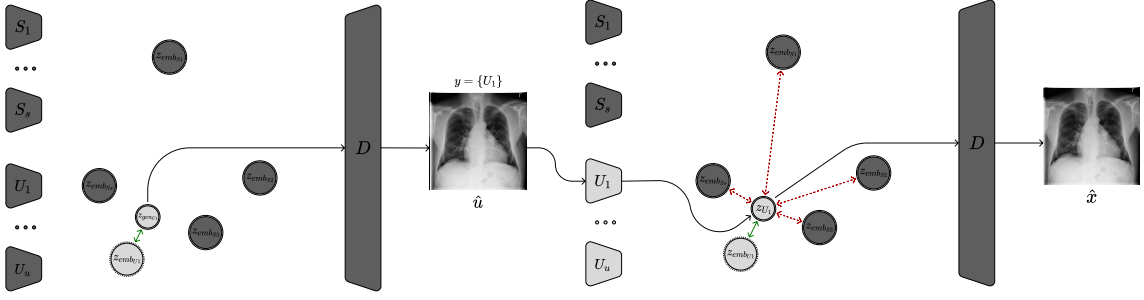


Figure 3.3: Block diagram of the proposed method’s second training phase. A sample similar to unseen class U_1 ’s semantic embedding is generated and decoded. Posteriorly, the sample is used to train the U_1 encoder. The seen class encoders are not used during this time, and the decoder is frozen.

By assuming this hypothesis, we are able to generate latent samples, z_{gen} , with high cosine similarity (≥ 0.95) to unseen class embeddings and decode them to generate images for the unseen classes. These single-label images are then used to train the encoders associated with each unseen class. During this training phase, the decoder is frozen in order to prevent it from unlearning what it acquired during its first training phase. This process is fully represented within Figure 3.3.

3.2.3 Inference

To classify a new sample, we pass the image through all $|C|$ encoders and compare the resulting encodings with the respective class embeddings via cosine similarity. If the similarity is greater than or equal to a threshold (we default to 0.5, as in [MJYK⁺23]), the sample is assumed to contain the class in question.

This is a slightly modified version of the inference procedure presented in Mahapatra et al. [MJYK⁺23], where instead of having a 1- $|C|$ comparison for each of the L encoders, we have a 1-1 comparison per encoding produced. An example is presented in Figure 3.4, where the sample is classified as belonging to classes S_1 and U_1 due to the high similarity between the latent vectors and the semantic embeddings.

3.3 On the Importance of Embeddings

Relying on a set of embeddings to shape the latent space of our models makes our model highly dependent on the quality of the embeddings. If the relationship between the embeddings does not properly reflect the relationship between the classes they represent, our model will not produce good results if our hypothesis proves to be correct.

In their work, Mahapatra et al. [MJYK⁺23] extract a set of embeddings from a BioBERT [LYK⁺19] model¹, a fine-tuned version of the BERT [DCLT19] model, which, in turn, is pre-trained on corpora encompassing a vast amount of topics. Therefore, while it may be specialised in the

¹ Available at <https://huggingface.co/dmis-lab/biobert-v1.1>

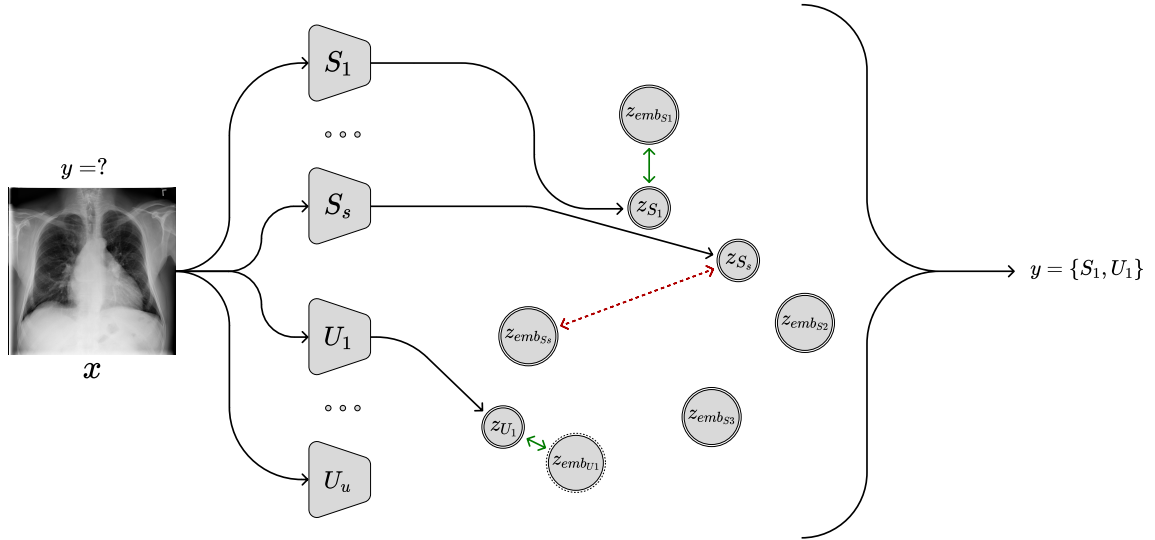


Figure 3.4: Test time example of our proposed method. A given sample is passed through each encoder, and the resulting latent output is compared with the corresponding class embedding via cosine similarity. For this example, classes S_1 and U_1 are signalled as present in the example.

biomedical domain, it also knows, to a lesser extent, about many other subjects. In Table 3.1, it is possible to see the cosine similarity between the embeddings of the BioBERT model for each of the pathologies addressed in CXR classification. We highlighted in blue the 5 lowest similarity scores found within the table and the 5 highest scores in red.

	Atl.	Card.	Cons.	Edema	Eff.	Emph.	Fibr.	Hernia	Inf.	Mass	No Find	Nodule	PT	Pne.	Pneu.
Atelectasis	1.000	0.839	0.926	0.918	0.656	0.988	0.772	0.990	0.927	0.925	0.494	0.704	0.790	0.986	0.889
Cardiomegaly	—	1.000	0.966	0.967	0.932	0.882	0.979	0.830	0.953	0.966	0.813	0.955	0.980	0.869	0.599
Consolidation	—	—	1.000	0.991	0.845	0.952	0.927	0.917	0.987	0.988	0.692	0.882	0.934	0.944	0.719
Edema	—	—	—	1.000	0.857	0.947	0.930	0.910	0.994	0.991	0.697	0.889	0.937	0.945	0.715
Effusion	—	—	—	—	1.000	0.705	0.963	0.646	0.841	0.851	0.915	0.982	0.951	0.700	0.400
Emphysema	—	—	—	—	—	1.000	0.820	0.985	0.953	0.951	0.544	0.754	0.832	0.992	0.863
Fibrosis	—	—	—	—	—	—	1.000	0.760	0.914	0.926	0.870	0.978	0.986	0.804	0.524
Hernia	—	—	—	—	—	—	—	1.000	0.919	0.915	0.476	0.696	0.777	0.986	0.908
Infiltration	—	—	—	—	—	—	—	—	1.000	0.986	0.676	0.873	0.921	0.954	0.732
Mass	—	—	—	—	—	—	—	—	—	1.000	0.701	0.884	0.937	0.945	0.720
No Finding	—	—	—	—	—	—	—	—	—	—	1.000	0.913	0.849	0.530	0.226
Nodule	—	—	—	—	—	—	—	—	—	—	—	1.000	0.969	0.744	0.453
Pleural_Thickening	—	—	—	—	—	—	—	—	—	—	—	—	1.000	0.817	0.539
Pneumonia	—	—	—	—	—	—	—	—	—	—	—	—	—	1.000	0.868
Pneumothorax	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1.000

Table 3.1: Cosine similarity matrix between BioBERT’s [LYK⁺19] embeddings. The 5 highest similarity scores (excluding similarity with self) are highlighted in red; the 5 lowest similarity scores are highlighted in blue.

We believe this forces all specialised domain knowledge into a corner of the latent space. Consequently, the representations of biomedical subjects, such as pathologies, inherently share similarities, even if, within their domain, they may differ a lot from each other. In turn, this results in coarser-grained features, which lead to worse performance when used to train a model.

To verify this hypothesis, we compare the BioBERT embeddings for the CXR classification pathologies to those of general words by reducing their dimensionality via t-SNE [vdMH08] and plotting them in their reduced form. Figure 3.5 shows the resulting plots for both 2D and 3D

vectors. In fact, CXR classification labels are mostly condensed into small spatial regions, which may lead to the high similarity scores found in Table 3.1.

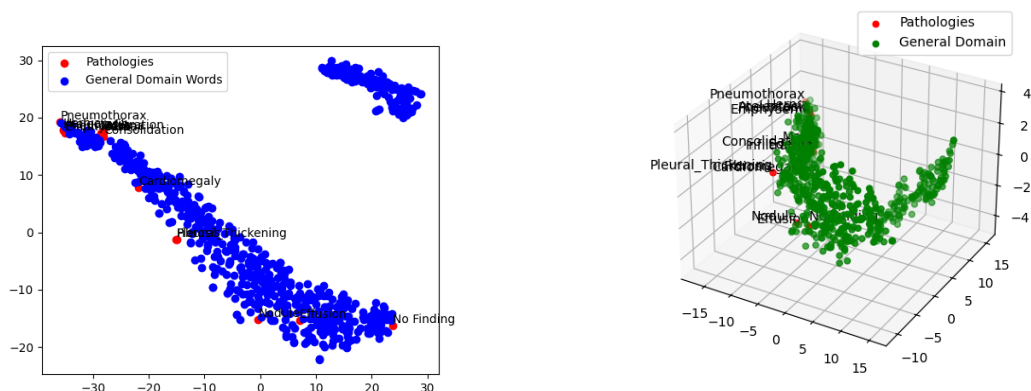


Figure 3.5: t-SNE [vdMH08] plot of BioBERT [LYK⁺19] embeddings for general words and CXR classification labels.

A second problem arises from the training domain. While Mahapatra et al. [MJYK⁺23] limit their use of the embeddings to the latent space, our approach leverages them to go from latent space to image space. While these class-representing vectors may map the semantic relationship between the pathologies, they may not account for visual differences between each disease as they were trained solely on textual data.

Therefore, while these embeddings provide a useful basis for comparison, we believe that multimodal embeddings, trained on both text and image spaces, are a more suitable candidate for the task.

3.3.1 Multimodal Embeddings

Ever since [RKH⁺21] introduced the CLIP model, there has been a significant focus on multimodal research combining vision and language [UBH⁺22]. Contrary to previous image models, CLIP jointly trains a text encoder and a vision encoder to predict a real (*image*, *text*) pair among fake pairs. This contrastive pre-training allows the model to learn a joint embedding space and associate visual concepts to their names. At test time, this can then be leveraged to match the encodings of textual labels to those of images, creating a zero-shot classifier from CLIP’s latent space.

However, CLIP is pre-trained on a massive dataset that contains a myriad of subjects [RKH⁺21]. As it aims to pick up general knowledge from the real world, there is no concern over finer-grained details that domain-specific models may have to pay attention to. Therefore, this approach is not transferable to the medical domain, where very similar text reports can point to entirely different diagnoses.

[WWAS22] propose training based on the semantic similarity of the multi-hot encodings of the images and texts instead. These vectors encapsulate the different pathologies present within a given data point. As illustrated by Figure 3.6, image-based vectors are created from their labels, and text-based ones are constructed using entity recognition to detect words related to the target classes.

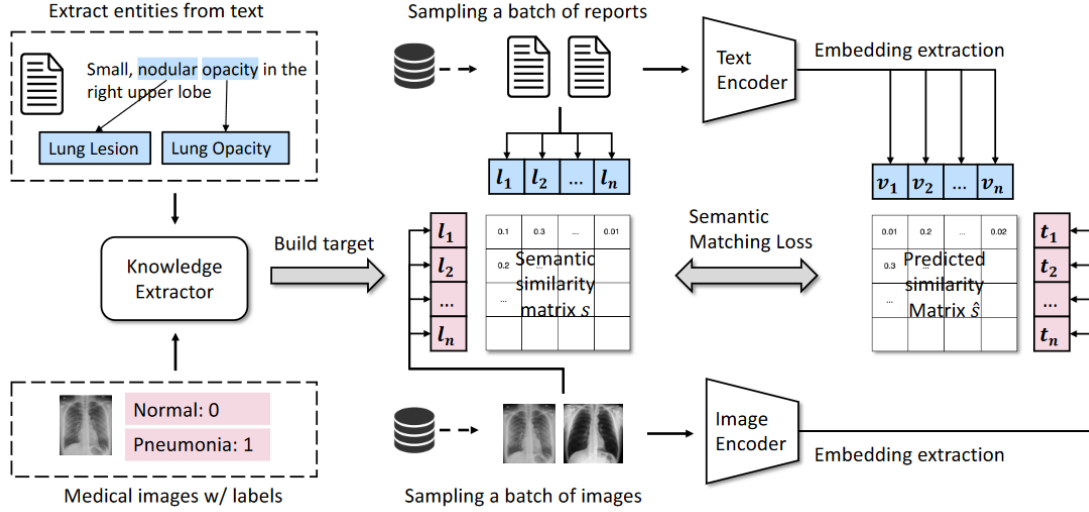


Figure 3.6: Training process for the MedCLIP model. A knowledge extractor first builds the multi-hot encodings for each entry to create a semantic similarity matrix. The text and image encoders are then trained to predict the similarity matrix between sample pairs. Taken from [WWAS22].

This allows the model to learn subtle differences between classes better, as false positives are mitigated through the usage of multi-hot vectors. At the same time, the need for image-text pairs datasets is eliminated, as it is now possible to mix textual and visual data from different datasets by converting both to vectors.

	Atl.	Card.	Cons.	Edema	Eff.	Emph.	Fibr.	Hernia	Inf.	Mass	No Find	Nodule	PT	Pne.	Pneu.
Atelectasis	1.000	0.575	0.766	0.706	0.770	0.722	0.749	0.760	0.737	0.729	0.759	0.688	0.667	0.733	0.746
Cardiomegaly	—	1.000	0.597	0.644	0.633	0.663	0.641	0.623	0.639	0.610	0.631	0.574	0.569	0.604	0.612
Consolidation	—	—	1.000	0.843	0.837	0.723	0.747	0.820	0.823	0.870	0.861	0.778	0.616	0.842	0.761
Edema	—	—	—	1.000	0.789	0.774	0.766	0.798	0.775	0.789	0.826	0.732	0.585	0.784	0.701
Effusion	—	—	—	—	1.000	0.713	0.721	0.748	0.823	0.776	0.852	0.749	0.626	0.751	0.793
Emphysema	—	—	—	—	—	1.000	0.837	0.867	0.732	0.763	0.749	0.744	0.738	0.765	0.711
Fibrosis	—	—	—	—	—	—	1.000	0.840	0.786	0.853	0.776	0.862	0.791	0.781	0.721
Hernia	—	—	—	—	—	—	—	1.000	0.789	0.835	0.829	0.792	0.735	0.858	0.717
Infiltration	—	—	—	—	—	—	—	—	1.000	0.766	0.837	0.741	0.651	0.751	0.755
Mass	—	—	—	—	—	—	—	—	—	1.000	0.841	0.915	0.701	0.842	0.742
No Finding	—	—	—	—	—	—	—	—	—	—	1.000	0.796	0.659	0.831	0.818
Nodule	—	—	—	—	—	—	—	—	—	—	—	1.000	0.750	0.784	0.714
Pleural_Thickening	—	—	—	—	—	—	—	—	—	—	—	—	1.000	0.663	0.707
Pneumonia	—	—	—	—	—	—	—	—	—	—	—	—	—	1.000	0.732
Pneumothorax	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1.000

Table 3.2: Cosine similarity matrix between MedCLIP’s [WWAS22] embeddings. The 5 highest similarity scores (excluding similarity with self) are highlighted in red; the 5 lowest similarity scores are highlighted in blue.

In sum, MedCLIP learns a joint embedding space trained on text and image data and is highly specialized in the domain-specific knowledge we require for the task at hand. Furthermore, as

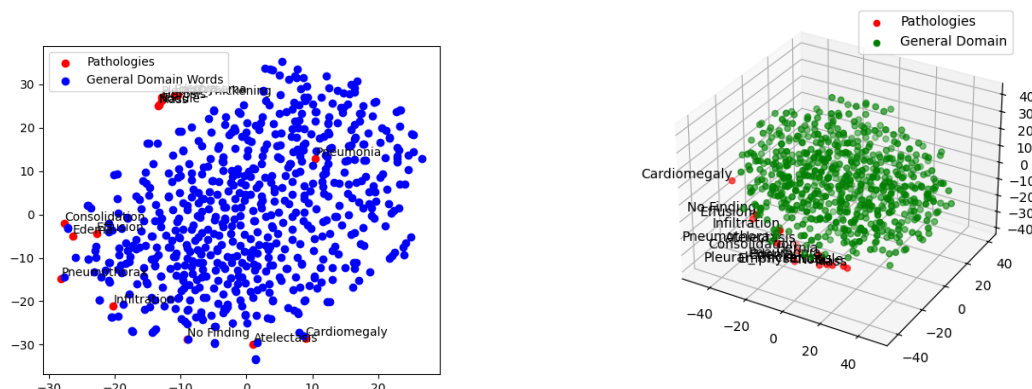


Figure 3.7: t-SNE [vdMH08] plot of MedCLIP [WWAS22] embeddings for general words and CXR classification labels.

it can be seen in Table 3.2, the similarity scores between pairs of pathologies are much more balanced, with scores rarely reaching values superior to 0.9. This means that MedCLIP can provide high-quality embeddings specialized in CXR pathologies trained on both text and image data, thus addressing all problems previously raised in Section 3.3.

A second comparison can be drawn by redoing the experiment presented in Figure 3.5 for MedCLIP. In our experiments, we default to the ResNet [HZRS16] version of the model². Figure 3.7 shows that a clearer separation between our target embeddings can be found when they are extracted from MedCLIP instead of BioBERT.

Therefore, we will compare the performance of our approach when using BioBERT and MedCLIP embeddings to verify whether changing to the latter’s provides any significant improvement.

3.4 Experimental Model Architectures

Section 3.2 has described the conceptual architecture for our proposed method. This section will describe two specific instances, illustrating how that conceptual architecture can be applied in practice.

3.4.1 Pooling Model

Following the footsteps of Mahapatra et al. [MJYK⁺23] once more, we take inspiration from the architecture they propose for their autoencoders to create our first model.

The encoders’ layers comprise a convolution followed by a max pooling [YSAF90] operation and are finished by a ReLU [Fuk69] activation function. After stacking 4 layers, whose outputs have 16, 32, 64, and 64 channels, we flatten the output of the final layer and reduce it to the

²Available at <https://github.com/RyanWangZf/MedCLIP>

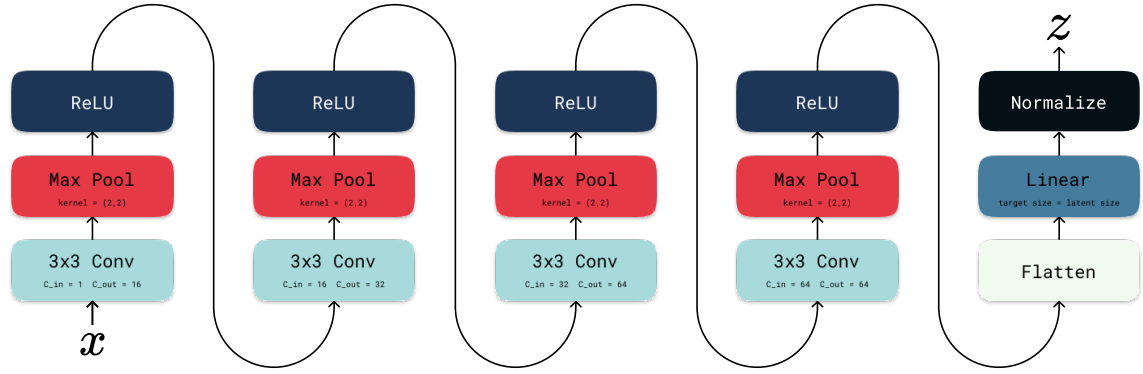


Figure 3.8: Encoder architecture for the first proposed model. Latent size here represents the number of dimensions of the class embeddings.

dimension of the class embeddings via a linear layer. Finally, the latent vector is normalized due to the semantic embeddings being on a normalized scale. This encoder architecture is best detailed in [Figure 3.8](#).

The decoder, presented in [Figure 3.9](#), follows a symmetric design for the most part – it has been extended (in comparison with the encoder) due to initial reconstruction experiments producing a lot of artefacts coming from the unpooling operations.

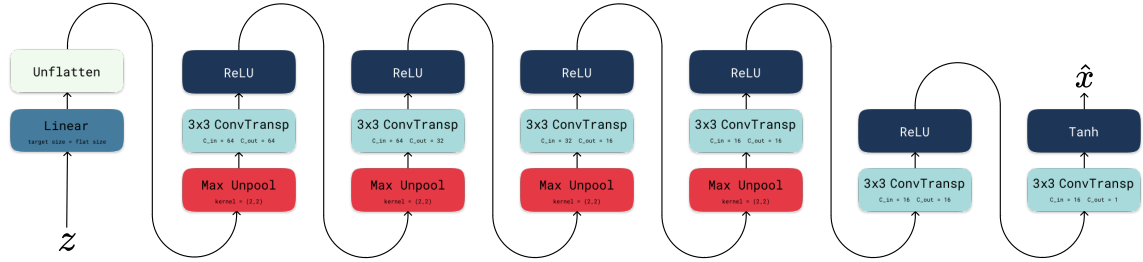


Figure 3.9: Decoder architecture for the first proposed model.

While this model type works for Mahapatra et al. [MJYK⁺23], it is incompatible with our proposed approach. Our second training phase, illustrated in [Figure 3.3](#), requires sampling from the latent space for subsequent decoding. However, since no image goes through any encoder, the indexes required for the decoder's max unpooling are not obtained, therefore making it impossible to reconstruct images from unseen classes and conclude the training process.

While it is possible to create random indexes and reconstruct images with them, the loss of quality is substantial. Therefore, we must resort to other operations to implement our approach.

3.4.2 Upsampling Model

To avoid the unpooling issue during the second training phase, we remove all pooling and unpooling operations and replace them with convolutional and upsampling layers instead. We set the convolutional operators' parameters to change the output's dimensions as if passed through the max pooling in [Figure 3.8](#). The resulting encoder architecture is presented in [Figure 3.10](#).

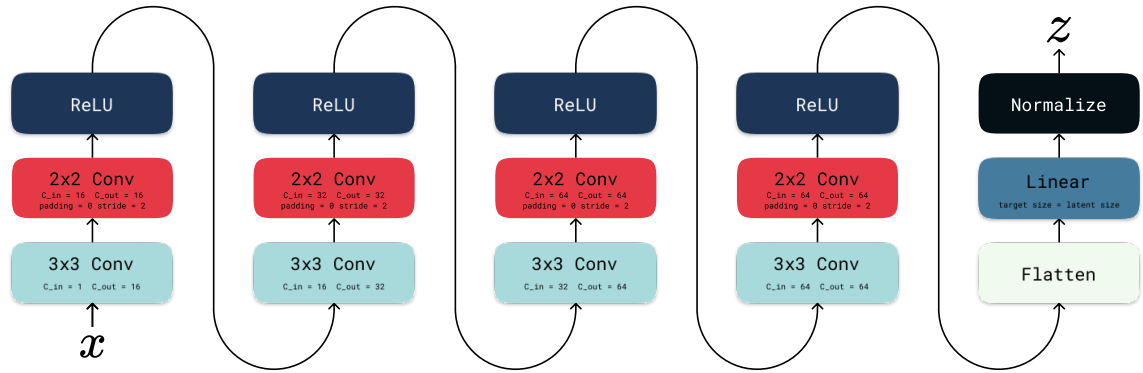


Figure 3.10: Encoder architecture for the second proposed model. Only the operations in red are modified, using convolutional operators instead of max pooling, as seen in Figure 3.8.

As for the decoder, showcased in Figure 3.11, we simply have to set the scale factor to 2 in order to reach similar output dimensions when compared to the max unpool operator. And, because we removed max unpool, we can also reduce the number of layers as we no longer have to deal with its artefacts.

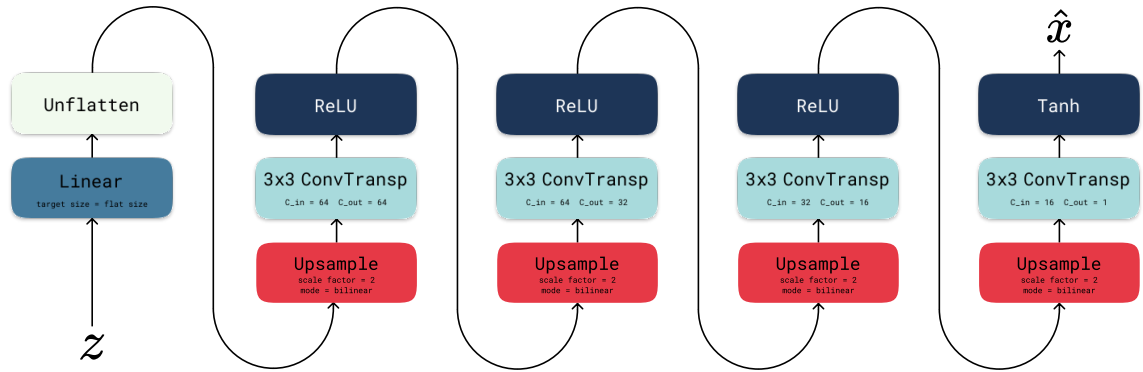


Figure 3.11: Decoder architecture for the second proposed model. Beyond the change from max unpooling to upsampling, the decoder is reduced as we no longer have to deal with unpooling artefacts.

This architecture solves the pooling indexes problem, but it puts the latent representation under a heavier weight, as the information carried over by the indexes now has to be incorporated into the latent vector in order to have a similar reconstruction quality.

Chapter 4

Results and Discussion

This chapter presents and discusses the results of implementing the method proposed in [Chapter 3](#). We introduce the dataset used in our experiments and analyse its class distribution. Following this, we examine the behaviour of different models and their variants during each training step. This analysis includes variables such as the loss evolution for the training and validation sets, the degree of class separation in the latent space, and the quality of reconstructed images. Finally, we evaluate each model using the test set and reflect on the obtained results.

4.1 Dataset

Our experiments are centred around the NIH dataset [\[WPL⁺17\]](#), first presented in [Subsection 2.1.1](#). As default in CXR classification literature, we consider only those with posteroanterior (PA) views of a patient’s thorax.

This leaves us with a total of 67310 samples, broken down into its classes within [Table 4.1](#). By further analysis of its class distribution, one can see two main problems within the dataset. Firstly, as expected from this classification task (see [Definition 3](#)), there is a significant class-overlap: excluding the *No Finding* class, which, by nature, cannot feature in multi-label samples, the percentage of single-label examples per-class ranges from 0.185 to 0.564.

All samples														
Atl.	Card.	Cons.	Edema	Eff.	Emph.	Fibr.	Hernia	Inf.	Mass	No Find	Nodule	PT	Pne.	Pneu.
5728	1563	1521	276	6589	1499	1408	192	9353	3567	39302	4177	2418	630	3407
Single-label samples														
Atl.	Card.	Cons.	Edema	Eff.	Emph.	Fibr.	Hernia	Inf.	Mass	No Find	Nodule	PT	Pne.	Pneu.
2210	746	346	51	2086	525	648	98	5270	1367	39302	1924	875	176	1506
Single-label samples percentage														
Atl.	Card.	Cons.	Edema	Eff.	Emph.	Fibr.	Hernia	Inf.	Mass	No Find	Nodule	PT	Pne.	Pneu.
0.386	0.478	0.228	0.185	0.317	0.350	0.460	0.511	0.564	0.382	1.000	0.460	0.362	0.280	0.442
Total number of samples														
67310														

Table 4.1: Class distribution for the NIH dataset [\[WPL⁺17\]](#). The high class-overlap in the dataset is shown by the low single-label samples percentage (at most 0.564, excluding *No Finding*, which appears alone by nature).

A second problem arises from the major disproportion in the number of samples across different classes. [Figure 4.1](#) better details how some classes dominate the dataset. For example, the *Infiltration* class (9353) has 48.7 times more samples than *Hernia* (192). While in some domains, this may be easily solvable by adding more data, we are limited by the natural occurrence of each pathology for this task.

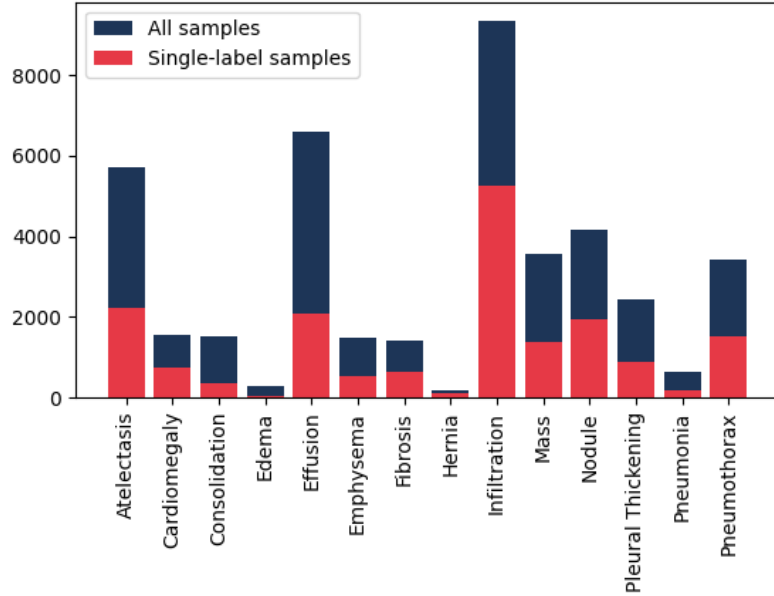


Figure 4.1: Class distribution for the NIH dataset [WPL⁺17], excluding *No Finding*. Beyond the high class-overlap, there is also a significant class imbalance problem, with some classes having more than 10 times the amount of samples than others (e.g., 9353 samples for *Infiltration* vs 192 for *Hernia*).

4.1.1 Training, Validation and Test Splits

Following Mahapatra et al. [MJYK⁺23], we create a 70/10/20 split for the training, validation, and test sets based on *Patient ID*. This approach ensures that different images of the same patient do not appear in multiple splits, thereby preventing data leakage and maintaining the integrity of the evaluation.

We define the set of seen classes, S , as *Atelectasis*, *Consolidation*, *Effusion*, *Infiltration*, and *Nodule*. These five classes comprise a large portion of the dataset (when excluding the *No Finding* class) while representing visually different pathologies. Both the training and validation sets are filtered to remove any samples which contain any class that isn't present within S . [Table 4.2](#) presents the two sets after filtering, which are to be used during the first training phase, described in [Subsection 3.2.2](#).

For the second training phase, described in [Subsection 3.2.2](#), we generate 1,000 samples for each unseen class, totalling 10,000 new examples for this new step.

Training Set					
	Atelectasis	Consolidation	Effusion.	Infiltration	Nodule
All samples	2786	633	2817	5166	2051
Single-label samples	1526	243	1457	3710	1338
Single-label samples percentage	0.548	0.384	0.518	0.718	0.652
Total number of samples	10647				
Validation Set					
	Atelectasis	Consolidation	Effusion.	Infiltration	Nodule
All samples	419	96	378	700	261
Single-label samples	244	33	198	521	178
Single-label samples percentage	0.582	0.343	0.524	0.744	0.682
Total number of samples	1482				

Table 4.2: Class distribution for the NIH dataset [WPL⁺17] training and validation splits.

Finally, we test our models against the test split, which contains both seen and unseen classes. Its details are described within Table 4.3.

All samples														
Atl.	Card.	Cons.	Edema	Eff.	Emph.	Fibr.	Hernia	Inf.	Mass	No Find	Nodule	PT	Pne.	Pneu.
1186	351	327	54	1333	341	309	34	1856	742	7890	891	474	123	743
Single-label samples														
Atl.	Card.	Cons.	Edema	Eff.	Emph.	Fibr.	Hernia	Inf.	Mass	No Find	Nodule	PT	Pne.	Pneu.
440	164	70	10	431	106	140	14	1039	284	7890	408	162	39	317
Single-label samples percentage														
0.371	0.467	0.214	0.185	0.323	0.311	0.453	0.412	0.560	0.383	1.000	0.458	0.342	0.317	0.427
Total number of samples														
13651														

Table 4.3: Class distribution for the NIH dataset [WPL⁺17] test split.

4.2 First Training Phase

We train each model for 50 epochs for the first training phase, using both losses right from the start. We batch samples into groups of 16 and decrease the learning rate by an order of magnitude in case the loss plateaus. A more complete list of hyperparameters and training details can be found Table A.1.

A detail worth noting is that the latent space dimensionality varies to match that of the semantic embeddings in use - for BioBERT [LYK⁺19], the latent space is comprised of 768 dimensions, while for MedCLIP [WWAS22] it is reduced to 512. Despite this being a necessity, it will impact the results in a non-negligible way.

4.2.1 Pooling Model

Figure 4.2 presents the evolution of the loss functions and the learning rate during the first training phase of our pooling model (detailed in Subsection 3.4.1). We showcase how the model

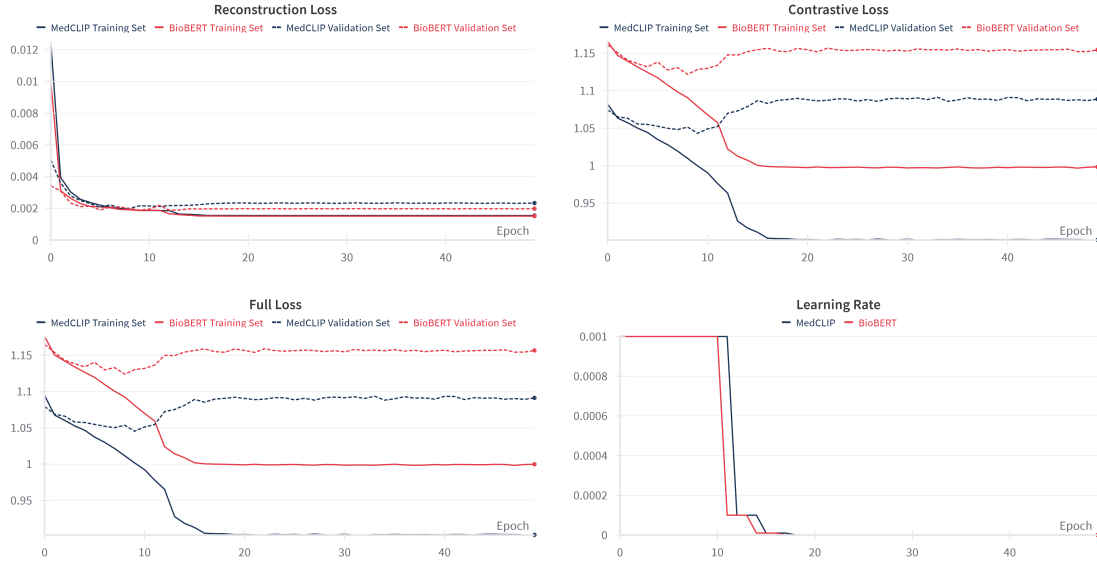


Figure 4.2: Loss evolution against training and validation sets for the pooling model during the first training phase. The loss behaves similarly for both variants, although using MedCLIP [WWAS22] embeddings leads to a marginal reduction of the contrastive loss.

performs against the training and validation sets while using either BioBERT [LYK⁺19] or MedCLIP [WWAS22] embeddings.

Upon analyzing the full loss, both MedCLIP [WWAS22] and BioBERT [LYK⁺19] versions exhibit similar behaviour, with MedCLIP [WWAS22] achieving marginally smaller loss values. The reconstruction loss is comparable for both versions, though slightly more favourable for BioBERT [LYK⁺19] – this is likely due to its embeddings having more dimensions than those of MedCLIP [WWAS22]. Conversely, MedCLIP [WWAS22] demonstrates superior performance on the contrastive loss, which may be attributed to the better separation of its embeddings, as seen in Table 3.2. Additionally, the learning rate for MedCLIP [WWAS22] plateaus later in the training process.

To verify whether the classes are correctly separated, we use t-SNE [vdMH08] visualizations of the latent encodings for each image by their respective class. Each image is encoded by the corresponding class encoders, resulting in multiple latent representations for visualization.

Figure 4.3 presents the display of the resulting 2D representations of each encoding for both model variants. While there is still significant overlap between some classes (like *Atelectasis* and *Effusion*), the separation is clear among others (e.g., *Atelectasis* and *Infiltration*). There isn't any meaningful distinction between the visualization of different variants, except for the BioBERT [LYK⁺19] one having some classes with slightly more variance.

Finally, as a sanity check, we manually verify the realism of the reconstructed images by selecting a random image from the training set, passing it through each encoder, and reconstructing each latent vector using the shared decoder. Figure 4.4 displays the resulting images produced by each model variant. Each image includes the label of the associated encoder and an indication of

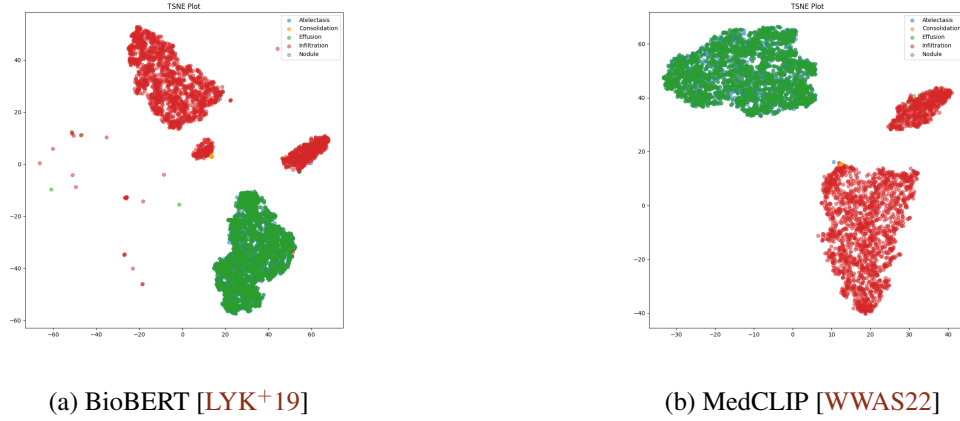


Figure 4.3: t-SNE visualizations of the latent encodings of the pooling model during the first training phase. No significant difference is found between variants.

whether that class appears in the original example. We observe no significant differences between the two sets of generated images, which aligns with the similar reconstruction loss observed for the two variants (as shown in Figure 4.2).

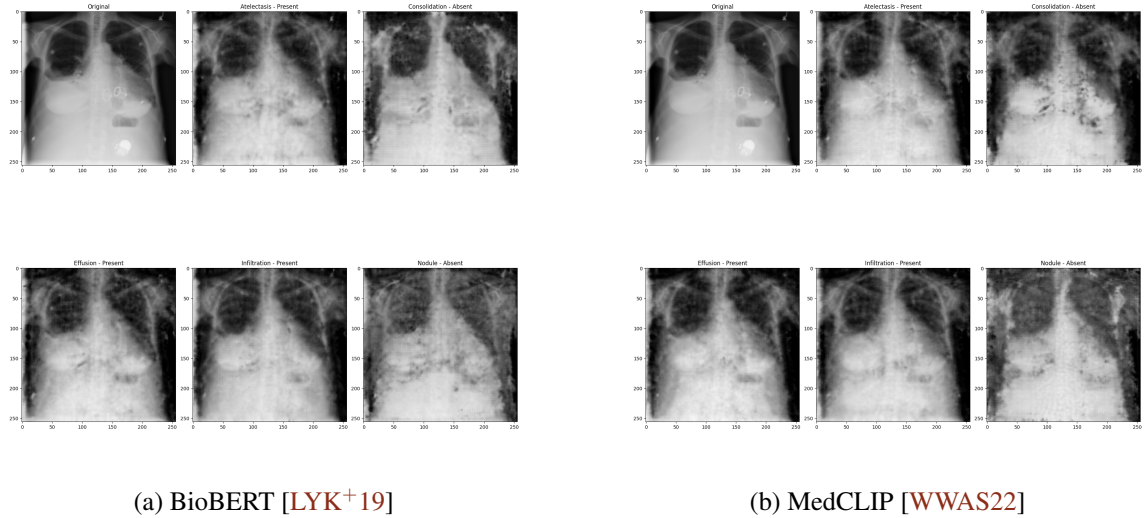


Figure 4.4: Reconstruction of a random training set sample by the pooling model. The labels at the top indicate the associated class encoder and whether or not that class appeared in the original sample.

4.2.2 Upsampling Model

Figure 4.2 presents the evolution of the loss functions and the learning rate during the first training phase of our upsampling model (detailed in Subsection 3.4.2). The conclusions established in Subsection 4.2.1 still hold for this model – MedCLIP [WWAS22] embeddings yield

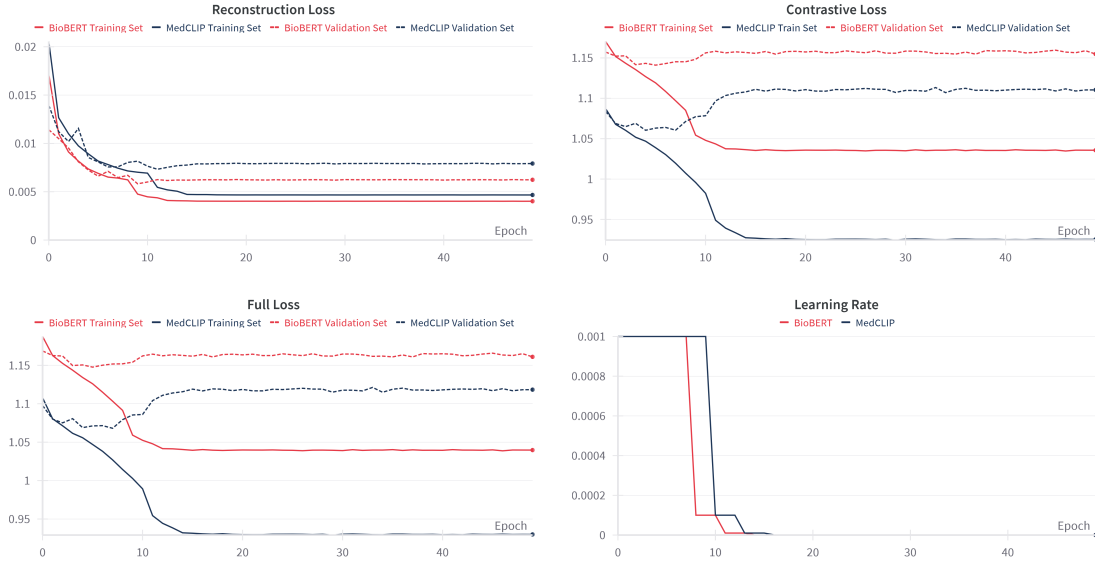


Figure 4.5: Loss evolution against training and validation sets for the upsampling model during the first training phase.

marginally better results on the contrastive loss, thus decreasing the overall loss, while BioBERT [LYK⁺19] has a minor advantage on the reconstruction loss.

A more relevant conclusion arises when measuring both models against one another – by directly comparing Figure 4.2 and Figure 4.5, one can verify that the pooling model outperforms the upsampling one across all losses.

For the t-SNE [vdMH08] experiments, shown for the upsampling model in Figure 4.6, the results are also similar to those of Subsection 4.2.1, albeit for the fact that both variants now show some minor variance on the latent representations for some of their classes.

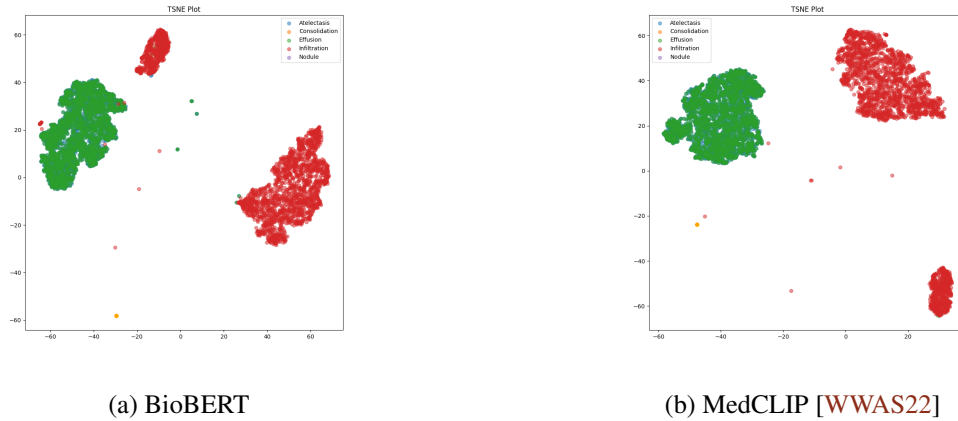


Figure 4.6: t-SNE visualizations of the latent encodings of the upsampling model during the first training phase. No significant difference is found between variants.

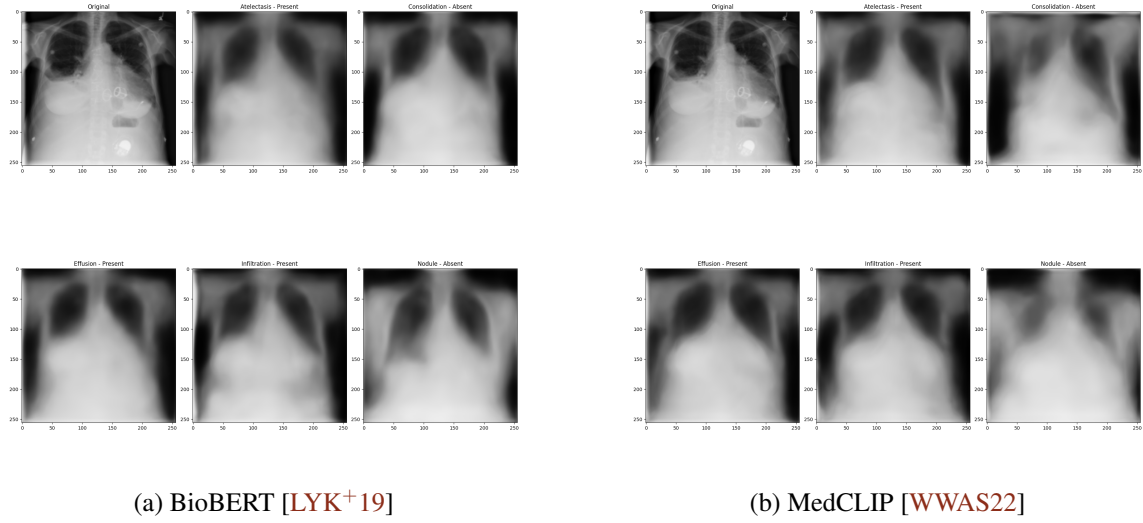


Figure 4.7: Reconstruction of a random training set sample by the upsampling model. The labels at the top indicate the associated class encoder and whether or not that class appeared in the original sample. In contrast with the pooling model, the upsampling model loses all details associated with its label set.

The key difference between the models is illustrated in Figure 4.7. Although their reconstruction losses differ slightly, the upsampling model loses significantly more detail than the pooling model, whose reconstructions are shown in Figure 4.4. This loss of detail is likely the primary reason for the upsampling model’s poorer overall performance. In turn, we believe this stems from removing the pooling operations, which forces the latent encodings to carry both semantic and visual information without increasing their dimensionality.

4.3 Second Training Phase

As per Subsection 3.2.2, we generate 1,000 samples for each unseen class using the shared decoder. We will call this set of images the generated set to avoid ambiguity. All examples within the generated set are single-label.

The hyperparameters for this second training phase remain mostly unchanged when compared to those of the first, found in Table A.1. The sole exceptions are the batch size, reduced to 12 to fit memory constraints when training the pooling model, and the starting learning rate, which was set to $7.5e^{-4}$ to match the batch size change.

4.3.1 Pooling Model

As indicated in Subsection 3.4.1, it is impossible for the pooling model to reconstruct images without having the indexes stemming from one of its encoders. Thus, in order to make image generation possible for this model, we generate random indexes for each max unpool operation

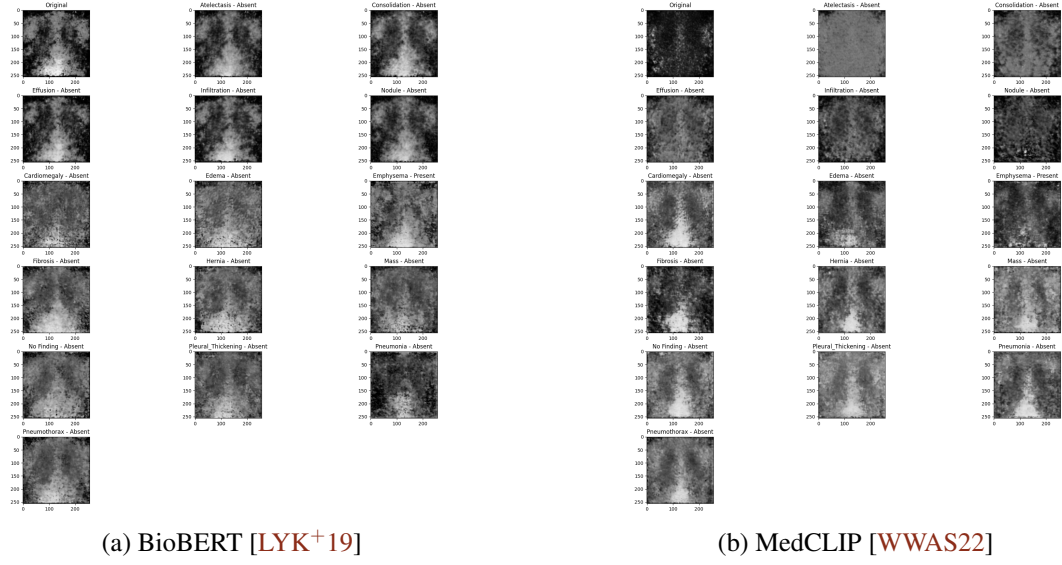


Figure 4.8: Reconstruction of a random sample from the generated set by the pooling model. Due to the use of random indexes, the pooling model creates heavily noisy images.

during the decoding process. As verifiable by Figure 4.8, this leads to poor image quality but allows us to proceed with training.

Figure 4.9 presents the evolution of the loss functions and the learning rate during the second training phase of our pooling model. Likely due to the poor quality of the generated images, the reconstruction loss did not improve during training. The contrastive loss, however, shows that better results are achieved for the BioBERT [LYK+19] model during this phase. Regardless, due to the weak reconstruction of images, we are unable to conclude the advantage of one variant over the other.

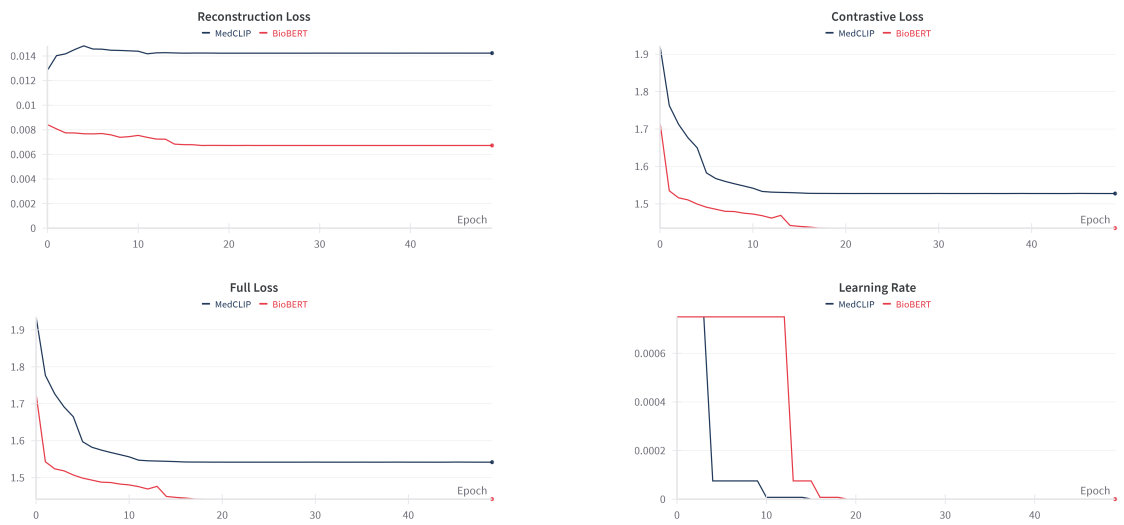


Figure 4.9: Loss evolution against the generated set for the pooling model during the second training phase.

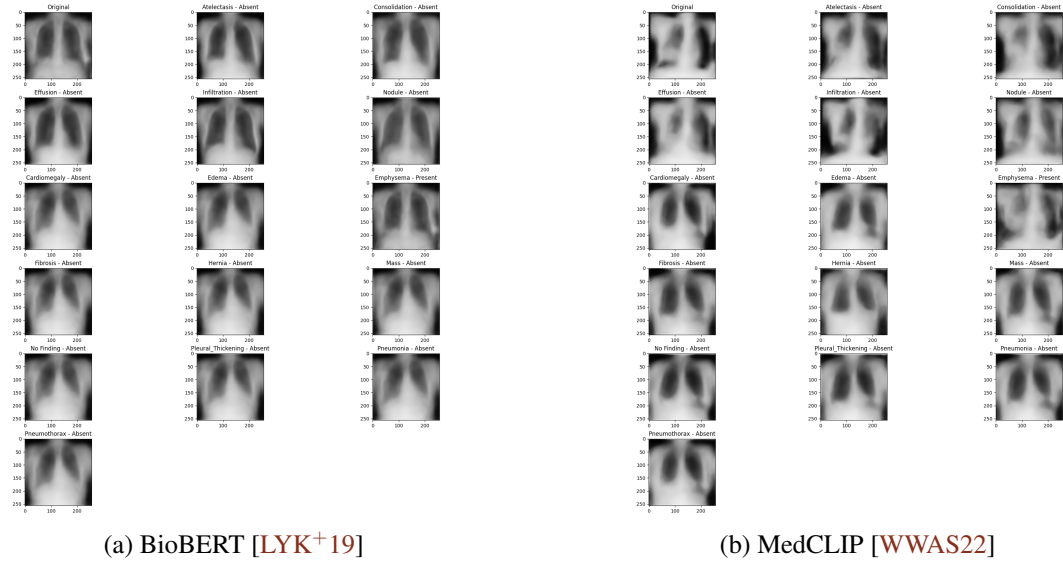


Figure 4.10: Reconstruction of a random sample from the generated set by the upsampling model.

4.3.2 Upsampling Model

Due to its architecture, the upsampling model poses no problem in image generation. Despite not having as much detail as the original dataset, the reconstructed images, seen in Figure 4.10, have structure to them, in contrast with those of the pooling model (found in Figure 4.8).

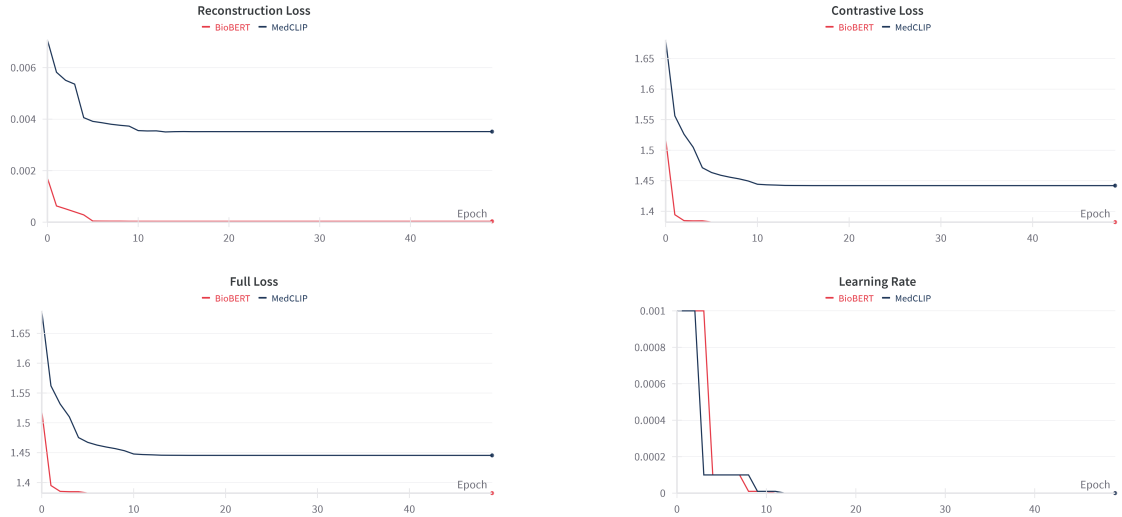


Figure 4.11: Loss evolution against the generated set for the upsampling model during the second training phase.

By looking at Figure 4.11, one can see that, in contrast with the pooling model (whose losses are found in Figure 4.9), the upsampling model's reconstruction loss decreases during training. This is to be expected as the training data no longer has a random nature to it, as the images reconstructed by the pooling model did, due to the use of random indexes.

The contrastive loss demonstrates superior results for BioBERT [LYK⁺19]. However, despite the improved quality of the generated images, crucial pathological details are lost during reconstruction. Therefore, these results should be interpreted cautiously – if all images lack any pathology, accurately distinguishing between classes based on the samples alone would be impossible.

4.4 Test Results

Finally, we present the results on the test set in Table 4.4. For inference, we use the approach described in Subsection 3.2.3. For each model and its variants, we report the average accuracy for both seen and unseen classes and their harmonic mean, facilitating comparison with other works.

	Pooling BioBERT	Pooling MedCLIP	Upsampling BioBERT	Upsampling MedCLIP
Accuracy _S	0.741	0.700	0.747	0.704
Accuracy _U	0.818	0.778	0.808	0.903
Harmonic Mean	0.778	0.737	0.777	0.792
Per-class Accuracy				
Atelectasis	0.767	0.704	0.772	0.716
Consolidation	0.975	0.973	0.976	0.976
Effusion	0.820	0.776	0.785	0.770
Infiltration	0.341	0.292	0.298	0.277
Nodule	0.801	0.752	0.904	0.782
Cardiomegaly	0.520	0.502	0.901	0.963
Edema	0.807	0.169	0.797	0.994
Emphysema	0.965	0.943	0.963	0.974
Fibrosis	0.925	0.969	0.825	0.974
Hernia	0.997	0.996	0.920	0.980
Mass	0.710	0.850	0.611	0.916
No Finding	0.429	0.472	0.424	0.423
Pleural Thickening	0.914	0.959	0.899	0.946
Pneumonia	0.965	0.977	0.955	0.926
Pneumothorax	0.945	0.939	0.788	0.937

Table 4.4: Test results, describing average accuracy for seen and unseen classes and their harmonic mean. Per-class accuracy is also presented, first for the seen classes and then for the unseen classes. Each line’s top score is highlighted in blue.

Overall, the MedCLIP [WAS22] variant of the upsampling model performed best – it reached the best harmonic mean between the seen and unseen classes accuracy and performed best on 6 out of the 15 target classes.

BioBERT [LYK⁺19] variants consistently demonstrate superior performance on the seen classes. This can be attributed to their higher-dimensional latent space, significantly enhancing their data

reconstruction ability. This improved reconstruction capability, in turn, leads to the generation of more robust features, particularly because we can utilize real data for these classes instead of synthetic data.

Since the generation process of the pooling model creates poor-quality data for its second training phase, we believe unseen class accuracy for its variants should not be considered. However, when analyzing the upsampling model, we can see that this metric significantly improves when swapping the BioBERT [LYK⁺19] for the MedCLIP [WWAS22] embeddings, with a slight decrease in seen class accuracy. We believe this to be a positive sign that MedCLIP [WWAS22] embeddings, despite their smaller dimensionality, are more suited for this problem.

	Accuracy _S	Accuracy _U	Harmonic Mean
Ours (Upsampling MedCLIP)	0.704	0.903	0.792
Mahapatra et al. [MJYK ⁺ 23]	0.862	0.850	0.856

Table 4.5: Comparison against state-of-the-art work by Mahapatra et al. [MJYK⁺23] (as reported in their original work). Each metric’s top score is highlighted in blue.

Table 4.5 shows a direct comparison against Mahapatra et al. [MJYK⁺23], from where our work takes inspiration. Despite the harmonic mean still being significantly lower than our target, our best model surpasses their unseen class accuracy by more than 5%. The scores from Mahapatra et al. [MJYK⁺23] are as reported, as we are unable to replicate their experiments.

Chapter 5

Conclusions and Future Work

Over the past decade, deep learning has made significant strides in tackling a wide range of medical diagnostic tasks, with its performance continually improving each year. This progress allows healthcare professionals to spend more time directly attending to patients as medical systems increasingly integrate advanced artificial intelligence technologies.

In this work, we focus on one application in this field, CXR classification. In this task, a model analyzes an image of a patient’s thorax and accurately identifies the pathologies present. While initially focused on single-label categorization, CXR classification research has moved to multi-label classification to account for the multiple simultaneous conditions that patients often present. Due to the frequent lack of labelled data, novel techniques have been developed to address this challenge within the framework of GZSL. In this setting, a model must accurately predict the labels of an image during testing, even if it has not encountered examples of those classes during training.

GZSL has been thoroughly explored within more general domains by leveraging attribute vectors (also known as semantic embeddings) – a set of features that unambiguously define a class. These methods usually learn how these attribute vectors relate to their respective classes in order to later exploit this relationship when dealing with unseen classes. This is usually done to augment the training set to contain samples from novel classes or to be able to approximate samples to representations of related classes.

However, it is not feasible to use attribute vectors for CXR classification as their handcrafted nature requires extensive amounts of time from medical professionals – which is the motivation for computer-aided CXR classification in the first place. Thus, recent works have focused on alternative approaches. Mahapatra et al. [MJYK⁺23] proposed the current state-of-the-art work on this task, which creates class-specific features and clusters them into centroids by disentangling the different classes through a series of autoencoders. They then generate synthetic features to compute centroids for the unseen classes and ensure their semantic relevance by approximating them to the relationship between BioBERT’s [LYK⁺19] embeddings. Their inference then relies on verifying the similarity between a sample’s encoding and the class centroids.

We took inspiration from this work and drew several guiding rules for our own. We believe that

the high-class overlap must be addressed and that clustering features within the latent space must follow two rules: similar images have to be aggregated to the same region, and semantic relevance has to be ensured by leveraging a known set of relationships between attribute vectors. We also found some concerns with Mahapatra et al.’s [MJYK⁺23] work, such as the mismatch between the number of encoders and the total number of classes and the usage of BioBERT [LYK⁺19] embeddings, who share high similarity scores amongst them.

Drawing from all this, we proposed our method: a two-phased training process which directly maps images into fixed locations of the latent space via a series of encoders and ensures the meaningfulness of the features through one single, shared decoder. The first step of our approach involves training with data from the seen classes, so the decoder may learn about the relationship between them. Then, we decode from regions of the latent space where semantic embeddings from unseen classes reside, in order to generate new samples to train the encoders associated with the unseen classes. Finally, we apply the same inference as Mahapatra et al. [MJYK⁺23], with the small caveat that we do a 1:1 comparison between the output of an encoder with the semantic embedding of the associated class.

On the other hand, to address our concerns with BioBERT [LYK⁺19], we propose instead the usage of MedCLIP [WWAS22] as the embedding provider, given its specialized domain knowledge and lesser inter-class similarity between embeddings.

We test our method using the two different model architectures, each with two variants (one per embedding provider), and find that, while BioBERT [LYK⁺19] provides a better reconstruction ground with its higher-dimensional embeddings, which in turn leads to better performance on seen classes, MedCLIP [WWAS22] better generalized to the unseen classes despite its smaller embeddings. Thus, we believe that MedCLIP is a better alternative to the CXR GZSL multi-label classification task, given that it better transposes the knowledge from seen to unseen classes.

Although our experiments allowed us to draw these conclusions, we believe that these should be taken with a grain of salt: for the second phase of training, both models were unable to reconstruct the original image with the details pertaining to the pathologies it contained. Therefore, it is unclear to us if our main hypothesis (described in Subsection 3.2.2) holds true. However, the results from the upsampling model, especially when dealing with the unseen classes, seem to be a positive signal in that direction. Therefore, this work represents an initial step towards simplifying the GZSL methodology for multi-label CXR classification. By leveraging embeddings right from the start, we present a more straightforward training process compared to the current state-of-the-art.

Based on our experiments, we recommend improving reconstruction quality for doubt-free verification of our hypothesis as future work. Moreover, if generalization for unseen classes performs as well as in our upsampling model experiments, better reconstruction should enhance performance with seen classes, bringing our approach closer to state-of-the-art results.

One research avenue may explore the distillation [HVD15] of the pooling model – which effectively reconstructs the original dataset images with their pathology details – into a model that

does not rely on unpooling operations, thus allowing for the second training phase to be carried out.

Another research direction could include the improvement of the embeddings used during training. Neither BioBERT [LYK⁺19] nor MedCLIP [WWAS22] perfectly separate all the classes into a completely disentangled latent space. It would be beneficial for this task to have a high-quality set of embeddings that distinguish between all classes unambiguously, allowing this problem to no longer require attribute vectors. As an example, instead of basing the embeddings used solely on the class label (as done per default in the literature), one could use the pathology description from the Unified Medical Language System [Bod04] to incorporate more information into these representations.

References

- [ACB17] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223. PMLR, 06–11 Aug 2017.
- [Act96] Accountability Act. Health insurance portability and accountability act of 1996. *Public law*, 104:191, 1996.
- [APHS13] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for attribute-based classification. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 819–826, 2013.
- [Bod04] Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32 Database issue:D267–70, 2004.
- [BPSdlIV20] Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria de la Iglesia-Vayá. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical Image Analysis*, 66:101797, December 2020.
- [CMM⁺20] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *CoRR*, abs/2006.09882, 2020.
- [DCLT19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.
- [FCS⁺13] Andrea Frome, Greg S. Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc’Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2, NIPS’13*, page 2121–2129, Red Hook, NY, USA, 2013. Curran Associates Inc.
- [FHXG15] Yanwei Fu, Timothy M. Hospedales, Tao Xiang, and Shaogang Gong. Transductive multi-view zero-shot learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(11):2332–2345, nov 2015.

- [Fuk69] Kunihiro Fukushima. Visual feature extraction by a multilayered network of analog threshold elements. *IEEE Transactions on Systems Science and Cybernetics*, 5(4):322–333, 1969.
- [GBR⁺12] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.
- [GPAM⁺14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.
- [HE20] Dat Huynh and Ehsan Elhamifar. A shared multi-attention framework for multi-label zero-shot learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8773–8783, 2020.
- [HLS21] Nasir Hayat, Hazem Lashen, and Farah E. Shamout. Multi-label generalized zero shot learning for the classification of disease in chest radiographs. In Ken Jung, Serena Yeung, Mark Sendak, Michael Sjöding, and Rajesh Ranganath, editors, *Proceedings of the 6th Machine Learning for Healthcare Conference*, volume 149 of *Proceedings of Machine Learning Research*, pages 461–477. PMLR, 06–07 Aug 2021.
- [HSMN12] Eric Huang, Richard Socher, Christopher Manning, and Andrew Ng. Improving word representations via global context and multiple word prototypes. In Haizhou Li, Chin-Yew Lin, Miles Osborne, Gary Geunbae Lee, and Jong C. Park, editors, *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 873–882, Jeju Island, Korea, July 2012. Association for Computational Linguistics.
- [HTHS17] Yao-Hung Hubert Tsai, Liang-Kang Huang, and Ruslan Salakhutdinov. Learning robust visual-semantic embeddings. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.
- [HVD15] Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*, 2015.
- [HWYW19] He Huang, Changhu Wang, Philip S. Yu, and Chang-Dong Wang. Generative dual adversarial network for generalized zero-shot learning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 801–810, 2019.
- [HZRS16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.
- [IRK⁺19] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghighi, 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.
- [JPB⁺19] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019.
- [KW14] Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.
- [KW17] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.
- [KXG17] Elyor Kodirov, Tao Xiang, and Shaogang Gong. Semantic autoencoder for zero-shot learning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4447–4456, 2017.
- [LBBH98] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- [LLWJ18] Shichen Liu, Mingsheng Long, Jianmin Wang, and Michael I. Jordan. Generalized zero-shot learning with deep calibration network. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems, NIPS’18*, page 2009–2019, Red Hook, NY, USA, 2018. Curran Associates Inc.
- [LYK⁺19] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 09 2019.
- [MBG21] Dwarikanath Mahapatra, Behzad Bozorgtabar, and Zongyuan Ge. Medical image classification using generalized zero shot learning. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 3337–3346, 2021.
- [MGR22] Dwarikanath Mahapatra, Zongyuan Ge, and Mauricio Reyes. Self-supervised generalized zero shot learning for medical image classification using novel interpretable saliency maps. *IEEE Transactions on Medical Imaging*, 41(9):2443–2456, 2022.
- [MJYK⁺23] Dwarikanath Mahapatra, Antonio Jose Jimeno Yepes, Shiba Kuanar, Sudipta Roy, Behzad Bozorgtabar, Mauricio Reyes, and Zongyuan Ge. Class specific feature disentanglement and text embeddings for multi-label generalized zero shot cxr classification. In Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu Salcudean, James Duncan, Tanveer Syeda-Mahmood, and Russell Taylor, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 276–286, Cham, 2023. Springer Nature Switzerland.
- [MRMM18] Ashish Mishra, Shiva Krishna Reddy, Anurag Mittal, and Hema A. Murthy. A generative model for zero shot learning using conditional variational autoencoders. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2269–22698, 2018.

- [PSB⁺20] Angshuman Paul, Thomas C. Shen, Niranjana Balachandrar, Yuxing Tang, Yifan Peng, Zhiyong Lu, and Ronald M. Summers. Come-see: Cross-modality semantic embedding ensemble for generalized zero-shot diagnosis of chest radiographs. In Jaime Cardoso, Hien Van Nguyen, Nicholas Heller, Pedro Henriques Abreu, Ivana Isgum, Wilson Silva, Ricardo Cruz, Jose Pereira Amorim, Vishal Patel, Badri Roysam, Kevin Zhou, Steve Jiang, Ngan Le, Khoa Luu, Raphael Sznitman, Veronika Cheplygina, Diana Mateus, Emanuele Trucco, and Samaneh Abbasi, editors, *Interpretable and Annotation-Efficient Learning for Medical Image Computing*, pages 103–111, Cham, 2020. Springer International Publishing.
- [PSL⁺21] Angshuman Paul, Thomas C. Shen, Sungwon Lee, Niranjana Balachandrar, Yifan Peng, Zhiyong Lu, and Ronald M. Summers. Generalized zero-shot chest x-ray diagnosis through trait-guided multi-view semantic embedding with self-training. *IEEE Transactions on Medical Imaging*, 40(10):2642–2655, 2021.
- [PSM14] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics.
- [PSP⁺22] Chinmay Prabhakar, Anjany Sekuboyina, Johannes C. Paetzold, Hongwei Li, Tamaz Amiranashvili, Jens Kleesiek, and bjoern menze. Improving generalized zero-shot learning for multi-label chest x-ray classification using knowledge graphs, 2022.
- [RIZ⁺17] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Yi Ding, Aarti Bagul, Curtis P. Langlotz, Katie S. Shpanskaya, Matthew P. Lungren, and Andrew Y. Ng. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *CoRR*, abs/1711.05225, 2017.
- [RKH⁺21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021.
- [RPT15] Bernardino Romera-Paredes and Philip Torr. An embarrassingly simple approach to zero-shot learning. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2152–2161, Lille, France, 07–09 Jul 2015. PMLR.
- [SC19] Mert Bulent Sariyildiz and Ramazan Gokberk Cinbis. Gradient matching generative networks for zero-shot learning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2163–2173, 2019.
- [SCD⁺19] 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, October 2019.

- [SES⁺19] Edgar Schonfeld, Sayna Ebrahimi, Samarth Sinha, Trevor Darrell, and Zeynep Akata. Generalized zero- and few-shot learning via aligned variational autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [SGMN13] Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-shot learning through cross-modal transfer. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.
- [SKP15] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.
- [SLC⁺22] Hongzu Su, Jingjing Li, Zhi Chen, Lei Zhu, and Ke Lu. Distinguishing unseen from seen for generalized zero-shot learning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7875–7884, 2022.
- [UBH⁺22] Shagun Uppal, Sarthak Bhagat, Devamanyu Hazarika, Navonil Majumder, Soujanya Poria, Roger Zimmermann, and Amir Zadeh. Multimodal research in vision and language: A review of current and emerging trends. *Information Fusion*, 77:149–171, 2022.
- [VAMR18] Vinay Kumar Verma, Gundeep Arora, Ashish Mishra, and Piyush Rai. Generalized zero-shot learning via synthesized examples. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4281–4289, 2018.
- [vdMH08] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.
- [VSP⁺17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.
- [WPL⁺17] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, July 2017.
- [WWAS22] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. pages 3876–3887, 2022. Publisher Copyright: © 2022 Association for Computational Linguistics.; 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022 ; Conference date: 07-12-2022 Through 11-12-2022.
- [WZZ⁺20] Jiamin Wu, Tianzhu Zhang, Zheng-Jun Zha, Jiebo Luo, Yongdong Zhang, and Feng Wu. Self-supervised domain-aware generative network for generalized zero-shot learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12764–12773, 2020.

- [XLSA18] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. Feature generating networks for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
- [XSSA19] Yongqin Xian, Saurabh Sharma, Bernt Schiele, and Zeynep Akata. F-vaegan-d2: A feature generating framework for any-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [YSAF90] Kouichi Yamaguchi, Kenji Sakamoto, Toshio Akabane, and Yoshiji Fujimoto. A neural network for speaker-independent isolated word recognition. In *Proc. First International Conference on Spoken Language Processing (ICSLP 1990)*, pages 1077–1080, 1990.
- [ZCDLP18] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.
- [ZWS19] Pengkai Zhu, Hanxiao Wang, and Venkatesh Saligrama. Generalized zero-shot recognition based on visually semantic embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Appendix A

Training hyperparameters

Hyperparameter	Value
Batch Size	16
BioBERT [LYK ⁺ 19] Latent Dimension	768
Image Size	256x256
Learning Rate	$1e^{-3}$
MedCLIP [WWAS22] Latent Dimension	512
MedCLIP [WWAS22] Variant	ResNet [HZRS16]
Number of Epochs	50
Number of Epochs with Loss Plateau before Reduction	2
Number of Seen Classes	5
Optimizer	Adam
Random Seed	42
t-SNE [vdMH08] Perplexity	100

Table A.1: Training hyperparameters.