FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Learning to write medical reports from EEG data

Ana Maria Amaro de Sousa

Mestrado em Bioengenharia

Supervisor: Prof. Dr Michel van Putten

Supervisor: Prof. Dr Luís Filipe Teixeira

Co-Supervisor: Eng.ª Catarina da Silva Lourenço

July 29, 2022

Learning to write medical reports from EEG data

Ana Maria Amaro de Sousa

Mestrado em Bioengenharia

July 29, 2022

Abstract

Electroencephalography (EEG) is an essential tool for the diagnosis and management of epilepsy, one of the most prevalent neurological disorders in the world, characterized by an increased likelihood of seizures. However these periods of abnormal brain activity (ictal EEG) may not occur often, so the diagnosis is usually made by visual analysis of the EEG in interictal periods. During this analysis the neurologist summarizes the findings and diagnosis in a clinical report. Although this is a current practice, it entails several disadvantages that motivate the development of automatic auxiliary algorithms that can streamline clinical workflow, reduce subjectivity and potentially improve diagnosis.

To address these shortcomings in EEG analysis and reporting, we apply deep learning (DL) methods adapting state-of-the-art image and video captioning approaches to generate automated preliminary clinical reports directly from the EEG signal. For that we develop several captioning architectures following the encoder-decoder approach. In all of them, a recurrent model is used as decoder and the encoder is a previously trained convolutional model (VGG16) for classification with 2s epochs of 869 EEG sessions with and without epileptic events from 340 patients. Implemented captioning models differ in the aggregation EEG embeddings methods used to extract and summarize the entire recording. This includes EEG average embedding-based, recurrent-based, attention-based, and multi-stream-based methods. The performance of the models was evaluated qualitatively by analyzing the reports and quantitatively by calculating common natural language processing metrics.

On classification of epileptic/no-epileptic EEG epochs, VGG16 yielded an AUC of 0.80, sensitivity of 82%, and specificity of 77% in the test set. On the other hand, with regard to report generation (EEG captioning), the average embedding-based and multi-stream-based models led to the best results, being able to detect and describe most EEG phenotypes through clinical reports using natural language. These models reached values in the range 50.5-56.3 in BLEU1, 20.7-23.3 METEOR, 42.5-46.0 ROUGE_L, 16.9-22.3 CIDEr and 17.1-19.3 SPICE.

We have shown that it is possible to generate reports from EEG data, but there is still space for improvement. We innovate by designing architectures for EEG captioning based on other architectures that have proven effective in other fields. There is great potential in the use of DL-based methods, but obtaining a model capable of diversifying language, dealing with the diversity clinical conditions ,and generating a report that effectively describing all EEG events are some challenges that must be tackled before implementing in the clinic.

Keywords: Epilepsy, EEG, Deep learning, Natural language processing, Signal captioning

Resumo

A eletroencefalografia (EEG) é uma ferramenta essencial para o diagnóstico e acompanhamento da epilepsia, um dos distúrbios neurológicos mais prevalentes no mundo, caracterizado pelo aumento da probabilidade de convulsões. No entanto, esses períodos de atividade cerebral anormal (EEG ictal) podem não ocorrem com frequência, portanto, o diagnóstico geralmente é feito pela análise visual do EEG nos períodos interictais. Durante esta análise, o neurologista resume os achados e o diagnóstico num relatório clínico. Embora esta seja uma prática atual, ela traz várias desvantagens que motivam o desenvolvimento de algoritmos auxiliares automáticos que podem agilizar o fluxo de trabalho clínico, reduzir a subjetividade e potencialmente melhorar o diagnóstico.

Para resolver essas deficiências a análisar e reportar EEGs, aplicámos métodos de *deep learning* (DL) adaptando abordagens de *captioning* de imagem e vídeo de última geração para gerar relatórios clínicos preliminares automáticos diretamente do sinal de EEG. Para isso desenvolvemos várias arquiteturas de *captioning* seguindo a abordagem codificador-descodificador. Em todos eles, um modelo recorrente é utilizado como descodificador e o codificador é um modelo convolucional previamente treinado (VGG16) para classificação com épocas de 2s em 869 sessões de EEG com e sem eventos epilépticos de 340 pacientes. Os modelos de *captioning* implementados diferem nos métodos de agregação de *embeddings do EEG* usados para extrair e resumir toda a gravação. Isso inclui métodos baseados na média dos *embeddings do EEG*, baseados em recorrência, baseados em atenção e baseados em múltiplos fluxos. O desempenho dos modelos foi avaliado qualitativamente por meio da análise dos relatórios e quantitativamente por meio do cálculo de métricas de processamento de linguagem natural comums.

Na classificação de épocas EEG epilépticas/não epilépticas, o modelo VGG16 produziu uma AUC de 0,80, sensibilidade de 82% e especificidade de 77% no conjunto de teste. Por outro lado, no que diz respeito à geração de relatórios (*captioning* de EEG), os modelos baseados em média dos *embeddings* e os baseados em múltiplos fluxos alcaçaram aos melhores resultados, sendo capazes de detectar e descrever a maioria dos fenótipos de EEG por meio de relatórios clínicos usando linguagem natural. Esses modelos atingiram valores na gama de 50,5-56,3 em BLEU1, 20,7-23,3 METEOR, 42,5-46,0 ROUGE_L, 16,9-22,3 CIDEr e 17,1-19,3 SPICE.

Mostramos que é possível gerar relatórios a partir de dados de EEG, mas ainda há espaço para melhorias. Inovámos desenhando arquiteturas para *captioning* de EEG com base em outras que provaram ser eficazes noutros campos. Há um grande potencial no uso de métodos baseados em DL, mas obter um modelo capaz de diversificar a linguagem, lidar com a diversidade de condições clínicas e gerar um relatório que efetivamente descreva todos os eventos de EEG são alguns desafios que devem ser enfrentados antes da implementação na prática clínica.

Keywords: Epilepsia, EEG, *Deep learning*, Processamento de linguagem natural, *Captioning* de sinal

Acknowledgements

Demorou, mas valeu tanto a pena...

Gostava de agradecer a todos que permitiram que assim fosse. Começando por toda a minha família, em especial aos meus pais, por serem as peças principais do puzzle, sem as quais nada seria possível. Milhões de obrigados não é suficiente para agradecer todo o carinho, apoio, por me darem condições e por, por vezes, acreditarem, mais em mim que eu própria.

Uma palavra de agradecimento aos meus orientadores, Prof. Dr. Michel van Putten pela oportunidade e por sempre disponibilizar recursos e condições necessárias, ao Prof. Luís Filipe Teixeira por me ter acompanhado ao longo do desenvolvimento da dissertação e em especial à Catarina Lourenço por ter sido muito mais que uma enorme orientadora. A estes agradeço pelas ideias, feedback, esforço e por terem criado um ambiente de trabalho tão honesto/humano e livre de julgamento, foi um prazer trabalhar convosco.

Esta dissertação é também o culminar de uma jornada. Uma verdadeira aventura repleta de desafios que me fizeram reinventar como estudante e pessoa. Muitas foram as pessoas que tive a sorte de conhecer ao longo destes 5 anos. Àquelas que apesar de *Perdidas* sei que posso sempre encontrar, às Darlings sempre *Plenas*, aos que da grolsch fizeram crescer raízes tão sólidas e genuínas (*Rutbeer*), aos viciados do Marketplace (*Fixhuis*), aos que com mais ou menos "Sauce" tornaram estes anos especiais, aos *Destruídos* e *Sobreviventes* manifesto a minha mais sentida palavra de apreço.

Agradeço também a todos que contribuíram para a minha formação académica e a toda a equipa do CNPH por me terem feito sentir bem vinda, apoiado desde o primeiro dia.

Às *Andorinha*s, *Faneca*s e todos os seres capazes de encurtar distância, prolongar se no tempo e elevar o significado de "amizade" para um outro nível. A todos estes (mesmo aquelas para as quais não há metáforas, Joana e Catarina) que se disfarçam de pseudo-primas, pseudo-irmãs e pseudo-madrinhas, obrigada pela empatia, paciência e por me acompanharem nesta aventura.

Por fim tenho de mencionar aquela que tem verdadeiros superpoderes, de me entender com um olhar, de me fazer sorrir sem sequer ter de falar e sentir sem o vivenciar. Obrigada Mana!

A Ti! e a todos aqueles que (sabendo ou não) contribuíram para que recorde estes anos com este sorriso no rosto, um sincero Obrigado!

Ana Maria Sousa

"The science of today is the technology of tomorrow."

Edward Teller

Contents

List of Figures

List of Tables

Abbreviations

ANN	Artificial neural network
AUC	Area Under the Curve
BCKG	Abnormal Background (Rhythmic activity)
BERT	Bidirectional Encoder Representations from Transformers
BioBERT	Bidirectional Encoder Representations from Transformers for Biomedical Text Mining
BLEU	Bilingual Evaluation Understudy
CBOW	Continuous Bag-of-Words
ClinicalBERT	Bidirectional Encoder Representations from Transformers for Clinical Text
CIDEr	Consensus-based Image Description Evaluation
CLARA	Clinical Report Auto-completion
CNN	Convolutional Neural Networks
DL	Deep Learning
DRLN	Deep Rectified Linear Network
ECG	Electrocardiogram
EEG	Electroencephalography
EHR	Electronic Health Record
ELMo	Embeddings from Language Models
EMG	Electromyogram
ESA	Explicit Semantic Analysis
GANs	Generative adversarial network
GloVe	Global Vectors for Word Representation
GPT	Generative Pre-trained Transformer
GRU	Gated Recurrent Unit
HIC	High Income Countries
HIPAA	Health Insurance Portability and Accountability Act
IEDs	Interictal Epileptiform Discharges
LMIC	Low/Middle Income Countries
LSA	Latent Semantic Analysis
LSTM	Long Short-Term Memory
METEOR	Metric for Evaluation of Translation with Explicit Ordering
MGH	Massachusetts General Hospital
ML	Machine Learning
MLC	Multi-label Classification
MLM	Masked Language Modeling
NLG	Natural language Generation
NLP	Natural Language Processing
NLU	Natural Language Understanding

PWE	People With Epilepsy
RNN	Recurrent Neural Networks
ROC	Receiver Operating characteristic Curve
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
SPICE	Semantic Propositional Image Caption Evaluation
SPSW	Spikes and/or Sharp Waves
TF-IDF	Term Frequency-Inverse Document Frequency
TUH	Temple University Hospital
ULMFIT	Universal Language Model FIne-Tuning
VDCNN	Very Deep Convolutional Neural Network
VGG	Visual Geometry Group
XAI	Explainable Artificial Intelligence

Chapter 1

Introduction

1.1 Motivation and Context

Epilepsy is the fourth most prevalent neurological disorder in the world [63]. It is a chronic brain disease characterized by an increased and permanent likelihood of unprovoked seizures [9, 37]. Epilepsy affects the quality of life of patients, for example, increasing the occurrence of physical injuries and psychological problems. This and the fact that more than 70% of patients could live normally (without seizures) if properly diagnosed and treated, it becomes evident that there is an urgent need to improve the diagnosis of epilepsy [27, 86].

Electroencephalography is a fundamental tool for diagnosing epilepsy and identifying the type of epilepsy. In this way, it aids in the management of epilepsy, planning the treatment and choosing an anti-epileptic medication if necessary [36]. Biological signals are typically analyzed by a clinician, who summarizes the information in a clinical report. In particular, EEGs provide crucial information about cortical brain activity that is crucial to characterize the health status of a patient.

However, the visual analysis of the EEG is time consuming and has disadvantages such as intra and interobserver variability and a long clinician training curve. Therefore, it would be helpful for clinicians to have access to a preliminary automated report based on the complete EEG record of the patient they are diagnosing and/or treating. In addition, this automation can lead to more objective and consistent EEG interpretations and potentially reduce 'inadvertent supervision' of medical conditions.

Report generation or, more generally, captioning, has been widely successful in images and video. More recently, the synergy between computer vision and natural language processing has started to extend to the medical field with captioning of medical images and representation learning of clinical texts [50]. However, these methods have not been successful when applied to time series such as EEG signals. This is mostly due to the variable dimensions (duration, channels) of the data, which are not typically supported, and the inherent structure of the clinical reports [14, 135].

1.2 Objectives

Ideally, it is possible to describe an EEG signal accurately and concisely in a report that correctly uses clinical terms and is easily understood by a neurologist. This would allow the implementation in the clinic of a useful tool to support EEG analysis and diagnosis of neurological diseases, particularly epilepsy. Thus, this work represents a contribution toward this ideal objective.

This dissertation aims to develop a method for generating clinical reports based on EEG signals, with a focus on the reliability of the generated description. Our contribution includes the review and adaptation of different state-of-the-art image, video and signal captioning approaches to develop a model that can clinically describe the EEG signal, in particular for the diagnosis of epilepsy.

1.3 Document Structure

This document is divided into 7 chapters. A brief introduction of topic, researching motivation, objectives and contributions are made in this initial Chapter.

Chapter 2 focuses on clinical background, starting by covering EEG signals and signal acquisition and analysis (section 2.1) and continuing with an overview of epilepsy (section 2.2), including the epidemiology, causes and manifestations of the disease, as well as diagnosis, through EEG analysis.

Chapter 3 clarifies some technical terms and techniques and presents current consensus and trends present in literature. Section 3.1 provides an overview of Natural Language Processing (NLP) and its applications as well as a brief historical summary. Furthermore it covers the different aspects and techniques for preparing and representing text with a particular focus on deep learning applications. In addition different decoding methods and NLP metrics are also assessed. On the other hand, section 3.2 describes the state of the art of NLP for clinical texts , image and signal captioning methods, with a greater focus on the medical field.

The methods that were implemented and used in this dissertation are described in Chapter 4 and their results are presented in Chapter 5.

Chapter 6 concerns the discussion of the results and in Chapter 7 presents the conclusions drawn from the dissertation and future work.

Chapter 2

Background

2.1 Electroencephalography

The electroencephalogram was performed for the first time in humans by the psychiatrist Hans Berger in 1929 [106]. It consists of the record of cortical electrical activity of the human brain. Given to its excellent temporal sensitivity, the EEG is very useful to evaluate the dynamic cerebral functioning and therefore, it is commonly used in the medical and research fields [59, 99].

When neurons are activated the differences of electrical potentials on the cells create electrical dipoles between the body of the neurons and the apical dendrites [99]. The synchronism of the dipoles generated originate local current flows on the extracellular medium that can be detected by the electrodes [110]. The EEG is the summation of excitatory and inhibitory postsynaptic potentials generated by groups pyramidal cells oriented perpendicularly to the brain's surface detected by through placement of electrodes on the scalp or surface of the brain. The recorded voltages are plotted against time, displaying a near real-time representation of ongoing cerebral activity [59, 99].

However, the EEG has several limitations as the captured signal corresponds to the potential generated only in the most superficial layers of the cortex. In the slightly more distant layers (in order of few square centimetres), a large population of neurons (10^8) must be activated so that the potential generated is sufficient to cause changes that can be registered by the electrodes [99, 106]. The spatial resolution is low, as the electrode placement pattern in the scalp does not cover the entire cortex, furthermore, the propagation of electrical activity can lead to a misleading impression of the location of the activity source. Thus, source location and size has a considerable influence on whether or not activity is detectable in the EEG [58, 106].

2.1.1 EEG Signal

EEG records both rhythmic and transient physiological and pathological activities. Unlike rhythmic activity, transient activities are unique and relatively rare events [110]. Those are characterized by high amplitude, normally associated with pathologies. However, they could be related to a normal physiological process. Rhythmic activity (background) are sinusoidal patterns with

characteristic frequencies that are usually related to specific regions of the brain. These can be classified as alpha (8-13 Hz), beta (13-30 Hz), theta (4-8 Hz) and delta (up to 4 Hz) waves. Interestingly, it has been proven that there is a link between these neuronal oscillations and different cognitive states, for example, in relaxation drowsiness alpha activity rises, however neuronal activity shifts to lower frequency bands if in sleep [59, 110]. Illustration of these rhythmic waves can be seen in the Figure 2.1.

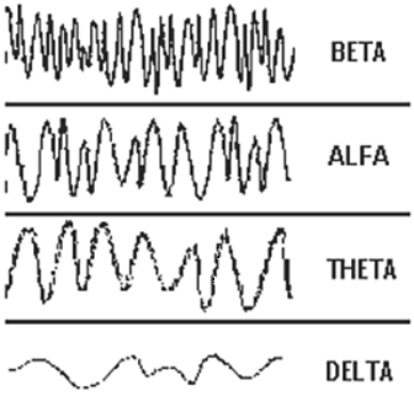


Figure 2.1: Brain waves samples belonging to beta, alpha, theta, and delta frequency band. Extracted from [110].

The large amount of information recorded in the EEG increases the difficulty of its interpretation. Furthermore, the low power and broad-frequency nature of neuronal activity allied with the existence of a large number of physiological structures and processes that occur in the brain make signal corruption by interference inevitable [59]. For instance, the normal EEG signal range is between 0.5-150 microvolts (μV), which is about 100 times lower than ECG signals [82, 110]. Muscle movement (EMG, Electromyogram), such as eye movement, or even cardiac activity (ECG, Electrocardiogram) can also be recorded on the EEG. In addition to patient noise sources, the AC power line, the contact between electrodes and skin, and impedance fluctuation are some technical artefact sources [110].

2.1.2 Signal Acquisition

The electrodes used in recording the EEG and their proper functioning are essential for the acquisition of a high-quality signal [59]. For this, can be used several types of electrodes that offer different characteristics. Needle electrodes are the most common in case of need to record invasive EEGs. However, in routine examinations, signal acquisition is performed non-invasively, usually with superficial Ag-AgCl electrodes [110].

The placement of electrode on scalp is usually done using electrode caps rather than individual electrode placement, and sometimes scalp abrasion is required to create a conductive interface.

The montage generally follows the standardized 10-20 system established in 1958 by the International Federation of Electroencephalography and Clinical Neurophysiology [58, 59]. In this system, 21 electrodes are placed and the head is divided into proportional distances, as shown in Figure 2.2. The electrodes are labeled with a letter according to the area of the skull in which they are placed: F (frontal), C (central), T (temporal), P (posterior), or O (occipital), and numbered with even numbers if on right side and odd if on the left side of the skull [58, 110]. As

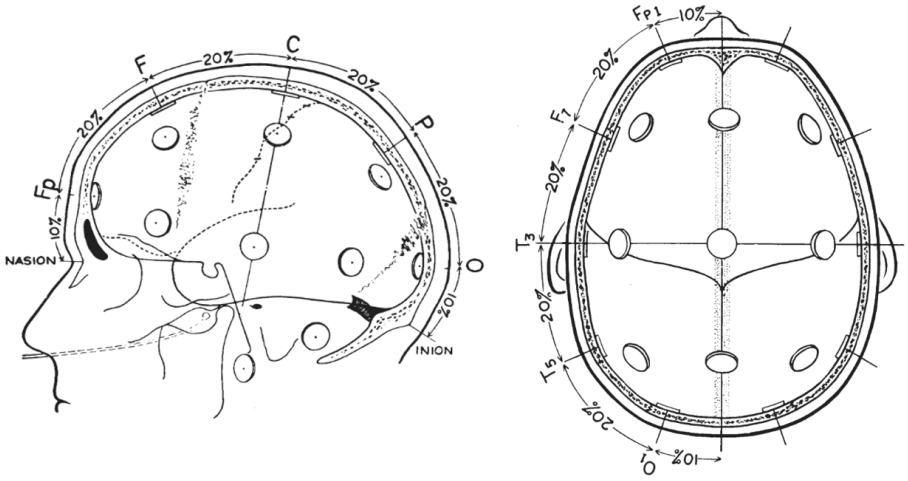


Figure 2.2: Lateral and Superior view of the standardized 10 ± 20 system for electrode placement. Adapted from [58].

the voltage ranges that characterize the EEG are of the order of microvolts (μV), it is necessary amplify them in order to become measurable. Therefore, the signal is amplified between 10^2 and 10^5 times [110].

To obtain the best possible signal-to-noise ratio, a filtering phase is follows, mitigating the impact of artefacts. A high pass filter is applied to remove low-frequency interference related to muscle movement and a notch filter to eliminate power line frequencies (50Hz-60Hz). Also, to avoid the aliasing phenomenon, higher frequencies are filtered and finally, the signal is digitized by an analog-to-digital (A/D) converter [59, 89, 110] .

2.1.3 EEG analysis

EEG signals consist of plotting recording voltages over time. [110]. Since the recordings are digital, experts can use tools or change settings to better analyze the signal [59, 99].

EEG is recognized as a safe and low-cost technique and, despite being extensively time-consuming, visual analysis remains the clinical gold standard for EEG interpretation. The wide variability of pattern and high subjectivity are the main limitations [5]. These issues motivated the development of tools to automate clinical decisions, however, they have not yet achieved a better performance than traditional visual analysis.

An alternative is the objective measurement of EEG signal characteristics. This quantitative assessment has advantages, it allows to reduce variability and time spent on the analysis [5, 78]. However, the definition and selection of relevant features are not trivial, thus the approaches based on handcraft features have been failing to cover all existing possibilities. Mainly because it does not require the prior definition of features, and can extract non-linear and complex relationships, deep learning emerges as a new and promising approach in the area [24, 79].

2.2 Epilepsy

Epilepsy is one of the most prevalent neurological disorders, affecting more than 70 million people worldwide. In Europe, an estimated 6 million people have active epilepsy and it costs more than €20 billion a year [27]. It affects people of any age, gender, country, ethnicity, and social background. However, it is slightly more prevalent in men than in women and there is a higher risk in children and older age groups. Furthermore, the prevalence and incidence are higher in low/middle income countries (LMIC) compared to high income countries (HIC). The incidence is approximately 140 per 100,000 people per year on LMIC and 50 per 100,000 people per year on HIC [9]. This difference is often attributed to poor hygiene and inadequate health systems, leading to an increased likelihood of infections [111].

The clinical definition of epilepsy has undergone changes and updates over the years. It is currently established, that people with epilepsy (PWE) have at least one of the following conditions: at least two unprovoked seizures separated by more than 24 hours; suffer an unprovoked seizure and have a higher probability (at least 60%) of having similar seizures over 10 years (equivalent to the likelihood after two unprovoked seizures), diagnosis of epilepsy syndrome [8, 40].

This disease is a symptom complex with multiple risk factors and a strong genetic predisposition. It is categorized into several syndromes that manifest in different ways and that are associated with several causes, symptoms, and possible treatments [36, 111]. In about half of the cases of diagnosed epilepsy, it is not possible to discriminate the causes. The cases in which this discrimination is possible causes range from genetic predisposition to the existence of brain lesions as a result of brain tumours, infections in the central nervous system, degenerative diseases or cerebral disabilities [86].

Epilepsy has a negative impact on the quality of life of the patients. It has cognitive consequences, PWE tends to suffer more physical injuries (20.6%), such as fractures and bruises, related to seizures and have higher rates of psychological problems (24%), including anxiety and depression (23.1%). In addition, the mortality associated with the disease is increased (2–3 times higher mortality rate), whether due to suicide (7,4%), sudden unexplained death in epilepsy or effects of seizures as some of the causes [27, 86].

2.2.1 Epileptic Seizures

A seizure is defined as a transient phenomenon resulting from atypical electrical brain activity, consisting of excessive or synchronous neuronal activity [41]. Seizures are generally unpredictable

and their occurrence causes neuronal electrical instability, making them more susceptible to the occurrence of future crises, becoming a recurrent phenomenon [10].

The abnormal brain state characterized by the presence of epileptic seizures is termed ictal. On the other hand, the interictal state is characterized by apparently normal brain activity, but which may present transient and repetitive patterns known as interictal epileptiform discharges (IEDs) and which suggest a greater epileptic predisposition in the patient. Epileptiform patterns essentially include spikes of high amplitude (up to 1000 μV) and short duration, often followed by slow waves [28, 41, 93].

The classification of seizures is crucial to evaluating the patient, diagnosing the epileptic syndrome, and guiding treatment. Currently, seizures can be classified as focal (localized, affects only a portion of the brain), or generalized (affects both hemispheres of the brain) and unknown onset. Regarding the types of epilepsy, Focal Epilepsy, Generalized Epilepsy, Unknown or even Combined Generalized and Focal Epilepsy [41, 100].

2.2.2 Diagnosis of Epilepsy

The diversity of the syndrome and symptoms of epilepsy and their similarity to the symptoms of other diseases make their distinction complicated [26, 130]. Also, according to the clinical definition of epilepsy, the first seizure is not synonymous of epilepsy as it may be the result of other conditions . However this sometimes leads to some confusion and misdiagnosis. It is estimated that 20% to 30% of patients diagnosed with epilepsy actually suffer from another disorder [130]. This incorrect diagnosis ends up in inadequate treatment for the patient's condition and possible worsening of it.

Considering the huge impact of epilepsy on quality of life and the fact that more than 70% of patients can live without seizures if properly diagnosed and treated, the importance of developing and improving the diagnosis, management and treatment of epilepsy becomes evident [27, 86].

2.2.2.1 Role of EEG in Epilepsy

Diagnosis of epilepsy is complex and is supported by considering a large set of patient data, including EEG and the patient's clinical history. The EEG remains a crucial tool since it handles a significant amount of useful information to characterize the patient clinical condition [83, 99].

Since epileptic seizures may not occur often, it is not common to have EEG recordings of this period (ictal EEG). Therefore, interictal EEG, acquired between seizures, is taken into consideration for diagnosis [94, 106]. The EEG is essential for the determination of the type of seizure and syndrome and the presence of transient patterns (IEDs) is a key aspect in the diagnosis of epilepsy [83]. However, although rare, these patterns can occur on the EEG of patients with other neurological disorders or even without any disorders. On the other hand, the absence of IEDs does not rule out the possibility of epilepsy, since about 10% of PWE do not have IEDs in the interictal EEG. This highlights the importance of considering all patient clinical data for diagnosis [94, 106].

To facilitate the diagnosis in the routine EEG exam, conditions are created that favor the appearance / detection of possible EEG abnormalities, such as sleep deprivation and application of activation techniques, as hyperventilation and intermittent photic stimulation [94, 99, 106].

The reduced duration of a routine EEG (20-30 min) can be limited, thus the patient can be monitored for a longer period (24h), which increases IEDs by about 20%. Furthermore, the timing of the EEG also influences the effectiveness of the diagnosis, and the appearance of IEDs is more likely if the examination is done within the first 24 hours after the seizure. Therefore, repeating the exam is recommended for greater confidence in the diagnosis [99, 106].

2.2.2.2 Analysis and Reporting

The expert analyzes the EEG, interprets it and writes a text report to summarize the findings. This analysis is subject to intra and inter-observer variability, requires experience, is time-consuming, error-prone and the report writing is laborious [5, 14].

Neurologists can use support systems for EEG visualization, which usually offer the possibility to change settings such as the time window, the EEG amplitude, the montage view and apply different filters. This is useful to better visualize hidden features and to facilitate the EEG analysis [59, 99]. Recently, tools have been developed that seek to reduce clinical errors and subjectivity in the analysis, providing additional information to the physician. Models for detection of spikes and seizures and automatic quantitative EEG analysis are some of theses examples [5, 29, 79].

Despite these advances, the writing of clinical reports still rests with the physician and analysis and reporting represents a considerable part of the neurologist's work [14].

Chapter 3

Literature Review

3.1 Natural Language Processing

Natural language processing is an interdisciplinary field of artificial intelligence that combines computational linguistics with statistical, machine learning (ML), and DL models for natural language analysis and manipulation. NLP encompasses a collection of techniques used to deal with understanding text and speech, not just as a sequence of characters or words, but as data with complex syntactic and phonological structures that carry a meaning that implies some interpretability [4, 17].

NLP can be distinguished into two sub-fields Natural language understanding (NLU) and Natural language Generation (NLG). NLU has a close relation with linguistic aspects which it uses to determine the meaning of a sentence. The word-formation (Morphology), sentence structure (Syntax), the intended meaning (Semantics) and Pragmatics are some of those linguistics aspects [88]. While NLU focuses on machine reading comprehension, NLG enables the machine to generate text given a data input. In addition to linguistic aspects, such as grammar, content selection and textual organization are crucial in the NLG [85].

3.1.1 Brief Historical Overview

The idea of natural language handled by computers is old. The first project addressing this subject was a computer translator,in the 1940s, by Weaver and Booth, during World War II. This inspired other projects that maintained an approach based on Weaver's suggestion to apply cryptography and information theory to language translation. These were based on a "dictionary look up" perspective, looking for the words in the target language and fitting them following some word-order rules. This approach turns out to be a very limited approach since the complexity, ambiguity and lexical aspect of language were completely ignored [64].

The poor results highlighted the need to understand more about linguistics. The boost in this field occurred in 1957 by the introduction of concepts of generative grammar by Chomsky in his publications [25]. After that, real advances have been achieved using linguistics to help in machine translation and in the other applications that then emerged as speech recognition [64].

In 1960s, the concern was how to develop meaningful computationally tractable solutions. This led to the development of new theories for the theoretical aspects and several prototype systems as Weizenbaum's ELIZA [115] which was a chatbot to simulate the conversation between a psychologist using a pattern matching and character substitution methodology and LUNAR, a question-answer system developed by Woods which was an interface system to the database about the Apollo 11 moon rocks using augmented transition network [64].

Working with natural language is challenging, as human language has vastly size and is highly ambiguous [17, 88]. The approaches until then relied purely on heuristics (handcrafted symbolic rules), which showed some limitations concerning the extraction meaning [4]. The paradigm shift in NLP came about by emerging statistical language processing. Statistical inference using machine learning models over text corpora revealed to be promising and with the advances in artificial intelligence, specifically in deep learning has allowed the NLP to grow rapidly and the models are being ever more powerful and robust [17, 85].

3.1.2 NLP Applications

The potential of NLP has been explored in the most diverse fields. Its ability to understand human language has proved to be a useful and powerful tool for managing and extracting information [17]. This has had a major impact at the business level, research/science, and in the development of systems and solutions to improve life, health care and to facilitate the tasks of clinicians. Thus providing better patient care and allowing for more continuous monitoring [4, 50].

Some of the typical problems address by NLP are Sentiment analysis, Document classification and Captioning. Sentiment analysis is particularly interesting for the market and businesses, since the ability to detect positive or negative feelings in the text reflects the customer's interest, allowing for an agreement management. On the other hand is the Document classification allows the categorization of emails and filtering of spam messages, for instance [12, 17]. Captioning is normally applied to images and consists of generating a textual description given to an image. It allows generating a clinical report given a medical image, for instance. This concept is now starting to be adjusted and tested to videos and signals as well [50, 52, 122].

3.1.3 Data Preparation

Language data cannot be used directly to the field in machine or deep learning models. In particular raw text must be first prepared and cleaned up by applying transformations. Initial text processing steps comprises operations such as removing punctuation and normalizing text by converting it to lowercase, for example [17, 68].

Preprocessing operations have a high impact on the performance of the natural language model, so they must be carefully chosen according to the task. This selection is not trivial, cleaning text is hard and full of trade-offs. The simpler the text data and the smaller the vocabulary, the easier it will be for the model to learn, however, with processing the text loses structure and

linguistic aspects such as syntax or semantics deteriorate. In some tasks such as document classification or sentiment analysis, in which it is not important to preserve all the structure of the text but just need to extract the content, this is not problematic and operations such as Stemming, Lemmatisation or Filtering out Stop Words can be applied [12, 17]. Stemming consists on converting each word to its root, removing variations such as the number, size or gender (e.g fishing, fished, fisher all would be reduced to the stem fish). Lemmatization is the process of grouping words and reducing them to the underlying concept, which may or may not be consistent with the stem (e.g "better" has "good" as its lemma), this is particularly useful for the model to be robust dealing with synonyms. Stop words are words that do not contribute to the deeper meaning of the phrase (e.g “a”, “as”, "is") [12, 17].

However, when considering tasks that involve text generation as translation, image captioning, and text summarization, it is essential that, in addition to the information, the structure, semantics and syntax of the text are preserved so that the model can learn and respect them during the execution of the final task [17].

After the text cleaning step, it is necessary to discretize it and encode it, associating an id to each part. This transformation is called Tokenization and consists of dividing the text into meaningful parts, called tokens [12, 38, 107]. What is actually a meaningful piece of text is, however an open problem, there are various approaches and types of tokenizer. One of the simplest and most common approaches is to consider each word as a token, however, this approach has some drawbacks as it end ups usually in a large dictionary because tokenizer is sensitive to little variations of the words and perceives for instance "dog" and "dogs" as two different tokens. The same happens with female/male, lowercase/uppercase letters, hence the need to clean and normalize the text previously. One way of control the dimension of the vocabulary is ignore rare words and encode those as outside the vocabulary (using a common token ,"**<unknown>**"), however this leads to an increased loss of information [39].

Characters tokenizer, solves some of the mentioned issues of the word tokenizer, forming a much smaller and flexible vocabulary that allows representing any word. However, this implies that each phrase is encoded by a large sequence of tokens, moreover, characters provide less information than words. Finally, the approach that underlies some of the models that achieve state-of-the-art results are subword tokenizers such as Wordpiece [44, 33], Unigram, Byte-Pair Encoding [101]. These tokenizers allow a type of composition of the less frequent words in meaningful and more frequent parts. This allows to preserve context-independent representations and obtain a reasonably sized vocabulary [39].

There are different packages and tools available to perform the aforementioned transformation in Python. One of the most used is the Natural Language Toolkit (NLTK) [95], but the scikit-learn, Tensorflow and Keras libraries also provide some NLP tools.

3.1.4 Text Vectorization

With the development of NLP and ML, several approaches emerged to encode text into numbers and vectors interpretable by algorithms. These feature extraction steps (or vectorization)

have been improved, obtaining increasingly compact and meaningful representations. Similar to the pre-processing step, the choice of representation technique is highly dependent on the NLP task at hand [17].

3.1.4.1 Bag-of-Words

Bag-of-Words is an algorithm that handles text describing the occurrence of words or n-grams by attributing a score, as represented in Figure 3.1. The tokens are interpreted as a group and information related to sentence structure and token sequence is lost. This property makes the application of the model limited. Furthermore, it tends to be used in traditional machine learning models rather than deep learning models. Some of the main applications are document classification and sentiment analysis [17, 24].

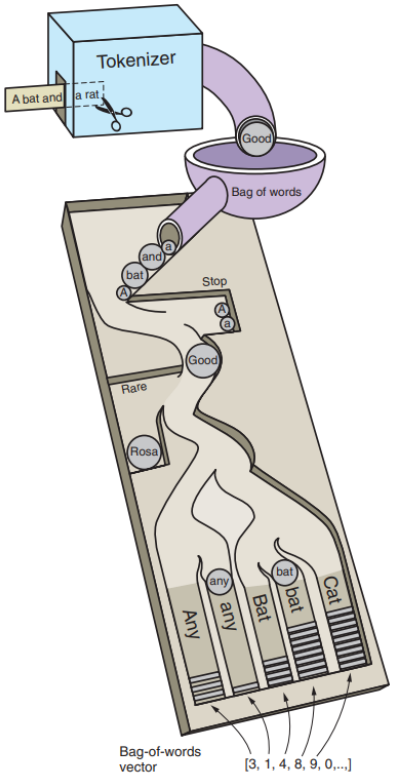


Figure 3.1: Illustration of text vectorization applying Bag-of-words.
In this, a word tokenizer is used and the words stop and rare are filtered out. The remains are considered by the algorithm. Extracted from [68].

There are many versions of the Bag-of-words model that differ in the base element that is considered (word or n-grams) and the method that is used in scoring. N-grams are normally associated with words but the same concept can be applied at the character level. N-grams are groups of consecutive words that can be extracted from a sentence, for instance in the previous sentence "A bat and a rat" the decomposition in 2-grams would be:

{"A", "A bat", "bat", "bat and", "and", "and a", "a", "a rat", " rat"}

The referred model that uses 2-grams is called Bag-of-2-grams same way a model of this family that is based on 3-grams is called Bag-of-3-grams. Those versions of the model that use n-gram instead of individual words are more sophisticated and can capture a little more meaning from the document [24].

Regarding the scoring attributed, it can be binary scoring (assigning 1 to present words and 0 to the absence words), a score reflecting the number of times each word appears in a document (Word hashing) or reflecting the frequency that each word (Term Frequency-Inverse Document Frequency (TF-IDF)). Word hashing consists of assigning a hash to each word and counting its occurrence in the document. The main challenge is to define the hash space that is sufficiently large to accommodate the entire vocabulary and simultaneously not too sparse and with a low probability of hash collision. On the other hand, the TF-IDF score reflects the frequency of words that occur considering their frequency in all documents. This re-scaling penalizes the most common words like "the", highlighting the distinct words, that contain useful information [17, 68].

Despite being simple to understand and implement, Bag-of-words has some limitations, such as the sparsity of representations that makes it harder to model, needing a careful vocabulary design and the fact that it ignores the order of the words hinders understanding [17].

3.1.4.2 One Hot Encoding

One Hot Encoding is a simple vectorization method that represents every token of the given text in a vector of 1 (one) and 0 (zeros). This allows the easy identification of each token by a unique vector. However this implies that similar words on spelling or meaning are represented as different vectors, thus there is a substantial loss of information.

In addition, these vectors have the vocabulary dimension. The phrases being the sequence of tokens end up being three-dimensional. Similarly to Bag-of-words, while using one-hot encoding strategy, it is crucial to previously perform careful text cleaning to limit the dimensionality and sparseness of the feature space [24, 68].

3.1.4.3 Word Embeddings

Word embeddings are text representation techniques that allow mapping each word to a fixed length vector of real numbers. This dense distributed representation preserves different degrees of similarity of the words, as semantics and meaning. Despite the high dimensionality of theses vectors, often tens or hundreds of dimensions, it is much smaller than vocabulary size or the thousands or millions of dimensions needed for sparse word representations, such as one-hot encoding [24].

Classical methods for word embedding like Latent Semantic Analysis (LSA) or Explicit Semantic Analysis (ESA), rely on statistical analysis and co-occurrence matrices [68]. However, there are many other approaches to represent words that instead consider embeddings matrices, with an embedding for each word, such as Word2Vec [32], GloVe (Global Vectors for Word Representation) [91], fastText [15]. More recently, deep belief networks or graphical generative models

such as ELMo (Embeddings from Language Models) [92], GPT (Generative Pre-trained Transformer) or BERT (Bidirectional Encoder Representations from Transformers) [33] have driven dense word representation.

Word2Vec is one of the most popular and efficient models and is has two possible model architectures: Continuous Bag-of-Words model (CBOW) and Continuous Skip-Gram model. In the training process represented in Figure 3.2, both models learn to map words according to the surrounding words. CBOW learns to predict the current word based on its neighbours while the skip-gram model learns predicting the surrounding window of words based on the current word [32].

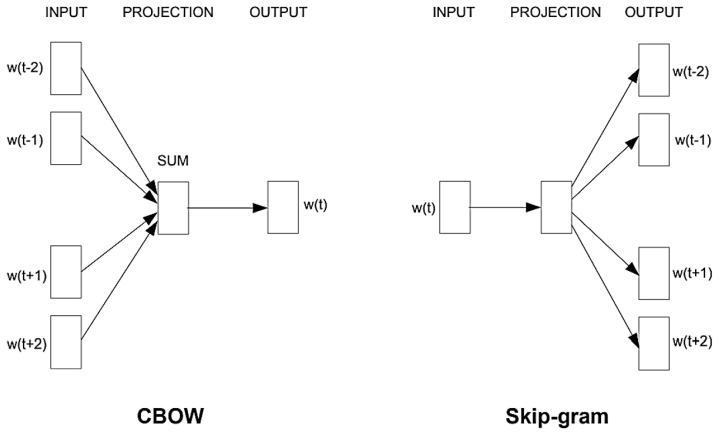


Figure 3.2: Representation of Word2Vec model architectures.
Extracted from [32].

Word2Vec can learn meaningful relationships by extracting the notions of relatedness across words, semantics, synonyms and analogies. This culminates in a meaningful representation of words in which semantic relationships are preserved and encoded as geometric transformations in the features space [17, 24]. For instance, Figure 3.3 illustrates a plausible feature space, where the distance between the word pairs cat-tiger and dog-wolf are similar and share the same transformations.

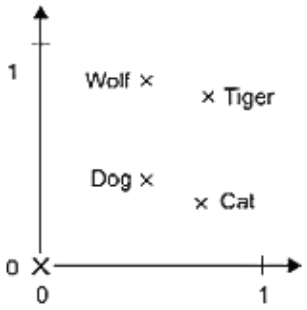


Figure 3.3: Word embedding space illustration.
Extracted from [24].

GloVe is an extension of the Word2Vec by combining with global statistics techniques as LSA [91]. Countering one of the biggest limitations of Word2Vec models, which is based on the local context of sentences [53]. fastText is also an extension of Word2Vec, but instead of creating a vector representation given a word, it decomposes words into n-grams allowing embedding to understand prefixes/suffixes. Unlike previous methods, fastText consides word morphology and can provide vector representations for words that were not seen in training [15].

Despite the ability of Word2Vec, GloVe, and fastText to capture syntactic and semantic information, these models are context-free in nature, as they generate a per-word vector representation that limits their ability to represent context-dependent information [68].

ELMo innovated by being the first context-sensitive model. It uses a bidirectional LSTM (Long Short-Term Memory) and considers the entire sentence before assigning an embedding to each word. ELMo is trained in the word prediction task, known as language prediction, by converting each token to character embeddings and combining them with the intermediate layer representations of the bi-LSTM [92].

ELMo by Peters et al. [92] and ULMFIT (Universal Language Model FIne-Tuning) by Howard et al. [49] were the first pre-trained language models to significantly improve the state-of-the-art of natural language understanding tasks. The success of this model motivated the development of new models that were also focused on the context. So quickly OpenAI and Google published pre-trained transformer-based language models called GPT [96] and BERT [33]. respectively.

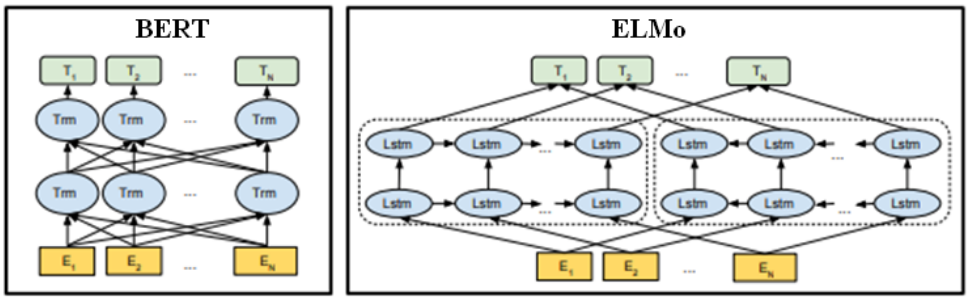


Figure 3.4: Differences of BERT and ELMo architectures.
Adapted from [33].

BERT architecture consists of a stack of transformers (12 layers in base version or 24 layers in their larger version). BERT is essentially an encoder model, which maps an input sequence to a contextualized encoded sequence, but it can be applied to a wide variety of tasks by adding an new layer, typically a feed-forward layer. Its robustness is also due to the training process that includes 2 tasks: Masked Language Modeling (MLM) and sentence prediction. This prediction task corresponds to predicting if given two sentences, one is likely to occur after the other and MLM consists of predicting masked words in sentences [33].

Transformer-based outperformed the LSTM-based language model, constituting the state of the art of NLP tasks. Figure 3.4 compares the architecture of both BERT and ELMo.

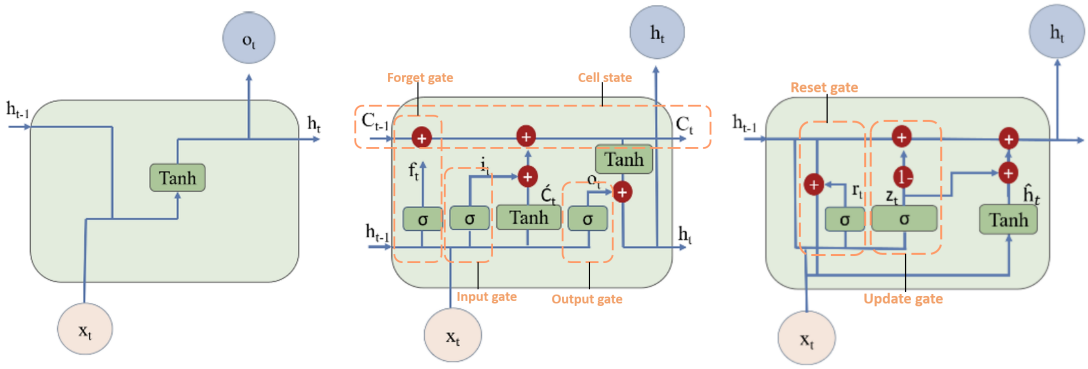
Since the release of these models, they have been used and adapted for different applications and new transformers and improved versions have been developed.

3.1.5 Natural Language Generation Models

Text representation is usually the start of NLG, it can be done by stacking an embedding layer that trains from scratch along with the model, or using a pre-trained model (Word2Vec, GloVe). Pre-trained models can be fine-tuned (i.e. BERT, ELMo) or used to extract the contextualized word embeddings to feed into the new model. Since word embeddings strongly impact the performance of the deep generative models used, those approaches guarantee an efficient algorithm for the representation of words [53].

In NLG the models needs to be capable to process text as sequence of tokens. Currently Neural-network-based language models have outperformed classical methods in tasks that require deeper language understanding. The improvement may be associated with more ability of Neural-network-based to generalize. These models can simultaneously learn a representation of the words and predict the probability of the next word given the previous ones [17, 68].

Recurrent Neural Networks (RNNs) are the typical Artificial Neural Networks (ANNs) used to deal with sequential data. However their sequential memory is limited, as the length of the sequences increases, the distance between the relevant information increases and the RNNs become unable to connect the information, suffering from vanishing gradient [47].



(a) RNN (b) LSTM (c) GRU

Figure 3.5: Standard architecture of RNN, LSTM and GRU.
Adapted from [97].

Over time, new RNNs have been proposed, with an architecture and underlying mechanism that allow mitigating this problem. Thus Hochreiter et al. [46] introduced the LSTM in 1995, and more recently, in 2014, Cho et al. [22] proposed the Gated Recurrent Unit (GRU). The architecture representation of each of the variant RNNs can be seen in Figure 3.5.

LSTMs have an internal mechanism, consisting of three gates, which decide which information should be stored or removed on each state and information output, allowing storing of information for longer periods. GRU is based on similar mechanisms but has only 2 gates (an update gate and a reset gate), for storing and forgetting information. This simplification of architectures

allows GRUs to use less memory and be faster than LSTM, whereas LSTM is more accurate on a larger dataset.

In theory, LSTMs and GRUs solve the issue of vanishing gradients. however, they still have a limitation to capture long-range dependencies. Transformers emerge as a promising solution to this problem by using self-attention and multi-headed attention mechanisms [113]. Its architecture is based on the stack of encoder and decoder layers as represented in Figure 3.6

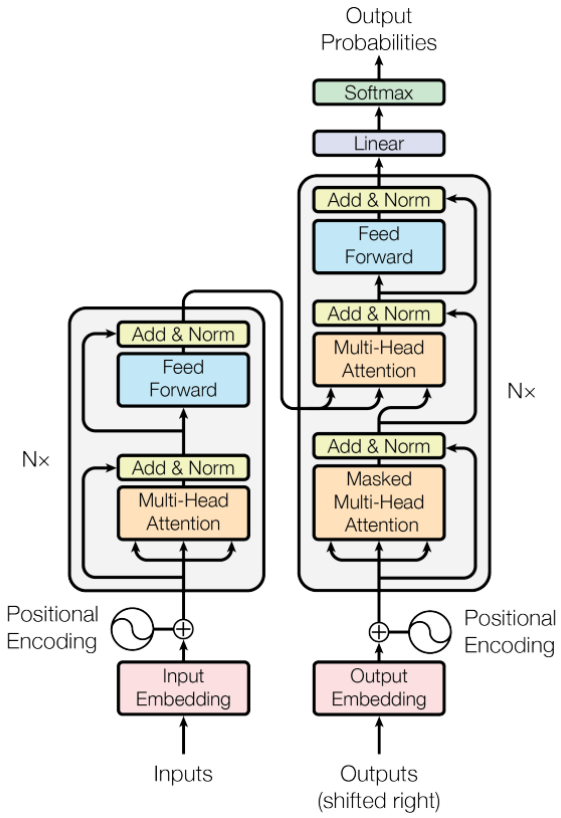


Figure 3.6: Representation of the original transformer model.
Its architecture includes several stacked layers (N) of encoder (left) and decoder (right) models. Each is composed by sub-layers of multi-head self-attention mechanism, followed by position-wise fully connected feed-forward network. The final output is predicted through linear transformation and a softmax function. Extracted from [113].

3.1.6 Decoding Methods

In text generation, the quality of the sequence that the model can generate depends not only on the success of model training and it's performance, but also on the decoding strategy. Decoding in NLG can be seen as a search problem, where the task is to find the most likely sentence (y) in and the infinite search space over potential sequences (V) that could be generated for a given input x:

$$y = arg\max_{y \in V} P(y/x) \tag{3.1}$$

The selection of decoding method and the definition of parameters that work better depends on the use case. Each method has its pros and cons, some allow to generate more variable and creative text, that may be useful for some applications, other methods just generate the most likely sequence [131].

Greedy decoding is the simplest strategy to approximate the likelihood objective, that at each decoding step chooses the token that has the highest conditional probability. It optimizes the probability of the sequence in an entirely local way, thus it can end up generating repeated or repetitive sequences [75]. Alternatively, the random sampling method tackles this issue by introducing some randomness, choosing a random token based on its probabilities.

A more widely used approach is Beam search, suggested by Lowerre [80], which is among the methods mentioned above, being able to balance Greedy decoding rightness and Random sampling variability. To do so, it considers a fixed number of candidates on the beam (k), at each step of the search procedure, adding words and keeping the k high-scoring candidates on the beam and pruning the remaining sequences. The final selected sequence corresponds to the most likely candidate considering the combine probability of all these tokens.

The key parameter is width (k), beam search with $k = 1$ is similar to greedy decoding which generates a single hypothesis with the most probable word given the previous words at each time step. On the other hand, beam search with $k = \infty$ is equivalent to a full breadth-first search.

To avoid a bias towards short generation outputs, beam search in neural generation requires a good stopping criterion or some way of normalizing scores between candidates. Over time, several versions and variants of this algorithm emerged, proposing different approaches and solving their biggest shortcomings [131].

3.1.7 Metrics

The generated text can be evaluated by human judgment, but this process involves the usual problems, such as subjectivity and difficulty in scaling. Thus, some metrics were developed in an attempt to automatically evaluate the generated text more objectively. The first metrics emerged for evaluating machine translation, however, they can also be applied for other NLP task including captioning [6]. The most common metrics are BLEU [87], METEOR [69], ROUGE-L [128] and more recently, CIDEr [114] and SPICE [3].

BLEU, or Bilingual Evaluation Understudy, was proposed by Kishore Papineni, et al. [87] and it is the most used metric because is simple to compute, easy to understand and is language independent. The evaluation consists of a comparison of the generated with the reference sentence and the score range between 1.0 for a perfect match and 0.0 for the worst case. However, it is often expressed in a range of 0-100. This comparison is done in phases by matching sentence segments (unigram to 4-grams, called respectively 1-BLEU to 4-BLEU) regardless the order. However the final score varies depending on the number of references and the size of the text generated and in practice, the perfect score is not possible, not even for humans [87].

BLEU does not consider meaning, so differences in function words are equally penalized as differences in content words. This contributes to its low correlation with human judgment.

The cumulative BLEU is calculated using the equation 3.2 that includes the weighted (W_n) geometric mean of modified precision (p_n) equation. 3.3 used to take into account the frequency of occurrence of each n-gram in the reference giving, less value to very abundant n-grams. Brevity Penality (BP) is used to penalize short sentences. BP is computed according to equation. 3.4 where c and g represent the length of the generated and referenced phrase, respectively [87].

$$BLEU\ score\ = BP * \exp \sum_{n=1}^{N} w_n \log(p_n) \tag{3.2}$$

$$Modified\ n-gram\ Precision(p_n) = \frac{\sum_{C \in Candidates} \sum_{n-gram \in C}\ Count_{clip}(n-gram)}{\sum_{C' \in Candidates} \sum_{n-gram' \in C'}\ Count_{clip}(n-gram')} \tag{3.3}$$

$$Brevity\ Penality\ (BP) = \begin{cases} 1 & c > g \\ e^{1-\frac{g}{c}} & c \leq g \end{cases} \tag{3.4}$$

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is most used to evaluate text summaries by comparing n-grams and word sequences with references. It calculates the precision (P), recall (R), and F1-measure based on matching of n-grams or longest common subsequence (LCS) between generated and ground truth sentences, ROUGE-N and ROUGE-L respectively [128].

METEOR (Metric for Evaluation of Translation with Explicit ORdering) on the other hand, calculates the BLEU1 and computes the F_{mean} of matching results, as shown in the equation 3.5. This can be extended to longer n-grams by adding some penalty, as different ordered words and synonyms are considered as possible matches [69]. This tackles the main drawbacks identified in the BLEU metrics ending up with a better correlation with the human assessment.

$$F_{mean} = \frac{10PR}{R+9P} \tag{3.5}$$

Unlike the previous ones, Consensus-based Image Description (CIDEr) and Semantic Propositional Image Caption Evaluation (SPICE) were specially developed for image captioning. CIDEr apply the tf-idf metric to attribute weight to n-grams accordingly with frequency across the dataset giving less importance to common words since those are less informative [114]. CIDEr compares and measures how often n-grams of the generated sentence are in the ground truth ones, to do so it considers the word stems.

$$CIDEr_{\ n} = \frac{1}{m} \sum_{j} \frac{g^n(c_i) g^n(g_{ij})}{||g^m(c_j)||\ ||g^m(g_{ij})||} \tag{3.6}$$

$$CIDEr\ score = \sum_{n=1}^{N} Wn * CIDEr_{\ n} \tag{3.7}$$

SPICE metric focuses in measuring semantic aspects, extracting relations between generated, ground truth captions and attributes. For that, it uses a graph-based representation of the text,

called scene graph [3]. In the mathematical perspective SPICE computes F1-measure for caption evaluation as shown in equation 3.8.

$$SPICE\ score = F_1 = \frac{2PR}{R+P} \tag{3.8}$$

Although correlated, no metric perfectly matches human judgment, each considers different linguistic aspects and each has its advantages and disadvantages. Thus it is recommended to combine several metrics to be able to assess various language dimensions. All but METEOR and SPICE are affected by word order. However, METEOR performs stemming and synonym matching but this impairs the assessment of semantics. BLEU and ROUGE are the most used, followed by CIDEr and METEOR. On the other hand, SPICE is still almost not used in the captioning of medical images, although it outperforms the other metrics in terms of agreement with human judgment [6].

3.2 State of the Art

Natural language processing methods have been widely used in clinical texts to address different tasks including assisting health professionals and clinical decision-making [4, 134].

The following subsections will focus on work that has been developed in the clinical field. Starting with text modelling, followed by a literature review on image captioning and finally on NLP for biological signals. In addition, in appendix A can be seen summary tables of state-of-the-art approaches referred in this chapter.

3.2.1 NLP for Clinical Texts

Clinical texts include expert insight into the patient's clinical condition and diagnosis. Clinical notes and structured data are complementary and extremely important since they contain the clinical history of patients, their condition, diagnosis and respective outcomes [4, 50]. Despite the richness of information provide by clinical notes, they are underused in relation to structured data.

Building models that can efficiently learn representations from clinical texts is challenging. Machine learning algorithms are suitable for running on structured data. The clinical text, on the other hand, has high dimensionality, is sparse and unstructured data [50]. Thus, it is necessary to resort to text pre-processing (section 3.1.3) and vectorization techniques (section 3.1.4) to represent the text data adequately for the application of the ML and DL algorithms [17, 24].

Modelling clinical text is complicated since clinical notes are long and words are interdependent. Therefore, typical word embedding models (Word2Vec, GloVe or fastText) that learn local word representations may not be ideal for the task. Thus, attention mechanisms, implicit in transformers, have been proposed in the literature as mechanisms that capture long-range information and relation and that allow obtaining some model explainability hints [24, 50].

One of the most used transformers in the field is the BERT, its design allows it to be pre-trained and easily adapt to a wide range of NLP tasks [33, 70, 104].

The main problems addressed in the clinical field are text classification (i.e., document classification [11], phenotyping [16]) and Relation Extraction which consists of the identification of relations between clinical concepts [81, 134].

A vast work has been developed in text classification, with very successful work developed by Biswal et al. for document classification of EEG reports. In [11], the authors use a supervised method to categorize text EEG reports detecting the ones that describe seizures and epileptiform discharges. The proposed pipeline consists of an initial filtering of the reports based on keywords, in which the reports that did not have any keywords or synonyms were classified as negative. The remaining reports follow a more careful classification. As EEG reports are not distinguishable by the mere frequency of occurrence of keywords, methods as Bag-of-words do not achieve satisfactory results in this task. Thus, Biswal et al. [11] proposes a new method to obtain more discriminative features called "elastic word sequence". This method takes into consideration keywords and the sequence of their occurrence, in a flexible manner. Those sequences are used as semantic features for the report classification. the classification was performed through the Naive Bayes model achieving an area under the receiver operating characteristic (AUC) of 0.975 [11].

Although not focusing on clinical text, many other authors address text classification. For this task, the key DL approach is to use word embeddings and convolutional neural network (CNN). However, recently, the very deep convolutional neural network model (VDCNN) has also been considered [17].

Electronic health record information (EHRs) hold clinically relevant data. However, that is not fully used in the clinical context as review becomes unpracticable [50]. Furthermore, there is undoubted clinical value in the corpus of the medical reports that are not being used. NLP has the potential to unlock this unused information by processing it and extracting clinical data [4].

Recently, Maldonado et al. [81] proposed a method to overcome the lack of the annotation of corpus's EEG reports by automatically identifying specific EEG and clinically relevant concepts. The annotation also includes the identification of brain signals' attributes such as Morphology, Frequency Band, Background, Magnitude, Recurrence, Dispersal, and Brain Location. This was achieved using two stacked LSTM networks to filter the parts of the report that include medical concepts. Then, the DRLN (Deep Rectified Linear Network) architecture was used for multi-task classification of the attributes and a self-attention neural model to carried out the distance relations between concepts, attributes, and relations in the EEG report.

Addressing a different problem, the authors in [50] use clinical notes to predict 30-day patient readmission. The patient evaluation is performed at various time points assigning a score of risk of readmission. For modelling the clinical notes, the authors adopt BERT to learn text embeddings.

Due to the scarcity of annotated clinical data and privacy concerns about sharing, pre-trained models and transfer learning are growing in popularity, playing an important role in enabling robust models to be obtained in medical studies [4, 17].

The framework implemented in [50] follows this trend. The BERT model is pre-trained using clinical notes (ClinicalBERT) and then fine-tuned for the readmission prediction task. ClinicalBERT has been compared with some of the relevant state-of-the-art models and outperformed in

all the metrics. Obtaining 0.674 AUC for predicting patient readmission between 12-24h versus 0.648 AUC and 0.649 AUC of the Bag-of-Words and Bi-LSTM models, respectively. Furthermore, when evaluating the similarity between learned embeddings and medical concepts, it is concluded that ClinicalBERT can learn more accurately the data distribution of clinical texts and, therefore, medical concepts, compared to the other models (Word2Vec and fastText) [50].

There is a differentiation in the choice of deep learning architectures through the task, for example, for text classification, the literature approach tends to follow CNN architectures, while for relation extraction RNN architectures are more common [134].

3.2.2 NLP for (Medical) Image Captioning

Image caption generation is the automatic task of describing the content of an image through natural language. Captioning is a complex task for a machine since, in addition to generating the human-readable textual description, it has to successfully extract the semantic content of the image. This goes beyond object detection, segmentation, and classification since it implies a deeper understanding of the content, including the relations between different objects and how they interact [6, 17].

Medical imaging, like ultrasound, is very noisy and the amount of annotated data is limited. To tackle this lack, a set of medical image datasets was created, and is used for research in this field: IU Chest X-Ray [132, 121], BCIDR [133], CheXpert [54], MIMIC–CXR [61], PadChest [19] and ICLEFCaption [34, 42].

Aside from the fact that image captioning is usually designed for short sentences, medical reports are descriptions with a variable number and longer sentences. This task becomes even more complex, as in the clinical field there is only one caption per image, while in classic captioning problems there are several options for descriptions per image, facilitating model training.

Despite this, over time, several computer-aided systems have been proposed based on different image captioning methods. The use of deep neural networks to generate medical image descriptions has replaced the traditional models, template-based and retrieval-based methods (or nearest-neighbor-based methods). In the template-based approach, image attributes are extracted, and the description is generated according to specified grammatical rules or sentence templates. Although it results in grammatically correct captions this approach is limiting, as it restricts description flexibility and diversity since the templates are predefined. Moreover, this approach lacks in scalability and is unable to generate variable-length captions. On the other hand, retrieval-based methods compare the input image to the database and retrieve the caption of the most similar images . The final description is selected from these pool of candidates captions or is a combination of many candidates.This approach ensures syntactically correct, but is not image-specific [1, 6, 48].

Thus, state-of-the-art methods usually rely on DL-based approaches that can generate new captions for each image, being more accurate than previous approaches. DL-based methods can be grouped according to several criteria. Figure 3.7 summarizes the taxonomy of these methods.

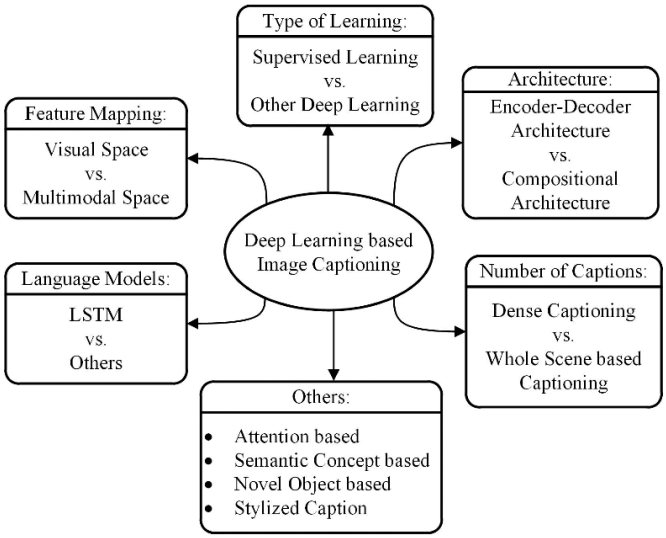


Figure 3.7: Overall taxonomy of deep-learning-based image captioning .
Extracted from [48].

Regarding the feature mapping, the image features and the caption can be interpreted as independent (Visual space-based methods) or can be mapped jointly into the same space (Multimodal space-based methods). Kiros et al. [65] started work in this area using a multimodal image+text representation for image captions generation. To do so, Kiros et al. proposed the Encoder-Decoder architecture inspired by machine translation.

The mentioned architecture includes feature extraction and language models that are trained jointly. In this end-to-end approach, the features extracted from the images by the encoder are used as input for the first of the decoder's time-step (language model) to generate a sentence. In contrast, in the compositional architecture the blocks are built independently, starting with extracting semantic concepts from the image, typically via a CNN, followed by generating a set of candidate captions based on previous semantic concepts and final caption selection according to multimodal similarity [48].

Most of the state-of-the-art methods use an Encoder-Decoder approach so this architecture will have greater focus. Feature extraction block entails obtaining high-level features to create an understanding of the image. Due to the large number and high dimensionality of image features, reduction methods are used to select the most relevant and summarize them in a fixed-length vector. CNNs can extract hierarchically the image features, being for this reason used as an encoder. However, CNN models need many examples in the training phase for effective learning features, which contrasts with the limited number of images in medical datasets. Thus, to deal with the scarceness of data, transfer learning and pre-trained models are frequently employed, as VGGNet [105], ResNet [45], InceptionV3 [109], GoogleNet [109] or AlexNet [66].

Language models predict the probabilities of the next words given a sequence of words until the **'<end>'** token be generated. They work as decoders that receive embedding vectors from the encoder and generate the description accordingly. Typically, language models are based on

recurrent neural networks such as LSTM [46] or GRU [23].

Shin et al. [103] was one of the first to use the encoder-decoder approach for chest X-rays image. In this work the author pre-trained the CNN model (GoogleNet) for classification and then retrained a CNN-RNN to generate the context of detected diseases. Despite the relevance of this architecture and the good results obtained, only sets of words were generated (based on Bag-of-words) to describe the context of the disease instead of a coherent text. Since then, many researchers have been using this architecture in the medical field, for example, in [90] Pelka predicted keywords using a CNN-LSTM model, fine-tuned through Inception V3. Also, Zeng et al. [132] used this architecture for ultra-sound image captioning. However, Zeng et al. [132] took a slightly different approach by dividing the main task into two subtasks: detecting the image region of interest using Faster-RCNN and generating the annotations of content focus area using an LSTM language model. Although this model performed better on both tasks than the individual models, it made some linguistic and classification errors and is only capable of generating short descriptions.

Recently, attention mechanisms have been introduced to improve performance, allowing the decoder to learn to place attention on certain pieces of information, such as the salient parts of the image [6]. Similarly to original encoder-decoder approaches, in their combination with attention mechanisms, the scientific community also strongly uses pre-trained models and fine-tuning.

The most common mechanism is authored by Xu et al. [124] and is a spatial-visual attention mechanism over image features. In his work, Xu [123] uses a ResNet to encode images and extract features, then soft visual attention was computed over the features. A simple LSTM was used as a decoder to create a description in a phased manner generating one word at a time based on hidden states, previous words, and context vectors.

On the other hand, You et al. [129] proposed a semantic attention mechanism, over tags of the images, and Jing et al. [60] used the properties of both mechanisms by combining them in a co-attention mechanism. The framework adopted by Jing et al. is presented in Figure 3.8.

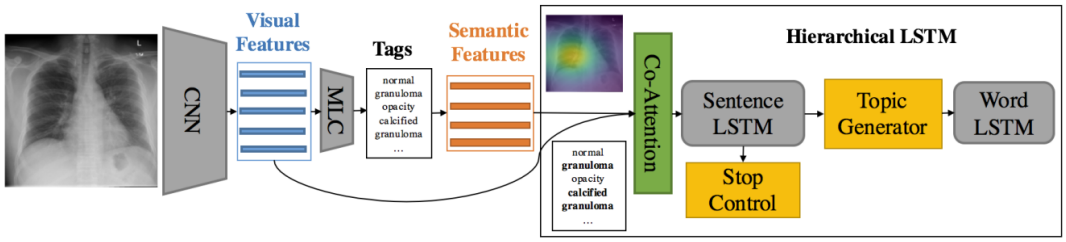


Figure 3.8: Illustration of the Jing et al. model. MLC denotes a multi-label classification network. Semantic features are the word embeddings of the predicted tags. The more provaveis tags "calcified granuloma" and "granuloma" are attended by the co-attention network . Extracted from [60].

They proved that co-attention is more beneficial for detecting and describing normal cases and abnormalities in chest X-ray images than the individual use of visual or semantic attention. In his work, the author approaches the problem by segmenting the image in the region and uses VGG19

to extract visual features. These features were used for a multi-label classification task where the model predicts tags for each region. Then, the description was generated through co-attention and using hierarchical LSTM, allowing the generation of long paragraphs.

Hierarchical LSTM consists of sentence-LSTM, which generates a topic for the input image, and word-LSTM, which generates, word for word, the corresponding description. This allows to generate high-level topics and from them write detailed descriptions. The proposed model outperforms the remaining models tested, including a traditional CNN-RNN approach without attention and a soft-attention model. However, this approach has some shortcomings as the generated reports have some repetitions.

The previously mentioned DL-based approaches produce captions based on the entire image, hence are known as whole scene-based methods [48]. The alternative approach refereed in Figure 3.7, called dense captioning, was proposed by Johnson et al. [62] and consists of generating captions by region. The relevant regions of the image are determined using a dense location layer.

Apart of pure encoder-decoder based approach, others have been developed, such as the generation of captions based on the combination of retrieval models with neural networks, building a retrieval model capable of generalizing to new examples [6]. Examples of this are the work developed by Biswal et al. [13] and Liang et al. [77]. In this work, Liang et al. also proved that the models performance is affected by caption lengths.

The authors of [13] compare the performance of their model with various chest X-ray caption methods, including the common CNN-RNN with visual attention [60], a template-based approach using a graph transformer-based neural network [74] and Li et al.'s [73] approach that uses reinforcement learning to decide when to retrieve a report from the database or generate a new report. Biswal's model [13] ended up outperforming all others, achieving 37,4 CIDEr and 48,9 BLEU1 in contrast to 34,3 CIDEr and 43,8 BLEU1 for [73], the second best result, and 27,7 CIDEr and 45,5 BLEU1 for [60].

Reinforcement learning is sometimes applied to drive model learning, but it has some drawbacks, such as the reward hacking problem that occurs when the reward function is overfitted. In that case, the function is maximized without improving caption quality ending up with unnatural sentences [13, 35]. To overcome this problem meta-learning methods can be used, as introduced in [76] which simultaneously optimizes the reward function (reinforcement task) and supervises the main task by taking gradient steps in both directions. Reinforcement learning and generative adversarial network (GANs) are some of the alternatives to classical supervised learning (Figure 3.7) approaches to captioning that have obtained very satisfactory results.

GANs are composed by the discriminator that distinguishes when it is a artificial caption or written by humans and the generator that produces the artificial captions [21, 35]. Through an adversarial game, the two sub-models are trained and improved for their tasks, with a gradual improvement of the artificial samples being visible, which become more and more realistic, until the Nash equilibrium is reached. This equilibrium occurs when the discriminator is no longer able to distinguish real samples from artificial ones, so the discriminator outputs the similar probabilities of classifying a samples as real or false [18, 43, 120].

Although GANs work well for images, text is not real value data, so it is difficult to apply back-propagation directly as operations are not differentiable [48].

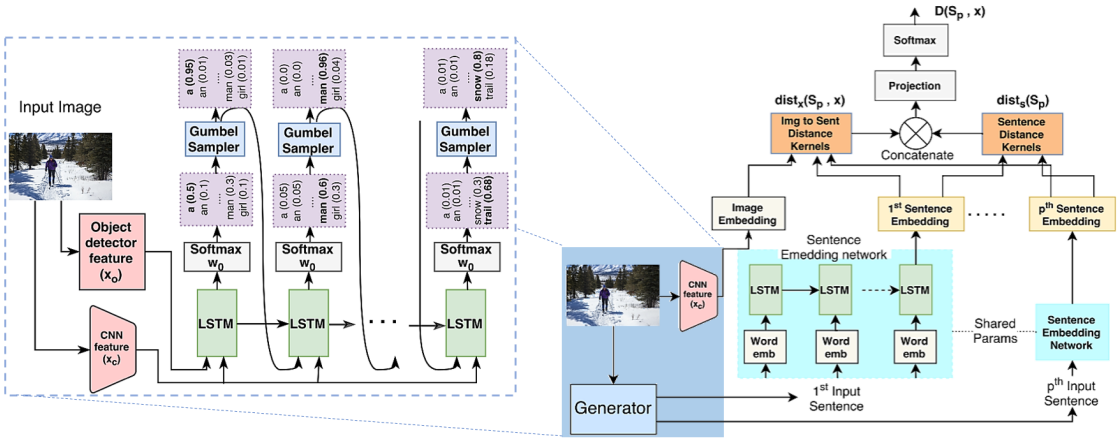


Figure 3.9: Represention of a GAN-based caption model.
In Caption generator model (left) visual features are input to an LSTM to generate a sentence. A Gumbel sampler is used to obtain soft samples from the softmax distribution, allowing for back-propagation. The Discriminator Network (right) scores the set as real/fake calculando the image to sentence ($dist_x(S_p,x)$) and sentence-to-sentence ($dists(S_p)$) distances. Adapted from [102].

Despite the challenges, some strategies have been proposed to overcome the limitations in the back-propagation of discrete data, such as using Policy Gradient or Gumbel sampler [48, 57]. This has allowed to develop successful GANs-based captioning model which, unlike conventional DL-based caption methods, can generate a diverse set of captions [30, 102]. Figure 3.9 outlines a GANs-based caption model proposed by Shetty et. al. [102].

3.2.3 NLP for Biological Signals

The application of NLP techniques to describe biological signals is relatively recent. This area arose from the adaptation of image captioning approaches to extract information and describe signals.

In the medical field, the goal of signal captioning is to generate a clinical description that reflects the patient state and the content of the biosignal. Kiyasseh et al. [135] contributed to bio-signal captioning by proposing a captioning model for 2-lead ECG. To do so, the author used a encoder-decoder architecture with both submodels being previously trained. To learn the representations of ECGs, the encoder (CNN) was trained in the classification of cardiac arrhythmia. Likewise, the decoder was pre-trained to learn word embeddings in specific languages by performing token prediction and masked token detection. The combination of both allows extracting the temporal features of the signal. Those can be used to implement either the standard visual attention mechanism [124] or multi-head attention, using a Transformer decoder. The results suggest that the models can generate ECG reports with good diversity and that reflect the high-level clinical information, capturing the general pathology and more specific aspects of the ECG content [135].

In the case of EEG, Biswal [14] proposed EEGtoText model (Figure 3.10) which uses CNN to extract shift-invariant and temporal patterns and then classifies key phenotypes. Biswal follows a modelling approach where phenotypes are used to generate reports filling the template report. A detailed explanation is generated using hierarchical LSTM (paragraph and sentence level) which combined with attention mechanism allows to locate of abnormal areas in the EEG and provide a consistent explanation of the extracted phenotypes [14].

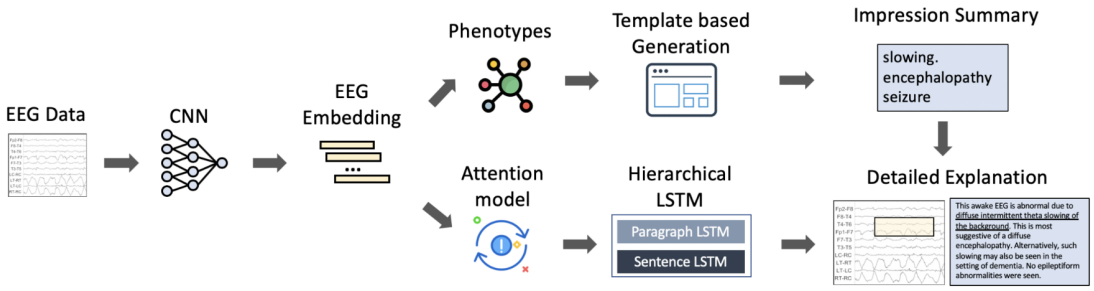


Figure 3.10: Overview of Biswal et al.'s framework for generating EEG reports. They use an CNN encoder to create EEG feature vectors, that are used for phenotype classification and also passed to the attention module to generate a final context vector for details decoder. The impression section is generated through a template approach in which the phenotypes are fed and the detailed explanation is generated by the decoder based on the impression section and context vector. Extracted from [14].

More recently, the same authors proposed an adaptation and combination of a neural network and retrieval models to generate fast and reliable clinical reports. Biswal et al. [13] propose CLARA (Clinical Report Auto-completion), an interactive method for generating reports by a method of retrieval at the sentence level, based on the extracted phenotypes, clinician inputs, and later editing. Furthermore, CLARA is an auto-completion model, resorting and relating anchor words and prefix text, with report text to generate more informative queries in the retrieval process. The query auto-completion technique was established by [20],initially only to select the relevant queries, more recently the application of LSTM [56] and hierarchical encoder-decoder allow generating unseen queries. The retrieved sentence is then edited to produce a new sentence, which will be part of the medical report, by applying a sequence-to-sequence model [108].

Biswal et al. compared the performance of both models [13, 14] in phenotype prediction and captioning with other models inspired by video captioning and machine translation approaches, for handling signal features produced by a CNN. Thus were consider, the Mean-pooling model (MP) [117] which uses mean pooling to combine signal features and then the LSTM based model to generate the textual description, the S2VT [116] model which uses a model based on LSTM to generate captions but is preceded by LSTM to process the resource vector, the Temporal Attention Network (TAM) [127] which resorts to 3D-CNN and attention mechanisms for explored motion features and finally the Soft Attention approach (SA) [7] that makes use of the corresponding attention mechanism to allow the decoder to focus on representations of EEG features.

CLARA [13] outperforms the remaining models in both phenotype classification and report regeneration in all the metrics. Obtaining an average accuracy of 0.834 in contrast to 0.793 the second best result that corresponds to the EEGtoText model [14]. At the reporting level, 45,5 CIDEr and 78,4 BLEU1 were obtained in the TUH EEG dataset, which corresponds to an increase in the performance of the previously best model (EEGtoText) which obtained 38,1 CIDEr and 75,2 BLEU1. The following (Table 3.1) summarizes the performance of each model on generation of impression section of the clinical report. Those models were tested on Massachusetts General Hospital EEG dataset (MGH) and the Temple University Hospital EEG dataset (TUH) .

Table 3.1: Model performance in generating the impression section based on EEG time series data (MGH and TUH).

EEG Dataset	Method	CIDEr	BLEU1	BLEU2	BLEU3	BLEU4
MGH	*MP*	36,7	71,4	64,4	56,3	44,3
	S2VT	31,9	74,1	62,8	52,9	46,2
	TAM	33,4	74,9	66,8	58,1	37,8
	SA	34,8	68,4	62,9	56,8	47,2
	EEGtoText	37,2	74,2	72,8	58,7	38,1
	CLARA	44,3	76,2	68,4	61,4	46,4
TUH	*MP*	36,3	64,5	57,8	45,9	36,1
	S2VT	36,4	72,4	61,3	54,3	43,8
	TAM	38,4	71,4	64,7	49,2	46,1
	SA	34,1	73,6	61,9	51,9	42,0
	EEGtoText	38,1	75,2	61,8	59,3	42,8
	CLARA	45,5	78,4	65,6	62,4	48,3

As previously seen, the work on signal captioning field has evolved mainly by the adaptations of the approaches already widely applied in other fields such as image but especially video captioning. Video captioning is comparatively the closest task in terms of the challenges that it presents. Videos, like biosignals, are temporal sequences and for description generation it is necessary that the model can effectively handle the variable length and dynamic nature of video.

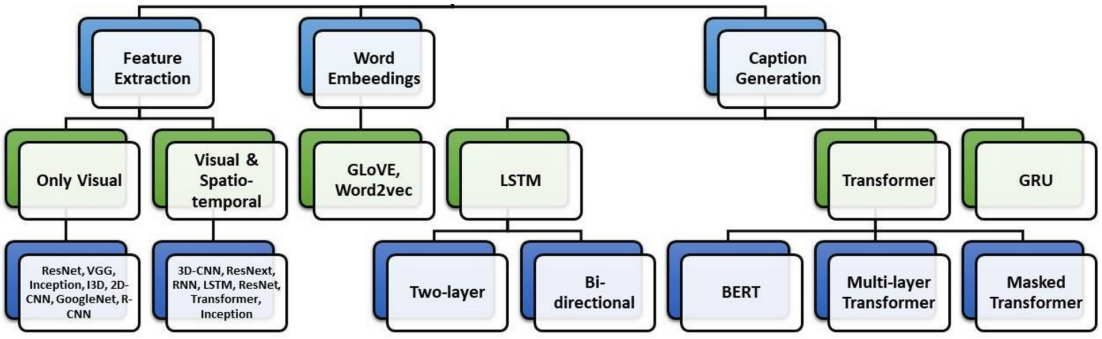


Figure 3.11: Overview of different deep learning methods in Video Captioning. Adapted from [55].

For that, one option is to use recurrent networks on the encoding stage, and resort to attention mechanisms, as in the aforementioned S2VT and SA models. Several other strategies have been considering different dimensions of the video by using multi-modal and multi-stream features [51]. Figure 3.11 present an overview of different DL methods in video captioning.

Multi-stream architectures can accurately exploit spatio-temporal features by having different (specialized) streams to capture objects, object actions, or interactions between different objects [119]. Those architectures usually considering static features extracted from a single frame selected from a fixed or random position, which focus on detecting the objects present in the video. And dynamic features usually extracted by 3D-CNN, which give a sense of movement, for instance express in optical flow frames [126, 127]. Likewise, an image captioning static and dynamic features can be both visual and semantic features [31, 72]. The combination of all streams is then considered for the description generation by the decoder state [125].

Chapter 4

Methods

The main focus of this study was to develop a model to generate reliable clinical EEG reports that describe the findings in a language as close as possible to natural human language. For this purpose, several approaches were studied for each of the steps involved in EEG captioning task. The present chapter describes the database and methodology used.

4.1 Dataset

We use the Clinical EEG database of Temple University Hospital (TUH EEG Corpus) publicly available from the Neural Engineering Data Consortium (www.nedcdata.org) [84]. The entire dataset is de-identified, ensuring compliance with the HIPAA Privacy Rule (Health Insurance Portability and Accountability Act). In total, it contains 16 986 sessions of 10 874 subjects. The age range of patients is wide, including individuals between 1 year and over 90 years old, with an average of 51.6 years. Regarding the patient gender, the database is fairly balanced, with 51% female and 49% male.

The database contains EEG records and respective clinical reports that were written by neurologists. For each session, the EEG was stored in one or more .edf (European Data Format) files, which is the standard format for storage medical time series. In addition, the corresponding clinical reports, were saved in .txt format.

Given the need for pre-training models evidenced in the state-of-the-art approaches (section 3.2) only annotated data were considered. Thus, the TUH EEG Seizure Corpus was used, which is a portion of the TUH EEG Corpus containing annotated sessions with seizure events and additional no-seizure sessions to balance the corpus. Despite this designation, the referred session called "seizure session" corresponds mostly to the interictal EEG with epileptic patterns, IEDs. Thus, from now on we will refer to them as epileptic and non-epileptic sessions.

All files contain the default channels expected in a 10-20 configuration, however, there is some variability in the corpus regarding the number of channels and configurations. The final training set considered contains 687 sessions following the average reference configuration (AR) and the test set consists of 182 sessions.

4.2 Data Preparation

4.2.1 Signal Pre-processing

The TUH EEG Seizure Corpus consists of samples with different sampling frequencies. Although most EEGs have a sampling frequency of 250 Hz, there are also samples sampled at 256 Hz, 400 Hz or 512 Hz. For standardization and dimensionality reduction, the signals were downsampled to 125Hz. Furthermore, to reduce the influence of artifacts, the EEG data were previously filtered to the range 0.5-35Hz.

Then, the signals were referenced to a longitudinal bipolar montage. Figure 4.1 represents scalp electrodes and the 18 channels, according to the longitudinal bipolar montage. Finally, the records were split into 2s epochs without overlap, obtaining a matrix of dimensions 18x250 (channels X time) for each epoch. Those epoch matrices were saved together with the information corresponding to the annotation at each instant.

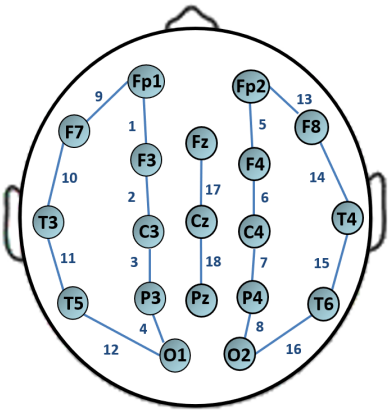


Figure 4.1: Representation of 18 channels (blue lines) in the longitudinal bipolar montage. Channels are connection between electrodes (represented as circles).

The annotations indicate an occurrence of Spikes and/or Sharp Waves (SPSW) which are transient epileptiform events typically present in patients with epilepsy. Epochs without seizure events were labeled as Background (BCKG).

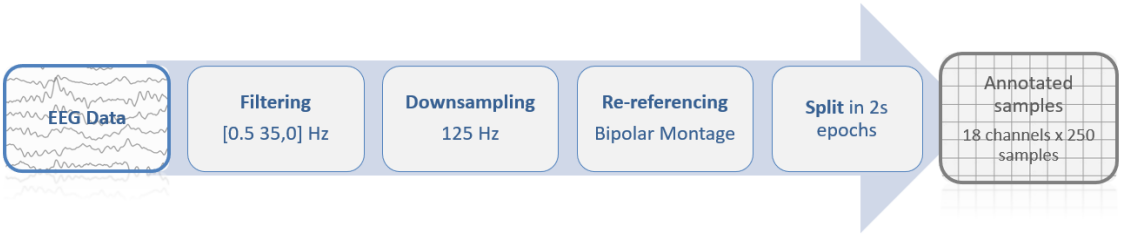
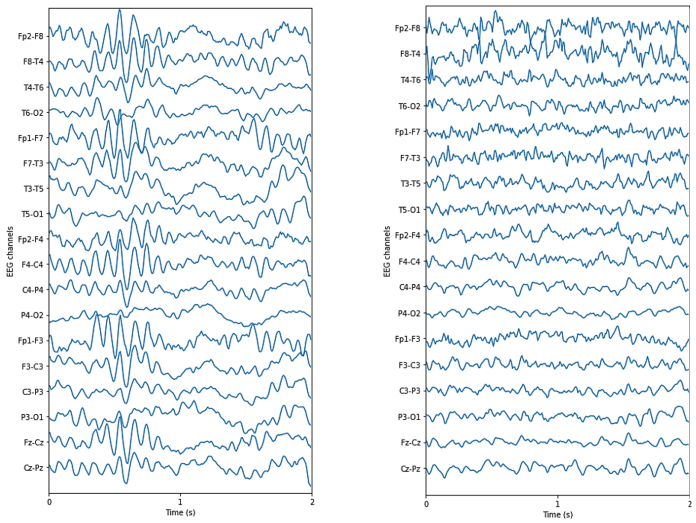


Figure 4.2: Summary of pre-processing steps applied to EEG data.



(a) SPSW sample (b) BCKG sample

Figure 4.3: Visualization of an EEG epoch of each class, epileptic (4.3a) and no-epileptic (4.3b).

An EEG epoch corresponding to each of the mentioned classes is shown in Figure 4.3, additional samples of each of the classes can be found in the Appendix B.2 and B.3 , respectively for SPSW and BCKG. Figure 4.2 presents a summary of all signal pre-processing steps. The pre-processing was implemented in Matlab R2021b (The MathWorks, Inc., Natick, MA) .

4.2.2 Report Preparation

Clinical reports are the official summary of the clinical impression containing the neurologist's insights and findings after EEG analysis. These reports are organized into several sections that describe the relevant patient medical history, medications, the EEG record and their correlation to a clinical condition.

Our goal is to automatically "translate" the relevant clinical aspects of the EEG signal into natural human language. Thus, similarly to the first part of Biswal. et al's [13] work, our focus was given to the "impression" section of the reports. The content of this section can be derived exclusively from the signal analysis and corresponds to a summary of the neurologist's findings. Thus, the impression section was isolated, extracted from the original reports and all sessions whose clinical report did not have an impression section (or equivalent) were discarded for the captioning task. Table 4.1 shows the TUH EEG Seizure Corpus samples and the portion of those we used for training and testing in the captioning task.

In addition, we standardized the reports and applied a series of text cleaning and normalization steps to prepare for use on models. These steps include removing punctuation and converting to lowercase. While creating the vocabulary we disregard tokens from the corpus if the frequency is less than two, being encoded with the special token **"<unknown>"**.

Table 4.1: TUH EEG Seizure Corpus

Set	Total AR Sessions	Sessions with impression section	Patients	Epochs (2s)	Times (hours)
Train	687	641	304	904640	251
Test	182	167	36	263307	146
Total	869	808	340	1167947	398

4.3 Implemented Pipelines

This section describes the proposed approaches to automatically generate clinical reports from EEG data. All the models were implemented in Python 3.8 using Keras 2.3 and a CUDA-enabled NVIDIA GPU (GTX-1080), running on CentOS 7.

In this project we focus on the reliability and clinical accuracy of the generated report. Thus, different encoder-decoder architectures and approaches to signal encoding were explored, in the attempt to obtain a meaningful and representative EEG embedding. Additionally, we explore the influence of different text representation techniques and decoding methods on the quality of reports.

4.3.1 Encoder

The encoder module is used to extract data embeddings from the input signal to guide the report generation. This is a CNN that compresses the original signal into a feature vector, keeping the most important information. The CNN used in all approaches was VGG16 and the features are extracted from the last convolutional layer, obtaining a feature map with size (9,2,512) which is then flattened and used as input to the decoder.

The VGG network was created by Karen Simonyan and Andrew Zisserman in 2014 [105] has been adapted for various applications. including for EEG. Particularly in 2019, it was adapted for IED detection by Lourenço et al. [79]. In that work, the authors reported that VGG16 yielded an AUC of 0.96, specificity of 99%, and sensitivity of 79%, outperforming state-of-the-art models. This proves that the VGG16 was able to successfully detect patterns and extract significant features from the EEG, evidencing the potential to be used as encoder. Details of the architecture of adapted VGG16 can be seen in the Appendix B.1.

Due to the scarcity of data, as suggested in the state of the art (section 3.2), we applied transfer learning techniques. To do so the encoder was previously trained on a classification task (SPSW / BCKG). In training, we applied five-fold cross-validation in the training/validation set, in which one of these partitions was used to validate the model and the others were used for training, changing the validation partition at each iteration. All EEG epochs from a patient in the iteration were used either for training or validation, similarly to what happens with patients for training and testing.

A sparse categorical cross entropy function was used to estimate the loss and a batch size of 64 was used. The stochastic optimization was performed using an Adam optimizer with a learning rate of $2x10^{-5}$, $\beta 1 = 0.91$, $\beta 2 = 0.999$, and $\varepsilon = 10^{8}$.

Given the large data imbalance (1:17), different class weights were applied, and the best performance corresponded to using class weights of twice the train imbalance. Note that the TUH EEG Seizure Corpus has a heterogeneous distribution of data, showing a lower class imbalance in the test dataset, which is characterized by a data imbalance of 1:11.

4.3.2 Text Vectorization

As mentioned in section 3.1.4, the representation of words strongly impacts the performance of the model, as it consists of all the language information that the model receives and whose distribution it needs to learn during the training process.

Thus, different approaches were used to represent the input text in feature vectors. The impact of each was evaluated through the final performance of the captioning model. The comparison includes the simple one-hot encoding strategy and several dense distributed representation methods such as the embedding layer stack to the model , Word2Vec and fastText. Those models were trained from scratch on our data and the weights were used as the initial state of embedding layer of the decoder model.

Additionally, the word embeddings mapping of Word2Vec and fastText were visualized through t-SNE. Furthermore, cosine distance was computed to evaluate the similarity between some word pairs showing some of the relationships that models are able to detect.

4.3.3 Decoder

An LSTM network is used as the decoder to generate the caption token by token. The decoding starts with the special token **<Start>** and at each step it receives the word embedding of each word of the caption and the weighted encoded signal as input to predict the next token. This process continues until the prediction of the token **<End>** or until the maximum number of time steps is reached.

On the CNN-LSTM approach represented in Figure 4.4 the global signal embeddings are fed in each time step to the decoder for the prediction of the next word. However, this may make the model more susceptible to overfitting and produce worse results. Alternatively, in Xu et al.'s [124] implementation, both the preceding word and the attention-weighted annotation vectors constitute the decoder input at each time step.

The training process involves Teacher Forcing, which is a strategy usually applied for training RNNs that at each step t uses ground truth at $t-1$ as input, instead of the model output at the previous step. This allows the RNN to converge faster and train more efficiently. However, it can end up bringing some instability and fragility to the model during inference [67].

4.3.4 Inference

Contrary to training, in inference the model does not use Teacher Forcing, instead the encoder receives the previously generated word at each time-step.

When generating text, encoder-decoder models require an additional decoding procedure to determine the output sequence, among the vast search space of potential sequences. The simpler approach is using the greedy method which chooses the word with the highest score in each time-step. However this method may fail to choose the best sequence. Thus other methods have been suggested in an attempt to find the optimum sequence [131].

As suggested by Vinyals et al., [118] we also implement the Beam search approach by defining the search width (k) as 3. In this approach, the model keeps a fixed number of candidate tokens, k, in each step and chooses the highest overall score among k candidate sequences.

Beam search requires a stopping criterion for searching and some way of normalizing scores between candidates in order to avoid a bias towards short sequences. So we normalize by dividing by the length of the sequence and attenuating by a factor $\alpha = 0.7$. Regarding the stopping criterion, the beam shrinkage strategy was used where the beam size is decreased each time a complete hypothesis is found, and the search ends when the beam size reaches 0 [7].

4.3.5 Implemented Architectures

In this project, several architectures were implemented, inspired not only by state-of-the-art approaches for image captioning, but also inspired by signal and video captioning approaches.

In all the attempts, a pre-trained encoder was used to extract EEG embeddings by epoch (f_t), then the following baselines were considered to obtained the global EEG embedding (f):

1. The average embedding suggested by Biswal et. This combination method consists on computing the average of all embedding epochs. al. [13].

$$f = \frac{1}{T}\sum_{t} f_t \tag{4.1}$$

This can be expressed by the equation 4.1, where T and t represent signal duration and epoch, respectively.

Once the EEG is summarized in a single embedding, it is possible to use image captioning approaches. Thus, within this method, we implemented an architecture inspired by the original Encoder-Decoder approach (CNN+LSTM) [65, 118] and an additional one inspired by Xu et al. [124] (CNN+Att+LSTM), who were the first to introduce an attention-based mechanism for captioning.

The illustration of the CNN+LSTM and CNN+Att+LSTM architectures are shown in Figures 4.4 and 4.5, respectively.

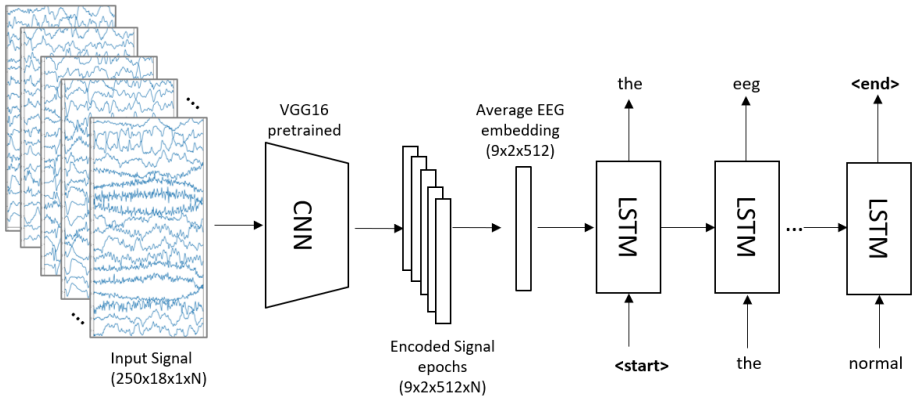


Figure 4.4: CNN-LSTM architecture.
The encoder (VGG16) receives the EEG signal, outputting the EEG epoch embeddings. These are averaged, forming a global EEG embedding which combined with the previous word consists of decoder input for text regeneration.

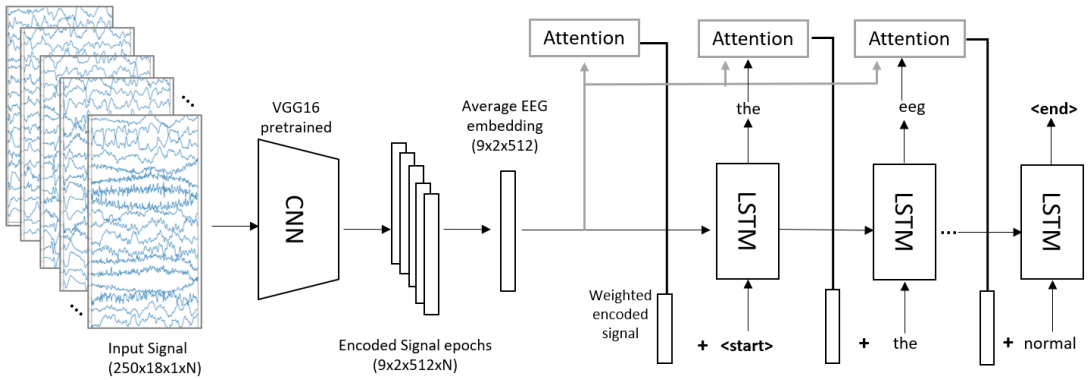


Figure 4.5: CNN-Att-LSTM architecture.
The encoder (VGG16) outputs are combined in a global average EEG embedding. An attention mechanism is applied over the global EEG embeddings allowing to create an weighted encoded signal. In each time step, the weighted encoded signal along with the previous word is used by the decoder to predict the next word.

2. Sequence to sequence model (Seq2Seq), which uses an LSTM to process the sequence of CNN outputs and another LSTM to generate text. This model was inspired by the implementation of Venugopalan et al. [116] for video captioning.

This approach allows exploring the temporal information of the recording, since the EEG representation used as input to the LSTM is a sequence of embeddings, each one corresponding to a (consecutive) interval of the signal.

In an attempt to minimize the loss of temporal information that can be crucial in the case of EEGs, we replaced the original 1:10 frame sampling proposed by Venugopalan et al. [116] with the averaging of short epoch sequences. The use of these local temporal features has proven useful for video [127]. Figure 4.6 presents the adapted Sequence to sequence model for EEG captioning.

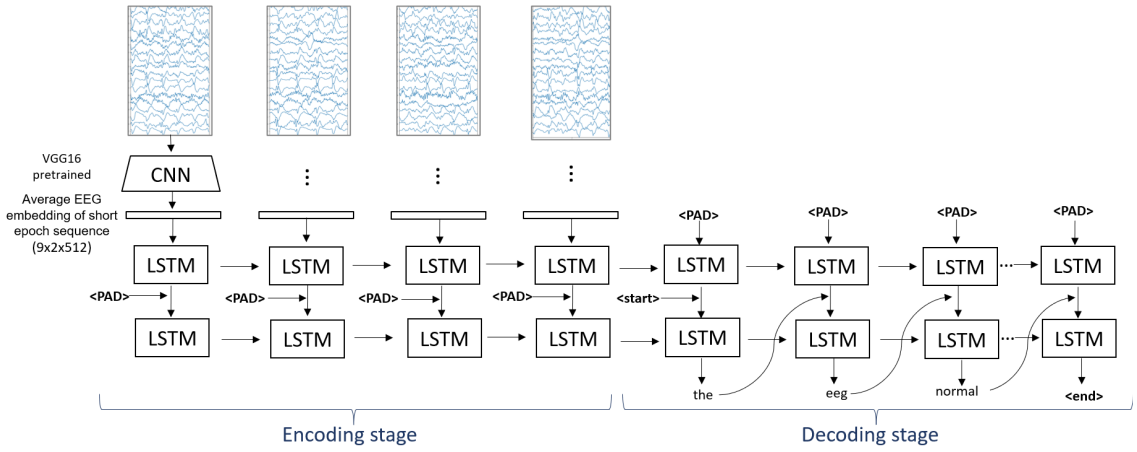


Figure 4.6: Sequence to Sequence model architecture (Seq2seq).
The two stack LSTMs learn to represent a sequence of embeddings into a sentence that describes the EEG. The top LSTM layer receives the embeddings and the second LSTM generates text given the input text and the hidden representation of the EEG.

3. Temporal attention mechanism (TAM) that allows to more efficiently explore the temporal evolution of the EEG, focusing on different segments of the EEG to produce the text report. This was inspired on several works by applying a soft-attention mechanism [7] over local temporal embedding [127].

Similarly to the previous method, those local temporal embeddings are the average embedding of short epoch sequences. The attention weights together with the previous word form the decoder input as shown on Figure 4.7.

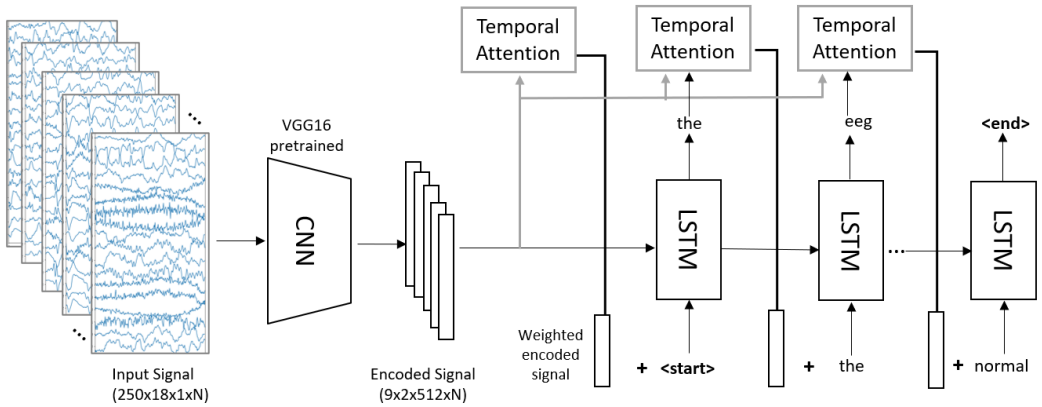


Figure 4.7: Temporal attention mechanism based model (TAM).
The outputs of encoder (VGG16) are combined in local average of short embedding sequences. Temporal attention mechanism is applied over those embeddings. The weighted encoded signal resulting from the attention mechanism emphasizes the most relevant EEG segments for text generation.

4. Multi-Stream models. We also explored new architecture inspired on previous approach and on multi-stream models applied on video captioning.

On the state-of-the-art presented in section 3.2 it is evident that in most cases the mere implementation of encoder-decoder architecture or the use of attention mechanisms on a single type of feature (e.g visual or semantic) is insufficient, and these methods are outperformed by co-attention mechanisms. This reveals the importance of using various feature domains. Using multi-stream architectures allows us to split feature extraction into streams specialized in capturing a specific type of features. For video captioning, the advantage of this approach is clear, since it allows to explore several feature domains usually static and dynamic features.

On the other hand, the advantage of applying those structures to time-series is not so evident since, unlike video, there are no objects or static features to capture. However, we hypothesize that it could be beneficial for the task to have information of overall signal and temporal variation, working analogously to the static and dynamic features of video respectively. Thus, inspired by previous architectures, we developed and tested two multi-stream models as shown on Figure. 4.8 and 4.9.

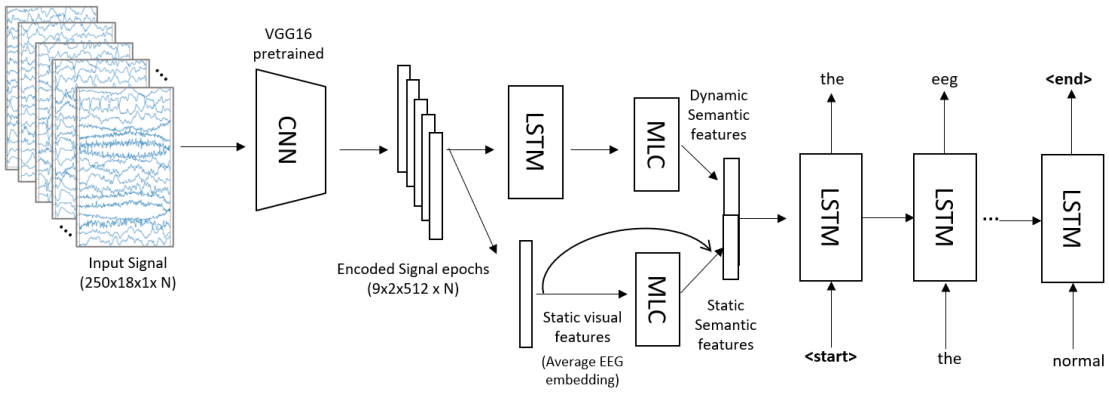


Figure 4.8: Multi-stream model architecture, using average EEG embedding (Multi-stream-Avg). The sequence of local temporal embeddings are used as based for the static and dynamic stream. In the dynamic stream, an LSTM receives the sequence of embedding, and it's output is process by fully connected layers to create dynamic semantic features. In the static stream, the sequence is summarized in global average EEG embedding. In this stream both, visual and semantic features are considered. The final EEG representation that is inputted into the decoder consists of a combination of static and dynamic features.

The first considers the average EEG embedding as the base of static streams and applies an LSTM on the dynamic stream. The average EEG embedding works as an overall summary of the signal. The dynamic stream is based on the extraction of features over time using a time-distributed layer and summarizes that temporal variation on a fixed length vector of features.

The second proposed architecture is an extension of the TAM model. In this architecture, the temporal features are primarily considered. Then, for the static stream, instead of selecting an EEG segment from a random time interval, as usual it is in video captioning, we select the segment with the highest attention weight.

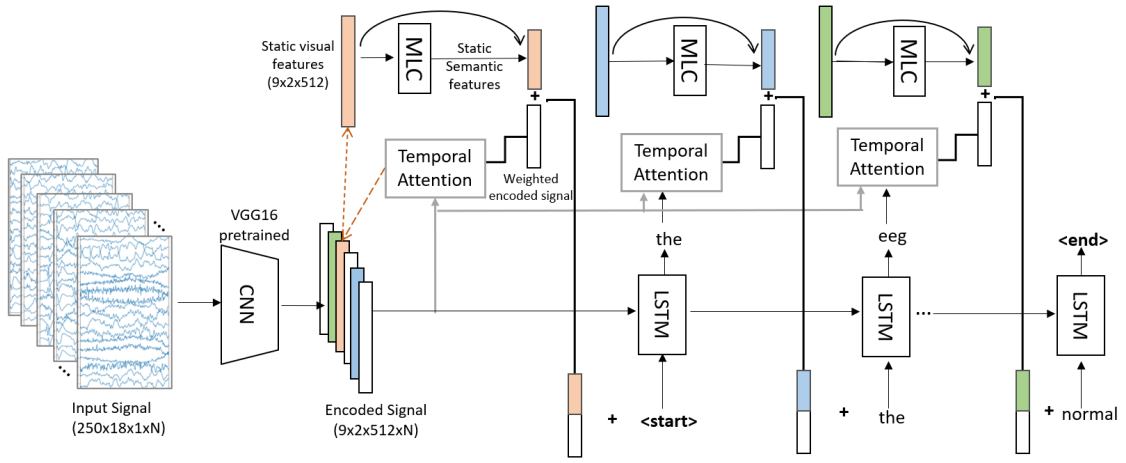


Figure 4.9: Multi-stream model architecture, using temporal attention (Multi-stream-TAM). A temporal attention mechanism is applied over the sequence of embeddings, allowing to obtain dynamic features. Based on attention weights, the most relevant EEG segment is selected and used as the basis for static stream. In this both, visual and semantic features are considered. The final EEG representation that is inputted into the decoder consists of a combination of static and dynamic features.

Ideally, features from different domains would be extracted by different models previously training them specifically to extract these features. However, in this case, this pre-training is not possible due to the absence of an explicit indication of the phenotypes or any other information that can be used for the pre-training of the second domain of features. Thus, for both attempts, we used the pre-trained VGG16 to extract the EEG embeddings from each epoch. These EEG embeddings were the basis for the streams, being then processed differently in each stream.

4.4 Training

To deal with limited data in a given domain, it is possible to apply transfer learning, reusing a trained network and applying it to a different problem. In all of the approaches mentioned, the pre-trained VGG16 is used as an encoder, rather than training a CNN from scratch.

All the captioning models were trained using the Adam optimizer and learning rate of $1x10^{-3}$, $\beta 1 = 0.9$, $\beta 2 = 0.999$, and $\varepsilon = 10^7$ and throughout the training process the encoder weights are not updated.

The training loss used was the categorical cross entropy which is applied over the word distribution, using:

$$Loss = -\log(\frac{\exp(x[word])}{\sum_j \exp(x[j])}) = -x[word] + \log(\sum_j \exp p(x[j])) \tag{4.2}$$

The input sequences have different lengths, so they are padded with zeros to create batch tensors of the same dimensions for training. To reduce the use of computational resources, the input

sequences were ordered by length. Ordering allows to reduce the waste of computing resources of computing the loss over the padded regions, for instance.

This strategy was applied over the captions in the approaches based on average signal embedding and over the signals on Seq2Seq, TAM and multi-stream methods. A representation of this batch creation strategy can be seen in Figure 4.10.

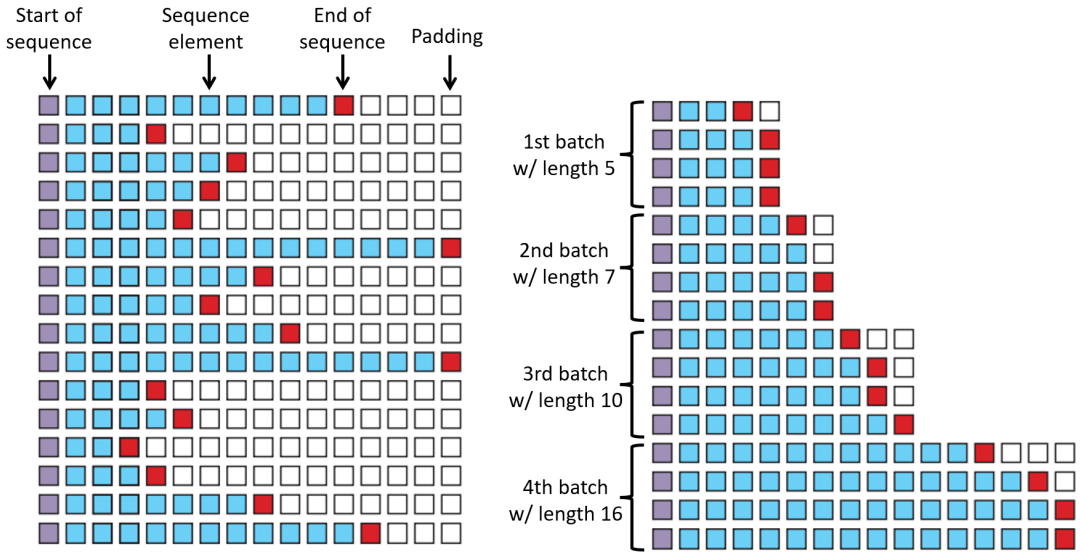


Figure 4.10: Bucketing applied to the sequence.
In batch creation the sequences are padded with zeros to create tensors with sequences of same length within the batch. In the optimized strategy (right) the sequences are ordered by length forming batches with shorter tensors. Extracted from [68].

4.5 Performance Evaluation

The performance evaluation was done using the benchmark metrics applied in NLP tasks mentioned in section 3.1.7. NLP metrics were calculated using a standard evaluation tool, *Microsoft COCO Caption Evaluation (pycocoevalcap)* [98].

The most common metrics used to report results are the cumulative BLEU1 to BLEU4 scores. However, taking into account that although none of the metrics corresponds perfectly to human judgment, but each allows the evaluation of different linguistic aspects, we decided to take advantage of all metrics, to have an overview as complete as possible of the quality of the generated text.

Thus, all models were evaluated through quantitative evaluation, using BLEU, METEOR, ROUGE_L, CIDEr and SPICE, and through qualitative assessment of the generated reports.

Chapter 5

Results

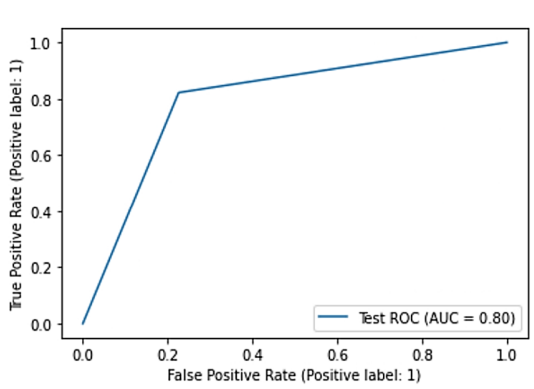
5.1 Performance Assessment of the Feature Extractor

In an attempt to ensure an extraction of significant features from the EEG, we pre-trained the VGG16 from scratch on the TUH EEG Seizure Corpus, and we applied transfer learning techniques to incorporate it on captioning.

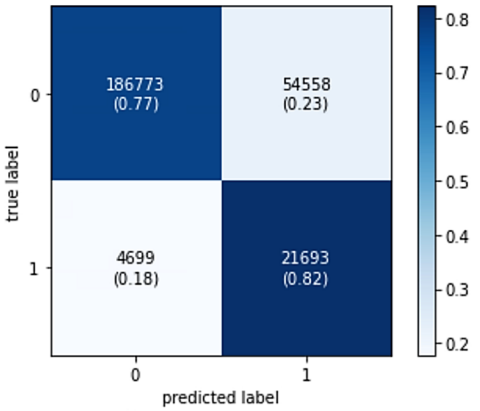
We obtained a classification model with an AUC of 0.80. Table 5.1 summarizes the model performance for each set.

Table 5.1: VGG16 Model performance for binary classification (SPSW and BCKG).

Set	AUC	Sensitivity	Specificity	Precision	F1-scores
Train	0,81	88%	75%	25%	0,39
Test	0,80	82%	77%	28%	0,42



(a) ROC curve on test set



(b) Confusion matrix (0-negative, 1-positive)

Figure 5.1: Result of the VGG16 model trained on TUH EEG Seizure Corpus

In addition, Figure C.1a shows the AUC of the VGG16 model on the classification of epileptic events (SPSW, label 1) and background (BCKG, label 0) over the test set. On Figure 5.1b, the confusion matrix is presented with additional indication of its normalized values.

5.2 Evaluation Impact of Text Representation Approaches

Firstly, all the architectures were tested applying the simplest strategy of text vectorization: one-hot encoding. The results obtained can be seen in Table 5.2.

Note that, as mentioned in Chapter 4, two configurations were tested for the multi-stream architecture. In this section, for simplicity, only the results corresponding to the Multi-stream-TAM configuration (Figure 4.9) are presented, since they present similar results. However the results for both architectures can be seen in section 5.3.

Table 5.2: Results obtained from the implemented architectures using one-hot encoding.

Method	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE_L	CIDEr	SPICE
CNN-LSTM	53,5	39,7	31,6	25,1	21,4	44,2	15,5	18,2
CNN-Att-LSTM	48,6	35,8	27,8	21,3	21,5	41,7	17,5	18,7
Seq2Seq	48,7	37,8	23,0	16,8	22,7	38,4	19,3	19,0
TAM	57,6	44,8	36,6	29,7	21,2	45,1	20,0	18,1
Multi-stream-TAM	**62,1**	**49,4**	**40,3**	**32,9**	**23,6**	**47,3**	**27,6**	**19,1**

To assess the impact of using different text representation methods on model performance and report quality, we selected the simplest architecture, CNN-LSTM, and tested different text representation approaches, including the use of embeddings that are initialized randomly and learned along training or pre-training models like Word2Vec, fastText to work as initial state. The results of this analysis are listed in Table 5.3.

Table 5.3: Results from CNN-LSTM using different word embeddings.

Approach	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE_L	CIDEr	SPICE
One-Hot encod.	53,5	**39,7**	31,6	25,1	21,4	44,2	15,5	**18,2**
Random init.	54,1	40,6	**32,4**	**25,7**	**21,6**	**44,3**	**22,3**	17,3
Word2vec	**54,6**	39,1	25,6	14,7	17,6	35,8	14,2	17,4
fasttext	49,5	37,1	25,4	15,7	17,8	37,2	14,0	16,8

In appendix C we show the mapping of Word2Vec (C.1a) and fastText (C.1b) embeddings obtained by t-SNE. Table C.1 presents the similarity between the model's embeddings of word pairs with different relationships.

5.3 Architecture Comparison

Since there is no significant improvement in using models to initialize the embedding layer, we proceed with the comparison of different architectures using random initialization. The training

process occurred using batch size of 64 and until stabilization of the loss function. Table 5.4 summarizes the quantitative results obtained for each of the aforementioned architectures, using the LSTM decoder. Note, that as mentioned in Chapter 4, two configurations were tested for the multi-stream architecture.

Table 5.4: Model performance of all captioning approaches with random embedding initialization.

Method	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE_L	CIDEr	SPICE
CNN-LSTM	54,1	40,6	32,4	25,7	21,6	44,3	**22,3**	17,3
CNN-Att-LSTM	50,5	38,9	31,3	25,0	20,7	42,5	16,9	17,1
Seq2Seq	51,2	35,9	26,9	19,9	17,6	40,1	16,4	13,6
TAM	52,3	39,8	32,0	25,7	20,0	44,1	16,5	16,9
Multi-stream-Avg	**56,3**	**43,4**	**35,0**	**28,3**	**23,3**	**46,0**	21,8	18,4
Multi-stream-TAM	54,8	41,4	32,9	26,2	23,2	44,7	21,5	**19,3**

Furthermore, a qualitative assessment was made through the analysis of a set of randomly chosen test samples. Table 5.5 shows one of these examples and the respective reports generated by each of the models. In this, the phenotypes correctly predicted by the models are highlighted (green).

This qualitative analysis also allowed the identification of cases which the models perform poorly. An example of this can be seen in the Table 5.6.

Table 5.5: Generated reports by all the models and respective true clinical report for qualitative assessment.

Model	**Generated report**
CNN+LSTM	this eeg is marked by the occurrence of sharp waves and **generalized slowing**.
CNN+Att+LSTM	this eeg is marked by the occurrence of several **seizures** , in addition to the **focal slowing** in the right posterior quadrant.
Seq2Seq	this eeg is remarkable for **focal** right **hemispheric slowing** , mild to moderate **diffuse slowing** and focal voltage attenuation
TAM	abnormal eeg due to the arising of multiple frequent **seizures** and moderate **diffuse slowing** .
Multi-stream-Avg	this eeg is marked by the **background slowing** , in addition to the occurrence of left temporal periodic lateralized epileptiform discharges and subclinical **focal** occipital **seizures**.
Multi-stream-TAM	this eeg is marked by the occurrence of several **seizures** and sharp waves in the left anterior temporal region , in addition to the severe to the severe **diffuse slowing** .
Ground Truth: abnormal eeg due to the arising of intermittent left occipital focal seizures, in addition to background slowing and focal left hemispheric slowing.	

Table 5.6: Models performing poorly, generating clinical report from a normal EEG.

Model	**Generated report**
CNN+LSTM	this eeg is remarkable for focal delta activity in the right hemisphere and marked background slowing.
CNN+Att+LSTM	this eeg is remarkable for focal slowing in the right temporal region.
Seq2Seq	abnormal eeg because of multiple left frontal electrographic seizures , in addition to focal slowing in the left hemisphere.
TAM	this eeg is marked by the occurrence of seizure , in addition to the focal slowing from the right hemisphere .
Multi-stream-Avg	this eeg is marked by the occurrence of seizure , in addition to continuous focal slowing and moderate diffuse background slowing.
Multi-stream-TAM	this eeg is marked by the occurrence of seizure , in addition to focal slowing in the left anterior temporal region with diffuse background slowing .
Ground Truth: eeg within the normal limits.	

Chapter 6

Discussion

The goal of this dissertation was to study the possibility of automatically generating the impression section of clinical reports from the EEG. Several captioning architectures were explored and implemented. Due to the shortage of large-scale EEG datasets with reports, the use of pre-trained models were also addressed.

In the present chapter, we discuss some limitations and the results obtained for the classification task, text vectorization and EEG captioning, with greater focus on the latter.

6.1 Feature Extrator

As shown in Table 5.1, the VGG16 was reasonably efficient on the classification task. However, the captioning model would benefit from the improvement of the feature extractor. An AUC of 0.80 on the test set corresponds to a model still with a notable difference to perfect fitting. However, it is a satisfactory performance given the scarcity of annotated data and the large difference between the amount of EEG epochs of each class.

Furthermore, Obeid et al. [84] admit that epochs with epileptic events are sometimes accompanied by background alterations. However those are only labeled as SPSW samples, therefore, this convention may have conditioned the training and robustness of the classifier.

Analyzing the results described in Table 5.1, we see that the model is able to generalize well, having an equivalent performance when applied to samples of new patients (test set). Achieving an AUC of 0.81 and 0.80, for train and test sets, respectively. The application of weight classes resulted in a model that is slightly biased towards the minority class (SPSW, label 1), showing great sensitivity (82-88%) at the expense of losing some specificity (75-77%). However, it was shown that a decrease in the weights class resulted in an abrupt decrease in the performance of the classifier, approaching that of a random classifier.

Although the methods and results are not directly comparable, our model has a higher sensitivity compared to others described in the literature for detecting IEDs, for example [79, 112], which have sensitivity of 79% and 80%, respectively. On the other hand, we also obtained lower

specificity and AUC compared to the best result in the literature (specificity of 99% and AUC of 0.96 obtained by Lourenço et al. [79]), therefore we obtained a higher rate of false positives.

6.2 Text Vectorization

The one-hot coding approach proved to be insufficient for most models, resulting in models with poor performance, with overfitting. Despite the high values of NLP metrics shown in Table 5.2, these were possible mainly due to the compromise in the quality of the generated captions.

Note that the metrics that exist for text evaluation are limited and do not correlate perfectly with human judgment. Most of the metrics were not designed for captioning and are not truly appropriate for report generation. Models can achieve high scores with missing information, or by repeating it several times [6]. This effect is particularly evident on BLEU metrics as it is based only on matching. However even in metrics especially designed for image captioning like SPICE this effect can be noticeable. This occurred in all methods while using one-hot encoding particularly on Seq2Seq, TAM and Multi-stream methods that obtained a high score by repeatedly outputting the same sequence along test samples. As mentioned in section 3.1, SPICE metric focus on extracting relations between generated, ground truth captions and attributes, thus facing an extremely unbalanced database with a predominance of "seizure", "shap waves" and "slowing" events a model that always generates a similar sentence referring to these phenotypes is able to achieve high metrics without truly meaning high quality in reliability of the reports.

This explains the discrepancy between quantitative (Table 5.2) and qualitative evaluation while using one-hot encoding.

Thus, the use of word embeddings as a form of text representation was tested. After retraining Word2Vec and fastText methods for the TUH Seizure Corpus, it was verified that both methods are able to efficiently detect word similarities, as seen in the embedding mappings in Figure C.1. As expected, fastText (Figure C.1b) unlike Word2Vec (Figure C.1a) , focuses on morphological similarities beyond semantics relations. Table C.1 shows that both models are able to map words according to their meaning, assigning similar embeddings (similarity close to 1) to words with similar meanings such as the word pair "arising/onset" with a similarity of 0.83 and 0.9 for embeddings from Word2Vec and fastText respectively. And lower similarity for words with very different meanings as in the example "spike/slowing" whose similarity was respectively of 0.12 and 0.24. However, when analyzing morphological relationships, the embeddings from fastText achieve much higher similarity values than Word2Vec. This facilitates the learning of relationships as singular/plural, for instance in the case of the word pair "wave/waves" but it may impair the mapping of other types of relationships, for example the word pair "rhythmic/arrthythmic" which, despite being antonyms, are understood to be very close by the fastText model.

Given the performance in the captioning task (Table 5.3) Word2Vec, compared to fastText, seems to be able to capture more efficiently the relationships between clinical terms. This was also demonstrated by Huang et al. in [50]. However, when comparing with random initialization, the text embeddings learned by Word2Vec and fastText do not seem to be advantageous. The

captioning model with random initialization outperforms the other on most metrics. This may be due to the nature of clinical notes whose words are relatively independent [50].

Note that defining the dimensionality of word embeddings has a big impact on the model's performance. Usually, performance does not improve substantially with increasing dimensionality, however a feature space with a dimensionality too small is limiting and may not be expressive enough. On the other hand, too large dimensionality leads to an increase in model complexity and lead to overfitting.

6.3 Architecture Comparison

6.3.1 Attention Mechanism

In the captioning, there is the potential scenario in which the clinical diagnosis of the report generated is incorrect. So, to avoid this, models have to create a correct understanding of the signal content. Adopting an encoder-decoder approach helps to reduce the risk of this scenario as it uses a multi modal signal+text features representation. However this association/mapping is sometimes complex. Thus the use of attention mechanisms allows to potentially improve the performance of the model allowing the decoder to learn to place attention on certain pieces of information [124].

However, the results obtained in Table 5.4 for the CNN-LSTM and CNN-Att-LSTM methods show that the models behaved similarly, reaching similar values for all metrics. In the qualitative analysis of the generated sentences, it appears that the model that uses the attention mechanism tends to have a slight improvement in the ability to capture EEG phenotypes. For example, in the reports shown in Table 5.5, the CNN-LSTM model wrongly indicates "sharp waves" while CNN-Att-LSTM correctly indicates the occurrence of "seizures". Despite this, both fail to provide a complete description.

In general, the use of visual attention, over the signal features did not result in a significant improvement in the reports. The attention mechanism is not always capable of linking the findings precisely with the corresponding words. This is in line with what was stated by Jing et at [60] which showed that the individual application of attention over the visual features or over semantic features is not sufficient, and the model leads to significant improvement by simultaneously applying attention over semantic and visual features (co-attention).

6.3.2 Type of Features

The crucial step of the captioning task is to capture meaningful features and efficiently summarize all EEG embeddings. This is not trivial given the complexity and wide variability of characteristics and lengths of the EEG recordings. For that, we initially use the average of EEG epoch embedding, forming a global representation of fixed size, allowing us to apply the approaches of image captioning. Despite this approach leading with loss of information, it was found that CNN-LSTM and CNN-att-LSTM were some of the most effective methods for capturing EEG abnormalities.

As reported by Biswal et. al. [13] the performance improvement while using more sophisticated EEG embeddings aggregations such as LSTM or attention mechanism is very limited. In the approach in which we used LSTM (Seq2Seq) and attention mechanism (TAM) in an attempt to track the information and temporal evolution of the signal, the models showed greater difficulty in correctly characterizing the EEG. On Seq2Seq, the performance decrease was more evident possibly given the long duration of EEG signals, which is normally >= 20min (600 epoch). Unlike the RNN, LSTMs can avoid the vanishing gradient problem. However its ability to learn long-range dependencies remains a major challenge, which justifies the poor performance of the Seq2Seq model [46, 113]. Note that as demonstrated in [13], the application of temporal attention (TAM) allowed to attenuate this difficulty allowing the models to focus more on particular intervals. Thus, in Table 5.4 we verify that TAM model outperforms Seq2Seq and achieves a performance comparable to models based on global average EEG embedding (CNN-LSTM and CNN-Att-LSTM).

Despite this, the TAM model, as well as the Seq2Seq model, continue to have a greater bias towards the most common abnormalities and difficulty in diversifying descriptions when compared to the CNN-LSTM and CNN-Att-LSTM models.

This difficulty in identifying rare phenotypes and a bias towards the most common are evident in Table 5.6, which presents the reports generated by the models, based on a normal EEG. It turns out that none of them was able to describe normality of that EEG session. This is justified by the fact that "normal EEGs" are an extremely rare phenotype in the dataset. On the other hand, and especially in the case of normality, signal corruption by noise can make it difficult or lead to misidentification of the EEG phenotype.

Finally we hypothesizes that the combination of both temporal and "visual" features could be supplementary and beneficial for the task. The multi-stream model design (Figures 4.8 and 4.9) outperforms the remaining methods in almost all of the metrics, however, while analyzing the generated report the improvement in diagnostics is not very significant.

This study also shows that there is a compromise between the complexity of the model and the extent and quality of the training dataset. Given the scarcity of data, more complex models tend to underfit the training data. Thus, in more complex models such as TAM, Seq2Seq and Multi-stream-TAM models, the linguist aspect is negatively affected, generating more frequently repeated words or sequences. For example, in the report generated by the Multi-stream-TAM shown in Table 5.5, there is a repetition of the expression "to the severe".

On the other hand, the low dimensionality of the selected embedding space also seems to be limiting the expressiveness of the models, particularly in the multi-stream methods. However an increase in this represents the addition of more parameters and therefore an increased need for a larger training set.

All implemented models fail to successfully indicate the location of anomalies, missing this information, making incomplete information available (e.i. indicating only the side/hemisphere correctly) or even indicating wrong skull regions. This behavior is explained by CNN's low ability to handle/deal with locations and the fact that there is a wide range of possible locations and few samples of each of the possibilities. Furthermore, this information is strongly dependent on the

physician's writing style, which can indicate the EEG channels in which the anomaly is visible or indicate the cranial region, which can be given in more or less detail. For instance, in Table 5.5 the location of the seizures could be indicated as "left occipital", "P3", on the other hand the indication of the location of focal slowing can vary between "left hemisphere", or in more detail referring to all regions included. Furthermore, abnormalities that manifest in a certain region can disturb neighboring regions, making it difficult for the clinician and captioning model to accurately spot the event [58, 106].

6.4 Limitations

The generation of clinical reports is challenging, as the clinical reports that are a source of information are extremely variable. The content, the detail of the description made in the report and the style of writing are very characteristic of each doctor and variable when compared to other experts.

Reports generation can be understood as providing a layer of interpretability, in the sense that captioning models learn to mimic expected judgments. Thus, the subjectivity and intra and interobserver variability of the EEG signal analysis make the task more complicated.

Furthermore, the nature of the EEG signal, which is also extremely variable and often corrupted by noise, makes it difficult to obtain an extractor feature robust enough to ensure the reliability of the generated reports.

In addition, despite the scientific importance of this database, being to the best of our knowledge, the only publicly available with both the EEG and the report, it offers major challenges. Firstly, due to the reduced volume of data and lack on the annotations. Since only a portion of the entire TUH EEG dataset is annotated (corresponding to the TUH EEG Seizure Corpus [84]) and it only has an indication of epileptic and non-epileptic events, instead of annotating all phenotypes included in clinical reports, such as normality, generalized and focal slowing, epileptiform discharges, seizures, abnormal delta. Furthermore, this database is characterized by considerable class imbalance, with a clear predominance of abnormal EEGs and "seizures" and "background slowing" phenotypes and rare examples of EEG recordings reported as normal.

Chapter 7

Conclusions and Future work

We show that it is possible to generate clinical reports from the EEG signal. However, the results also reveal that EEG captioning models still have some limitations and should be further explored and improved so that they can be applied in the clinic.

Feature extraction and summarization in global embeddings are one of the major challenges of the task. Multi-stream architecture, in particular Multi-stream-Avg, seems to be promising. However, it should be more widely tested and studied. For example, being tested in other datasets for captioning and/or in other tasks that allow proving the effectiveness of the extraction and aggregation of EEG embeddings by these models, such as multi-label classification. Unlike captioning, classification allows us to directly assess the efficiency of feature extraction, allowing clearer conclusions to be drawn. Both multi-stream structures would benefit from pre-training the streams to capture specific features either dynamic or static, semantic or visual. The pre-training of the network on semantic streams would be possible using a database with an explicit indication of the EEG phenotypes. In addition, the use of 3D-CNN instead of applying 2D-CNN over time could potentially improve the effectiveness of dynamic feature extraction.

Multi-stream-Avg and CNN-LSTM were the models that performed better and were generally able to capture EEG phenotypes and describe them using natural language. However, these are restricted regarding the diversity of phenotypes that they can capture and the variability of the reports they generate, which is justified by the aforementioned annotation limitation (section 6.4) and small volume of the training dataset.

Thus, the models should be trained in the future with a more comprehensive database, which considers more EEG abnormalities, to obtain models capable of dealing with the diversity of clinical conditions and thus, be incorporated into clinical practice for the generation of preliminary clinical reports.

Furthermore, training with a substantially larger volume of data would allow more effective learning of the models, allowing the encoder to generate meaningful EEG embeddings and the decoder to describe them accurately through a natural and diversified language.

For incorporation into clinical practice, it would be important to evaluate a set of generated reports by a group of experts for clinical validation. In the current scenario, and especially in healthcare, it is important to build a model that is understandable to humans to increase user acceptance and so that the neurologist can take a critical position in relation to the decisions of the model.

Explaining what triggered a model's decision is challenging, and there has been a growing interest in this area, particularly in the application of DL models. Explainable Artificial Intelligence (XAI) is a research field in machine learning that focuses on the issue of explainability and transparency of models. There are already a wide variety of methods that allow obtaining tips, visual explanations of different process stages (Pre-Model, In-Model, Post-Model), or even textual explanations.

In models that use attention mechanisms, such as CNN-Att-LSTM, TAM and Multi-stream-TAM, the attention weights and their visualization can be used for model interpretability, since it allows us to know where or in which EEG intervals the model is valuing more for outputting each part of the report. These XAI techniques add a degree of interpretability to the model, on the one hand increasing the acceptability of the model, allowing a faster EEG review by detecting the relevant intervals, and on the other hand facilitating the detection of abnormal or unwanted behaviour of the model.

Despite the challenges, in this work we developed models that could accurately describe EEG recording, however, the test of the different models also revealed the compromise between the amount of data and the complexity of the models. Complex models either rely more on the language information rather than focusing on the EEG content and getting more bias towards common phenotypes, or deterioration in the quality of the text they generate.

Regarding the quality of the report language, there is still space for improvement. We have shown that the context-free models (Word2Vec and fastText) are not the most suitable for clinical text. However, since most implemented captioning models are based on multimodal signal+text representation, it is admitted that there is also a compromise between the complexity of EEG features and text. So, in the future, it would be interesting to test the tuning of Word2Vec, fastText and test other available pre-trained models, as context-sensitive models as ELMo or BERT, in the chosen captioning model in order to assess which representation technique fits better the degree of complexity of the problem at hand and the capacity of the model.

In the future, it would also be relevant to explore the benefit of using different structures, particularly the replacement of RNN by a transformer-based architecture. The use of transformers has been proven to be advantageous in a state-of-the-art approach, as these structures are able to generalize well to sequential data. Transformer-based models are computationally heavy, so it was not possible to explore this research path further due to limited computational resources. However an interesting approach would be to use a decoder consisting of multiple stack transform layers as proposed by Kiyasseh et. al. [135] for ECG captioning, in which the CNN encoder process the EEG and outputs signal embeddings with transformer-compatible dimensions. These transformer-based models contain a huge amount of parameters so they need a lot of examples to

train. In case of a small dataset, as TUH EEG Seizure Corpus, an alternative that would be advantageous is to adapt publicly available pre-trained models for EEG captioning, such as BERT and, in particular, its already trained version for clinical or biomedical text, for example, ClinicalBERT or BioBERT [2, 50, 71].

Another issue was the generation of incomplete reports, which did not describe all EEG events. To tackle this, one possible approach would be to adopt a dense captioning model. Those region-based models have been proven to bring out a more complete understanding, producing more objective and detailed than global description. Furthermore, as these methods are region-based, they could reduce the difficulty that models have to correctly describe the event region [48, 62].

A potential research path is to explore the advantage of generative adversarial neural networks (GAN) for signal captioning. This strategy presents many challenges, as the discrete nature of the data requires the adoption of methods such as the Gumbel sampler to enable back-propagation. Also, the GANs' training process is very unstable, the Nash equilibrium is difficult to achieve, and the model may not converge or end up in model collapse. However, GANs may potentially be able to generate diverse and high-quality captions [48].

There are many open questions and much to improve in this challenging task of finding the best model for EEG signal reporting. The main limitations are related to the dataset, but despite that, it was proved that it is possible to generate clinical reports from the EEG. Unfortunately, it was not possible to improve and arrive at a system that was accurate and comprehensive enough to be incorporated into the clinic. Thus, the results and conclusions of this work should be considered and further explored in future projects.

Appendix A

State-of-the-art Summary Tables - Chapter 3

Table A.1: Summary of the studies presented on NLP for Clinical Text

Ref	Author	Year	Task	Approach
[11]	Biswal	2016	Document classification of EEG reports	"Elastic word sequence" and Naive Bayes
[81]	Maldonado	2020	Annotation of corpus's EEG reports	bi-LSTM; DRLN and Self-attention
[50]	Huang	2019	Prediction of 30-day patient readmission	ClinicalBERT (Transformer)

Table A.2: Summary of the studies that use encoder-decoder approach for Image captioning

Deep Neural Networks approach						
Ref	Author	Year	Additional Processing	Encoder	Decoder	Attention Mechanism
[65]	Kiros	2014	No	OxfordNet	LSTM	No
[103]	Shin	2015	Image classification and detection	GoogleNet	LSTM, GRU	No
[90]	Pelka	2017	No	Inception V3	LSTM	No
[132]	Zeng	2020	Region detection, Classification	VGG-16	LSTM	No
[123]	Xu	2019	Concepts detection	ResNet-101	LSTM	Visual attention
[129]	You	2016	Classification / Word Detection	GoogleNet	LSTM	Semantic attention
[60]	Jing	2018	Multi-label Classification to predict tags	VGG-19	Hierarchical LSTM (Sentence, Word)	Co-attention (visual + semantic)

Table A.3: Summary of the studies that use Hybrid approach for Image captioning

Hybrid between DL and retrieval approach				
Ref	Author	Year	Feature Extraction	Description
[13]	Biswal	2020	CNN	Retrieval template sentence based on extracted phenotypes and physician's inputs (prefix text or anchor words) and sentence edition using sequence to sequence model.
[77]	Liang	2017	CNN	RNN-LSTM to generate a caption. Decide between the predicted and a retrieve caption using a Euclidean distance threshold (between extracted and retrieved features)
[73]	Li	2018	CNN	Reinforcement learning to either generate a text report or retrieve a report from a template database.

Table A.4: Summary of the studies presented on NLP for Biosignals

						Report Generation	
Ref	Author	Year	Biosignal	Additional Processing	Feature extraction	Approach	Description
[14]	Biswal	2019	EEG	Extracted phenotypes	CNN	Hybrid (DL and Template-based)	Fill a Template report based on extracted phenotypes (to generate impression section). Use hierarchical LSTM (paragraph, sentence) and attention to generate detailed explanations in the impression section)
[135]	Kiyasseh	2021	ECG	Signal classification, Language token prediction	CNN	DL-based	Decoder composed by transformer layers to generate multilingual reports
[13]	Biswal	2020	EEG	Extracted phenotypes	CNN	Hybrid (DL and retrieval-based)	Retrieval template sentence based on extracted phenotypes and physician's inputs (prefix text or anchor words) and sentence edit using sequence to sequence model.

Appendix B

Supplements to Methods - Chapter 4

VGG16					
Layers	Parameters	output shape	Layers	Parameters	output shape
First Block			Fourth Block		
Input		250,18,1	2D Zeropadding	padding (1,1)	34,5,256
2D Zeropadding	padding (1,1)	252,20,1	2D convolution	512 units kernel (3,3) activation "relu"	32,3,512
2D convolution	64 units kernel (3,3) activation "relu"	250,18,64	2D Zeropadding	padding (1,1)	34,5,512
2D Zeropadding	padding (1,1)	252,20,64	2D convolution	512 units kernel (3,3) activation "relu"	32,3,512
2D convolution	64 units kernel (3,3) activation "relu"	250,18,64	2D Zeropadding	padding (1,1)	34,5,512
2D max pooling	pooling (2,2) stride (2,2)	125,9,64	2D convolution	512 units kernel (1,1) activation "relu"	34,5,512
Second Block			2D max pooling	pooling (2,2) stride (2,2)	17,2,512
2D Zeropadding	padding (1,1)	127,11,64	Fifth Block		
2D convolution	128 units kernel (3,3) activation "relu"	125,9,128	2D Zeropadding	padding (1,1)	19,4,512
2D Zeropadding	padding (1,1)	127,11,128	2D convolution	512 units kernel (3,3) activation "relu"	17,2,512
2D convolution	128 units kernel (3,3) activation "relu"	125,9,128	2D Zeropadding	padding (1,1)	19,4,512
2D max pooling	pooling (2,2) stride (2,2)	62,4,128	2D convolution	512 units kernel (3,3) activation "relu"	17,2,512
Third Block			2D Zeropadding	padding (1,1)	19,4,512
2D Zeropadding	padding (1,1)	64,6,128	2D convolution	512 units kernel (1,1) activation "relu"	19,4,512
2D convolution	256 units kernel (3,3) activation "relu"	62,2,256	2D max pooling	pooling (2,2) stride (2,2)	9,2,512
2D Zeropadding	padding (1,1)	64,6,256	Sixth Block		
2D convolution	256 units kernel (3,3) activation "relu"	62,2,256	Flatten		9216
2D Zeropadding	padding (1,1)	64,6,256	Dense	4096 units activation 'relu'	4096
2D convolution	256 units kernel (1,1) activation "relu"	64,6,256	Droupout	0,5	4096
2D max pooling	pooling (2,2) stride (2,2)	32,3,256	Dense	4096 units activation 'relu'	4096
			Droupout	0,5	4096
			Dense	2 units activation 'softmax'	2
			Output		2

Figure B.1: VGG16 architecture

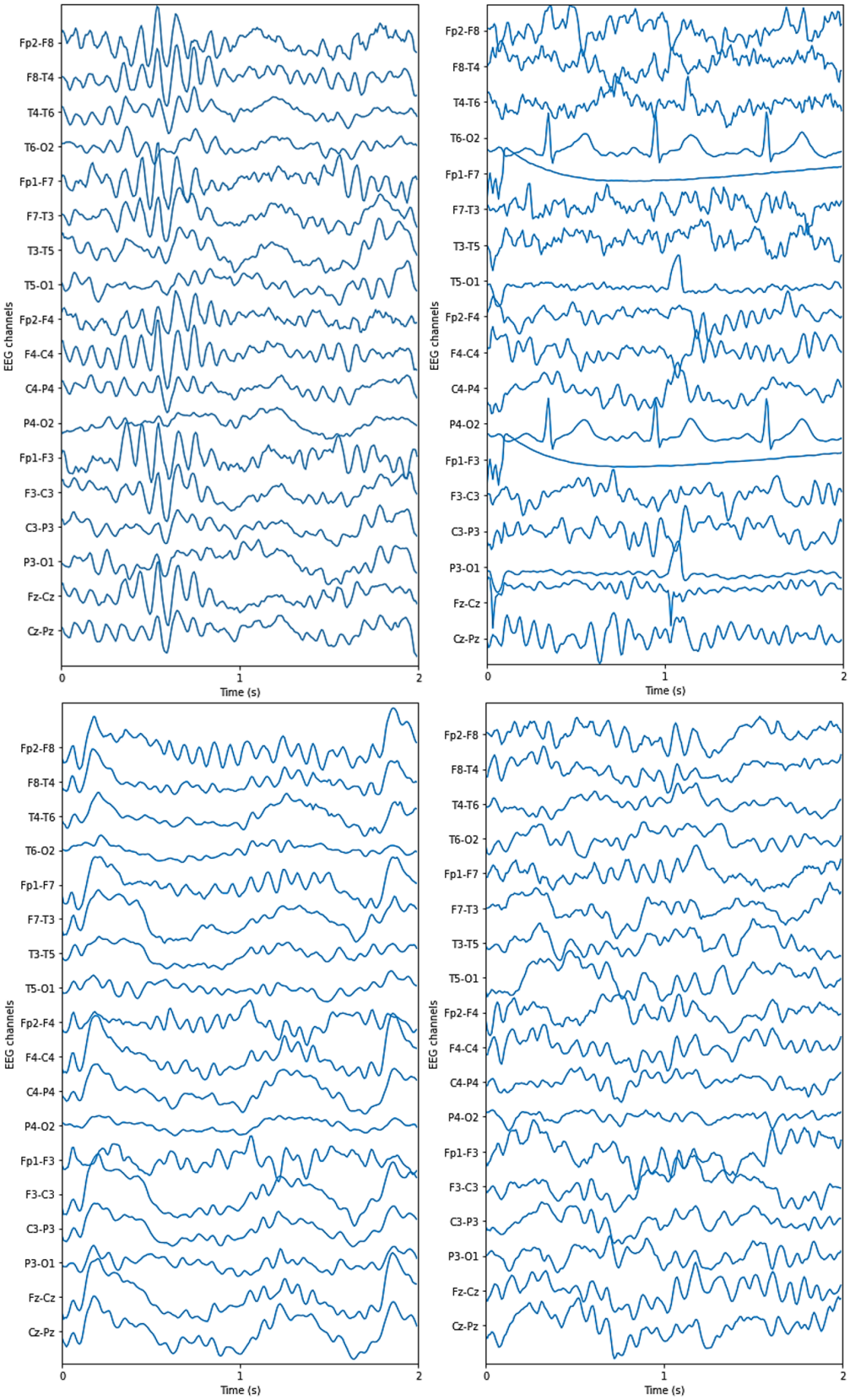


Figure B.2: Visualization of EEG epochs with spikes and/or sharp waves (SPSW).

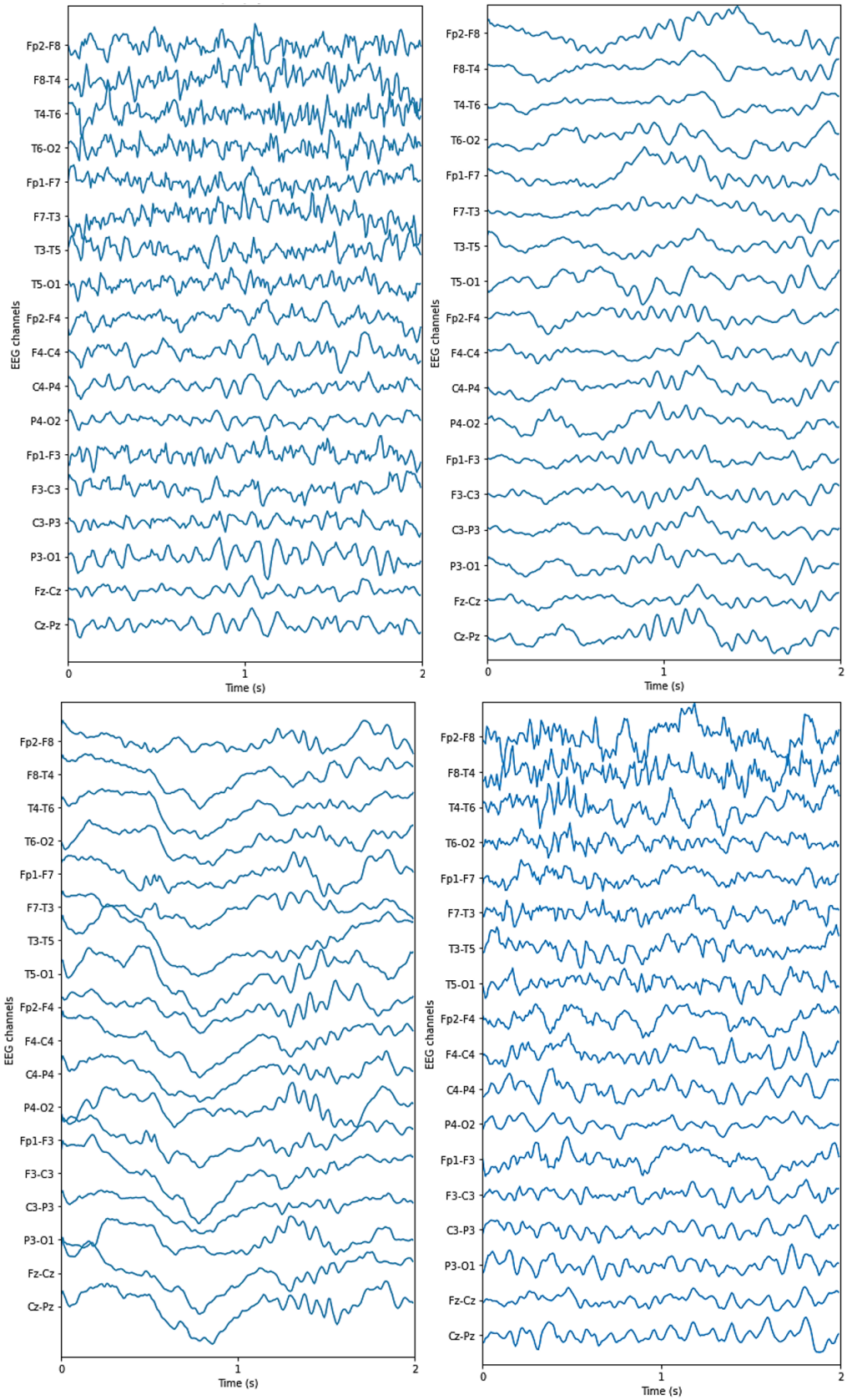


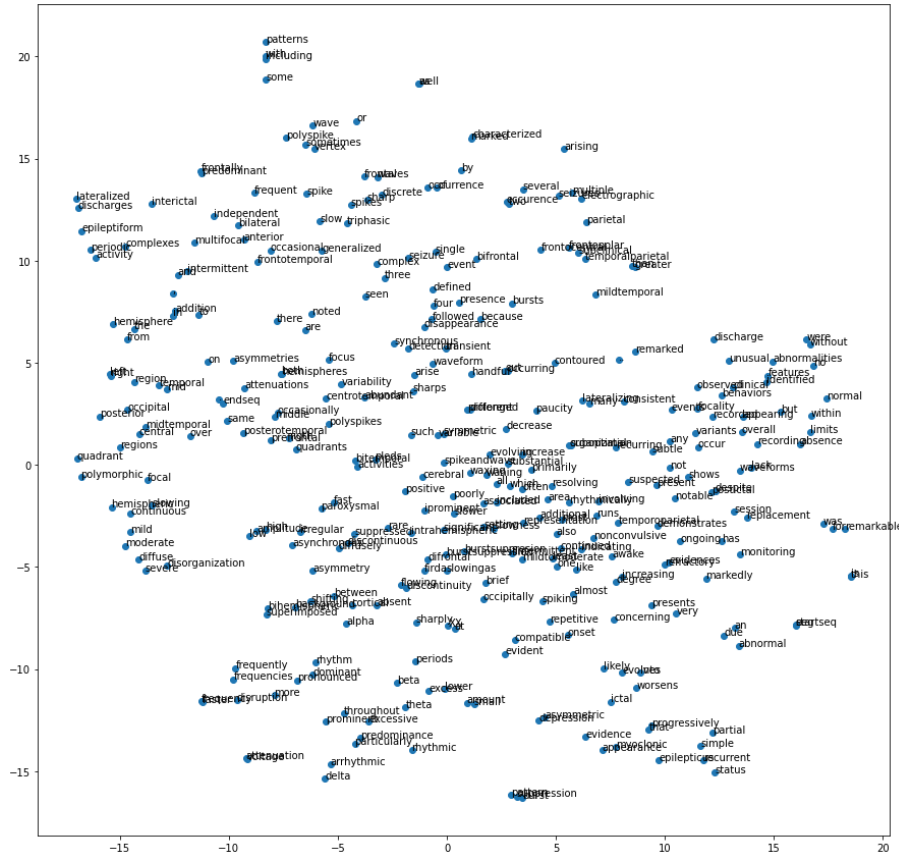
Figure B.3: Visualization of EEG epochs with abnormal background activity (BCKG).

Appendix C

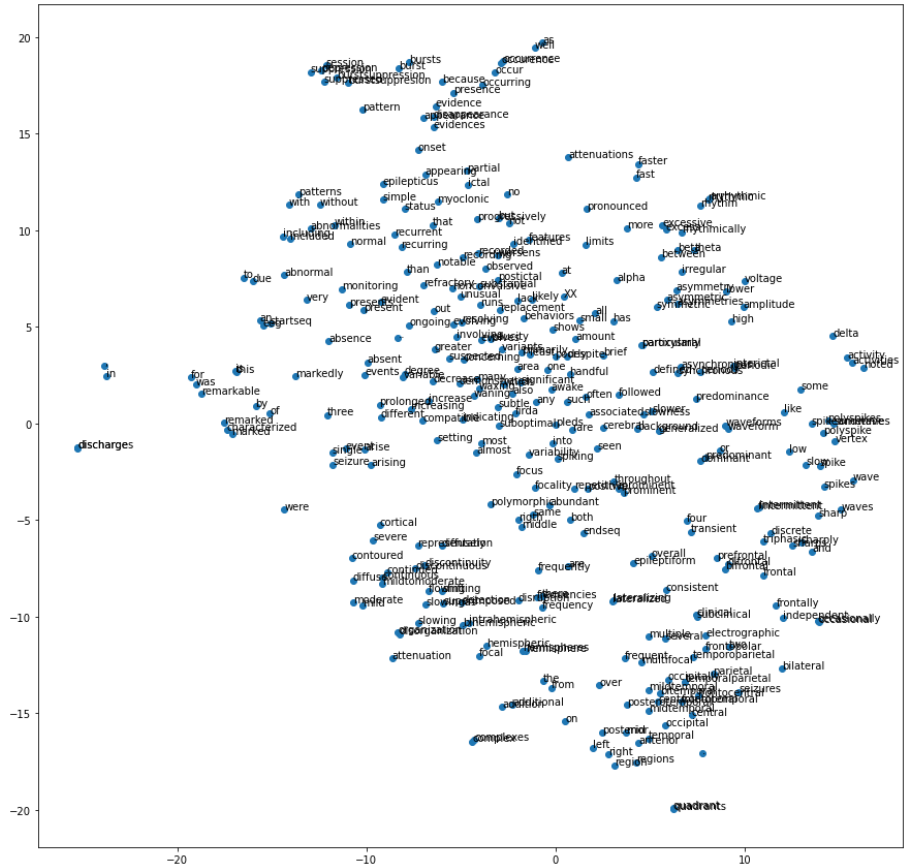
Supplements to Results - Chapter 5

Table C.1: Word embedding Similarity

Word pairs		**Methods**	
similar meaning		*Word2Vec*	*fastText*
'mid'	'moderate'	**0,6**	0,56
'without'	'no'	**0,93**	0,76
'arising'	'onset'	0,83	**0,9**
diferent meanings			
'focal '	'generalize'	0,35	**0,16**
'left'	'rigth'	**0,55**	0,77
'spike'	'slowing'	**0,12**	0,24
similar morfolgy			
'frontal'	'frontotemporal'	0.48	**0.86**
'wave'	'waves'	0.74	**0.83**
'activity'	'activities'	0.51	**0.92**
'rhythmic'	'arrhythmic'	**0.47**	0.94



(a) Word2Vec



(b) fastText

Figure C.1: Word embedding mapping of context-free models

References

[1] Asma Ben Abacha, AG Seco De Herrera, Soumya Gayen, Dina Demner-Fushman, and Sameer Antani. Nlm at imageclef 2017 caption task, 2017.

[2] Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*, 2019.

[3] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *European conference on computer vision*, pages 382–398. Springer, 2016.

[4] Locke Saskia; Bashall Anthony; Al-Adely Sarah; Moore John; Wilson Anthony and Gareth B Kitchen. Natural language processing in medicine: A review. *Trends in Anaesthesia and Critical Care*, pages 38:4–9, 2021.

[5] Shaun S. Lodder; Jessica Askamp and Michel J.A.M. van Putten. Computer-assisted interpretation of the eeg background pattern: A clinical evaluation. *PLoS ONE*, pages 9(1):1–8, Jan 2014.

[6] Hareem Ayesha, Sajid Iqbal, Mehreen Tariq, Muhammad Abrar, Muhammad Sanaullah, Ishaq Abbas, Amjad Rehman, Muhammad Farooq Khan Niazi, and Shafiq Hussain. Automatic medical image interpretation: State of the art and future directions. *Pattern Recognition*, 114:107856, 2021.

[7] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate, 2016.

[8] Michel Baulac, Hanneke De Boer, Christian Elger, Mike Glynn, Reetta Kälviäinen, Ann Little, Janet Mifsud, Emilio Perucca, Asla Pitkänen, and Philippe Ryvlin. Epilepsy priorities in europe: A report of the ilae-ibe epilepsy advocacy europe task force. *Epilepsia*, 56(11):1687–1695, 2015.

[9] Ettore Beghi. The epidemiology of epilepsy. *Neuroepidemiology*, pages 54(2):185–191, December 2019.

[10] Anne T Berg. Risk of recurrence after a first unprovoked seizure. *Epilepsia*, 49:13–18, 2008.

[11] Siddharth Biswal; Zarina Nip; Valdery Moura Junior; Matt T. Bianchi and Eric S Rosenthal. Automated information extraction from free-text eeg reports, 2016.

[12] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

[13] Siddharth Biswal, Cao Xiao, Lucas M. Glass, Brandon Westover, and Jimeng Sun. Clara: Clinical report auto-completion. In *Proceedings of The Web Conference 2020*, page 541–550, New York, NY, USA, 2020. Association for Computing Machinery.

[14] Siddharth Biswal, Cao Xiao, M. Brandon Westover, and Jimeng Sun. Eegtotext: Learning to write medical reports from eeg recordings. In *Proceedings of the 4th Machine Learning for Healthcare Conference*, volume 106 of *Proceedings of Machine Learning Research*, pages 513–531. PMLR, 09–10 Aug 2019.

[15] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

[16] Licong Cui; Satya S. Sahoo; Samden D. Lhatoo; Gaurav Garg; Prashant Rai; Alireza Bozorgi and Guo-Qiang Zhang. Complex epilepsy phenotype extraction from narrative clinical discharge summaries. *Journal of Biomedical Informatics*, pages 272–279, 2014.

[17] Jason Brownlee. Deep learning for natural language processing. [Online], 2017.

[18] Jason Brownlee. Generative adversarial networks with python. *Mach. Learn. Mastery*, pages 1–654, 2019.

[19] 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, Dec 2020.

[20] Fei Cai and Maarten de Rijke. A survey of query auto completion in information retrieval. *Found. Trends Inf. Retr.*, 10(4):273–363, sep 2016.

[21] Chen Chen, Shuai Mu, Wanpeng Xiao, Zexiong Ye, Liesi Wu, and Qi Ju. Improving image captioning with conditional generative adversarial nets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8142–8150, 2019.

[22] Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. n the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, Doha, Qatar, October 2014. Association for Computational Linguistics.

[23] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, October 2014. Association for Computational Linguistics.

[24] FRANÇOIS CHOLLET. *Deep Learning with Python*. Manning Publications, 2018.

[25] Noam Chomsky. *Syntatic structures*. The Hague: Mouton & Company, 1957.

[26] FA Chowdhury, L Nashef, and RDC Elwes. Misdiagnosis in epilepsy: a review and recognition of diagnostic uncertainty. *European journal of neurology*, 15(10):1034–1042, 2008.

[27] J Helen Cross. Epilepsy in the who european region: fostering epilepsy care in europe. *Epilepsia*, 52(1):187–188, 2011.

[28] Fernando Lopes Da Silva, Wouter Blanes, Stiliyan N Kalitzin, Jaime Parra, Piotr Suffczynski, and Demetrios N Velis. Epilepsies as dynamical diseases of brain systems: basic models of the transition between normal and epileptic activity. *Epilepsia*, 44:72–83, 2003.

[29] Catarina da Silva Lourenço, Marleen C. Tjepkema-Cloostermans, and Michel J.A.M. van Putten. Machine learning for detection of interictal epileptiform discharges. *Clinical Neurophysiology*, 132(7):1433–1443, 2021.

[30] Bo Dai, Sanja Fidler, Raquel Urtasun, and Dahua Lin. Towards diverse and natural image descriptions via a conditional gan. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2989–2998, 2017.

[31] Eleftherios Daskalakis, Maria Tzelepi, and Anastasios Tefas. Learning deep spatiotemporal features for video captioning. *Pattern Recognition Letters*, 116:143–149, 2018.

[32] Tomas Mikolov; Kai Chen; Greg Corrado; Jeffrey Dean. Efficient estimation of word representations in vector space, 2013.

[33] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *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.

[34] Carsten Eickhoff, Immanuel Schwall, Alba Garcia Seco De Herrera, and Henning Müller. Overview of imageclefcaption 2017–image caption prediction and concept detection for biomedical images, 2017.

[35] Ahmed Elhagry and Karima Kadaoui. A thorough review on recent deep learning methodologies for image captioning. *arXiv preprint arXiv:2107.13114*, 2021.

[36] Athanasios Covanis Emilio Perucca and Tarun Dua. Eeg in the diagnosis, classification, and management of patients. *Journal of Neurology, Neurosurgery & Psychiatry*, page 76(2):ii2–ii7, 2005.

[37] Athanasios Covanis Emilio Perucca and Tarun Dua. Commentary: Epilepsy is a global problem. *Epilepsy*, page 55(9):1326–1328, August 2014.

[38] Hugging Face. Preprocess. https://huggingface.co/docs/transformers/preprocessing, 2020.

[39] Hugging Face. Summary of the tokenizers. https://huggingface.co/docs/transformers/tokenizer_summary, 2020.

[40] Robert S Fisher, Carlos Acevedo, Alexis Arzimanoglou, Alicia Bogacz, J Helen Cross, Christian E Elger, Jerome Engel Jr, Lars Forsgren, Jacqueline A French, Mike Glynn, et al. Ilae official report: a practical clinical definition of epilepsy. *Epilepsia*, 55(4):475–482, 2014.

[41] Robert S Fisher, Walter Van Emde Boas, Warren Blume, Christian Elger, Pierre Genton, Phillip Lee, and Jerome Engel Jr. Epileptic seizures and epilepsy: definitions proposed by the international league against epilepsy (ilae) and the international bureau for epilepsy (ibe). *Epilepsia*, 46(4):470–472, 2005.

[42] Alba Garcia Seco De Herrera, Carstern Eickhof, Vincent Andrearczyk, and Henning Müller. Overview of the imageclef 2018 caption prediction tasks, 2018.

[43] 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.

[44] Google. The wordpiece algorithm in open source bert., 2018.

[45] 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.

[46] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 11 1997.

[47] J J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558, 1982.

[48] Md. Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CSUR)*, 51:1 – 36, 2019.

[49] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*, 2018.

[50] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *CoRR*, abs/1904.05342, 2019.

[51] Vladimir Iashin and Esa Rahtu. Multi-modal dense video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 958–959, 2020.

[52] IBM. Natural language processing (nlp). https://www.ibm.com/cloud/learn/natural-language-processing, July 2020.

[53] Touseef Iqbal and Shaima Qureshi. The survey: Text generation models in deep learning. *Journal of King Saud University-Computer and Information Sciences*, 2020.

[54] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):590–597, Jul. 2019.

[55] Saiful Islam, Aurpan Dash, Ashek Seum, Amir Hossain Raj, Tonmoy Hossain, and Faisal Muhammad Shah. Exploring video captioning techniques: A comprehensive survey on deep learning methods. *SN Computer Science*, 2(2):1–28, 2021.

[56] Aaron Jaech and Mari Ostendorf. Personalized language model for query auto-completion, 2018.

[57] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

[58] George H. Klem; Hans Otto LuÈders; H.H. Jasper and C. Elger. The ten±twenty electrode system of the international federation. *Electroencephalography and clinical neurophysiology*, pages 52:3–6, July 1999.

[59] MZ Koubeissi; WE Lievens; EM Pestana-Knight; Jeffrey W Britton; Lauren C Frey; JL Hopp; P Korb and EK Louis S. *Electroencephalography (EEG): An introductory text and atlas of normal and abnormal findings in adults, children, and infants*. American Epilepsy Society, Chicago, 2016.

[60] Baoyu Jing, Pengtao Xie, and Eric Xing. On the automatic generation of medical imaging reports. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2577–2586, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[61] Alistair E.W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R.Greenbaum, Matthew P. Lungren, Chih ying Deng, RogerG. Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Sci Data*, page 6:317, 2019.

[62] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[63] Theo Vos Joshua A Salomon and Hogan DR et al. Common values in assessing health outcomes from disease and injury: disability weights measurement study for the global burden of disease study. *Lancet*, Dec 2012.

[64] Sethunya Joseph; Hlomani Hlomani; Keletso Letsholo; Freeson Kaniwa and Kutlwano Sedimo. Natural language processing: A review. *International Journal of Research in Engineering and Applied Sciences*, pages 6(3):2249–3905, 2016.

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

[66] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90, may 2017.

[67] Alex M Lamb, Anirudh Goyal ALIAS PARTH GOYAL, Ying Zhang, Saizheng Zhang, Aaron C Courville, and Yoshua Bengio. Professor forcing: A new algorithm for training recurrent networks. *Advances in neural information processing systems*, 29, 2016.

[68] H Lane, C Howard, and H Hapke. *Natural Language Processing in Action: Understanding, analyzing, and generating text with Python. Manning Publications Co*. 2019.

[69] Alon Lavie and Abhaya Agarwal. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[70] 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.

[71] 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, 2020.

[72] Sujin Lee and Incheol Kim. Video captioning with visual and semantic features. *Journal of Information Processing Systems*, 14(6):1318–1330, 2018.

[73] Christy Y. Li, Xiaodan Liang, Zhiting Hu, and Eric P. Xing. Hybrid retrieval-generation reinforced agent for medical image report generation, 2018.

[74] Christy Y. Li, Xiaodan Liang, Zhiting Hu, and Eric P. Xing. Knowledge-driven encode, retrieve, paraphrase for medical image report generation, 2019.

[75] Jiwei Li, Will Monroe, and Dan Jurafsky. A simple, fast diverse decoding algorithm for neural generation. *arXiv preprint arXiv:1611.08562*, 2016.

[76] Nannan Li, Zhenzhong Chen, and Shan Liu. Meta learning for image captioning. In *AAAI*, 2019.

[77] Sisi Liang, Xiangyang Li, Yongqing Zhu, Xue Li, and Shuqiang Jiang. Isia at the imageclef 2017 image caption task. In *CLEF (Working Notes)*, 2017.

[78] Shaun S. Lodder and Michel J.A.M. van Putten. Quantification of the adult eeg background pattern. *Clinical Neurophysiology*, pages 124(2):228–237, 2013.

[79] Catarina Lourenço, Marleen C Tjepkema-Cloostermans, Luís F Teixeira, and Michel JAM van Putten. Deep learning for interictal epileptiform discharge detection from scalp eeg recordings. In *mediterranean conference on medical and biological engineering and computing*, pages 1984–1997. Springer, 2019.

[80] BT Lowerre. The harpy speech recognition system[ph. d. thesis]. 1976.

[81] Ramon Maldonado and Sanda Harabagiu. The language of brain signals: Natural language processing of electroencephalography reports. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2268–2275, Marseille, France, May 2020. European Language Resources Association.

[82] J. W. C. Medithe and U. R. Nelakuditi. Study of normal and abnormal eeg. *in 2016 3rd International Conference on Advanced Computing and Communication Systems*, page 1–4, Jan 2016.

[83] Soheyl Noachtar and Jan Rémi. The role of eeg in epilepsy: a critical review. *Epilepsy & Behavior*, 15(1):22–33, 2009.

[84] Iyad Obeid and Joseph Picone. The temple university hospital eeg data corpus. *Frontiers in Neuroscience*, 10, 2016.

[85] Prakash M Nadkarni; Lucila Ohno-Machado and Wendy W Chapman. Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, pages 18(5):544–551, 2011.

[86] World Health Organization et al. *Epilepsy: a public health imperative*. World Health Organization, 2019.

[87] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[88] Reshamwala Alpa; Mishra Dhirendra; Pawar and Prajakta. Review on natural language processing. *Engineering Science and Technology: An International Journal*, pages 3(1):2250–3498, 2013.

[89] Lin ZHU; Haifeng CHEN; Xu ZHANG; Kai GUO; Shujing WANG; Yu WANG; Weihua PEI and Hongda CHEN. Design of portable multi-channel eeg signal acquisition system. *In Biomedical Engineering and Informatics, 2nd International Conference on*, pages 1–4, 2009.

[90] Obioma Pelka and Christoph M Friedrich. Keyword generation for biomedical image retrieval with recurrent neural networks. In *CLEF (Working Notes)*, 2017.

[91] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics.

[92] ME Peters, M Neumann, M Iyyer, M Gardner, C Clark, K Lee, and L Zettlemoyer. Deep contextualized word representations. arxiv 2018. *arXiv preprint arXiv:1802.05365*, 12, 1802.

[93] Jyoti Pillai and Michael R Sperling. Interictal eeg and the diagnosis of epilepsy. *Epilepsia*, 47:14–22, 2006.

[94] Jyoti Pillai and Michael R. Sperling. Interictal eeg and the diagnosis of epilepsy. *Epilepsia*, 47(s1):14–22, 2006.

[95] NLTK Project. Natural language toolkit, 2019.

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

[97] Muhammad Sajjad, Zulfiqar Ahmad Khan, Amin Ullah, Tanveer Hussain, Waseem Ullah, Mi Young Lee, and Sung Wook Baik. A novel cnn-gru-based hybrid approach for short-term residential load forecasting. *Ieee Access*, 8:143759–143768, 2020.

[98] salaniz/pycocoevalcap:. Python 3 support for the ms coco caption evaluation tools. https://github.com/salaniz/pycocoevalcap.git, 2020.

[99] W.O.Tatum; G.Rubboli; P.W.Kaplan; S.M.Mirsatari; K.Radhakrishnan; D.Gloss; L.O.Caboclo; F.W.Drislane; M.Koutroumanidis; D.L.Schomer; D.Kasteleijn-Nolst Trenite; Mark Cook; S.Beniczky. Clinical utility of eeg in diagnosing and monitoring epilepsy in adults. *Clinical Neurophysiology*, pages 129(5):1056–1082, May 2018.

[100] Ingrid E Scheffer, Samuel Berkovic, Giuseppe Capovilla, Mary B Connolly, Jacqueline French, Laura Guilhoto, Edouard Hirsch, Satish Jain, Gary W Mathern, Solomon L Moshé, et al. Ilae classification of the epilepsies: position paper of the ilae commission for classification and terminology. *Epilepsia*, 58(4):512–521, 2017.

[101] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, August 2016. Association for Computational Linguistics.

[102] Rakshith Shetty, Marcus Rohrbach, Lisa Anne Hendricks, Mario Fritz, and Bernt Schiele. Speaking the same language: Matching machine to human captions by adversarial training. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4135–4144, 2017.

[103] Hoo-Chang Shin, Le Lu, Lauren Kim, Ari Seff, Jianhua Yao, and Ronald M Summers. Interleaved text/image deep mining on a very large-scale radiology database. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1090–1099, 2015.

[104] Yuqi Si, Jingqi Wang, Hua Xu, and Kirk Roberts. Enhancing clinical concept extraction with contextual embeddings. *Journal of the American Medical Informatics Association*, 26(11):1297–1304, 07 2019.

[105] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[106] S J M Smith. Eeg in the diagnosis,classification, and management of patients with epilepsy. *Journal of Neurology, Neurosurgery, and Psychiatry*, page 2(2):ii2–ii7, July 2005.

[107] Xinying Song, Alex Salcianu, Yang Song, Dave Dopson, and Denny Zhou. Fast wordpiece tokenization. *arXiv preprint arXiv:2012.15524*, 2020.

[108] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *NIPS*, 2014.

[109] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015.

[110] M. Teplan. Fundamentals of eeg measurement. *MEASUREMENT SCIENCE REVIEW*, pages 2(2):1–11, 2002.

[111] Roland D Thijs, Rainer Surges, Terence J O'Brien, and Josemir W Sander. Epilepsy in adults. *The Lancet*, 393(10172):689–701, 2019.

[112] John Thomas, Jing Jin, Prasanth Thangavel, Elham Bagheri, Rajamanickam Yuvaraj, Justin Dauwels, Rahul Rathakrishnan, Jonathan J Halford, Sydney S Cash, and Brandon Westover. Automated detection of interictal epileptiform discharges from scalp electroencephalograms by convolutional neural networks. *International journal of neural systems*, 30(11):2050030, 2020.

[113] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[114] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.

[115] JosEPh VEIZENBA. Eliza a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, pages 9(1):36–45, 1966.

[116] Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. Sequence to sequence-video to text. In *Proceedings of the IEEE international conference on computer vision*, pages 4534–4542, 2015.

[117] Subhashini Venugopalan, Huijuan Xu, Jeff Donahue, Marcus Rohrbach, Raymond J. Mooney, and Kate Saenko. Translating videos to natural language using deep recurrent neural networks. *CoRR*, abs/1412.4729, 2014.

[118] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. *CoRR*, abs/1411.4555, 2014.

[119] Hao Wang, Longyin Wen, Libo Zhang, and Tiejian Luo. Collaborative three-stream transformers for video captioning. 2021.

[120] Kunfeng Wang, Chao Gou, Yanjie Duan, Yilun Lin, Xinhu Zheng, and Fei-Yue Wang. Generative adversarial networks: introduction and outlook. *IEEE/CAA Journal of Automatica Sinica*, 4(4):588–598, 2017.

[121] 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)*, pages 3462–3471, 2017.

[122] Zuxuan Wu, Ting Yao, Yanwei Fu, and Yu-Gang Jiang. Deep learning for video classification and captioning, 2017.

[123] Jing Xu, Wei Liu, Chao Liu, Yu Wang, Ying Chi, Xuansong Xie, and Xiansheng Hua. Concept detection based on multi-label classification and image captioning approach - damo at imageclef 2019. In *CLEF*, 2019.

[124] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. 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 2048–2057, Lille, France, 07–09 Jul 2015. PMLR.

[125] Ning Xu, An-An Liu, Yongkang Wong, Yongdong Zhang, Weizhi Nie, Yuting Su, and Mohan Kankanhalli. Dual-stream recurrent neural network for video captioning. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(8):2482–2493, 2018.

[126] Chenggang Yan, Yunbin Tu, Xingzheng Wang, Yongbing Zhang, Xinhong Hao, Yongdong Zhang, and Qionghai Dai. Stat: Spatial-temporal attention mechanism for video captioning. *IEEE transactions on multimedia*, 22(1):229–241, 2019.

[127] Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, and Aaron Courville. Describing videos by exploiting temporal structure, February 2015.

[128] Chin yew Lin. Rouge: a package for automatic evaluation of summaries, 2004.

[129] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4651–4659, 2016.

[130] Amir Zaidi, Peter Clough, Paul Cooper, Bruce Scheepers, and Adam P Fitzpatrick. Misdiagnosis of epilepsy: many seizure-like attacks have a cardiovascular cause. *Journal of the American College of Cardiology*, 36(1):181–184, 2000.

[131] Sina Zarrieß, Henrik Voigt, and Simeon Schüz. Decoding methods in neural language generation: A survey. *Information*, 12(9):355, 2021.

[132] Xianhua Zeng, Li Wen, Banggui Liu, and Xiaojun Qi. Deep learning for ultrasound image caption generation based on object detection. *Neurocomputing*, 392:132–141, 2020.

[133] Zizhao Zhang, Yuanpu Xie, Fuyong Xing, Mason McGough, and Lin Yang. Mdnet: A semantically and visually interpretable medical image diagnosis network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3549–3557, 2017.

[134] Stephen Wu; Kirk Roberts; Surabhi Datta; Jingcheng Du; Zongcheng Ji; Yuqi Si; Sarvesh Soni; Qiong Wang; Yang Xiang; Bo Zhao and Hua Xu Qiang Wei. Deep learning in clinical natural language processing. *Journal of the American Medical Informatics Association*, pages 27(3):457– 470, 2020.

[135] Dani Kiyasseh; Tingting Zhu and David Clifton. Let your heart speak in its mother tongue: Multilingual captioning of cardiac signals. preprint arXiv:2103.11011.