University of Milano - Bicocca

Department of Informatics, Systems and Communication

Master's degree in Data Science

Brain-to-image: Image classification and retrieval based on EEG signals

Supervisor: Prof. Paolo Napoletano

Co-supervisor: Prof. Simone Bianco

Master's Thesis of:

Gianluca Scuri

886725

Academic Year 2022-2023

Contents

Chapter 1

Introduction

With its 86 billions of interconnected neurons [20], whose interactions change from millisecond to millisecond, the human brain is the most complex organ of the human body [49]. For years scientists from diverse fields, including neurobiology, neuroscience and artificial intelligence have tried to decode the human brain and gain insights on its functioning. To achieve this goal, many invasive and non-invasive tools have been designed to extract the biosignals generated by neurological processes. EEG is one of the most popular non-invasive brain recording methods and has been used vastly in studies related to understanding the brain activities. This neuroimaging technique is capable of recording time series of the electrical activity produced by the sum of an electrical potential difference across groups of neurons. The analysis of these non-linear and non-stationary signals in recent years has been dominated by machine learning and artificial intelligence because of their ability to recognize patterns and detect content. In particular, these techniques are demonstrating great potential in decoding the signals involved in visual perception, the process of receiving and interpreting stimuli from the retina. This area is of great interest because it has several possible applications, it can be used: to assess and treat visual dysfunction, to assist patients who have difficulties communicating due to psychological trauma or disability, to extend the potential of Brain-Computer Interface, or to enhance experiences and immersion in VR/AR systems.

Several papers have been published in recent years regarding the analysis of EEG signals with deep learning techniques but only recently it has been applied for the classification of stimuli associated with natural images belonging to a high number of classes. Early works are from Spampinato et. al. [55] [28] [59] [39] in which they were able to classify EEG signals by 40 classes of images shown with a claimed accuracy of 83%. In addition, they also proposed a model of image generation from EEG signals with conditional GAN. Lately [24] and [29] improved the classification accuracy and reconstruction using respectively a visual-

guided convolutional neural network and a Geometric Deep Network-based GAN. However, Li et al. [33] pointed out that these and other works are based on a dataset which present a design pitfalls that inflates the results.

For this reason, in 2022 a new large-scale EEG dataset, collected with Rapid Serial Visual Presentation RSVP technique, has been proposed to model human visual recognition and decode objects pairwisely [10]. The uniqueness of this dataset is the very high number of image classes used and the very high-frequency protocol. This project is entirely based on this dataset and, at the beginning of the development of this project, no paper had used it yet or had ever attempted to classify such a high number of classes. However, the goal of this project is not only the classification of stimuli with respect to the image classes but also the reconstruction of the seen image. To do this, an approach never tested in the literature for this application is used, which involves extracting a pool of most similar images from an existing dataset. This image retrieval approach, in contrasts with the trend of generative models, could allow overcoming some of their critical issues: visual artifacts, training complexity, and high computational cost. Both classification and retrieval tasks relies on 3 key concepts that need to be demonstrated: EEG signals can contain information about the image content, EEG features can be extracted from this signals and that these features can be used for classification or retrieval.

To perform these tasks several tests are performed to identify the best models and techniques. Regarding the classification of signals, this was performed on both the concepts and categories of the images shown. There are 1654 concepts in the dataset indicating the object in the image while the 27 categories indicate families of concepts. The first type of classification is a Single-Label Multiclass and in this project its performance are evaluated by considering different subsets sizes of these concepts. The second type, on the other hand, is a Multi-Label Multiclass classification since each concept can belong to 0,1 or more categories. To carry out both the classification tasks 3 different models were developed to deal with EEG signals and these were compared with the well-known EEGNet model [32]. All the proposed models are based on deep learning and process data differently. What emerged from the experiments was that the proposed models, in several cases, yielded superior results to the EEGNet model. In fact, with regard to signal classification, the model that allowed the highest accuracy for signal classification with respect to 1654 concepts is the LSTM+CONV, a model composed of an LSTM layer to process temporal informations and a convolutional layer for electrodes disposition. This model obtained top-1 and top-5 accuracy values of 0.94% and 3.01%, respectively. These results, although quite low, are still 15 times higher than chance, showing that it is possible to extract visual information even from stimuli presented with a high-frequency protocol. As for

the classification of the 27 categories what is obtained is an F1 value of 9.37% and an accuracy of 20.82%. Furthermore, many other tests on classification have been done in this project. A new EEG preprocessing, different from that offered by the authors of the dataset, is proposed that allows to improve accuracies of 60% for some models. In addition, comparisons were made between 5 subjects of the dataset, showing that EEG signals can vary greatly from person to person.

Regarding the image retrieval task from EEG signals, several steps are necessary. It is first necessary to perform feature extraction for EEG and images. To do this, the best trained model obtained at the classification task and the pre-trained ResNet18 model on ImageNet are used, respectively. These features are then reduced to 100 components through PCA and scaled in the range [-1,1]. Second, it is necessary to align the space of these two features in order to obtain images features from the EEGs. This step is performed with a linear model and two different loss functions are tested: Mean Squared Error (MSE) loss and Contrastive Language–Image Pre-training (CLIP) loss. Finally, it is necessary to define the image dataset on which to perform image retrieval and it is necessary to represent it in such a way that similarity metrics can be evaluated. To do this, 1500 images are selected from the 100 classes on which the encoder and mapper were trained. These, after also being preprocessed with ResNet18 and PCA, were represented in a KDTree. This data structure allows to be queried and retrieve the k most similar images with respect to a distance metric. Because the chosen image dataset contains also the image that generated the stimulus and other images of that same class, it was possible to calculate quantitative metrics. By extracting k=10 images it was in fact possible to obtain a 2.13% probability of extracting the correct image and 14.11% probability of extracting images of the same class, which are higher than 0.65% and 9.38% corresponding to chance. Also for this task, the results are quite low and mainly due to the poor ability of the EEG encoder to extract semantic information from EEG signals.

The experiments showed that this pioneering project on this dataset allowed to: demonstrate the possibility of classifying EEG signals based on classes present in the images shown and the possibility to reconstruct the original input image using image retrieval. However, this project also demonstrated some limitations that capped accuracies: the extreme experimental paradigm, the inherent limitations of EEG signals and the rigid approach used based only on EEG and images.

This thesis is composed of several parts. In Chapter 2 the EEG signals discussed are contextualized with a brief theoretical introduction. In Chapter 3 the related works on both EEG signals classification by image classes and image reconstruction are presented. In Chapter 4 are presented all the methods proposed while in Chapter 5 all the experiments performed. Finally, Chapter 6 contains the conclusion and possible future developments of this project.

Chapter 2

Theoretical foundations

2.1 Human brain

2.1.1 Neuroanatomy and neuroscience

The nervous system extends throughout the body and the main part of it is the central nervous system CNS that consists of the brain and spinal cord. The cerebrum, the largest part of the human brain, is the core where all the information are processed. It is divided longitudinally in two cerebral hemispheres with broadly similar shape and functions. Each of these are made up of white matter in the core and gray matter on the outer surface, the cerebral cortex. The main task of the brain is to collect and integrate information from other organs and the environment, plan appropriate responses and transmit them to the rest of the organism.

The brain, and more specifically each of the two hemispheres, are conventionally divided into four lobes (Figure 2.1) and each of these addresses specific functions of the organism [56].

- The frontal lobe is associated with executive functions, including self-control, planning, reasoning, and abstract thought [31].
- The temporal lobe, located on the sides, controls auditory and visual memories, language, and some hearing and speech [31].
- The parietal lobe, located behind the frontal lobe, is responsible for sensory integration, location awareness and learning movements [31].
- The occipital lobe, which is located in the back of the head, is the visual processing center. It's responsible for visual reception, visual-spatial processing, motion perception, and color recognition [31].

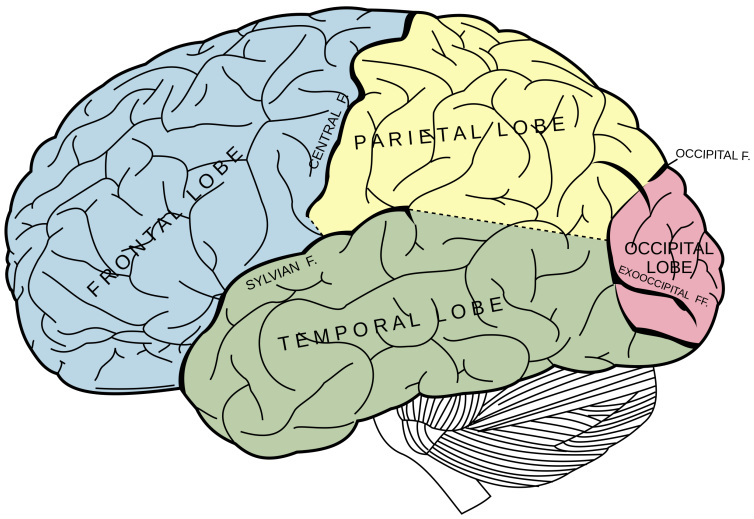

Figure 2.1: *Principal fissures and lobes of the cerebrum viewed laterally.*

These brain activities are made possible by the interconnections of neurons and through neurotransmission process. There are more than 86 billion neurons in the brain [20] connected to form neural pathways, neural circuits, and elaborate network systems. A neuron is comprised of a cell body, axon, and dendrites. Dendrites have branching structures that receive signals from other neurons' axon terminals. These signals can trigger an action potential, an electrochemical signal, which travels along the neuron's axon to reach axon terminals, where it can connect to another neuron's dendrites or cell body. The initiation of an action potential happens at the axon's initial segment, housing specialized proteins. When the action potential reaches the axon terminal, it prompts the release of a neurotransmitter at a synapse, which affects the target cell.

The brain's electrical charge is sustained through the activity of billions of neurons. Neurons possess an electric charge, referred to as "polarization," which is established by specific membrane transport proteins responsible for moving ions across their membranes. Neurons continually exchange ions with their external environment, a process essential for maintaining their resting potential and facilitating the transmission of action potentials. Since ions with the same charge tend to repel one another, when numerous ions are simultaneously expelled from multiple neurons they push neighboring ions, which in turn push their neighbors creating a propagating wave. This phenomenon is known as volume conduction.

2.1.2 Visual perception

The sensory nervous system is involved with the reception and processing of sensory information. The brain receives and interprets information from the five senses of vision, smell, hearing, taste and touch. This information travel through the cranial nerves, through tracts in the spinal cord, and directly at centers of the brain.

Visual perception is the ability to interpret the surrounding environment using light reflected by objects in the environment. Vision is generated by light in the range of wavelengths between 370nm and 730nm, that reaches the retina of the eyes. There the photo-receptors in the retina are designed to decode photon stimuli into a series of electrochemical signals to be transmitted to the brain. The signal is then sent to the visual cortex in the occipital lobe through optic nerves. Along this nerve, at the level of the hypothalamus, there is the optic chiasm, where the optic nerves cross. Here optic nerve fibers from one eye cross to the opposite sides joining the fibers from the opposite retinas to form the optic tracts. The arrangements of the eyes' optics and the visual pathways mean vision from the left visual field is received by the right half of each retina, is processed by the right visual cortex, and vice versa as shown in Figure 2.2.

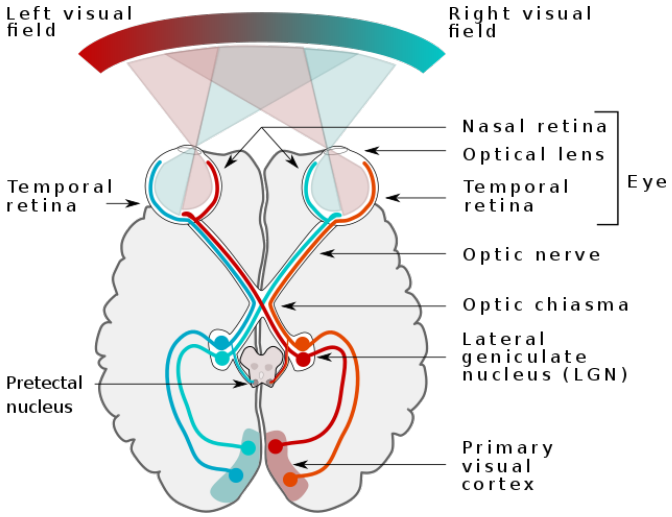

Figure 2.2: *Visual pathway with optic chiasm (X shape).*

Then, after the signal reaches the occipital lobe, the visual processing occurs through two distinct 'streams' of information [12]. One stream, sometimes called the What Pathway, is involved in recognizing and identifying objects. The other stream, sometimes called the Where Pathway, concerns object movement and location, and so is important for visually guided behavior. The human brain achieves

visual object recognition through multiple stages of linear and nonlinear transformations taking a mere 13 milliseconds [43].

2.2 Neuroimaging

2.2.1 Biosignals

Electrical biosignals, or bioelectrical time signals, usually refers to the change in electric current produced by the sum of an electrical potential difference across a specialized tissue, organ or cell system like the nervous system. Biosignals, if measured and analyzed, contain useful information that can be used to understand the underlying physiological mechanisms of a specific biological event or system. Regarding the brain, every cerebral activity, as generated by electrochemical impulses, creates electrical biosignal. Thus, by analyzing these signals it is possible to study the functioning of the brain during different brain processes.

2.2.2 Brain waves

Brain activity shows oscillatory behavior occurring at various frequencies called neural oscillations or brain waves. Many of these oscillations exhibit distinct frequency ranges, spatial patterns, and are linked to different states of brain function, such as wakefulness and various sleep stages. Brain waves consist of a mixture of diverse base frequencies which have been arranged on five different frequency bands. Each frequency band is associated to a different state, that can change depending on age, and is generated by different locations. Here are presented the bands and the associated activities for adults.

- Delta (0.5–4 Hz): is normally associated with slow-wave sleep (or deep sleep)
- Theta (4–8 Hz): generated in a drowsy or idling state, also associated with relaxed, meditative, and creative states.
- Alpha (8–12 Hz): normally associated to a relaxed/reflecting state or closing the eyes
- Beta (12–35 Hz): generally associated with active, calm, intense or stressed state and generated during active thinking, focus, high alert, anxious
- Gamma (>35 Hz): Displays during cross-modal sensory processing and usually appear during learning and problem-solving tasks

2.2.3 Recording techniques

There are many techniques that allow neuroscientists to study the brain activity. Here a few of them are listed. Each of these techniques come with their own strengths and limitations and exploit different phenomena.

- Positron Emission Tomography (PET): measures brain metabolism and the distribution of exogenous radio-labeled chemical agents.
- Near Infra-Red Spectroscopy (NIRS): measures regional level parameters such as oxygenation and cerebral tissue blood flow.
- Magnetoencephalogram (MEG): recording magnetic fields produced by brain electrical currents using very sensitive magnetometers.
- Functional Magnetic Resonance Imaging (fMRI): relies on the fact that cerebral blood flow and neuronal activation are coupled. When an area of the brain is in use, blood flow to that region also increases.
- Electroencephalography (EEG): records the electrical activity produced by the brain's neurons through the use of electrodes that are placed around the research participant's head.

The latter technique record an electrogram of the spontaneous electrical activity of the brain using a series of electrodes placed in many position on the head and a differential amplifier, which registers the difference between two electrodes. It is typically a non-invasive technique, with the EEG electrodes placed along the scalp (commonly called "scalp EEG"), but there is also an invasive variant in which the electrodes are surgically placed on the exposed surface of the brain (intracranial EEG). The data used in this project were recorded using scalp EEG, for this reason in the next section only this technique will be explored in depth.

2.3 Electroencephalography EEG

The history of EEG (electroencephalography) began in the early 20th century when the physiologist and psychiatrist Hans Berger pioneered the recording of human electrical brain activity. It quickly became a crucial tool for diagnosing neurological disorders like epilepsy, brain damage from head injuries, brain tumors and many more. In the last years it played a vital role in neuroscience research and brain-computer interface development.

2.3.1 Mechanism

Electroencephalography records voltage fluctuations using a bio-amplifier and a series of electrodes placed on the scalp. As the electric potential generated by an individual neuron is far too small to be picked up by EEG, the electrical activity monitored reflects the summation of the synchronous activity of thousands or millions of neurons [37]. When the wave of ions reaches the electrodes on the scalp, they can push or pull electrons on the metal in the electrodes. This difference in push or pull voltages between any two electrodes can be measured by a voltmeter and creates EEG. However, as voltage field gradients fall off with the square of distance, not all neurons will contribute equally to an EEG signal, with an EEG predominately reflecting the activity of cortical neurons near the electrodes on the scalp [25]. For this reason the recordings made by the electrodes on the surface of the scalp vary in accordance with their orientation and distance to the source of the activity. Furthermore, the value recorded is distorted by intermediary tissues and bones, which act in a manner akin to resistors and capacitors in an electrical circuit. Deep structures within the brain further away from the electrodes will not contribute directly to an EEG.

2.3.2 Recording technique

In standard scalp EEG, the recording is achieved by positioning electrodes on the scalp's surface using a conductive gel or paste. Often, this follows the preparation of the scalp area, which may involve gentle abrasion to minimize impedance caused by dead skin cells. Many EEG systems employ individual wires connected to electrodes, while others opt for caps or nets that incorporate embedded electrodes. The latter approach is particularly prevalent when high-density arrays of electrodes are necessary.

There are different standards in the electrodes placement, the most used for clinical and research applications is the international 10–20 system (Figure 2.3). The "10" and "20" refer to the fact that the actual distances between adjacent electrodes are either 10% or 20% of the total front–back or right–left distance of the skull. Measurements are taken based on specific landmarks on the skull. These common methods were created to establish standardized testing procedures, guaranteeing that a subject's research results could be compiled, reproduced, and effectively analyzed and compared through the scientific method. This system relies on the connection between the electrode's placement and the corresponding region of the brain, particularly the cerebral cortex. This standardized system ensures also that the naming of electrodes is consistent across laboratories and is usually defined by a letter, to identify the lobe or area of the brain, and a numbers, even on the right side and odd on the left side of the head.

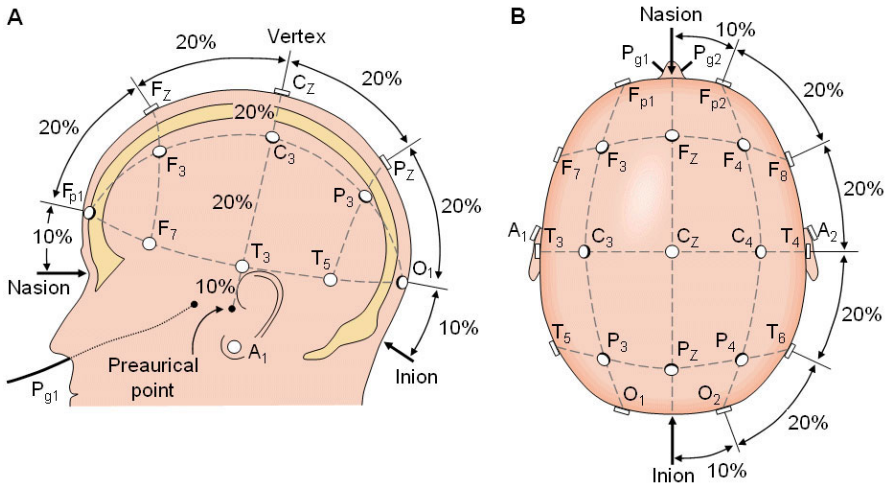

Figure 2.3: *Electrode locations of International 10-20 system for EEG recording.*

Each electrode is connected to an input of a differential amplifier (with one amplifier for each pair of electrodes). A shared reference electrode is connected to the other input of each differential amplifier. A typical adult human EEG signal is about 10 µV to 100 µV in amplitude when measured from the scalp [2]. These amplifiers boost the voltage disparity between the active electrode and the reference, typically amplifying it by a factor of 1,000 to 100,000 times, which translates to a voltage gain of 60 to 100 decibels (dB).

Because an EEG voltage signal reflects the voltage difference between two electrodes, the presentation of the EEG for the interpreting neurologist can be configured in various ways. This depiction of EEG channels is commonly referred to as a montage. So the term channel refers to the potential difference between an electrode of interest and a reference.

- Sequential montage: each channel illustrates the voltage difference between two adjacent electrodes. The complete montage comprises a series of these channels throughout the array of electrodes.

- Referential montage: each channel shows the voltage difference between a specific electrode and a designated reference electrode. There isn't a fixed standard position for this reference electrode, but it is distinct from the "recording" electrodes. Often, mid line positions are utilized to ensure that the signal isn't biased towards one hemisphere over the other.

- Average reference montage: the outputs of all amplifiers are summed and averaged, and this averaged signal serves as the common reference for each channel.

- Laplacian Montage: each channel represents the voltage difference between an electrode and a weighted average of the surrounding electrodes.

2.3.3 EEG strengths and weaknesses

EEG is one of the first methods designed to analyze brain activity and is still used nowadays. In this subsection the main advantages and disadvantages of EEG over the other methods are exposed.

The first strength of this method is its non-invasiveness and simplistic fidelity. The equipment is compact, silent and significantly cheaper than most of the other techniques. This allowed a high research throughput and the creation of numerous datasets. Secondly, EEG can readily have a high temporal resolution (although sub-millisecond resolution generates less meaningful data) with sampling rates between 250 and 2000 Hz in clinical and research settings. Moreover it is relatively tolerant of subject movement because different method exists for minimizing and even eliminating movement artifacts (in Figure 2.4 are reported some of the possible artifacts).

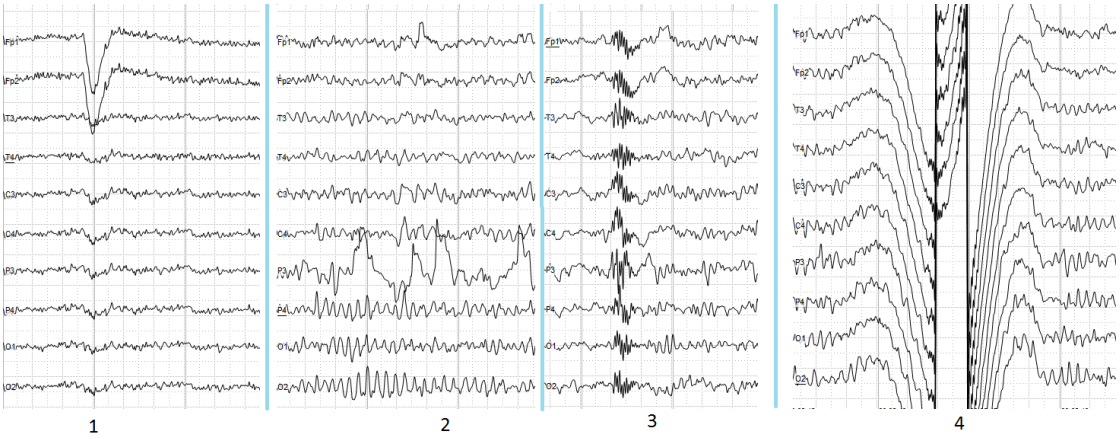

Figure 2.4: *EEG artifact: 1. caused by the excitation of eyeball's muscles, 2. caused by bad contact between electrode and skin (and thus bigger impedance), 3. caused by swallowing, 4. caused by bad contact between reference electrode and skin.*

Despite the advantages, EEG comes with some disadvantages. First of all it takes a long time to connect a subject to EEG, as it requires precise placement of dozens of electrodes around the head with the use of a conductive gel. Furthermore many variants of the international standard system are used hindering the reproducibility of the experiments. Secondly, EEG has a low spatial resolution because it records the brain activity from the scalp. For this reason it introduce a blurring effects of the head volume conductor and prevents measures of the neu-

ral activity that occurs below the cortex. Moreover EEG is recorded with a low signal-to-noise ratio and sophisticated data processing is needed to extract useful information. The reason for this is that a large population of cells in synchronous activity is necessary to cause a significant deflection on the recordings. Lastly, the signals collected with this instrument are very personal and they can vary greatly from person to person and even from session to session. Some of these discrepancies may be due to different conformation of the skull and skin, slight uncertainties in electrode placement, different corebrain activations.

Chapter 3

Related works

The electroencephalogram is a recording instrument that has been used in the medical and research fields for 100 years now. Since its invention, EEG analysis has brought significant advancements in studies of diagnosis and treatment of various neurological brain conditions and in Brain-Computer Interfaces (BCIs). BCI is a computer-based system that allow a direct communication pathway between the brain's electrical activity and an external device. It's main purpose is assisting, augmenting, or repairing human cognitive or sensory-motor functions. Because of its many advantages (reported in the Section 2.3.3), EEG has been extensively used to study the functioning of the human mind in general and visual perception.

This chapter reports some of the work and methods that have arisen to deal with EEGs , to be able to classify them with respect to the visual stimulus and to reconstruct the image used as the stimulus. Both papers employing traditional methods and more recent papers carried with machine learning and deep learning models are shown. In particular, the latter techniques are the ones that have shown the greatest potential; it is possible to see the growth of interest in them in Figure 3.1.

3.1 Feature Extraction

3.1.1 EEG

Feature extraction refers to the process of transforming raw data into numerical features that can be processed while preserving the information in the original data set. It is an essential step for this type of signals and usually yields better results than applying machine learning directly to the raw data. In this section are reported some of the most common features extraction methodologies that have been applied to EEG signals divided in one-dimensional and multi-dimensional

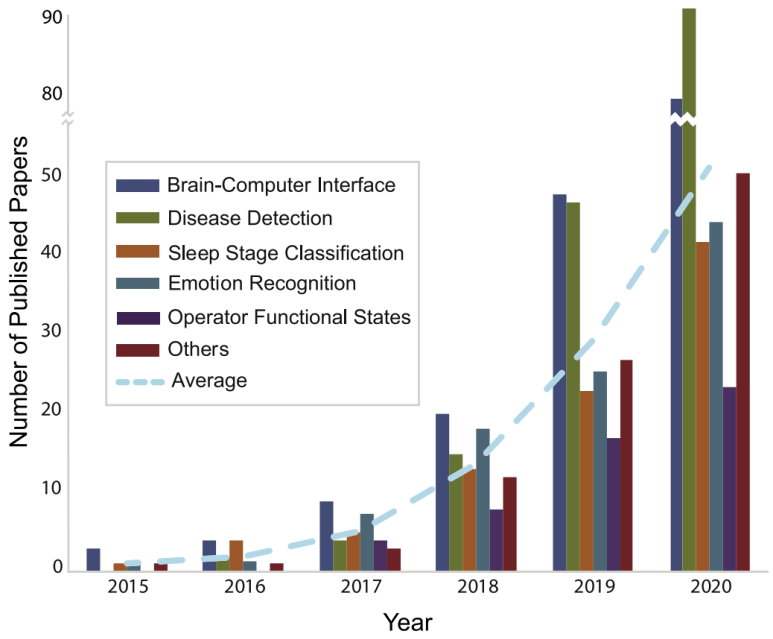

Figure 3.1: *Numbers of the published papers, per category related to EEG signals analyzed with deep learning models, in each year. Note that numbers before 2015 are omitted because of rare papers. [11]*

[52]. To the first category belong the following methods.

- Time domain: traditional techniques to study time series such as autoregressive AR models [3] or descriptive statistics (e.g. mean, standard deviation, abstolute deviation, ...).
- Frequency/spectral domain: models that enable the contributions of different brain waves to be evaluated. Among them are Fourier transform, Power Spectral Density and Band Power.
- Decomposition domain: these methods allow simultaneous feature extraction and filtering of the signal. In this category belong Adaptive Hermite decomposition (AHD) and wavelet transform.

While among the multi-dimensional feature extraction techniques there are the following methods.

- Joint time-frequency domain: these models exploit spectral information at different time instants. The main method in this category is Short-time Fourier Transform (STFT) that allow to convert EEG channels into 2D images.

- Spatial domain: in this domain there is the Common Spatial Pattern (CSP), a supervised spatial filter, that converts the brain waves into a unique space. In this unique space, the variance of one group is magnified, and a lower variance is seen in the remaining group [45].

As an alternative to these feature extraction methods, there are deep learning models. Usually are end-to-end model that integrate feature extraction and classification or clustering. These models have shown great promise in helping make sense of EEG signals due to their capacity to learn good feature representations from raw data [47]. These are very powerful because they are able to extract features, optimized for the task of interest, from training data.

3.1.2 Image

Also for images there are different features extraction techniques, both traditional and deep learning based. In the first category there are the different models as: Scale Invariant Feature Transform (SIFT), Features from accelerated segment test (FAST) and Histogram of Oriented Gradients (HOG). The first one allows estimation of scale-space extrema followed by key-point localization, orientation and subsequently computation of local image descriptor for each key point [35]. The second one is a corner detection method to extract feature points [46] while the third one allow the occurrence of edge orientations to be measured [36].

However, in recent years, deep learning models have emerged as the principal method for extracting image features, which enable learning the best representation based on training data. Feature extractors in this category are often obtained by truncating pre-trained neural networks on large image datasets to extract the latent space representation of images. The most widely used models for this purpose are: VGG [51], ResNet [18] and DenseNet [22] depending on application. These differs for the amount of parameters and consequently for the performance.

3.2 EEG classification based on image class

Research in neuroscience and neuroimaging [17] has shown that human cognitive processes related to human perception (and visual perception in particular) can be decoded through non-invasive imaging techniques such as fMRI, EEG, MEG. Early works related to studying visual perception in humans by EEG focused on recognizing when, where and how intensely different areas of the brain were activated. In particular, the studies focused on the study of Event-Related Potentials (ERPs) stimulated by light pulses, colors, geometric shapes or natural images. In Figure 3.2 a common potential curve is shown characterized by different peaks.

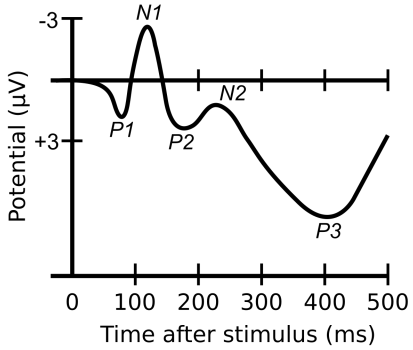

Figure 3.2: *A waveform showing several ERP components, including the N100 (labeled N1) and P300 (labeled P3). The ERP is plotted with negative voltages upward, a common, but not universal, practice in ERP research.*

Each of these peaks depict a specific phase in brain processing and is characterized by polarity (positive or negative going voltage), timing, scalp distribution, and sensitivity to task manipulations. Studing the ERP components elicited by different images can be used to classify EEG signals via category-specific ERP patterns.

This methodology had been used in a number of cognitive neuroscience studies and, by identifying specific regions of visual cortex, have demonstrated that up to a dozen of object categories can be decoded in event-related potential (ERP) amplitudes recorded through EEG [7]. Here are reported a few of the papers that leveraged this method. In [50] they classified 12 categories comparing the signal evoked by its spoken name, its visual representation and its written name on 20 subjects. In [4] they studied how the visual system abstracts object category across variations in retinal location. In [62] they classified 4 categories (human faces, buildings, cats and cars) with an accuracy of 50% across 3 subjects. In [26] they trained a classifier able to distinguish EEG brain signals evoked by twelve different object classes, with an accuracy of about 29%. All of these works performed the classification leveraging traditional machine learning models.

However, ERP features contain little information for complex image classification or image generation tasks. For this reason deep learning models then began to be used to classify EEG signals. The availability of large datasets containing unprecedented numbers of examples is often mentioned as one of the main enablers of deep learning research in the early 2010s. Deep learning is an enhanced variant of traditional neural network, which is thought to be established based on the inspiration of hierarchical structure existing in visual cortex of the human brain. In paper [5] they realized a review of 90 works on EEG classification tasks that have been explored with deep learning. "All the considered studies fell into six groups:

emotion recognition (16%), motor imagery (22%), mental workload (16%), seizure detection (14%), and sleep stage scoring (9%), event related potential detection (10%), and other studies (13%), which include Alzheimer's classification, bullying indices detection, depression, gait pattern classification, gender classification, detection of abnormal EEG, and transcranial stimulus treatment effectiveness".

Regarding work related to visual stimuli, the works that started to exploit the potential of deep learning for raw EEG classification are the following [27] presented a visual category recognition method using EEG signals, and they classified images into three object categories (animals, faces and inanimate). In [34] they obtained an average classification rate of 82.70% using a dataset containing 400 images and brain signals. [40] they presented an EEG-based image annotation system trained on 2500 images that obtained F1-score of 0.88. These researches above demonstrate that it is feasible to classify the visual scenes using brain signals, however, they only processed a small number of categories, and their classification accuracies for visual scenes are not high enough. For this reason in 2017 [55] explored the capabilities of deep learning in modeling visual stimuli–evoked EEG with more object classes than previous methods. This is the beginning of the research on the multi-class visual object EEG signals. In the same paper they also investigated how to project images into an EEG-based manifold in order to allow machines to interpret visual scenes according to human brain processes. In the classification task they reported a 83% accuracy on their 40 classes dataset. Based on the same dataset they also released 2 other papers in which they proposed a generative model [28] [59] to recreate the image shown and siamese training configuration [39] to maximizes the compatibility measure between visual features and brain representations. The dataset they created and used has become one of the most famous and used because of the high number of EEG-image pairs. However, [33] demonstrated that the dataset used in all these works had been recorded incorrectly with a block design paradigm where all stimuli of a given class are presented together. This design, if not handled properly, leads to classification of arbitrary brain states based on block-level temporal correlations that are known to exist in all EEG data, rather than stimulus-related activity. In particular, arbitrary temporal artifacts of the data instead of brain activity are classified and inflates the accuracy values.

After these works other papers tried to classify EEG evoked by visual stimuli on a corrected version of this datset or on new datasets. In 2022 two new datasets were proposed to overcome the limited number of categories in the previous ones. [10] and [15] proposed two EEG datasets based on a 1854 concepts image dataset [19] using Rapid Serial Visual Presentation (RSVP). RSVP task is one in which a participant detects a single target image in a rapidly refreshing image stream at the same location. With this method stimuli are all still processed by the visual

system and their neural representations can co-occur, however cerebral responses can overlap each other and they are not fully captured. In the presentation paper they also demonstrated the feasibility of image-to-EEG encoding and category separability. In particular, they trained a full end-to-end DNN model that output EEG responses for arbitrary input images. The first paper published on the first dataset [10] is [8] which uses multi-modal learning of brain-visual-linguistic features. They focused on modeling the relationships between brain, visual and linguistic features via multi-modal deep generative models. In particular for every image they exploited state-of-the-art image captioning model and Wikipedia articles for 1000 categories to obtain linguistic features. With this approach they obtained a 5.82% top-1 accuracy and 17.45% top-5 accuracy on 200 unseen classes in a zero-shot classification. Currently, [54] obtained the state-of-the-art performance on the first dataset [10], classifying 200 concepts in a zero-shot fashion with a top-1 accuracy of 15.6% and a top-5 accuracy of 42.8%. In order to achieve this result natural language supervision was used to pre-train a visual model. Image text pairs replace the original hard labels to drive self-supervised learning for better feature representation and achieve good results in zero-shot fashion.

3.3 From EEG to image

This section reports on papers that have been published to obtain the image shown to subjects only through the use of EEG signals. The image reconstruction process can be performed with different objectives: reconstruct an image with similar semantic to the original image or reconstruct an image that is visually as close as possible to the original. Thus, the former aims to include the main subjects in the image without regard to their positions and sizes (e.g. a brown dog running). The latter aims to recreate the original composition of the image by identifying shapes with correct proportions and positions. These two goals converge when a high accuracy is obtained: in the former the more semantic information extracted and the more visually similar the image becomes, and in the latter the greater the detail of the shapes and the easier it is to recognize objects.

In addition, regarding the evaluation of the generated images, there are several metrics that can be evaluated. Different aspects can be considered: the independent image quality, the image quality with respect to the real world, and the image quality with respect to a reference. Therefore, many metrics have been defined that take into account different definitions of similarity. The utility of these metrics will vary based on the nature of the task, here a few of them are presented.

- Inception Score: "How realistic is this image relative to the pretrained model?"

- FID Score: "How similar is this group of images relative to this other group of images?"
- LPIPS Score: "Did the structure of this patch change, and by how much?"
- SSIM, MSSIM, PSNR: "How noisy is this generated image as compared to the ground truth image?"

Given the wide variety of metrics and targets that can be used by researchers here are papers on the topic reporting only the models used.

3.3.1 Image generation

[64] and [9] tried to reconstructing geometrical shapes from brain activities focusing primarily in generating precise edges and other low-level details to construct natural shapes at pixel-level.

The first works that proposed a generative model applied on natural images was [28] in 2017 and [59] in 2018. They performed attempts at image generation from a latent feature space using Deep Generative Models (DGMs). This family of methods raised from deep learning and is formed through the combination of generative models and deep neural networks. In this paper the two most popular DGMs are tested: Variational Autoencoders (VAEs) and conditional Generative Adversarial Networks (cGANs). With both the models they managed to obtain recognizable images. GANs generated very sharp-looking images but with some artifacts while VAEs lack sharpness, because of the noise introduction during the intermediate representation, but is more stable in training. Their best model is ToughtViz, results are shown in Figure 3.3. However, also these works were affected by the flawed dataset as decribed in Section 3.2 and reported in [33].

Then other similar works were proposed. In [65] a model composed of LSTM + CNN combined with Spectral Normalization Generative Adversarial Network (SNGAN) was proposed to yield seen images from EEG encodings. [30] proposed conditional Progressive growing of GANs (CProGAN) to develop perceived images and showed higher inception than previous related works. [63] proposed a contrastive self-supervised approach that has been shown to maximize the mutual information between visual stimulus and corresponding EEG latent representations. [53] used a contrastive learning method to extract features from EEG signals and synthesize the images from extracted features using conditional GAN. This work is focused on generation with small datasets (result in Figure 3.3).

Regarding image reconstruction with focus on visual similarity of output, [29] proposed a Geometric Deep Network-based GAN (GDN-GAN). This model is trained to map the EEG signals to the visual saliency maps corresponding to each

(a) ThoughtViz [13] (b) EEG2Image (Ours)

Figure 3.3: *Qualitative comparison between the images generated by EEG signals using the ThoughtViz method (left) and the EEG2Image method (right). Images in the red bounding box are the sample images from the Object test dataset. These images are visualized by the participants, and the respective brain activity EEG signals is used here for reconstruction. [53]*

image (Figure 3.4). The saliency map is then used to guide the black and white image reconstruction of the original picture.

3.3.2 Image retrieval

A second approach that can be used to obtain images from EEG signals is to extract from an image dataset the one that is most similar to the input image. Similarity in this case is defined by the type of features which describe images and by the type of query used in the search. The advantage of this approach is to obtain artifact-free images but with the limitation that the extracted images belong to a specific and finite set. The only papers in literature leverage EEG signals to improve textual image search is [60]. However, in this case EEG signals are only used to refine the results of image retrieval from textual search. In fact, EEG is used with RSVP of a small pool of retrieved images to detect the optimal one. This model is trained using targets and distractors.

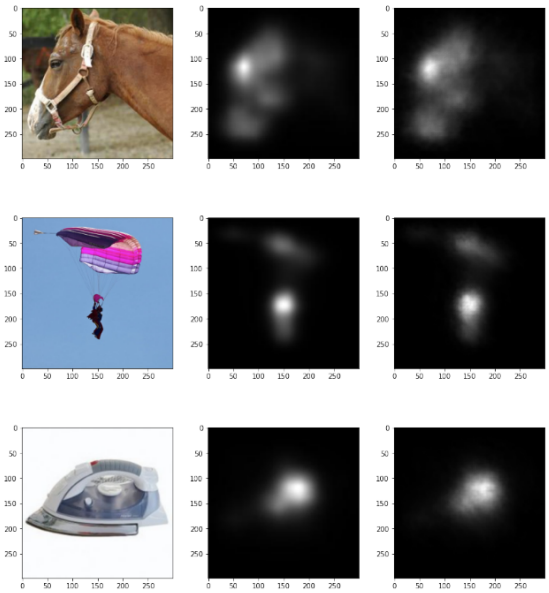

Figure 3.4: *Saliency reconstruction obtained in the paper [29]. From left to right: Image used as stimulation, fixation map, EEG-based reconstructed saliency map.*

Chapter 4

Methodology

4.1 Overview

In this project a framework to classify and retrieve images from EEG signals, obtained with Rapid Serial Visual Presentation (RSVP) technique, is presented. The overall framework is depicted in Figure 4.1.

The first task aims to classify EEG signals with respect to concepts and high-level categories into which the images shown can be classified. In particular this project deals with Single-Label Multiclass (SLMC) classification with a high number of concepts and Multi-Label Multiclass (MLMC) for a low number of higher-level categories. For this task, different models have been developed and compared to extract visual and semantic information from EEG signals.

The second task aims to retrieve the image most similar to the one shown. For this task it is first necessary to align the representation of EEG signals and images. Therefore in the training process, image and EEG signals are input into the framework as stimulus-response pairs. Subsequently, 1D feature vectors are extracted using the model that performed best on first task as EEG encoder and a computer vision model pre-trained on ImageNet as image encoders. These are then processed with PCA in order to obtain an equal number of values. Linear mapping is then applied to the same size output to align the image and EEG modal representations. Lastly it is necessary to define a database from which to extract the most similar images and create a space-partitioning data structure for organizing points in a k-dimensional space. The structure used is a KDTree which, in the inference phase, is queried to obtain the most similar vector by evaluating the distance in space with a nearest neighbor search (NN) algorithm.

Both these 2 tasks relies on three key intuitions that will need to be assessed:

- EEG signals, recorded during a visual task, convey feature-level and cognitive-level information about the image content;

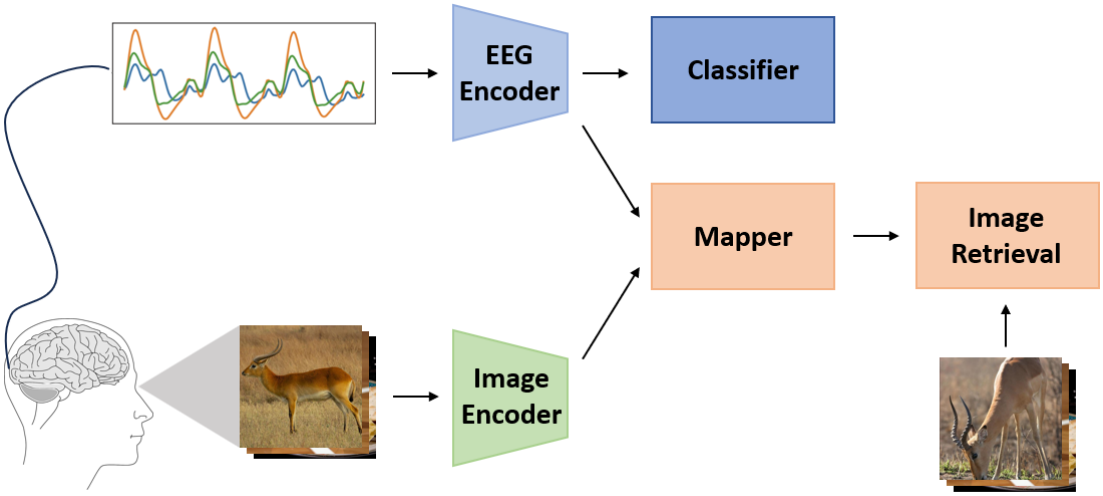

Figure 4.1: *Overall framework composed of image classification and reconstruction from EEG signals. First EEG signals are classified according to image classes. Then, image-EEG pairs are input and processed by an image encoder (pre-trained on Image-Net) and an EEG encoder (pre-trained in classification). The mapper perform a regression task and align EEG representation and image representation. The image representations obtained are then queried on an binary structures image dataset to retrieve the most similar matches.*

- A low-dimensional manifold exists and can be extracted from EEG signals to obtain a 1D representation (EEG features);

- EEG features are assumed to mainly encode visual data, thus it is possible to extract the corresponding image descriptors for automated classification or retrieval.

4.2 EEG classification by concept and category

Classification of concepts and high-level categories involves categorizing EEG signals with respect to concepts and categories into which the images shown can be grouped. To accomplish this, a classification model should be trained to: extract semantic information of the image shown from the EEG signals and identify the most likely class. The following section contains the models developed to carry out these tasks.

4.2.1 Classification models

The models developed for these classification tasks are 4 and are all based on neural networks.

1. LSTM + CONV1D
2. Temporal and Spatial CONV2D
3. EEGNet [32]
4. Spectrogram CONV2D

3 of these models are original while EEGNet [32] is a well-known network for EEG analysis in different fields of BCI, and in this project is used as a benchmark. They differ in the number and type of layers they are composed of and in the type of input they receive. The first 3 models receive the raw EEG in input, usually arranged into 2D matrix with C $\times$ T dimension, in which C denotes electrode channels and T denotes time samples. The last model receives the spectrogram of the EEG signals in input with C x N x M dimension, in which C denotes electrode channels, N denotes the number of frequency bins and M denotes the time bins.

The first model is composed of two layers to carry out a temporal and spatial analysis. This deep learning architecture combines the strengths of Long Short-Term Memory (LSTM) layer for modeling temporal dependencies and convolutional layers for spatial feature extraction. In this model the EEG in input is treated as a 1D array of size T with C channels. The architecture of the model is presented in Table 4.1. After the LSTM and convolutional layers the Rectified Linear Unit (ReLU) is applied to obtain non-linearity.

Layer	In	Out	Kernel	Stride	Dimension
LSTM	C	k1			(b,k1,T)
Avg. Pool 1D	k1	k1	m1	s1	(b,k1,T/s1)
Conv. 1D	k1	k2	m2	s2	(b,k2,T/s1)
Avg. Pool 1D	k2	k2	m3	s3	(b,k2,T/(s1*s3))
Flatten&Linear					(b, n_classes)

Table 4.1: *Architecture of LSTM + CONV1D model. C represent the number of EEG channels, T the number of EEG time points, b is the size of batches.*

The second model proposed combines two convolutional layers to perform operations on both temporal and spatial axis. Convolution along the two dimensions is widely used in deep learning-based EEG analysis models as EEG encoder [54].

In this case the inputs are treated as 2D arrays of size C x T with 1 channel. The network architecture is very concise and is presented in Table 4.2. After each convolutional layer, batch normalization and Exponential Linear Units (ELU) activation functions are used for improve training and add non-linearity.

Layer	In	Out	Kernel	Stride	Dimension
Conv. 2D	1	k	(1,m1)	(1,s1)	(b,k,C,T-m1+1)
Avg. Pool 2D	k	k	(1,m2)	(1,s2)	(b,k,C,(T-m1-m2+2)/s2)
Conv. 2D	k	k	(C,1)	(1,1)	(b,k,1,(T-m1-m2+2)/s2)
Flatten&Linear					(b, n_classes)

Table 4.2: *Architecture of Temporal and Spatial CONV2D model. C represent the number of EEG channels, T the number of EEG time points, b is the size of batches.*

The third model is a well-known architecture, proposed by [32], and used as benchmark model in many EEG applications. EEGNet is a compact convolutional neural network model for EEG-based BCIs. This model generalizes well across BCI paradigms and achieved performance comparable with the reference algorithms when only limited training data is available [32]. Its architecture is presented in Table 4.3. After every convolutional layer batch normalization is applied and, before the two average pooling layers, ELU activation function is applied.

Layer	In	Out	Kernel	Stride	Dimension
Conv. 2D	1	k1	(1,m1)	(1,1)	(b,k1,C,T)
Conv. 2D w/ const.	k1	k2	(C,1)	(1,1)	(b,k2,1,T)
Avg. Pool 2D	k2	k2	(1,m2)	(1,s1)	(b,k2,1,(T-m2+1)/s1)
Conv. 2D	k2	k2	(1,m3)	(1,1)	(b,k2,1,(T-m2+1)/s1)
Conv. 2D	k2	k3	(1,1)	(1,1)	(b,k3,1,(T-m2+1)/s1)
Avg. Pool 2D	k3	k3	(1,m2)	(1,s1)	(b,k3,1,(T-m2+1)/(s1*s1)
Flatten&Linear					(b, n_classes)

Table 4.3: *Architecture of EEGNet model [32]. C represent the number of EEG channels, T the number of EEG time points, b is the size of batches.*

The fourth model uses a different approach. For each channel of the EEG signal, the spectrogram is calculated, obtaining a matrix C x N x M in which C denotes electrode channels, N denotes the number of frequency bins and M denotes the time bins. This approach allows a detailed evaluation of the brain

waves frequency ranges. Then two convolutional layers and average pooling are applied to this input as shown in Table 4.4.

Layer	In	Out	Kernel	Stride	Dimension
Conv. 2D	C	k1	(m1,m1)	(1,1)	(b,k1,N,M)
Avg. Pool 2D	k1	k1	(m2,m2)	(s1,s1)	(b,k1,N/s1,M/s1)
Conv. 2D	k1	k2	(m3,m3)	(1,1)	(b,k2,N/s1,M/s1)
Avg. Pool 2D	k2	k2	(N/s1,N/s1)	(1,1)	(b,k2,1,1)
Flatten&Linear					(b, n_classes)

Table 4.4: *Architecture of Spectrogram CONV2D model. C represent the number of EEG channels, T the number of EEG time points, b is the size of batches.*

4.2.2 Classification by concepts

The first task of classification is performed on image concepts. The concept defines the main object represented in the natural image, and each image is associated with one and only one concept. This type of classification is called Single-Label Multiclass (SLMC) and to perform this task the models presented in Section 4.2.1 are used. In all these models, the output linear layer is added as a projector and is set to transform the features to the same size as the number of classes to be classified. In addition, a softmax function is added to the output to obtain the vector of probabilities, essentially a probability distribution over the output classes.

4.2.3 Classification by categories

Classification is also performed by category. Categories are a higher-level classification of the concepts and are defined by groups of concepts. This type of classification is called Multi-Label Multiclass (MLMC) and is a variant of the classification problem where multiple nonexclusive labels may be assigned to each instance. The models presented in Section 4.2.1 are used for this task as well, adjusting the generally lower number of classes in the output accordingly. Instead of softmax function, for this type of classification, sigmoid activation is used because it is treated as a binary classification problem for each possible category.

4.3 Image retrieval based on EEG

The second task aims to retrieve the image most similar to the proposed one among a defined set. This approach, although it has the limitations of being able to return only existing images, has the advantage of overcoming the complexity and limitations of image generation, e.i. avoids the generation of unrealistic images. To accomplish this task, however, it is necessary to be able to map signals of different nature, EEG and images, together. To do this it is necessary to first obtain a representation for these two digital signals. Then, once the model for calculating the representation of the EEG signal in image representation has been obtained, it is necessary to build a data structure that allows searching via multidimensional search key. These steps are described below.

4.3.1 Feature extraction

As anticipated, this step is necessary to obtain a representation of the two signals and allow their alignment. In particular, encoders must be defined to obtain features for both the EEG and the images, e.i. vectors that summarize the semantic component of the associated signal. The type of information contained depends mainly on the type of encoder used and is a key aspect in defining the type of image will be returned in retrieval.

Image encoder

Also for images it is necessary to identify an encoder that allows to obtain a vector of features that contain semantic information on the image itself. In this case, pre-trained computer vision models on large image datasets are exploited. In particular, the ResNet18 model was chosen as it is relatively small but highly performing. By removing the final dense layer used to train the model on the classification of 1000 ImageNet classes it is possible to obtain a vector of 512 values. In this case, since it is a pre-trained model it is necessary to preprocess the inputs like those used in training (resize with right resolution/interpolation, apply inference transforms, rescale the values etc.). Then PCA is also performed for these features to keep only the most significant M components and subsequently different normalization and scaling techniques are tested.

To assess the ideal number of components to be preserved and the transformation to be performed, features are tested in an image concept classification task. Specifically, with regard to feature scaling, several techniques such as Z-score 4.1, min-max scaling 4.2, and mean scaling 4.3 are tried.

$$x' = \frac{x - \overline{x}}{\sigma} \tag{4.1}$$

$$x' = \frac{x - min(x)}{max(x) - min(X)} \tag{4.2}$$

$$x' = \frac{x - \overline{x}}{max(x) - min(X)} \tag{4.3}$$

EEG encoder

To extract the semantic component of EEGs the EEGNet model, described in Section 4.2.1, is used. This is the model that, in the previous task, demonstrated good consistency in the different classification tasks (more details in Chapter 5). In fact, the greater the accuracy, the greater the model's ability to extract semantic information from the EEG signal. The model is therefore used with the weights trained in the classification task and, to turn it a feature extractor, the dense output layer is removed.

The C x T EEG matrices (C channels, T time instant) processed by this network produce a vector of N features. All the feature vectors obtained are then subjected to Principal Component Analysis (PCA), fitted only on the training test images and applied to all the features. Data points are projected into a new coordinate system and only a subset M of the coordinates that explains the most variance are considered. In this way it is possible to limit the course of dimensionality.

4.3.2 EEG-Image alignment

This phase is very important as it allows to map EEG signals into images. To do this, it is necessary to define an appropriate model and loss function. As described in Chapter 2, the brain performs linear and nonlinear processes to accomplish different tasks. In this case, the nonlinear components are extracted from the encoder, which has more layers and activation functions. To map signals instead, as is common to do in this field, a simple linear model is used. It receives M input values, obtained from feature extraction and PCA of the EEGs, and produces M output values obtained from feature extraction and PCA of the images.

Since this is essentially a regression task it is necessary to select an appropriate loss function. This particular project seeks to compare the use of two different functions: mean square error loss (MSE) and an adaptation of Contrastive Language-Image Pre-training loss (CLIP). The former allows to measures the mean squared error (squared L2 norm) between each element in the input. It can be thought as a (normalized) distance between the vector of predicted values and the vector of observed values. The latter exploits an adaptation of the contrastive training method proposed in [44]. In their case it was designed to align the feature

space of images with textual features, but here it is used to align images and EEG. The operation requires that for each batch supplied, the model learns to minimize the cosine similarity of one vector with its corresponding one and maximize it for all others. This is the principle of contrastive learning, obtain similar values for similar stimuli and vice versa.

4.3.3 Data structure

Once a structure has been made to obtain an image feature vector from the EEG signal, an attempt can be made to retrieve images as similar as possible to the one shown. This can be done by training a conditioned image generation model or by searching for the most similar image in a given dataset. In this project the second path is taken as it has never been tested in literature. To achieve this it is first necessary to identify a dataset of images suitable for the objective and is necessary to perform the same preprocessing and encoding of the images used in feature alignment. In particular it requires the same image encoder, PCA transformation and normalization. Then this dataset must be represented in a space that allows to search for the most similar images. The chosen structure is KDTree, a space-partitioning data structure for organizing points in a k-dimensional space. In this case the k-dimension correspond to the M features of the image representation.

On this tree, once created, is it possible to query for k-nearest neighbors using the scored image feature. The idea behind nearest neighbor methods is to find a predefined number of training samples closest in distance to the new point.

Chapter 5

Experiments and results

5.1 Dataset: THINGS-EEG2

5.1.1 Dataset selection

The aim of this project is to study the functioning of visual perception in healthy subjects. Specifically, the focus is to analyze by EEG how the cortex responds to vision of natural images. Many public datasets of EEG signals are available online for research but only a few of these meet the requirements of the project. In Table A.1 there are most of the public available datasets of EEG signals obtained in response to visual stimuli. In this project the THINGS-EEG2[10] dataset is used.

5.1.2 Images with concepts and categories

THINGS-EEG2 is one of the dataset developed within the project THINGS. The THINGS project is an initiative with the aim of bringing together researchers using the same image database to collect and share behavioral and neuroscience data in object recognition and understanding. The dataset is entirely based on the THINGS[19] dataset which is composed of 26,107 high-quality naturalistic images, with natural background and cropped to square size. Every image of the dataset belong to 1854 object concepts (e.g., antelope, strawberry, t-shirt) and each of these concepts contain 12 or more images. The 1854 object concepts have been categorized in few higher-level categories (e.g., animal, food, clothing). The original paper proposed 27 high-level categories (Figure 5.1) while in a more recent paper[57] proposed a 53 high-level division. These high-level categorizations are provided by human raters and were chosen trying to maximize the concepts included and minimize the overlap of classes.

THINGS-EEG2 pseudo-randomly divided the 1854 THINGS object concepts into non-overlapping 1654 training and 200 test concepts under the constraint

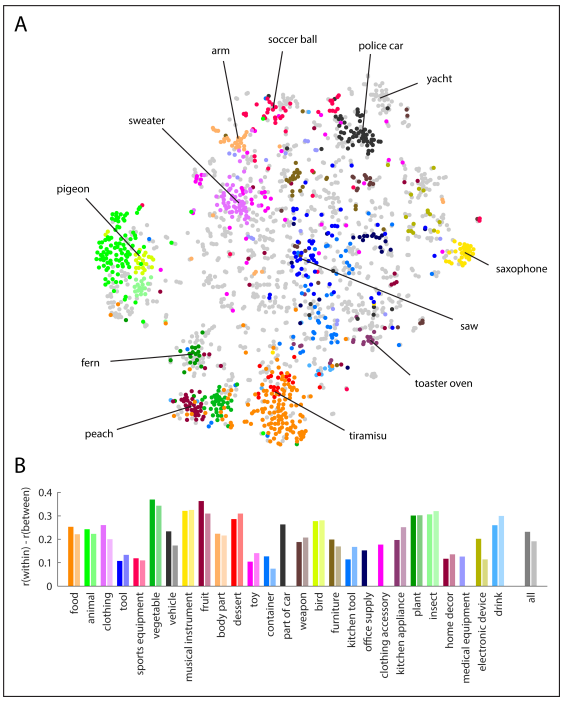

Figure 5.1: *A. Visualization of the semantic relationship of the 1,854 object concepts applying t-SNE to the semantic embedding, with the 27 core object categories depicted in different colors and example concepts highlighted. B. Selectivity of the 27 core object categories, separately for bottom-up categorization (left bars, darker shades) and top-down categorization (right bars, lighter shades). Category selectivity was quantified by the difference in correlation of semantic embedding vectors of concepts within each category as compared to the correlation with concepts outside of the category.[19]*

that the same proportion of the 27 higher-level categories had to be kept in both partitions. Regarding the training partition, 10 images were taken for each concept (1654 training object concepts x 10 images per concept = 16,540 training image conditions) while 1 image per concept was chosen for the test partition (200 test object concepts x 1 image per concept = 200 test image conditions).

5.1.3 Acquisition paradigm

In order to assess the brain's response to visual stimuli, 10 patients were selected. They were 8 females and 2 males of mean age 28.5 years (SD=4) in good health and with normal or corrected-to-normal vision. To study their response to visual stimuli, they were made to sit in front of a screen while wearing an EEG recorder cap with the standard 10–10 system (more details in Section 5.1.4). On the screen appeared the pictures described in Section 5.1.2 according to the Rapid Serial

Visual Presentation (RSVP) paradigm. This approach is very time-efficient and allows many visual stimuli to be shown in a short time. This, along with a simple target detection task, ensure the maximum attention from the subjects Every participant successfully completed 4 identical experimental sessions, yielding 10 datasets (1 per subject). Each dataset comprised 16,540 training images (10 per concept) repeated 4 times and 200 test images (1 per concept) repeated 80 times. In total, for every subject there are 82,160 trials.

Every session lasted 5 minutes and comprised 19 runs. In each of the first 4 runs 200 test image conditions were shown through 51 rapid serial sequences of 20 images, for a total of (4 test runs $\times$ 51 sequences per run $\times$ 20 images per sequence) = 4080 image trials. In each of the following 15 runs 8270 training image conditions (half of all the training image conditions, as different halves were shown on different sessions) were shown through 56 rapid serial sequences of 20 images, for a total of (15 training runs $\times$ 56 sequences per run $\times$ 20 images per sequence) = 16,800 image trials.

As shown in Figure 5.2 every rapid serial sequence started with 750 ms of blank screen, then each of the 20 images was presented for 100 ms and a stimulus onset asynchrony (SOA) of 200 ms, and it ended with another 750 ms of blank screen. After every rapid sequence there were up to 2s during which the participants were instructed to first blink (or make any other movement) and then report, with a key-press, whether the target (image of Buzz Lightyear) appeared in the sequence. This reduced the chances of eye blinks and other artifacts during the image presentations. Moreover a central bull's eye fixation target was present on the screen throughout the entire experiment to limit eye movements. The images were presented in a pseudo-randomized order, and the target image appeared in 6 sequences per run.

5.1.4 EEG raw data

The EEG data were recorded using a 64-channel EASYCAP and a Brainvision actiCHamp amplifier. The electrodes were arranged in accordance with the standard 10–10 system [38], an extension of the 10-20 system (more details in Section 5.1.4), and referenced to the Fz electrode. The data were sampled at the frequency of 1000Hz and online filtering was performed between 0.1Hz and 100Hz. All the raw EEG data were saved on different files with regard to subject, session and train/test partition (10 subjects x 4 sessions x 2 partitions).

5.1.5 EEG preprocessed data

The THINGS-EEG2 dataset not only provides the files with the raw data, but also provides a version of the data already preprocessed, the steps performed are

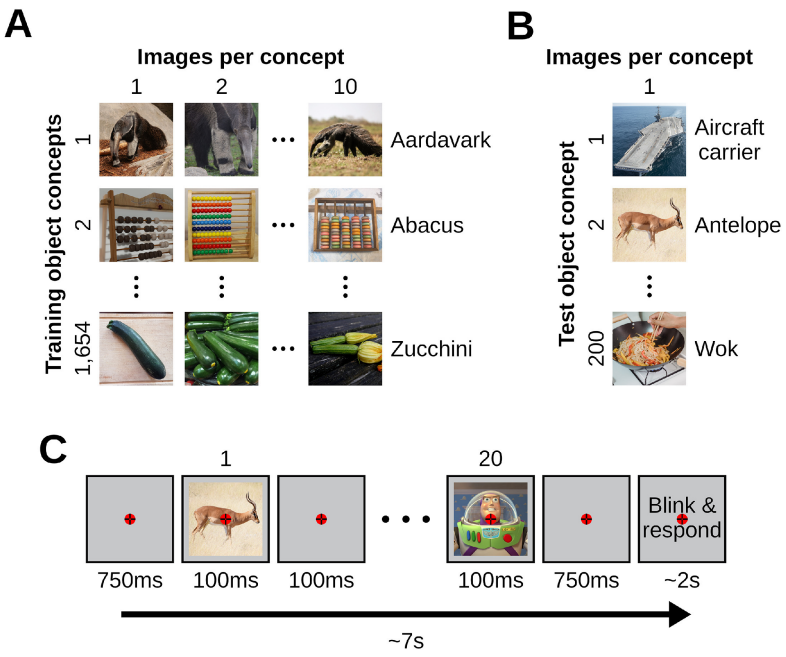

Figure 5.2: *Stimuli images and experimental paradigm. A. The training image partition (1654 object concepts x 10 images) B. The test image partition (200 object concepts x 1 image) C. Images were presented using a RSVP paradigm. Every sequence started with 750ms of blank screen, then each image was presented centrally for 100ms and a SOA of 200ms, and it ended with another 750ms of blank screen. After every rapid sequence there were up to 2s during which we instructed participants to first blink and then report, with a key-press, whether the target image appeared in the sequence. We asked participants to gaze at a central bull's eye fixation target present throughout the entire experiment.[10]*

the following. First of all, the continuous EEG data recording was divided into as many sections as the number of images shown. Every section range from 200 ms before stimulus onset to 800 ms after stimulus onset. Then only the 17 channels covering occipital e parietal cortex were considered, because are the one involved in visual perception. Thus, the channels considered are: O1, Oz, O2, PO7, PO3, POz, PO4, PO8, P7, P5, P3, P1, Pz, P2, P4, P6, P8.

Secondly, the signals were sub-sampled to 100 Hz and applied baseline correction by subtracting the mean of the pre-stimulus interval for each trial and channel separately. Then, to reduce the noise, multivariate noise normalization [16] was applied independently to the data of each recording session.

For each participant, the preprocessing resulted in a training data matrix of shape (16,540 training image conditions $\times$ 4 condition repetitions $\times$ 17 EEG channels $\times$ 100 time points) and a test data matrix of shape (200 test image conditions

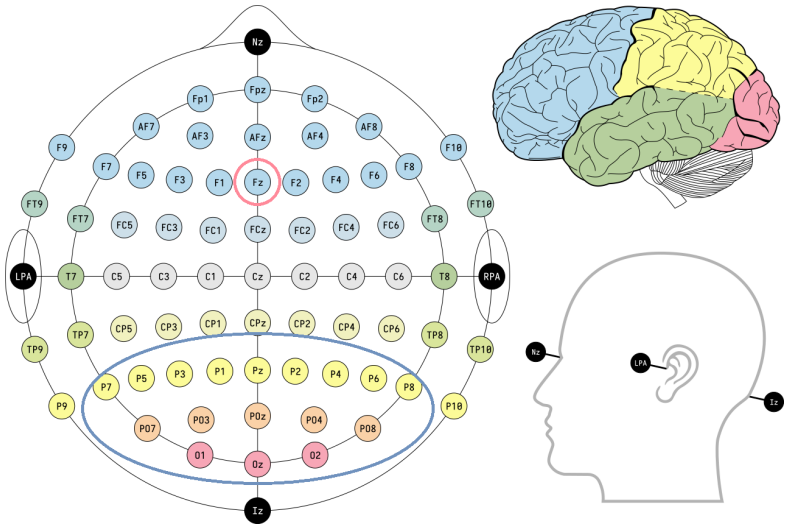

Figure 5.3: *The picture represent the 10-10 electrodes placement system used to record THINGS-EEG2 dataset. The red circle marks the reference electrode Fz used to define the channels while the blue circle marks the 17 channels considered in the pre-processing phase.*

× 80 condition repetitions × 17 EEG channels × 100 time points).

5.1.6 Other data

In addiction to the EEG data recorded during the vision of the images, the dataset provides EEG data recorded in resting state. In particular, for every subject, there are 5 minutes of recording at the beginning and at the end of the sessions where the subjects were instructed to fixate a central bull's eye fixation target presented on a gray background. During this period the were asked to blink as little as possible, and to refrain from other facial or bodily movements.

Moreover, the data relative to the target image detection task are provided. It's composed of 1 vector for the ground truth values and 1 vector for the responses. The results of this task resulted in a accuracy of 99.55%, which means that the image viewing time of 100ms was sufficient to distinguish their content.

5.2 Analysis tools

The project was developed entirely on Python and was run on a NVIDIA GeForce GTX 1650 Ti GPU. In addition, the following libraries were mainly used to perform the preprocessing and to realize the models:

- MNE Library [13]: package for exploring, visualizing, and analyzing human neurophysiological data;
- Pytorch [41]: machine learning library used for developing and training neural network based deep learning models;
- Scikit-Learn [42]: machine learning library built on top of SciPy

5.3 EEG preprocessing

Preprocessing of EEG data is a fundamental and critical step. EEG signals are valuable for understanding brain function, but they are inherently noisy and contain various artifacts that can distort the true neural signals. To overcome the low signal-to-noise ratio filtering and artifacts removal are required. As described in Section 5.1.5 the THINGS-EEG2 dataset provides a version of the data already preprocessed. However for this work a new data preprocessing based on neuroscience is proposed and compared with the original one in terms of performance.

The first step of this preprocessing is channel selection. As the original preprocessing only the 17 channels of the 64 channels, belonging to occipital (O) and posterior (P) area, are considered because are the one involved in visual perception. In particular the used channels are: O1, Oz, O2, PO7, PO3, POz, PO4, PO8, P7, P5, P3, P1, Pz, P2, P4, P6, P8. Secondly, epoching is performed in order to obtain the record related to every event. The epoching consider 500ms of recording starting at the time-locked event (i.e. from the instant of the appearance of the image) instead of 1000ms window starting from -200ms used proposed from the authors. Considering that each image is shown for 100ms followed by 100ms of black, the stimuli of multiple images are present in each epoch, as shown in Figure 5.4. In this step there is also the target event removal (images of Buzz Lightyear). Baseline correction is then performed for each of the sections based on the values from 0 to 200ms, duration of a complete image-blank cycle. Correction is applied to each epoch and channel individually in the following way: calculate the mean signal of the baseline period and subtract this mean from the entire epoch. Then frequency downsampling is performed in order to limit useless information. The frequency is reduced to 200Hz, instead of the 100Hz proposed by the dataset, to preserve signals over 50Hz belonging to gamma waves. The EEG data of each session, according to the image conditions, are then sorted and arranged in a matrix of shape: Image conditions $\times$ EEG repetitions $\times$ EEG channels $\times$ EEG time points.

Lastly, to reduce noise and enhance signal quality, Multivariate Noise Normalization (MVNN) [16] is applied independently to the data of each session. In the context of EEG, each electrode on the scalp represents a channel, and EEG

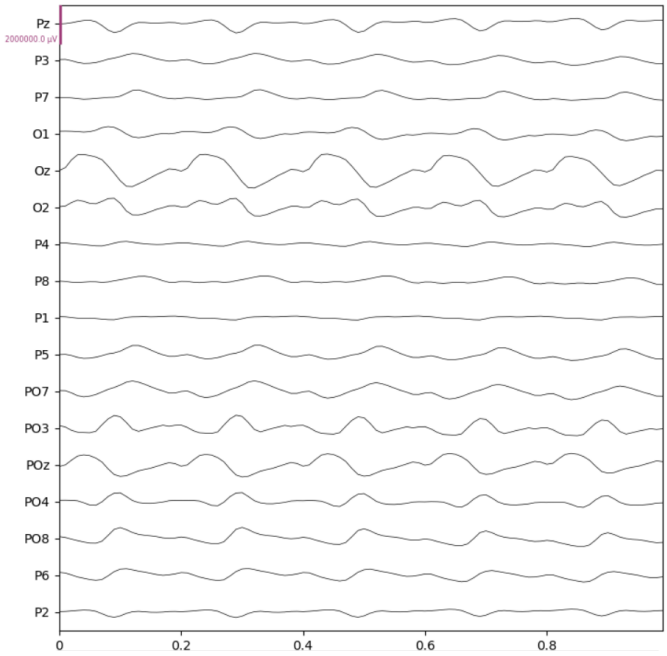

Figure 5.4: *EEG recording preprocessed according to [10]. For each trial, the recording window considered is 1 second. Since the visual stimuli are shown for 100ms every 200ms it is possible to see that in all 17 channels considered there are 5 peaks corresponding to the 5 visual stimuli present in each trial.*

data is inherently multivariate as it consists of recordings from multiple electrodes. To perform MVNN the covariance matrices of the EEG data (calculated for each time-point or epoch/repetitions of each image condition) is computed, and then its values are averaged across image conditions and data partitions. The inverse of the resulting averaged covariance matrix is used to whiten the EEG data (independently for each session).

As for the images, these were also transformed before they could be processed by the encoder. In fact, since it is a model pre-trained on another dataset, it is able to effectively analyze images with similar specifications. So firstly the images were resized using bilinear interpolation so that smaller edge of the image was equal to 256. Subsequently a central crop reduced the images shapes to squares of 224x224. Finally the values are first rescaled to [0.0, 1.0] and then normalized using mean=[0.485, 0.456, 0.406] and std=[0.229, 0.224, 0.225].

Once the EEG and image data has been preprocessed the division into training, validation and test set is performed. Also this phase was carried out differently from that proposed by the dataset. As described in detail in Section 5.1.2 every image in THING-EEG2 has a label defining the concept represented for a total of 1854 concepts (e.g., antelope, strawberry, t-shirt). THINGS-EEG2 pseudo-randomly di-

vided the 1854 THINGS object concepts into non-overlapping partitions under the constraint that the same proportion of the 27 higher-level categories had to be kept in both partitions: 1654 training and 200 test. However, to make it possible to train a classifier that exploits only the original concept labels, it is necessary for the same concepts to be present in both the training and test partitions. For this reason a new partition splitting was done using the data from the original training partition only. The original training partition contains 10 images for each of the 1654 concept that are shown 4 times each, resulting in 40 trials per concept. Therefore for each concept the new division has 8 images (32 trials) in the training set, 1 image (4 trials) in the validation set, and 1 image (4 trails) in the test set. This type of division makes it possible not to introduce bias into the evaluation partitions because all repetitions of the same stimulus belong to different sets. So in total there are 52928 trials in the new training set to classify 1654 concepts, with 6616 trials in both validation and test sets to assess performance.

5.4 Experimental classification results

This section reports the results of the two different types of classification. The comparisons of the different preprocessing, models and number of classes considered are presented.

5.4.1 Classification by concepts

To verify which is the best system for classifying EEG signals with respect to concepts, various tests are carried out. For all the models a common evaluation environment is defined. All models are optimized to predict the 1654 classes and the cross entropy loss is employed as the objective function. The metrics used for performance comparison are top-1 and top-5 accuracy. All the models are developed on Pytorch and best models are saved when the validation top-1 accuracy reaches a minimum in 100 epochs in the training process.

As a result from grid search the best training parameters, used for all models, are Learning Rate (LR) of 0.0001 with a scheduler function that reduce LR of a factor 0.5 after 7 epochs if no improvements in the validation loss. In addition, the dataset with EEG-label pairs is processed in batches of 64 items and the optimizer used is Adam, a stochastic gradient descent method that is based on adaptive estimation of first-order and second-order moments. The two beta values for this optimizer are set at 0.0002 and 0.5.

Regarding the model parameters given in Section 4.2.1, the values are as follows. For LSTM model 4.1 the parameters chosen for the layers are respectively 32 and 64 for k1 and k2, 2, 10 and 5 for m1, m2 and m3, 2, 1, and 5 for s1, s2

and s3. For Temporal and Spatial CONV2D model 4.2 the parameters chosen for the first layer are k=40, m1=25 and s1=1, while for the average pooling layer are m2=51 and s2=5. For EEGNet model 4.3 the original formulation propose k1=8 and m1=64 for the first layer, k2=16 for the second layer, m2=4 and s1=4 for the 2 average pooling layers, m3=16 for the third convolutional layer, k3=16 for the fouth. For Spectral model 4.4 k1 and m1 for the first layer are set to 32 and 3, m2 and s1 are both set to 2, k2 and m3 are set to 64 and 3, while the last kernel is set to N/s1. The spectrograms computed are composed of 12 frequency bins from 0Hz to 100Hz and 10 temporal bins every 50ms for every channel.

EEG preprocessing comparison

First, it is evaluated whether the developed preprocessing, different from the one proposed from dataset authors, brings some benefits in classification. To do this, a subset of 100 concepts is selected and the classification with the 4 models is tested averaging the results of 3 sessions per model. The results of the comparison are shown in Table 5.1.

	LSTM		TSconv		EEGNet		Spectr	
	Top1	Top5	Top1	Top5	Top1	Top5	Top1	Top5
THING-EEG2 preprocessing	5.64	16.84	5.56	16.23	7.38	23.61	**2.43**	7.78
Proposed preprocessing	**7.64**	**23.96**	**9.20**	**24.65**	**9.11**	**25.19**	1.99	**8.51**

Table 5.1: *This table contains the top-1 and top-5 accuracy values obtained classifying EEG signals by 100 concepts with 4 models. The first row shows the accuracies obtained using the preprocessing proposed by [10] that considers a time window of 1000ms and a frequency of 100Hz. The second shows the values obtained using the proposed preprocessing that considers 500ms and a frequency of 200Hz.*

As can be seen from Table 5.1, halving the observation window and doubling the sampling frequency leads to an increase in accuracy for almost all models. In particular, for the value of top-1 accuracy there is a 35% increase for LSTM model, 65% increase for Temporal and Spatial convolution model, 23% increase for EEGNet model and 18% reduction for Spectral model. These improvements may be due to a greater resolution of the signal with the consequent possibility of evaluating brain waves at higher frequencies. Furthermore, the reduction of the time window leads to a greater attention of the models on the signal immediately after the stimulus. The only model negatively affected by the proposed preprocessing is the model that analyzes the signal spectrogram. However, even without preprocessing, this model performs poorly compared to the others. In this project then, EEG signals will be preprocessed according to the proposed method.

Subjects comparison

As described in Section 2.3.3 the EEG signals can vary greatly from subject to subject and from session to session. Therefore, the models in this project are all trained and tested on individual subjects. A comparison of the classification accuracy of 100 classes for different subjects and different models is given in Table 5.2.

	Subj. 1		Subj. 2		Subj. 3		Subj. 4		Subj. 5	
	Top1	Top5	Top1	Top5	Top1	Top5	Top1	Top5	Top1	Top5
LSTM	7.64	23.96	5.21	14.76	**8.07**	23.70	5.04	15.62	4.60	13.63
TSconv	**9.20**	24.65	**7.55**	**19.70**	7.03	23.35	**6.68**	**23.18**	**8.07**	**23.18**
EEGNet	9.11	**25.19**	6.25	19.01	7.99	**24.48**	6.51	20.05	5.82	17.80
Spectr	1.99	8.51	2.95	8.25	2.69	7.98	1.91	6.16	3.56	9.81

Table 5.2: *This table contains the top-1 and top-5 accuracy values obtained classifying EEG signals by 100 concepts with 4 models and 5 subjects.*

Table 5.2 shows how each subject exhibits different but consistent accuracy values across the 4 models. In fact, the EEG signals of subjects 1 and 3 appear to be the most interpretable by all models. As for the model comparison, for the 100-class classification, the Temporal Spatial convolutional model is the one that consistently provides higher accuracies. In this project then, for computational reasons, only EEG signals from subject 1 are considered.

Subset size comparison

The performance of a classifier intrinsically depends on the number of classes considered. In previous comparisons we were limited to evaluating a subset of 100 classes. In this section, different subset sizes are analyzed to identify under which circumstances the models classify best.

Overall, the accuracy values obtained are quite low for all subsets considered but still higher than chance. In fact, the models achieve higher accuracy than chance by a factor of 4/5 with a few classes and by a factor of 15/16 with many classes. As can be seen from Table 5.3, the models perform differently with respect to the quantity of classes. The Temporal Spatial convolutional model obtains higher performance in the range 20-100, the EEGNet model in the range 50-200 while the LSTM model obtains the best results when all classes are considered. The Spectral model, on the other hand, struggles the most to extract information from EEG signals.

One thing to take into consideration with these models is that, as the classes increase, the number of trainable parameters in each model also increases. This is

	20 classes		50 classes		100 classes		200 classes		1654 classes	
	Top1	Top5	Top1	Top5	Top1	Top5	Top1	Top5	Top1	Top5
LSTM	16.14	45.84	11.46	30.21	7.64	23.96	4.47	12.98	**0.94**	**3.01**
TSconv	**17.19**	**48.44**	10.59	33.51	**9.20**	24.65	4.65	12.80	0.81	2.59
EEGNet	14.58	48.43	**11.81**	**37.33**	9.11	**25.19**	**5.08**	**14.37**	0.62	1.94
Spectr	14.58	34.90	5.52	18.05	1.99	8.51	1.39	4.47	0.16	0.41

Table 5.3: *This table contains the top-1 and top-5 accuracy values obtained classifying EEG signals by different subsets of concepts with 4 models. On columns a subset of 20, 50, 100, 200 and 1654 concepts is considered with corresponding top-1 chance accuracy of 5%, 2%, 1%, 0.5% and 0.06%, respectively.*

due to the change in size of the last dense layer. Table 5.4 shows the number of trainable parameters for each model as the subset of classes varies.

	20 classes	50 classes	100 classes	200 classes	1654 classes
LSTM	39552	58752	90752	154752	1085312
TSCONV	33160	48360	52360	76360	425320
EEGNet	9056	20576	39776	78176	636512
SPECTR	24608	26528	29728	36128	129184

Table 5.4: *This table contains the number of trainable parameters for every model according to the number of cconcepts that are classified. The number varies because of the variable size of the final linear layer.*

From Table 5.4 can be seen how each model has a different number of parameters and how these increase differently as the number of classes increases. The growth rate is due to the number of values in output from the second last layer and the output of the linear layer, which correspond to the number of classes. Furthermore, from this table and Table 5.3 it can be seen that in the case of the 1654 concepts classification there is a correlation between the accuracy values obtained and the number of parameters that compose the model.

5.4.2 Classification by categories

To verify which is the best system for classifying EEG signals with respect to categories a common evaluation environment is defined. All models are optimized to predict the 27 high-level categories and binary cross entropy loss with logits is employed as objective function. The metrics used for performance comparison are precision, recall, F1 and accuracy scores are evaluated. All the models are developed on Pytorch and best models are saved when the validation F1 score

reaches a minimum in 100 epochs in the training process. As a result from grid search the best training parameters, used for all models, are Learning Rate (LR) of 0.0001 with a scheduler function that reduce LR of a factor 0.5 after 7 epochs if no improvements in the validation loss. In addition, the dataset with EEG-label pairs is processed in batches of 64 items and the optimizer used is Adam. The two beta values for this optimizer are set at 0.0002 and 0.5. Regarding the model parameters given in Section 4.2.1, the values are as reported in Section 5.4.1.

Models comparison

Evaluating the performance of a Multi-Label classifier in usually less intuitive than a Single-Label classifier. There are several metrics that can be evaluated, and the choice of which to maximize depends on the practical case. In this project, as a balance between precision and recall metrics is pursued, the F1 score is used to select the model. Since the categories in this dataset are composed of different numbers of concepts, and consequently different numbers of trials, it is necessary to weight the categories differently. If this were not done the less represented categories would get a lower recall value. For this purpose 27 weights are calculated using the formula:

$$w_i = \frac{N_{TOT}}{n_i}$$

where n_i is the number of concepts for i-th category while N_{TOT} is the total number of concepts. These weights are thus used by the objective function where a higher value of w_i increase the recall of that class.

Furthermore, to calculate the metrics, since the predicted values processed by the sigmoid are continuous values in the range [0,1], it is necessary to define a threshold to be used to binarize the values. For the case of TSconv and EEGNet classifiers this is set to 0.95, for the LSTM network to 0.1 and for Spectr to 0.5. These values are not standard and depend on the activation functions in the model. In this case the choice is driven by attempting to maximize the F1 metric. Table 5.5 collects the metrics averaged over 3 training runs of the 4 models.

	F1	Precision	Recall	Accuracy
LSTM	3.66	2.92	4.97	20.33
TSCONV	8.78	7.51	**11.01**	19.43
EEGNet	**9.37**	**8.31**	10.75	**20.82**
SPECTR	5.68	4.71	7.35	19.07

Table 5.5: *This table contains the F1, Precision, Recall and Accuracy values obtained classifying EEG signals by 27 categories with 4 models.*

As can be seen from Table 5.5, again the EEGNet and Temporal Spatial convolutional models are the ones that allow more information to be extracted from the EEG signal. In fact, these return the highest values of the F1 metric and consequently also of precision and recall. In this case there is no correlation between the metrics and the number of trainable parameters of the model since the LSTM model has 44032 while the EEGNet model has 11744.

Once again, the results obtained are not very high despite the much lower number of classes compared to the previous task. This may be due to: a higher semantic complexity of the task and to the imbalance of classes. Regarding the first point, the classification is based on much higher level classes that might not necessarily be found in EEG signals. In fact, the high frequency of the RSVP protocol reduces the time for the brain to process such information. In addition, some of the 27 categories include a wide range of concepts with high intra-class diversity, e.g. "tool", "part of car", "container" (the complete list can be seen in Table B.1). As for the second point, in Table B.1 class-by-class metrics are presented for the EEGNet model, which reported an average F1 value of 9,47. From here it is possible to see that the recall for minority classes is generally lower. This is an inherent limitation of unbalanced datasets and it shows that the weights used were not sufficient to overcome the problem. The definitive solution would be dataset balancing but it is hampered by the fact that it is a multi-class problem and each concept can be assigned to multiple categories, therefore, it is not easy to identify which concept to undersample or oversample.

5.5 Experimental image retrieval results

In addition to classification, this project wants to investigate the possibility of extracting more information from EEG signals and obtaining an image as similar as possible to the seen one. As explained in the methods Section 4.3, obtaining the image can be done in different ways: through generation or through retrieval. This project focuses on evaluating the potential of obtaining images by searching for the most similar within a dataset. However, for both the approaches the key step is to obtain: a representation for both EEG signals and images and a representation mapping model.

5.5.1 Feature extraction

In order to convert from a set of time series in matrix form C x T (C=17 EEG channels, T=100 time instants) to an image consisting of 224 x 224 pixels, it is necessary for both to find a latent representation. A latent representation is a simplified representation of the input data that still contains information about it.

Image encoder

To extract semantic information from images a pre-trained model is used. The model chosen is the ResNet18 model pre-trained on ImageNet since it offers good performance and a small number of features in latent space. This model in fact, by removing the dense output layer, produces a vector of 512 values per image in range [0,+inf]. Then the number of features per signal is further reduced by performing PCA. The value of these features after this procedure is in the range [-inf,+inf]. PCA is fitted on the training data partition and then is extended to the other two partitions.

To define the minimum number of features to be kept without having excessive loss of information, the classification of image features with respect to the concepts they represent is performed. In this way it is possible to evaluate how the number of features affects the classification accuracy. A model consisting of a single linear layer is used for classification with input dimension the number of components considered and in output all 1654 concepts. For training, all image-label pairs are processed in batches of 32 items and is run for 30 epochs. The loss function is cross entropy and Adam with an LR of 0.0005 is used as the optimizer. Top-1 and top-5 accuracy metrics are evaluated in the classification. The results of this analysis are shown in Table 5.6.

N. components	Top1	Top5
512	67.48	86.76
200	66.34	86.57
100	63.64	85.15
50	58.19	82.00

Table 5.6: *This table contains top-1 and top-5 accuracy values obtained classifying image features by concepts with a viable number of features. The features are extracted with ResNet18 model, pre-trained on Image-Net, that provides 512 values per image. The number of component are then reduced using PCA.*

The first thing that can be noticed is that the accuracies are high and very close to the value of 69.8% which is the one obtained in the pretrain phase on 1000 ImageNet classes, so fine tuning is not performed. As is to be expected as the components decrease also the information contained in the features decreases. However, the decrease in accuracy at considering 200 components is only one percentage point, at 100 components it is 4% while at 50 it is 9%. It is therefore chosen to consider 100 components as a good compromise between parsimony and accuracy.

As mentioned earlier, the values processed by the network and transformed

with PCA belong to the range of values [-inf,+inf]. Various techniques for feature scaling are tested below, the results are shown in Table 5.7.

Feature scaling	Range val.	Top1	Top5
None	[-inf,+inf]	63.64	85.15
Z-score 4.1	[-inf,+inf]	64.97	85.64
Max-min scal. 4.2	[0,1]	17.73	35.02
Mean scal. 4.3	[-1,1]	58.94	82.72

Table 5.7: *This table contains top-1 and top-5 accuracy values obtained classifying image features by concepts with different feature scaling. The features are extracted with ResNet18 model, pre-trained on Image-Net, that provides 512 values and only 100 of these are selected with PCA. The range values column shows the range in which the numerical values of the features are scaled.*

As the table shows, the type of scaling used greatly influences classification accuracy. These discrepancies are most likely given by the optimization algorithm and the initialization values of the weights in the models. It can be seen that standardization leads to a slight increase in classification accuracy but still remains unbounded. On the other hand, as far as scaling in a finite range of values is concerned, the mean scaling method is definitely the one that preforms best. The latter method is chosen because it simplifies the mapper's training without losing much in accuracy.

EEG encoder

To extract semantic information from an EEG signal a model that is trained to process EEGs with similar characteristics (200hz frequency, 0.5 seconds long, and 17 specific channels) is required. Given the high specificity of the problem, the model used is the same one used for signal classification. This in fact, because it can be pre-trained on the classification task, has weights suitable for extracting semantic information from these EEG signals. Different levels of abstraction occur in the different layers of the model, so it is possible to truncate the network at each point to obtain a reduced semantic representation of the signal. Among the 4 models, EEGNet is selected because it is the one that exhibits good consistency in the different classification tasks.

Compared with the original architecture presented in Table 4.3, the model is truncated after the flattening function before the final dense layer. What results from processing the data with this model is an array of 240 values with possible values in the range [-1,+inf] because of the ELU activation function. Processing the three data partitions defined in Section 5.3, features with mean, max and min

values equal to 0.21, 232.62 and -1.0 respectively are obtained. Then the number of features per signal is further reduced by performing PCA. This is fitted on the training data partition and then is extended to the other two partitions. To match the number of values in the image features 100 components, out of 240, are preserved.

5.5.2 Alignment

Once the representations for both objects of the EEG-image pair have been obtained, it is necessary to align the spaces of these two representations. The model consisting of a single linear layer is used to align the 100 EEG features with the 100 image features. Training is performed on pairs belonging to 100 concepts by processing batches of 128 elements for 100 epochs. Adam with LR of 0.001 is used as the optimizer.

The two loss functions are tested to train the regressor and the results the results obtained in terms of RMSE on the validation set are 1.03 for MSE loss and 0.76 for CLIP loss. However, these values are not very indicative to evaluate the goodness of alignment, so it is necessary to evaluate the performance in the actual task.

5.5.3 Image retrieval

Data structure

For the realization of the data structure that can be queried for the most similar image, a subset of the THINGS dataset [19] containing the first 100 classes is chosen as the dataset. This dataset is the one on which THINGS-EEG2 [10] is based (more details in Section 5.1) and contains about 15 images per concept for a total of 1535 images. So it contains the same classes and even some of the same images used in the encoder and mapper training. This choice was made to allow quantitative metrics on retrieval to be evaluated. The images used, in order to be in a form similar to that obtained from the EEG mapping process, are processed from the encoder, subjected to PCA and scaled (same procedure as described in Section 5.5.1).

The KDTree data structure is then realized using the euclidean metric and a leaf size value of 40. Thus the structure obtained allows to return the k most similar images. For this project a value of k=10 is chosen which allows to obtain a good pool of similar images.

Loss functions comparison

In this section the quality of the retrieval is evaluated. Specifically, considering that the dataset forming the tree includes both the stimulus image and other images from the same category, we assess the accuracy of find these images among the retrieved 10. The outcomes are documented in Table 5.8. The comparison is between the two loss functions used and the case of random extraction, obtained from the combinatorics, is also reported.

	Same image accuracy	Same class accuracy
Chance	0.65	9.38
MSE loss	0.81	10.02
CLIP loss	2.13	14.11

Table 5.8: *This table contains accuracy of image retrieval from EEG signals. The images are retrieved from a dataset of 1535 images which contains the same image concepts on which the encoder and mapper were trained. In addition, the 200 images shown in validation are present within the dataset. Ten images are then extracted for each EEG signal, and the values reported show the accuracy of extracting the exact image or an image of the same concept.*

The results obtained with the two loss functions are quite different. The function that uses a contrastive approach allows to obtain a more accurate retrieval both in extracting the exact image and in extracting the exact class. However, the results are still generally quite low, the limiting factor is once again the encoder of the EEG signal which does not allow much semantic information to be extracted from the EEG signals.

Chapter 6

Conclusion and future work

Researchers have worked on decoding human brains for a long time allowing us to deepen our understanding of the complex mechanisms that drive it using EEG and fMRI. In particular, in recent years, some efforts have been put into interpret visual perception, the complex mechanism that allows us to perceive and make sense of the light signals that reach our eyes. Some studies have achieved good results exploiting fMRI imaging technique to classify and reconstruct the visual stimulus [21] [48]. However, this instrument has limitations as it is very expensive, slow and has poor temporal resolution. For this reason, many studies [55] [28] [59] [39] [24] [29] in the past 5 years have focused on the analysis of EEG signals, a viable alternative with lower cost and higher time resolution. However, much of these papers turned out to be compromised because based on a flawed dataset that inflated their results [33]. So in 2022 a new large-scale EEG dataset, collected with Rapid Serial Visual Presentation RSVP technique, was proposed to model human visual recognition and decode objects pairwisely [10].

This entire project is therefore based on this new dataset composed of more than 66,000 EEG recordings of subjects observing 16,000 natural images cataloged into 1854 concepts and 27 categories. At the beginning of the development of this project, no paper had used it yet or had ever attempted to classify such a high number of classes. In this project an extra step was then taken to try to reproduce the image that generated a certain stimulus. To do so, it was decided to consider the possibility of obtaining the stimulus seen performing image retrieval in a images dataset arranged in a tree structure. Also this approach had never been attempted in literature.

6.1 Summary of key findings

6.1.1 EEG classification

This project first evaluated the classification of the signals with respect to image concepts and categories. To accomplish this 4 deep learning models were tested, 1 of these is a well-known model, already used for other BCI tasks, while the other 3 were developed for this project. The well-known model is the EEGNet [32], and it processes raw EEG signals by exploiting multiple levels of convolution. The first proposed model is a network composed of an LSTM layer, to analyze the temporal component, and a convolutional layer, for the spatial component of the electrodes. The second model, called Temporal Spatial convolution, treats the signal as a two-dimensional matrix and applies two convolution layers for the temporal and spatial part. The third model, on the other hand, consists of two convolutional layers to process spectrograms of 12 frequency intervals for 10 time intervals, obtained for every EEG channel. All of these 4 models were provided with a linear output layer to allow classification of both the 1654 concepts using a softmax function and for the 27 categories using a sigmoid function.

Classification by concepts

The experiments regarding concept classification were aimed at comparing: 2 different types of data preprocessing, 5 different subjects, 4 models, and different subsets of concepts.

Regarding the preprocessing comparison, it was evaluated whether it is possible to improve the preprocessing proposed by the authors of the dataset. To accomplish this the 4 models, in classification of 100 concepts, were used to compare preprocessing. It can be seen from the accuracy results that the new preprocessing procedure, that consider a higher sampling frequency and a shorter time window, leads to improvements of up to 60% in top-1 accuracy. These improvement may be due to a greater resolution of the signal with the consequent possibility of evaluating brain waves at higher frequencies. Furthermore, the reduction of the time window leads to a greater attention of the models on the signal immediately after the stimulus.

In terms of inter-subject comparison, important differences in accuracy between subjects were noted. This is a standard behaviour of EEG signals since they are very personal and can even vary from session to session due to possible differences in electrodes placement. This is also corroborated by the fact that the results among all models are in agreement. Given this characteristic, therefore, in the project it was decided to consider only subject 1 and optimize the metrics for him. This is a widely used approach in this field.

To compare the models, a classification of subsets with different sizes was made in order to assess the best prediction range for each model. Classification of 20, 50, 100, 200, and 1654 concepts was then performed, and top-1 and top-5 metrics were evaluated for each model. With the best model for each subset resulted top-1 accuracies of 17,2%, 11,8%, 9.2%, 5.1% and 0.94%. These correspond to improvements from 4/5 times over chance for a few classes up to 15/16 times with all 1654 classes. The highest accuracies for the classification of all classes were obtained with the LSTM model, which is also the one with the most trainable parameters, while in the cases with fewer classes the best model turned out to be EEGNet and Temporal Spatial convolution. Thus, the models have been shown to be able to capture class information from EEG signals even if somewhat limited.

Classifying such a large number of classes is an inherently very complicated task. For this type of signals it is also particularly complex since the signal-to-noise ratio is very low given the limitations of this technique and given the RSVP paradigm. Having in fact very short and close stimuli there is a lot of spurious signal information since each 500ms trial associated with an image actually contains the projection of at least 3 distinct images. Furthermore, classification is further hampered by the fact that there are structural differences in the signal depending on where the signal is located in the presentation block. Signals at the beginning of the block contain no influences from previous images while those at the end of the block contain only the stimuli from 1 or 2 images.

Comparing the results of the classification with the results of recent papers [8] [54], posthumous to the development of this project, published on this dataset shows that their multimodal approach that includes text extraction from the image is the most robust one and allows not only to be able to handle a high number of classes but to do so in a zero-shot fashion.

Classification by category

The 3 models developed and the benchmark model were also used for classification with respect to high-level categories of images. To perform this type of Multi-Label Multiclass classification binary cross entropy was evaluated so that each category was considered as a binary realization.

The metrics obtained by classifying this dataset were fairly balanced between recall ability and precision, thanks in part to the weights used to handle class imbalance. In this case, the model that generally performed best is the EEGNet, which allowed for a precision of 8.3% and recall of 10.7%, which sum up to an F1 value of 9.4% while the accuracy is 20.8%. Also in this case, the results are not very high. The reasons again are similar to those for classification by concepts. Howerver, this may also be due to a higher semantic complexity of the task. In fact this classification is based on much higher level classes that might not necessarily

be found in EEG signals because of the RSVP protocol. Its high frequency reduces the time for the brain to process such a high-level information. In addition, some of the 27 categories include a wide range of concepts with high intra-class diversity, e.g. "tool", "part of car", "container" (the complete list can be seen in Table B.1). An additional cause can also be the presence of an imbalance in the dataset that even the calculated weights can't overcome. However, a possible correction of the distribution is difficult since each concept can belong to more than one class.

6.1.2 Image retrieval

As for the image retrieval experiment from EEG signals, several steps were necessary. First, feature extraction of both images and EEG signals was performed. The first was done with the pre-trained ResNet18 model, whose 512 output values for each image were reduced to 100 with PCA and scaled to a range of values [-1,1]. The quality of these representations was evaluated with a classification task with respect to image concepts obtaining a top-1 accuracy value of 58.94 and top-5 accuracy value of 82.72. Instead, feature extraction for EEG signals was performed with the EEGNet pre-trained in the concept classification task. Features were extracted before the last linear layer and reduced from 240 to 100 with PCA. These representations were then aligned with a linear model that was trained with two loss functions: MSE loss and CLIP loss. The latter is the one that returns the lowest RMSE value of 0.76 while the MSE loss a value of 1.03.

As for the actual image retrieval step, a tree data structure was first constructed using 1535 images, duly pre-processed, belonging to the same concepts on which the encoder and mapper were trained. Then this was queried with the features predicted from the EEG signals to extract top 10 images most similar to the one shown. The accuracy of this process was evaluated by assessing the number of times the exact image with which the stimulus was obtained or images belonging to the same class were returned in the pool of 10 images. The results obtained are somewhat higher than the random extraction corresponding to 0.65% for the exact image and 9.38% for the exact class. Specifically, what was obtained is 0.81% and 10.02% with the MSE loss function and 2.13% and 14.11% with the CLIP loss function. Also for this task, the results are quite low and mainly due to the poor ability of the EEG encoder to extract semantic information from these signals.

However, in conclusion, this project assessed the three key intuitions declared in Chapter 4: EEG signals convey feature-level and cognitive-level information about the image content, features can be extracted from EEG signals and image descriptors can be extracted from them. However the accuracies obtained are quite limited mainly due to the paradigm used. In fact, the amount of data needed to allow good classification and retrieval accuracy with this rigid type of approach, that includes only EEGs and images, is very high and infeasible

to collect. Therefore, an approach that integrates textual information would be preferable to this. In this way it is possible to soften the rigid constraints of a traditional classification task and also allow for greater generalization with zero-shot classification.

6.2 Original contributions of the thesis

This project was developed before any paper based on this dataset was published. Thus, it is a pioneering work that attempted for the first time to analyze a dataset of EEG signals in response to natural imagery stimuli composed of a very large number of classes. In particular, until before the two THINGS-EEG datasets were published [15] [10], the RSVP paradigm was mainly used to investigate the detection of the target image among a set of non-target images. In this case, however, the detection task is secondary to the image observation task and is only exploited to keep subjects attention high while being shown a very high number of images per session. In this project, 3 original deep learning models were developed that allowed the classification of a very high amount of EEG signal classes with accuracies about 15 times higher than chance. This project also explored for the first time in the literature the possibility of performing image retrieval from EEG signals. This approach, despite the low accuracies obtained in this project, has the potential to be a valuable tool, e.i. to assist patients with difficulties communicating due to psychological trauma or disability. It also overcomes complexity, artifacts, and expense of the most commonly used generative methods.

6.3 Future developments

This project lends itself to many future developments, as there are relatively few publications in this area. The first possible development to improve the results of this project, which has been explored in part in the papers [8] and [54], is to extract and exploit textual information from the images to achieve better feature allignment and enable zero-shot classification. Indeed, an alignment of EEG, image and text representations could be achieved by increasing their semantic content. This would lead to an improvement of both the classification task and the retrieval task.

Another possible area of interesting research could be to evaluate how to improve the encoder of the EEG signal. In particular, this could be done by more thoroughly considering the 3D arrangement of electrodes on the skull. By constructing the arrangement graph considering the electrodes as nodes and the distance between them in the 3 dimensions as arcs it becomes possible to better

evaluate the correlation between channels and make the models more robust to slight electrode misplacements. Consequently, however, it becomes necessary to develop models that can handle this data structure, and it is therefore necessary, for example, to use convolutional layers that can operate on the graphs. This type of analysis has already been tested on EEG signals but not in this field.

Another future development, which is not so much addressed in the literature, is the study of the different meanings of similarity in generation or retrieval. In fact, much work is based on obtaining semantically similar images to the source images, trying to reconstruct the scene with the main subjects and not evaluating their location and size. As for the concept of visual similarity, i.e., obtaining an image with figures and shapes placed correctly in space, only the paper [29] addressed it by using a GAN to recreate the saliency image with respect to the original image. However, there is still much room for improvement by attempting, for example, to recreate image object masking or segmentation map. This could be used to enhance the generation of the real scene. Visual similarity can also be investigated using other types of encoders not trained on classification tasks, since space-invariant, but with autoencoders.

Appendix A

Name (University)	Year	Specs	Limitations
University of Catania [55]	2017	6 subj., 40 classes, 128 ch.	Single block paradigm
Gdansk University of Technology [23]	2018	10 subj., 45 images, 14 ch.	Commercial EEG recorder
MindBigData ImageNet [61]	2018	1 subj., 14000 images, 5 ch.	Commercial EEG recorder
Sidney University [14]	2018	x subj., 50 classes, n ch.	No natual images
Bucharest University [6]	2019	6 subj., 6 classes, 14 ch.	Single block paradigm
Purdue University [1]	2020	1 subj., 40 classes, 96 ch.	NA
MindBigData MNIST [58]	2021	1 subj., 10 numbers, 64 ch.	Not natural images
THINGS-EEG [15]	2021	50 subj., 1854 concepts, 64 ch.	NA
THINGS-EEG2 [10]	2021	10 subj., 1854 concepts, 64 ch.	NA

Table A.1: *Most of the public available EEG datasets collected to assess visual perception. For each dataset, the presentation paper, characteristics of the collected signals and possible application limitations for this specific project are given. The dataset selected for this project is the last one.*

Appendix B

	F1	Precision	Recall	N concepts
animal	32.33233	36.29213	29.15162	177
bird	7.272727	6.756757	7.874016	27
body part	19.05405	11.66253	52.02952	34
clothing	22.48829	18.16653	29.5082	108
clothing accessory	5.340454	4.123711	7.575758	38
container	6.940639	11.875	4.903226	105
dessert	4.819277	5.025126	4.62963	37
drink	3.212851	3.508772	2.962963	19
electronic device	10.04785	8.467742	12.35294	74
food	28.85797	38.44828	23.09684	295
fruit	7.260726	5.612245	10.28037	34
furniture	7.903403	5.564142	13.63636	39
home decor	8.180536	7.591623	8.868502	45
insect	1.652893	3.030303	1.136364	17
kitchen appliance	2.711864	2.614379	2.816901	20
kitchen tool	NA	0	0	27
medical equipment	5.629139	4.038005	9.289617	27
musical instrument	2.762431	3.597122	2.242152	33
office supply	6.530612	4.347826	13.11475	25
part of car	4.395604	3.317536	6.511628	30
plant	12.91364	8.705114	25	47
sports equipment	4.018547	5.627706	3.125	64
tool	4.604052	6.493506	3.566334	107
toy	3.571429	5.357143	2.678571	34
vegetable	6.18047	4.725898	8.928571	42
vehicle	7.721046	7.828283	7.616708	70
weapon	1.126761	3.846154	0.660066	48

Table B.1: *This table reports the metrics of EEG signal classification by 27 image category. For every category the number of concepts included, F1, Precison and Recall are reported. The average value of these metrics is F1=9,473, Precision=8,393, Recall=10,872.*

Bibliography

[1] Hamad Ahmed et al. "Object classification from randomized EEG trials". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 3845–3854.

[2] H Aurlien et al. "EEG background activity described by a large computerized database". In: *Clinical Neurophysiology* 115.3 (2004), pp. 665–673.

[3] George EP Box et al. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[4] Thomas A Carlson et al. "High temporal resolution decoding of object position and category". In: *Journal of vision* 11.10 (2011), pp. 9–9.

[5] Alexander Craik, Yongtian He, and Jose L Contreras-Vidal. "Deep learning for electroencephalogram (EEG) classification tasks: a review". In: *Journal of neural engineering* 16.3 (2019), p. 031001.

[6] Nicolae Cudlenco, Nirvana Popescu, and Marius Leordeanu. "Reading into the mind's eye: Boosting automatic visual recognition with EEG signals". In: *Neurocomputing* 386 (2020), pp. 281–292.

[7] Mohammad Reza Daliri, Mitra Taghizadeh, and Kavous Salehzadeh Niksirat. "EEG signature of object categorization from event-related potentials". In: *Journal of medical signals and sensors* 3.1 (2013), p. 37.

[8] Changde Du et al. "Decoding visual neural representations by multimodal learning of brain-visual-linguistic features". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023).

[9] Ahmed Fares, Sheng-hua Zhong, and Jianmin Jiang. "EEG-based image classification via a region-level stacked bi-directional deep learning framework". In: *BMC medical informatics and decision making* 19.6 (2019), pp. 1–11.

[10] Alessandro T Gifford et al. "A large and rich EEG dataset for modeling human visual object recognition". In: *NeuroImage* 264 (2022), p. 119754.

[11] Shu Gong et al. “Deep learning in EEG: Advance of the last ten-year critical period”. In: *IEEE Transactions on Cognitive and Developmental Systems* 14.2 (2021), pp. 348–365.

[12] Melvyn A Goodale and A David Milner. “Separate visual pathways for perception and action”. In: *Trends in neurosciences* 15.1 (1992), pp. 20–25.

[13] Alexandre Gramfort et al. “MEG and EEG Data Analysis with MNE-Python”. In: *Frontiers in Neuroscience* 7.267 (2013), pp. 1–13. DOI: 10.3389/fnins.2013.00267.

[14] Tijl Grootswagers, Amanda K Robinson, and Thomas A Carlson. “The representational dynamics of visual objects in rapid serial visual processing streams”. In: *NeuroImage* 188 (2019), pp. 668–679.

[15] Tijl Grootswagers et al. “Human EEG recordings for 1,854 concepts presented in rapid serial visual presentation streams”. In: *Scientific Data* 9.1 (2022), p. 3.

[16] Matthias Guggenmos, Philipp Sterzer, and Radoslaw Martin Cichy. “Multivariate pattern analysis for MEG: A comparison of dissimilarity measures”. In: *Neuroimage* 173 (2018), pp. 434–447.

[17] John-Dylan Haynes and Geraint Rees. “Decoding mental states from brain activity in humans”. In: *Nature reviews neuroscience* 7.7 (2006), pp. 523–534.

[18] Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[19] Martin N Hebart et al. “THINGS: A database of 1,854 object concepts and more than 26,000 naturalistic object images”. In: *PloS one* 14.10 (2019), e0223792.

[20] Suzana Herculano-Houzel. “The remarkable, yet not extraordinary, human brain as a scaled-up primate brain and its associated cost”. In: *Proceedings of the National Academy of Sciences* 109.supplement_1 (2012), pp. 10661–10668.

[21] Tomoyasu Horikawa and Yukiyasu Kamitani. “Generic decoding of seen and imagined objects using hierarchical visual features”. In: *Nature communications* 8.1 (2017), p. 15037.

[22] Gao Huang et al. “Densely connected convolutional networks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 4700–4708.

[23] Patryk Jasik and Adrian Kastrau. *The set of 22 sessions of 14-channel eeg signals recorded during watching pictures*. 2019. DOI: 10.34808/1e5c-pp74. URL: https://mostwiedzy.pl/en/open-research-data/the-set-of-22-sessions-of-14-channel-eeg-signals-recorded-during-watching-pictures,110408381611163-0.

[24] Zhicheng Jiao et al. "Decoding EEG by Visual-guided Deep Neural Networks." In: *IJCAI*. Macao. 2019, pp. 1387–1393.

[25] James W Kalat. *Biological psychology*. Cengage Learning, 2015.

[26] Blair Kaneshiro et al. "A representational similarity analysis of the dynamics of object processing using single-trial EEG classification". In: *Plos one* 10.8 (2015), e0135697.

[27] Ashish Kapoor, Pradeep Shenoy, and Desney Tan. "Combining brain computer interfaces with vision for object categorization". In: *2008 IEEE conference on computer vision and pattern recognition*. IEEE. 2008, pp. 1–8.

[28] Isaak Kavasidis et al. "Brain2image: Converting brain signals into images". In: *Proceedings of the 25th ACM international conference on Multimedia*. 2017, pp. 1809–1817.

[29] Nastaran Khaleghi et al. "Visual saliency and image reconstruction from EEG signals via an effective geometric deep network-based generative adversarial network". In: *Electronics* 11.21 (2022), p. 3637.

[30] Sanchita Khare et al. "NeuroVision: perceived image regeneration using cProGAN". In: *Neural Computing and Applications* 34.8 (2022), pp. 5979–5991.

[31] Bryan Kolb and Ian Q Whishaw. *Fundamentals of human neuropsychology*. Macmillan, 2009.

[32] Vernon J Lawhern et al. "EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces". In: *Journal of neural engineering* 15.5 (2018), p. 056013.

[33] Ren Li et al. "The perils and pitfalls of block design for EEG classification experiments". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.1 (2020), pp. 316–333.

[34] Rouby El-Lone et al. "Visual objects categorization using dense EEG: A preliminary study". In: *2015 International Conference on Advances in Biomedical Engineering (ICABME)*. IEEE. 2015, pp. 115–118.

[35] David G Lowe. "Distinctive image features from scale-invariant keypoints". In: *International journal of computer vision* 60 (2004), pp. 91–110.

[36] David Monzo et al. “Precise eye localization using HOG descriptors”. In: *Machine Vision and Applications* 22.3 (2011), pp. 471–480.

[37] C S Nayak and A C Anilkumar. “EEG Normal Waveforms”. In: *StatPearls. Treasure Island* (2023).

[38] Marc R Nuwer et al. “IFCN standards for digital recording of clinical EEG”. In: *Electroencephalography and clinical Neurophysiology* 106.3 (1998), pp. 259–261.

[39] Simone Palazzo et al. “Decoding brain representations by multimodal learning of neural activity and visual features”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.11 (2020), pp. 3833–3849.

[40] Viral Parekh et al. “An EEG-based image annotation system”. In: *Computer Vision, Pattern Recognition, Image Processing, and Graphics: 6th National Conference, NCVPRIPG 2017, Mandi, India, December 16-19, 2017, Revised Selected Papers 6*. Springer. 2018, pp. 303–313.

[41] Adam Paszke et al. “Pytorch: An imperative style, high-performance deep learning library”. In: *Advances in neural information processing systems* 32 (2019).

[42] Fabian Pedregosa et al. “Scikit-learn: Machine learning in Python”. In: *the Journal of machine Learning research* 12 (2011), pp. 2825–2830.

[43] Mary C Potter et al. “Detecting meaning in RSVP at 13 ms per picture”. In: *Attention, Perception, & Psychophysics* 76 (2014), pp. 270–279.

[44] Alec Radford et al. “Learning transferable visual models from natural language supervision”. In: *International conference on machine learning*. PMLR. 2021, pp. 8748–8763.

[45] Mamunur Rashid et al. “Current status, challenges, and possible solutions of EEG-based brain-computer interface: a comprehensive review”. In: *Frontiers in neurorobotics* (2020), p. 25.

[46] Edward Rosten and Tom Drummond. “Machine learning for high-speed corner detection”. In: *Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006. Proceedings, Part I 9*. Springer. 2006, pp. 430–443.

[47] Yannick Roy et al. “Deep learning-based electroencephalography analysis: a systematic review”. In: *Journal of neural engineering* 16.5 (2019), p. 051001.

[48] Guohua Shen et al. “Deep image reconstruction from human brain activity”. In: *PLoS computational biology* 15.1 (2019), e1006633.

[49] Zhongzhi Shi. *Intelligence science: Leading the age of intelligence*. Elsevier, 2021.

[50] Irina Simanova et al. “Identifying object categories from event-related EEG: toward decoding of conceptual representations”. In: *PloS one* 5.12 (2010), e14465.

[51] Karen Simonyan and Andrew Zisserman. “Very deep convolutional networks for large-scale image recognition”. In: *arXiv preprint arXiv:1409.1556* (2014).

[52] Anupreet Kaur Singh and Sridhar Krishnan. “Trends in EEG signal feature extraction applications”. In: *Frontiers in Artificial Intelligence* 5 (2023), p. 1072801.

[53] Prajwal Singh et al. “EEG2IMAGE: Image reconstruction from EEG brain signals”. In: *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2023, pp. 1–5.

[54] Yonghao Song et al. “Decoding Natural Images from EEG for Object Recognition”. In: *arXiv preprint arXiv:2308.13234* (2023).

[55] Concetto Spampinato et al. “Deep learning human mind for automated visual classification”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 6809–6817.

[56] Susan Standring et al. “Gray’s anatomy: the anatomical basis of clinical practice”. In: *American journal of neuroradiology* 26.10 (2005), p. 2703.

[57] Laura M Stoinski, Jonas Perkuhn, and Martin N Hebart. “THINGSplus: New norms and metadata for the THINGS database of 1854 object concepts and 26,107 natural object images”. In: *Behavior Research Methods* (2023), pp. 1–21.

[58] *The Visual MNIST of Brain Digits*. `https://mindbigdata.com/opendb/visualmnist.html`.

[59] Praveen Tirupattur et al. “Thoughtviz: Visualizing human thoughts using generative adversarial network”. In: *Proceedings of the 26th ACM international conference on Multimedia*. 2018, pp. 950–958.

[60] Marija Ušćumlić, Ricardo Chavarriaga, and José del R Millán. “An iterative framework for EEG-based image search: robust retrieval with weak classifiers”. In: *PloS one* 8.8 (2013), e72018.

[61] David Vivancos and Felix Cuesta. “MindBigData 2022 A Large Dataset of Brain Signals”. In: *arXiv preprint arXiv:2212.14746* (2022).

[62] Changming Wang et al. “Combining features from ERP components in single-trial EEG for discriminating four-category visual objects”. In: *Journal of neural engineering* 9.5 (2012), p. 056013.

[63] Zesheng Ye et al. “See what you see: Self-supervised cross-modal retrieval of visual stimuli from brain activity”. In: *arXiv preprint arXiv:2208.03666* (2022).

[64] Xiang Zhang et al. “Multi-task generative adversarial learning on geometrical shape reconstruction from eeg brain signals”. In: *arXiv preprint arXiv:1907.13351* (2019).

[65] Xiao Zheng et al. “Decoding human brain activity with deep learning”. In: *Biomedical Signal Processing and Control* 56 (2020), p. 101730.