Deciphering Beyond the View: A Brain Decoding Approach to Language Processing Tasks

Thesis submitted in partial fulfilment
of the requirements for the degree of

Master of Science
in
Computational Linguistics by Research

by

Jashn Arora
2018114006
jashn.arora@research.iiit.ac.in

International Institute of Information Technology
Hyderabad - 500 032, INDIA
May 2023

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled "Deciphering Beyond the View: A Brain Decoding Approach to Language Processing Tasks" by Jashn Arora, has been carried out under my supervision and is not submitted elsewhere for a degree.

_____ _____

Date Advisor: Prof. R. S. Bapi

To family, friends and
the almighty

Acknowledgments

As I reflect on my thesis journey, I am filled with gratitude for the many individuals who have played a crucial role in helping me to achieve this milestone. First and foremost, I must express my sincere thanks to God for blessing me with the strength, motivation, and resilience to complete this undertaking.

My deepest thanks go to my thesis advisor, Prof. Bapi Raju. His knowledge, expertise, and encouragement have been invaluable throughout my research journey. He has pushed me to challenge myself, helped me to navigate difficult moments, and made me a better scholar. I owe him a debt of gratitude that I will never be able to repay.

I also want to thank my constant mentor and guide, Subbareddy Oota. His guidance, advice, and feedback have been instrumental in shaping my research and giving it direction. Without his support, I would not have been able to complete this project. I would also like to extend my heartfelt gratitude to Prof. Manish Gupta, whose expertise in the field of linguistics has been essential to the success of my research. I am truly grateful for his generosity in sharing his expertise and knowledge.

Of course, I must also thank my family, who have been a constant source of encouragement and support. My parents have always believed in me, and my sister, Guncha Arora, has been a rock for me through life's ups and downs.

Last but not least, I want to give a big shout-out to my friends, Sajal, Nikunj, Akshit, Anubhav, Manas, Guru and Kushagra. You guys have been my pillars of support throughout this journey, always there to offer a helping hand, a listening ear, or a much-needed break from the grind of academia. Tanish, your support and patience in clearing the dumbest of my doubts throughout my college journey have been invaluable to me, and I cannot thank you enough for all the help you have provided. You have been a true mentor and a friend, always there to guide me through difficult times. A huge thanks to Tisha, who has supported me through this process with her unwavering faith and constant encouragement.

Thank you all for being there for me and for making this journey one that I will never forget. I could not have done it without all of you, and for that, I am truly grateful.

Abstract

Brain decoding involves the reconstruction of stimuli from brain recordings. These recordings can be obtained by presenting stimuli to a subject in various forms, such as text, image, and speech. Despite extensive research on brain decoding, important questions remain unanswered. Can we develop multi-view decoders capable of decoding concepts from brain recordings of any view, including picture, sentence, or word cloud? Can we build a system that can use brain recordings to automatically generate descriptions of what a subject is viewing using keywords or sentences? How about a system that can automatically extract important keywords from sentences that a subject is reading? Answering these questions requires innovative approaches to brain decoding, as traditional methods have not yet been proven adequate.

Previous brain decoding efforts have focused only on single-view analysis and hence cannot help us build such systems. As a first step toward building such systems, inspired by Natural Language Processing literature on multi-lingual and cross-lingual modelling, this thesis proposes novel brain decoding setups: (1) Multi-view Decoding (MVD), (2) Cross-view Decoding (CVD), and (3) Abstract v/s Concrete Decoding. In MVD, the goal is to build an MV decoder that can take brain recordings for any view as input and predict the concept. In CVD, the goal is to train a model which takes brain recordings for one view as input and decodes a semantic vector representation of another view. Specifically, this thesis studies practically useful CVD tasks like image captioning, image tagging, keyword extraction, and sentence formation. In Abstract v/s Concrete Decoding, the goal is to build a decoder trained on concrete concepts and test it on both abstract and concrete concepts and similarly build a decoder trained on abstract concepts and test it in both types of concepts.

Extensive experiments lead to MVD models with $\sim$0.68 average pairwise accuracy across view pairs and CVD models with $\sim$0.8 average pairwise accuracy across tasks. It was found that the decoder trained on concrete concepts can decode both abstract and concrete objects with great and better accuracy than the model trained on abstract objects. Analysis of the contribution of different brain networks reveals exciting cognitive insights: (1) Models trained on picture or sentence view of stimuli are better MV decoders than a model trained on word cloud view. (2) Our extensive analysis across 9 broad brain regions, 11 language sub-regions, and 16 visual sub-regions of the brain helped us localize, for the first time, the parts of the brain involved in cross-view tasks like image captioning, image tagging, sentence formation, and keyword extraction. (3) The visual brain network is very important for processing Word+Picture stimuli for concrete concepts. Surprisingly, this is not the case for abstract concepts where voxels from the language and DMN brain network are more activated.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The human brain is a complex organ that plays a vital role in the functioning of the human body. It is responsible for processing information received through various stimuli and generating appropriate responses. For centuries, scientists and researchers have been trying to understand the mechanisms by which the brain processes and responds to different stimuli, such as visual, text or auditory cues.

The process of mapping brain responses to different stimuli has been an area of active research for many years. In recent years, this area of research has become even more critical as scientists seek to understand how the brain responds to various external stimuli and how this response can be harnessed to improve human health.

Language is a crucial component of human communication, and understanding how the brain processes linguistic stimuli is of significant interest to researchers in the field of cognitive neuroscience. The human brain has a remarkable ability to interpret and understand language, but the underlying mechanisms by which this occurs are not yet fully understood.

Studying how the brain processes linguistic stimuli has been an area of active research for many years. Various techniques have been used to investigate how different parts of the brain respond to linguistic stimuli, including electroencephalography (EEG), functional magnetic resonance imaging (fMRI), and magnetoencephalography (MEG) [33]. These techniques have enabled researchers to identify specific regions of the brain that are associated with language processing, such as Broca's and Wernicke's areas [27].

Understanding how the brain processes linguistic stimuli has important implications for fields such as linguistics, psychology, and neuroscience. It can provide insights into how we acquire language, how we understand it, and how we use it to communicate. In addition, this research has practical applications in areas such as speech therapy, language education, and artificial intelligence.

In recent years, a new approach called "brain decoding" has emerged, which aims to decode the neural activity patterns in the brain associated with linguistic stimuli to reconstruct insights about the stimuli. The brain decoding approach uses machine learning algorithms to analyze the neural activity patterns associated with stimuli. By identifying the neural patterns associated with specific linguistic

features, such as syntax or semantics, researchers can develop models that can predict the stimuli from brain activity patterns [56].

This approach has the potential to improve our understanding of how the brain processes linguistic stimuli and can provide insights into how we acquire language and how we use it to communicate. For example, recent studies have used brain decoding to investigate how the brain processes sentences with different levels of syntactic complexity [30].

This thesis will examine the current state of research on how the brain processes stimuli, focusing on the brain decoding approach. The thesis will explore the Cross-modal functioning of our brain, finding how even visual stimuli can activate the linguistic processing of the brain and how we can exploit that to perform Cross-View tasks like Image captioning and improve the accuracy of Abstract concept decoding. Additionally, the thesis will discuss the challenges and limitations of brain decoding, including the difficulty of interpreting the complex neural patterns associated with language processing and the need for more data to train machine learning algorithms effectively. Finally, the thesis will identify areas where further research is needed to advance our understanding of how the brain processes linguistic stimuli and how brain decoding can be applied to other areas of cognitive neuroscience.

1.1 Motivation for current Thesis

The human brain's ability to integrate information from different sensory modalities to form a coherent perception of the world is a fascinating and complex process. Understanding how the brain achieves this cross-modal processing is an important question in cognitive neuroscience with implications for various fields, such as psychology, linguistics, and artificial intelligence. One intriguing example of cross-modal processing is the way in which the brain integrates visual and linguistic information. When we see an image, our brain automatically tries to caption it, and when we read text, we try to imagine the narrative in our minds. Investigating how the brain achieves this integration of visual and linguistic information can provide insights into the fundamental mechanisms of cross-modal processing.

Therefore, the motivation for this thesis is to explore the cross-modal functioning of the brain, focusing specifically on the integration of visual and linguistic information. Using brain decoding techniques to analyze the neural activity patterns associated with visual and linguistic stimuli, the thesis aims to identify the neural mechanisms that enable the brain to integrate information from different sensory modalities. This research can help shed light on the cognitive processes underlying cross-modal integration and may have implications for various fields, such as natural language processing, computer vision, and assistive technologies for individuals with sensory disabilities. Ultimately, this thesis aims to contribute to our understanding of the complex, fascinating processes of the human brain and how language is represented and processed in the brain.

1.2 Major Contributions

The major contribution of the thesis are as follows:

- We propose three novel brain decoding settings: Multi-view decoding, Cross-view decoding and Abstract-Concrete brain decoding.
- We build decoder models using Transformer-based methods and analyze brain network contributions across multi-view and cross-view tasks.
- We augment the popular Pereira et al.'s dataset [72] with pairwise-view relationships and use it to demonstrate the efficacy of our proposed methods. We make the code and augmented dataset publicly available.[1]

1.3 Organisation of Thesis

This dissertation is organized in the following way:

Chapter 2 details how stimuli are processed in the brain, the brain imaging recording modalities, different approaches to how language and vision-based stimuli are represented in AI and finally, how the brain activations and stimuli representations can be bridged.

Chapter 3 introduces the concept of Brain Decoding and related works in this domain. It puts forwards the dataset, the process of brain network and voxel selection, the model architecture to build the decoder and the evaluation metrics to evaluate the results. All these are common across multiple thesis chapters.

Chapter 4 supports the first hypothesis that a decoder unique for a particular view can be used to decode fMRI corresponding to other views as well. It shows that the decoder trained on sentence view or picture view can be used as a universal decoder to decode other views as well, putting forward the fact that there is enough language processing in the brain, despite the view, to help decode other views. This work is part of our paper titled "Multi-view and Cross-view Brain Decoding" and was presented at COLING 2022.

Chapter 5 takes motivation from the results of the previous chapter to explore Cross-view tasks like Image Captioning, Sentence Formation etc. Great results in this study show that our brain works in a Cross-view fashion, i.e. whenever we see a visual stimulus, we try to caption it automatically in our brain and vice versa. When we read a text, our brain tries to form a possible image of it. This chapter also puts forwards a new dataset appended to the previous publicly available data, specially constructed for the study. This work is part of our paper titled "Multi-view and Cross-view Brain Decoding" and was presented at COLING 2022.

Chapter 6 puts forward one of our exciting studies about how Brain activations differ when we try to imagine a concrete object and an abstract object. The study also focused on identifying and

[1] https://tinyurl.com/MVCVBD

constructing a single decoder that could decode abstract and concrete objects accurately. This work is part of our paper titled "Brain Decoding for Abstract versus Concrete Concepts" and was presented at ACCS9.

Chapter 7 discusses two of my additional works on Brain Encoding. The first one studies what different NLP tasks are being performed by the brain while reading and listening. This work is part of our paper titled "Neural Language Taskonomy: Which NLP Tasks are the most Predictive of fMRI Brain Activity?." [62] and was presented at NAACL 2022. Second, focuses on studying how well Image and Multi-modal transformers perform Brain Decoding as compared to CNNs. This work is part of our paper titled "Visio-Linguistic Brain Encoding" and was presented at COLING 2022.

Chapter 8 finally summarizes the main findings of the works described in the thesis and puts forwards the limitations and possible future directions of our work.

Chapter 2

Background and Literature

2.1 Stimuli in Brain

2.1.1 Processing of Stimuli in Brain

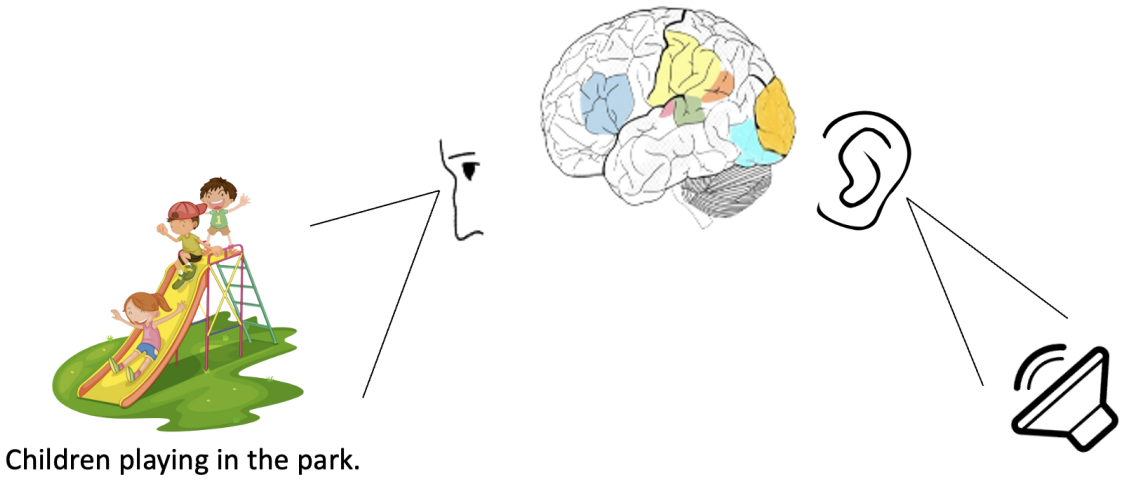


Figure 2.1: Langauge Processing induced in Brain due to different types of language stimuli.

Language is a systematic use of speech, text and gestures by humans to communicate ideas and feelings. Language processing refers to how humans process and understands language. It involves the ability to comprehend and produce spoken and written language. Language processing is a uniquely human ability that our brain performs so easily that we do not even realize it. We, humans, evolved at a much greater pace as compared to other living beings because of our ability to convey our thoughts, teach and learn skills and acquire knowledge through language. Human language is unique among all known systems of animal communication in a way that it has many modes of transmission (i.e. speech,

text, sight etc.), changes culturally, and geographically, and even diversifies over time. Thus, language plays a vital role in the development of human beings.

Language processing in the brain is a complex process that involves several brain regions and cognitive functions. The use of language for communication could be through audio stimuli (speech), wherein the auditory cortex of our brain is responsible for processing sounds, or it could be through visual stimuli (text or actions), wherein the primary visual cortex would be responsible for processing visual information. These brain regions are connected to other regions involved in language processing. The areas of the brain that work for different linguistic processes may vary. Some regions may be specific to lexical tasks, while some regions might perform the grammatical or the syntactical tasks, while some may deal with the semantic tasks during language processing at the same time. More specific details are listed below.

Language processing in the brain involves a complex series of neural mechanisms. Language processing involves several distinct stages or modules, each of which is responsible for a specific aspect of language comprehension and production.

- **Phonological processing:** The first stage of language processing involves the analysis of sounds and their combinations to form words. This involves the activation of regions in the left superior temporal gyrus (STG) and the posterior superior temporal sulcus (pSTS) [41].
- **Lexical and syntactic processing:** After phonological processing, the brain extracts meaning from the sounds and begins to build a syntactic structure for the sentence. This involves the activation of the inferior frontal gyrus (IFG) and the posterior middle temporal gyrus (pMTG) [37].
- **Semantic processing:** Once the sentence structure is built, the brain assigns meaning to the words and constructs a representation of the sentence's overall meaning. This involves the activation of the anterior temporal lobe(ATL) and the posterior cingulate cortex (PCC) [13].
- **Pragmatic processing:** Finally, the brain uses contextual information to interpret the meaning of the sentence in the broader context of the discourse. This involves the activation of the medial prefrontal cortex (mPFC) and the posterior superior temporal gyrus (pSTG) [103].

2.1.2 Brain Regions and Networks

Brain Regions and Brain Networks are intimately related to each other. Brain regions are collections of neurons that are anatomically and functionally connected, and they can be defined based on criteria such as location or functional specialization. Brain networks, on the other hand, are collections of brain regions that are functionally connected and work together to perform specific cognitive or perceptual functions [83].

The relationship between brain regions and brain networks can be thought of as a hierarchical organization, with brain regions forming the basic building blocks of brain function and brain networks representing higher-order functional units that emerge from the interactions between brain regions [17, 83].

Each brain network is composed of a set of brain regions that are functionally connected, and the properties of the network emerge from the interactions between these regions. Conversely, the properties of individual brain regions are shaped by their connections to other regions in the brain, and their functional role is influenced by the networks to which they belong.

2.1.3 Identifying Brain Networks

Brain Atlases are one of the important tools in neuroscience that helps researchers understand the complex structure and organization of the brain and identify specific regions involved in various functions. Brain atlases divide the brain into distinct regions based on various criteria such as anatomical location, function or connectivity.

The main purpose of brain atlases is to provide a standardized reference for researchers to use when studying the brain, making it easy for them to compare results across different studies and investigate the relationships between different brain regions.

For our work, we used the Automated Anatomical Labeling (AAL) Atlas [94] that is used to parcellate the human brain into regions of interest based on anatomical landmarks, as seen in MRI scans.

Some of the important Brain Networks [84] in AAL atlas are:

- **Language Network:** The Language Network is primarily composed of two regions in the left hemisphere: Broca's area and Wernicke's area. Broca's area, located in the left inferior frontal gyrus, is responsible for language production and speech output, while Wernicke's area, located in the left superior temporal gyrus, is responsible for language comprehension. Other regions involved in the language network include the middle and superior temporal gyri, the angular gyrus, and the supramarginal gyrus. Tzourio-Mazoyer et al. [94] used the AAL atlas to map the neural substrates of language processing in the brain.

- **Visual Network:** The Visual Network is primarily located in the occipital lobe and is responsible for processing visual information. The primary visual cortex (V1) is responsible for processing basic visual information, such as the orientation and location of visual stimuli. The secondary visual cortex (V2) and the visual association cortex (V3-V5) are involved in processing more complex visual information, such as object recognition and spatial awareness. Tzourio-Mazoyer et al. [94] mapped the neural substrates of visual processing in the brain.

- **Default Mode Network:** The Default Mode Network (DMN) is a set of brain regions that are active when an individual is at rest and not engaged in any particular task. The DMN includes the medial prefrontal cortex, the posterior cingulate cortex, the precuneus, the inferior parietal lobule, and the medial temporal lobe. This characterization is based on research by Raichle et al. [78], which used the AAL atlas to map the neural substrates of the DMN. The DMN is thought to be involved in a range of cognitive processes, including self-reflection, introspection, and mind-wandering.

- **Task Positive Network:** The Task Positive Network (TP) is a set of brain regions that are active when an individual is engaged in a particular task, such as working memory, attention, or decision-making. The TP Network includes the dorsolateral prefrontal cortex, the lateral parietal cortex, and the anterior cingulate cortex. Fox et al. [32] used the AAL atlas to map the neural substrates of the TP Network. The TP Network is thought to be involved in cognitive control and executive functions, and is often anti-correlated with the DMN.

Other than the main brain networks that are important for this work and detailed above, the AAL atlas has various other Brain Networks as well. Some of them are named below:

- **Somatomotor Network:** Involved in motor planning and execution.
- **Limbic Network:** Involved in emotion, motivation, and memory.
- **Attentional Network:** Involved in directing and maintaining attention.
- **Auditory Network:** Involved in processing auditory information, including speech and music.
- **Sensory Network:** Involved in processing tactile, thermal, and pain information.
- **Salience Network:** Involved in detecting and prioritizing relevant information in the environment.
- **Frontoparietal Network:** Involved in executive functions such as working memory and decision-making.

2.1.4 Extracting Brain Representations

There have been several techniques to record brain activations. These methods serve as a window into the inner workings of the brain. This thesis analyzes data collected mainly from functional magnetic resonance imaging (fMRI) technique, however for completeness some of the popular techniques are described as follows:

2.1.4.1 fMRI

fMRI [44] stands for Functional Magnetic Resonance Imaging (fMRI). It is a non-invasive imaging technique that detects changes in blood flow to measure brain activity. The principle behind fMRI scans is that there is higher blood flow and oxygen consumption in active brain regions than the inactive regions. The reason for the popularity of fMRI in Brain Decoding tasks is its high spatial resolution that helps us accurately identify brain activity in specific brain regions.

2.1.4.2 EEG

EEG [59] stands for electroencephalography. It is also a non-invasive imaging technique that measures brain activity by detecting electrical activity on the scalp. It is based on the principle that the brain generates electrical activity that can be measured on the scalp. EEG has low spatial resolution making it less accurate at identifying brain activity in specific brain regions but has high temporal resolution making it accurate at measuring brain activity in real time.

2.1.4.3 MEG

MEG [7] stands for Magnetoencephalography. It is also a non-invasive technique that measures brain activity by detecting magnetic fields generated by brain activity. It is based on the principle that the brain generates magnetic fields that can be measured outside the head. MEG has high temporal resolution and can accurately measure brain activity in real-time, but it has lower spatial resolution than fMRI.

2.1.4.4 TMS

TMS [38] stands for Transcranial magnetic stimulation. TMS is a non-invasive technique that uses magnetic fields to stimulate specific brain regions. It is based on the principle that brain activity can be influenced by the application of magnetic fields. TMS is often used in combination with other techniques, such as fMRI or EEG, to study the causal relationship between brain activity and behaviour.

2.2 Language in AI

Language in AI approaches is represented using two major techniques, symbolic and statistical approaches. Symbolic approaches use explicit rules and structures such as formal grammar to represent language. Whereas, Statistical approaches use ML algorithms to automatically learn patterns and relationships in large text datasets.

Some of the most common statistical approaches are Word Embeddings, RNNs and Transformers. These techniques are detailed below.

2.2.1 Word-level Representations

Word embeddings are low-dimensional, continuous-valued vector representations of words which are large amounts of text data using unsupervised learning methods. These vector representations capture the semantic and syntactic relationships between words, allowing AI systems to understand the meaning of language and make predictions based on it. Some of the approaches to get the word embeddings are described below:

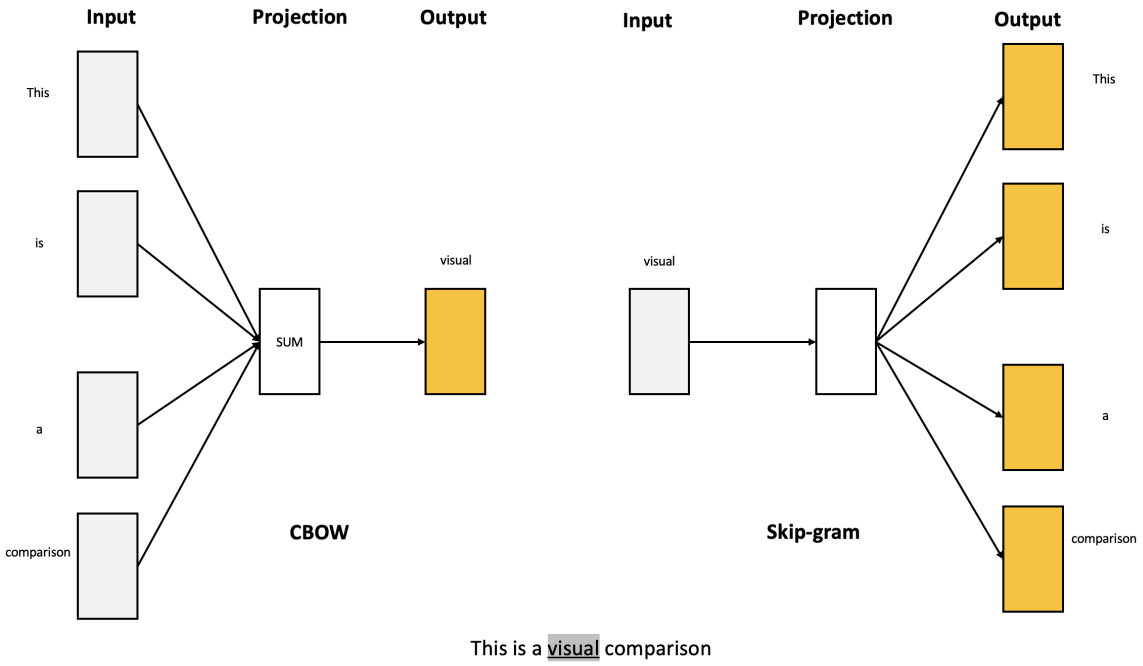


Figure 2.2: Training algorithms for word2vec (Adapted from a Blog titled "A Beginner's Guide to Word Embedding with Gensim Word2Vec Model")

2.2.1.1 Word2vec

Word2vec [54] is a method of learning word embeddings that was developed by Google researchers in 2013. The word vectors are learned through a neural network trained on a large amount of text data, such as a collection of news articles or a web crawl.

There are two main training algorithms for word2vec:

- **continuous bag-of-words (CBOW):** This algorithm tries to predict a word given its context.
- **skip-gram:** This model tries to predict the context given the centre word.

Fig. 2.2 summarizes the above two training algorithms. These models are trained by minimizing the negative log-likelihood of the training data, which effectively maximizes the probability of observing the training data given the model parameters.

Word2vec word embeddings were successful at improving the performance of various NLP tasks. However, there are several issues with the original Word2vec approach, such as:

- **Limited context:** Word2vec takes into account only a small window of words surrounding the central word during training. This can result in suboptimal embeddings for words that have multiple meanings or are used in diverse contexts.

- **Lack of global information:** Word2vec does not considers the global co-occurrence statistics of words in the training data, which can lead to poor generalization of new tasks.
- **Expensive Training:** Word2vec requires multiple iterations over the training corpus, which can be time-consuming and computationally expensive.

2.2.1.2 GloVe

To address the issues with Word2vec embeddings, Pennington et al. [68] introduced GloVe (Global Vectors for Word Representation) with the idea of using a count-based method for generating word embeddings that combines the advantages of global and local information.

The authors proposed constructing a co-occurrence matrix by counting the number of times each word occurs within a fixed context window and normalizing it to get the probability of observing two words together. The goal is to capture the ratio of co-occurrence probabilities, rather than the probabilities themselves. This allows GloVe to capture both global and local information, and produce embeddings that are more accurate and generalizable. GloVe has been shown to outperform word2vec on a range of natural language processing tasks, including word similarity and analogy tasks, as well as text classification and language modelling.

While GloVe have been an effective method for generating word embeddings, they have some limitations, including capturing non-linear relationships between words and handling variable-length inputs.

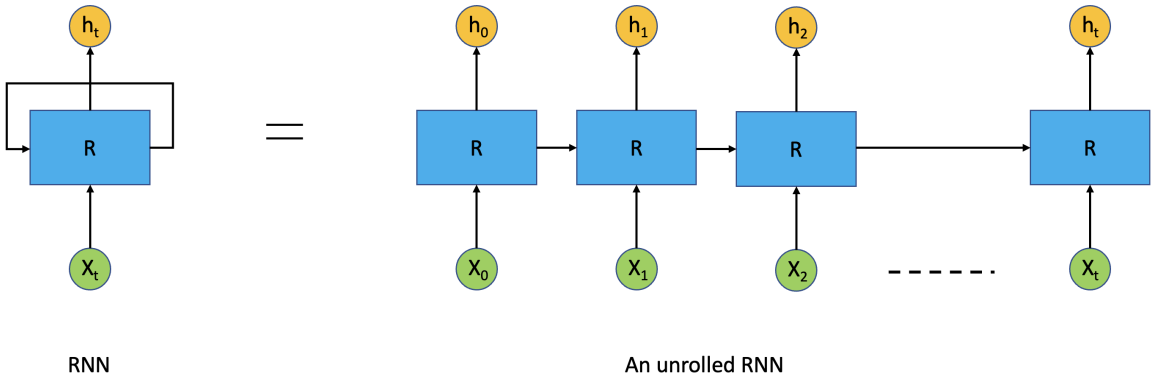


Figure 2.3: Recurrent Neural Network (Reproduced from Colah's Blog)

2.2.2 RNNs

Recurrent Neural Networks (RNNs) [55] are a type of neural network that is well suited for language modelling as they are designed to capture sequential dependencies in data. In language, the meaning of a word is often dependent on the context in which it appears. An RNN consists of a chain of repeating

modules, called cells which allow the network to maintain a ”memory” of previous inputs and use that information for future prediction. This makes RNNs useful for processing sequential data, such as text. Fig: 2.3 describes how an RNN structure looks like.

RNNs can also be used for tasks where the length of the input and output sequences can vary. They can handle variable-length sequences because they process each input one at a time and maintain a state vector that summarizes the information seen so far. This allows them to produce output sequences of varying lengths.

However, RNNs suffers from the problem of vanishing gradients, where the gradients that are back-propagated through the network get smaller and smaller as they move backward. This makes it difficult for hte network to learn long-term dependencies.

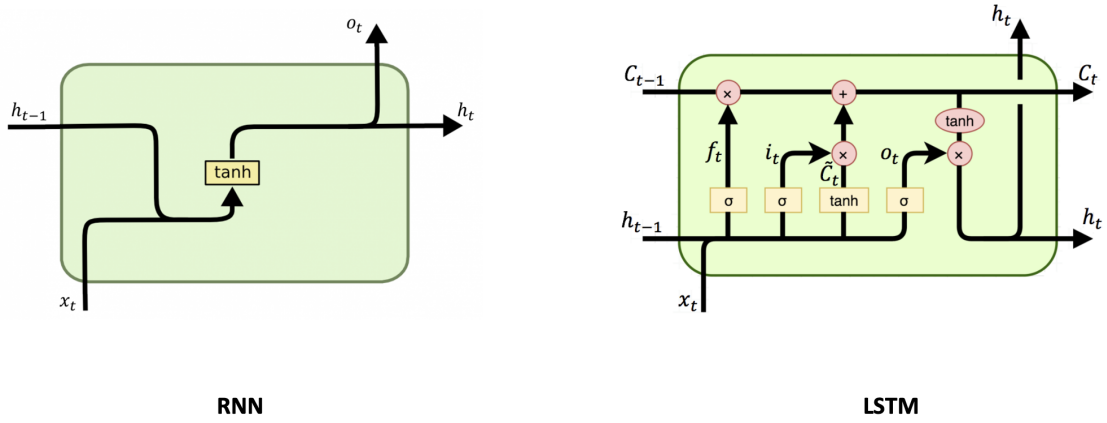


Figure 2.4: Architechture of RNN vs LSTM (Reproduced from Colah's Blog)

2.2.2.1 LSTM

Long short-term memory (LSTM) [42] is a type of RNN architecture that was specially designed to address the issue of vanishing gradients of RNNs. LSTMs use a special type of cell that can store information in a ”memory” cell and control the flow of information using ”gates”, which makes it possible to store information over a longer period of time and selectively forget or remember information based on the input.

In addition to vanishing gradients, RNNs can also suffer from the problem of exploding gradients, where the gradients become too large and cause the weights of the network to diverge. LSTMs address this problem by introducing a gradient clipping mechanism that limits the size of the gradients during training. Fig. 2.4 shows the difference between the architecture of a RNN to that of a LSTM.

Overall, LSTMs are a significant improvement over traditional RNNs and are widely used in a variety of applications, including natural language processing, speech recognition, and time series prediction.

However, Both LSTMs and RNNs are computationally expensive to train, because the process sequences sequentially, which they cannot parallelize across multiple processors. This leads to a higher training time.

2.2.3 Transformers

Vaswani et al. [95] came up with the Transformer architecture to address the limitations using self-attention mechanisms and a parallelizable architecture.

The transformer model consists of an encoder and a decoder, each composed of multiple layers. The model uses self-attention to compute a weighted sum of the input sequence, allowing the model to attend to different parts of the sequence at different times. This makes it easier for the model to capture long-term dependencies. Fig. 2.5 shows the architecture of the Transformer model.

The model also uses multi-head attention, which allows it to attend to different parts of the input sequence at multiple levels of abstraction, improving its ability to capture complex patterns. Its ability to capture long-term dependencies and its computational efficiency have made it a popular choice for many applications. There are many transformer-based models currently available, like BERT, RoBERTa, GPT, BART.

2.2.3.1 BERT

BERT (Bidirectional Encoder Representations from Transformers) [25] is an encoder-only Transformer model developed by Google in 2018 for natural language processing tasks. BERT is pre-trained using a large corpus of text, such as the entire Wikipedia or the BooksCorpus dataset [110]. During pre-training, BERT is trained on two tasks:

- **Masked Language Modelling:** Given a sentence with some of the tokens randomly masked out as input, the task for BERT is to predict the missing tokens based on the surrounding tokens. This allows BERT to learn representations that capture contextual information.
- **Next Sentence Prediction:** Given a pair of sentences, the task for BERT is to predict whether the second sentence follows the first sentence in the original text. This allows BERT to learn relationships between sentences and improve its ability to handle tasks such as Natural Language Inference (NLI).

After pre-training, BERT can be fine-tuned for specific NLP tasks. The task-specific output layer is added to the encoder to generate task-specific outputs. Fine-tuning can be done with relatively small amounts of task-specific data, as the pre-trained representations capture a large amount of information about the structure of natural language.

The input format for BERT is a sequence of tokens. The input sequence is augmented with special tokens, such as [CLS] and [SEP], which are used to indicate the beginning and end of the sequence and to separate different segments of the input sequence. In the case of pre-trained BERT, we get the final

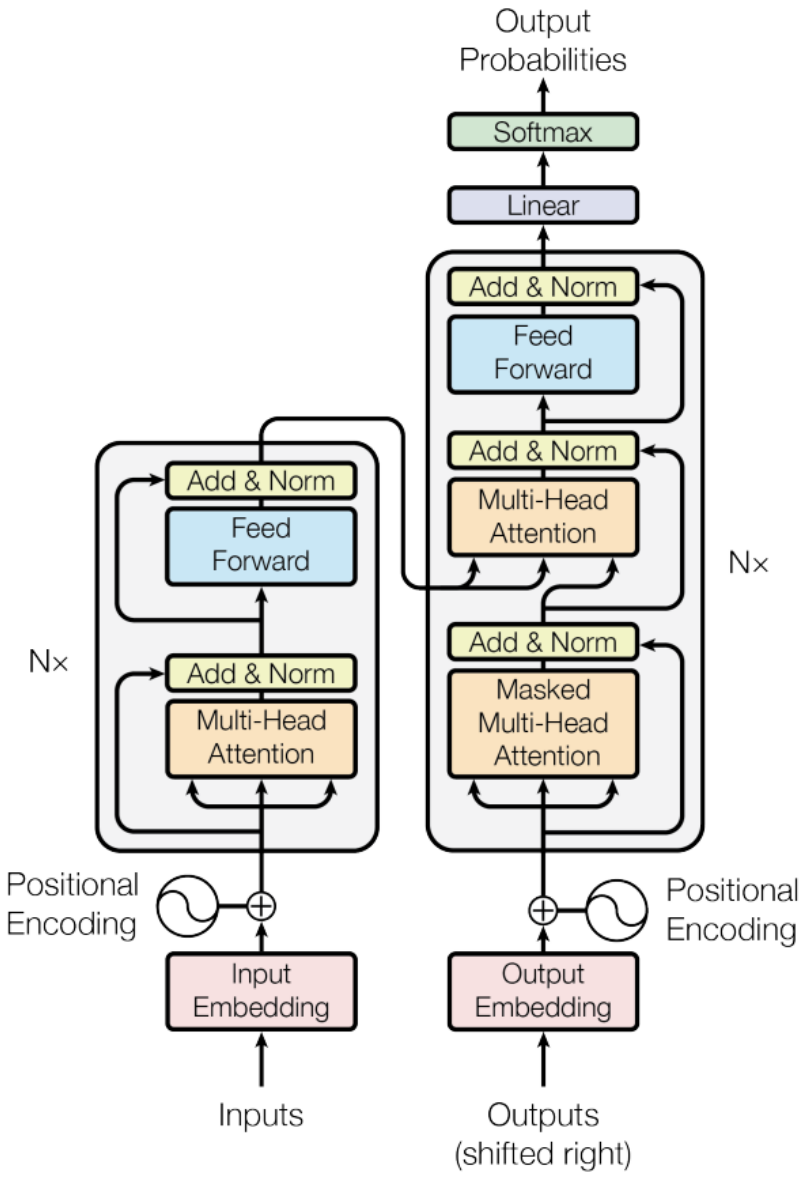


Figure 2.5: Model Architecture of Transformer (Reproduced from Vaswani et al. [95])

embedding for each input token, including the [CLS] and [SEP] tokens. Fig. 2.6 shows the input and output format for BERT.

There are two ways in which we can get the embedding that represents the whole of the sequence. First, we can take the embedding corresponding to the [CLS] token, as it is assumed to capture information of the whole sentence and is used to finetune BERT for classification tasks. Second, we can take the average of the final embeddings corresponding to all the tokens, excluding [CLS] and [SEP] tokens. We call this representation for the sequence "Average polled embedding".

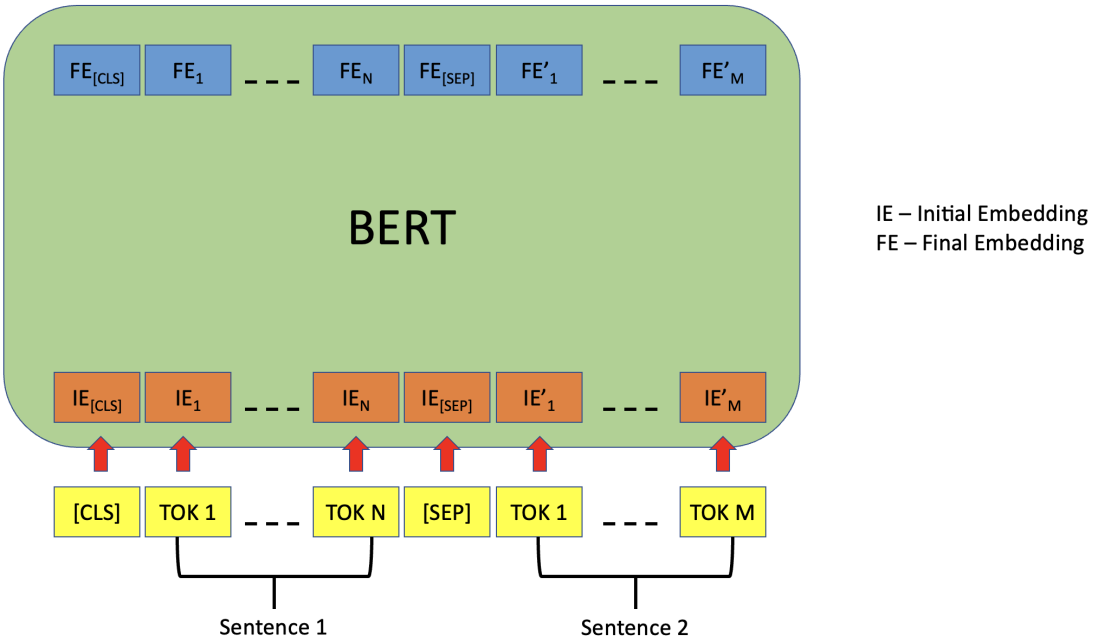


Figure 2.6: Input and Output format for BERT (Adapted from Devlin et al. [25])

2.3 Mapping Brain to AI

2.3.1 Brain Decoding

Brain decoding is the process of reconstructing stimuli presented to a person using machine learning algorithms [72, 100].

In this approach, researchers present a subject with a series of stimuli, such as text or image and the brain activity of the subject is recorded using techniques such as functional magnetic resonance imaging (fMRI). The brain activity is then used as input to a Machine learning based decoder, which is trained to identify patterns in the brain activity that correspond to specific stimuli and predict/reconstruct the stimuli presented earlier.

Once the decoder has been trained, it can be used to reconstruct the stimuli from the person's brain activity. This is done by presenting the person with a new stimulus and measuring their brain activity. The decoder then processes the brain activity and attempts to reconstruct the presented stimulus.

This approach to brain decoding has a wide range of applications, including the development of brain-machine interfaces (BMIs) and the study of brain function and development. It allows researchers to gain insight into the mental processes and states of an individual and can be used to understand how the brain processes and encodes information.

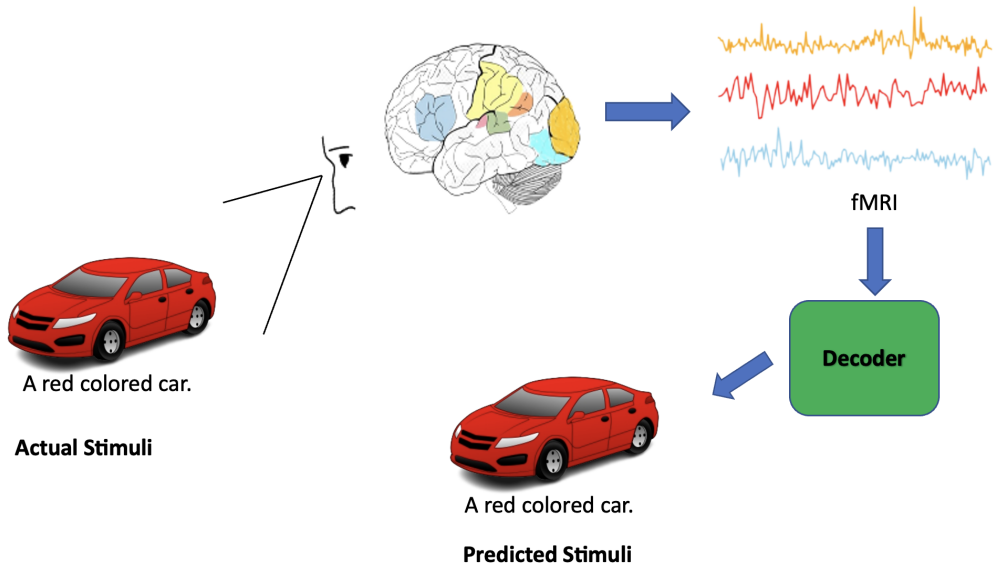


Figure 2.7: Brain Decoding

2.3.2 Brain Encoding

Brain encoding is the process of using Machine Learning models to predict the patterns of neural activity that are evoked by a specific stimulus presented to the subject [34, 80].

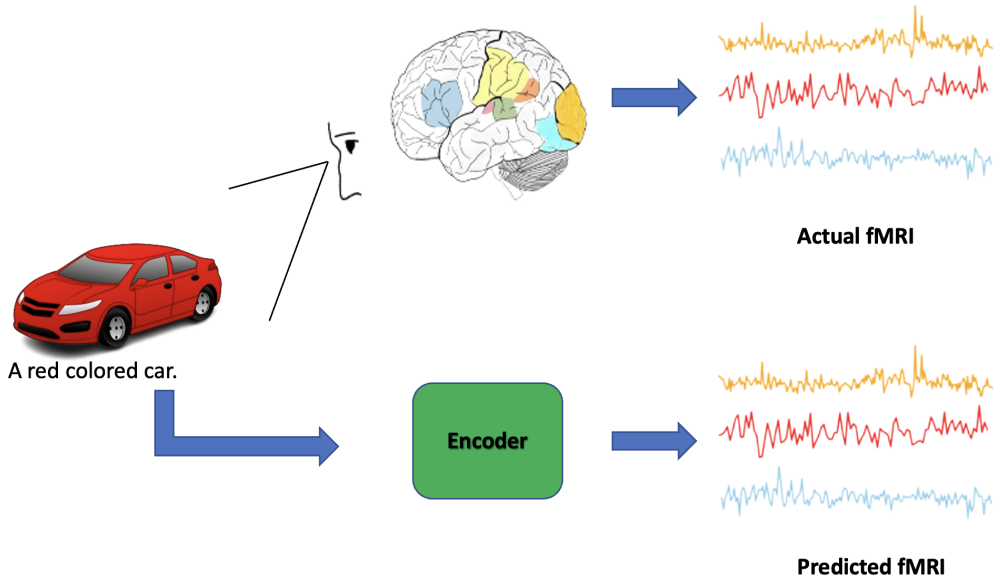


Figure 2.8: Brain Encoding

The goal of brain encoding is to understand how the brain processes information by identifying the neural code that underlies perception and cognition.

In this approach, researchers present a subject with a series of stimuli, such as text or image and the brain activity of the subject is recorded using techniques such as fMRI. Once the neural activity has been measured, ML models can be trained to decode the patterns in brain activity and relate them to the presented stimulus. Now, these models can be used to predict Brain Activity for a subject corresponding to a new stimulus presented to the subject.

Brain encoding has a wide range of applications, from understanding basic perceptual and cognitive processes to developing brain-computer interfaces and neuroprosthetics. By identifying the neural code that underlies perception and cognition, brain encoding can help us better understand how the brain works and develop new technologies that interface directly with the brain.

Chapter 3

Brain Decoding

3.1 Introduction

Brain decoding models aim to understand what a subject is thinking, seeing, and perceiving by analyzing neural recordings. Thus, in the context of language, it may be beneficial to learn mappings between linguistic representation and the associated brain activation, and how we compose the linguistic meaning from different stimuli such as text [72, 100], images [29, 11], videos [46, 61], or speech [109] by analyzing the evoked brain activity. Also, decoding the functional activity of the brain has numerous applications in education and healthcare.

Brain recordings can be obtained by providing stimuli to a subject in various forms. For example, a concept (like *apartment*) can be presented using: (1) Word Picture (WP) view: picture along with the concept word, (2) Sentence (S) view: sentence containing the word, or (3) Word cloud (WC) view: word cloud containing the word along with other semantically related words. Recent studies have made much progress using functional magnetic resonance imaging (fMRI) brain activity to reconstruct semantic vectors corresponding to linguistic items, including words [56, 72], phrases, sentences, and paragraphs [100]. However, all such studies have been limited to single-view analysis. Separate models are trained to process different views. Also, the decoding target is typically a semantic vector of the concept word.

In the Natural Language Processing (NLP) community, researchers have recently started focusing on building multi-lingual and cross-lingual systems [22, 21, 105]. Multi-lingual systems improve accuracy for low-resource languages and enable applications even in the absence of training data for low-resource languages. Cross-lingual systems take input in one language and produce output (e.g., summary) in another language. Inspired by this multi-lingual/cross-lingual shift in NLP, we propose two novel brain decoding setups: multi-view decoding (MVD) and cross-view decoding (CVD). Such setups are critical to build MV decoders which can decode concepts from brain recordings corresponding to any view (picture, sentence, word cloud) of stimuli or systems that can automatically describe using sentences or keywords what a subject is watching or automatically extract important keywords from sentences that a subject is reading.

In Multi-View Decoding, the goal is to build a Multi-View decoder that can take brain recordings for any view as input and predict the concept. Fig. 4.1 shows examples of using a Multi-View decoder. Such a Multi-View decoder can be trained on data for any specific view. Multi-lingual models have shown huge zero-shot accuracy improvements for inference on low-resource language inputs across many NLP tasks [21]. Similarly, can we improve decoding accuracy using a Multi-View decoder model for some views?

In Cross-View Decoding, the goal is to train a model which takes brain recordings for one view as input and decodes a semantic vector representation of another view. Fig. 5.1 shows examples of four such Cross-View Decoding tasks. Given an fMRI activation corresponding to a picture view of the stimuli, how accurately can we decode a sentence representing the picture? Which parts of the brain are involved in Cross-View Decoding tasks like image captioning, image tagging, keyword extraction, and sentence formation?

Historically, the fMRI brain activity has been decoded to a semantic vector representation of a view (word picture, sentence, word cloud) using a ridge-regression decoder [72, 85]. In particular, earlier brain decoding works focused on hand-crafted features to train such decoder models [56, 100] , which suffer from these drawbacks: (1) cannot address word sense disambiguation, (2) limited in terms of vocabulary, (3) inability to extract signals for abstract stimuli, and (4) inability to capture the context and sequential aspects of a sentence. Recently, many studies have shown accurate results in mapping the brain activity using neural distributed word embeddings for linguistic stimuli [4, 72, 65, 60, 85]. To represent meaning, these studies use either word or sentence level embeddings extracted from the models trained on large corpora. Unfortunately, none of these addresses the open questions around multi-view decoding and cross-view decoding. Recently, Transformer-based models have been explored for brain encoding [43], which inspires us to harness Transformer-based models like BERT [25] for our brain decoding tasks.

3.2 Related Works

Advances in functional neuroimaging tools such as fMRI have made it easier to study the relationship between language/visual stimuli and functions of brain networks [23, 90, 31]. In the past two decades, researchers have leveraged fMRIs to understand how the brain represents language and semantics.

Initial brain decoding experiments studied the recovery of simple concrete nouns and verbs from fMRI brain activity [56, 67, 61, 70] where the subject watches either a picture or a word. Unlike the earlier work,Wehbe et al. [100] and Huth et al. [46] built a model to decode the text passages instead of individual words. However, these studies used either simple or constrained sets of stimuli, which poses a question of generalization of these models. Recently, Pereira et al. [72] explicitly decoded both words and sentences when subjects were shown both concrete and abstract stimuli. Affolter et al. [1] reconstructed the sentences along with categorizing words or predicting the semantic vector representa-

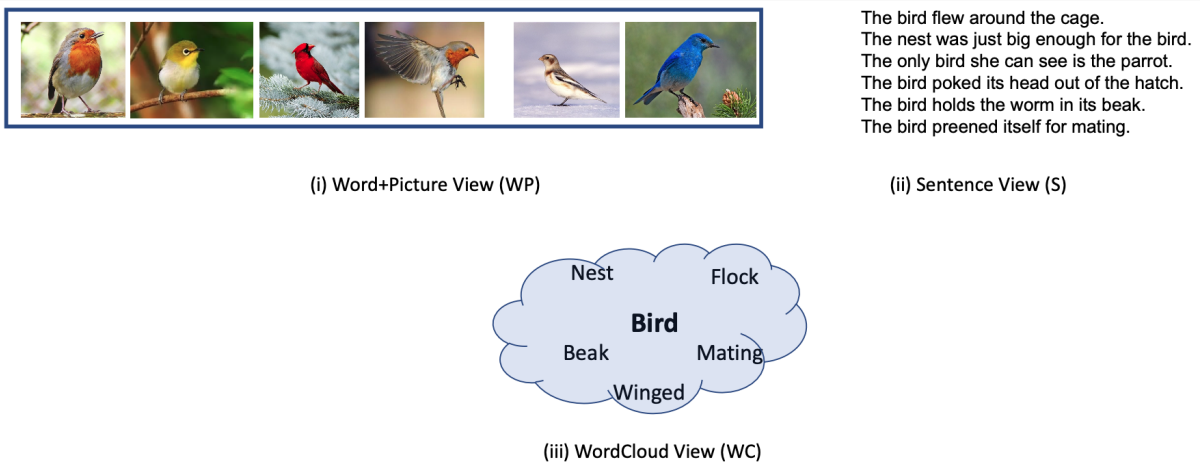


Figure 3.1: Stimuli example from Dataset [72]

tion from fMRI brain activity. Schwartz et al. [81] and Wang et al. [98] focused on understanding how multiple tasks activate associated regions in the brain.

To train ridge regression decoder models, earlier works focused on hand-crafted features [56, 100], which suffer from various drawbacks like inability to capture the context and sequential aspects of a sentence, inability to extract signals for abstract stimuli, etc. With the success of deep learning-based word representations, multiple researchers have used distributed word embeddings for brain decoding models in place of carefully hand-crafted feature vectors [46, 4, 72, 65, 60, 85, 99]. Using the distributed sentence representations, Wehbe et al. [102], Jain et al. [48], and Sun et al. [85] demonstrated that neural sentence representations are better for decoding whole sentences from brain activity patterns.

Recently, the success of contextual and Transformer based language models has raised the question of whether these models might be able to make an association between brain activation and language. Beinborn et al. [10] showed the success of the ELMo language model [73] in predicting the fMRI brain activation of several datasets. Also, Gauthier et al. [34] and Toneva et al. [92] tried to decode the fMRI activations to improve the latent representations of language stimuli using BERT [25]. In contrast to earlier works, Affolter et al. [1] described language generation with GPT-2 using brain activities. We take inspiration from these pieces of work and experiment with BERT for various multi-view and cross-view brain decoding tasks.

3.3 Dataset

We experiment with the popular dataset from Pereira et al. [72]. It is obtained from 11 subjects (P01, M01, M02, M04, M07, M09, M10, M13, M15, M16, M17) where each subject read 180 concept words (abstract + concrete) in three different paradigms or views while functional magnetic resonance images (fMRI) were acquired. These contain 128 nouns, 22 verbs, 29 adjectives and adverbs, and 1 function

word. In paradigm-1 (WP), participants were shown concept word along with picture with an aim to observing brain activation when participants retrieved relevant meaning using visual information. In paradigm-2 (S), the concept word presented in a sentence allows us to probe activity in the language areas associated with contextual information and meaning of a sentence. In paradigm-3 (WC), the concept word was presented in a word cloud format, surrounded by five semantically similar words. These paradigms provide brain representation of 180 concepts in three different views.

For each of the 180 concepts, the dataset contains five pictures, six sentences each containing the concept word, and a word cloud. For example, Fig. 3.1 shows all the three views for the concept word 'bird'. The dataset has (1) a picture p showing a red bird sitting on a tree branch, (2) sentence s like "A green bird flying in the sky", and (3) word cloud c with words "bird, purple, flock, winged, nest, beak". The dataset also has fMRIs for each of these three views.

3.4 Extracting Brain Representations

Several techniques are used for recording brain activations, but we used the functional magnetic resonance imaging (fMRI) dataset.

Functional magnetic resonance imaging (fMRI) [44] is a type of brain imaging that uses magnetic fields and radio waves to measure blood flow in the brain. This technique can be used to map brain activity by detecting changes in the amount of oxygenated blood flowing to different regions of the brain.

fMRI data is typically collected by placing a person inside an MRI machine and having them perform a specific task or simply lie still while the machine takes multiple images of the brain. The data is then processed using specialized software to generate maps of brain activity, which can be used to study brain function.

The extracted data is an image that indicates the brain activity at each time point, giving us the ability to see which specific brain regions are active during a task. This brain activity is correlated to increased blood flow, where active neurons will consume more oxygen and make the blood flow to this region more prominent. The image that we get using this is called BOLD (blood-oxygen-level dependent) images which is a contrast mechanism for MRI.

For each of the three views, the stimuli were shown to each subject, and fMRI was recorded. Specific details about this can be found in a previous work by Pereira et al. [72].

3.5 Informative Voxel Selection

Inspired by the voxel selection method in Pereira et al. [72], we chose the informative voxels for our linear regression models as follows. The regression models are trained on each voxel and its 26 neighbouring voxels to predict the semantic vector representation. For each voxel in the training part,

the mean correlation was calculated between "true" (text-derived) and predicted representations, and the voxels corresponding to the top 5000 mean correlation values were selected as informative voxels.

For each of the experiments, the informative voxels were chosen by the above method for the train dataset and the exact same voxel loaction (or voxels) were taken for the test dataset.

3.6 Brain Network Selection

Inspired by Pereira et al. [72] and based on the resting-state functional networks, we focused on four brain networks:

- **Default Mode Network (DMN):** linked to the functionality of semantic processing [16, 13].
- **Language Network:** related to language processing, understanding, word meaning, and sentence comprehension [30].
- **Task Positive Network:** related to attention, salience information [13, 28, 75].
- **Visual Network:** related to the processing of visual objects, object recognition [16, 75].

We report the distribution of 5000 informative voxels across the four brain networks across various experiments in Section 4.2.2. Across all participants, voxel distribution across networks is as follows: 4670 (Language), 6490 (DMN), 11630 (TP), and 8170 (Visual). Note that the reported distributions in Section 4.2.2 do not add up to 1 because the contribution of the remaining brain networks is not considered.

3.7 Extracting Language Representation

We used BERT for extracting semantic representations for the stimuli. BERT has been explained in sec. 2.2.3. BERT-pooled output was taken as the stimuli embeddings in our experimentation.

3.8 Model Architecture

We trained a ridge regression based decoding model to predict the semantic vector representation associated with the fMRI informative voxels for a type (view) of each language stimulus. Each dimension is predicted using a separate ridge regression model. Formally, we are given the informative voxel matrix $X \in \mathbb{R}^{N \times V}$ and stimuli vector representation $Y \in \mathbb{R}^{N \times D}$, where N denotes the number of training examples, V denotes the number of informative voxels (we fix it to 5000), and D denotes the embedding dimension of language stimuli. For BERT, D=768.

The ridge regression objective function is:

$$f(X_i) = \min_{W_{io}} \|Y_o - X_i W_{io}\|_F^2 + \lambda \|W_{io}\|_F^2 \tag{3.1}$$

where, X_i denotes the input voxels for view i (out of {word+picture, sentence, wordcloud}), Y_o denotes the matrix with embeddings o (out of {word, sentence, word cloud}), W_{io} denotes the learned weight coefficients for each input view i and output embedding o, $\|.\|_F$ denotes the Frobenius norm, and $\lambda > 0$ is a tunable hyper-parameter representing the regularization weight. Besides ridge regression, of course, various other models could be used. However, the goal of this thesis is to analyze novel decoding setups using the most popular decoding model in neuro-science literature, namely, ridge regression. We leave exploration of complex models as part of future work.

Hyper-parameter Settings: We used sklearn's ridge regression with default parameters, 18-fold cross-validation, Stochastic-Average-Gradient Descent Optimizer, Huggingface for BERT, MSE loss function and L2-decay (λ):1.0.

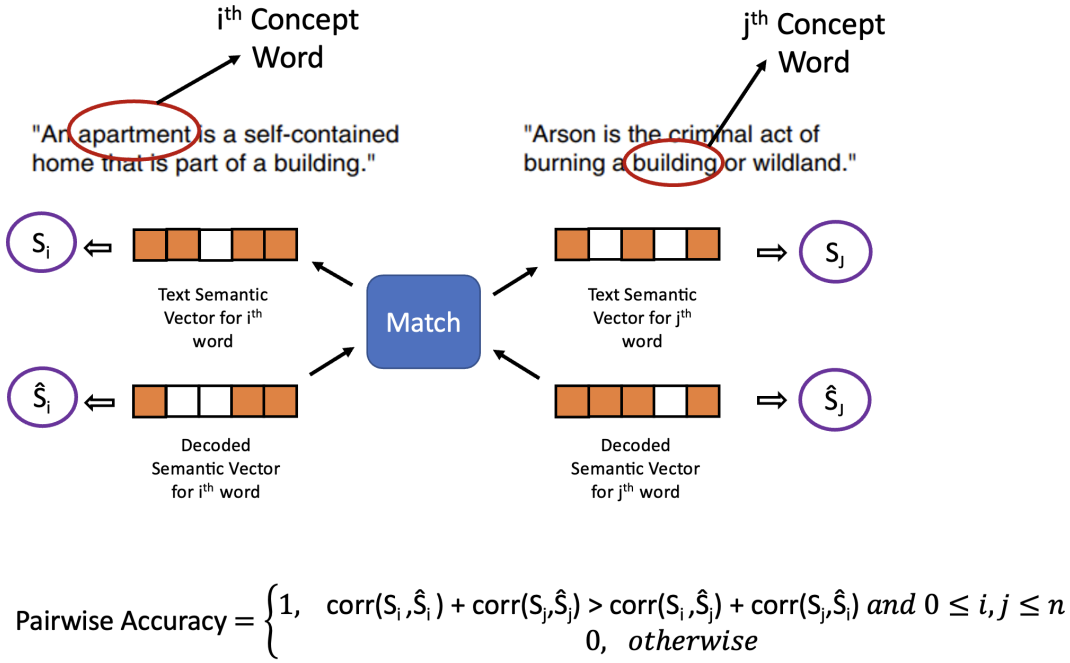


Figure 3.2: Pairwise Accuracy

3.9 Evaluation Metric

We use the popular pairwise and rank accuracy metrics for evaluation.

3.9.0.1 Pairwise Accuracy

To measure the pairwise accuracy, the first step is to predict all the test stimulus vector representations using a trained decoder model. Let S = [S_0, S_1,$\cdots$,S_n], $\hat{S}$ = [$\hat{S}_0$, $\hat{S}_1$,$\cdots$,$\hat{S}_n$] denote the "true" (text-derived) and predicted stimulus representations for n test instances resp. Given a pair (i, j) such that $0 \leq i, j \leq n$, score is 1 if *corr*(S_i,$\hat{S}_i$) + *corr*(S_j,$\hat{S}_j$) $>$ *corr*(S_i,$\hat{S}_j$) + *corr*(S_j,$\hat{S}_i$), else 0. Here, *corr* denotes the Pearson correlation. Final pairwise matching accuracy per participant is the average of scores across all pairs of test instances. Fig. 3.2 pictorially explains how to calculate the pairwise accuracy.

3.9.0.2 Rank Accuracy

We compared each decoded vector to all the "true" text-derived semantic vectors and ranked them by their correlation. The classification performance reflects the rank r of the text-derived vector for the correct word: $1 - \frac{r-1}{\#instances-1}$. The final accuracy value for each participant is the average rank accuracy across all instances.

Chapter 4

Multi-View Brain Decoding

4.1 Methodology

4.1.1 Task Description

We train the decoder regression models on 5000 informative voxels selected from fMRI brain activations and evaluate all the models using pair-wise accuracy and rank-based decoding. Details of the informative voxel selection, the regression model, and metrics are discussed in the subsequent sections.

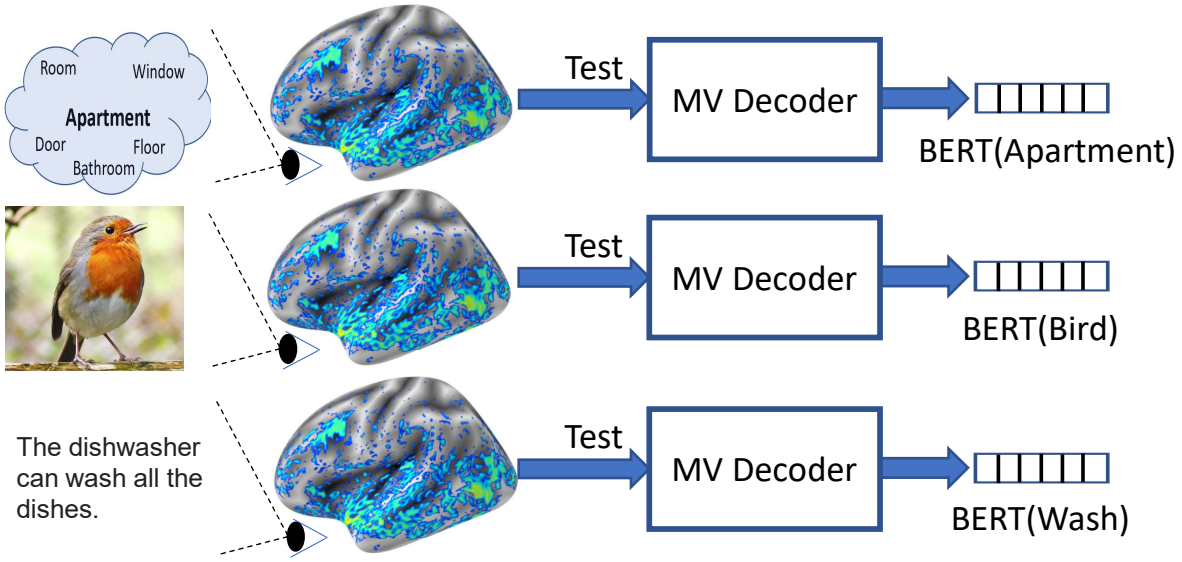


Figure 4.1: A multi-view decoder can be used to decode concepts using brain recordings for any view. Target is BERT representation of the concept word.

The main goal of each decoder model is to predict a semantic vector representation of the stimuli in each experiment. The input view (word+picture, sentence, or word-cloud) and output representation (word, sentence, or word-cloud) differ across experiments. We follow K-fold cross-validation, in which

all the data samples from K-1 folds were used for training, and the model was tested on samples of the left-out fold. We use the BERT-pooled output for obtaining output semantic representations. We also experimented with RoBERTa, but the results were very similar to BERT, and hence we omit them for lack of space.

For each subject in the dataset, for each of the three input views, we trained K=18 models (one for each fold) where each model is trained on the brain activity of 170 concepts and tested on left-out 10 concepts to predict vector representation of the concept word. The 5000 informative voxels were selected for 170 concepts in each fold, and the same voxel locations were chosen for test datasets. At test time, the input to each model can belong to any of the three views. Thus, for each subject, for each fold, we perform (1) three same-view train-test experiments and (2) six multi-view zero-shot train-test experiments with different input views at train and test time. Target is always fixed as a vector representation of the concept word. We use pairwise accuracy to report results.

4.1.2 Informative Voxel Selection

Informative Voxel selection has been explained in section 3.5. The target semantic representations are word(concept) embeddings for multi-view zero-shot concept decoding experiments.

4.2 Results and Cognitive Insights

Since we are the first to propose multi-view and cross-view tasks, unfortunately, there are no baselines to compare with. For the sake of comparison, we design a "chance-level" BERT (Random) baseline where models are trained using BERT embeddings of randomly chosen words as a target rather than BERT embeddings of the actual target word. For same-view experiments, our results are in line with that reported in Pereira et al. [72]. We also performed the experiments using the GloVe embeddings. Comparison of results for BERT and GloVe embeddings are detailed in Appendix A.

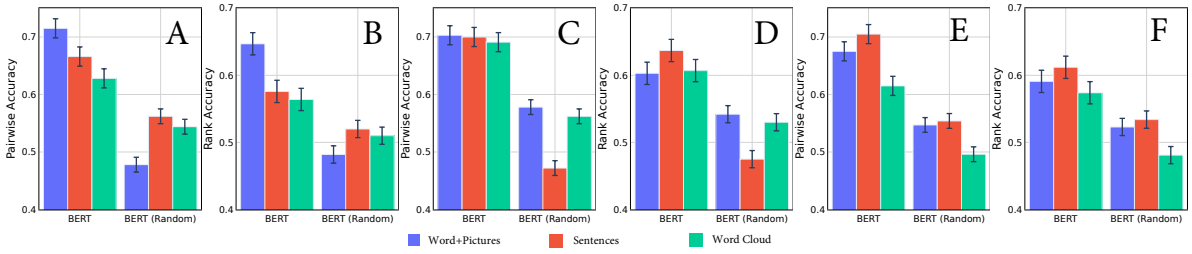


Figure 4.2: Model trained on *Word+Pictures* (A and B), *Sentences* (C and D), and *Word-Cloud* (E and F) view. MVD Pairwise and Rank accuracy when tested on Word+Picture/Sentence/Word-cloud views, averaged across all subjects.

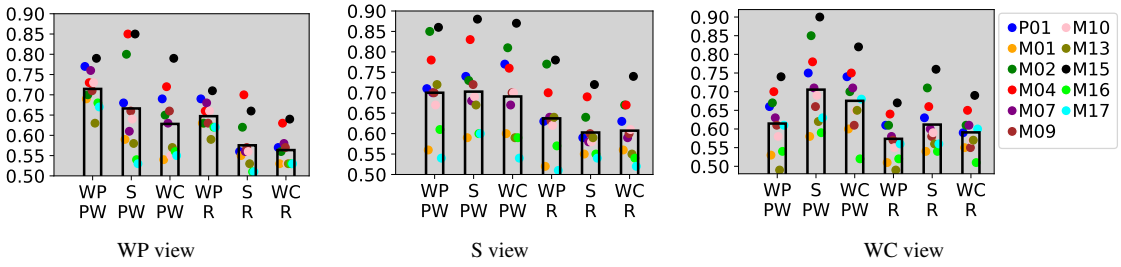


Figure 4.3: Model trained on *Word+Pictures* view (left), *Sentences* view (middle), *Word-Cloud* view (right). MVD Pairwise (PW) and Rank (R) accuracy when tested on Word+Picture (WP)/Sentence (S)/Word-cloud (WC) views. Each colored dot represents a subject. The bar plot shows averages.

4.2.1 Pairwise and Rank Accuracy Results

Fig. 4.2 and Table 4.1 show detailed results for models trained on word+picture (WP), sentence (S), and word-cloud (WC) views and tested on each of the three views. Specifically, Fig. 4.2(A) shows pairwise accuracy results when we train using the WP view but infer using voxels corresponding to any of the three views. Ground-truth is the BERT embedding vector. In comparison to the "chance-level" BERT (Random) baseline with random target vectors, our proposed BERT embedding-based method is much better. Fig. 4.3 shows subject wise results.

Test↓/Train→	WP	S	WC
WP	0.72/0.65	0.70/0.60	0.68/0.59
S	0.67/0.58	0.70/0.64	0.71/0.61
WC	0.63/0.56	0.69/0.61	0.62/0.57

Table 4.1: Multi-View Zero-shot Concept Decoder Results (Pairwise/Rank Accuracy)

4.2.1.1 Same view versus MV zero-shot

In most cases, same-view results are better than multi-view zero-shot results. However, this does not hold for the WC view, where a model trained on sentence view performs better (Left green bars in Fig. 4.2 (C and D) vs. Fig. 4.2 (E and F)).

4.2.1.2 Can we train MV decoders that can decode concepts from brain recordings for any view?

We experimented with three different MV decoders, each trained on one of the three views. Fig. 4.2 and the statistical significance test results in Table 4.2 show that either of the WP and sentence (S) views can be used to train MV decoders. This means that if we train a model with WP or S view fMRIs, and

test it using any of the three views, the results are better or equivalent to any other model. This does not hold for the WC view. Thus, an MV decoder trained with a WC view is not very effective.

Setting 1	Setting 2	p-value
Train(WP)-Test(WP)	Train(S)-Test(WP)	0.098
Train(WP)-Test(WP)	Train(WC)-Test(WP)	0.026*
Train(S)-Test(WP)	Train(WC)-Test(WP)	0.474
Train(WP)-Test(S)	Train(S)-Test(S)	0.485
Train(WC)-Test(S)	Train(S)-Test(S)	0.469
Train(WP)-Test(S)	Train(WC)-Test(S)	0.420
Train(WP)-Test(WC)	Train(WC)-Test(WC)	0.691
Train(WP)-Test(WC)	Train(S)-Test(WC)	0.134
Train(S)-Test(WC)	Train(WC)-Test(WC)	0.045*

Table 4.2: p-values for measuring if setting 1 is stat significantly better than setting 2. Only rows with * mark denote statistically significant improvements.

4.2.2 Cognitive Insights based on Distribution of Informative Voxels

Table 4.3 and Fig. 4.4 show the distribution of informative voxels among four brain networks for various MV models. In this figure, (WP, D) means input view=WP (Word+picture), brain network=DMN (D). The figure clearly shows that a lot of informative voxels belong to the visual brain region for the WP view. Also, for sentence view, a large percentage of informative voxels are from the language region.

Figs. 4.5 to 4.7 show more distribution details by zooming further into language and visual regions. When the model is trained on the WP view (unlike other views), Table 4.3 and Fig. 4.5 show that most informative voxels (about 53%) lie in the visual brain network, which is expected for the predominantly visual information-driven task.

	Word+Picture	Sentence	Word-Cloud
DMN	0.162	0.222	0.137
Visual	0.534	0.202	0.161
Language	0.177	0.246	0.192
Task-Positive	0.064	0.135	0.145

Table 4.3: Distribution of informative voxels among four brain networks for various Multi-View models

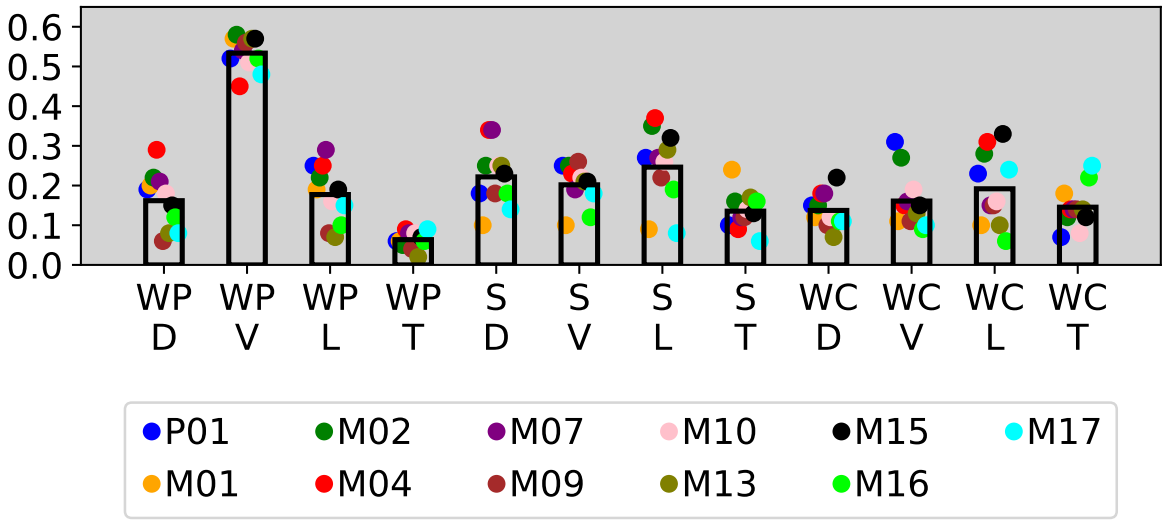


Figure 4.4: Distribution of informative voxels among four brain networks: DMN (D), Visual (V), Language (L), Task Positive (T). Models trained on Word+Picture (WP), Sentence (S) or Word-Cloud (WC) views.

We also observe that DMN and Language network voxels are higher in the sentence view than in the word cloud view. Compared to the model trained on WP view, the distribution of voxels among the four brain networks shows that the model trained on sentence view has a higher percentage of voxels among the Language, DMN, and Task-positive networks and lower in the visual network. This is in line with our understanding that linguistic and attention skills are essential for understanding sentence stimuli. As for the model trained on the WC view compared to other views, we see that the informative voxels are spread equally among all the networks. From Fig. 4.5, we observe that in all the views, the region corresponding to language processing in the left hemisphere (Language_LH) has higher informative voxels than that of the right hemisphere (Language_RH). This is in line with the left hemisphere dominance for language processing [13]. When the visual network dominates as in the case of WP view, the majority of these are located in the object processing area, followed by face and body processing areas. In the following, we investigate these two regions in detail.

In the language network, the distribution of informative voxels in the sub regions (LPTG, LMTG, LATG, LFus, LPar, LAngG, LIFGorb, LIFG, LaMFG, LpMFG, and LmMFG) are shown in Fig. 4.6. We find that regions in the posterior (LPTG), middle (LMTG), and anterior (LATG) temporal gyrus share a higher percentage of informative voxels than other regions in the language network, such as those in the middle and inferior frontal areas. This indicates that the language functions sub-served by the temporal cortex, such as comprehension and semantic processing, are critical for processing sentences as well as multi-modal integration and thus are important for decoding across multiple views. Further, brain

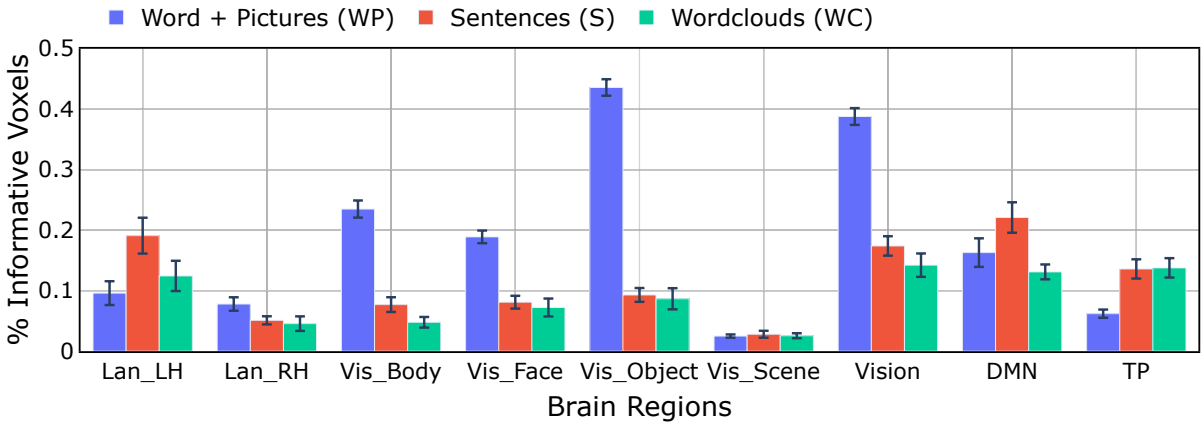


Figure 4.5: Distribution of informative voxels among nine brain regions for Multi-view Decoding

regions in the angular gyrus (LAngG) and parietal (LPar) each have >5% of informative voxels. These areas may be involved in attention, self-processing, and visio-linguistic integration.

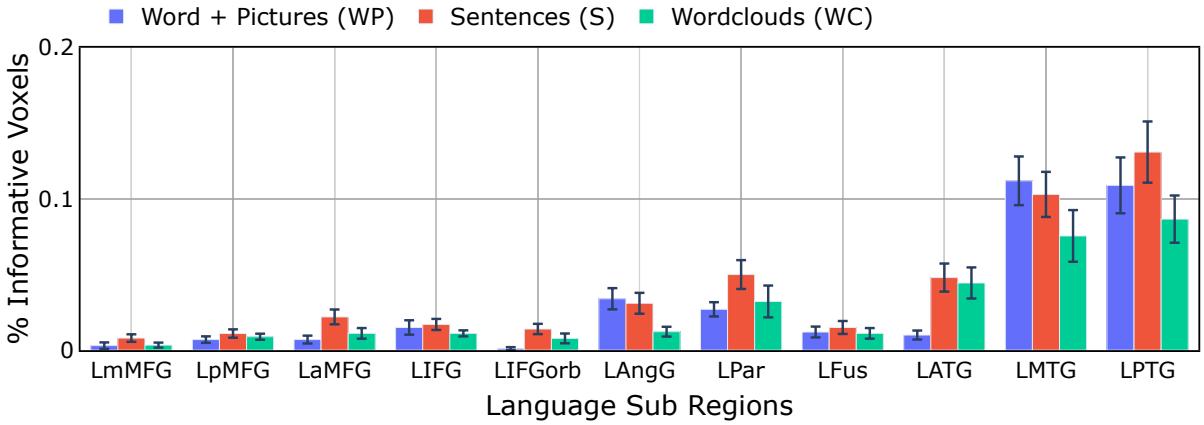


Figure 4.6: Distribution of informative voxels among eleven sub regions of Language network for MVD

Similarly, we explored the distribution of informative voxels across sub regions of the visual network, as shown in Fig. 4.7. In the visual sub regions, voxels in the bilateral occipital cortex (LLOC and RLOC) have more informative voxels than in other sub regions. In particular, the scene regions in the parahippocampal place area (such as RSC and PPA) display very few informative voxels, while the bilateral body area (REBA and LEBA) captures more voxels in the WP view. Interestingly, activation in the superior temporal sulcus (RSTS and LSTS) in all views point out its role in visio-linguistic integration. Lastly, Fig. 4.8 shows the spatial distribution of informative voxels (plotted using nilearn Python library) across models trained on different forms of stimuli (WP, S, and WC). The value of each voxel is the fraction of 11 participants for whom that voxel was among the 5000 most informative.

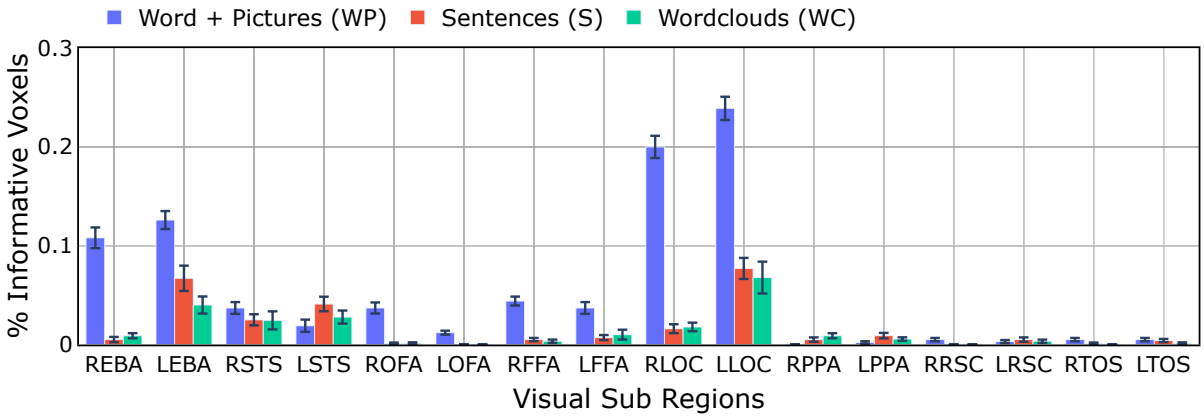
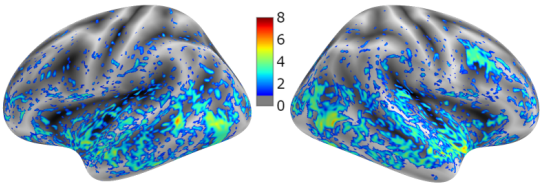
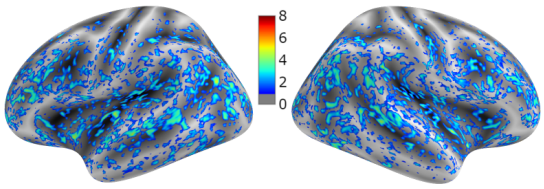


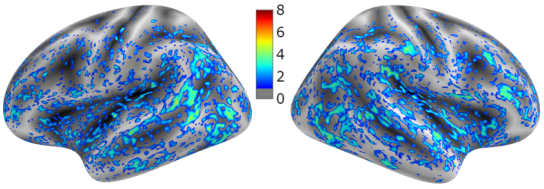
Figure 4.7: Distribution of informative voxels among sixteen sub regions of Visual network for MVD



(a) WP view



(b) S view



(c) WC view

Figure 4.8: Brain Maps for Multi-View Decoding Tasks (plotted using nilearn Python library).

4.2.3 Informative Voxel Overlap across Views

	DMN	Visual	Language	Task Positive
WP-S	0.24/0.17	0.11/0.29	0.25/0.17	0.09/0.05
WC-S	0.25/0.16	0.25/0.20	0.30/0.22	0.07/0.07
WP-WC	0.14/0.16	0.08/0.25	0.15/0.15	0.06/0.03

Table 4.4: For each pair of views and each brain network, we show coverage ratios (second task on first/first task on second) of the voxels.

Given the distribution of informative voxels across four brain networks, we further examine how these voxels from one view overlap with those from another view. Table 4.4 shows that:

- The language network has a very high overlap compared to other brain networks in the WC-S pair.
- 29% (and 25%) of visual voxels for the S (and WC) view are shared with visual voxels of the WP view. This makes sense since a large percentage of informative voxels for WP view are from the visual network.

4.3 Conclusion

We studied brain decoding in the context of Multi-view decoding tasks. Our experiments lead us to really interesting insights. Models trained on picture or sentence view are better MV decoders than models trained on word cloud view. Surprisingly, the MV decoder trained on sentence view leads to a zero-shot accuracy for word cloud stimuli, which is better than that obtained using the same-view word cloud model.

Chapter 5

Cross-View Brain Decoding

5.1 Dataset

This dataset described in section 3.3 was meant for single-view decoding and hence follows a star schema (concept at the center and specific views like word+picture, sentence, and word cloud around it). Clearly, we cannot use this dataset as is for cross-view decoding (CVD). For example, for the image captioning CVD task, it is wrong to take an fMRI with the stimuli being a picture showing a red bird sitting on a tree branch, and use it to decode a sentence "A green bird flying in the sky".

To enable cross-view decoding tasks, it was critical to build direct pairwise-view relationships (picture-sentence, picture-word cloud, sentence-word cloud, and word cloud-sentence). In other words, it was necessary to have captions and tags for image-view, keywords for sentence-view, and 3-4 sentences corresponding to wordcloud-view. Hence, we augment the dataset in Pereira et al. [72] by obtaining target annotations manually. For example, for the fMRI associated with picture p, we manually annotated it with target sentence s'="A red bird sitting on a tree branch". Pairs like (p, s') are then used to train model for image captioning. Note that these manual annotations do not involve obtaining more fMRIs.

Task	Input	Output (View type)
Image captioning	Word+Picture fMRI	Caption (Sentence)
Image tagging	Word+Picture fMRI	Image tags (Word Cloud)
Keyword extraction	Sentence fMRI	Keywords (Word Cloud)
Sentence formation	Word-cloud fMRI	Sentence

Table 5.1: Cross-View Decoding Task Definitions

5.2 Methodology

5.2.1 Task Description

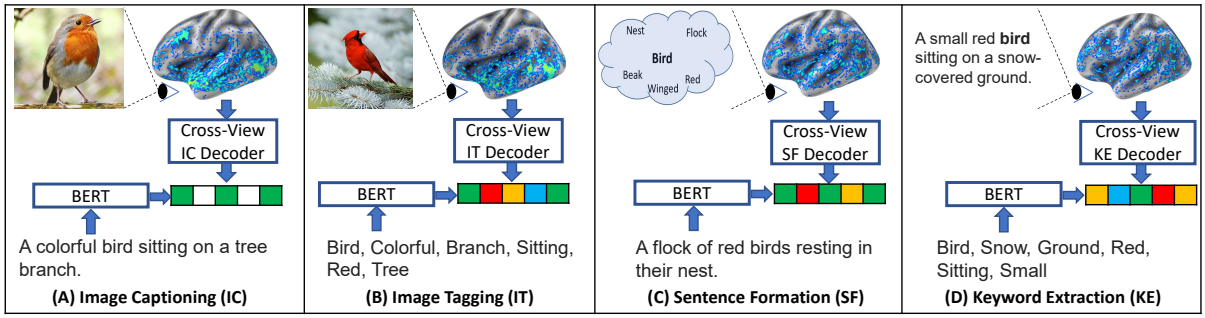


Figure 5.1: Cross-View Decoding Task (Input, output) Examples.

We train the decoder regression models on 5000 informative voxels selected from fMRI brain activations and evaluate all the models using pair-wise accuracy and rank-based decoding. Details of the informative voxel selection, the regression model, and metrics are discussed in the subsequent sections. The main goal of each decoder model is to predict a semantic vector representation of the stimuli in each experiment. The input view (word+picture, sentence, or word-cloud) and output representation (word, sentence, or word-cloud) differ across experiments. We follow K-fold cross-validation, in which all the data samples from K-1 folds were used for training, and the model was tested on samples of the left-out fold. We use the BERT-pooled output for obtaining output semantic representations. We also experimented with RoBERTa, but the results were very similar to BERT, and hence we omit them for lack of space.

For each subject in the dataset, we learn models for the four cross-view decoding tasks (IC, IT, KE, SF) using 18 fold cross-validation. The input and output for each of these tasks is shown in Table 5.1. Fig. 5.1 shows an example for each task. As before, we use 5000 informative voxels, computed separately for each of the 11 subjects and each of the four tasks. The regression target is semantic vector representation.

5.2.2 Informative Voxel Selection

Informative Voxel selection has been explained in section 3.5. Target semantic representations are ‘word or sentence or word-cloud’ embedding for cross-view decoding experiments depending on type of task.

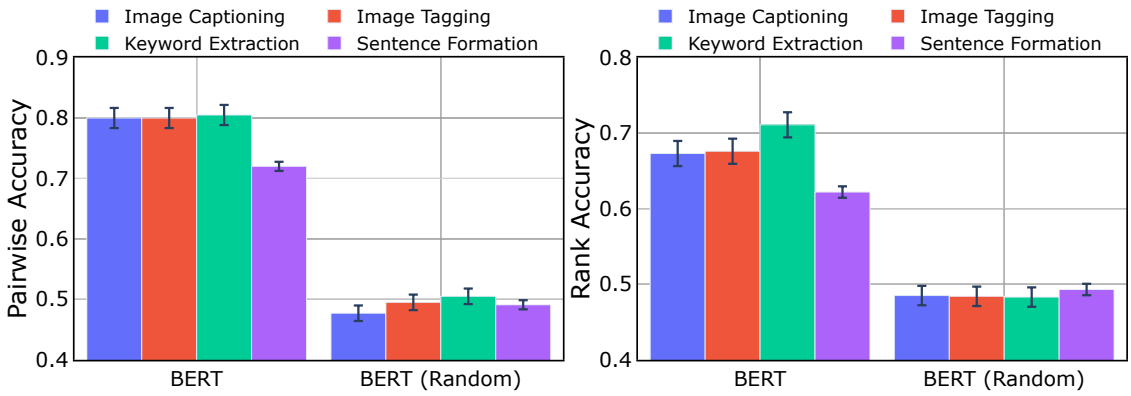


Figure 5.2: Cross-View Decoding Pairwise and Rank accuracy for Image Captioning (IC), Image Tagging (IT), Keyword Extraction (KE), and Sentence Formation (SF) averaged across all the subjects.

5.3 Results and Cognitive Insights

Since we are the first to propose multi-view and cross-view tasks, unfortunately, there are no baselines to compare with. For the sake of comparison, we design a "chance-level" BERT (Random) baseline where models are trained using BERT embeddings of randomly chosen words as a target rather than BERT embeddings of the actual target word. For same-view experiments, our results are in line with that reported in Pereira et al. [72]. We also performed the experiments using the GloVe embeddings. Comparison of results for BERT and GloVe embeddings are detailed in Appendix B.

5.3.0.1 Pairwise and Rank Accuracy Results

Fig. 5.2 illustrates pairwise and rank accuracy for Image Captioning (IC), Image Tagging (IT), Sentence Formation (SF), and Keyword Extraction (KE). Subject wise results are reported in Fig. 5.3. We observe that:

- Our proposed BERT embedding-based method is much better compared to the "chance-level" baseline with random target vectors.
- For all the four tasks, pairwise accuracy is $\sim$80%, and rank-based accuracy is $\sim$70% (except for SF), which shows that CVD is possible with good accuracy.

5.3.0.2 Cognitive Insights based on Distribution of Informative Voxels

Fig. 5.4 shows the distribution of informative voxels among nine brain regions across all four tasks. As expected, a high percentage of visual voxels are involved in IC and IT tasks, and a high percentage of language voxels are involved in the SF and KE tasks, especially in the left hemisphere. Further, from

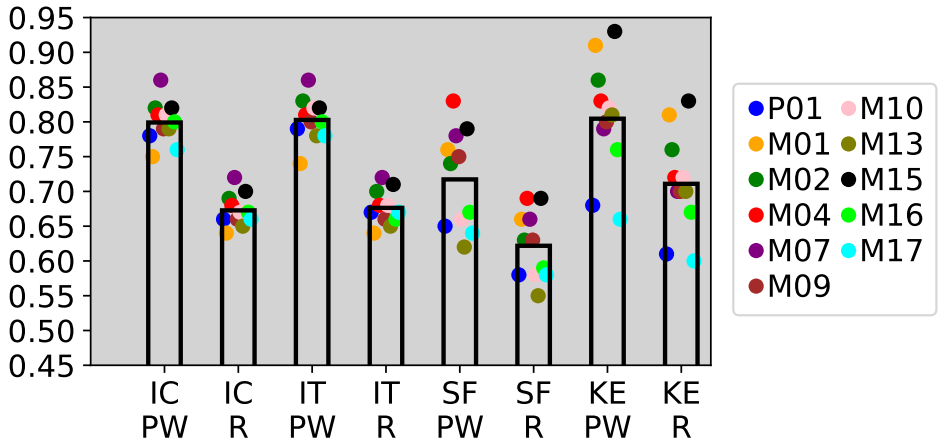


Figure 5.3: CVD Pairwise (PW) and Rank (R) accuracy for IC, IT, KE and SF tasks. Each colored dot represents a subject. The bar plot shows averages.

	IC	IT	SF	KE
DMN	0.114	0.067	0.152	0.214
Visual	0.572	0.736	0.154	0.236
Language	0.116	0.081	0.182	0.275
Task Positive	0.045	0.007	0.141	0.118

Table 5.2: Distribution of informative voxels among four brain networks for all 4 CVD Tasks.

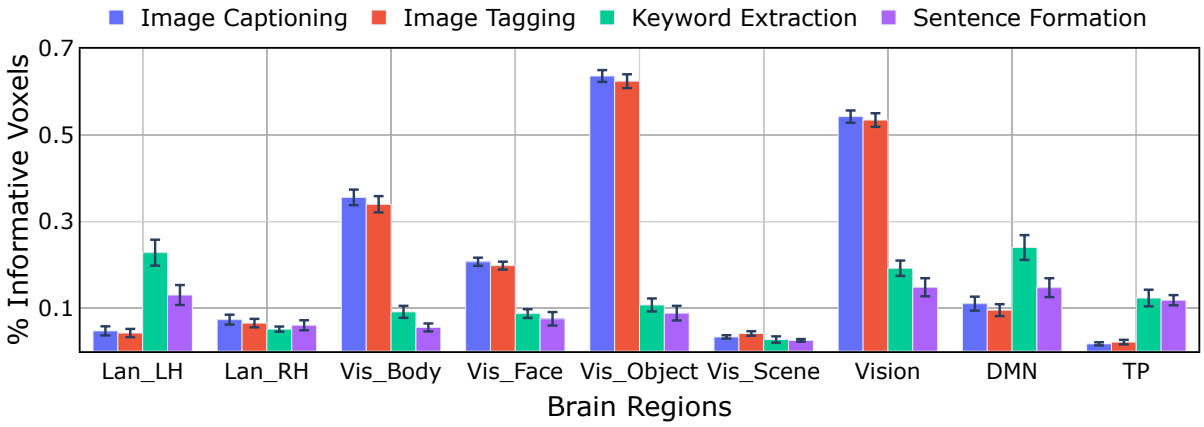


Figure 5.4: Distribution of informative voxels among nine brain regions for CVD tasks.

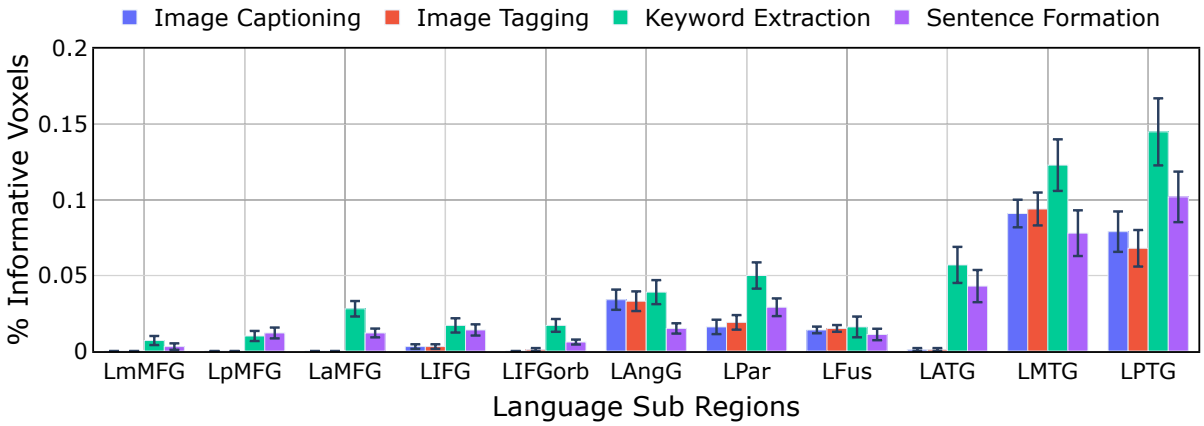


Figure 5.5: Distribution of informative voxels among 11 sub regions of Language network for CVD tasks.

Table 5.2, and Fig. A.4, we observe that IC involves relatively higher language voxels compared to IT. This could be because generating a caption involves a higher level of language (sequence) skills than generating a set of keywords.

To further investigate the informative voxel distribution across Language and Visual networks, we display the sub region voxels distribution for the Language network in Fig. 5.5, and for the Visual network in Fig. 5.6. In all the tasks, the left hemisphere language network activation is dominated by activity in the temporal gyrus (middle: LMTG and posterior: LPTG) but more in the KE task. This clearly demonstrates the importance of language comprehension and semantic process common across the cross-view tasks. Further, the common activation in the angular gyrus (LAngG) in all tasks points out the role of visio-linguistic integration critical for all the tasks. The activation profile of the vision network, in contrast, shows distinct activation differences across the tasks (IC & IT vs. KE & SF). IC and IT tasks are related to a higher proportion of informative voxels in the primary visual regions in the

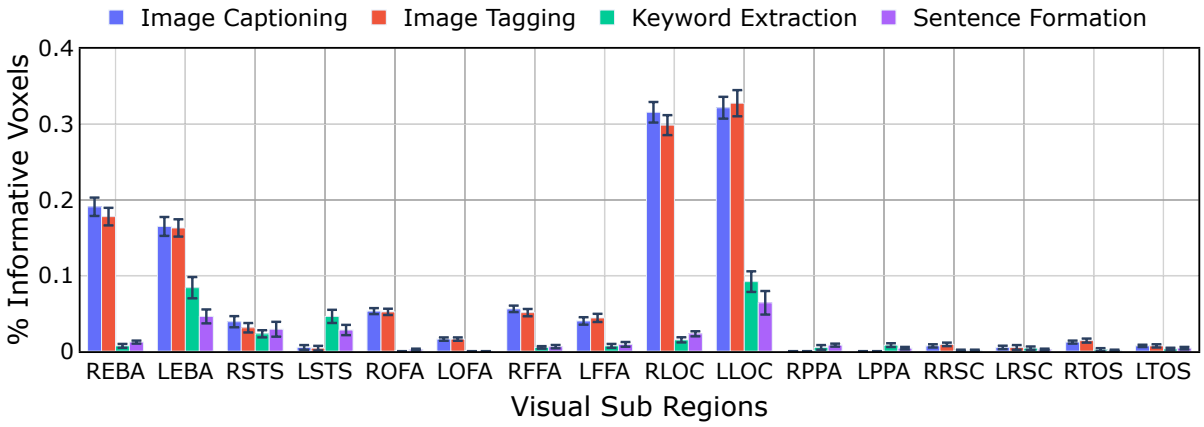


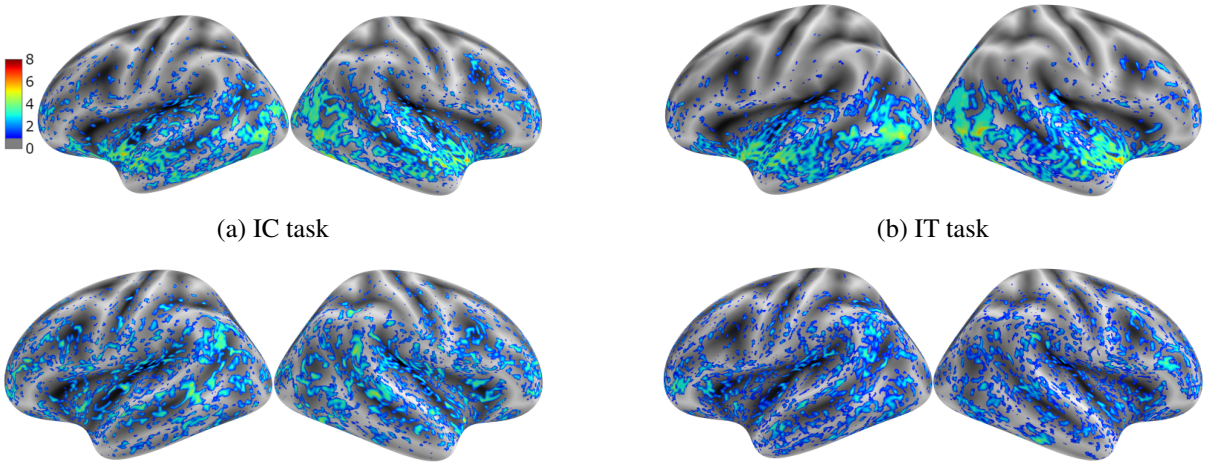
Figure 5.6: Distribution of informative voxels among 16 sub regions of Visual network for CVD tasks.

lateral occipital areas (LLOC, RLOC) and bilateral extrastriate body-related areas (REBA and LEBA). Domination of activation in the vision network in captioning and tagging tasks (IC and IT) as compared to predominantly sentence processing based tasks (KE and SF) is along expected lines.

	DMN	Visual	Language	Task Positive
IC-IT	0.27/0.44	0.70/0.54	0.32/0.45	0.07/0.32
IC-KE	0.31/0.17	0.11/0.27	0.28/0.12	0.12/0.05
IC-SF	0.16/0.12	0.07/0.25	0.14/0.09	0.08/0.03
IT-KE	0.27/0.08	0.08/0.25	0.22/0.07	0.05/0.01
IT-SF	0.13/0.05	0.06/0.27	0.10/0.05	0.04/0.00
KE-SF	0.19/0.26	0.20/0.29	0.22/0.32	0.09/0.08

Table 5.3: For each pair of CVD tasks and each brain network, we show coverage ratios (second task on first/first task on second) of the voxels.

The brain maps (see Fig. 5.7) corresponding to the IC and IT tasks clearly activate the visual cortex and the temporal cortex, the areas known for visual processing and object identification. On the other hand, the brain maps of KE and SF exhibit diffuse activation that includes the temporal and frontal regions known to be related to the sentence semantics. None of the maps show a left-hemisphere bias, which is often found in such semantic-related maps. Lack of frontal-lobe activation and the concentration of informative voxels in the sensory cortex suggest that the cross-view embedding may rely on some non-abstract domain-specific encoding rather than higher-level semantic concept encoding.



(a) IC task

(b) IT task

(c) KE task

(d) SF task

Figure 5.7: Brain Maps for Cross-View Decoding Tasks (plotted using nilearn Python library).

5.3.0.3 Informative Voxel Overlap across Tasks

Given the distribution of informative voxels across four brain networks, we further examine how these voxels from one task overlap with those from another task. Table 5.3 shows that:

- Many voxels overlap across different brain networks for IC and IT tasks. This is expected since the two tasks are very related. Interestingly, 44% of DMN voxels needed for IT are shared with IC. Similarly, as high as 70% of visual voxels needed for IC are shared with IT.
- Similarly, KE and SF share a very good overlap across different brain networks, which is expected given the textual nature of the two tasks.

5.4 Conclusion

We studied brain decoding in the context of cross-view decoding tasks. We studied four cross-view decoding tasks: image captioning, image tagging, sentence formation, and keyword extraction. We show that cross-view decoding is feasible with good accuracy. Brain network distribution analysis reveals insights about the importance of various parts of the brain for each of these tasks.

Chapter 6

Abstract-Concrete Brain Decoding

6.1 Introduction

Neuroimaging techniques such as fMRI record brain activation while participants experience a stimulus. The concreteness of concepts defines how well our brain is able to imagine them. We hypothesise that brain activation would be distinctly different when participants view stimuli corresponding to concrete words versus abstract words. Specifically, we expect primary sensory areas to engage more in the former. To study this, we used a multi-view dataset with each concept being displayed as a picture, as a word in a sentence and also as a wordcloud [72]. We investigate how well a machine learning-based decoder trained on concrete concepts can decode both concrete and abstract concepts and vice versa for the model trained on abstract concepts. We find that a model trained on stimuli related to concrete concepts yields better accuracy than the model trained on abstract concepts. We also explore the contribution of voxels from different brain regions in this decoding process. Analysis of the contribution of different brain networks reveals exciting cognitive insights.

The motivation for this experiment is to investigate how the brain perceives linguistic meaning only from abstract or concrete concepts across different views. Experimentation with the same-paradigm decoding, i.e., "abstract-train and abstract-test" or "concrete-train and concrete-test" has already been done, but the work was limited to specific parts of speech and not the complete set of concepts [99]. Additionally, the experiments reported here also implement cross-paradigm decoding, i.e., "abstract-train and concrete-test" and vice-versa for the complete set of concepts. Further, we experimented with a balanced dataset with (randomly chosen) 64 concrete and 64 abstract concepts and found similar results.

6.2 Methodology

6.2.1 Task Description

We train the decoder regression models on 5000 informative voxels selected from fMRI brain activations and evaluate all the models using pair-wise accuracy and rank-based decoding. Details of the informative voxel selection, the regression model, and metrics are discussed in the subsequent sections.

The main goal of each decoder model is to predict a semantic vector representation of the stimuli in each experiment. The input view (word+picture, sentence, or word-cloud) and output representation (word, sentence, or word-cloud) differ across experiments. We follow K-fold cross-validation, in which all the data samples from K-1 folds were used for training, and the model was tested on samples of the left-out fold. We use the BERT-pooled output for obtaining output semantic representations. We also experimented with RoBERTa, but the results were very similar to BERT, and hence we omit them for lack of space.

Out of 180 concepts in Dataset, 116 are concrete concepts, and others are abstract. For each subject in Dataset, for each of the three views, we first train a regression model on fMRI activations of all abstract concepts and test it on concrete concepts data. Similarly, we use all concrete concept fMRIs for training and infer on the abstract concepts data.

The 5000 informative voxels were selected for abstract and concrete only concepts separately, and the same voxel locations were chosen for train and test datasets. The regression target is vector representation of the concept word. The motivation for this experiment is to investigate how the brain perceives linguistic meaning only from abstract or concrete concepts across different views.

6.2.2 Informative Voxel Selection

Informative Voxel selection has been explained in section 3.5. The target semantic representations are word(concept) embeddings for Abstract-Concrete decoding experiments.

6.3 Results and Cognitive Insights

BERT (Bidirectional Encoder Representations from Transformers) is a popular Transformer based natural language processing model which has been shown to perform extremely well across many NLP tasks [4]. Hence, we use BERT for generating our word representations. The results for all four experiments are presented per view using BERT word embeddings in Table 6.1. We report average pairwise accuracy across all participants. We also performed the experiments using the GloVe embeddings. Comparison of results for BERT and GloVe embeddings are detailed in Appendix C.

We observe that the model trained on concrete concepts (the last two columns of Table 6.1) provides better accuracy than the model trained on abstract concepts (the first two columns of Table 6.1) while decoding both abstract and concrete concepts. Interestingly, the model trained on concrete concepts

	Abs2Abs	Abs2Conc	Conc2Conc	Conc2Abs
Word+Picture	0.647	0.648	0.764	0.697
Sentence	0.570	0.604	0.735	0.642
WordCloud	0.520	0.522	0.581	0.587

Table 6.1: Pairwise Accuracy for the following models: Abs2Abs: Model trained on abstract concepts and tested on abstract concepts. Abs2Conc: Model trained on abstract concepts and tested on concrete concepts. Conc2Conc: Model trained on concrete concepts and tested on concrete concepts. Conc2Abs: Model trained on concrete concepts and tested on abstract concepts.

	DMN	Visual	Language	Task Positive
Word+Picture	0.14	0.15	0.12	0.17
Sentence	0.16	0.10	0.15	0.17
WordCloud	0.10	0.10	0.09	0.17

Table 6.2: Distribution of informative voxels among four brain networks for model trained on abstract concepts

	DMN	Visual	Language	Task Positive
Word+Picture	0.08	0.45	0.08	0.1
Sentence	0.15	0.14	0.14	0.16
WordCloud	0.12	0.10	0.10	0.19

Table 6.3: Distribution of informative voxels among four brain networks for model trained on concrete concepts

outperformed the model trained on abstract concepts for decoding abstract concepts. The distribution of informative voxels among the four brain networks for both the decoder models are presented in Tables 6.2 and 6.3, leading to the following cognitive insights.

1. As expected, the visual brain network is very important for processing Word+Picture stimuli for concrete concepts. Surprisingly, this is not the case for abstract concepts.

2. For both Word+Picture and Sentence views, language and DMN brain networks are more important for processing abstract concepts than concrete concepts.

3. For the WordCloud view, the contribution of the task network is relatively larger compared to other networks, possibly because processing a word cloud recruits more attentional resources.

6.4 Conclusion

We studied brain decoding in the context of concreteness of concepts being visualised using three different types of view. We investigate the accuracy of brain decoding models for abstract versus concrete concepts with respect to different views. We find that models trained on concrete concept stimuli lead to better decoding results than models trained on abstract concept stimuli. We also report the contribution of voxels from different brain areas while processing these views and compare them based on the abstractness or concreteness of the concepts.

Chapter 7

Brain Encoding

7.1 Language Taskonomy[1]

7.1.1 Introduction

Brain encoding aims at constructing neural brain activity given an input stimulus. Since the discovery of the relationship between language stimuli and functions of brain networks using fMRI [for ex., Constabel et al.[23]], researchers have been interested in understanding how the neural encoding models predict the fMRI brain activity. Several brain encoding models have been developed to (i) understand the ventral stream in biological vision [108, 50, 9], and (ii) to study the higher-level cognition like language processing [34, 80, 81].

Recently, Transformer [95] based models like BERT [25] have been found to be very effective across a large number of natural language processing (NLP) tasks. These Transformer based models have been pretrained on millions of text instances in an unsupervised manner and further finetuned to specialize for various NLP tasks. Natural language understanding requires integrating several cognitive skills like syntactic parsing of the language structure, identifying the named entities, capturing the word meaning in the context, coreference resolution, etc. Learning from massive corpora enables these models to excel at cognitive skills required for language understanding. Interestingly, such Transformer-based neural representations have been found to be very effective for brain encoding as well [80].

Recently, a study using multiple computer vision tasks has shown that 3D vision task models predict better fMRI brain activity than 2D vision task models [96] for visual stimuli. Inspired by the success of correlations in the vision field [96], and brain encoding study of a variety of language Transformer models [80, 19, 18], we build neural language taskonomy models for brain encoding and aim to find NLP tasks that are most explanatory of brain activations for reading and listening tasks.

[1] **This is part of the additional work and not the main focus of the thesis**

7.1.2 Related Works

Older methods for text-based stimulus representation include text corpus co-occurrence counts [56, 69, 45], syntactic and discourse features [101]. In recent times, both semantic and experiential attribute models have been explored for text-based stimuli. Semantic representation models include distributed word embeddings [71, 4, 72, 92, 43, 99], sentence representation models [85, 92, 86], recurrent neural networks [48, 66], and Transformer-based language models [34, 92, 81, 63, 64]. Experiential attribute models represent words in terms of human ratings of their degree of association with different attributes of experience, typically on a scale of 0-6 [2, 5, 12, 49, 18, 6] or binary [39, 97]. Fine-grained details such as lexical, compositional, syntactic, and semantic representations of narratives are factorized from Transformer-based models and utilized for training encoding models. The resulting models are better able to disentangle the corresponding brain responses in fMRI [18].

7.1.3 Task Description

To explore how and where contextual language features are represented in the brain when reading sentences and listening to stories, we extract different features spaces describing each stimulus sentence and use them in an encoding model to predict brain responses. Our reasoning is as follows. If a feature is a good predictor of a specific brain region, information about that feature is likely encoded in that region. In this work, for both datasets(Pereira [72] and Narratives-Pieman [58]), we train fMRI encoding models using Ridge regression on stimuli representations obtained using a variety of NLP tasks. The main goal of each fMRI encoder model is to predict brain responses associated with each brain region given a stimuli. In all cases, we train a model per subject separately. Following literature on brain encoding [19, 91], we choose to use a ridge regression model instead of more complicated models. We plan to explore more such models as part of future work.

We follow K-fold (K=10) cross-validation. All the data samples from K-1 folds were used for training, and the model was tested on samples of the left-out fold.

7.1.4 Findings

For **Pereira Dataset(Reading sentences)**, we observe that tasks such as Coreference Resolution(CR), Named Entity Resolution(NER), Semantic Role Labelling(SRL), and Shallow Syntax(SS) appear to have a better correlation to the brain responses compared to the other tasks. These results demonstrate that when reading a sentence, information processing operations related to recognizing named entities, labeling semantic roles to the constituents of a sentence, identifying the references from a sentence to the given topic (concept), and syntactic processing may be engaged.

Further, we observe that the ROI corresponding to language processing in the left hemisphere (Language_LH) has higher encoding performance than that of the right hemisphere (Language_RH). This is in line with the left hemisphere dominance for language processing [13]. Also, lateral visual ROIs such as Vision_Object, Vision_Body, Vision_Face, and Vision ROIs display higher correlation with the

language tasks associated with named entities (NER), relating the entities (CR), and syntax processing (SS). Higher correlations with all the visual brain regions point to the possible alignment of visual and language regions for semantic understanding [74] in a reading task.

For **Narratives-Pieman Dataset (Listening Stories)**, Tasks such as Paraphrase Detection(PD), Summarization(Sum), and Natural Language Inference(NLI) seem to yield better performance in predicting the brain responses than the other NLP tasks across all the ROIs. We observe that the profiles of performance show low scores in the early auditory cortex (EAC), auditory association cortex (AAC); average scores in TPOJ and DFL; and superior scores in PMC. This aligns with the known language hierarchy for spoken language understanding [57].

Further, we see that the bilateral posterior medial cortex (PMC) associated with higher language function exhibits a higher correlation among all the brain ROIs. ROIs, including bilateral TPOJ and bilateral DFL, yield higher correlations with the five NLP tasks, which is in line with the language processing hierarchy in the human brain.

In summary, different and distinct language Taskonomy features seem to be related to the encoding performance in reading versus listening tasks. CR, NER, SRL, and SS perform better for reading. PD, Sum, and NLI perform better for listening. While listening the subject is cognitively more involved in the activity compared to reading [15]. Thus, it makes sense that shallow tasks like NER and SS are useful for reading while more complex NLP tasks like PD, Sum and NLI are effective for encoding listening stimuli.

7.1.5 Conclusion

In this work, we studied the effectiveness of task-specific NLP models for brain encoding. We observe that building individual encoding models and exploiting existing relationships among models can provide a more in-depth understanding of the neural representation of language information. Our experiments on Pereira and Narrative datasets lead to interesting cognitive insights.

7.2 Visio-Linguisitic Brain Encoding[2]

7.2.1 Introduction

Brain encoding aims at constructing neural brain activity recordings given an input stimulus. The two most studied forms of stimuli include vision and language. Since discovering of the relationship between language/visual stimuli and functions of brain networks [23, 90], researchers have been interested in understanding how the neural encoding models predict the fMRI (functional magnetic resonance imaging) brain activity. Recently, several brain encoding models were developed to (i) understand the ventral stream in biological vision [108, 50, 9] and (ii) study higher-level cognition like language pro-

[2]**This is part of the additional work and not the main focus of the thesis**

cessing [34, 80, 81]. Previous work has mainly focused on independently understanding vision and text stimuli. However, the biological systems perceive the world by simultaneously processing high-dimensional inputs from diverse modalities such as vision, auditory, touch, and proprioception [47]. In particular, how the brain effectively processes and provides its visual understanding through natural language and vice versa is still an open question in neuroscience.

Earlier studies mainly were related to neural encoding models that predict brain activity using representations of single-mode stimuli: visual or text. Convolutional neural networks (CNNs) were known to encode semantics from visual stimuli effectively. Interestingly, intermediate layers in deep CNNs trained on the ImageNet [24] categorization task can partially account for how neurons in intermediate layers of the visual system respond to any given image [106, 108, 36, 107, 96]. Similar to CNN based visual encoding models, various studies leveraged neural models like deep recurrent neural networks (RNNs), Transformer [95] based language models such as BERT [25], RoBERTa [53], and GPT-2 [77] to predict the brain activity corresponding to semantic vectors of linguistic items, including words, phrases, sentences, and paragraphs [34, 80].

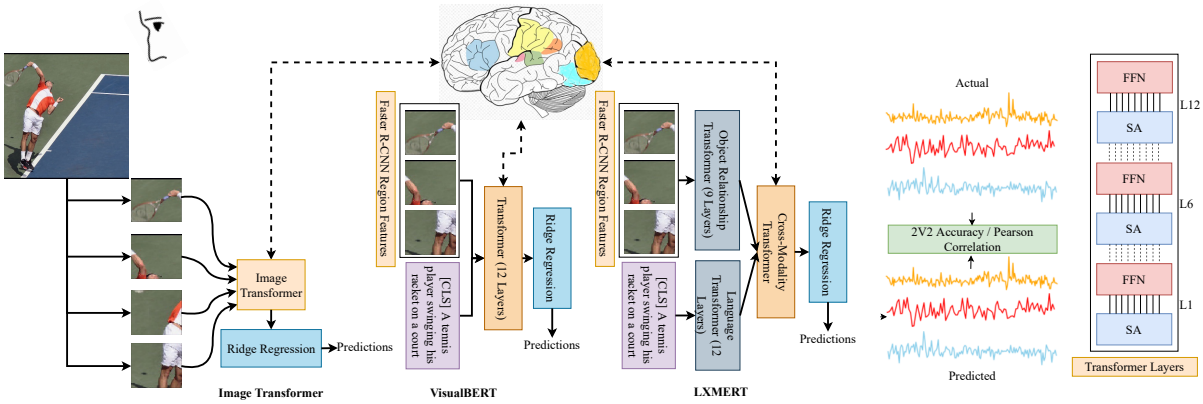


Figure 7.1: Logical architecture of the proposed approach: We use features from image/multi-modal Transformers (like ViT, VisualBERT, and LXMERT) as input to the regression model to predict the fMRI activations for different brain regions. We evaluate the brain encoding results by computing 2V2 accuracy and Pearson correlation between actual and predicted activations. We also perform layer-wise correlation analysis between transformer layers and brain regions.

Unlike previous studies, which focus on single-modality (visual or language stimuli), some authors demonstrated that multi-modal models formed by combining text-based distributional information with visual representations provide a better proxy for human-like intelligence [3, 66]. However, these methods extract representations from each mode separately (image features from CNNs and text features from pretrained embeddings) and then perform a simple late-fusion. Thus, they cannot effectively exploit semantic correspondence across the two modes at different levels. Such late-fusion-based multi-modal models are the closest to our work, and our experiments show that our models outperform them.

Recently, Transformer-based models were found to be very effective than CNNs, in all language and image-related tasks [25]. Image-based transformer models like ViT [26], DEiT [93], and BEiT [8] have been shown to provide excellent results compared to traditional CNNs on image classification tasks. Also, multi-modal Transformers like VisualBERT [52], LXMERT [88], and CLIP [76] have shown excellent results on visio-linguistic tasks like visual question answering, visual common-sense reasoning. Inspired by the success of language, image, and multi-modal Transformers, we build multi-modal transformer models to learn the joint representations of image content and natural language and use them for brain encoding. Overall, in this work, we investigate whether *image-based and multi-modal Transformers* can accurately perform fMRI encoding on the *whole brain*. Fig. 7.1 illustrates our method for brain encoding.

7.2.2 Task Description

We train fMRI encoding models using Ridge regression on stimuli representations obtained using various models for both datasets (Pereira [72] and BOLD5000 [20]), as shown in Fig. 7.1. The main goal of each fMRI encoder model is to predict fMRI voxel values for each brain region given stimuli. In all cases, we train a model per subject separately. Different brain regions are involved in processing stimuli involving objects and scenes. Similarly, some regions specialize in understanding vision inputs while others interpret linguistic stimuli better.

To evaluate the generalizability of our models across objects vs. scenes understanding, we also perform cross-data experiments where the train images belong to one sub-dataset, and the test images belong to the other sub-dataset. Thus, for each subject, we perform (1) three same-sub-dataset train-test experiments and (2) six cross-sub-dataset train-test experiments.

Full dataset fMRI Encoding: Whenever we train and test on the same dataset, we follow K-fold (K=10) cross-validation. All the data samples from K-1 folds were used for training, and the model was tested on samples of the left-out fold.

Cross-data fMRI Encoding: In the BOLD5000 dataset, we have three sub-datasets: COCO, ImageNet, and Scenes. ImageNet images mainly contain objects. Scenes images are about natural scenes, while COCO images relate to both objects and scenes. For each of the three sub-datasets, we perform K-fold (K=10) cross-validation within the sub-dataset.

We used the following models to extract stimuli features:

Pretrained CNNs: We extract the layer-wise features from different pretrained CNN models such as VGGNet19 [82], ResNet50 [40], InceptionV2ResNet [87], and EfficientNetB5 [89], and use them for predicting fMRI brain activity. We use adaptive average pooling on each layer to get features for each image.

Pretrained text Transformers: RoBERTa [53] builds on BERT's language masking strategy and has been shown to outperform several other text models on the popular GLUE NLP benchmark. We use the average-pooled representation[3] from RoBERTa to encode text stimuli.

[3] Average-pooled representation gave us better results compared to using the CLS representation.

Image Transformers: We used three image Transformers: Vision Transformer (ViT), Data Efficient Image Transformer (DEiT), and Bidirectional Encoder representation from Image Transformer (BEiT). Given an image, image Transformers output two representations: pooled and patches. We experiment with both representations.

Late-fusion models: In these models, the stimuli representation is obtained as a concatenation of image stimuli encoding obtained from pretrained CNNs and text stimuli encoding obtained from pretrained text Transformers. Thus, we experiment with these late-fusion models: VGGNet19+RoBERTa, ResNet50+RoBERTa, InceptionV2ResNet+RoBERTa and EfficientNetB5+RoBERTa. These models do not incorporate real information fusion but do concatenation across modalities.

Multi-modal Transformers: We experiment with these multi-modal Transformer models: Contrastive Language-Image Pre-training (CLIP), Learning Cross-Modality Encoder Representations from Transformers (LXMERT), and VisualBERT. These Transformers take both image and text stimuli as input and output a joint visio-linguistic representation. Specifically, the image input for these models comprises region proposals as well as bounding box regression features extracted from Faster R-CNN [79] as input features. These models incorporate information fusion across modalities at different levels of processing using co-attention and hence are expected to result in high-quality visio-linguistic representations.

7.2.3 Findings

We calculate the 2V2 accuracy and Pearson correlation results for models trained with different input representations (extracted from the best-performing layer of every pretrained CNN model and the last output layer of the Transformer model) on the two datasets: BOLD5000 and Pereira .

BOLD5000: We make the following observations:

- On both 2V2 accuracy and Pearson correlation, VisualBERT is better across all the models.
- Other multi-modal Transformers such as LXMERT and CLIP perform as good as pretrained CNNs. We observed that image Transformers perform worse than pretrained CNNs. Late fusion models and RoBERTa has the least performance.
- Late visual areas such as OPA (scene-related) and LOC (object-related) display a higher Pearson correlation with multi-modal Transformers, which is in line with the visual processing hierarchy. A higher correlation with all the visual brain ROIs with multi-modal Transformers demonstrates the power of jointly encoding visual and language information.
- The patch representation of image Transformers shows an improved 2V2 accuracy and Pearson correlation compared to the Pooled representation.
- Both InceptionV2ResNet and ResNet-50 have better performance among uni-modality models.

Pereira: We make the following observations:

- Similar to BOLD5000, multi-modal Transformers such as VisualBERT and LXMERT perform better.
- Lateral visual areas such as Vision_Object, Vision_Body, Vision_Face, and Vision areas display higher correlation with multi-modal Transformers. A higher correlation with all the visual brain regions, language regions, DMN, and TP with multi-modal Transformers, demonstrates that the alignment of visual-language understanding helps.

7.2.4 Conclusion

We studied the effectiveness of multi-modal modeling for brain encoding. We found that VisualBERT, which jointly encodes text and visual input using cross-modal attention at multiple levels, performs the best. Our experiments on BOLD5000 and Pereira datasets lead to interesting cognitive insights. These insights indicate that fMRIs reveal reliable responses in scenes and object selection visual brain areas, which shows that cross-view decoding tasks like image captioning or image tagging are practically possible with reasonable accuracy. We plan to explore this as part of future work. We also plan to explore correlations between brain voxel space and representational feature space in the future. Finally, the combined strength of joint (audio, vision, and text) modalities remains to be investigated.

Chapter 8

Conclusion and Future Directions

8.1 Summary of the Main Findings

Brain decoding is a fascinating area of research that involves analyzing brain activity data to decode the information encoded in it. Previous efforts have mostly focused on single-view analysis, which limits their ability to build complex brain decoding systems. Inspired by the multi-lingual and cross-lingual modelling techniques used in natural language processing, this thesis proposes novel brain decoding setups for multi-view decoding, cross-view decoding, and abstract versus concrete decoding. These setups aim to take a step forward in building complex brain decoding systems that can recognize and decode concepts across different views.

The experiments conducted in this study led to some interesting and exciting findings. Firstly, the multi-view decoding (MVD) models achieved an average pairwise accuracy of $\sim$0.68 across view pairs. In comparison, the cross-view decoding (CVD) models achieved an average pairwise accuracy of $\sim$0.8 across tasks. The results indicate that cross-view decoding is feasible with good accuracy and is a promising area for further research.

The analysis of the contribution of different brain networks revealed exciting cognitive insights. Models trained on picture or sentence view were found to be better multi-view decoders than models trained on word cloud view. In addition, the MV decoder trained on sentence view achieved accuracy for word cloud stimuli, which was better than that obtained using the same-view word cloud model. This finding suggests that the semantic information in the sentence view is more generalized and can be used to decode other views, even when the model is not explicitly trained on them.

Moreover, the study of cross-view decoding for image captioning, image tagging, sentence formation, and keyword extraction revealed insights into the importance of various parts of the brain for each of these tasks. By analyzing the contribution of voxels from different brain areas, this study localized, for the first time, the parts of the brain involved in these cross-view tasks. This finding could pave the way for building more accurate and effective brain decoding models for these tasks.

Finally, the analysis of abstract versus concrete decoding for three different types of views showed that models trained on concrete stimuli led to better decoding results than models trained on abstract

stimuli. The study also reported the contribution of voxels from different brain areas while processing these views and compared them based on the abstractness or concreteness of the concepts. These findings provide insights into how the brain processes and decodes different types of concepts.

In conclusion, this thesis proposes novel brain decoding setups for multi-view decoding, cross-view decoding, and abstract versus concrete decoding. The findings of this study show that these setups are promising for building complex brain decoding systems that can recognize and decode concepts across different views. The study also provides insights into the importance of different brain areas for these tasks, which could be useful for building more accurate and effective brain decoding models in the future.

8.2 Limitations

Brain decoding techniques, which use neural activity to decode the content of mental representations, have gained popularity in recent years as a powerful tool for investigating the neural basis of cognitive processes. However, like any methodology, Brain Decoding has its limitations, which must be taken into account when interpreting results.

Here are some possible limitations of Brain Decoding techniques:

- **Generalization across individuals:** One potential limitation of brain decoding is that neural activity patterns can vary across individuals. This means that a decoding model trained on one individual's brain may not generalize well to other individuals, which could limit the applicability of the results to the broader population.

- **Decoding the Actual Stimuli:** Brain Decoding techniques are still limited to decoding the representation of the stimuli rather than the stimuli themselves. For example, in this thesis, all the experiments aimed to produce text embeddings rather than the actual text. One of the possible reasons for this is the limited availability of datasets which is discussed in the next point.

- **Limited Dataset:** One of the major challenges in brain decoding is the limited amount of data available for analysis. Collecting brain activity data is an expensive and time-consuming process that often requires specialized equipment and trained personnel. This limits the amount of data that can be collected for a given study and, in turn, limits the ability to train accurate and robust machine learning models for brain decoding tasks.

- **Interpretation of decoding results:** Another limitation of Brain decoding is the difficulty of interpreting decoding results. One of the primary issues with interpreting decoding results is the question of what features of the brain activity data are being used to make the predictions. Decoding models can be highly accurate, but it is often unclear which specific brain regions or patterns of activity are driving the predictions. This can make it difficult to draw conclusions about the underlying neural mechanisms and cognitive processes.

- **Limited spatial and temporal resolution:** The spatial and temporal resolution of brain imaging techniques like fMRI, EEG, and MEG is limited, which means that it can be challenging to pinpoint the exact location and timing of neural activity accurately. This can impact the accuracy of decoding models and limit the conclusions that can be drawn from the results.

In summary, Brain Decoding techniques have limitations that must be taken into account when interpreting results. These limitations include generalization across individuals, limited datasets, the difficulty of interpreting decoding results, and limited spatial and temporal resolution of brain imaging techniques. Despite these limitations, Brain Decoding remains a valuable tool for investigating the neural basis of cognitive processes.

8.2.1 Possible solution to the Limitations

To address some of the limitations of Brain Decoding, researchers can employ several strategies:

- **Generalization across individuals:** To ensure that decoding models generalize across individuals, researchers can use a larger sample size when collecting data and apply statistical methods that account for individual variability. Additionally, it may be useful to use transfer learning, which involves training a decoding model on data from one or more participants and then fine-tuning it on data from another participant.

- **Decoding the Actual Stimuli:** One solution is to collect larger and more diverse datasets. This could involve combining data from multiple studies or sources to increase the size and diversity of the dataset or collecting new data using more efficient or cost-effective methods. Advances in natural language processing and computer vision can potentially allow decoding the actual stimuli rather than just their representation. Could techniques for data augmentation, like using GAN [35] or VAE [51] to generate fake data, be useful to increase the amount of dataset to train the models that could decode/reconstruct the stimuli?

- **Limited Dataset:** Collecting more brain activity data can help increase the amount of data available for analysis and improve the accuracy of decoding models. Additionally, sharing and creating open-source datasets can enable researchers to train more robust and accurate models.

- **Interpretation of decoding results:** To improve the interpretability of decoding models, researchers can use methods such as feature selection and visualization to identify which specific brain regions or patterns of activity are driving the predictions. Additionally, combining brain decoding with other techniques, such as neurophysiology, neuroimaging, and behavioural testing, can provide additional insights into the underlying neural mechanisms and cognitive processes.

- **Limited spatial and temporal resolution:** Advances in brain imaging techniques, such as high-density EEG [59] and functional near-infrared spectroscopy (fNIRS) [14], can provide higher spatial and temporal resolution than traditional techniques like fMRI. Additionally, combining

multiple imaging techniques can allow for more accurate localization and timing of neural activity patterns.

In summary, while Brain Decoding has several limitations, these can be mitigated by employing various strategies, such as collecting larger and more diverse datasets, employing transfer learning using data augmentation techniques, improving interpretability through feature selection and visualization, combining brain decoding with other techniques, and using advances in brain imaging techniques with higher spatial and temporal resolution. By using these strategies, researchers can gain a more detailed understanding of the neural mechanisms in the brain.

8.3 Future Works

Certainly, there are many exciting avenues for future research in the area of Brain Decoding. Here are a few possibilities:

- **Multilingual brain decoding:** Most current decoding studies have been conducted in a single language, typically English. However, Multilingual brain decoding studies [104] could help researchers gain insights into the neural mechanisms underlying multilingualism, such as the degree to which different languages are represented in the brain and the cognitive processes involved in switching between languages. Multilingual brain decoding has the potential to improve our understanding of language processing and cognition in general and could have applications in areas such as second language learning and bilingual education.
- **Reconstructing Imagination:** Current Brain Decoding techniques are focussed on decoding the stimuli that the subject has seen or heard, but could it be possible to decode what the subject has just imagined?
- **Clinical applications:** Brain Decoding could have important clinical applications, such as developing brain-computer interfaces for people with motor disabilities or developing new diagnostic tools for neurological disorders. For example, decoding neural activity associated with language comprehension could provide new insights into language disorders such as aphasia.
- **Deep Learning Approaches:** Future research could investigate the use of more advanced machine learning techniques, such as deep learning and reinforcement learning, to improve the accuracy and robustness of decoding models.

These are just a few examples of the many possible future directions for research in this area. As technology and methods continue to advance, we can expect to see exciting new developments in Brain Decoding in the coming years.

Appendix A

Multi-View Brain Decoding

In this technical appendix, results are compared for both GloVe and BERT for all the experiments in the main chapter for Multi-View Brain Decoding (chapter 4).

We present the pairwise and rank accuracy for models trained on word+picture (WP), sentence (S) and word-cloud (WC) views in Figs. A.1, A.2 and A.3 respectively. Specifically, Fig. A.1 shows results when we infer using voxels corresponding to each of the three views. Ground-truth is GloVe (G) or BERT (B) embedding vector. Thus, WP_B_R means input view=WP (Word+picture), embedding=BERT, and metric=Rank (R) accuracy. Overall averaged (across subjects) accuracy results are summarized in Table A.1. Table A.2 shows distribution of informative voxels across the four brain networks. In this figure, WP_G_D means input view=WP (Word+picture), embedding=GloVe, and brain network=DMN (D). Individual level statistics can be found in Fig. A.4.

Train→	WP		S		WC	
Test↓	GloVe	BERT	GloVe	BERT	GloVe	BERT
WP	.74/.65	.72/.65	.71/.60	.70/.60	.66/.58	.68/.59
S	.65/.57	.67/.58	.69/.63	.70/.64	.67/.59	.71/.61
WC	.62/.55	.63/.56	.67/.60	.69/.61	.61/.56	.62/.57

Table A.1: Multi-View Decoder Summary Results (Pairwise/Rank Accuracy)

A.1 Train on WP view

We make the following observations from Fig. A.1 and Tables A.1&A.2:

- For test on WP view, wrt pairwise accuracy, GloVe model (0.74) is better than BERT (0.72) (one-sample t-test, 0.05 significance level, p=0.024).

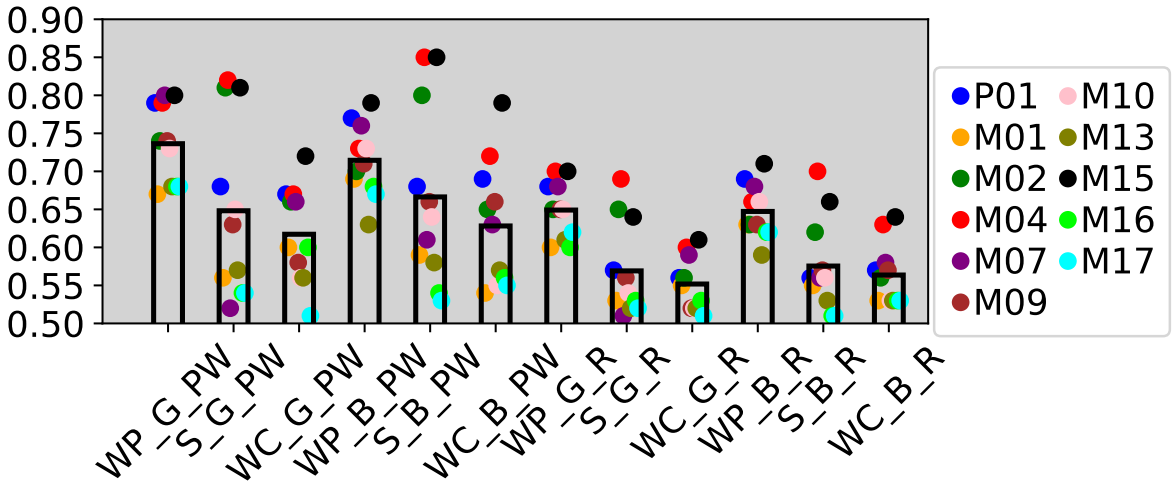


Figure A.1: Model trained on *Word+Pictures* view. Multi-View Decoding Pairwise (PW) and Rank (R) accuracy when tested on Word+Picture (WP)/Sentence (S)/Word-cloud (WC) views using GloVe (G) and BERT (B). Each colored dot represents a subject. The bar plot shows averages.

- For the test on S or WC views, BERT shows better performance than GloVe across both metrics. This can be explained by analyzing the brain network distribution differences in the following points.
- We observe that BERT captures a higher percentage of language informative voxels (18%) and DMN voxels (16%) compared to GloVe (12%, 13%), demonstrating the better language understanding with transformer based representations. This result has p=0.003 for language voxels and p=0.021 for DMN using a t-test with 0.05 significance.
- When the model is trained on WP view (unlike other views), for both embeddings, most informative voxels (about 53%) lie in the visual brain network, which is expected. Also, the location of these voxels was consistent across participants.

A.2 Train on S view

We make the following observations from Fig. A.2 and Tables A.1&A.2:

- For zero-shot test on WP view, wrt pairwise accuracy, GloVe model (0.71) is better than BERT (0.70) but we observed that the improvements are not significant (p=0.608).
- Accuracy of the model trained on S view and tested on WC view is better than same-view accuracy of the model trained and tested on WC view. This matches our observation that DMN and Language network voxels are higher in the S view than the WC view.

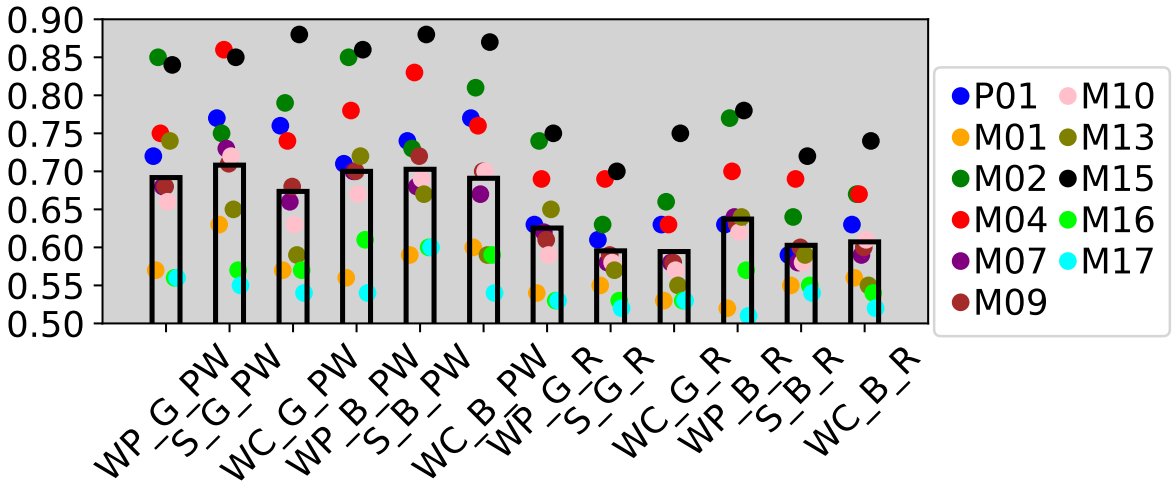


Figure A.2: Model trained on *Sentences* view. Multi-View Decoding Pairwise (PW) and Rank (R) accuracy when tested on Word+Picture (WP)/Sentence (S)/Word-cloud (WC) views using GloVe (G) and BERT (B) embeddings. Each colored dot represents a subject. The bar plot shows averages.

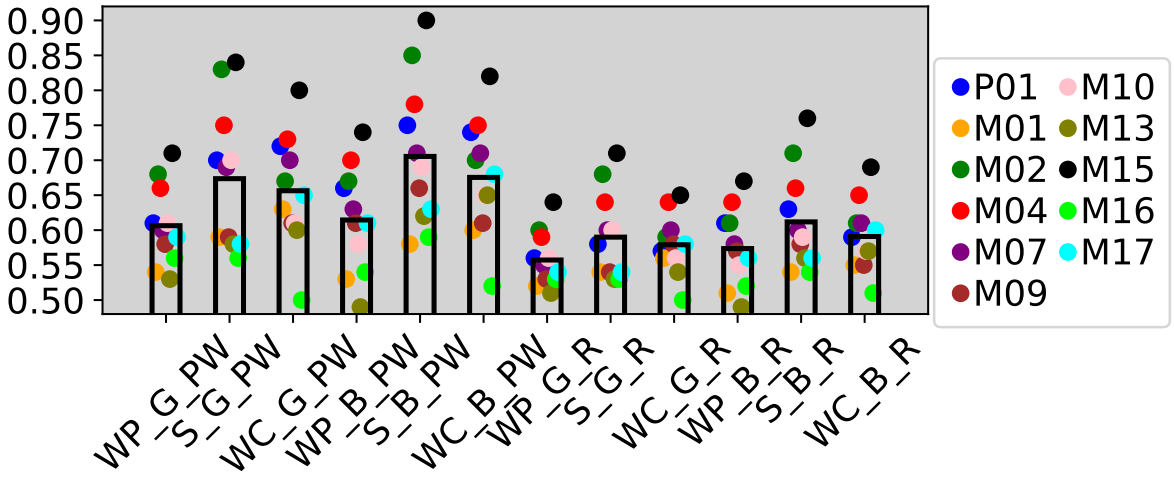


Figure A.3: Model trained on *Word-Cloud* view. Multi-View Decoding Pairwise (PW) and Rank (R) accuracy when tested on Word+Picture (WP)/Sentence (S)/Word-cloud (WC) views using GloVe (G) and BERT (B). Each colored dot represents a subject. The bar plot shows averages.

- For the test on S or WC views, BERT shows slightly better zero-shot performance than GloVe across both metrics. Results are not significant (p=0.251) for the S view, but they are significant for the WC view (p=0.021).

	WP		S		WC	
	G	B	G	B	G	B
D	.125	.162	.191	.222	.115	.137
V	.537	.534	.160	.202	.115	.161
L	.119	.177	.203	.246	.123	.192
T	.055	.064	.145	.135	.165	.145

Table A.2: Distribution of informative voxels among four brain networks: DMN (D), Visual (V), Language (L), Task Positive (T). Embeddings: GloVe (G), BERT (B). Input views: Word+Picture (WP), Sentence (S), Word-Cloud (WC)

- Compared to the model trained on WP view, distribution of voxels among the four brain networks shows that the model trained on S view has a higher percentage of voxels among the Language and DMN networks and lower in the visual network. Further, for the model trained on S view, BERT captures more informative voxels among the four brain networks compared to GloVe.
- Compared to the WP view, for the model trained on S view, informative voxels in the language and task brain network are much higher. This is in line with our understanding that linguistic and attention skills are important to understand sentence stimuli.

	DMN	Visual	Language	Task Positive
WP-S	.24/.17	.11/.29	.25/.17	.09/.05
WC-S	.25/.16	.25/.20	.30/.22	.07/.07
WP-WC	.14/.16	.08/.25	.15/.15	.06/.03

Table A.3: For each pair of views and each brain network, we show coverage ratios (second task on first/first task on second) of the voxels.

A.3 Train on WC view

We make the following observations from Fig. A.3 and Tables A.1&A.2:

- BERT performs better than GloVe. Results are not significant with p=0.401 for test on WC view. Results for test on WP and S views are significant with p=0.002, 0.014 using t-test, and 0.05 significance level.

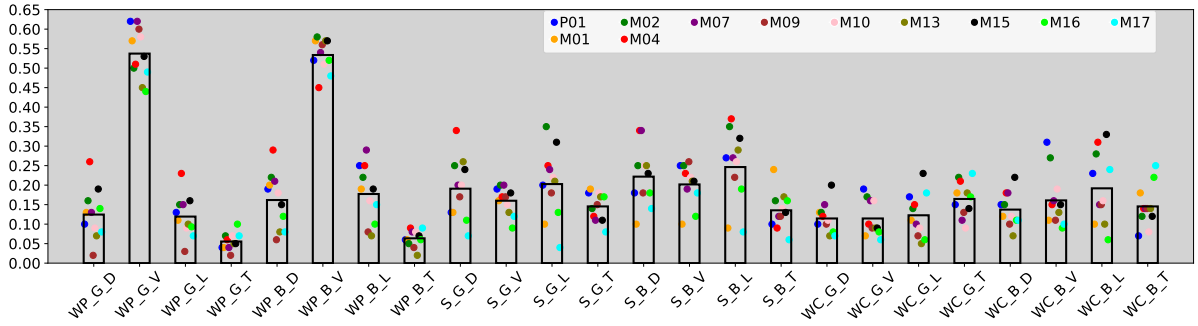


Figure A.4: Distribution of informative voxels among four brain networks: DMN (D), Visual (V), Language (L), Task Positive (T). Embeddings: GloVe (G), BERT (B). Input views: Word+Picture (WP), Sentence (S), Word-Cloud (WC)

- The supremacy of BERT can be explained by observing that BERT captures a higher percentage of informative voxels from the DMN (14%), Language (19%), and Visual (16%) networks when compared to GloVe (DMN - 11.5%, Language - 12%, Visual - 11.5%) when trained on WC view.

Appendix B

Cross-View Brain Decoding

In this technical appendix, results are compared for GloVe and BERT for all the Cross-View Brain Decoding tasks discussed in the main chapter for Cross-View Brain Decoding (chapter 5).

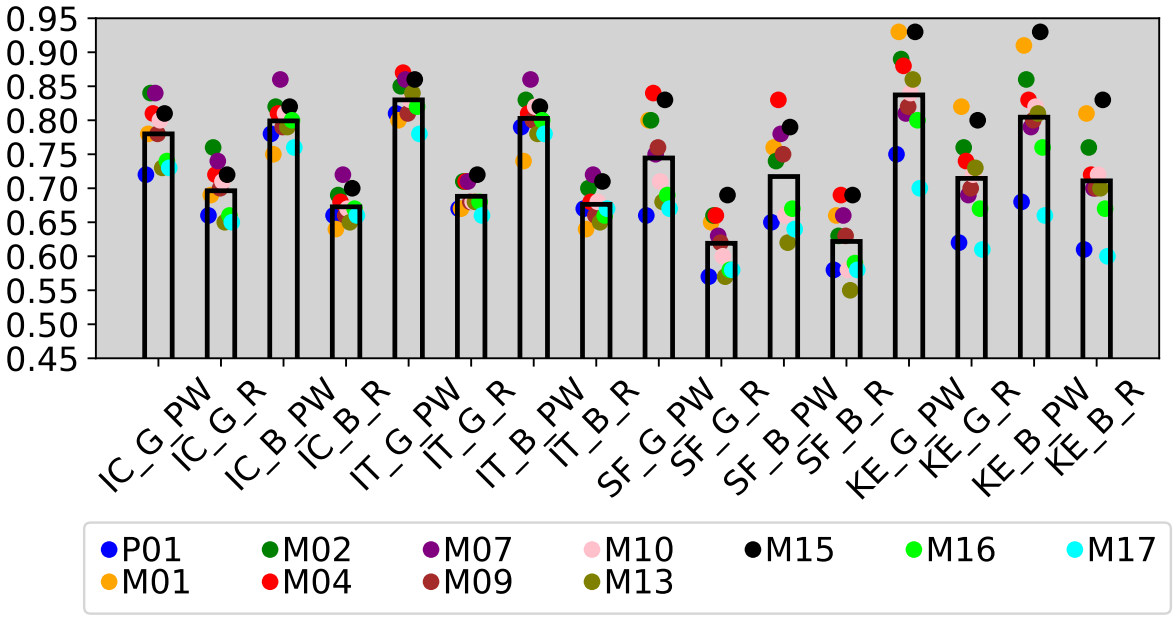


Figure B.1: Cross-View Decoding Pairwise (PW) and Rank (R) accuracy for Image Captioning (IC), Image Tagging (IT), Sentence Formation (SF) and Keyword Extraction (KE) using GloVe (G) and BERT (B) embeddings. Each colored dot represents a subject. The bar plot shows averages.

Appendix C

Abstract-Concrete Brain Decoding

In this technical appendix, results are compared for GloVe and BERT for all the experiments discussed in the main chapter for Abstract-Concrete Brain Decoding (chapter 6). Individual level statistics are also presented in this appendix.

The results for the abstract-train-concrete-test and concrete-train-abstract-test decoder models are presented per view using two embeddings in Table C.1. Detailed subject-level results are in Figs. C.1 and C.2. Fig. C.3 shows distribution of informative voxels among four brain networks using BERT embeddings for both the decoder models.

	Abs2Conc				Conc2Abs			
	PW		R		PW		R	
	G	B	G	B	G	B	G	B
WP	.648	.648	.574	.575	.698	.697	.591	.591
S	.598	.604	.548	.550	.641	.642	.560	.560
WC	.522	.519	.506	.505	.577	.587	.534	.528

Table C.1: Abs2Conc: Model trained on abstract concepts and tested on concrete concepts. Conc2Abs: Model trained on concrete concepts and tested on abstract concepts. Views: Word+Picture (WP)/Sentence(S)/Word-cloud (WC). Pairwise (PW) and Rank (R) accuracy when using GloVe (G) and BERT (B) embeddings.

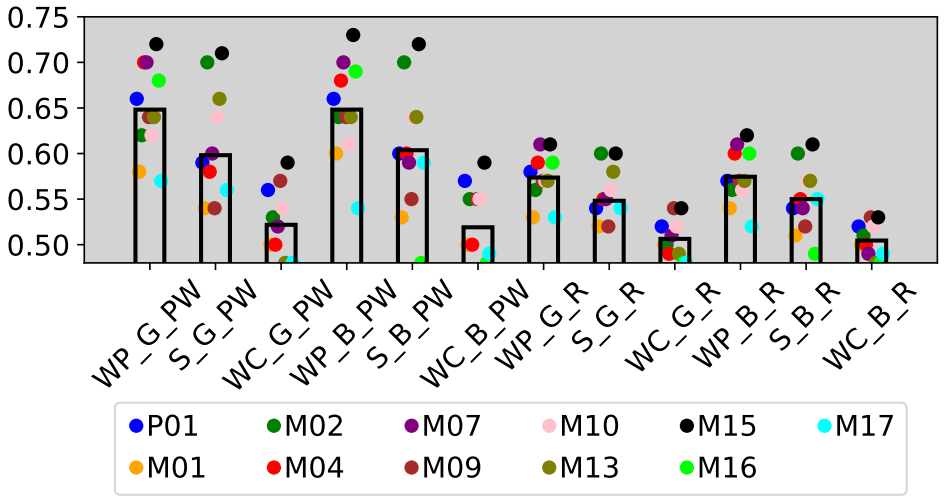


Figure C.1: Model trained on abstract concepts and tested on concrete concepts for Word+Picture (WP)/Sentence(S)/Word-cloud (WC) views. Pairwise (PW) and Rank (R) accuracy when using GloVe (G) and BERT (B) embeddings. Each colored dot represents a subject. Bar plot shows averages.

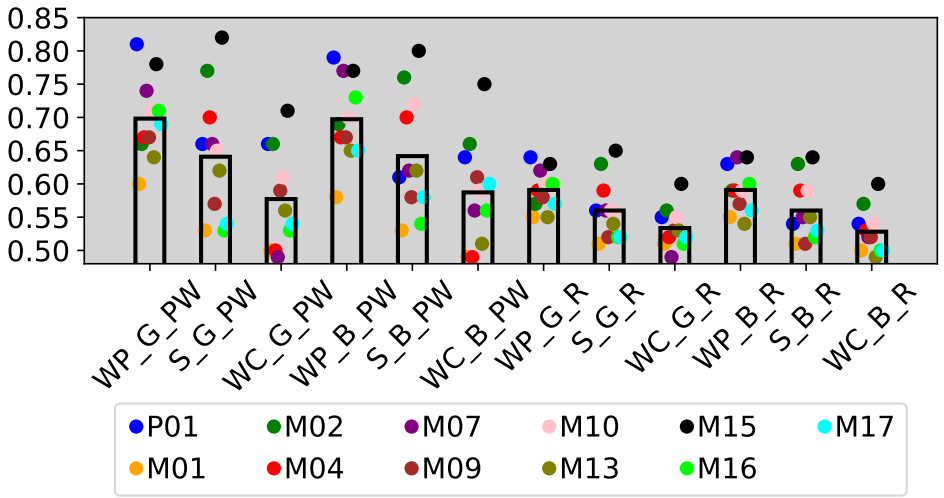


Figure C.2: Model trained on concrete concepts and tested on abstract concepts for Word+Picture (WP) / Sentence(S) / Word-cloud (WC) views. Pairwise (PW) and Rank (R) accuracy when using GloVe (G) and BERT (B) embeddings. Each colored dot represents a subject. Bar plot shows averages.

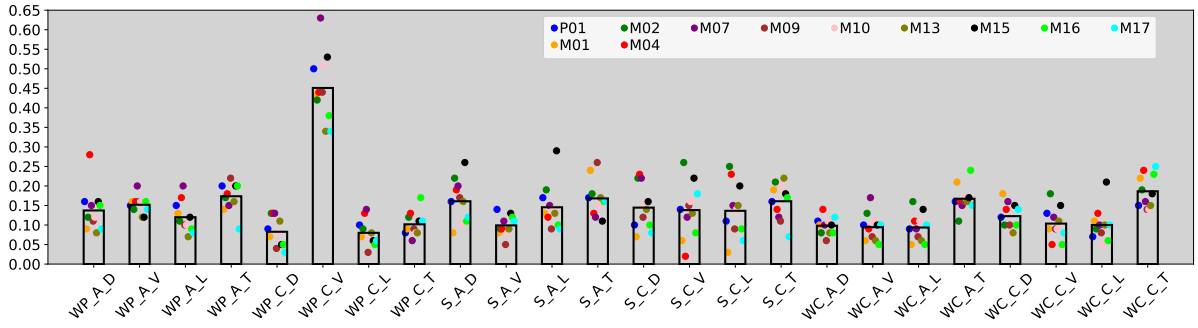


Figure C.3: Distribution of informative voxels among four brain networks: DMN (D), Visual (V), Language (L), Task Positive (T). Using BERT Embeddings. Input views: Word+Picture (WP), Sentence (S), Word-Cloud (WC). Decoders: abstract-train-concrete-test (A) and concrete-train-abstract-test (C).

Related Publications

1. Subba Reddy Oota[†], **Jashn Arora**[†], Manish Gupta, and Raju S. Bapi. "Multi-view and Cross-view Brain Decoding". In Proceedings of the 29th International Conference on Computational Linguistics, pages 105–115, Gyeongju, Republic of Korea. International Committee on Computational Linguistics. (COLING 2022).

2. **Jashn Arora**, Subba Reddy Oota, Manish Gupta, and Raju S. Bapi. "Brain Decoding for Abstract versus Concrete Concepts." 9th Annual Conference of Cognitive Sciences. 2022. (ACCS9 2022).

3. Subba Reddy Oota[†], **Jashn Arora**[†], Veeral Agarwal, Mounika Marreddy, Manish Gupta, and Bapi Surampudi. "Neural Language Taskonomy: Which NLP Tasks are the most Predictive of fMRI Brain Activity?." In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 3220–3237, Seattle, United States. Association for Computational Linguistics. (NAACL 2022).

4. Subba Reddy Oota, **Jashn Arora**, Vijay Rowtula, Manish Gupta, and Raju S. Bapi. 2022. "Visio-Linguistic Brain Encoding." In Proceedings of the 29th International Conference on Computational Linguistics, pages 116–133, Gyeongju, Republic of Korea. International Committee on Computational Linguistics. (COLING 2022).

[†] Shared First Authors

Bibliography

[1] N. Affolter, B. Egressy, D. Pascual, and R. Wattenhofer. Brain2word: Decoding brain activity for language generation. *arXiv preprint arXiv:2009.04765*, 2020. 19, 20

[2] A. J. Anderson, J. R. Binder, L. Fernandino, C. J. Humphries, L. L. Conant, R. D. Raizada, F. Lin, and E. C. Lalor. An integrated neural decoder of linguistic and experiential meaning. *Journal of Neuroscience*, 39(45):8969–8987, 2019. 45

[3] A. J. Anderson, E. Bruni, A. Lopopolo, M. Poesio, and M. Baroni. Reading visually embodied meaning from the brain: Visually grounded computational models decode visual-object mental imagery induced by written text. *NeuroImage*, 120:309–322, 2015. 47

[4] A. J. Anderson, D. Kiela, S. Clark, and M. Poesio. Visually grounded and textual semantic models differentially decode brain activity associated with concrete and abstract nouns. *Transactions of the Association for Computational Linguistics*, 5:17–30, 2017. 19, 20, 45

[5] A. J. Anderson, K. McDermott, B. Rooks, K. L. Heffner, D. Dodell-Feder, and F. V. Lin. Decoding individual identity from brain activity elicited in imagining common experiences. *Nature communications*, 11(1):1–14, 2020. 45

[6] R. Antonello, J. Turek, V. Vo, and A. Huth. Low-dimensional structure in the space of language representations is reflected in brain responses. *arXiv preprint arXiv:2106.05426*, 2021. 45

[7] S. Baillet. Magnetoencephalography for brain electrophysiology and imaging. *Nature neuroscience*, 20(3):327–339, 2017. 9

[8] H. Bao, L. Dong, and F. Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 48

[9] P. Bao, L. She, M. McGill, and D. Y. Tsao. A map of object space in primate inferotemporal cortex. *Nature*, 583(7814):103–108, 2020. 44, 46

[10] L. Beinborn, S. Abnar, and R. Choenni. Robust evaluation of language-brain encoding experiments. *International Journal of Computational Linguistics and Applications*, pages to–appear, 2019. 20

[11] R. Beliy, G. Gaziv, A. Hoogi, F. Strappini, T. Golan, and M. Irani. From voxels to pixels and back: Self-supervision in natural-image reconstruction from fmri. *arXiv preprint arXiv:1907.02431*, 2019. 18

[12] J. Berezutskaya, Z. V. Freudenburg, L. Ambrogioni, U. Güçlü, M. A. van Gerven, and N. F. Ramsey. Cortical network responses map onto data-driven features that capture visual semantics of movie fragments. *Scientific reports*, 10(1):1–21, 2020. 45

[13] J. R. Binder, R. H. Desai, W. W. Graves, and L. L. Conant. Where is the semantic system? a critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral cortex*, 19(12):2767–2796, 2009. 6, 22, 29, 45

[14] D. A. Boas, C. E. Elwell, M. Ferrari, and G. Taga. Twenty years of functional near-infrared spectroscopy: introduction for the special issue. *NeuroImage*, 85:1–5, 2014. Celebrating 20 Years of Functional Near Infrared Spectroscopy (fNIRS). 53

[15] A. Buchweitz, R. A. Mason, L. Tomitch, and M. A. Just. Brain activation for reading and listening comprehension: An fmri study of modality effects and individual differences in language comprehension. *Psychology & neuroscience*, 2(2):111–123, 2009. 46

[16] R. Buckner, J. Andrews-Hanna, and D. Schacter. The brain's default network: anatomy, function, and relevance to disease. *Annals of the New York Academy of Sciences*, 1124:1–38, 2008. 22

[17] E. Bullmore and O. Sporns. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, 10(3):186–198, 2009. 6

[18] C. Caucheteux, A. Gramfort, and J.-R. King. Disentangling syntax and semantics in the brain with deep networks. In *International Conference on Machine Learning*, pages 1336–1348. PMLR, 2021. 44, 45

[19] C. Caucheteux, A. Gramfort, and J.-R. King. Model-based analysis of brain activity reveals the hierarchy of language in 305 subjects. *arXiv preprint arXiv:2110.06078*, 2021. 44, 45

[20] N. Chang, J. A. Pyles, A. Marcus, A. Gupta, M. J. Tarr, and E. M. Aminoff. Bold5000, a public fmri dataset while viewing 5000 visual images. *Scientific data*, 6(1):1–18, 2019. 48

[21] A. Conneau and G. Lample. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32, 2019. 18, 19

[22] A. Conneau, R. Rinott, G. Lample, A. Williams, S. Bowman, H. Schwenk, and V. Stoyanov. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, 2018. 18

[23] R. T. Constable, K. R. Pugh, E. Berroya, W. E. Mencl, M. Westerveld, W. Ni, and D. Shankweiler. Sentence complexity and input modality effects in sentence comprehension: an fmri study. *NeuroImage*, 22(1):11–21, 2004. 19, 44, 46

[24] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 47

[25] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the*

Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, 2019. x, 13, 15, 19, 20, 44, 47, 48

[26] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020. 48

[27] N. F. Dronkers, D. P. Wilkins, R. D. Van Valin Jr, B. B. Redfern, and J. J. Jaeger. Lesion analysis of the brain areas involved in language comprehension. *Cognition*, 92(1-2):145–177, 2004. 1

[28] J. Duncan. The multiple-demand (md) system of the primate brain: mental programs for intelligent behaviour. *Trends in cognitive sciences*, 14(4):172–179, 2010. 22

[29] M. Eickenberg, A. Gramfort, G. Varoquaux, and B. Thirion. Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage*, 152:184–194, 2017. 18

[30] E. Fedorenko, M. K. Behr, and N. Kanwisher. Functional specificity for high-level linguistic processing in the human brain. *Proceedings of the National Academy of Sciences*, 108(39):16428–16433, 2011. 2, 22

[31] E. Fedorenko, P.-J. Hsieh, A. Nieto-Castañón, S. Whitfield-Gabrieli, and N. Kanwisher. New method for fmri investigations of language: defining rois functionally in individual subjects. *Journal of neurophysiology*, 104(2):1177–1194, 2010. 19

[32] M. D. Fox, A. Z. Snyder, J. L. Vincent, M. Corbetta, D. C. Van Essen, and M. E. Raichle. The human brain is intrinsically organized into dynamic, anticorrelated functional networks. *Proceedings of the National Academy of Sciences*, 102(27):9673–9678, 2005. 8

[33] A. D. Friederici. The brain basis of language processing: from structure to function. *Physiological reviews*, 91(4):1357–1392, 2011. 1

[34] J. Gauthier and R. Levy. Linking artificial and human neural representations of language. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 529–539, 2019. 16, 20, 44, 45, 47

[35] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 53

[36] U. Güçlü and M. A. van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27):10005–10014, 2015. 47

[37] P. Hagoort. On broca, brain, and binding: a new framework. *Trends in cognitive sciences*, 9(9):416–423, 2005. 6

[38] M. Hallett. Transcranial magnetic stimulation: a primer. *Neuron*, 55(2):187–199, 2007. 9

[39] G. Handjaras, E. Ricciardi, A. Leo, A. Lenci, L. Cecchetti, M. Cosottini, G. Marotta, and P. Pietrini. How concepts are encoded in the human brain: a modality independent, category-based cortical organization of semantic knowledge. *Neuroimage*, 135:232–242, 2016. 45

[40] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 48

[41] G. Hickok and D. Poeppel. The cortical organization of speech processing. *Nature reviews neuroscience*, 8(5):393–402, 2007. 6

[42] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 12

[43] N. Hollenstein, A. de la Torre, N. Langer, and C. Zhang. Cognival: A framework for cognitive word embedding evaluation. In *Proceedings of The SIGNLL Conference on Computational Natural Language Learning 2019*, 2019. 19, 45

[44] S. A. Huettel, A. W. Song, G. McCarthy, et al. *Functional magnetic resonance imaging*, volume 1. Sinauer Associates Sunderland, 2004. 8, 21

[45] A. G. Huth, W. A. De Heer, T. L. Griffiths, F. E. Theunissen, and J. L. Gallant. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600):453–458, 2016. 45

[46] A. G. Huth, T. Lee, S. Nishimoto, N. Y. Bilenko, A. T. Vu, and J. L. Gallant. Decoding the semantic content of natural movies from human brain activity. *Frontiers in systems neuroscience*, 10:81, 2016. 18, 19, 20

[47] A. Jaegle, F. Gimeno, A. Brock, A. Zisserman, O. Vinyals, and J. Carreira. Perceiver: General perception with iterative attention. *arXiv preprint arXiv:2103.03206*, 2021. 47

[48] S. Jain and A. G. Huth. Incorporating context into language encoding models for fmri. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 6629–6638, 2018. 20, 45

[49] S. Jat, H. Tang, P. Talukdar, and T. Mitchel. Relating simple sentence representations in deep neural networks and the brain. In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, pages 5137–5154. Association for Computational Linguistics (ACL), 2020. 45

[50] T. C. Kietzmann, C. J. Spoerer, L. K. Sörensen, R. M. Cichy, O. Hauk, and N. Kriegeskorte. Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences*, 116(43):21854–21863, 2019. 44, 46

[51] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 53

[52] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 48

[53] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 47, 48

[54] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. 10

[55] T. Mikolov, M. Karafiát, L. Burget, J. Cernockỳ, and S. Khudanpur. Recurrent neural network based language model. In *Interspeech*, volume 2, pages 1045–1048. Makuhari, 2010. 11

[56] T. M. Mitchell, S. V. Shinkareva, A. Carlson, K.-M. Chang, V. L. Malave, R. A. Mason, and M. A. Just. Predicting human brain activity associated with the meanings of nouns. *science*, 320(5880):1191–1195, 2008. 2, 18, 19, 20, 45

[57] S. A. Nastase, Y.-F. Liu, H. Hillman, K. A. Norman, and U. Hasson. Leveraging shared connectivity to aggregate heterogeneous datasets into a common response space. *NeuroImage*, 217:116865, 2020. 46

[58] S. A. Nastase, Y.-F. Liu, H. Hillman, A. Zadbood, L. Hasenfratz, N. Keshavarzian, J. Chen, C. J. Honey, Y. Yeshurun, M. Regev, et al. Narratives: fmri data for evaluating models of naturalistic language comprehension. *bioRxiv*, pages 2020–12, 2021. 45

[59] E. Niedermeyer and F. L. da Silva. *Electroencephalography: basic principles, clinical applications, and related fields*. Lippincott Williams & Wilkins, 2005. 9, 53

[60] S. Nishida and S. Nishimoto. Decoding naturalistic experiences from human brain activity via distributed representations of words. *Neuroimage*, 180:232–242, 2018. 19, 20

[61] S. Nishimoto, A. T. Vu, T. Naselaris, Y. Benjamini, B. Yu, and J. L. Gallant. Reconstructing visual experiences from brain activity evoked by natural movies. *Current biology*, 21(19):1641–1646, 2011. 18, 19

[62] S. R. Oota, J. Arora, V. Agarwal, M. Marreddy, M. Gupta, and B. Surampudi. Neural language taskonomy: Which NLP tasks are the most predictive of fMRI brain activity? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3220–3237, Seattle, United States, July 2022. Association for Computational Linguistics. 4

[63] S. R. Oota, J. Arora, M. Gupta, and R. S. Bapi. Cross-view brain decoding. *arXiv preprint arXiv:2204.09564*, 2022. 45

[64] S. R. Oota, J. Arora, V. Rowtula, M. Gupta, and R. S. Bapi. Visio-linguistic brain encoding. *arXiv preprint arXiv:2204.08261*, 2022. 45

[65] S. R. Oota, N. Manwani, and R. S. Bapi. fMRI Semantic Category Decoding Using Linguistic Encoding of Word Embeddings. In *International Conference on Neural Information Processing*, pages 3–15. Springer, 2018. 19, 20

[66] S. R. Oota, V. Rowtula, M. Gupta, and R. S. Bapi. Stepencog: A convolutional lstm autoencoder for near-perfect fmri encoding. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019. 45, 47

[67] M. Palatucci, D. Pomerleau, G. E. Hinton, and T. M. Mitchell. Zero-shot learning with semantic output codes. In *NIPS*, 2009. 19

[68] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 11

[69] F. Pereira, M. Botvinick, and G. Detre. Using wikipedia to learn semantic feature representations of concrete concepts in neuroimaging experiments. *Artificial intelligence*, 194:240–252, 2013. 45

[70] F. Pereira, G. Detre, and M. Botvinick. Generating text from functional brain images. *Frontiers in human neuroscience*, 5:72, 2011. 19

[71] F. Pereira, B. Lou, B. Pritchett, N. Kanwisher, M. Botvinick, and E. Fedorenko. Decoding of generic mental representations from functional mri data using word embeddings. *bioRxiv*, page 057216, 2016. 45

[72] F. Pereira, B. Lou, B. Pritchett, S. Ritter, S. J. Gershman, N. Kanwisher, M. Botvinick, and E. Fedorenko. Toward a universal decoder of linguistic meaning from brain activation. *Nature communications*, 9(1):1–13, 2018. x, 3, 15, 18, 19, 20, 21, 22, 26, 33, 35, 40, 45, 48

[73] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, 2018. 20

[74] S. F. Popham, A. G. Huth, N. Y. Bilenko, F. Deniz, J. S. Gao, A. O. Nunez-Elizalde, and J. L. Gallant. Visual and linguistic semantic representations are aligned at the border of human visual cortex. *Nature Neuroscience*, 24(11):1628–1636, 2021. 46

[75] J. D. Power, A. L. Cohen, S. M. Nelson, G. S. Wig, K. A. Barnes, J. A. Church, A. C. Vogel, T. O. Laumann, F. M. Miezin, B. L. Schlaggar, et al. Functional network organization of the human brain. *Neuron*, 72(4):665–678, 2011. 22

[76] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. *Image*, 2:T2, 2021. 48

[77] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 47

[78] M. E. Raichle, A. M. MacLeod, A. Z. Snyder, W. J. Powers, D. A. Gusnard, and G. L. Shulman. A default mode of brain function. *Proceedings of the National Academy of Sciences*, 98(2):676–682, 2001. 7

[79] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99, 2015. 49

[80] M. Schrimpf, I. Blank, G. Tuckute, C. Kauf, E. A. Hosseini, N. Kanwisher, J. Tenenbaum, and E. Fedorenko. The neural architecture of language: Integrative reverse-engineering converges on a model for predictive processing. *PNAS*, Vol:To appear, 2021. 16, 44, 47

[81] D. Schwartz, M. Toneva, and L. Wehbe. Inducing brain-relevant bias in natural language processing models. *arXiv preprint arXiv:1911.03268*, 2019. 20, 44, 45, 47

[82] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 48

[83] O. Sporns. Structure and function of complex brain networks. *Dialogues in clinical neuroscience*, 15(3):247, 2013. 6

[84] O. Sporns. *Networks of the Brain*. MIT press, 2016. 7

[85] J. Sun, S. Wang, J. Zhang, and C. Zong. Towards sentence-level brain decoding with distributed representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7047–7054, 2019. 19, 20, 45

[86] J. Sun, S. Wang, J. Zhang, and C. Zong. Neural encoding and decoding with distributed sentence representations. *IEEE Transactions on Neural Networks and Learning Systems*, 32(2):589–603, 2020. 45

[87] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*, 2017. 48

[88] H. Tan and M. Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, 2019. 48

[89] M. Tan and Q. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019. 48

[90] B. Thirion, E. Duchesnay, E. Hubbard, J. Dubois, J.-B. Poline, D. Lebihan, and S. Dehaene. Inverse retinotopy: inferring the visual content of images from brain activation patterns. *Neuroimage*, 33(4):1104–1116, 2006. 19, 46

[91] M. Toneva, O. Stretcu, B. Póczos, L. Wehbe, and T. M. Mitchell. Modeling task effects on meaning representation in the brain via zero-shot meg prediction. *Advances in Neural Information Processing Systems*, 33, 2020. 45

[92] M. Toneva and L. Wehbe. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). *arXiv preprint arXiv:1905.11833*, 2019. 20, 45

[93] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021. 48

[94] N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, and M. Joliot. Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain. *Neuroimage*, 15(1):273–289, 2002. 7

[95] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *NIPS*, 2017. x, 13, 14, 44, 47

[96] A. Wang, M. Tarr, and L. Wehbe. Neural taskonomy: Inferring the similarity of task-derived representations from brain activity. *Advances in Neural Information Processing Systems*, 32:15501–15511, 2019. 44, 47

[97] J. Wang, V. L. Cherkassky, and M. A. Just. Predicting the brain activation pattern associated with the propositional content of a sentence: Modeling neural representations of events and states. *Human brain mapping*, 38(10):4865–4881, 2017. 45

[98] S. Wang, J. Zhang, N. Lin, and C. Zong. Probing brain activation patterns by dissociating semantics and syntax in sentences. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9201–9208, 2020. 20

[99] S. Wang, J. Zhang, H. Wang, N. Lin, and C. Zong. Fine-grained neural decoding with distributed word representations. *Information Sciences*, 507:256–272, 2020. 20, 40, 45

[100] L. Wehbe, B. Murphy, P. Talukdar, A. Fyshe, A. Ramdas, and T. Mitchell. Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PloS one*, 9(11):e112575, 2014. 15, 18, 19, 20

[101] L. Wehbe, B. Murphy, P. Talukdar, A. Fyshe, A. Ramdas, and T. Mitchell. Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *in press*, 2014. 45

[102] L. Wehbe, A. Vaswani, K. Knight, and T. Mitchell. Aligning context-based statistical models of language with brain activity during reading. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 233–243, 2014. 20

[103] D. Wilson and D. Sperber. *Meaning and Relevance*. Cambridge University Press, 2012. 6

[104] M. Xu, D. Li, and P. Li. Brain decoding in multiple languages: Can cross-language brain decoding work? *Brain and Language*, 215:104922, 2021. 54

[105] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, 2021. 18

[106] D. Yamins, H. Hong, C. Cadieu, and J. J. DiCarlo. Hierarchical modular optimization of convolutional networks achieves representations similar to macaque it and human ventral stream. 2013. 47

[107] D. L. Yamins and J. J. DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3):356–365, 2016. 47

[108] D. L. Yamins, H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, and J. J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, 111(23):8619–8624, 2014. 44, 46, 47

[109] S. Zhao, X. Jiang, J. Han, X. Hu, D. Zhu, J. Lv, T. Zhang, L. Guo, and T. Liu. Decoding auditory saliency from fmri brain imaging. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 873–876, 2014. 18

[110] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27, 2015. 13