Testing the Alignment of Multi-Modal Neural Network Models to Human Brain Areas

Master Thesis

Narek Alvandian

August 23, 2024

Carried out in the Kreiman lab, Harvard Medical School
Advisors: Prof. G. Kreiman, Prof. M. Schrimpf

School of Computer and Communication Sciences, EPFL

Abstract

Studying brain function often involves controling stimuli and recording neural responses. This approach has significantly advanced our understanding of sensory processing. Recently, artificial neural networks (ANNs) optimized for naturalistic tasks (such as object detection and text generation) have been used to model brain signals. These models however, are typically unimodal, focusing on one sensory modality at a time. Importantly, our perception of the world is inherently multisensory, relying on the integration of various sensory modalities. Sensory inputs in the brain are not processed in isolation, and information from different modalities can be both redundant and unique.

We propose that using multimodal neural network models could improve the modeling of both single sensory circuits and sensory integration. In this work we set the foundation for future research with audio-visual Neural Networks as Models of the brain. We prepared a dataset of auditory-visual language task, and introduced a set of tools for statistical analysis of modeling performance.

Unexpectedly, our preliminary results on visual parts of the data suggest that single-modality models, like ResNet-50 and AlexNet, are not able to model visually-selective electrodes.

Acknowledgements

First of all, I want to thank my supervisors, Prof. Kreiman and Prof. Schrimpf, for taking the time and effort to share their knowledge and experience with me and for doing so with great attention to detail. I am grateful for their patience when I wasn't making sense and for creating and managing laboratories where I felt comfortable asking questions and sharing ideas. To Prof. Kreiman, for being an inspiration with his broad intellectual curiosity, and to Prof. Schrimpf, for encouraging collaborative science through his personal example.

I would also like to thank all members of Prof. Kreiman's lab, where I spent the past five months: Pranav Misra, Chenguang Li, Elisa Pavarino, Bastien Le Lan, Ravi Srinivasan, Victor Gillioz, Morgan Talbot, Spandan Madan, Dianna Hidalgo, and Alliyah Steele, for providing invaluable feedback during weekly lab meetings, answering my endless questions every single day, and, most importantly, for being supportive and fun people to be around.

I am deeply grateful to my parents, Lilit Hakobyan and Arshak Alvandian, thank you for providing me with the opportunity to study and explore my curiosity. To my grandparents, and brothers for their unconditional love and support.

I thank my wonderful girlfriend Daria for the great deal of love, support, and patience she has afforded to me over the three years of my absence.

Finally, I extend my gratitude to my friend Claude for proofreading the early drafts of this thesis.

Contents

Chapter 1

Introduction

Our day-to-day experience is highly multisensory [37]. We understand and interact with the world around us through different modalities, and the loss of any of the senses would result in a drastically different perception of the world. Different sensory modalities provide unique information, which together make our perception so rich.

The first and simplest organisms had no separation of modalities—no muscles, nerves, or axons. For them, chemical and mechanical gates played redundant roles [18] [14]. Once a signal reached the organism, it would spread throughout the whole body, affecting all the cells. This is what single-sensory perception looked like. It was observed that the behavioral reaction remained the same regardless of the input modality. Consider touching a eukaryote to trigger mechanical signal gates, or changing the chemical concentration of the surrounding liquid to activate chemical gates. It is hypothesized that our vision later evolved from these mechanical "touch" receptors (think of photons "touching" the eye). These simpler organisms had a single "modality," a very limited view of the world, and identical reactions regardless of the type of input stimuli, which clearly constrained their behavior as well as adaptivity, and chances of survival.

In mammals, we have complicated and specialized processing pathways. Different sensory inputs (vision, audition, touch, smell, etc.) are first captured by peripheral organs (eyes, ears, skin, nose, etc.), then processed in the brain's primary processing areas, and finally integrated in unknown ways to represent the world as we know it, forming a (mostly) consistent perception of reality.

The standard way of studying the neuronal mechanisms underlying cognition and behavior is to control the incoming stimuli as much as possible and to perform recordings from the brain (e.g., ECoG, sEEG, fMRI). This approach allows us to study the effects of particular inputs on the activity of specific circuits or neurons. It has proven successful and has yielded much of our knowledge about brain function.

Today, a promising approach to modeling the signals recorded from the brain involves using representations from Artificial Neural Networks, optimized for naturalistic tasks such as object detection, text generation, and speech understanding.

Such an approach is very effective for modeling separate circuits. However, it is insufficient for modeling the brain as a whole because it ignores everything else happening in the natural environment beyond the modality of interest. Activity across the brain is never uniquely determined by changes in a single sensory inputs, nor did our circuits evolve to process sensory information separately without further integration.

Different sensory inputs can contain both redundant information (e.g., the sound of a cat and the image of a cat both point to the same object) as well as unique information specific to each modality (e.g., it is impossible to create an exact visual percept of a particular image through sound alone). This suggests that the brain must have mechanisms specific to processing each modality separately, as well as mechanisms for integrating information across modalities (e.g., associating a sound with an image).

We argue that using Multi-Modal (or multi-sensory) Neural Network Models will be beneficial for both 1. modeling single sensory circuits and 2. building models of sensory integration.

Current state-of-the-art models of visual areas are single-modality Neural Network models [34]. However, vision in complex organisms did not develop in isolation; it formed alongside and under constraints from many other sensory inputs. Therefore, it would be more biologically plausible to build visual representations that were developed in conjunction with representations from other modalities.

Additionally, certain phenomena in “traditionally visual” areas, such as the inferior temporal (IT) cortex, cannot be explained solely by visual inputs. Studies have shown the existence of neurons in cats’ [38] and monkeys’ [13] [30] ”visual” areas that respond to auditory stimuli. Similarly, some “orientation-tuned neurons” in visual areas respond to somatosensory signals when these align with the neurons’ preferred visual orientation [23]. Regardless of the underlying mechanism—whether it be memory, imagination, or world-modeling—these phenomena cannot be modeled by vision-only models, as visual input is not the sole driver of cell activity. A vision-only model of the IT cortex lacks any notion of inputs from other modalities.

To develop a complete model of any “sensory processing” area, we must integrate all relevant modalities.

Going beyond single-sensory processing, some perceived concepts are entirely amodal. For example, concepts such as the number of repetitions, intensity, and possibly even spatial location can be mapped from any sensory input.

Our brain also excels at identifying different sensory manifestations of the "same object." For example, we can associate the smell of an orange with its appearance and the expected tactile sensation.

Interestingly, language is also a multi-modal phenomenon. We read, hear, speak, and can even feel language through our skin (e.g., Braille, which is used by blind individuals).

Developing algorithms that process multiple modalities could allow us to model the mechanisms in the brain that enable these phenomena.

The first question we wanted to investigate was whether visual representations formed alongside language were better models of activity in visual areas. After reviewing the literature [41] [8] [7] [39], we came to a conclusion that the current evidence suggests the answer is negative. We provide more details on this in section 2.6.

Consequently, we shifted our focus to building audio-visual models that could explain visual and auditory processing separately, as well as brain areas involved in vision-auditory integration, as an alternative to using separate visual and auditory models.

The neural data we aim to model consists of recordings from 654 stereo-electroencephalography (sEEG) [5] [28] electrodes placed in various brain areas of 8 human subjects with pharmacologically intractable epilepsy. The participants performed a task involving separate visual and auditory signals.

To assess whether audio-visual models offer any advantages, we first need to measure the performance of single-modality models. At present, we only report the modeling performance of a single-sensory vision model.

We find that, across all electrodes, a few are selective to visual stimuli. For these electrodes, we applied statistical significance testing techniques to quantify the modeling performance of ResNet-50 and AlexNet models. The results are detailed in section 4.2.

Chapter 2

Related work

2.1 Visual Processing in the Brain. The Two-streams hypothesis

Currently, our understanding of the neuroscience of vision follows a model that explains the processes of visual information through two distinct pathways: the ventral stream and the dorsal stream Figure 2.1.

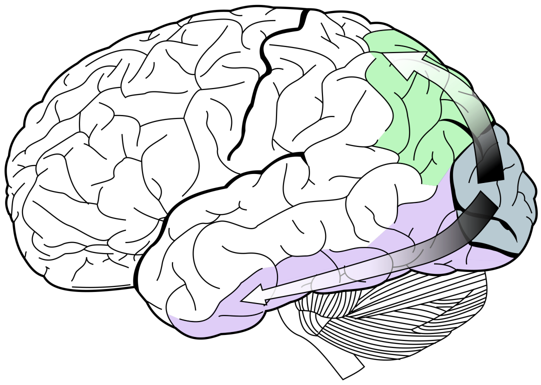


Figure 2.1: Ventral stream (the "what" pathway) in purple, Dorsal stream (the "when" pathway) in green, the primary vision area (the occipital lobe) in blue

Ventral Stream ("What" Pathway) is primarily responsible for object recognition and form representation. It extends from the primary visual cortex (V1) in the occipital lobe to the inferior temporal (IT) cortex.

This pathway processes detailed visual information about the identity of objects, such as their color, shape, and size. Because of this, it's often referred to as the "what" pathway. The ventral stream enables us to recognize and categorize objects, helping us understand what we are looking at.

It processes images through a series of stages, starting from the retina and ending in the inferior temporal (IT) cortex. This complex system involves many interconnected brain regions working together to process visual information. In the retina, light is converted into electrical signals by photoreceptor cells. These signals then travel through the optic nerve to the lateral geniculate nucleus (LGN) in the thalamus. The LGN acts as a relay station, filtering visual information and integrating feedback from higher cortical areas to focus attention on specific parts of the visual field. From the LGN, signals go to the primary visual cortex (V1) which extracts features like edges, orientations, and spatial frequencies. The information then flows through several areas: V2, V3, V4, each specializing in processing different aspects of visual information, such as contours, shapes, motion, color, and form. Finally, the visual information reaches the inferior temporal (IT) cortex, which is crucial for object recognition. IT neurons are selective for complex visual features and object categories, and their responses are thought to be the basis of our ability to recognize and categorize objects.

Studies have shown that the IT cortex is the highest point of visual processing and is central to object recognition. The hierarchical organization of the visual system, ending in the IT cortex, allows for the creation of increasingly complex and abstract representations of visual information. While early visual areas extract basic features, the IT cortex integrates these into more sophisticated representations that don't change much with viewing conditions, enabling robust object recognition.

Dorsal Stream ("Where" or "How" Pathway) is involved in spatial awareness and the coordination of actions. It projects from the primary visual cortex to the posterior parietal cortex. This pathway processes information related to the location, movement, and spatial relationships of objects, earning it the nickname "where" pathway. It also integrates visual information with motor functions, guiding actions like reaching or grasping, which is why it's sometimes referred to as the "how" pathway.

Together, these streams allow the brain to process visual information in a comprehensive manner, with the ventral stream helping to identify objects and the dorsal stream providing the spatial context and guiding interactions with those objects. The two streams work in parallel but are also interconnected, allowing for integrated visual perception and action.

2.2 Auditory Processing in the Brain

Sound waves enter the ear and are transmitted through the outer ear to the eardrum, causing it to vibrate. These vibrations are transferred to the middle ear, where three small bones (ossicles) amplify the sound and transmit it to the cochlea in the inner ear. The cochlea, a fluid-filled structure, contains hair cells that convert mechanical vibrations into electrical signals [17]. These hair cells are organized tonotopically, meaning different parts of the cochlea respond to different frequencies.

After few more stations (i.e. from Auditory Nerve to Brainstem to Midbrain to Medial Geniculate Body) the signal reaches the Auditory Cortex located in the temporal lobe of the brain. The auditory cortex is involved in the higher-level processing of sounds, such as identifying and interpreting complex sounds like speech and music [15]. The auditory cortex is organized in a tonotopic manner, meaning neurons are arranged based on the frequencies they respond to, similar to the organization in the cochlea.

2.3 Non-Activity in IT Caused By Non-Visual Inputs

In a study [13] examining the sensory modality specificity of neural activity related to memory in visual cortex, authors provide compelling evidence for cross-modal influences on delay period activity in inferotemporal (IT) cortex. Using a series of delayed match-to-sample tasks involving both visual and auditory stimuli, they found that many IT neurons exhibited selective delay period activity that depended on the sample stimulus, particularly in cross-modal tasks. Notably, 26% of neurons showed selective delay activity in a visual-to-auditory task, and 23% in an auditory-to-visual task, compared to lower percentages in unimodal tasks. Authors report correlation between delay selectivity in the two cross-modal tasks, suggesting a long-term representation of learned cross-modal associations. Furthermore, in one animal, delay activity in the auditory-to-visual task predicted behavioral performance. These findings indicate that neurons in IT contribute to abstract memory representations that can be activated by different sensory modalities while remaining specific to visual behaviors. This work highlights the presence of cross-modal interactions in a cortical area traditionally considered to be primarily visual, demonstrating the complexity of sensory integration in higher-order cognitive processes.

Similar results have been demonstrated [23] for a visual-tactile cross-modal task.

There has been other evidence that IT neurons can be activated by auditory stimuli [30]

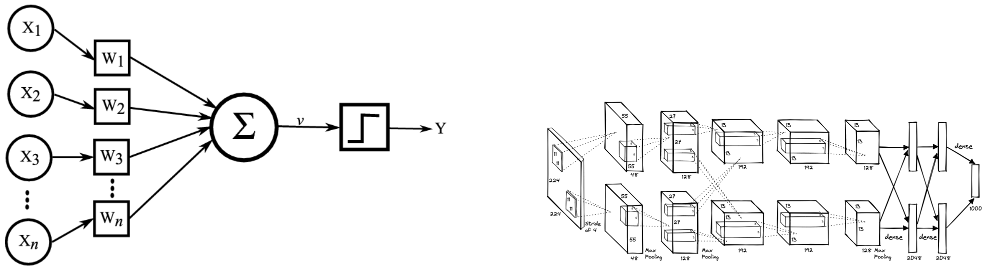
2.4 Modality Integration Illusions

The McGurk illusion is a perceptual phenomenon that demonstrates the interaction between auditory and visual information during speech perception. Discovered by [24], the illusion occurs when conflicting visual and auditory stimuli are presented, leading to a perception of a sound that differs from the actual auditory input. For example, when the sound of the syllable /ba/ is paired with the visual articulation of the syllable /ga/, observers often perceive a third syllable, /da/. This illusion highlights the brain's reliance on both auditory and visual cues for interpreting speech, revealing that speech perception is not solely based on auditory information but is a multimodal process involving integration across sensory modalities.

The ventriloquist illusion is a perceptual phenomenon in which the perceived location of a sound is shifted towards a concurrent visual stimulus, such as a moving mouth or object, despite the actual sound source being elsewhere. This illusion is most famously demonstrated in ventriloquism, where a ventriloquist speaks without moving their lips while manipulating a puppet's mouth. Observers perceive the puppet as the source of the voice, even though the sound is coming from the ventriloquist. This effect illustrates the brain's tendency to rely more on visual information than auditory information when determining the location of a sound, a process known as visual capture. Studies have shown that the ventriloquist illusion arises from the integration of sensory information, where the brain resolves conflicting spatial cues by prioritizing the more reliable visual input [2]

2.5 Artificial Neural Networks

Artificial Neural Networks (ANNs) are computational models inspired by biological neural networks. They consist of layers of interconnected nodes (neurons), where each connection (synapse) has an associated weight. Neurons receive input, process it by applying an activation function, and pass the output to subsequent layers.



(a) Perceptron. Model takes a weighed combination of inputs and passes the result through a non-linear "activation" function

(b) Alexnet. Consecutive application of convolutional layers with the last layers being fully-connected.

Figure 2.2: Artificial Neural Networks

The concept of ANNs originated with the perceptron, introduced by [31], which is a simple linear classifier that serves as the foundation for more complex architectures. It's visualised on Figure 2.2a. Modern ANNs typically include an input layer, one or more hidden layers, and an output layer. During training, the network learns to perform a task by optimizing an objective function (e.g., minimizing classification error). This optimization is carried out using gradient descent, where the model iteratively updates the weights based on the gradient of the loss function with respect to the weights. Backpropagation, introduced by [32], is used to efficiently compute these gradients across the network by applying the chain rule.

Specialized forms of ANNs, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have been developed to handle specific types of data. CNNs [22] are designed for processing grid-like data structures, such as images, and have been particularly successful in computer vision tasks. RNNs, formalized by [12], are tailored for sequential data and have been widely used in natural language processing.

For our research we will be working with two architectures of CNNs:

AlexNet [21] is a deep convolutional neural network that won the 2012 ImageNet competition, marking a breakthrough in computer vision. It consists of five convolutional layers followed by three fully connected layers and popularized techniques like ReLU [1] activations and dropout [36], which helped achieve state-of-the-art performance at the time. The schema of it's architecture is on Figure 2.2b.

ResNet (Residual Network) [16] is a deep neural network architecture that introduced skip connections, allowing information flow to bypass layers and enabling the training of very deep networks without performance degradation.

More recently, the development of Transformers [40] has revolutionized many fields, particularly in natural language processing, allowing for parallel processing and greater scalability compared to RNNs.

An interesting historical observation is that many of these architectures were inspired from our knowledge of brain function. The perceptron is inspired after the biological neuron and the hierarchical structure of CNNs is similar to the hierarchy of the visual pathway.

2.6 ANNs as models of the brain

Vision was the first are, where representations learned by neural networks showed promising modeling performance.

[42] Introduced the idea of using Convolutional Neural Networks as models of the higher level visual area the IT cortex in monkeys. Although there are active debates on the degreee of bilogical plausibility of DNNs, they have some feature that allow us to make this comparison. The main one being the

hierarchical processing of visual inputs, alongside activation functions and convolutions.

Their study aimed to develop computational models that could accurately predict neural responses in higher visual cortex areas, specifically the inferior temporal (IT) cortex and V4. Researchers used deep convolutional neural networks (CNNs) that were optimized for performance on a real-world object recognition task. These models, when trained on natural images, were found to be good predictors of neural responses in IT cortex and V4 of macaque monkeys, explaining around 30% of the variance (normalized by the self-consistency)in the electrode recordings. The study showed that the representations learned by an artificial neural network were in some sense similar to those found in the primate visual system, particularly in higher visual areas. An interesting observation was that the performance-optimized CNNs outperformed other models, including those based solely on visual features or randomly initialized networks, in predicting neural responses. This suggested that the constraints imposed by object recognition tasks may shape the representations in higher visual cortex areas.

The "Brain-Score" paper [34] built upon and extended the work of [42] in several important ways. Brain-Score introduced a more comprehensive benchmark for comparing artificial neural networks to the primate visual system. While Yamins et al. focused primarily on IT and V4, Brain-Score included neural predictivity for three key areas of the ventral visual stream: V4, IT, and primary visual cortex (V1) and added a behavioral component, testing models on their ability to predict human object recognition behavior, going beyond just neural responses. The study also evaluated a much wider range of convolutional neural network architectures, including many state-of-the-art models developed after 2014 and this way provided insights into which architectural features lead to better brain-like representations.

Conwell et al. [9] further explore the relationship between artificial neural networks and biological visual processing. Authors went beyond CNN architectures, benchmarking Visual Transformers [11] as well. They also used fMRI data from human brains and not electrode recordings from monkeys. Authors examined how various properties of neural networks (e.g., depth, width, skip connections, normalization) influence their similarity to brain responses. It also investigated how training on different tasks affects the brain-likeness of network representations, extending beyond just object recognition. Interestingly, they show that the training diet of the model had a much higher influence on the brain-likeness than any other feature like the number of layers, activation function or even the architecture as a whole.

Another work [10] delves deeper into the emergence of functionally specialized regions within deep neural networks, drawing striking comparisons to the organization observed in the primate visual cortex. Researchers demonstrate that when trained on naturalistic visual tasks, deep neural networks spontaneously develop distinct regions that specialize in processing specific visual features as well as share lower-level ones, mirroring the functional organization found in the primate visual system. This emergent specialization occurs without explicit guidance or constraints aimed at replicating biological structures, suggesting that it may arise as a natural consequence of optimizing for diverse visual tasks in a hierarchical processing system. By mapping the activity patterns of these artificial networks to neuroimaging data from human subjects, the study reveals similarities in the spatial organization of feature selectivity. Notably, the networks exhibit specialized regions for processing faces, words, and scenes, analogous to the human brain. It suggests that the functional architecture of the visual cortex may be, to some extent, an inevitable outcome of the computational demands of visual processing, rather than solely a product of evolutionary or developmental constraints specific to biological systems.

Additionally, Kell et al. [20] demonstrate that the approach of using task-optimized neural networks to model brain function can be successfully applied beyond the visual system, auditory processing in this case. Authors developed a deep neural network optimized for speech and music recognitiontasks and found that it could accurately predict human behavioral responses to auditory stimuli. Moreover, the network's internal representations showed strong correlations with brain responses recorded via functional MRI, particularly in regions of the early auditory cortex. Importantly, this work revealed a hierarchical organization in the auditory cortex that mirrors the layer-wise organization of the artificial neural network and showed that the network optimized for two tasks develops shared and task-specific features. This work showed the task-specialization before [10]. Lower layers of the network corresponded to activity in primary auditory areas, while higher layers matched responses in more advanced speech processing regions. It suggests that similar computational principles may underlie the processing of diverse sensory inputs in the brain, offering a unifying framework for understanding sensory cognition.

Same principles have been applied to our understanding of language processing in the brain. This is interesting because unlike vision and audio - language is not a sensory modality. We can encode and decode language visually, auditorily as well as through tactile inputs (with Braille language). [33] reveal that language models such as Transformers, when mapped onto human brain activity, show striking similarities to the neural architecture of language processing. The study identifies a hierarchical organization in the language network, with different levels of linguistic abstraction represented across distinct brain regions. Importantly, the research highlights the central role of predictive processing in language comprehension, aligning with influential theories in cognitive neuroscience. The work demonstrates that the most accurate predictions of brain activity during language tasks come from models that incorporate both bottom-up and top-down processing, mirroring the predictive coding framework proposed for human cognition. This suggests that the brain's language system may operate on similar principles of prediction and error correction as implemented in these artificial models.

Wang et al. [41] explore the potential benefits of multimodal representations in predicting neural responses in the high-level visual cortex. The hypothesis was that neural network models incorporating both visual and linguistic inputs would better predict brain activity in response to complex visual stimuli, such as real-world scenes. To test this, they compared the predictive power of visual encoders taken from multimodal models like CLIP (which processes both images and associated captions) against unimodal models like ResNet and BERT. Using voxel-wise encoding models and fMRI data from the Natural Scenes Dataset [3], they found that multimodal embeddings, particularly those from CLIP, outperformed unimodal embeddings in predicting neural responses in high-level visual areas. The CLIP model was also better at capturing contextual and semantic similarities even when visual similarities were absent. Additionally, the principal semantic dimensions of the CLIP encoding model aligned well with core organizational axes in the brain. However, it's important to note that these findings should be interpreted cautiously, as the study did not control for various factors such as training data size and type, batch size, architecture, loss function, optimizer, and number of training steps, which could potentially influence the results. Later revision of this work showed that indeed in the more controlled setting the differences are much less pronounced and obvious.

Similarly, Conwell et al [8] trained different versions of a SLIP model, including SLIP-SimCLR (unimodal), SLIP-CLIP (multimodal), and SLIP-Combo to control for architecture and trainning data [6]. The results showed that language alignment in multimodal models **did not** provide significant advantages in predicting neural responses in the ventral stream of the visual cortex. This finding held true not only for higher-level visual areas but also for early visual cortex and, surprisingly, even for the visual word form area, which is typically associated with written language processing. However, it is worth noting that the only form of multimodal language information used was CLIP-style [27] vision-language alignment, suggesting that this type of multimodal integration may not be as beneficial for predicting neural responses in visual processing areas as initially thought, and leaves the question of whether other types of vision-language integration might yield better results.

Finally [39] leverage multimodal neural networks, trained on both images and text, to reveal brain regions involved in integrating sensory inputs. By comparing these models' representations with sEEG recordings from patients watching movies, the authors find areas that are better explained by multimodal networks compared to single-sensory vision and language networks. This research went through a set of single and multi modal networks of multiple architectures and ways of combining modalities, as well as controlled for architecture, similarly to [8].

Chapter 3

Methods

3.1 Neural Data

Our neural data [26] [25] consist of intracranial field potentials from 674 electrodes implanted in 8 patients with pharmacologically intractable epilepsy, via stereoelectroencephalography (sEEG).

All electrode locations are shown in Figure 3.1. A bipolar reference was used, and for further analysis, we considered only field potential signals filtered in the high gamma frequency band (65-150 Hz), as this band has been reported

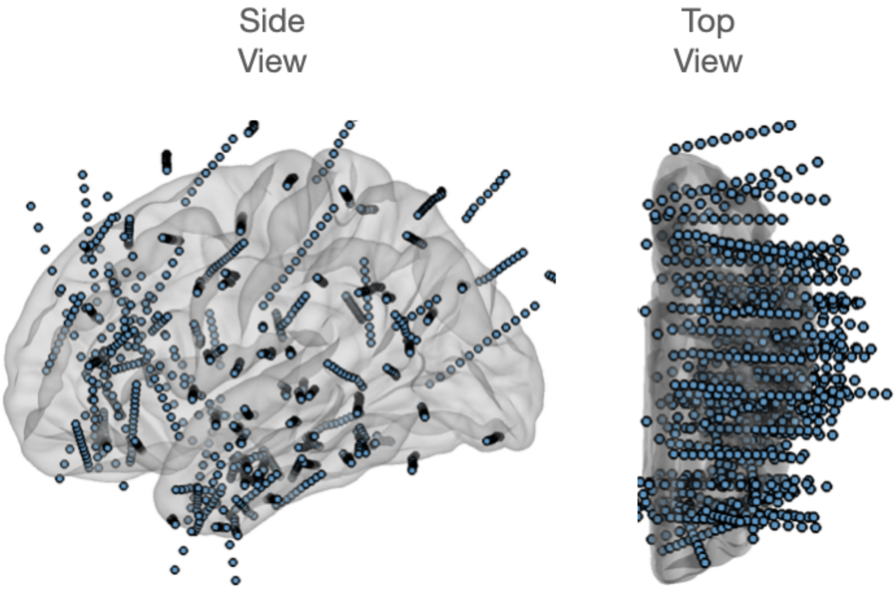


Figure 3.1: Visualisation of electrode placement for all 654 electrodes across 8 patients

to be associated with the high-level phenomena that we are interested in [19] [29].

3.1.1 Task Design

Participants either heard (auditory modality) or read (visual modality) four-word sentences that were sequentially presented (see Figure 3.2). There were three types of sentences: semantic (e.g., "the girls ate cakes," called GS sentences), non-semantic (e.g., "the cakes ate girls," called NS sentences), and ungrammatical (e.g., "the ate girls cake," called NG sentences), and two modalities of presentation (visual and auditory).

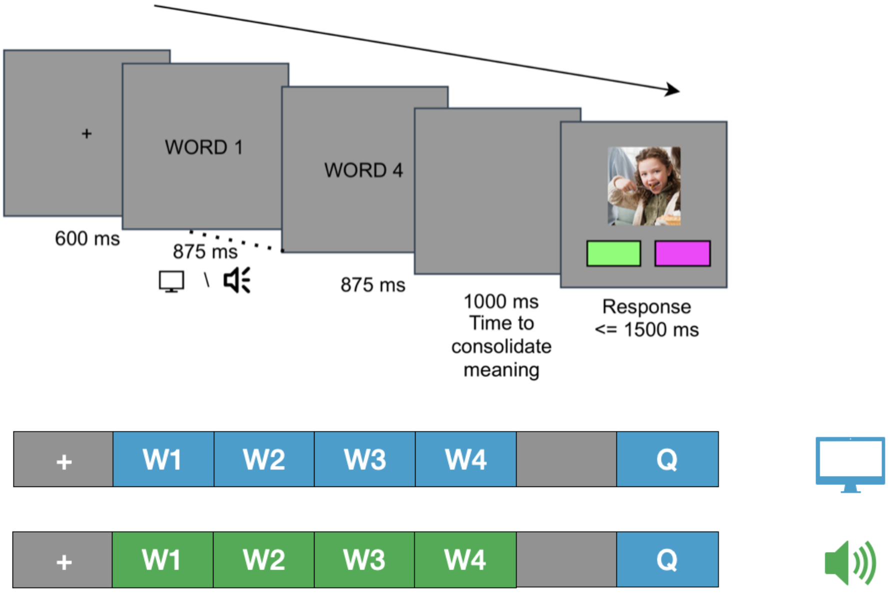


Figure 3.2: An illustration of the task during which the neural data was collected. Patients were presented with four consecutive words, one at a time. After the words were displayed, there was a delay period of 1 second. Following the delay, an image was shown. The subjects were required to specify via a button press whether the image matched the sentence formed by the words or not.

To assess comprehension, participants were asked to indicate whether the sentence adequately described an image that appeared after a 1,000 ms interval following the last word. Participants performed the task correctly on 86±13% of the trials.

The task was coded and displayed in MATLAB using the Psychophysics Toolbox.

3.1.2 Splitting the Experiment into Image, Visual Word, and Auditory Word Datasets

The initial task was designed to investigate language, not vision or auditory processing, so we had to split the original dataset into segments relevant to our analysis.

Since we were interested in how well neural networks could explain visual and auditory data, we divided the initial dataset into three separate datasets: a dataset of image presentations, a dataset of auditory word presentations, and a dataset of visual word presentations.

For each of these presentations, we considered a 1500 ms activity window around the presentation onset (-400 ms to 1100 ms).

The Image (image_v) dataset consists of images presented at the end of each trial alongside two buttons, and the corresponding electrode signals. We have 151 unique images (examples shown in figure), each repeated exactly 6 times (3 for each sentence type: GS/NS/NG * 2 for sentence modality presentations).

The Visual Word (w2_v) dataset consists of images of words and the corresponding electrode activities. It includes 236 unique words, with the number of repetitions varying from 1 to 24. The first word was always "the," and all 236 words appeared in the second position, so we decided to build our word dataset with the electrode signals corresponding to the presentation of the second word. Alternatively, we could have extracted signals from all four word positions, but we opted to minimize the effect of word order on electrode responses, even if the word itself was the same.

However, it may be worth exploring a variant of this dataset that includes repetitions of the same word across all four positions. This would increase the number of repeats and help us identify electrodes that are responsive and selective to the stimuli of interest.

The Auditory Word (w2_a) dataset consists of 1-2 second spoken speech recordings of words and the corresponding electrode activities. The only difference from the Visual Word dataset is that here, the stimuli are in audio format; otherwise, all statistics are the same as in the Visual Word dataset.

3.1.3 Normalization by the Baseline Period

Electrode signals at any given time point can represent information from previous events or even reflect the brain's complex inner state. What we are interested in, however, is the activity caused by the stimulus presentation. One way to decouple these factors is to normalize the activity within the window of interest by subtracting the mean and dividing by the standard deviation of the activity during the baseline period. In our case, the baseline period was chosen as the 400 ms window before the first word was presented.

Our experiments were conducted using both raw and normalized versions of electrode signal recordings, and we found that the raw recordings provided us with better signals (more on this in section 4.2).

3.2 Electrode Reliability with Split-Half Consistency

Split-half consistency is a measure used to assess the reliability of data by evaluating the agreement between two halves of the dataset. In this work, split-half consistency was applied to electrode recordings in response to two types of stimuli: images and audio recordings.

The data from the electrode recordings were divided into two equal halves, and the correlation between the values of these two halves was calculated. This approach allows us to assess how consistently the brain's response to the stimuli is captured by the electrodes, providing a measure of the reliability and stability of the neural recordings. High split-half consistency would indicate that the electrode recordings are dependable and that the observed neural responses are robust across different subsets of the data.

Since we make our estimate with a half of a dataset, the correlation for split-half reliability needs to be adjusted to account for the fact that we have half as much data. This adjustment, known as the Spearman-Brown correction [35] [4], can be computed using the following formula:

$$r^* = \frac{2 \cdot r}{1 + r}$$

Note that in cases where $r < -\frac{1}{3}$, we may obtain $r^* < -1$, which might be unexpected at first glance, but is perfectly legitimate under this formula.

We performed four different half-splits, allowing us to calculate the mean and standard deviation for our estimates of self-consistency.

3.3 Explaining Electrode Activity with Neural Network Representations

In this chapter, we describe the approach taken to compare human brain activity, recorded via intracranial electrodes (sEEG), with neural network representations from multiple models: ResNet-50, AlexNet, and a regression model trained on pixels of the image.

Following the approach established in previous studies [42] [34], we align the activations from these models with the recorded brain responses to assess how well they explain the patterns of brain activity elicited by visual stimuli.

In our setting, we explain each electrode separately using the following algorithm:

1. Pass the stimulus through a neural network;
2. Extract features from a specific layer;
3. Reduce the number of features to 1,000 components using PCA;
4. Train a ridge regression on features from all train-split stimuli to predict the average activity of an electrode during the presentation of the same stimuli;
5. Measure the correlation r between the predicted mean activity and the actual activity on the validation set.

An important detail is that for our regression training, we only consider stimuli that have been repeated at least four times. We also conducted experiments allowing stimuli with only two repetitions; however, the results were noticeably noisier.

3.4 Permutation Test

We want to understand whether the correlation values we obtain are better than random predictions. To quantify this, we performed a permutation test on our mapping function (the ridge regression).

For each electrode, we trained 1,000 regressions on bootstrapped datasets with randomly shuffled labels. For example, this means that we might extract features from a neural network when an image of a "cat" was passed through, but predict the electrode's mean activity during the presentation of an image of a "dog." The 1,000 resulting correlation values form an approximately normal distribution (due to bootstrapping).

If the correlation from our actual model (the one trained on the original dataset) is greater than 95% of the values in the permutation distribution, we assign a p-value of 0.05. This means that the probability of obtaining an r value this high by pure chance is 5%.

Chapter 4

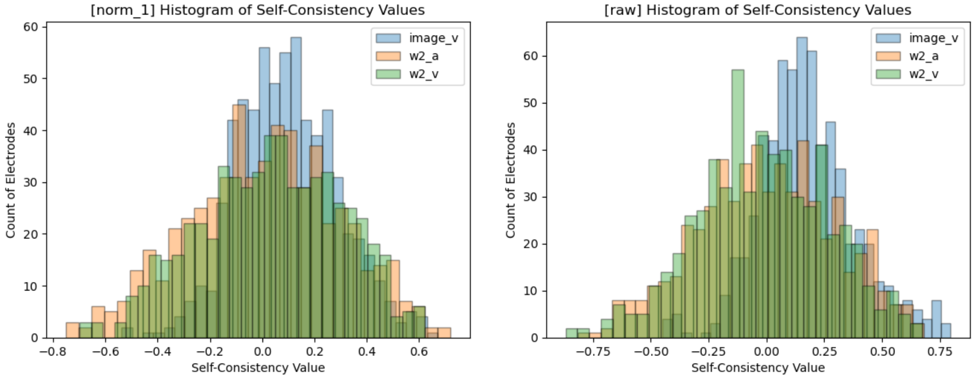
Results

4.1 Signal or Noise? Self-Consistency of Electrodes

Before comparing different models, we need to ensure that the data we're working with contains information relevant to the phenomena we want to study.

Given that we have recordings from many electrodes, placed across the brain's surface, we expect that most electrodes will not capture signals related to the task. Therefore, our first step is to identify electrodes that are reactive and selective to the visual and auditory stimuli.

One way to approach this is by finding electrodes that exhibit self-consistency.



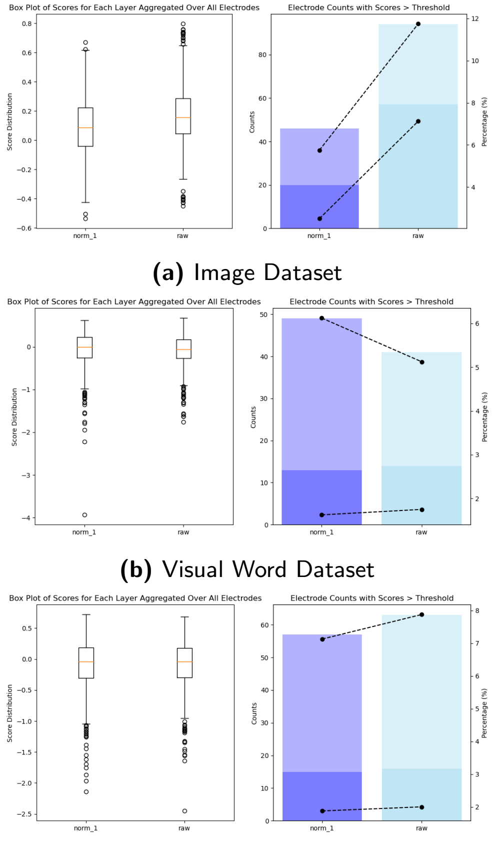
(a) Baseline period normalization

(b) No normalization

Figure 4.1: Histograms of distributions of Half-Split self consistency values with Spearman-Brown correction for all electrodes.

Figure 4.1 shows the distributions of Spearman-Brown corrected self-consistency values for all electrodes, across all three tasks, and both normalization strategies. Interestingly, some of our self-consistency values have absolute values less than -1. This occurs because the r value before Spearman-Brown correction is less than $-\frac{1}{3}$, which is unlikely in high-repetition settings but entirely possible with a low number of repetitions due to chance. For our histogram, we filtered out those electrodes.

We believe that this variability is expected since the electrodes were placed throughout the brain. Not all electrodes are expected to consistently react to the visual or auditory stimuli we control for. Most are expected to be irrelevant to the task.



(a) Image Dataset

(b) Visual Word Dataset

(c) Audio Word Dataset

Figure 4.2: Comparison of Self-Consistency values for the three tasks (a)(b)(c) depending on neural recording normalization type. Sub-figures on the left show the distribution of self consistency values, sub-figures on the right show the counts of electrodes with mean self-consistency above 0.4 (lighter shades of each color) and the count of electrodes with mean - std of self-consistency being above 0.4

Next, we need to determine which normalization type (norm_1 or raw) yields more consistent electrodes. Figure 4.2 shows box plots with distributions of scores, and the count of electrodes with self-consistency (SC) values above 0.4.

We also visualized two thresholding approaches: $\mu_i \geq 0.4$ and $\mu_i - \sigma_i \geq 0.4$. While the difference is not substantial, for the image data, the raw electrode recordings perform slightly better. For all future analyses, we will only consider electrodes i such that $\mu_i - \sigma_i \geq 0.4$.

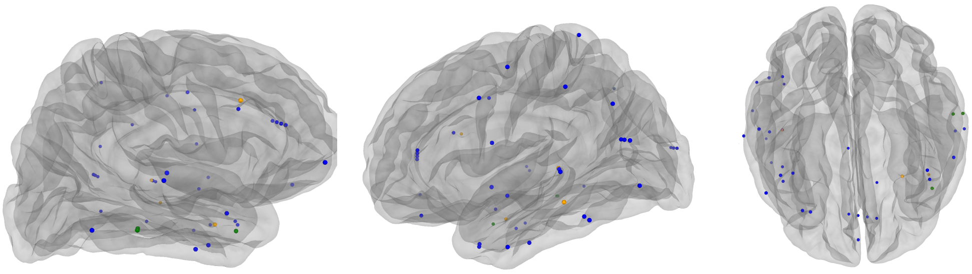


Figure 4.3: Visualisation of electrode location for electrodes that have self-consistency values above 0.5. Colors represent different tasks, blue for image, green for word_visual, yellow for word_auditory

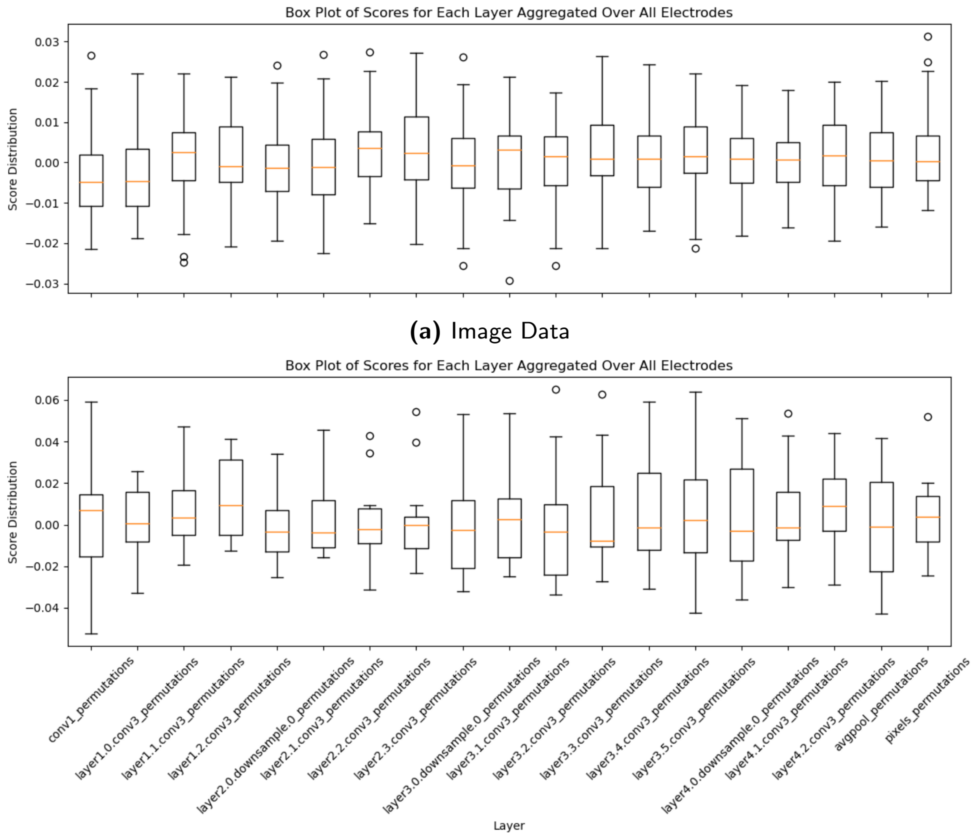
Figure 4.3 shows the locations of electrodes with self-consistency values above 0.5. Most of these electrodes come from the image-viewing dataset.

Interestingly, 2/4 of the word-auditory consistent electrodes are located in what is believed to be the primary auditory cortex, and 3/3 of the word-visual consistent electrodes are in the inferior temporal area, which is part of the ventral visual pathway.

4.2 Permutation Statistics

Results presented in following sections show the performance of the Resnet-50 model with a pixel baseline for reference. Same statistics are available for Alexnet, and are in Appendix A. Alexnet did not yield qualitatively different results.

We performed 1000 permutations on each layer of a resnet for each electrode. Figure 4.4 visualises distribution of mean (averaged across 1000 splits) r values. Each box-plot is a distribution with the number of points equal to the number of self-consistent electrodes (which is same within a task). Notice that all the box-plots are centered around 0. Although this was expected, it didn't have to be this way. This permutation test gets rid of per-sample information, however it preserves the biases of the network itself. If we had the case where the architectural biases were enough to explain some variance we would see these boxplots being centered aroung a value that is higher than 0.



(a) Image Data

(b) Visual Word Data

Figure 4.4: Box-plots of the average (across 1000 splits) r values for permutation test of each layer. Each box-plot has number of points equal to self-consistent electrodes.

4.3 Explaining the Variance of Self-Consistent Electrodes

We wanted to determine if any layer from our neural network could explain the self-consistent electrodes better than random predictions. We compared two distributions:

1. The distribution of mean correlations averaged over 5 regressions trained on random permutations
2. The distribution of mean correlations averaged over 5 cross-validated regressions from features of each layer

To determine if our models perform better than random, we conducted a Mann-Whitney test on the two distributions. Since we have multiple electrodes, we applied a Bonferroni correction to the p-value by dividing the desired p-value by the number of electrodes. The distributions are shown in

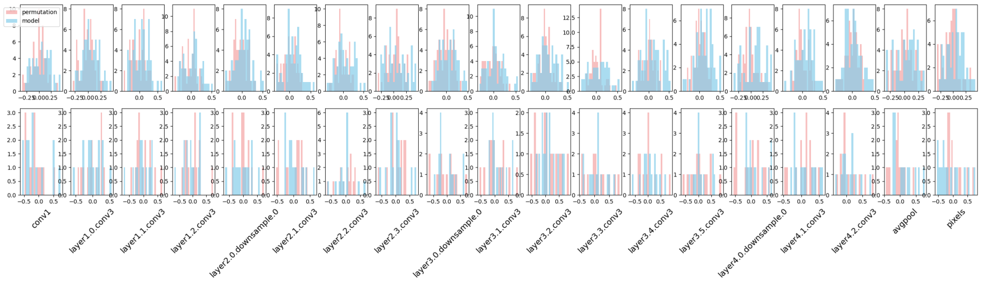


Figure 4.5: Histograms of correlation scores for predicting Self-Consistent electrodes that each layer of a resnet-based model achieves (blue) compared to regressions trained on random permutations of resnet features (red). None of the layers could explain the variance significantly better than the permutation test. Top row is the Image dataset, bottom row is the visual word dataset.

Figure 4.5.

None of the layers have explained the variance better than the permutation test with a p-value of 0.05.

Although none of the layers explains all the self-consistent electrodes better than the random permutation, it doesn't exclude the possibility of a specific electrode being explained significantly better than random.

Figure 4.6 shows the number of electrodes that are explained 1 and 2 standard deviations away from the permutation distribution's mean. For the image dataset, we have a single electrode predicted 2-sigma away, indicating that there is a 5% chance of obtaining such a value if the null hypothesis is true.

Interestingly, the only reliably explained electrode is in the IT cortex. Figure 4.7

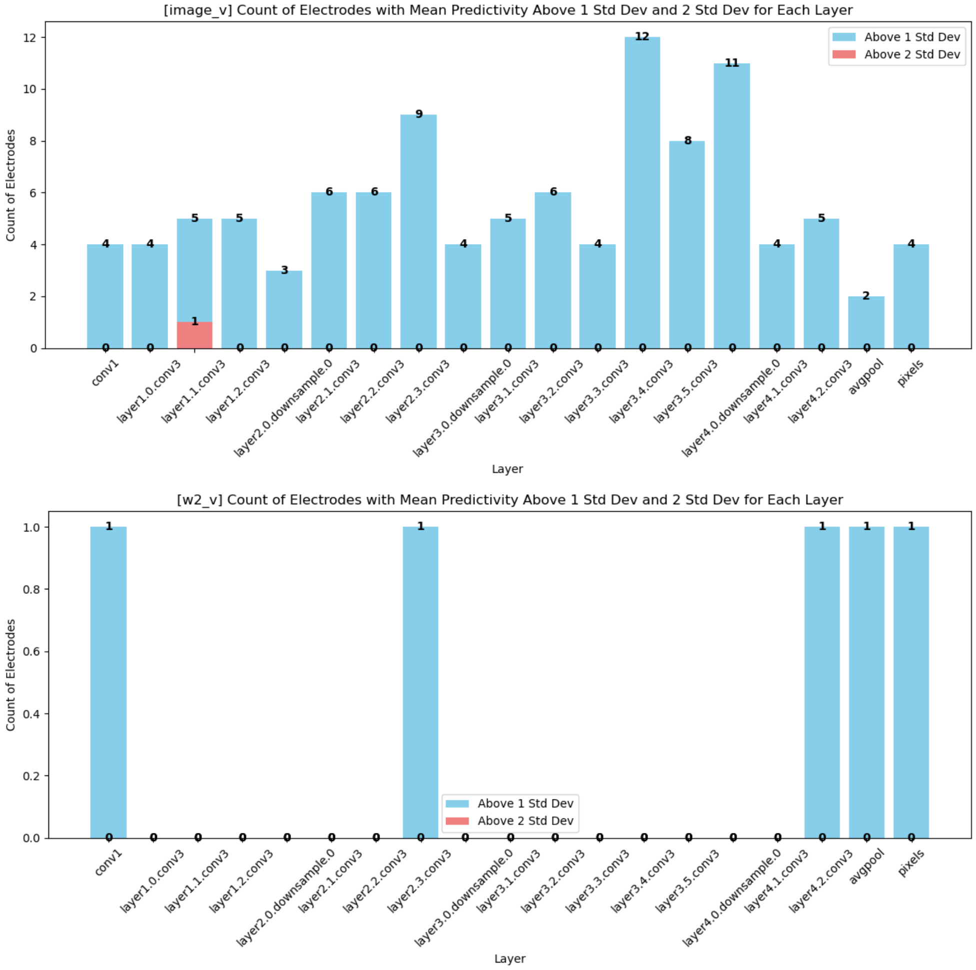


Figure 4.6: Bar plots showing how many of the electrodes can be predicted (by each layer of a resnet) above 1 (blue) and 2 (red) standard deviations away from the mean permutation value for image viewing task (top tow) and visual word task (bottom row)

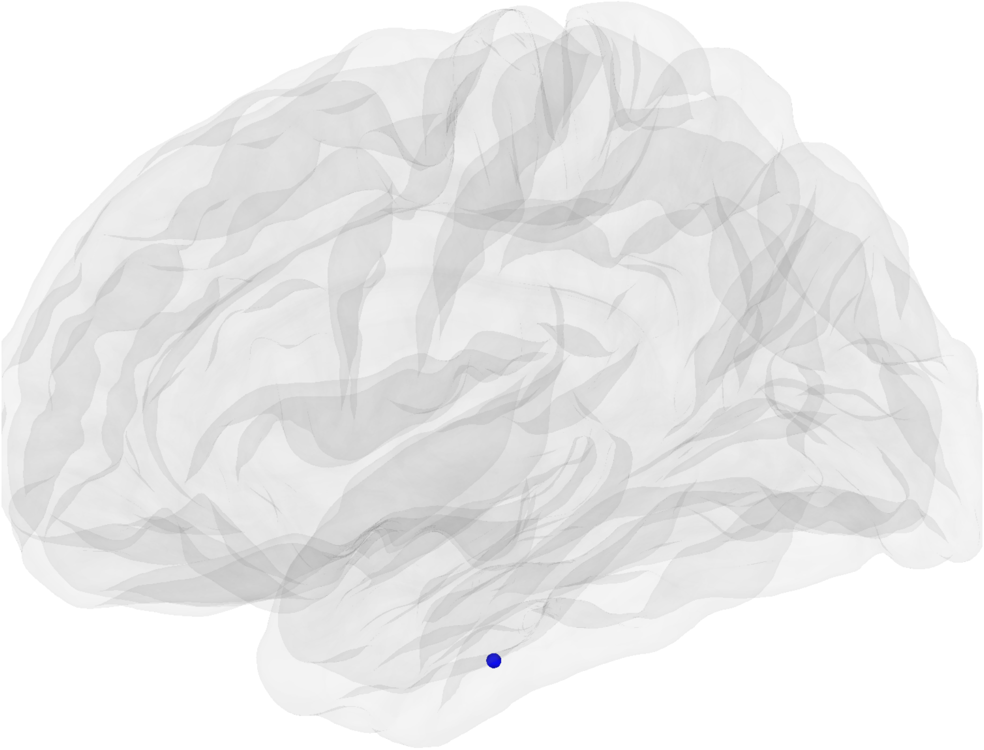


Figure 4.7: Location of the only visual electrode we could explain significantly for the image viewing task

Chapter 5

Discussion and Future Work

Current results point to one of the two possibilities: 1) the neural signal we are recording may not be sensitive (selective) enough to visual stimuli, or 2) the CNNs we chose are not suitable models of our data. Our uncertainty stems primarily from the limited amount of data we have. We have few unique stimuli (words and images), making it difficult to train the regressions, and few repetitions of each stimulus, which complicates the estimation of the signal-to-noise ratio in each electrode.

5.1 Confidence Intervals

An important statistical test missing in our current analysis is the estimation of confidence intervals for the regressions we trained and compared with the permutation regressions.

In previous figures, we only considered point estimates. The results could differ if we compared the bootstrapped confidence interval of the actual model's performance to the permutation distribution.

This could be done by resampling 1000 versions of our dataset with replacement, training 1000 regressions, and comparing the resulting distribution with the null hypothesis of the permutation distribution.

5.2 Noisy Recordings or Bad Models?

The fact that we observe 50 electrodes with half-split consistency values above 0.4 (for the image dataset) supports the idea that these electrodes are visually selective. However, the somewhat arbitrary threshold of 0.4 is far from perfect, so we may need to use other tools to determine whether the electrodes encode the signals we are interested in.

One approach would be to train a linear decoding model to decode which image is being shown based on the electrode recordings. This could be done either through an n-way classification (where n is the number of unique images) or through n binary classifiers, where the i-th model classifies whether it is the i-th image or anything else (one-vs-all). Successfully decoding the image with a linear model would suggest that visually selective information is present and that the chosen CNN representations are insufficient models.

However, failure to decode the images would not imply that there is no information. It could indicate that the space in which the signal can be separated is not linear.

Alternatively, we could apply statistical tests to compare the average activity of an electrode by grouping repetitions of the same stimulus, i.e., comparing whether voltage in an electrode for a word ″girl″ differs significantly from the signal for a ″house.″

5.3 More Data

To get more reliable estimates we would need to add more data to our regressions and statistical tests.

For the visual and auditory word datasets we could consider words on positions 3 and 4, and group the activity for each unique word. This would give us more repetitions of the same word, which would be noticeably improvement, considering the extremely low repetition number.

Initially, this dataset has been collected from 17 subjects, with a total of 1573 electrodes. Adding more patient to the study might also be useful.

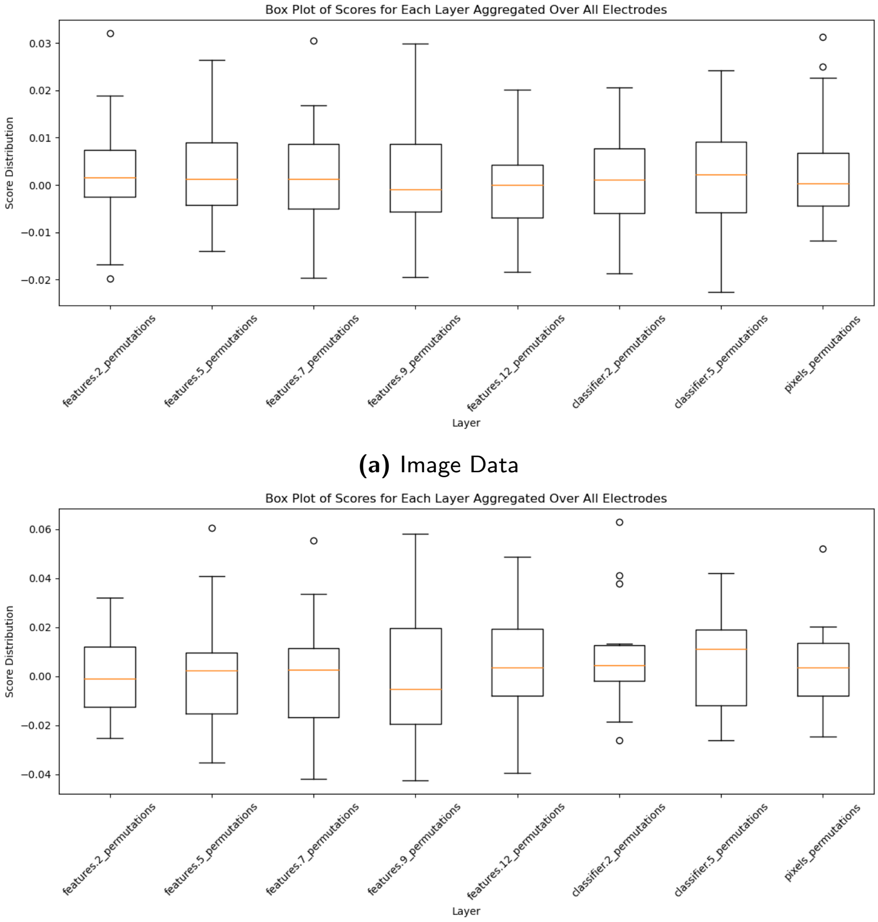
5.4 Different Similarities

Finally, it might be worth considering similarity measures other than correlation. Previous works have also considered Representational difference analysis (RDA) which is less sensitive to the low-repeatition setting that we have.

Finding best metrics for comparison, however, in itself is a topic of active discussion and research.

Appendix A

Alexnet as a Model of Vision



(a) Image Data

(b) Visual Word Data

Figure A.1: [Alexnet] Box-plots of the average (across 1000 splits) r values for permutation test of each layer. Each box-plot has number of points equal to self-consistent electrodes.

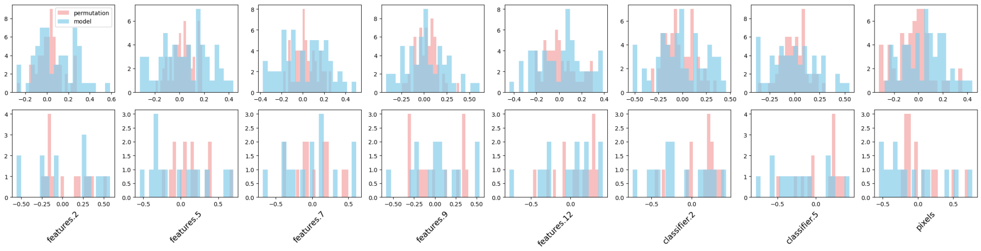


Figure A.2: Histograms of correlation scores for predicting Self-Consistent electrodes that each layer of a alexnet-based model achieves (blue) compared to regressions trained on random permutations of alexnet features (red). None of the layers could explain the variance significantly better than the permutation test. Top row is the Image dataset, bottom row is the visual word dataset.

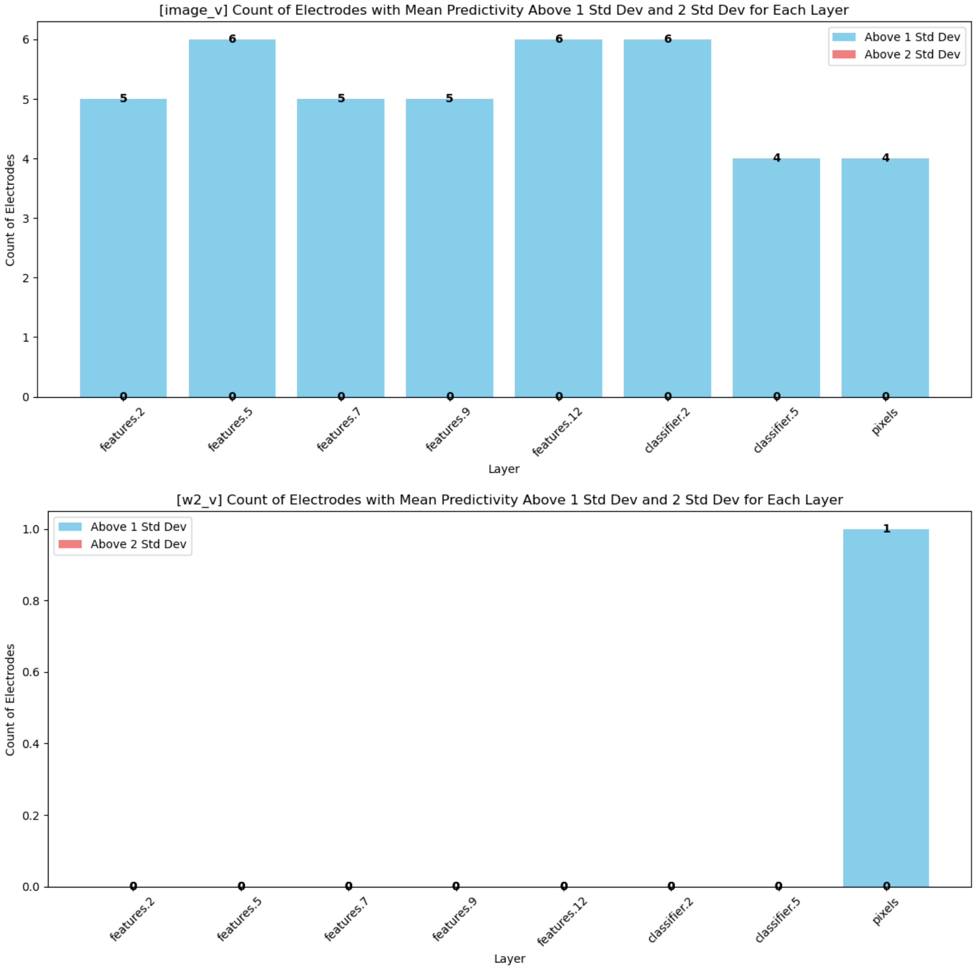


Figure A.3: Bar plots showing how many of the electrodes can be predicted (by each layer of an Alexnet) above 1 (blue) and 2 (red) standard deviations away from the mean permutation value for image viewing task (top tow) and visual word task (bottom row)

Appendix B

Statistics on Neural Datasets

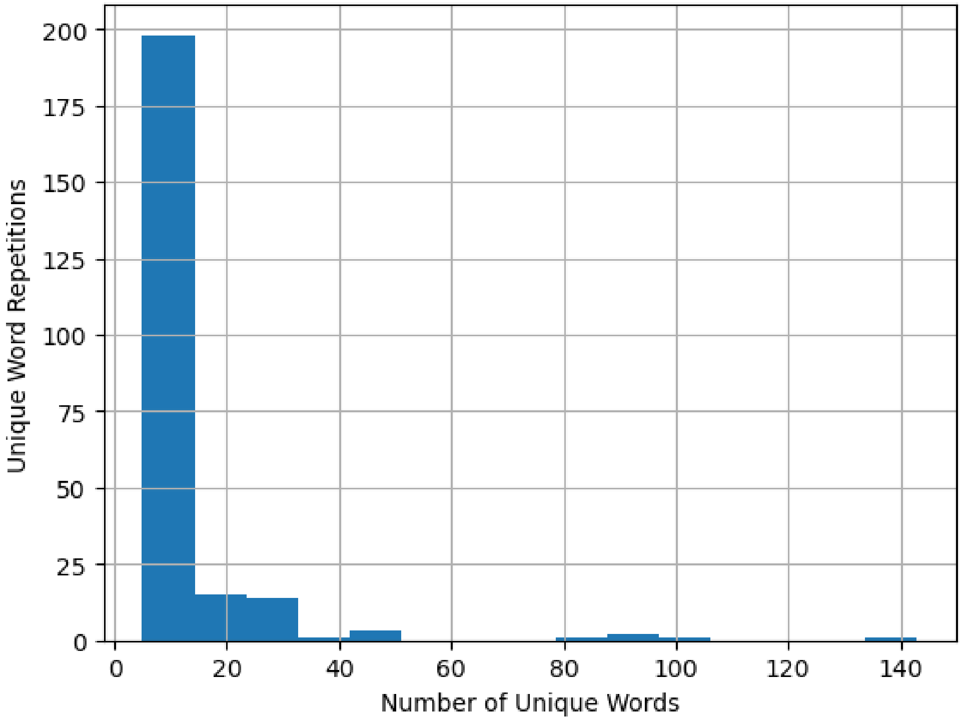
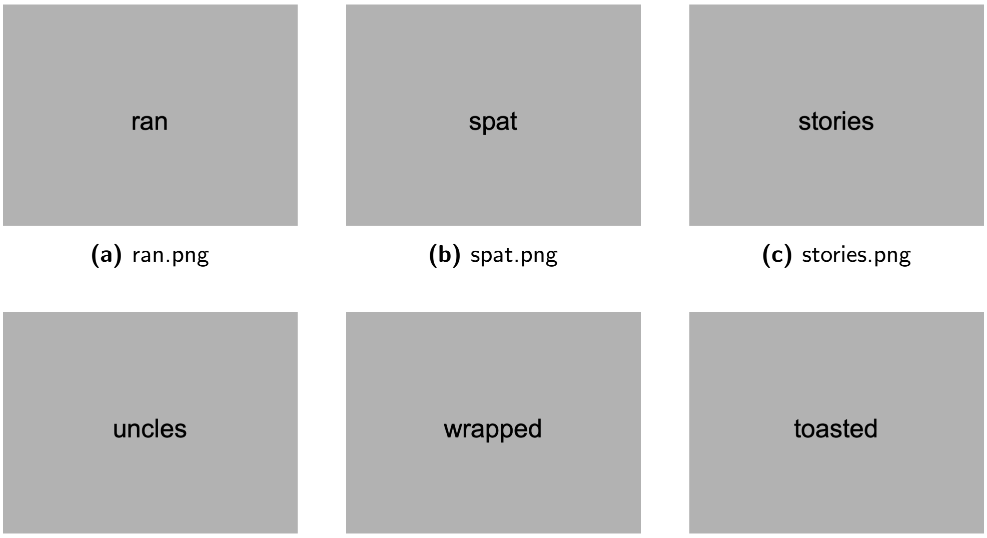


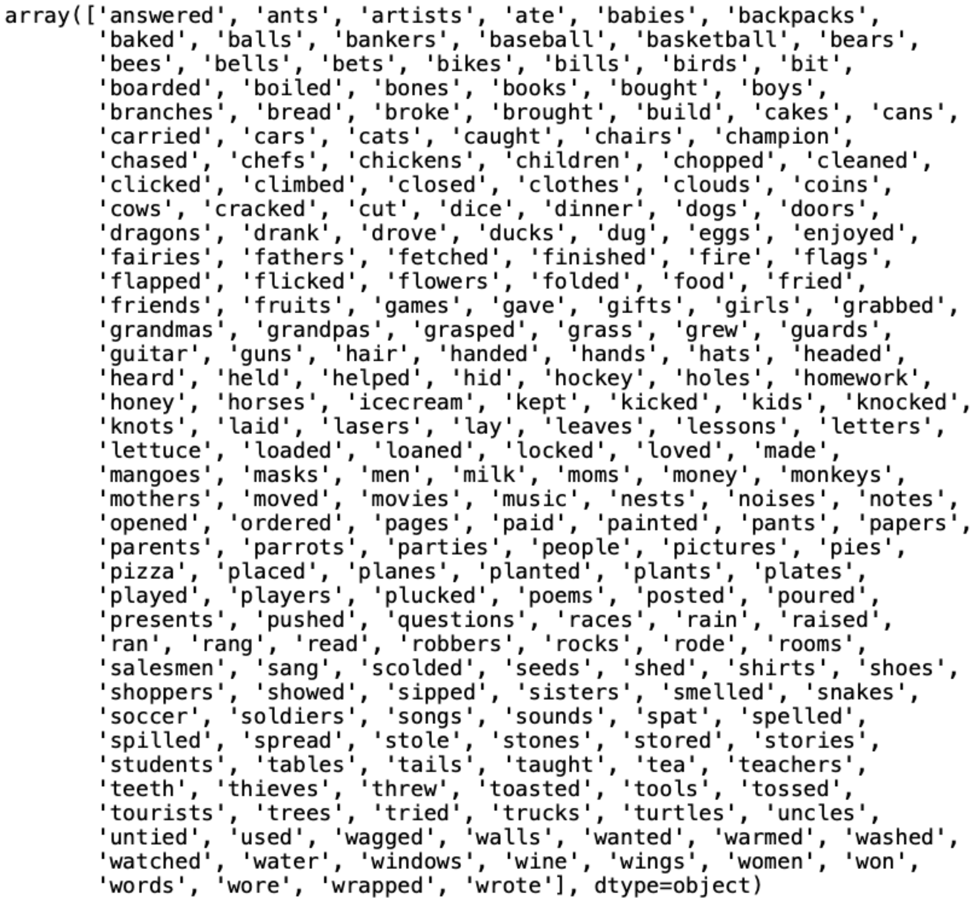
Figure B.1: Histograms of counts



(a) ran.png **(b)** spat.png **(c)** stories.png

(d) uncles.png **(e)** wrapped.png **(f)** toasted.png

Figure B.2: Example visual word presentations



```
array(['answered', 'ants', 'artists', 'ate', 'babies', 'backpacks',
       'baked', 'balls', 'bankers', 'baseball', 'basketball', 'bears',
       'bees', 'bells', 'bets', 'bikes', 'bills', 'birds', 'bit',
       'boarded', 'boiled', 'bones', 'books', 'bought', 'boys',
       'branches', 'bread', 'broke', 'brought', 'build', 'cakes', 'cans',
       'carried', 'cars', 'cats', 'caught', 'chairs', 'champion',
       'chased', 'chefs', 'chickens', 'children', 'chopped', 'cleaned',
       'clicked', 'climbed', 'closed', 'clothes', 'clouds', 'coins',
       'cows', 'cracked', 'cut', 'dice', 'dinner', 'dogs', 'doors',
       'dragons', 'drank', 'drove', 'ducks', 'dug', 'eggs', 'enjoyed',
       'fairies', 'fathers', 'fetched', 'finished', 'fire', 'flags',
       'flapped', 'flicked', 'flowers', 'folded', 'food', 'fried',
       'friends', 'fruits', 'games', 'gave', 'gifts', 'girls', 'grabbed',
       'grandmas', 'grandpas', 'grasped', 'grass', 'grew', 'guards',
       'guitar', 'guns', 'hair', 'handed', 'hands', 'hats', 'headed',
       'heard', 'held', 'helped', 'hid', 'hockey', 'holes', 'homework',
       'honey', 'horses', 'icecream', 'kept', 'kicked', 'kids', 'knocked',
       'knots', 'laid', 'lasers', 'lay', 'leaves', 'lessons', 'letters',
       'lettuce', 'loaded', 'loaned', 'locked', 'loved', 'made',
       'mangoes', 'masks', 'men', 'milk', 'moms', 'money', 'monkeys',
       'mothers', 'moved', 'movies', 'music', 'nests', 'noises', 'notes',
       'opened', 'ordered', 'pages', 'paid', 'painted', 'pants', 'papers',
       'parents', 'parrots', 'parties', 'people', 'pictures', 'pies',
       'pizza', 'placed', 'planes', 'planted', 'plants', 'plates',
       'played', 'players', 'plucked', 'poems', 'posted', 'poured',
       'presents', 'pushed', 'questions', 'races', 'rain', 'raised',
       'ran', 'rang', 'read', 'robbers', 'rocks', 'rode', 'rooms',
       'salesmen', 'sang', 'scolded', 'seeds', 'shed', 'shirts', 'shoes',
       'shoppers', 'showed', 'sipped', 'sisters', 'smelled', 'snakes',
       'soccer', 'soldiers', 'songs', 'sounds', 'spat', 'spelled',
       'spilled', 'spread', 'stole', 'stones', 'stored', 'stories',
       'students', 'tables', 'tails', 'taught', 'tea', 'teachers',
       'teeth', 'thieves', 'threw', 'toasted', 'tools', 'tossed',
       'tourists', 'trees', 'tried', 'trucks', 'turtles', 'uncles',
       'untied', 'used', 'wagged', 'walls', 'wanted', 'warmed', 'washed',
       'watched', 'water', 'windows', 'wine', 'wings', 'women', 'won',
       'words', 'wore', 'wrapped', 'wrote'], dtype=object)
```

Figure B.3: List of all the words presented for both visual and auditory presentations.

(a) the-clouds-brought-rain.jpg

(b) the-girls-ran-races.jpg

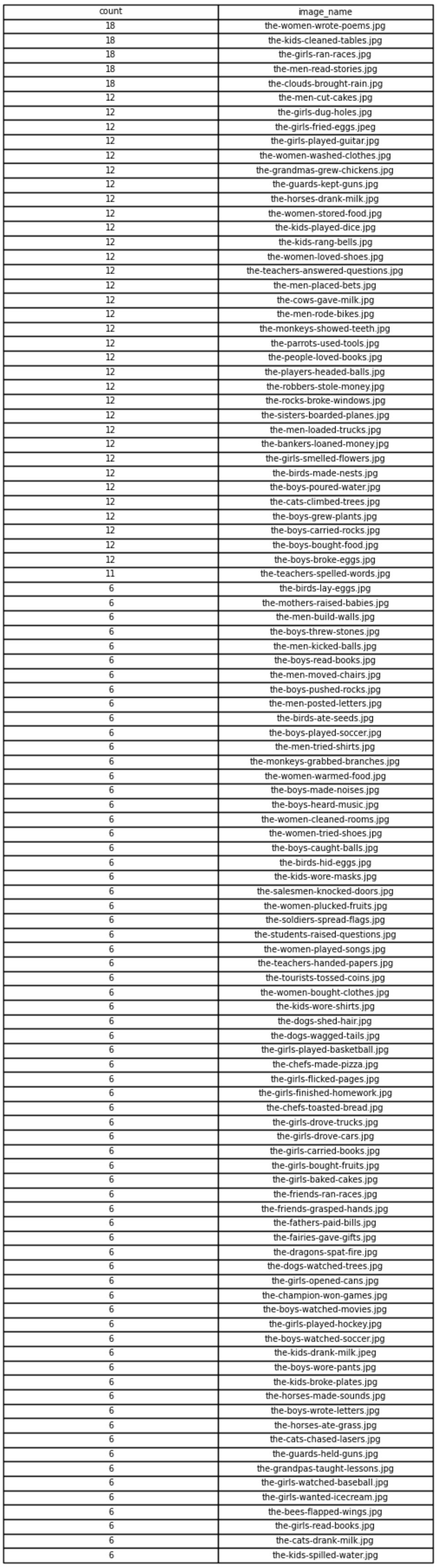
(c) the-kids-cleaned-tables.jpg

(d) the-men-cut-cakes.jpg

(e) the-men-read-stories.jpg

(f) the-women-wrote-poems.jpg

Figure B.4: Example images.



count	image_name
18	the-women-wrote-poems.jpg
18	the-kids-cleaned-tables.jpg
18	the-girls-ran-races.jpg
18	the-men-read-stories.jpg
18	the-clouds-brought-rain.jpg
12	the-men-cut-cakes.jpg
12	the-girls-dug-holes.jpg
12	the-girls-fried-eggs.jpeg
12	the-girls-played-guitar.jpg
12	the-women-washed-clothes.jpg
12	the-grandmas-grew-chickens.jpg
12	the-guards-kept-guns.jpg
12	the-horses-drank-milk.jpg
12	the-women-stored-food.jpg
12	the-kids-played-dice.jpg
12	the-kids-rang-bells.jpg
12	the-women-loved-shoes.jpg
12	the-teachers-answered-questions.jpg
12	the-men-placed-bets.jpg
12	the-cows-gave-milk.jpg
12	the-men-rode-bikes.jpg
12	the-monkeys-showed-teeth.jpg
12	the-parrots-used-tools.jpg
12	the-people-loved-books.jpg
12	the-players-headed-balls.jpg
12	the-robbers-stole-money.jpg
12	the-rocks-broke-windows.jpg
12	the-sisters-boarded-planes.jpg
12	the-men-loaded-trucks.jpg
12	the-bankers-loaned-money.jpg
12	the-girls-smelled-flowers.jpg
12	the-birds-made-nests.jpg
12	the-boys-poured-water.jpg
12	the-cats-climbed-trees.jpg
12	the-boys-grew-plants.jpg
12	the-boys-carried-rocks.jpg
12	the-boys-bought-food.jpg
12	the-boys-broke-eggs.jpg
11	the-teachers-spelled-words.jpg
6	the-birds-lay-eggs.jpg
6	the-mothers-raised-babies.jpg
6	the-men-build-walls.jpg
6	the-boys-threw-stones.jpg
6	the-men-kicked-balls.jpg
6	the-boys-read-books.jpg
6	the-men-moved-chairs.jpg
6	the-boys-pushed-rocks.jpg
6	the-men-posted-letters.jpg
6	the-birds-ate-seeds.jpg
6	the-boys-played-soccer.jpg
6	the-men-tried-shirts.jpg
6	the-monkeys-grabbed-branches.jpg
6	the-women-warmed-food.jpg
6	the-boys-made-noises.jpg
6	the-boys-heard-music.jpg
6	the-women-cleaned-rooms.jpg
6	the-women-tried-shoes.jpg
6	the-boys-caught-balls.jpg
6	the-birds-hid-eggs.jpg
6	the-kids-wore-masks.jpg
6	the-salesmen-knocked-doors.jpg
6	the-women-plucked-fruits.jpg
6	the-soldiers-spread-flags.jpg
6	the-students-raised-questions.jpg
6	the-women-played-songs.jpg
6	the-teachers-handed-papers.jpg
6	the-tourists-tossed-coins.jpg
6	the-women-bought-clothes.jpg
6	the-kids-wore-shirts.jpg
6	the-dogs-shed-hair.jpg
6	the-dogs-wagged-tails.jpg
6	the-girls-played-basketball.jpg
6	the-chefs-made-pizza.jpg
6	the-girls-flicked-pages.jpg
6	the-girls-finished-homework.jpg
6	the-chefs-toasted-bread.jpg
6	the-girls-drove-trucks.jpg
6	the-girls-drove-cars.jpg
6	the-girls-carried-books.jpg
6	the-girls-bought-fruits.jpg
6	the-girls-baked-cakes.jpg
6	the-friends-ran-races.jpg
6	the-friends-grasped-hands.jpg
6	the-fathers-paid-bills.jpg
6	the-fairies-gave-gifts.jpg
6	the-dragons-spat-fire.jpg
6	the-dogs-watched-trees.jpg
6	the-girls-opened-cans.jpg
6	the-champion-won-games.jpg
6	the-boys-watched-movies.jpg
6	the-girls-played-hockey.jpg
6	the-boys-watched-soccer.jpg
6	the-kids-drank-milk.jpeg
6	the-boys-wore-pants.jpg
6	the-kids-broke-plates.jpg
6	the-horses-made-sounds.jpg
6	the-boys-wrote-letters.jpg
6	the-horses-ate-grass.jpg
6	the-cats-chased-lasers.jpg
6	the-guards-held-guns.jpg
6	the-grandpas-taught-lessons.jpg
6	the-girls-watched-baseball.jpg
6	the-girls-wanted-icecream.jpg
6	the-bees-flapped-wings.jpg
6	the-girls-read-books.jpg
6	the-cats-drank-milk.jpg
6	the-kids-spilled-water.jpg

Figure B.5: List of all the images presented with according counts. Each word has been presented at least 6 times (3 sentence types GS/NG/NGNS * 2 modalities)

Bibliography

[1] AF Agarap. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*, 2018.

[2] David Alais and David Burr. The ventriloquist effect results from near-optimal bimodal integration. *Current biology*, 14(3):257–262, 2004.

[3] Emily J Allen, Ghislain St-Yves, Yihan Wu, Jesse L Breedlove, Jacob S Prince, Logan T Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, et al. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, 25(1):116–126, 2022.

[4] William Brown. Some experimental results in the correlation of mental abilities 1. *British Journal of Psychology, 1904-1920*, 3(3):296–322, 1910.

[5] Francesco Cardinale, Massimo Cossu, Laura Castana, Giuseppe Casaceli, Marco Paolo Schiariti, Anna Miserocchi, Dalila Fuschillo, Alessio Moscato, Chiara Caborni, Gabriele Arnulfo, et al. Stereoelectroencephalography: surgical methodology, safety, and stereotactic application accuracy in 500 procedures. *Neurosurgery*, 72(3):353–366, 2013.

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[7] Colin Conwell. *Is visual cortex really "language-aligned"? Perspectives from Model-to-Brain Comparisons in Human and Monkeys on the Natural Scenes Dataset*. PhD thesis, Harvard Medical School.

[8] Colin Conwell, Jacob S Prince, Christopher J Hamblin, and George A Alvarez. Controlled assessment of clip-style language-aligned vision

models in prediction of brain & behavioral data. In *ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models*, 2023.

[9] Colin Conwell, Jacob S. Prince, Kendrick N. Kay, George A. Alvarez, and Talia Konkle. What can 1.8 billion regressions tell us about the pressures shaping high-level visual representation in brains and machines? *bioRxiv*, 2023.

[10] Katharina Dobs, Joanne Yuan, Julio Martinez, and Nancy Kanwisher. Behavioral signatures of face perception emerge in deep neural networks optimized for face recognition. *Proceedings of the National Academy of Sciences*, 120(32):e2220642120, 2023.

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[12] Jeffrey L Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990.

[13] Jay R Gibson and John HR Maunsell. Sensory modality specificity of neural activity related to memory in visual cortex. *Journal of Neurophysiology*, 78(3):1263–1275, 1997.

[14] Karl Gottlieb Grell. *Protozoology*. Springer Science & Business Media, 2013.

[15] Troy A Hackett. Information flow in the auditory cortical network. *Hearing research*, 271(1-2):133–146, 2011.

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

[17] AJ Hudspeth. How the ear's works work: mechanoelectrical transduction and amplification by hair cells. *Comptes rendus biologies*, 328(2):155–162, 2005.

[18] Herbert Spencer Jennings. *Behavior of the lower organisms*. Columbia University Press, 1931.

[19] Ole Jensen, Jochen Kaiser, and Jean-Philippe Lachaux. Human gamma-frequency oscillations associated with attention and memory. *Trends in neurosciences*, 30(7):317–324, 2007.

[20] Alexander JE Kell, Daniel LK Yamins, Erica N Shook, Sam V Norman-Haignere, and Josh H McDermott. A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, 98(3):630–644, 2018.

[21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[22] Yann LeCun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.

[23] John HR Maunsell, Gary Sclar, Tara A Nealey, and Derryl D DePriest. Extraretinal representations in area v4 in the macaque monkey. *Visual neuroscience*, 7(6):561–573, 1991.

[24] Harry McGurk and John MacDonald. Hearing lips and seeing voices. *Nature*, 264(5588):746–748, 1976.

[25] Pranav Misra. Robust and multimodal signals for language in the brain. *graduate thesis*, 2024.

[26] Pranav Misra, Yen-Cheng Shih, Hsiang-Yu Yu, Daniel Weisholtz, Joseph R Madsen, Stone Sceillig, and Gabriel Kreiman. Invariant neural representation of parts of speech in the human brain. *bioRxiv*, pages 2024–01, 2024.

[27] Xuran Pan, Tianzhu Ye, Dongchen Han, Shiji Song, and Gao Huang. Contrastive language-image pre-training with knowledge graphs, 2022.

[28] Josef Parvizi and Sabine Kastner. Promises and limitations of human intracranial electroencephalography. *Nature neuroscience*, 21(4):474–483, 2018.

[29] Supratim Ray, Ernst Niebur, Steven S Hsiao, Alon Sinai, and Nathan E Crone. High-frequency gamma activity (80–150 hz) is increased in human cortex during selective attention. *Clinical Neurophysiology*, 119(1):116–133, 2008.

[30] James L Ringo and STEPHEN G O'Neill. Indirect inputs to ventral temporal cortex of monkey: the influence of unit activity of alerting auditory input, interhemispheric subcortical visual input, reward, and the behavioral response. *Journal of neurophysiology*, 70(6):2215–2225, 1993.

[31] Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

[32] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

[33] Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118, 2021.

[34] Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J. Majaj, Rishi Rajalingham, Elias B. Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, Kailyn Schmidt, Daniel L. K. Yamins, and James J. DiCarlo. Brain-score: Which artificial neural network for object recognition is most brain-like? *bioRxiv preprint*, 2018.

[35] Charles Spearman. Correlation calculated from faulty data. *British journal of psychology*, 3(3):271, 1910.

[36] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[37] Barry E Stein and M Alex Meredith. *The merging of the senses*. MIT press, 1993.

[38] Barry E Stein, Terrence R Stanford, and Benjamin A Rowland. Multisensory integration and the society for neuroscience: Then and now. *Journal of Neuroscience*, 40(1):3–11, 2020.

[39] Vighnesh Subramaniam, Colin Conwell, Christopher Wang, Gabriel Kreiman, Boris Katz, Ignacio Cases, and Andrei Barbu. Revealing vision-language integration in the brain with multimodal networks. *arXiv preprint arXiv:2406.14481*, 2024.

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

[41] Aria Y Wang, Kendrick Kay, Thomas Naselaris, Michael J Tarr, and Leila Wehbe. Better models of human high-level visual cortex emerge from

natural language supervision with a large and diverse dataset. *Nature Machine Intelligence*, 5(12):1415–1426, 2023.

[42] Daniel L. K. Yamins, Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James 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.