

Encoding and Decoding Brain Signals in the Primate Visual Cortex Using Deep Learning

著者	伊達 裕人
学位授与機関	Tohoku University
学位授与番号	11301甲第19344号
URL	http://hdl.handle.net/10097/00130234

TOHOKU UNIVERSITY
Graduate School of Information Sciences

**Encoding and Decoding Brain Signals in the
Primate Visual Cortex Using Deep Learning**
(深層学習を用いた霊長類視覚皮質における脳活動の符
号化と復号化)

A dissertation submitted for the degree of Doctor of Philosophy
(Information Science)

Graduate School of Information Sciences

by

Hiroto Date

January 14, 2020

Encoding and Decoding Brain Signals in the Primate Visual Cortex Using Deep Learning

Hiroto Date

Abstract

When a person sees an image, complex brain activities occur in diverse scales of the brain, from single-neuron spikes to synchronous oscillations of neural populations. Researchers have investigated the property of diverse brain activities using experimental and theoretical methods to understand how the visual system in the brain works. Cognitive neuroscience studies the underlying mechanism of visual perception. To develop quantitative models, various methods have been proposed for modeling the relationship between perceptual information and brain activities. One of the most classic method is to test whether brain activities significantly change between different conditions (e.g., face and non-face images).

Recently, various machine learning methods have been used as brain *encoding* and *decoding* models, as the result of the growth of the machine learning literature and the increase of available computational resources. Encoding models are used to predict brain activities from perceptual information. On the other hand, decoding models are used to predict perceptual information from brain activities. The flexibility of encoding and decoding models make them important tools for cognitive neuroscience. Furthermore, developing better encoding and decoding models in hard conditions is an important problem for achieving real-world brain-computer interface (BCI) systems.

In this thesis, we study brain encoding and decoding methods using *deep learning*. Deep learning has achieved state-of-the-art performance on various tasks in artificial intelligence (AI), such as general object recognition, machine translation, speech recognition, and automatic game playing. Using deep learning for brain encoding and decoding is promising, because both brain activities and perceptual information are complex spatiotemporal data. In Chapter 2, to investigate how visual selectivity differs across frequency bands, we analyze frequency-specific activities in ECoG signals recorded from the macaque inferior temporal cortex (ITC) using rich hierarchical visual representations extracted from a deep convolutional neural network (CNN). In Chapter 3, we develop deep learning-based models that reconstruct diverse natural images from brain signals. To investigate what kind of models are effective for the task, we trained and evaluated multiple state-of-the-art image restoration models in deep learning. In Chapter 4, we develop deep learning methods for channel-agnostic brain decoding across multiple subjects. Inspired from multi-instance learning, we propose a novel decoder architecture that can handle a variable number of channels, has permutation invariance to the order of channels, and can capture inter-channel relationships.

Our results in this thesis indicate the importance of deep learning in encoding and decoding complex brain signals. Furthermore, we believe that our proposed methods are effective tools for analyzing and reading brain signals in a lot of future cognitive neuroscience research and real-world BCI applications.

Contents

1	Introduction	1
1.1	Background	1
1.2	Goals	3
1.3	Recording brain activities	6
1.3.1	Neuron	6
1.3.2	Techniques for recording brain activities	7
1.4	Existing methods for analysing brain activities	13
1.4.1	Classical statistical hypothesis testing methods	13
1.4.2	Encoding and decoding methods	14
1.5	Deep Learning	18
1.5.1	Neural networks and deep learning	18
1.5.2	Perceptrons and Multilayer Perceptrons	20
1.5.3	Deep Learning	20
1.5.4	Convolutional Neural Networks	21
1.5.5	Deep Learning for Brain Activity Analysis	24
1.6	Thesis structure	24
2	Encoding and Analyzing Frequency-Specific ECoG Signals Using Hierarchical Visual Features	25
2.1	Introduction	25
2.2	Background	27
2.3	Related work	29
2.3.1	Hierarchical relationships in the primate visual cortex and CNNs	30

2.3.2	Representation similarity between the primate visual cortex and CNNs	31
2.3.3	Spatial relationship: anatomical hierarchy of the primate visual cortex and CNNs	32
2.3.4	Temporal relationship: latency of brain activities and CNNs	33
2.3.5	Different time-frequency bands for distinct visual information	33
2.3.6	Different time-frequency bands for complementary roles in cortical information processing	34
2.3.7	Open problems and the purpose of this work	35
2.4	Materials and methods	37
2.4.1	Image set	37
2.4.2	Subjects	37
2.4.3	Details of our ECoG system	37
2.4.4	ECoG recording	39
2.4.5	Time-frequency decomposition	40
2.4.6	Extracting hierarchical visual features from convolutional neural networks	42
2.4.7	Encoding ECoG features from CNN features	43
2.5	Experiments	45
2.5.1	Theta and gamma bands are better predicted from CNN features	45
2.5.2	Theta and gamma bands are better predicted from higher and lower CNN layers, respectively	47
2.5.3	Theta and gamma bands are better predicted in later and earlier time windows, respectively	50
2.5.4	Visualizing the selectivity of theta- and gamma-band encoding models	50
2.6	Discussion and Conclusion	51
3	Natural Image Reconstruction from ECoG Signals Using Deep Learning	54

3.1	Introduction	54
3.2	Related work	55
3.2.1	Image identification from fMRI data	56
3.2.2	Image reconstruction from fMRI data	57
3.2.3	Image reconstruction from EEG signals	57
3.2.4	Open problems and the purpose of this work	58
3.3	Methods	58
3.3.1	Background	58
3.3.2	Models	60
3.3.3	Network architecture	62
3.3.4	Stabilizing GAN training	63
3.4	Experiments	64
3.4.1	Training	64
3.4.2	Reconstruction results	69
3.4.3	Quantitative results	70
3.5	Discussion and Conclusion	71
4	Deep Learning for Channel-Agnostic Brain Decoding across Multiple Subjects	74
4.1	Introduction	74
4.2	Related work	76
4.2.1	Multi-subject decoding for fMRI data	76
4.2.2	Multi-subject decoding for EEG signals	77
4.2.3	Open problems and the purpose of this work	77
4.3	Methods	78
4.3.1	Channel-agnostic brain decoding as multi-instance learning	78
4.3.2	Channel-wise transform	79
4.3.3	Across-channel transform	80
4.3.4	Multi-channel pooling	80
4.3.5	Baselines	81

4.4	Experiments	84
4.4.1	Training	84
4.4.2	Classification results: Across-channel transform	85
4.4.3	Classification results: Multi-channel pooling	86
4.4.4	Visualization of self-attention weights	88
4.5	Discussion and Conclusion	88
5	Conclusions	90

List of Figures

1-1	An illustrative diagram of a neuron. (Source: Figure by BruceBlaus / CC BY 3.0)	6
1-2	An example scene of EEG recording. (Source: Photo by Tim Sheerman-Chase / CC BY 2.0)	8
1-3	An example of ECoG array implantation. (Source: Matsuo <i>et al.</i> [1] / CC BY 4.0)	9
1-4	An example scene of MEG recording. (Image source: National Institute of Mental Health, National Institutes of Health, Department of Health and Human Services)	10
1-5	An example of fMRI activities. (Source: Lizette <i>et al.</i> [2] / CC BY 3.0)	11
1-6	Brain encoding and decoding.	14
1-7	Computation of a perceptron (unit)	19
1-8	Multilayer perceptrons (MLPs)	19
1-9	Basic architecture of CNNs	21
1-10	Computation of a convolutional layer	22
1-11	Computation of a pooling layer	22

2-1	Encoding frequency-specific ECoG activities from CNN features. We trained encoding models that predict frequency-specific ECoG activities given visual features extracted from a pretrained CNN. We recorded ECoG signals from the macaque ITC while presenting natural images, and extracted frequency-specific amplitude using time-frequency decomposition. Using the same image set, we extracted visual features from each convolution and fully-connected (FC) layer of a pretrained CNN. For convolutional layers, which have three dimensions (width, height, channels), we downsampled features over the width and height (global average pooling). After feature extraction, we trained ridge regression models that predict ECoG amplitude at a specific site, frequency, and time window from CNN features at a specific layer. . . .	26
2-2	Visual processing in the primate visual cortex	29
2-3	Examples images from each class in our image set. From top to bottom, building, body part, face, fruit, insect, and tool.	36
2-4	Lateral view of the macaque brain with an ECoG electrode implanted (the right hemisphere of Subject 1). Reconstructed with post-mortem observations. Pink dots indicate the position of the electrode contacts. The scale bar indicates 5 mm. Among total 128 contacts, 108 visible ones from this view are shown. The other 20 contacts are located on the ventral or medial surface of the cortex (not visible from this view).	38
2-5	Stimulus presentation in ECoG recording.	39
2-6	A visualized event-related spectral perturbation (ERSP).	40
2-7	The architecture of VGG-16 network.	41

2-8	Comparison of the prediction performance. In the test set, the prediction performance was measured as Pearson’s correlation coefficient between ground truth and predicted values. (A) The prediction performance over the frequencies and time windows. For each site, the maximum performance over the CNN layers was extracted. The average performance over sites that showed better performance than the significance threshold ($p < 0.0001$ in the permutation test) is shown here. (B) The prediction performance over the frequencies. Red dots indicate the prediction performance of each ECoG site. For each site, the maximum performance over the time windows and CNN layers was extracted. Only results above the significance threshold are shown here ($p < 0.0001$ in the permutation test). Blue line indicate the mean prediction performance over the ECoG sites. Blue error bars indicate the standard error of the prediction performance over the ECoG sites. . .	44
2-9	Assignments of the CNN layers. For each site, the maximum performance over the time windows is extracted. Only sites above the significance threshold are shown here ($p < 0.0001$ in the permutation test). (A) Topographical visualizations of assigned time windows. The top, bottom, right, and left side of each electrode map corresponds to the dorsal, ventral, anterior, and posterior part of the macaque brain, respectively. The color at each site indicates the assigned layer. (B) Proportion of each CNN layer in assignments.	46
2-10	Assignments of the time windows. For each site, the maximum performance over the CNN layers is extracted. Only sites above the significance threshold are shown here ($p < 0.0001$ in the permutation test). (A) Topographical visualizations of assigned time windows. The top, bottom, right, and left side of each electrode map corresponds to the dorsal, ventral, anterior, and posterior part of the macaque brain, respectively. The color at each site indicates the latency of the assigned time window. (B) Proportion of each time window in assignments. . .	48

2-11	Examples of optimized (maximize, minimize) and preferred (top, bottom) images for theta- and gamma-band encoding models. Optimized images were produced by updating randomly-initialized images so as to maximize or minimize the predicted value of each encoding model. Preferred images were selected based on predicted values of each encoding model on the test set.	49
3-1	Image reconstruction models (generator: G, discriminator: D)) in our experiments. (a) The L1 model is trained only with the pixel-wise error (L1 loss). (b) The L1-VGG-GAN model is trained with a weighted combination of L1, perceptual, and adversarial losses. We used a pre-trained VGG-16 network for computing perceptual loss. (c) The conditional GAN (cGAN) model is trained with the conditional formulation of GAN. In this case, the discriminator receives both an image (reconstruction or ground truth) and brain signals.	59
3-2	Image reconstruction from brain signals. Single-trial brain signals are first transformed into a vector by a temporal (1D) convolution network (TCN). Then, the vector is used to produce a reconstruction by a convolutional neural network (CNN).	62
3-3	Example reconstructions for each subject and model. The first row shows ground truth images. The second to fourth rows show reconstruction results for Subject 1. The fifth to seventh rows show reconstruction results for Subject 2.	65
3-4	Example reconstructions with ECoG signals (downsampling width: 300 ms). The first row shows presented images (ground truth). The second to fourth rows show reconstruction results for Subject 1. The fifth to seventh rows show reconstruction results for Subject 2.	66

3-5	Example reconstructions with downsampled ECoG signals (downsampling width: 100 ms). The first row shows presented images (ground truth). The second to fourth rows show reconstruction results for Subject 1. The fifth to seventh rows show reconstruction results for Subject 2.	67
3-6	Example reconstructions for novel classes (building, body part, tool). The first row shows presented images (ground truth). The second to fourth rows show reconstruction results for Subject 1. The fifth to seventh rows show reconstruction results for Subject 2.	68
4-1	Proposed decoder architecture based on channel-wise temporal convolutional networks (TCNs), across-channel self-attention, and multi-channel pooling.	78
4-2	Comparison of each across-channel transform modules.	82
4-3	Visualized self-attention weights extracted from a trained model (transform: self-attention, pool: mean). The numbers on the horizontal and vertical axes indicate channel indices in each subject’s ECoG electrode. For Subject 1, all 128 channels were implanted on the inferior temporal cortex. For Subject 2, channels 1-128 were implanted on the inferior temporal cortex, and channels 129-192 were implanted on the prefrontal cortex.	87

List of Tables

1.1	Comparison of recording techniques.	12
3.1	The network architecture of the generator (reconstruction) network. .	62
3.2	The network architecture of the discriminator network for L1-VGG-GAN and cGAN models. The conditional part is used only in cGAN models.	63
3.3	Quantitative results of each subject and model. For PSNR and SSIM, higher values are better. For FID, lower values are better.	71
4.1	Comparison of classification accuracy between three across-channel transform modules. The results with the best multi-channel pooling function are shown for each transform module. For each model, we conducted eight runs with different weight initialization, and the average and standard error of classification accuracy over the eight runs are reported here.	85
4.2	Comparison of classification accuracy between multi-channel pooling functions. The results for the multi-head self-attention module are shown here. For each model, we conducted eight runs with different weight initialization, and the average and standard error of classification accuracy over the eight runs are reported here.	86

Chapter 1

Introduction

1.1 Background

When a person sees an image, neuronal activities occur in diverse scales and regions in the brain. Understanding the relationship between complex brain activities and visual perception is an important goal in visual neuroscience. To examine the relationship, we can use statistical or machine learning models in two contrasting ways [3, 4]. One way is called *brain encoding*, where encoding models are constructed to map visual features into brain activities. The other way is called *brain decoding*, where decoding models are constructed to map brain activities into visual features. Because brain encoding and decoding are computationally opposite models, their roles in understanding the brain mechanism is different. Brain encoding models can be used to test specific computational models of the brain, by evaluating how well each encoding model predicts actual brain responses. On the other hand, brain decoding models help researchers investigate what kind of visual features are related to brain activities, by evaluating how well each decoding model predicts specific visual features from brain activities.

Both brain encoding and decoding models have been used in the literature of cognitive and computational neuroscience. While most older studies used simple statistical or machine learning methods such as linear models and support-vector machines

(SVMs)[5], a number of recent studies used deep learning [6, 7] for brain encoding or decoding tasks, following successful applications of deep learning to various tasks in computer vision [8, 9, 10, 11]. Deep learning is a field where large, differentiable computational graphs, called *neural networks*, are employed to solve diverse kinds of tasks using the back propagation algorithm and large-scale training datasets. In brain encoding, several recent studies [12, 13, 14] analyzed brain activities using convolutional neural networks (CNNs) that were pre-trained on a large-scale visual object classification task. They compared visually-evoked brain activities and visual representations in CNNs, and found a similarity between the hierarchical organization of the primate visual cortex and the layer hierarchy of CNNs. In brain decoding, [15] compared the performance of decoding CNN features from human functional magnetic resonance imaging (fMRI) data. They also observed a similarity between the hierarchical organization of the human visual cortex and the layer hierarchy of CNNs. Several other studies [16, 17, 18] used deep learning for reconstructing presented images from human brain activities.

Deep learning is a rapidly growing field in machine learning, and has been an important method in analyzing brain activities. However, most applications of deep learning in neuroscience are on either neuronal spiking activities (single-unit activity: SUA, multi-unit activity: MUA) or fMRI data; few studies have applied or developed deep learning methods for analyzing meso-scopic brain activities, such as electroencephalography (EEG), magnetoencephalography (MEG), and electrocorticography (ECoG). While SUA and MUA have far better spatial resolution than the other recording techniques, it is not straightforward to use them to record activities from a large part of the brain or to continue brain recording for days or months. On the other hand, fMRI can cover the whole brain, its temporal resolution is on the level of seconds, which is far slower than the sub-millisecond temporal resolution of neuronal activities. Meso-scopic brain activities can be recorded with sub-millisecond precision, and its recording channels can cover the entire surface of the brain or the scalp. Furthermore, accurate decoding of various features from meso-scopic brain activities is a crucial component for real-world brain-computer interface (BCI) applications

[19, 20]. Therefore, it is crucial to develop better brain encoding and decoding methods to elucidate the relationship between visual experiences and complex neuronal activities and to advance the potential of meso-scopic brain recording for future BCI devices.

Towards this end, we study methods for encoding and decoding ECoG signals using deep learning. We prepare a large-scale ECoG dataset by recording brain signals from the macaque visual cortex while presenting visual stimuli to subjects. In brain encoding, we construct models that predict frequency-specific brain activities using hierarchical visual features extracted from pre-trained CNNs. This experiment is important for understanding how different time-frequency bands in brain activities are related to diverse visual features, because little has been known about information content of frequency-specific brain activities [21]. In brain decoding, we construct models that directly reconstruct presented images solely from ECoG signals. This image reconstruction method is crucial to analyze the importance of rich temporal dynamics in ECoG signals for representations of diverse visual features in the brain. To the best of our knowledge, our reconstruction method is the first to show successful reconstructions of natural images from meso-scopic brain signals. Furthermore, we study channel-agnostic multi-subject decoding methods towards more versatile brain decoding in real-world scenarios. More specifically, we formulate multi-subject decoding as multi-instance learning, and propose a novel channel-agnostic brain decoding model, which can be applied to subjects who have different number of recording channels. Channel-agnostic decoding methods are crucial for developing real-world BCI devices that can be employed over multiple subjects.

1.2 Goals

Towards (1) understanding the relationship between complex brain activities and diverse visual features and (2) developing practical brain encoding and decoding methods for real-world BCI tasks, we study methods for encoding and decoding brain signals using deep learning. In collaboration with Prof. Keisuke Kawasaki (Graduate

School of Medical and Dental Sciences, Niigata University), we prepare a large-scale ECoG dataset by recording high-temporal resolution responses from the macaque inferior temporal cortex while presenting diverse natural images to subjects.

First, to develop a brain encoding method for analyzing the relationship between rich temporal dynamics in neuronal activities and visual features, we conduct experiments to compare frequency-specific ECoG signals and hierarchical visual features extracted from pre-trained CNNs (Chapter 2). Existence of hierarchical visual representations in the ventral visual pathway is well known in neuroscience [22], and, in several previous studies [12, 13, 14, 15], the similarity between the hierarchical organization of the ventral visual pathway and the layer hierarchy of CNNs was suggested by results of comparing human fMRI data and visual features extracted from CNNs. While results from a number of previous studies [23] suggest that neuronal activities in several time-frequency bands have different roles and selectivity in visual processing, the information content of these activities are not well investigated. We use hierarchical visual features of CNNs to analyze ECoG signals in each time-frequency band to investigate whether and how different time-frequency bands are related to diverse visual features.

Second, to develop a brain decoding method that can reconstruct diverse natural images from meso-scopic brain signals, we study deep learning methods for reconstructing natural images from ECoG signals (Chapter 3). A number of studies have proposed reconstruction methods for various types of images, such as binary contrast patterns [24], characters [25], colors [26], faces [27], and natural movies [28]. Following successful applications of deep learning in computer vision, several recent studies used deep learning for reconstructing face images [14] or natural images [17, 18]. Most previous studies on image reconstruction proposed methods for human fMRI data. Although fMRI can cover a broad part of the brain, its hemodynamic responses inherently limit the temporal resolution of recorded signals. In the brain, neuronal activities continuously change in the sub-millisecond level. Therefore, to elucidate the relationship between visual experiences and complex neuronal activities, it is crucial to develop image reconstruction method for high-temporal-resolution electrophysiology.

ical recordings, such as EEG, MEG, and ECoG. Furthermore, accurate decoding of various stimuli from electrophysiological signals is crucial for real-world BCI applications [19, 20]. Towards this end, we study deep learning methods for reconstructing diverse natural images from ECoG signals. Then, we construct and evaluate several deep learning models to show (1) the possibility of natural image reconstruction from meso-scopic brain activities and (2) the importance of utilizing rich temporal dynamics in brain signals in this task.

Third, to develop a brain decoding method that can adapt to multiple subjects, we study channel-channel-agnostic brain decoding (Chapter 4). While multi-subject or transfer learning tasks have been considered in a number of previous studies [29, 30, 31, 32], few studies have investigated channel-agnostic multi-subject decoding methods. In practice, when considering brain decoding across diverse subjects, it is not straight to record brain activities using a technique that has a same number of recording channels. Moreover, if a decoding model is not channel-agnostic, its decoding ability is not applicable to other subjects' data that has a less or greater number of recording channels. Therefore, developing channel-agnostic brain decoding methods is important for applying decoding methods in various practical scenarios, such as multi-subject data analysis, real-world BCI applications, and collaborative BCI tasks [33]. To develop a channel-agnostic brain decoding method, we consider this problem as a multiple instance learning task [34]. In multiple instance learning, inputs are considered as a set of independent instances (bags), and each task is considered as a weakly supervised learning task where only one label is annotated for each input bag. This formulation naturally fits channel-agnostic brain decoding. By formulating channel-agnostic brain decoding as a multiple instance learning task and incorporating recently developed set-based neural network architectures, we develop channel-agnostic brain decoding methods that can decode visual features given ECoG signals from multiple subjects who have different recording channels.

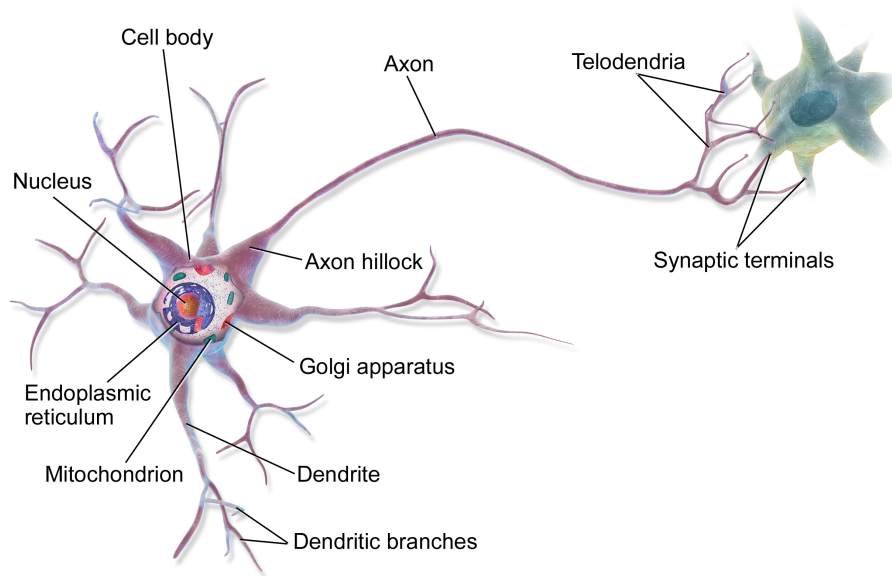


Figure 1-1: An illustrative diagram of a neuron. (Source: Figure by BruceBlaus / CC BY 3.0)

1.3 Recording brain activities

1.3.1 Neuron

Neurons (nerve cells) are electrically-excitabile cells, which are the most important and well-studied subject in the brain. Neurons play a crucial role in detecting patterns and transmitting signals to other cells, by generating electrical signals in response to chemical and other input signals. Typically, a neuron consists of a cell body (called *soma*), *dendrites*, and an *axon* (See Figure 1-1 for illustration). Each neuron receives input signals at dendrites from other neurons, and transmits output signals at the axon. At the farthest tip of the axon's branches, there are terminals, where the neuron transmits output signals across the synapse to other cells. Neurons control ionic flows across their cell membrane, where ions (e.g., sodium: Na^+ , potassium: K^+ , calcium: Ca^{2+} , chloride: Cl^-) move into and out of the cell body through ion channels. These ionic flows change in response to voltage changes and internal/external signals.

The most relevant signal to neurons is the difference of electrical potentials between the internal of neurons and the surrounding, extracellular medium. When a

neuron does not detect any certain signal that the neuron prefers, the electrical potential inside the neuronal cell membrane is around -70 mV relative to the neuron's surrounding bath. In this condition, the neuron is said to be *polarized*. In *hyperpolarization*, the electric potential of the neuronal cell membrane becomes more negative when positively-charged ions flow out of the cell membrane or negatively-charged ions flow into the cell membrane. On the other hand, in *depolarization*, current flowing into the neuronal cell membrane makes the membrane potential less negative or even positive values.

When a neuron is depolarized sufficiently enough to raise the neuronal membrane potential above a certain threshold in a short time interval, the neuron generates an *action potential* (also known as *spike*). Each action potential is an around 100 mV change of the electrical potential across the neuronal cell membrane, and lasts for about 1 ms. The generation of action potentials also depends on the recent history of the neuron's action potentials.

1.3.2 Techniques for recording brain activities

Single-unit activities

In cognitive neuroscience, one of the most widely-used recording technique is *single-unit activities* (SUAs). SUAs measure electrophysiological responses of single neurons using microelectrode-based recording. In recording SUAs, a microelectrode is inserted into the brain to record the extracellular voltage change near the neuronal cell membrane. Typically, the number of action potentials in a defined time window (the *firing rate*) is extracted from original time series of SUAs, and used for further analyses. Rather than the firing rate, we can extract high-frequency time-series by applying band-pass filtering to original data with a bandwidth from 300 to 6000 Hz. These high-frequency signals are called *multi-unit activities* (MUAs), which are thought to be related to summed activities of local neuronal populations. Lower-frequency signals than MUAs (8-200 Hz) are called *local field potentials* (LFPs), which are thought to be related to summed and synchronized activities of local neuronal populations.



Figure 1-2: An example scene of EEG recording. (Source: Photo by Tim Sheerman-Chase / CC BY 2.0)

Although these recording techniques are invaluable for measuring high-resolution activities of single neurons, they are not suitable for recording from healthy human subjects, long-term recording, or mobile brain-computer interface (BCI) systems, because they require inserting microelectrodes into the brain.

Electroencephalography

In contrast to SUA, MUA, and LFP, which measure activities in the brain, electroencephalography (EEG) measures voltage changes with multiple electrodes on the scalp (Figure 1-2). EEG has been used not only for cognitive neuroscience studies, but also for the diagnosing of epilepsy and sleep disorders, BCI devices, among others. While the hardware cost of EEG is significantly lower than other recording techniques, EEG has high temporal resolution (mostly in the level of milliseconds), which makes

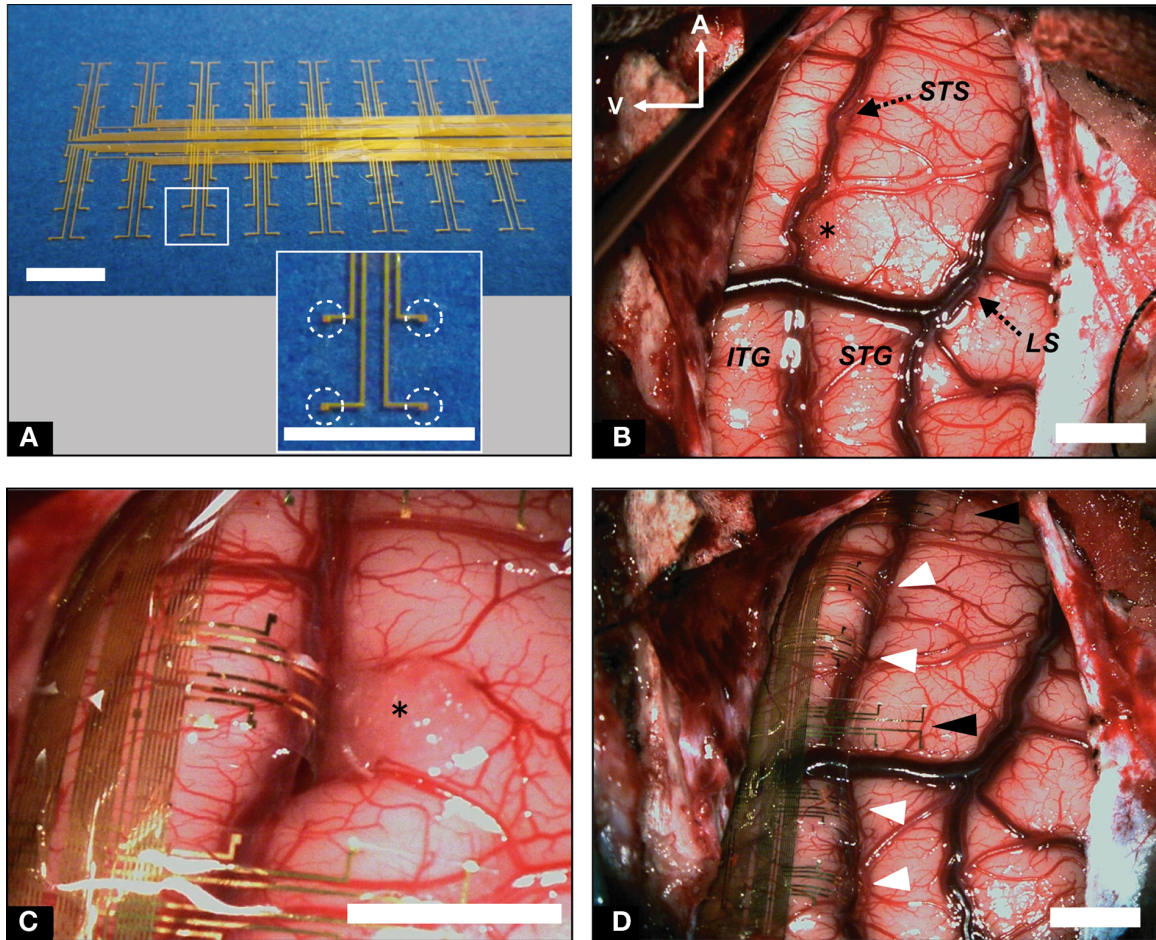


Figure 1-3: An example of ECoG array implantation. (Source: Matsuo *et al.* [1] / CC BY 4.0)

EEG an useful technique for studying complex temporal neuronal responses. EEG is also a non-invasive technique, so applicable for a wide range of subjects, studies, and applications. However, EEG has significantly lower spatial resolution than other recording techniques, because its signals must pass through the skull and the scalp, which attenuates original neuronal activities inside the brain.

Electrocorticography

Electrocorticography (ECoG), or intracranial electroencephalography (iEEG), is an electrophysiological recording technique that measures voltage changes with multiple electrodes on the surface of the brain. ECoG has a similar level of high temporal resolution as EEG, but has better spatial resolution than EEG thanks to its electrode



Figure 1-4: An example scene of MEG recording. (Image source: National Institute of Mental Health, National Institutes of Health, Department of Health and Human Services)

location. As is the case with SUA, MUA, and LFP, ECoG is an invasive technique, because the implantation of ECoG electrodes requires a craniotomy (Figure 1-3). However, ECoG is suitable for long-term recording, and applications in real-world BCI systems is an active area of research and development.

Magnetoencephalography

Magnetoencephalography is a non-invasive recording technique that measures the change of magnetic fields occurred by electric neuronal activities in the brain. Similar

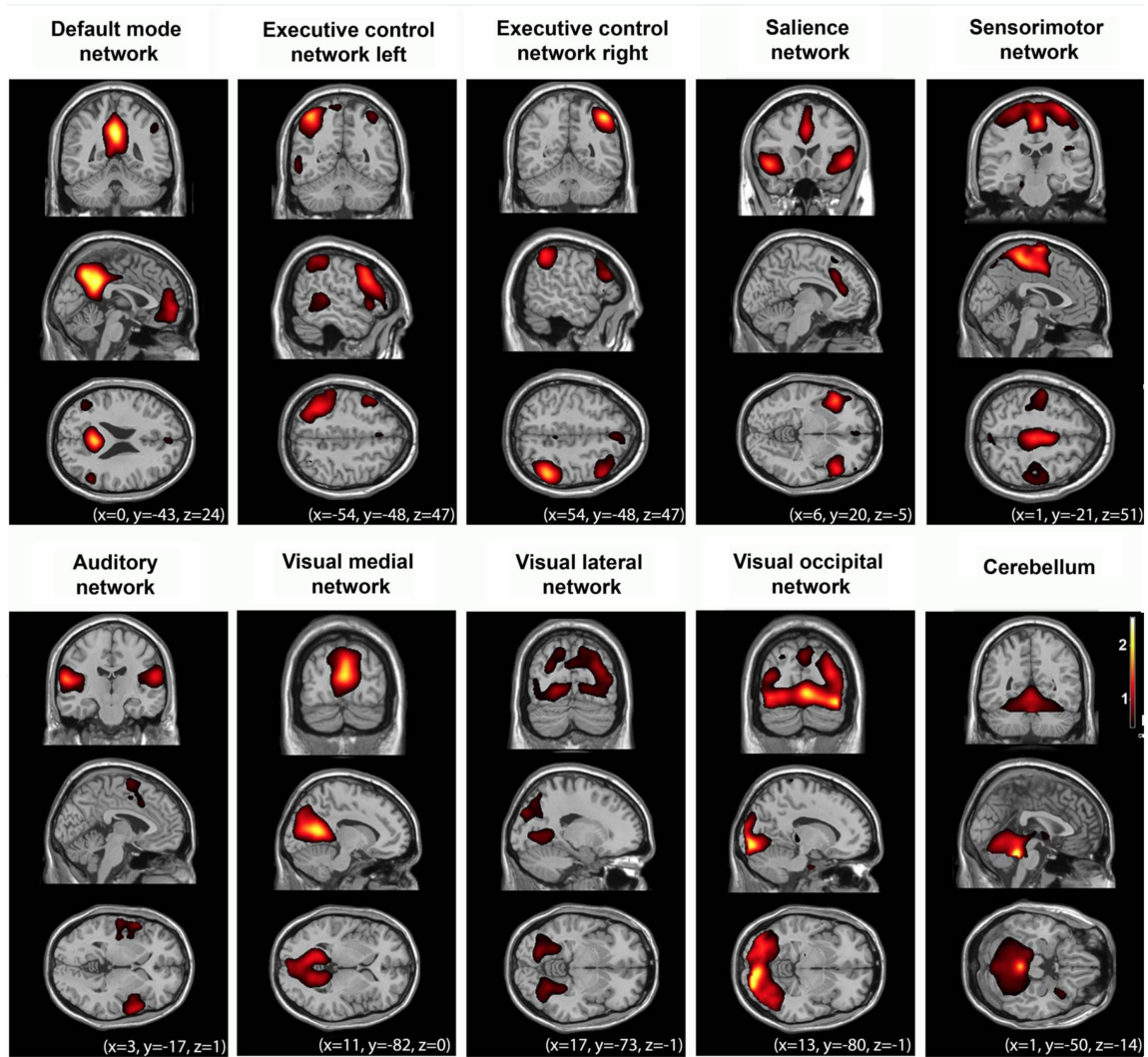


Figure 1-5: An example of fMRI activities. (Source: Lizette *et al.* [2] / CC BY 3.0)

to EEG and ECoG, MEG has high temporal resolution, and the spatial resolution of MEG signals is better than EEG (but slightly worse than ECoG). However, preparing equipment for MEG recording is far more expensive than EEG and ECoG; MEG requires not only specialized equipment but also shielded areas. Moreover, MEG signals are more easily distorted by surrounding signals in the recording environment.

Functional magnetic resonance imaging

Functional magnetic resonance imaging (fMRI) is one the most widely-used recording techniques in cognitive neuroscience. Based on the fact that blood flows in the

Table 1.1: Comparison of recording techniques.

Technique	Signal type	Temporal resolution	Spatial resolution	Invasiveness	Portability
SUA	Electrical	< 1 ms	10-30 μ m	Invasive	Non-portable
EEG	Electrical	50 ms	10 mm	Non-invasive	Portable
ECoG	Electrical	30 ms	1 mm	Invasive	Portable
MEG	Magnetic	50 ms	5 mm	Non-invasive	Non-portable
fMRI	Hemodynamic	1 s	1 mm	Non-invasive	Non-portable

brain and neuronal activities are strongly coupled, fMRI measures activities in a wide range of brain regions using blood oxygen level-dependent (BOLD) contrast imaging. fMRI allows researchers to record activities from the whole brain volume, and its spatial resolution is better than EEG. However, the temporal resolution of fMRI data (typically on the order of seconds) is far worse than other recording techniques. This significantly slow processing time of fMRI limits its use for experiments that study brain processes lasting for more than a few seconds.

Comparison of recording techniques

Considering both research uses and applications (e.g., BCI devices), we can compare the above recording techniques in terms of the signal type, temporal resolution, spatial resolution, invasiveness, and portability. First, on the signal type, techniques that record electrical or magnetic signals are suitable for analysing neuronal activities, because neurons in the brain process and communicate information via local and meso-scopic electric signals. Although hemodynamic measurements of fMRI is known to be correlated with local neuronal activities, they are not direct measurements of neuronal electric communications. Second, on the temporal resolution, techniques with better temporal resolution is better for analysing complex temporal dynamics of neuronal activities and for reading out various information from brain activities

in finer details. Third, on the spatial resolution, although techniques with a better spatial resolution is useful for detecting more local neuronal activities, covering a wider range of brain regions is also important. Fourth, on the invasiveness, SUA is the most invasive technique and not suitable for longitudinal research or applications. While ECoG is also an invasive technique, it can be used for longitudinal cases (e.g., days, months). Fifth, on the portability, SUA, MEG, and fMRI require their specific lab equipment and not portable.

In this thesis, we study ECoG signals recorded in the primate visual cortex. As described above, ECoG has better spatial and temporal resolution than EEG, and it has been a popular recording technique for cognitive/clinical research uses and applications such as real-world BCI devices. We make use of the rich temporal dynamics of ECoG signals for analysing visual selectivity of complex brain signals and for decoding various perceptual information.

1.4 Existing methods for analysing brain activities

1.4.1 Classical statistical hypothesis testing methods

In cognitive neuroscience, researchers have traditionally studied brain activities for synthetic stimuli that differ along a specific attribute of interest, such as the spatial frequency, orientation, color, the existence of face, and object classes. For example, let us say a researcher is to study whether a neuron of interest significantly changes its firing rate between face and non-face images. They record responses of the neuron for both sets of images. After recording, they compute cross-trial statistics over multiple trials independently for face and non-face images. Then, the hypothesis is tested by statistical testing methods, such as chi-squared test, Student's t-test, and analysis of variance (ANOVA).

While this framework is simple to use and its results are easy to interpret, it has several shortcomings [35]. First, using a tightly-controlled stimulus set limits the research target, because researchers needs to decide which specific attribute of

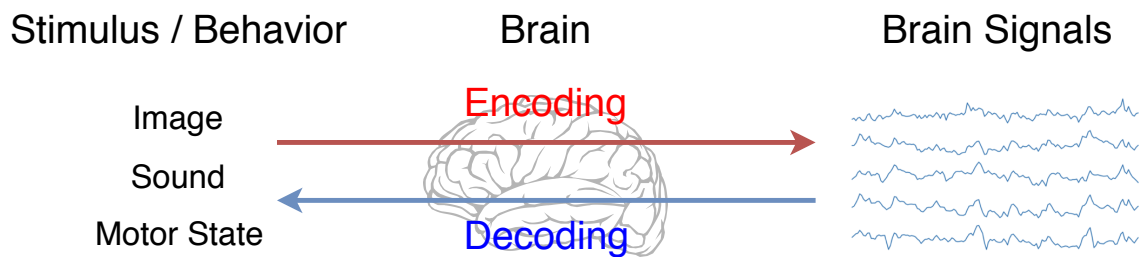


Figure 1-6: Brain encoding and decoding.

stimuli to investigate, and the attribute of interest must be clear. This problem makes conducting experiments with this framework time-consuming when researchers want to investigate a diverse set of stimuli. Second, this framework can lead to biased, artificial stimulus designs by manipulating stimuli for a specific hypothesis, moving stimuli away from those encountered in the real world. This problem appears most notably when researchers want to investigate regions where neurons are hypothesized to have strong selectivity to complex perceptual information, such as natural scenes, abstract concepts, and dynamic motions. Third, this framework requires researchers to fix their research hypothesis before stimulus preparation and brain recording. Since these hypotheses are often based on simplified experiments and artificial stimuli, the results are need easily transferred to more realistic, complex conditions.

1.4.2 Encoding and decoding methods

An alternative framework is multi-variate, predictive modeling between brain activities and stimuli [3, 4, 35]. In this framework, researchers can construct a diverse family of models using statistical or machine learning models, such as classification and regression models. As the result of the increase of available computational resources and the rapid progress of the machine learning literature, a lot of recent studies in cognitive and computational neuroscience have employed more flexible, learning-based methods: *brain encoding* and *decoding* (Figure 1-6).

Brain encoding considers the relationship between brain activities and stimuli by predicting brain activities from stimuli or their features. In contrast, brain decoding models predict stimuli or their features from brain activities. Brain encoding models

can be used to test specific computational models of the brain, by evaluating how well each encoding model predicts actual brain responses. On the other hand, brain decoding models help researchers investigate what kind of visual features are related to brain activities, by evaluating how well each decoding model predicts specific visual features from brain activities.

Encoding models

If we record brain activities while changing stimuli to the subject, we can construct an encoding model that considers the probabilistic model of brain activities given stimulus: $p(y|x)$, where y is recorded brain activities (output) and x is a stimulus (input). One way to construct an encoding model is to approximate the true probabilistic model with a Gaussian: $p(y|x) \approx \mathcal{N}(y|\mu(x; \theta), \sigma^2)$, where $\mu(x; \theta)$ is a parametric function that predicts the first-order moment of the Gaussian distribution given x , and σ is the standard deviation. Then, given training data, $\mathcal{D}_{train} = \{(x_1, y_1), \dots, (x_N, y_N)\}$, we can estimate the parameter θ of the encoding model based on maximum likelihood estimation (MLE):

$$\begin{aligned}
 \theta^* &= \arg \max_{\theta} \prod_{i=1}^N p_{\theta}(y|x) \\
 &= \arg \max_{\theta} \prod_{i=1}^N \mathcal{N}(y|\mu(x; \theta), \sigma) \\
 &= \arg \max_{\theta} \sum_{i=1}^N \log \mathcal{N}(y|\mu(x; \theta), \sigma) \\
 &= \arg \max_{\theta} \sum_{i=1}^N -\frac{1}{2\sigma^2}(y_i - \mu(x_i; \theta)) - \log \sigma + \text{const.} \\
 &= \arg \min_{\theta} \sum_{i=1}^N \frac{1}{2\sigma^2}(y_i - \mu(x_i; \theta)) + \log \sigma,
 \end{aligned}$$

which corresponds to the minimization of the sum-of-squares error function with a penalty term.

Encoding models can take a variety of inputs, such as raw stimuli (e.g., images,

audio, motor states), high-level features, statistics, and even discrete attributes (e.g., existence or absence of face in the image). This is in contrast to the classic, hypothesis testing framework, where a set of complementary hypotheses must be defined for the experiment. Although the hypothesis testing framework is useful for investigating the property of neuronal activities in the lower-level sensory cortex, where most neurons have strong selectivity to simple perceptual patterns, it is not straightforward to investigate the property of complex spatiotemporal activities in mid- or higher-level sensory cortices.

Decoding models

In the opposite direction to encoding models, we can construct a decoding model that predicts a category from brain activities: $p(\mathcal{C}|y)$, where $\mathcal{C} = 1, \dots, K$ is the category (output) and y is recorded brain activities (input). We can construct a decoder model by assuming that the true probabilistic model can be approximated with a parameterized softmax function:

$$p(\mathcal{C}|y) \approx \frac{\exp(f^{(\mathcal{C})}(y; \theta))}{\sum_{k=1}^K \exp(f^{(k)}(y; \theta))}.$$

In the similar way to encoding models, we can estimate the parameter θ of the decoding model as:

$$\begin{aligned} \theta^* &= \arg \max_{\theta} \prod_{i=1}^N p_{\theta}(\mathcal{C}_i | y_i) \\ &= \arg \max_{\theta} \prod_{i=1}^N \frac{\exp(f^{(\mathcal{C}_i)}(y_i; \theta))}{\sum_{k=1}^K \exp(f^{(k)}(y_i; \theta))} \\ &= \arg \max_{\theta} \sum_{i=1}^N \log \frac{\exp(f^{(\mathcal{C}_i)}(y_i; \theta))}{\sum_{k=1}^K \exp(f^{(k)}(y_i; \theta))} \\ &= \arg \max_{\theta} \sum_{i=1}^N f^{(\mathcal{C}_i)}(y_i; \theta) - \text{LSE}_k f^{(k)}(y_i; \theta), \end{aligned}$$

where LSE is the logsumexp function

Depend on the type of output features, decoding models are divided into two

classes: classification and regression models. In classification models, as described in the above example, an input sample of brain activities is identified as belonging to one of a pre-defined set of possible event classes (e.g., face or non-face, several visual objects, motion directions). The output space of classification models is a discrete space and must be defined before training the model. Therefore, we cannot apply classification models to a dataset that has novel event classes.

On the other hand, regression models predict continuous features from brain activities. The target features can be raw stimuli (e.g., image, audio, motion state) or high-level features/statistics (e.g., image statistics, spectrogram). Regression models for raw stimuli are sometimes called *reconstruction* models.

Benefits of encoding and decoding models

Compared with the classic statistical testing framework, the predictive modeling framework (encoding, decoding) has several key benefits [35]. First, the classical statistical testing framework compares the cross-trial average of brain activities for each hypothesis. The statement on the statistical significance is based on the error of the point estimates, such as the standard error of the averaged responses. On the other hand, in the predictive modeling framework, the model is first trained (i.e., parameter estimation) on the training dataset, and then the model's generalization ability (prediction performance) is evaluated on the independent test dataset. Thus, the predictive modeling framework can measure *how well* each encoding/decoding model predicts the target, while the classical statistical testing framework only measure whether the null hypothesis would be rejected or not with a certain significance threshold. This property of predictive models is useful when we compare multiple encoding or decoding models for the same brain activity dataset; The analysis can show us what kind of inputs or models are effective for predicting the target features.

Second, the predictive modeling framework allows us to explore a wide range of models and brain regions for investigating research hypotheses on the representation/importance of different model properties and brain regions. For example, we can explore multiple levels of visual complexity by comparing the performance of

encoding models for each level of complexity. We can also compare the performance of face-or-non-face classification by comparing classification models for each brain region of interest. This is possible because encoding and decoding models are flexible enough to take a variety of input/output feature types.

Third, because encoding/decoding models can take multi-variate inputs and outputs, successful models can capture internal structures and variable relationships that may not be investigated using the univariate, statistical testing framework. For example, classification models for brain signals that have the space (recording channels) and time (recording latency) domains, we can analyze what kind of spatiotemporal structures are key to successful classification using trained models.

Fourth, the statistical testing framework typically requires separation or down-sampling of the spatial and temporal dimension of brain activities. For example, if recorded brain activities consist of 500 time steps (e.g., ms) for each trial, we need to separate each time step or take the downsampled activity over the time steps. This is problematic because neuronal activities change in the level of millisecond. The flexibility of encoding and decoding models allows us to model complex spatiotemporal structures and relationships from data.

1.5 Deep Learning

1.5.1 Neural networks and deep learning

Before deep learning models were well established, achieving human-level visual object recognition with machine learning algorithms had been thought quite hard. However, current state-of-the-art deep learning models have achieved human-level or even greater performance on some computer vision tasks such as object category recognition, object detection and semantic segmentation. The success was mainly lead by efficient training methods, large labeled dataset for supervised learning, open-sourced deep learning libraries, and participation of a huge number of researchers from diverse research fields.

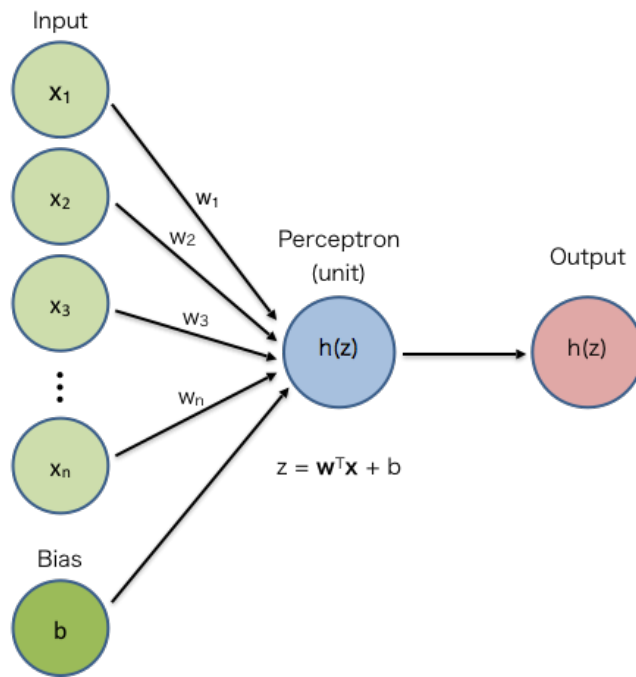


Figure 1-7: Computation of a perceptron (unit)

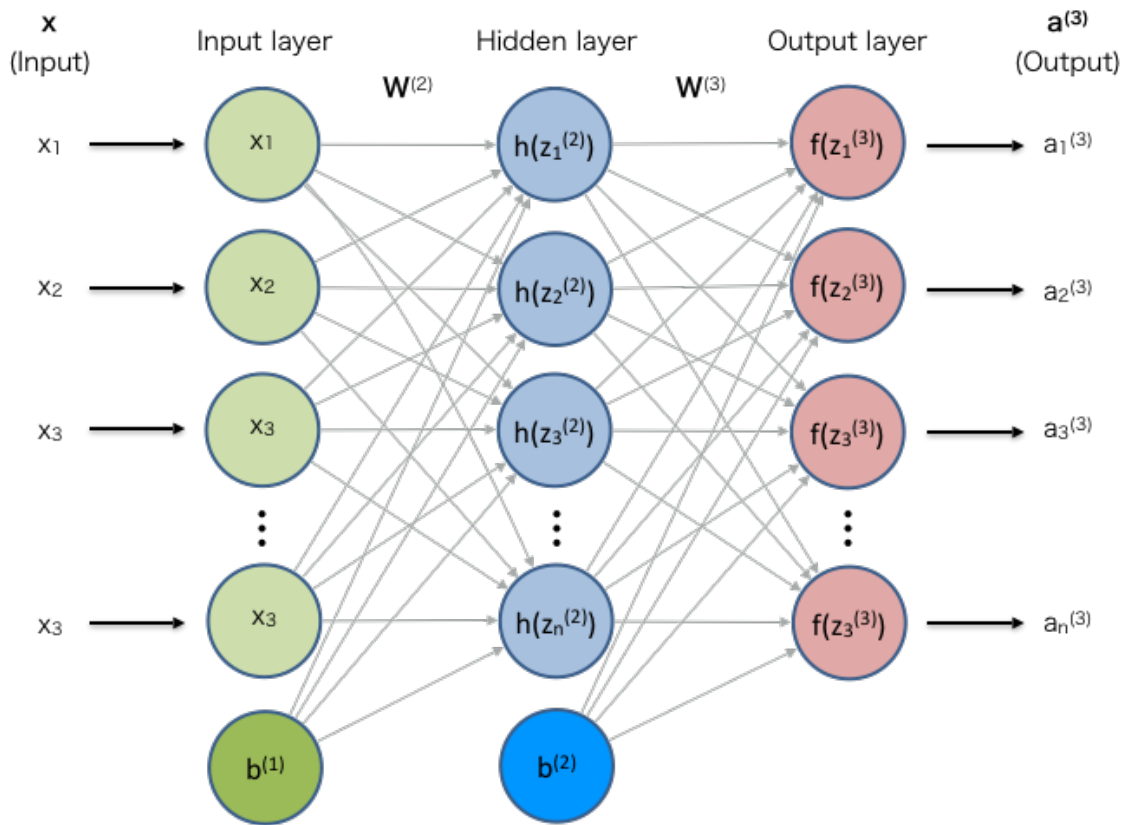


Figure 1-8: Multilayer perceptrons (MLPs)

1.5.2 Perceptrons and Multilayer Perceptrons

Perceptron (often simply called *unit*, Figure 1-7) is a simplified computational model of biological neurons in the brain. It receives a number of inputs, computes the weighted sum of them, and outputs a value computed by a specified activation function. The most classical activation function is sigmoid function

$$h_{\text{sig}}(x) = \frac{1}{1 + e^{-x}}.$$

Recent deep learning models typically use rectified linear unit (ReLU)

$$h_{\text{relu}}(x) = \max(0, x)$$

for avoiding several learning problems.

Multilayer perceptrons (MLPs, Figure 1-8), also known as feedforward neural networks, consist of stacked layers of perceptrons. MLPs receive an input, repeat transforming it in the hidden layers, and finally return an output. Each perceptron in MLPs computes outputs by summation of its weighted inputs followed by non-linear activation functions. Although the computations are relatively simple, MLPs can learn powerful non-linear transformations. In fact, with enough number of perceptrons in the hidden layers, they can represent arbitrarily complex but smooth functions and can be a universal approximator.

Roughly speaking, all the deep learning models are just derivations of MLPs with several modifications on the numbers of layers and units, architecture, and connections between layers/units.

1.5.3 Deep Learning

Although neural networks were proved to have their huge representational capacity, most researchers/practitioners used conventional, non-deep-learning models for various pattern recognition problems. This is because neural networks are hard to train

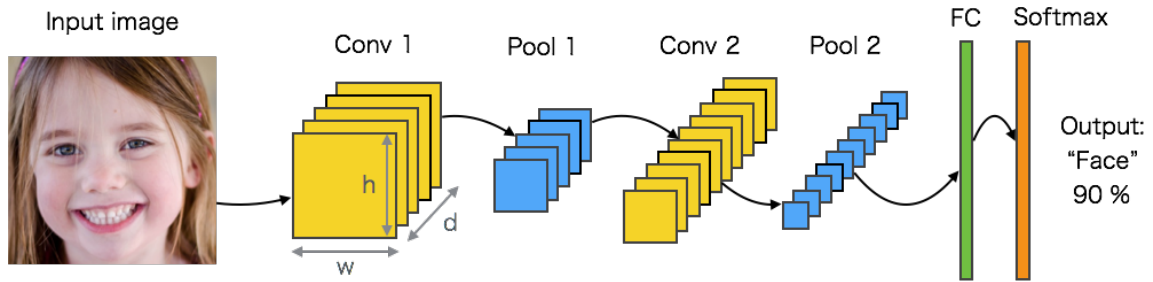


Figure 1-9: Basic architecture of CNNs

due to 1) the over-fitting problem, 2) lack of enough computational resources for training deep neural networks with a huge number of parameters, and 3) the lack of efficient architectures and training methods. Conventional models are relatively simpler than neural networks, however, they require careful engineering and considerable domain-specific knowledge for efficiently designing features that are used as inputs for classifiers/regressors. For example, for visual object recognition, conventional models typically use hand-crafted/designed feature-extraction methods such as scale-invariant feature transform (SIFT) [36], speeded up robust features (SURF) [37], and others.

In the ImageNet Large Scale Visual Recognition Competition (ILSVRC) 2012, a deep CNN model, so-called *AlexNet* [8], outperformed conventional models with more than 10-percent margin on the runner-up for 1000-category image classification task. This result drastically changed the view of researchers/practitioners. Since the success of AlexNet, various deep learning models and their applications have been proposed, and these models have been the state-of-the-art for many problems in computer vision, natural language processing, and other domains.

1.5.4 Convolutional Neural Networks

The basic architecture of CNNs is structured as a series of specific computation stages (Figure 1-9). The first stage typically repeats convolution and pooling layers.

Convolution layers (Figure 1-10) are used for the detection of specific feature patterns, and units in these layers are organized as features maps. A unit in a feature

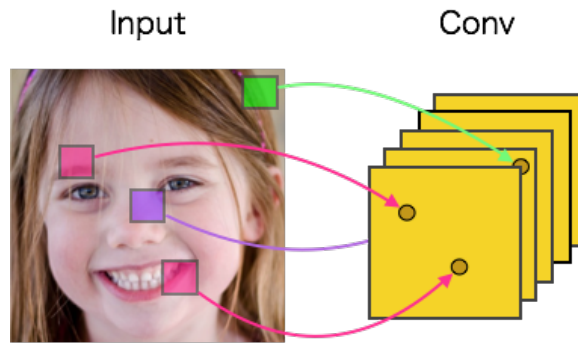


Figure 1-10: Computation of a convolutional layer

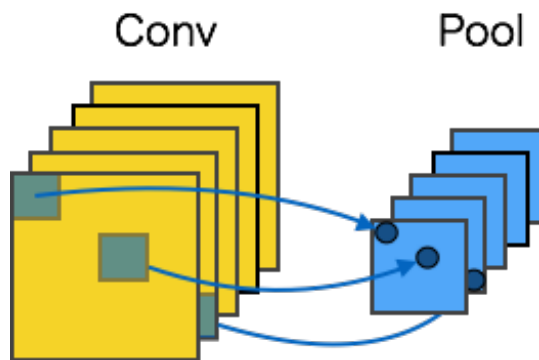


Figure 1-11: Computation of a pooling layer

map is related to some units in a specific region of the previous layer, called receptive fields. Units in convolution layers receive a volumetric input with size $\mathbb{R}^{w \times h \times d}$, where w , h and d are the width, height and depth (number of channels) of the input, respectively. The units compute the weighted sum of the input, and then output activations computed from activation functions.

Pooling layers (Figure 1-11) receive activations from the previous convolution layers. These layers are used for reducing the width and height of the activations by sub-sampling. Similar to convolution layers, units in these layers also have specific receptive fields, and they typically take maximum values in the regions for each channel.

The second stage of CNNs consists of fully-connected layers. Units in a fully-connected layer don't have partial receptive fields as convolution/pooling layers. They receive inputs from all the units in the previous layer. The number of units in the final fully-connected layer is defined as the number of object categories for a specific

classification task.

There are several architectures of CNNs commonly used in the literature. The first successful and most classical model is LeNet [38] developed by Yann LeCun in 1990's. AlexNet developed by Krizhevsky et al. [8] extended LeNet and made CNNs as the standard for general object recognition by winning the ILSVRC 2012. AlexNet is, roughly speaking, a deeper and bigger model of LeNet. VGGNet [39], the runner-up in ILSVRC 2014, was developed by Simoyan et al. The success of VGGNet showed the effectiveness of the depth, the number of CNN layers. Although, VGGNet is relatively expensive to train by its huge number of parameters, the pre-trained model has been used for various problems such as other object recognition tasks, object detection, and so on. ResNet [40] developed by He et al. was the winner of the ILSVRC 2015. ResNet employed residual learning module for making CNNs deeper while avoiding inefficient training. Currently, ResNet is thought as the state-of-the-art CNN model and the common choice for using CNNs in practical applications.

Several elements or ideas of CNNs were inspired by previous neurophysiological findings on the visual cortex. Convolution and pooling operations used in CNNs were directly inspired by the classic notions of simple and complex cells proposed by two Noble Prize-winning neuroscientist: David H. Hubel and Torsten Wiesel [41, 42]. They showed that, in the visual cortex, there are *simple cells* that have strong selectivity for relatively specific orientations in smaller receptive fields, and *complex cells* that have relatively loose and spatially-invariant selectivity in larger receptive fields. The hierarchical architecture of CNNs was also strongly influenced by the hierarchy of the primate ventral pathway. With these notions, Fukushima first proposed a basic idea of CNNs, so-called Neocognitron [43]. However, Neocognitron doesn't have an end-to-end learning algorithm for classification tasks. Later on, LeCun et al. applied backpropagation for supervised learning and proposed the first classical model of CNNs for hand-written digit classification [38].

1.5.5 Deep Learning for Brain Activity Analysis

While most older studies used simple statistical or machine learning methods such as linear models and support-vector machines (SVMs)[5], a number of recent studies used deep learning [6, 7] for brain encoding or decoding tasks, following successful applications of deep learning to various tasks in computer vision [8, 9, 10, 11].

In brain encoding, several recent studies [12, 13, 14] analyzed brain activities using convolutional neural networks (CNNs) that were pre-trained on a large-scale visual object classification task. They compared visually-evoked brain activities and visual representations in CNNs, and found a similarity between the hierarchical organization of the primate visual cortex and the layer hierarchy of CNNs.

In brain decoding, [15] compared the performance of decoding CNN features from human functional magnetic resonance imaging (fMRI) data. They also observed a similarity between the hierarchical organization of the human visual cortex and the layer hierarchy of CNNs. Several other studies [16, 17, 18] used deep learning for reconstructing presented images from human brain activities.

1.6 Thesis structure

This thesis is organized as follows. In Chapter 2, we review the literature. We first introduce the fundamentals of biological visual object recognition, then explain the fundamentals of visual object recognition using CNNs, and finally review related work on the relationship between the brain and CNNs. In Chapter 3, we first describe the detail of our experimental setup and the results. In Chapter 4, we discuss our results with relating the literature, and conclude this thesis. And Chapter 5 concludes this thesis.

Chapter 2

Encoding and Analyzing Frequency-Specific ECoG Signals Using Hierarchical Visual Features

2.1 Introduction

One important goal in cognitive neuroscience is to understand the relationship between brain activities and sensory information. From single-neuron spikes to mesoscopic brain signals, brain activities can be measured in a wide range of scales of the brain. In the literature, most previous studies have investigated the change of the neuronal firing rate to various synthetic or natural stimuli. However, there has been increasing evidence that indicates the importance of neuronal oscillatory activities in cognition [44, 45].

Neuronal oscillatory activities can be measured by local field potentials (LFPs), electrocorticography (ECoG), electroencephalography (EEG), and magnetoencephalography (MEG). It is thought that raw brain signals contain aggregated activities in several time-frequency bands, such as delta, theta, alpha, beta, and gamma bands [46]. In the visual cortex, several previous studies suggest that brain signals in specific time-frequency bands have stronger selectivity to visual stimuli [47, 48, 49, 50],

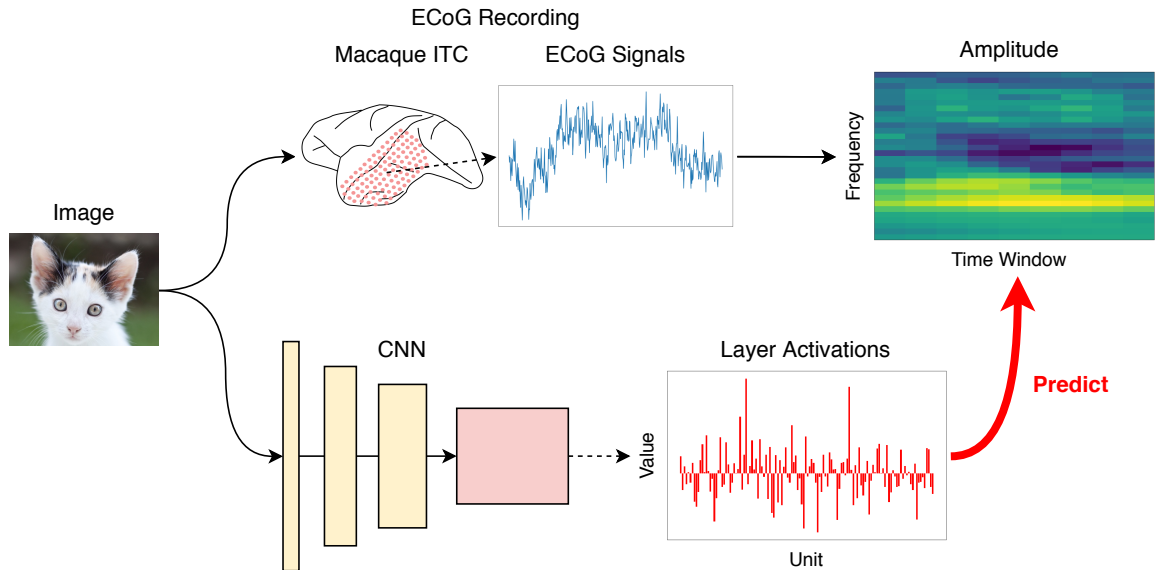


Figure 2-1: Encoding frequency-specific ECoG activities from CNN features. We trained encoding models that predict frequency-specific ECoG activities given visual features extracted from a pretrained CNN. We recorded ECoG signals from the macaque ITC while presenting natural images, and extracted frequency-specific amplitude using time-frequency decomposition. Using the same image set, we extracted visual features from each convolution and fully-connected (FC) layer of a pretrained CNN. For convolutional layers, which have three dimensions (width, height, channels), we downsampled features over the width and height (global average pooling). After feature extraction, we trained ridge regression models that predict ECoG amplitude at a specific site, frequency, and time window from CNN features at a specific layer.

and that low- and high-frequency bands are related to distinct visual information [51, 23, 52]. Furthermore, different frequency bands are thought to play complementary roles in inter-areal feedforward and feedback processing [53, 54, 55, 56]. However, few studies have investigated the detailed visual selectivity of each frequency band, especially in the primate inferior temporal cortex (ITC), which is the highest-level area in the ventral visual pathway.

In this work, to investigate how visual selectivity differs across frequency bands, we analyze frequency-specific activities in ECoG signals recorded from the macaque ITC using rich hierarchical visual representations extracted from a deep convolutional neural network (CNN). CNNs have achieved state-of-the-art performance on various computer vision tasks [8, 9, 10, 11]. Furthermore, CNNs enable us to extract opti-

mized hierarchical visual features. Previous studies indicate that units in lower, mid, and higher CNN layers represent lower-, mid-, and higher-level visual features, respectively [57]. Several recent studies in neuroscience used visual features in CNNs for analyzing the similarity of the layer hierarchy of CNNs and the anatomical hierarchy of the ventral visual pathway [12, 14]. In our experiments, we trained and evaluated encoding models that predict frequency-specific ECoG activities from visual features extracted at a specific layer of a pretrained CNN (Figure 2-1). We found that two specific frequency bands, theta (around 5 Hz) and gamma (around 20-25 Hz) bands, were better predicted from CNN features than the other bands. Furthermore, these two bands were better predicted from higher and lower CNN layers, respectively. Our visualization analysis using CNN-based encoding models qualitatively showed that theta- and gamma-band encoding models had selectivity to higher- and lower-level visual features, respectively.

2.2 Background

Primates such as humans and monkeys can easily recognize objects with their vision systems, even under active motions and complex conditions, such as viewpoint/scale variations, occlusions, deformations, and illuminations changes.

Thanks to previous neuroanatomical studies of non-primates, we know relatively much about the anatomical organization of the primate visual system (Figure 2-2). The system can be grouped into five groups: 1) retinal sensors, the lateral geniculate nucleus and the thalamus, 2) primary visual cortex, 3) the ventral "what" pathway, 4) the dorsal "where" pathway, and 5) higher-level regions. When we see, photons first arrive at the retina, and the light signals are transformed into neural electrical signals by the photoreceptors, Horizontal, bipolar and amacrine cells receive the signals followed by retinal ganglion cells that send signals to the lateral geniculate nucleus (LGN). The thalamus receives signals from LGN, and then sends to primary visual cortex (V1). Neurons in V1 have strong selectivity to specific low-level visual features such as color, contrast, orientation, and spatial frequency. Neural signals

containing these low-level visual features are then sent to two complementary pathways: the ventral "what/object" pathway and the dorsal "where/action" pathway. The ventral pathway also has hierarchical structure in itself, and neurons in the pathway are thought to be involved in the recognition of various types of objects. On the other hand, neurons in the dorsal pathway are thought to be involved in the spatial localization of objects within their environments and in guiding action towards those objects. These two pathways have multiple interactions each other, so they are not independent. High-level visual features processed in these two pathways are finally sent to the medial temporal lobe (MTL) and to the prefrontal cortex (PFC).

Rich literature in neurology and neurophysiology has shown that such fast, high-level, and robust visual object recognition is achieved by a special visual system in the brain: *the ventral pathway* [22]. The ventral pathway is a series of hierarchically-organized visual areas spanning from the back to the temporal side of the brain. It has been known that, after we see an object, low-level visual features are first processed in lower visual areas such as V1 and V2, they are then transferred to intermediate V4 area, and finally reach at the highest visual area, the inferior temporal cortex. Response properties of single-neuron firing rates in the ventral pathway have been well studied, and we know neurons in lower and higher visual cortices have strong selectivity for lower- and higher-level visual information, respectively. However, it has been little known how neurons in the ventral pathway represent diverse visual information in spatial and temporal domains, because the problem requires us to establish biologically plausible computational models of visual processing in the brain.

In the ventral stream, especially the inferior temporal cortex (ITC) is thought to have the most fundamental role in visual object recognition. This is supported by many previous neurological and neurophysiological studies [58, 59]. In neurological studies, it is repeatedly reported that lesions in the ITC of macaque monkeys produce severe deficits in visual object recognition. For the human brain, patients with prosopagnosia have normal sight but cannot recognize faces from visual stimuli [60, 61]. These subjects can recognize many other object types. They typically show lesions in the ITC. There are other evidences of human patients who show severe

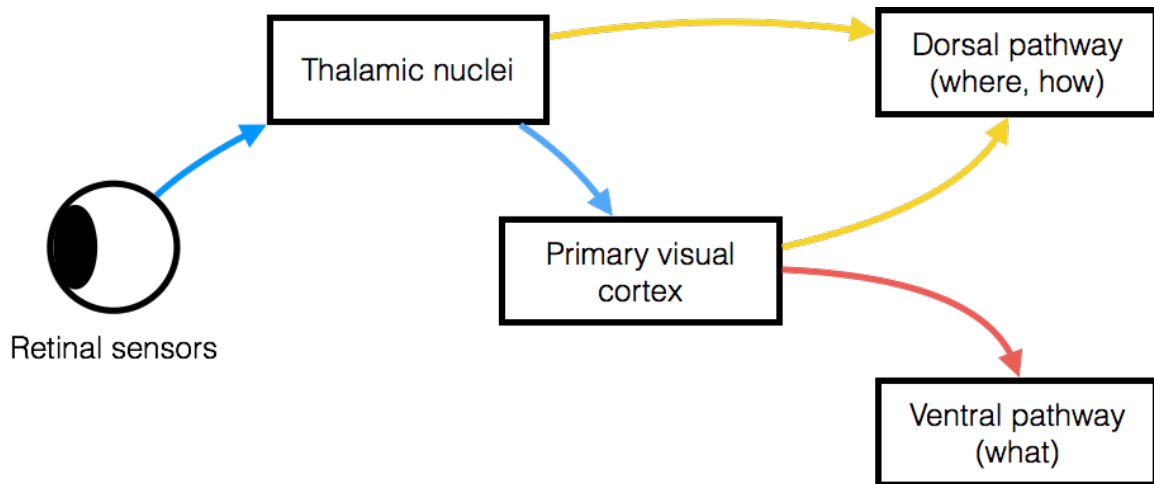


Figure 2-2: Visual processing in the primate visual cortex

deficits for other categories beyond faces [62, 63]. Electrophysiological recording of single neurons in the macaque ITC has revealed that most of the neurons have strong selectivity to specific types of complex objects such as face, body part, building, and tool. Moreover, they maintain strong robustness for various visual condition changes. They typically show strong invariance to scale and location changes, eye movements, shape, rotation and others. These studies clearly indicate the fundamental role of the ITC in visual object recognition [64, 65, 66].

While much has been known about how anatomical regions are hierarchically organized in the ventral pathway and what visual features make neural spiking rate increase in the primary visual cortex and the ITC, the whole picture of biological visual object recognition is almost unresolved [22]. That is, we don't have how neurons, neural populations and anatomical regions function together for rapidly recognizing objects. This requires a computational model which sufficiently explains the spatial and temporal variability of neural activities and performs animal-level visual object recognition.

2.3 Related work

In computer vision, deep learning models, in particular convolutional neural networks (CNNs), have achieved impressive results on various tasks, such as general visual ob-

ject recognition [67, 40], semantic segmentation [10], and object detection [68, 69]. The success of CNNs on computer vision tasks is intriguing for cognitive neuroscience, because (1) the fundamental design of CNNs is inspired from anatomical and physiological findings about biological visual object recognition in the brain [43, 38], and (2) the performance of state-of-the-art CNN architectures on popular tasks (e.g., object classification) has been closed to the human performance. Motivated by the success of CNNs, there have been an increasing number of cognitive neuroscience studies that investigate brain activities using CNNs.

2.3.1 Hierarchical relationships in the primate visual cortex and CNNs

The ventral pathway in the primate visual cortex has a hierarchical structure ranging from lower visual cortex (V1, V2) to higher visual cortex (V4, ITC) [22]. Previous neurophysiological studies have revealed the selectivity of neurons in the ventral pathway by measuring spiking activities after stimulus presentation. Neurons in lower visual cortex show strong selectivity for low-level visual features such as bars with specific orientation, color and contrast. Neurons in V4 show strong selectivity for mid-level visual features such as specific shapes and surface properties. Neurons in the ITC show strong selectivity for complex, high-level visual features such as object categories.

Interestingly, similar properties were observed for units in CNNs trained for visual object recognition. Zeiler *et al.* [57] proposed a visualization method for investigating the selectivity of units in CNNs. Employing deconvolutional networks [70], they investigated what kind of visual features each unit in CNNs represents. Similar to neurons in the ventral pathway, units in lower layers showed strong selectivity for low-level visual features such as color, contrast, orientation, and those in higher layers showed strong selectivity for complex, high-level visual features such as specific textures, shapes, animal/human faces, etc.

2.3.2 Representation similarity between the primate visual cortex and CNNs

Recent studies compared the representation similarity between the primate visual cortex and CNNs. That is, how much we can predict neural responses in the visual cortex from representations in CNNs, and whether and how much they have similarity for various images.

Although it has been known that neurons in higher visual cortex (V4, IT) have strong selectivity for mid-/high-level visual features, there had not been much successful computational models that can predict their responses for various images. That is because V4 and IT neurons have robust and invariant representations for complex visual object recognition tasks, and, before the establishment of CNNs, there had not been any successful machine learning/computer vision model for such tasks. Yamins *et al.* [12] compared various computational models on the prediction of IT neuron responses. For a real-world, complex image set, they collected features from various models such as SIFT, V1-like Gabor-based model [71], V2-like conjunction-of-Gabors model [72], HMAX [73, 74], and CNN models with different hyper-parameters. They then compared the relationship between the explained variance of IT-neuron responses from these model-specific features and each model’s classification performance. Their results showed the strong correlation between models’ classification performance and the IT-neuron-predictivity, and performance-optimized CNNs was the best model for both classification and IT-neuron-predictivity.

Yamins *et al.* [12] also investigated the problem using a method called representation similarity analysis (RSA [67]). RSA compares the similarity between the representation dissimilarity matrix (RDM) of subjects or models. RDM is computed as the magnitude of negative correlation between stimulus-specific responses/features of a subject/model. Thus, it enables us to directly compare the representation similarity between different subjects/models that may have different measurements or the number of feature dimension. Their results showed that performance-optimized CNNs achieved better representation similarities of IT-neuron responses than other

models. That is, the behavior/trend of representations of the CNNs and IT neurons is strongly correlated, although the CNNs were trained only for classification of object categories.

2.3.3 Spatial relationship: anatomical hierarchy of the primate visual cortex and CNNs

In addition to the representation similarity, there have been several studies that indicate there exists a spatial similarity between the ventral pathway and CNNs. That is, the hierarchical organization of the ventral pathway and the layer hierarchy of CNNs may be similar.

Güçlü *et al.* [14] compared the prediction accuracy of BOLD (blood-oxygen-level dependent) signals recorded in the human ventral pathway. They conducted the experiment from 8 layers of a CNN model. The prediction models not only successfully predicted neural responses from lower to higher visual areas, but also showed a hierarchical similarity along the ventral pathway. Comparing the prediction accuracy between the measured locations in the ventral pathway and the layer hierarchy of CNNs, responses in lower and higher visual cortex were respectively better predicted from lower and higher layers in CNNs.

Similar results were also observed in [12]. They compared the prediction accuracy of single-unit neural responses in V4 and the ITC. They observed that the highest layer in CNNs best predicted the neural responses in the ITC, and that the intermediate layers in CNNs better predicted the neural responses in V4 than the lowest and highest layers.

Cichy *et al.* [75] analyzed the representation similarity between BOLD activities in human ventral and dorsal pathways and hierarchical representations in CNNs. Using functional magnetic resonant imaging (fMRI), they recorded BOLD activities from diverse regions in the ventral and dorsal pathways. Similar to the above studies, they found a spatial, hierarchical similarity between the ventral pathway and CNNs. Moreover, they obtained similar results for the dorsal pathway, while previ-

ous neurophysiological studies suggest that the regions are used for motion or location perception. The CNN models used for their experiments were trained for object categorization. Therefore, their results suggest that the roles of the dorsal pathway should include not only spatial cognition but also general object recognition.

2.3.4 Temporal relationship: latency of brain activities and CNNs

In the temporal domain, Cichy *et al.* [75] investigated the relationship between the time course of visual processing in the visual cortex and the layer hierarchy of CNNs. They recorded millisecond-resolved magnetoencephalography (MEG) signals from the human visual cortex. Then, they compared the representation similarity between MEG signals at a specific time window (-100 to +1000 ms with respect to image onset) and a layer in CNNs. Although the trend was modest, their results indicate there may exist a similarity between the layer hierarchy of CNNs and the peak latency of layer-specific representations.

2.3.5 Different time-frequency bands for distinct visual information

It has been hypothesized that, in the brain, neural activities in different time-frequency bands have complementary information and they couple/cooperate together [76].

While there has been several work investigating the relationship between the visual cortex and CNNs, they focused on either the spatial or temporal aspect of neural activities in the brain. Neural activities in the brain may not be a combination of independent modules in space and time. Rather, information processing in the brain can simultaneously emerge both in space and time [77, 44, 78]. Moreover, even only in the temporal domain, previous studies on neural oscillatory activities suggest that there exist several time-frequency bands where different neural representation and information processing may occur [46, 79, 80, 81, 54].

Jasobs *et al.* [49] investigated ECoG signals from human patients who study lists

of letters in a working memory task. Using a decoding/classification approach, they observed that 1) gamma band (25-128 Hz) activities in the occipital regions were informative for the task, 2) the gamma band activities may be related to not specific types of letter but lower-level visual features of letters ,and that 3) the gamma band activities were strongly coupled to the phase of theta (4-8 Hz) band activities.

There have also been several studies showing different frequency bands may have complementary visual information. Belitski *et al.* [51] recorded local field potentials (LFPs) from the primary visual cortex of anesthetized macaques with presenting color movies. They analyzed mutual information between visual features in the movies and the power of the LFPs. Comparing the time-frequencies, they observed that the most informative frequency bands are 1-8 and 60-100 Hz. Moreover, their results of mutual information and correlations between frequency-specific signals suggest that these low and high frequency bands have complementary information. These studies indicate that there may be low and high frequency bands that contain complementary information.

2.3.6 Different time-frequency bands for complementary roles in cortical information processing

In addition to the possibility of information contents of different frequency bands, there have been several studies indicating that different frequency bands have complementary roles in visual processing.

van Kerkoerle *et al.* [82] recorded LFPs and multi-unit activities (MUAs) from the macaque visual cortex. They investigated how low-frequency alpha band and high-frequency gamma band activities are characterized by directions of information flow in the laminar profile. They found that, in V1, gamma band oscillations are initiated in input layer 4 and propagate to the deep and superficial layers of the cortex. On the other hand, alpha band oscillations propagate in the opposite direction. Moreover, simultaneously recording neural activities in both V1 and V4, they observed that gamma and alpha band oscillations respectively propagate in feedforward and

feedback directions between the two regions.

Similar results were obtained by other studies. Bastos *et al.* [54] recorded ECoG signals from the macaque visual cortex, and analyzed frequency-specific directed influences among 28 pairs of visual areas. They observed that feedforward influences are mainly carried by theta (< 4 Hz) and gamma (< 60 -80 Hz) band activities, whereas feedback influences by beta band (< 14 -18 Hz) activities. Their results suggest that the primate brain uses distinct frequency bands for regional feedforward and feedback processing.

2.3.7 Open problems and the purpose of this work

As reviewed, there exists spatial and temporal hierarchies in the ventral pathway, and different time-frequency bands seem to be multiplexed in both the spatial and temporal domains. However, few studies have investigated these problems with unified image set, subjects and measurement. Moreover, little has been known about what kind of visual information is represented in different time-frequency bands. Here, towards this end, we recorded spatiotemporal neural activities from the macaque ITC, and analyzed the complex, multiplexed activities with hierarchical visual features of deep CNNs. This approach enables us to understand not only information contents of different time-frequency bands, but also how these band-specific activities emerge in spatial and temporal domains in the brain.

Moreover, previous studies on neural oscillations [46, 44] suggest that mesoscopic neural oscillatory signals contain mixed activities from different time-frequency bands that have complementary selectivity and functional roles. Therefore, an analysis using spatiotemporal neural activities for diverse time-frequency bands is required for understanding the detailed relationship between neural activities in the ventral pathway and CNNs, and for developing more biologically plausible models.

Building



Body part



Face



Insect



Fruit



Tool



Figure 2-3: Examples images from each class in our image set. From top to bottom, building, body part, face, fruit, insect, and tool.

2.4 Materials and methods

2.4.1 Image set

We prepared diverse natural images from six object classes: building, body part, face, fruit, insect, and tool (Figure 2-3). First, 60,000 (10,000 per class) candidate images were selected by keyword search on *Flickr*. These images were then screened by a web-based survey using *Amazon Mechanical Turk*. For each class, the participant was given a brief instruction on image selection (mostly regarding image quality), and was presented six good and three bad example images. Each candidate image was evaluated by three participants. Images selected by all the three participants were considered as verified. For each class, if more than 1,000 images were verified, randomly-chosen 1,000 images among verified ones were used for our experiment. For stimulus presentation, original images were cropped to 512×512 pixels.

2.4.2 Subjects

We recorded ECoG signals from two female macaques (*Macaca fuscata*, Subject 1: 6.1 kg, Subject 2, 5.1 kg). All animal procedures were complied with the National Institute of Health Guide for the Care and Use of Laboratory Animals, and the Guide of the National BioResource Project "Japanese Monkeys" of the Ministry of Education, Culture, Sports and Technology (MEXT), Japan. The Niigata University Institutional Animal Care and Use Committee approved the experimental protocols.

2.4.3 Details of our ECoG system

For recording from the macaque inferior temporal cortex (ITC), we designed a 128-channel electrode grid that covers an area of $20 \text{ mm} \times 40 \text{ mm}$ with a inter-grid distance of 2.5 mm. The electrode was fabricated on a 20 μm -thick flexible Parylene-C film using micro-electro-mechanical systems technology. One side of the each square contact was 1 mm.

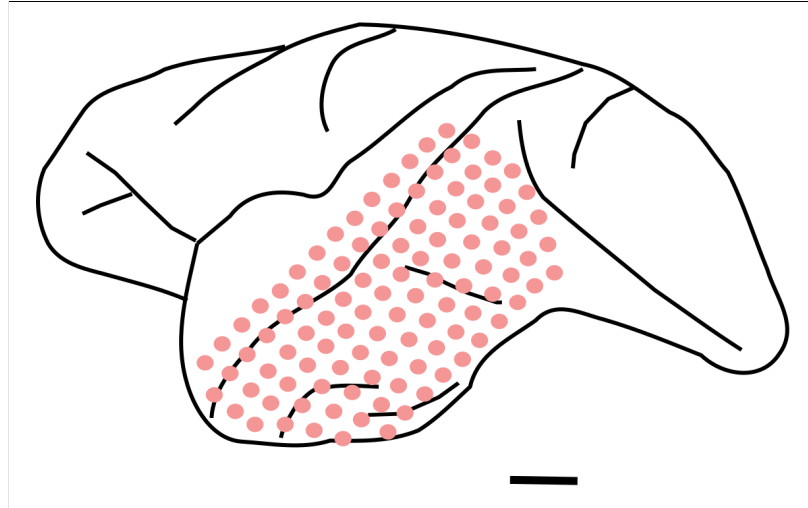


Figure 2-4: Lateral view of the macaque brain with an ECoG electrode implanted (the right hemisphere of Subject 1). Reconstructed with post-mortem observations. Pink dots indicate the position of the electrode contacts. The scale bar indicates 5 mm. Among total 128 contacts, 108 visible ones from this view are shown. The other 20 contacts are located on the ventral or medial surface of the cortex (not visible from this view).

The ECoG electrode was subdurally implanted under an aseptic conditions (see Figure 2-4 for a visualization of the electrode locations). After premedication with ketamine (50 mg/kg) and medetomidine (0.03 mg/kg), each subject was intubated with an endotracheal tube of 6 or 6.5 mm and connected to an artificial respirator (A.D.S.1000, Engler engineering corp., FL, USA). The venous line was secured using lactated Ringer's solution, and ceftriaxone (100 mg/kg) was dripped as a prophylactic antibiotic. Body temperature was maintained to keep around 37 °C using an electric heating mat. A vacuum fixing bed (Vacuform, B.u.W.Schmidt GmbH, Garbsen, Germany) was used to maintain the position of the body. The oxygen saturation, heart rate, and end-tidal CO₂ were continuously monitored (Surgi Vet, Smiths medical PM inc., London, UK) throughout the surgery to adjust the level of anesthesia. The skull was fixed with a 3-point fastening device (Integra Co., NJ, USA) with a custom-downsized attachment for macaques. The target location and the size of craniotomy were determined using preoperative magnetic resonance imaging. In the intra-dural operation we used a microscope (Ophthalmo-Stativ S22, Carl Zeiss Inc.,

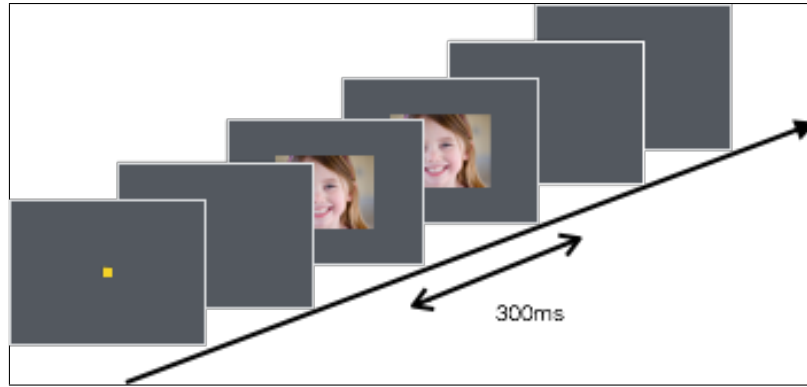


Figure 2-5: Stimulus presentation in ECoG recording.

Oberkochen, Germany) with a CMOS color camera (TS-CA-130MIII, MeCan Imaging Inc., Saitama, Japan). The electrode grid was carefully attached onto the cortical surface, and the dura was closed with water tight suturing to prevent cerebrospinal fluid leakage. The electrode lead, microconnectors (Omnetics, MN, USA), and a custom-made plastic connector chamber (Vivo, Hokkaido, Japan) were fixed onto the bone with resin.

2.4.4 ECoG recording

The subjects were trained with a visual fixation task to keep their gazes within ± 1.5 degree of visual angle around the fixation target. Eye movements were captured with an infra-red camera system (i-rec) with a sampling rate of 60 Hz.

Images were presented on a 15-inch CRT monitor (NEC, Tokyo, Japan) with a viewing distance of 26 cm. After 450 ms of stable fixation, each image was presented for 300 ms, followed by a 600-ms blank interval. Two to five images were successively presented as one single session. The subjects were rewarded with a drop of apple juice for maintaining their fixations over the entire duration of each session. The long axis of each image subtended 6 degrees of the visual angle. Images were presented with a PC running a custom-written OpenGL-based stimulation program. Behavioral control (timing, synchronization) was conducted by a network of interconnected PCs.

Signals were differentially amplified using a 128-channel amplifier (Plexon, TX,

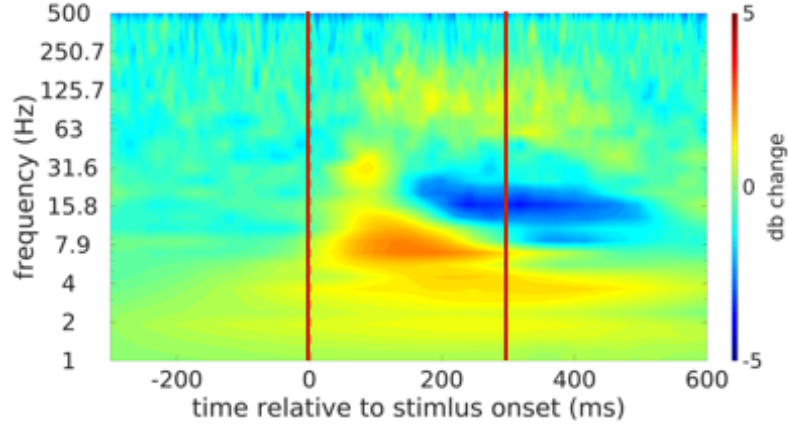


Figure 2-6: A visualized event-related spectral perturbation (ERSP).

USA or Tucker Davis Technologies, FL, USA) with high- and low-cutoff filters at 300 Hz and 1.0 Hz, respectively. All subdural electrodes were referenced to a titanium screw that was attached directly to the dura at the vertex area. Recording was conducted at a sampling rate of 1 kHz per channel. Recorded signals were online-monitored and stored on a PC-based system (NSCS, Niigata, Japan).

2.4.5 Time-frequency decomposition

As preprocessing, we first eliminated line noise in raw ECoG signals by applying a third-order Butterworth filter at 50 Hz. Then, we rereferenced signals at all the channels by taking bipolar derivatives between neighboring channels. Because electric potentials recorded by ECoG often contain noise from a non-cortical reference site or non-local signals, this rereferencing procedure can help us extract more local electric activities on the cortical surface. In total, we used ECoG signals at 112 sites extracted from the original 128 channels.

We computed the analytic amplitude at 30 frequencies using complex Morlet wavelet convolution. The central frequencies were logarithmically sampled from 1 to 250 Hz. For each central frequency f (Hz), we constructed a complex Morlet wavelet:

$$W(f, t) = \frac{1}{(\sigma\sqrt{\pi})^{1/2}} e^{i2\pi ft} e^{-t^2/2\sigma^2}, \quad (2.1)$$

Layers (num of channels)
Input (3: RGB)
conv1_1 (64)
conv1_2 (64)
pool1 (64)
conv2_1 (128)
conv2_2 (128)
pool2 (128)
conv3_1 (256)
conv3_2 (256)
conv3_3 (256)
pool3 (256)
conv4_1 (512)
conv4_2 (512)
conv4_3 (512)
pool4 (512)
conv5_1 (512)
conv5_2 (512)
conv5_3 (512)
pool5 (512)
fc (4096)
fc (4096)
fc (1000)
soft-max

Figure 2-7: The architecture of VGG-16 network.

where t is a time step (ms), $\sigma = n_f/(2\pi f)$ is the standard deviation, and n_f is the number of wavelet cycles. The number of wavelet cycles n_f for each central frequency was logarithmically sampled from 3 to 14. We computed the analytic amplitude as the absolute value of the convolution results between preprocessed ECoG signals and the wavelet:

$$A(f, t) = |W(f, t) * s(t)|. \quad (2.2)$$

After extracting the analytic amplitude, we conducted postprocessing for coping with trial-by-trial and temporal differences. We first normalized each trial's activities with the average amplitude in the baseline (-500 to -201 ms relative to the stimulus onset) using decibel conversion:

$$Z(f, t) = 10 \log_{10} \frac{A(f, t)}{\frac{1}{T_{baseline}} \sum_{t' \in baseline} A(f, t')}, \quad (2.3)$$

where $T_{baseline}$ is the number of time steps in the baseline (300 ms). After baseline normalization, we took the cross-trial average of normalized decibel changes over five trials for each stimulus. Finally, we downsampled multi-trial activities in nine sliding time windows (1-100, 51-150, 101-200, 151-250, 201-300, 251-350, 301-400, 351-450, and 401-500 ms relative to the stimulus onset). As the result, we obtained frequency-specific ECoG activities for 112 sites, 30 central frequencies, and 9 time windows for each monkey.

2.4.6 Extracting hierarchical visual features from convolutional neural networks

Deep convolutional neural networks (CNNs) have achieved state-of-the-art performance on diverse computer vision tasks, such as object recognition, semantic segmentation, object detection, and video recognition [8, 9, 10, 11]. Several previous studies [83, 84, 85] showed that representations in pretrained CNNs are efficiently applicable on novel image sets and tasks (e.g., object category classification, scene recognition, fine grained recognition, attribute detection and image retrieval). Furthermore, CNNs enable us to extract hierarchical visual features, since CNNs have its layer hierarchy and previous studies indicate that lower, mid, and higher CNN layers represent lower-, mid-, and higher-level visual features, respectively [57]. Therefore, pretrained CNNs are useful for extracting optimized hierarchical visual features. Interestingly, several recent studies in cognitive neuroscience investigated brain activities using hierarchical visual features from CNNs, and found the similarity between the layer hierarchy of CNNs and the anatomical hierarchy of the primate ventral stream [12, 14].

In this work, we used a pretrained CNN for analysing frequency-specific ECoG activities. We employed a VGG-16 network [39] that was pretrained on the ILSVRC2012 object classification task [86]. We fed our image set to the pretrained CNN, and extracted visual features at each layer. VGG-16 consists of 13 convolution layers, 5 max pooling layers, 3 fully-connected (FC) layers, and the final classification (softmax) layer. We extracted visual features from all the convolution and FC layers,

resulting in 16 layers used in total.

While features at FC layers are vectors, those at convolution layers are three-dimensional tensors that have width, height, and depth (channels). In our experiment, we downsampled the output of convolution layers by taking the spatial average (global average pooling) over the width and height.

2.4.7 Encoding ECoG features from CNN features

To compare the prediction performance over the frequency bands and CNN layers, we trained and evaluated encoding models that predict frequency-specific ECoG activities from visual features at a specific CNN layer. More specifically, each encoding model was trained to predict ECoG activities at a specific site, frequency, and time window, given CNN features at a specific layer as input:

$$\hat{y}_i = \mathbf{w}^\top \phi_l(\mathbf{x}_i) + b, \quad (2.4)$$

where $\mathbf{x} \in \mathbb{R}^{3 \times 224 \times 224}$ is the image, $\phi^l(\mathbf{x}_i) \in \mathbb{R}^{C_l}$ is a CNN feature at l -th layer, $\mathbf{w} \in \mathbb{R}^{C_l}$ is a weight vector that projects CNN features into a scalar, and $b \in \mathbb{R}$ is a bias term. The parameters, \mathbf{w} and b , were optimized using ridge regression:

$$\mathbf{w}^* = \arg \min_{\mathbf{w}} \frac{1}{N_{train}} \sum_{i=1}^{N_{train}} (y_i - \hat{y}_i)^2 + \lambda \|\mathbf{w}\|^2, \quad (2.5)$$

where N_{train} is the number of training samples, and λ is a hyperparameter to control the penalty term. Encoding models were independently trained for each combination of ECoG site, frequency, time window, and CNN layer.

We split the image set so that no image in the validation and test set is included in the training set. In monkey 1’s dataset, the training, validation, and test set contains 2431, 808, and 808 stimuli, respectively. In monkey 2’s dataset, the training, validation, and test set contains 2441, 813, and 813 stimuli, respectively. Before training encoding models, CNN features in the training set were standardized so that they have zero mean and unit variance. CNN features in the validation and test

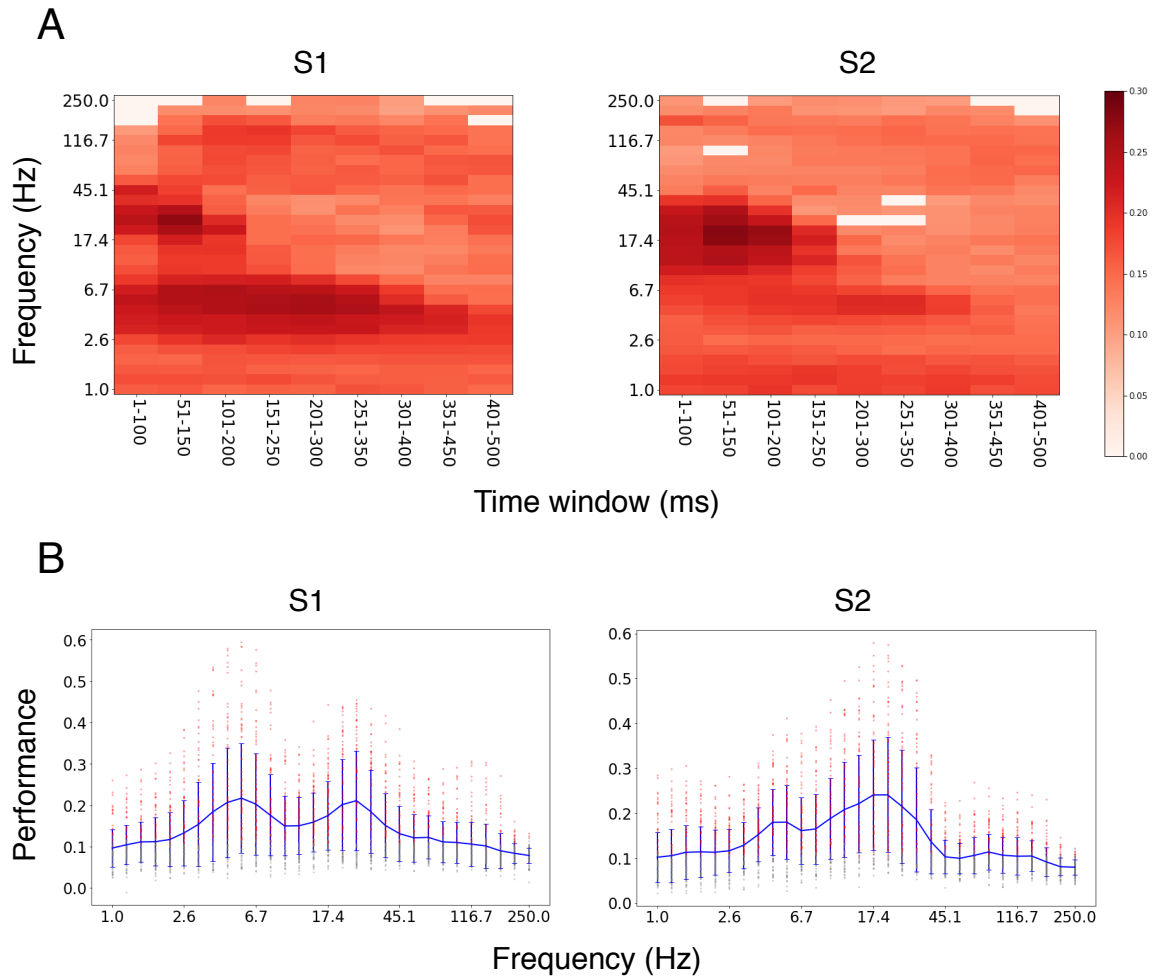


Figure 2-8: Comparison of the prediction performance. In the test set, the prediction performance was measured as Pearson's correlation coefficient between ground truth and predicted values. (A) The prediction performance over the frequencies and time windows. For each site, the maximum performance over the CNN layers was extracted. The average performance over sites that showed better performance than the significance threshold ($p < 0.0001$ in the permutation test) is shown here. (B) The prediction performance over the frequencies. Red dots indicate the prediction performance of each ECoG site. For each site, the maximum performance over the time windows and CNN layers was extracted. Only results above the significance threshold are shown here ($p < 0.0001$ in the permutation test). Blue line indicate the mean prediction performance over the ECoG sites. Blue error bars indicate the standard error of the prediction performance over the ECoG sites.

sets were normalized with the mean and standard deviation in the training set. We implemented our encoding models with *PyTorch* [87]. We optimized our encoding models using the *Adam* optimizer [88], with a learning rate of 10^{-4} , $\beta_1 = 0.9$, $\beta_2 = 0.999$, a weight decay (λ) of 10^{-6} , and a batch size of 128. The maximum training epoch was 100 epochs, but we stopped training when the validation performance was not improved for successive 10 epochs.

In the test set, we evaluated each encoding model’s prediction performance as Pearson’s correlation coefficient between ground truth and predicted values. To eliminate results that can occur by chance, we determined the significance threshold of Pearson’s correlation coefficient using the permutation test. For each encoding model, we computed the correlation between ground truth and randomly-permuted predictions. We repeated the procedure for 1,000 permutations, and took the largest value as the threshold.

2.5 Experiments

2.5.1 Theta and gamma bands are better predicted from CNN features

First, we compare the difference of the prediction performance over the frequency bands. Comparing the prediction performance over the frequency bands shows us which frequency bands are more related to visual features extracted from a pretrained CNN. Our hypothesis here is that specific frequency bands show better performance than the other bands. That is, specific frequency bands might show stronger selectivity to visual features in CNNs.

Figure 2-8 shows the comparison of the prediction performance over the frequency bands. For each ECoG site, we took the maximum performance over the time windows and CNN layers for visualization. For both monkeys, the average prediction performance over the sites was peaked at theta (around 5 Hz) and gamma bands (around 20-25 Hz), showing that ECoG activities in the low- and high-frequency

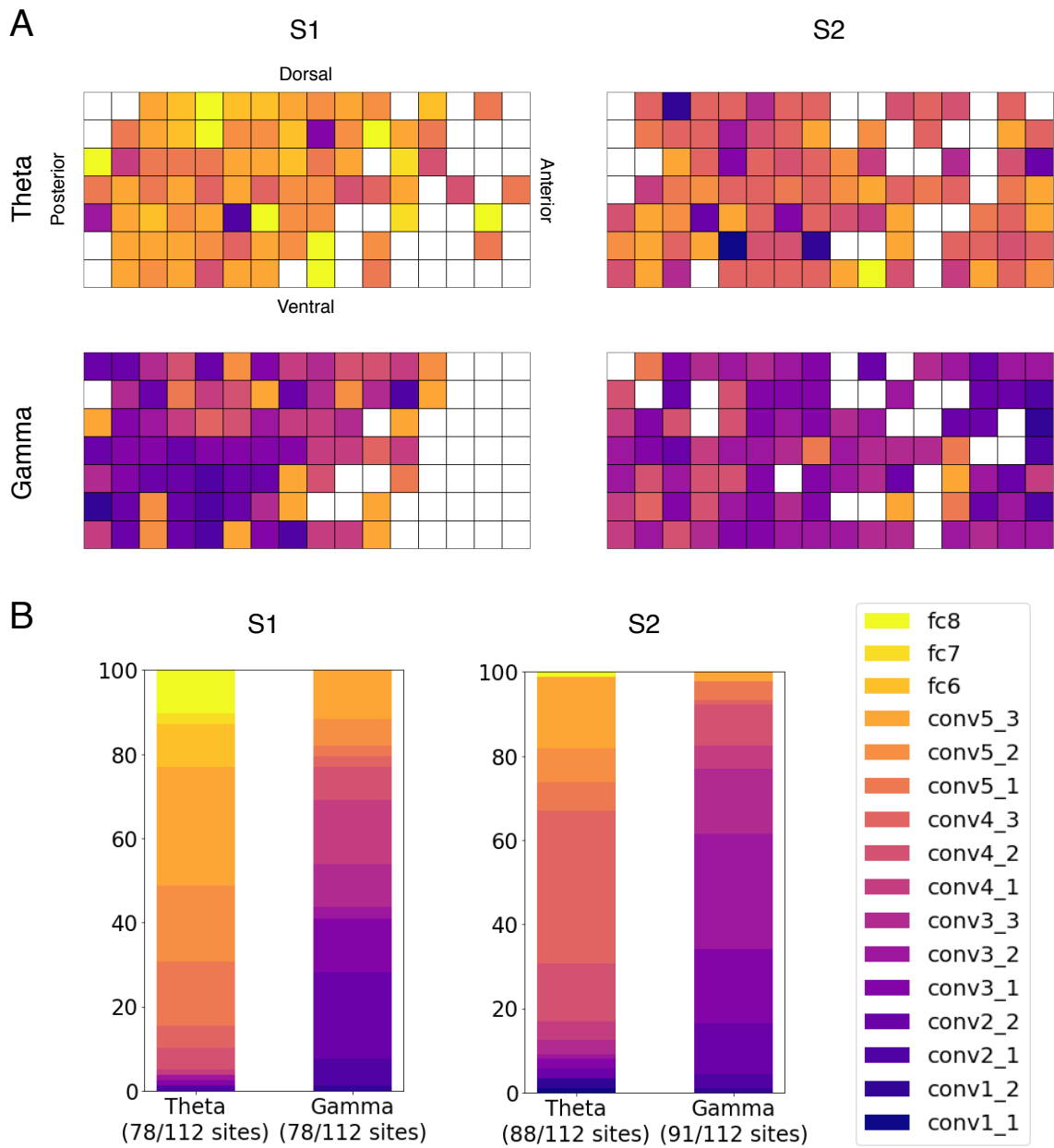


Figure 2-9: Assignments of the CNN layers. For each site, the maximum performance over the time windows is extracted. Only sites above the significance threshold are shown here ($p < 0.0001$ in the permutation test). (A) Topographical visualizations of assigned time windows. The top, bottom, right, and left side of each electrode map corresponds to the dorsal, ventral, anterior, and posterior part of the macaque brain, respectively. The color at each site indicates the assigned layer. (B) Proportion of each CNN layer in assignments.

bands were better predicted from CNN features than the other frequency bands. In delta (1-2 Hz) and high-gamma bands (100-200 Hz), several sites showed the prediction performance around 0.3, but the curve of the mean prediction performance did not peak at these frequency bands.

2.5.2 Theta and gamma bands are better predicted from higher and lower CNN layers, respectively

In visual object recognition, CNNs receive input images, perform convolution and subsampling (pooling) operations at each layer, and finally output object-level classification results. Therefore, units in higher CNN layers have larger receptive fields. Furthermore, several previous studies analyzed the visual selectivity of units in each layer, and showed that lower and higher layers have selectivity to lower- (e.g., color, orientation, grating, texture) and higher-level (e.g., shape, object part, animal face) visual features [57, 89]. Since visual features in higher CNN layers have larger receptive fields and are correlated with higher-level visual patterns, we can investigate the difference of theta- and gamma-band activities in terms of the hierarchy of visual features in CNNs. Here, we visualized the relationship between the CNN layers and the frequency bands.

Figure 2-9 shows the comparison of layer assignments for theta and gamma bands. For both monkeys, in gamma band, most sites were assigned lower convolution layers. On the other hand, in theta band, most sites were assigned higher convolution and FC layers. Several sites in theta band were assigned the highest layer (FC8), which outputs class-wise logits for classification, but no site in gamma band was assigned the FC8 layer. These results show the distinct visual selectivity of theta- and gamma-band ECoG activities in terms of the layer hierarchy of CNNs. While our ECoG electrode covers from the posterior to anterior ITC, we did not observe strong spatial modulations of assigned CNN layers in the posterior-to-anterior direction.

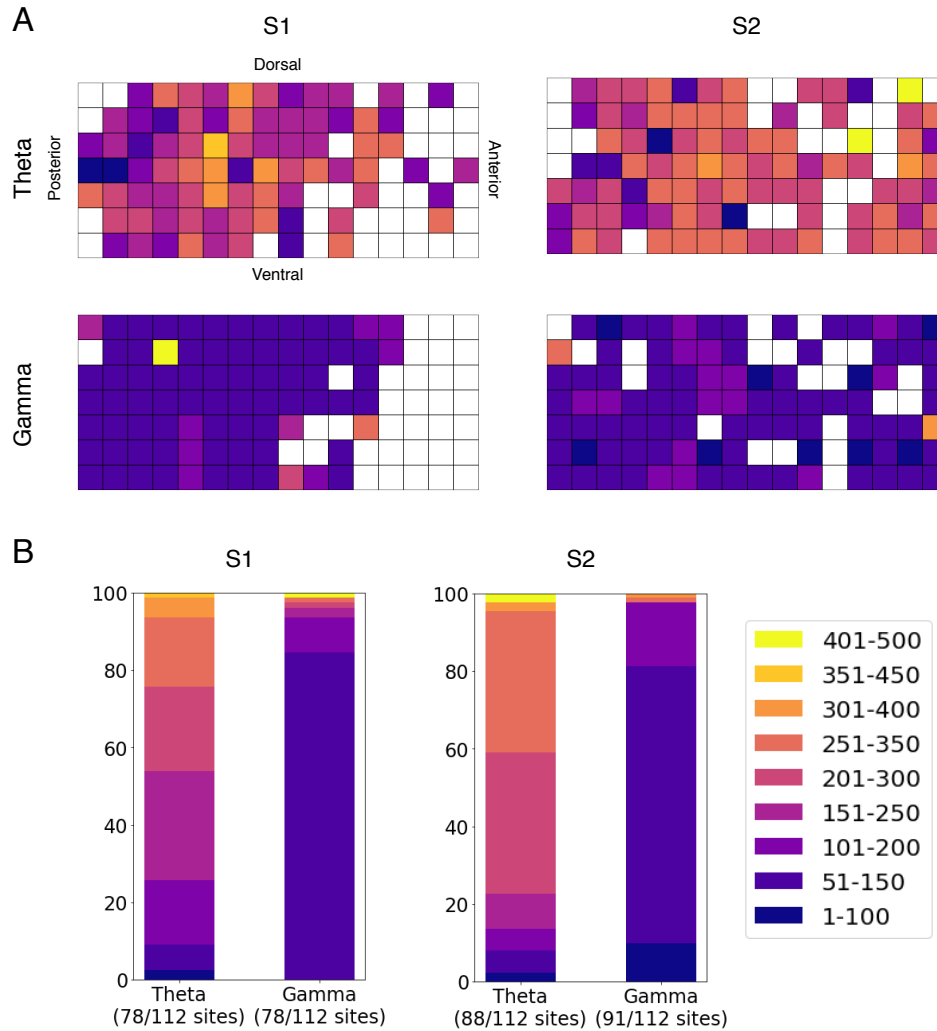
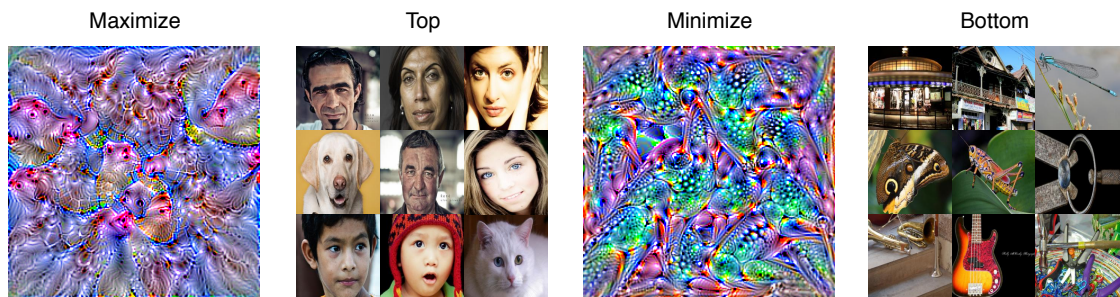


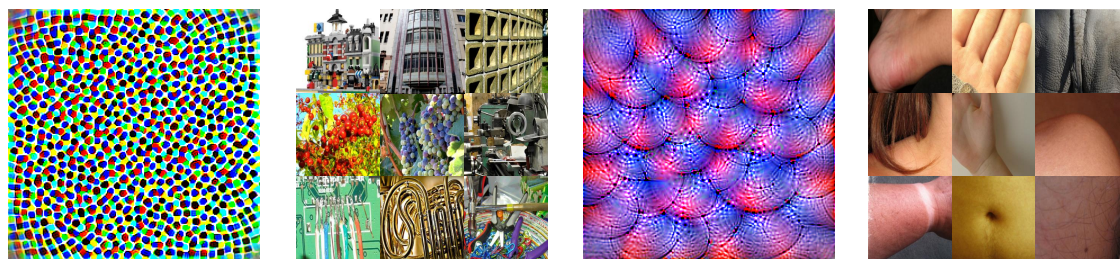
Figure 2-10: Assignments of the time windows. For each site, the maximum performance over the CNN layers is extracted. Only sites above the significance threshold are shown here ($p < 0.0001$ in the permutation test). (A) Topographical visualizations of assigned time windows. The top, bottom, right, and left side of each electrode map corresponds to the dorsal, ventral, anterior, and posterior part of the macaque brain, respectively. The color at each site indicates the latency of the assigned time window. (B) Proportion of each time window in assignments.

S1

Theta (site 58)

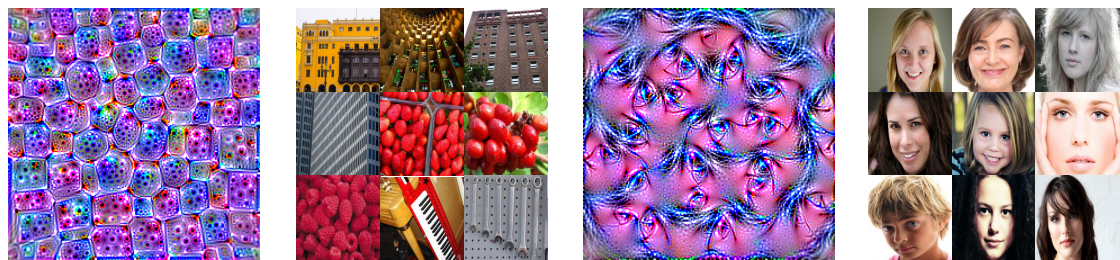


Gamma (site 98)



S2

Theta (site 103)



Gamma (site 74)

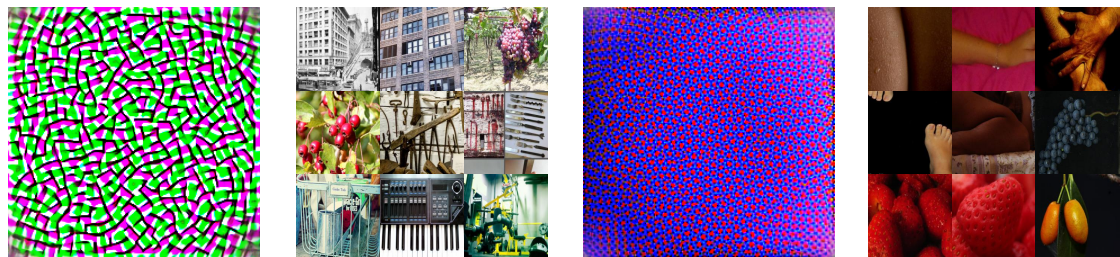


Figure 2-11: Examples of optimized (maximize, minimize) and preferred (top, bottom) images for theta- and gamma-band encoding models. Optimized images were produced by updating randomly-initialized images so as to maximize or minimize the predicted value of each encoding model. Preferred images were selected based on predicted values of each encoding model on the test set.

2.5.3 Theta and gamma bands are better predicted in later and earlier time windows, respectively

Next, we examined the difference of preferred time windows between theta and beta bands. If theta- and beta-band activities are respectively related to higher- and lower-level visual features, in view of serial information processing, theta and beta bands should respectively prefer later and earlier time windows, because lower-level visual features should arrive first in passive visual perception.

Figure 2-10 shows the comparison of assigned the time windows for theta and gamma bands. For gamma band, most sites were assigned earlier time windows (around 1-200 ms). On the other hand, for theta band, most sites were assigned later time windows (around 201-500 ms). Therefore, as we expected, the prediction performance of gamma-band activities peaked at earlier time windows, and then the prediction performance of theta-band activities peaked at later time windows. Similar to CNN layer assignments, we did not observe strong spatial modulations of assigned time windows in the posterior-to-anterior direction.

2.5.4 Visualizing the selectivity of theta- and gamma-band encoding models

Next, we examined more detailed visual selectivity of theta and gamma bands using optimized encoding models. Each encoding model first transforms an image with convolution and pooling operations in the pretrained CNN, and then projects CNN features into frequency-specific prediction value. Starting from random images, we can optimize the image so as to maximize the prediction value with gradient descent [90, 91]:

$$\mathbf{x}^* = \arg \max_{\mathbf{x}} \mathbf{w}^T \phi_l(\mathbf{x}) + b, \quad (2.6)$$

where ϕ_l is a feature extractor at a specific CNN layer given an image. On the other hand, in minimization, we optimize the image so as to minimize the prediction value:

$$\mathbf{x}^* = \arg \min_{\mathbf{x}} \mathbf{w}^\top \phi_l(\mathbf{x}) + b. \quad (2.7)$$

With this method, we can qualitatively investigate what kind of visual patterns are preferred by theta- and gamma-band encoding models. Using the pretrained CNN and frequency-specific encoding models, we optimize a randomly-initialized image using the Adam optimizer with a learning rate of 0.01, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. To avoid noisy results, we stochastically transformed the image before feeding into the CNN [91]. More specifically, we stochastically jittered (0-8 pixels), rotated (0-45 degree), and scaled (1.0-1.8 times) the image. We also extracted nine top and bottom images based on test-set predictions of each encoding model.

Several examples of optimized and preferred images are shown in Figure 2-11. For theta-band models, preferred images appeared to be strongly related to higher-level visual features (e.g., face), and optimized images tended to contain higher-level, more complex visual patterns. In contrast, for gamma-band models, preferred images appeared to be related to lower-level visual features, and optimized images tended to contain lower-level, more local visual patterns.

2.6 Discussion and Conclusion

In this work, we have analyzed frequency-specific ECoG activities in the macaque ITC using hierarchical visual features extracted from a pretrained CNN. By analyzing the prediction performance and visual selectivity of our encoding models, we were able to quantitatively and qualitatively investigate the frequency-specific visual selectivity of ECoG activities. Comparing the prediction performance of our encoding models, we observed that ECoG activities in two time-frequency bands, theta and gamma bands, were better predicted from CNN features than the other bands. In the comparison of assigned CNN layers, we found that theta- and gamma-band activities were better predicted from higher and lower layers, respectively. Furthermore, our visualization analysis qualitatively showed that theta- and gamma-band activities were related to higher- and lower-level visual features, respectively. Our results indicate that distinct

levels of visual information are multiplexed in low- and high-frequency activities in brain signals.

There have been several previous studies on the visual selectivity of frequency-specific neuronal activities in the visual cortex. For example, Belitski *et al.* [51] analyzed local field potentials (LFPs) and multi-unit activities (MUAs) recorded from the primary visual cortex of anaesthetized monkeys. Their results indicate that: (1) lower (< 12 Hz) and high-gamma (60-100 Hz) bands are more related to visual information, and (2) these two bands are not correlated with each other, implying that these two time-frequency bands represent distinct visual features. Similar results were also observed in [23], where human EEG activities in multiple time-frequency bands are analyzed on a face image set. They found that lower- and higher-frequency bands were related to distinct facial features. Jacobs *et al.* [49] analyzed ECoG signals recorded from neurosurgical patients. Their results indicate that: (1) gamma band (25-128 Hz) is more informative than the other bands for classifying letter identities, and (2) gamma band has strong phase-amplitude coupling with theta band (4-8 Hz). Recently, similar to our work, Kuzovkin *et al.* [92] analyzed frequency-specific ECoG activities in the human visual cortex using a pretrained CNN. They employed the *representation similarity analysis* (RSA) [93] to compare CNN layers and frequency-specific ECoG activities. They observed that, the layer hierarchy of CNNs matched the anatomical hierarchy of the human ventral visual pathway for gamma-band (30-150 Hz) ECoG activities. In this work, using hierarchical representations extracted from CNNs, we observed that theta- and gamma-band activities were more related to visual features than the other bands, and that these two time-frequency bands have distinct visual selectivity. It is noteworthy that we found the distinct visual selectivity of theta and gamma band activities on a more diverse set of natural images.

Our findings are related to recent studies on frequency-specific roles in feedforward and feedback processing in the primate visual cortex. Bastos *et al.* [54] analyzed ECoG signals recorded from rhesus monkeys to investigate whether inter-areal influences between visual cortical areas are subserved differentially by different frequency bands. Their results indicate that feedforward influences are related to theta (~ 4

Hz) and gamma band ($\sim 60-80$ Hz), and that feedback influences by beta band ($\sim 14-18$ Hz). Similarly, van Kerkoerle *et al.* [82] reported that, in the monkey visual cortex, alpha (5-15 Hz) and gamma (40-90 Hz) bands are related feedback and feedforward processing between V1 and V4. These studies indicate that several different frequency bands are related to frequency-specific roles in feedforward or feedback processing between visual cortical areas. Combining our results with these previous studies on frequency-specific roles in inter-areal communications, it could be possible that theta and gamma bands carry distinct visual information for different roles in inter-areal communications in the visual cortex.

Chapter 3

Natural Image Reconstruction from ECoG Signals Using Deep Learning

3.1 Introduction

When a person sees an image, complex neuronal activities occur in diverse scales of the brain. The goal of brain decoding [94, 3] is to predict visual features from recorded brain activities. Developing better decoding methods is important for understanding the underlying mechanism of visual perception [95, 96, 97]. Also, accurate decoding of visual features is crucial for real-world brain-computer interface (BCI) applications [98].

In the literature of brain decoding, most previous studies are on classification tasks, where the goal is to predict a class label given brain activities as input. In classification tasks, models map brain activities into the most likely class in a predefined class set. Because outputs are limited in the class level, it is difficult to analyze the relationship between more fine-grained visual features and brain activities with classification models. Therefore, for understanding how complex brain activities are related to diverse visual features, it is important to develop image reconstruction methods, which directly predict presented images solely from brain signals.

A number of studies have proposed reconstruction methods for various types of

images, such as binary contrast patterns [24], characters [25], colors [26], faces [27], and natural movies [28]. Following successful applications of deep learning in computer vision, several recent studies used deep learning for reconstructing face images [16] or natural images [17, 99, 100, 18].

Most previous studies on image reconstruction proposed various methods for human functional magnetic resonance imaging (fMRI) data. Although fMRI can cover a broad part of the brain, its hemodynamic responses inherently limit the temporal resolution of recorded signals. In the brain, neuronal activities continuously change in the submillisecond level. Therefore, to elucidate the relationship between visual experiences and complex neuronal activities, it is crucial to develop image reconstruction method for high-temporal-resolution electrophysiological recordings, such as electroencephalography (EEG), magnetoencephalography (MEG), and electrocorticography (ECoG). Furthermore, accurate decoding of various stimuli from electrophysiological signals is crucial for real-world brain-computer interface (BCI) applications [19, 20].

Towards this end, we study natural image reconstruction from ECoG signals using deep learning. We first recorded ECoG signals from the inferior temporal cortex (ITC) of two macaque monkeys while presenting diverse natural images. Then, we trained three deep learning models that reconstruct presented images from ECoG signals. We quantitatively and qualitatively evaluated our reconstruction models. Our results suggest the possibility of reconstructing diverse natural stimuli from high-temporal-resolution electrophysiological recordings.

3.2 Related work

In the literature of brain decoding, most previous studies are on classification tasks, where the goal is to predict a class label given brain activities as input. In classification tasks, models map brain activities into the most likely class in a predefined class set. Because outputs are limited in the class level, it is difficult to analyze the relationship between more fine-grained visual features and brain activities with classification models. Therefore, for understanding how complex brain activities are

related to diverse visual features, it is important to develop image reconstruction methods, which directly predict presented images solely from brain signals.

A number of studies have proposed reconstruction methods for various types of images, such as binary contrast patterns [24], characters [25], colors [26], faces [27], and natural movies [28]. Following successful applications of deep learning in computer vision, several recent studies used deep learning for reconstructing face images [16] or natural images [17, 99, 100, 18].

3.2.1 Image identification from fMRI data

One way to predict an image from brain activities is to identify the most probable image from a pre-defined image set. Kay *et al.* [101] proposed an image identification for fMRI data in the primary visual cortex (V1, V2, V3). They first estimate receptive-field models (space, orientation, spatial frequency) that predict each voxel's responses. Then, given a set of candidate images, they compute predictive voxel responses for each image using estimated receptive-field models. They compare Pearson's correlation coefficient between predicted and actual voxel responses, and select the best candidate image as their answer. Naselaris *et al.* [3] extended the framework proposed in [101], and proposed a Bayesian image identification method that combines voxel-wise receptive-field models, a larger candidate image set, and semantic annotations. For reconstructing natural videos, Nishimoto *et al.* [28] proposed a motion-energy encoding models that separately describes fast visual motion information and much slower, visually-evoked hemodynamic responses. Combining their encoding models and the Bayesian image identification framework in [3], they produced a series of the most probable video clip or an average of several most probable clips as reconstructions.

Although identification-based methods are easy to apply, they have several problems. First, they limit the output image space, because they require a pre-defined image set as candidate reconstructions. This is problematic when models are used for reconstruction of a novel class of images or perceived scenes, which cannot be explicitly defined by researchers beforehand. Moreover, the computational complexity

of identification-based methods increase with the size of brain activities (if encoding-based identification is employed) and the number of candidate images.

3.2.2 Image reconstruction from fMRI data

Most previous studies on image reconstruction proposed various methods for human functional magnetic resonance imaging (fMRI) data. Miyawaki *et al.* [24] proposed a reconstruction method based on a linear combination of multiple local image bases at multiple scales. The weight of each image basic is estimated from their fMRI dataset, and they used their proposed method for reconstruction of binary contrast patterns. Güçlütürk *et al.* [16] proposed a method using CNNs and principal component analysis (PCA) for reconstructing aligned face images. They first predict dimension-reduced CNN features from fMRI data, and then reconstruct face images from the predicted features. They applied a generative adversarial network (GAN) in the reconstruction part. Seeliger *et al.* [17] used pretrained deep convolutional GANs (DC-GANs) [102] for reconstructing images of various objects. Specifically, they trained a model that predicts the latent code of a DCGAN, so that DCGAN's output image becomes close to the presented image. Shen *et al.* [100] employed a similar approach to the *activation maximization* method [90, 103], which has been a popular method for visualizing the selectivity of units and layers of CNNs. They first trained a decoding model that predicts intermediate features of a pretrained CNN from fMRI data. Then, they estimate an image reconstruction by minimizing the difference between decoded and reconstruction-induced features. Shen *et al.* [18] proposed an end-to-end method that directly reconstructs natural images from fMRI data. Their models are based on CNNs, and trained with a weight combination of pixel-wise loss, perceptual loss, and adversarial loss.

3.2.3 Image reconstruction from EEG signals

Although fMRI can cover a broad part of the brain, its hemodynamic responses inherently limit the temporal resolution of recorded signals. In the brain, neuronal

activities continuously change in the submillisecond level. Therefore, to elucidate the relationship between visual experiences and complex neuronal activities, it is crucial to develop image reconstruction method for high-temporal-resolution electrophysiological recordings, such as electroencephalography (EEG), magnetoencephalography (MEG), and electrocorticography (ECoG). Furthermore, accurate decoding of various stimuli from electrophysiological signals is crucial for real-world brain-computer interface (BCI) applications [19, 20]. Recently, Rashkov *et al.* [104] proposed a deep learning method for reconstructing natural images (visual illusion, waterfall, human face, Goldberg machine, sport) from human EEG signals. They computed latent representations for reconstructions and ground truth, and their reconstruction model is trained so that the distance between these latent representations. They also proposed a novel feedback loop for enhancing reconstructions; the subject is presented the current reconstruction for producing better reconstructions.

3.2.4 Open problems and the purpose of this work

In the vision domain, most previous work in brain decoding has focused on classification tasks, and few studies have studies image reconstruction from brain activities, especially for brain signals (EEG, ECoG, MEG) under natural stimuli. We developed deep learning-based models that reconstruct diverse natural images from ECoG signals recorded in the primate inferior temporal cortex (ITC). To investigate what kind of models are effective for the task, we trained and evaluated multiple state-of-the-art image restoration methods in deep learning.

3.3 Methods

3.3.1 Background

In brain recording, we obtain brain signals $y \in \mathbb{R}^{C \times T}$ while presenting an image $x \in \mathbb{R}^{H \times W \times 3}$ to a subject. Here, C is the number of recording channels, T is the total length of brain recording, and H and W are respectively the height and width

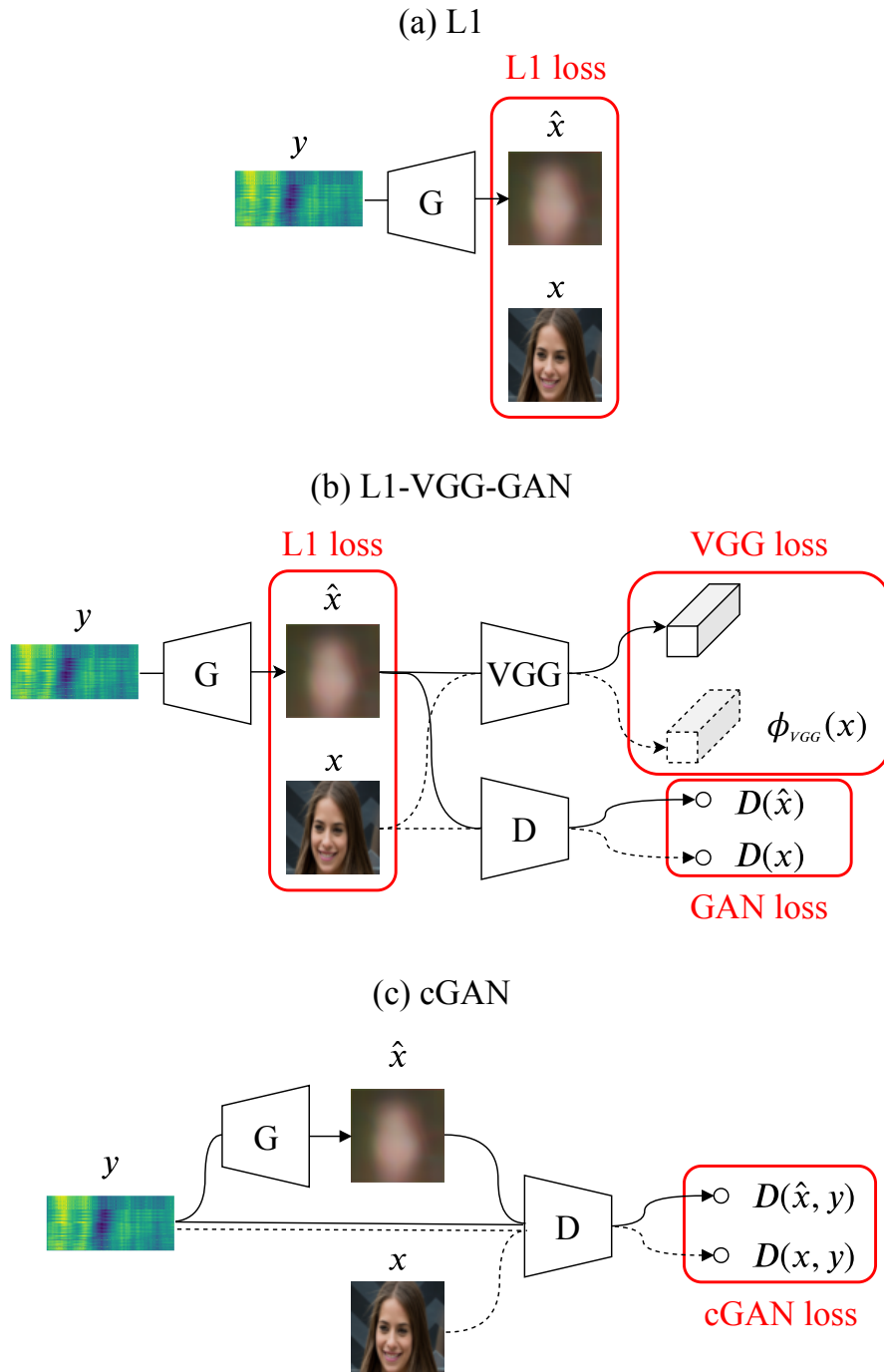


Figure 3-1: Image reconstruction models (generator: G, discriminator: D) in our experiments. (a) The L1 model is trained only with the pixel-wise error (L1 loss). (b) The L1-VGG-GAN model is trained with a weighted combination of L1, perceptual, and adversarial losses. We used a pretrained VGG-16 network for computing perceptual loss. (c) The conditional GAN (cGAN) model is trained with the conditional formulation of GAN. In this case, the discriminator receives both an image (reconstruction or ground truth) and brain signals.

of the image. A reconstruction model (generator) G , which is parameterized with θ , outputs a reconstruction \hat{x} given y . Our goal is to optimize the model’s parameters θ with the dataset so as to obtain as better reconstructions as possible for novel trials.

3.3.2 Models

Most previous work on image reconstruction methods for human fMRI activities [105, 106, 18, 17, 99] used a combination of pixel-level distortion losses (e.g. L1, L2), perceptual losses [107], and adversarial losses [108]. In this work, we compare the reconstruction performance of the following three models: L1, L1-VGG-GAN, and conditional GAN (cGAN). Figure 3-1 illustrates each model’s architectural diagram.

L1

The L1 model is trained only with the L1 loss, which is the sum of absolute pixel-wise errors between reconstructions and presented images:

$$\mathcal{L}_{L1} = \mathbb{E}_{x,y}[\|x - G(y)\|_1]. \quad (3.1)$$

L1-VGG-GAN

The L1-VGG-GAN model is trained with a weighted combination of the L1, perceptual, and adversarial losses. The perceptual loss is the difference of reconstructions and presented images in terms of inter-mediate features of a pretrained classification model:

$$\mathcal{L}_{VGG} = \mathbb{E}_{x,y}[\|\phi_{VGG}(x) - \phi_{VGG}(G(y))\|_1], \quad (3.2)$$

where $\phi_{VGG}(x)$ and $\phi_{VGG}(G(y))$ are features extracted from the conv5_3 layer of a VGG16 [109] network (pretrained on the ILSVRC2012 classification task).

The adversarial loss (generative adversarial network: GAN) [108, 110] forces reconstructions to look similar to natural images by simultaneously training with the discriminator network D . The discriminator is trained to distinguish between re-

constructed/fake \hat{x} and presented/real images x . GAN is trained with the following objective:

$$\mathcal{L}_D(G, D) = \mathbb{E}_x[\log D(x)] + \mathbb{E}_y[\log(1 - D(G(y)))] \quad (3.3)$$

$$\mathcal{L}_G(G, D) = -\mathbb{E}_y[\log(1 - D(G(y)))] \quad (3.4)$$

Putting the L1, perceptual, and adversarial losses together, the L1-VGG-GAN model is trained with:

$$\mathcal{L}_{\text{L1-VGG-GAN}} = \lambda_{L1}\mathcal{L}_{L1} + \lambda_{VGG}\mathcal{L}_{VGG} + \lambda_G\mathcal{L}_G, \quad (3.5)$$

where λ_{L1} , λ_{VGG} , and λ_G control the influence of each loss term.

Conditional GAN

Conditional GAN (cGAN) is the conditional formulation of GAN, and has recently been a popular method in various computer vision tasks, such as class-conditional image generation [111], single-image super resolution [112], text-to-image synthesis [113, 114], image-to-image synthesis [115], and video-to-video synthesis [116]. In cGAN, the discriminator is trained to distinguish between real pairs (x, y) and fake pairs (\hat{x}, y) , and the generator is trained to generate a realistic image conditioned by y . cGAN is trained with the following objective:

$$\begin{aligned} \mathcal{L}_D(G, D) = \mathbb{E}_{x,y}[\log D(x, y)] + \\ \mathbb{E}_{x,y}[\log(1 - D(G(y), y))] \end{aligned} \quad (3.6)$$

$$\mathcal{L}_G(G, D) = -\mathbb{E}_{x,y}[\log(1 - D(G(y), y))]. \quad (3.7)$$

With the above objective, the generator parameters are trained to mimic the image distribution conditioned with brain signals $p_{data}(x|y)$, while the discriminator parameters are trained to distinguish fake and real images conditioned with brain signals.

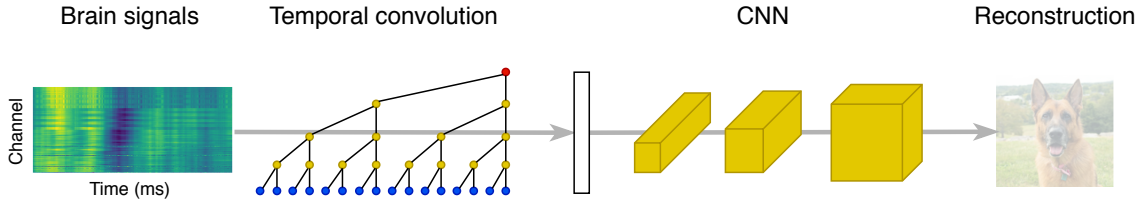


Figure 3-2: Image reconstruction from brain signals. Single-trial brain signals are first transformed into a vector by a temporal (1D) convolution network (TCN). Then, the vector is used to produce a reconstruction by a convolutional neural network (CNN).

Table 3.1: The network architecture of the generator (reconstruction) network.

$y \in \mathbb{R}^{128 \times 500}$
Conv 1D (256)
ResBlock 1D (256) \times 6
FC (256)
FC ($4 \times 4 \times 1064$)
Up, ResBlock 2D (512)
Up, ResBlock 2D (256)
Up, ResBlock 2D (128)
Up, ResBlock 2D (64)
Up, ResBlock 2D (64)
BatchNorm, ReLU, Conv 2D (3), Tanh

3.3.3 Network architecture

Figure 3-2 shows a diagram of our model for reconstructing images from brain signals, and Table 3.1 shows the network architecture of the generator (reconstruction) network. The generator first converts brain signals y into a hidden vector with a temporal convolutional network (TCN) [117, 118] that has one 1D convolution layer and six residual blocks. Then, a residual network (ResNet) [40] with five residual blocks transforms the hidden vector to generate a reconstruction \hat{x} . In the residual blocks of both TCN and ResNet, batch normalization [119] and rectified linear unit (ReLU) activation are applied to inputs before each of two convolutional layers.

Table 3.2: The network architecture of the discriminator network for L1-VGG-GAN and cGAN models. The conditional part is used only in cGAN models.

$x \in \mathbb{R}^{64 \times 64 \times 3}$
Conv 2D (64)
ResBlock 2D (64), Down
ResBlock 2D (128), Down
ResBlock 2D (256), Down
ResBlock 2D (512), Down
ResBlock 2D (1024), Down
FC (1)
(cond) $y \in \mathbb{R}^{128 \times 500}$
Conv 1D (256)
ResBlock 1D (256) $\times 6$
FC (1024 \times 4 \times 4)

Table 3.2 shows the network architecture of the discriminator for L1-VGG-GAN and cGAN models. The discriminator first processes an image with a ResNet with five residual blocks. The feature map of the final residual block (1024 \times 4 \times 4) is reshaped into a vector. Then, the final FC layer computes the discriminator’s unconditional output. In cGANs, we employ the projection-based discriminator architecture [112]. In the conditional path, brain signals are fed into a TCN that has the identical architecture with the one in the generator. The final FC layer in the conditional path maps the TCN’s output to a vector that has the same number of units as the feature map in the final residual block of the unconditional path.

3.3.4 Stabilizing GAN training

In our implementation, we apply two methods for stabilizing the training of GANs. First, as suggested in [108], we use the *non-saturating loss* for training the generator:

$$\mathcal{L}_G(G, D) = \mathbb{E}_y[\log D(G(y), y)]. \quad (3.8)$$

Second, we regularize the discriminator by applying the zero-centered gradient penalty [120, 121]:

$$R_{real}(\psi) := \lambda_{GP} \mathbb{E}_{p_{\mathcal{D}}(x)} [\|\nabla D_{\psi}(x)\|^2] \quad (3.9)$$

$$R_{fake}(\psi) := \lambda_{GP} \mathbb{E}_{p_{\mathcal{D}}(y)} [\|\nabla D_{\psi}(G(y))\|^2], \quad (3.10)$$

where λ_{GP} is a hyperparameter for controlling the strength of each penalty term.

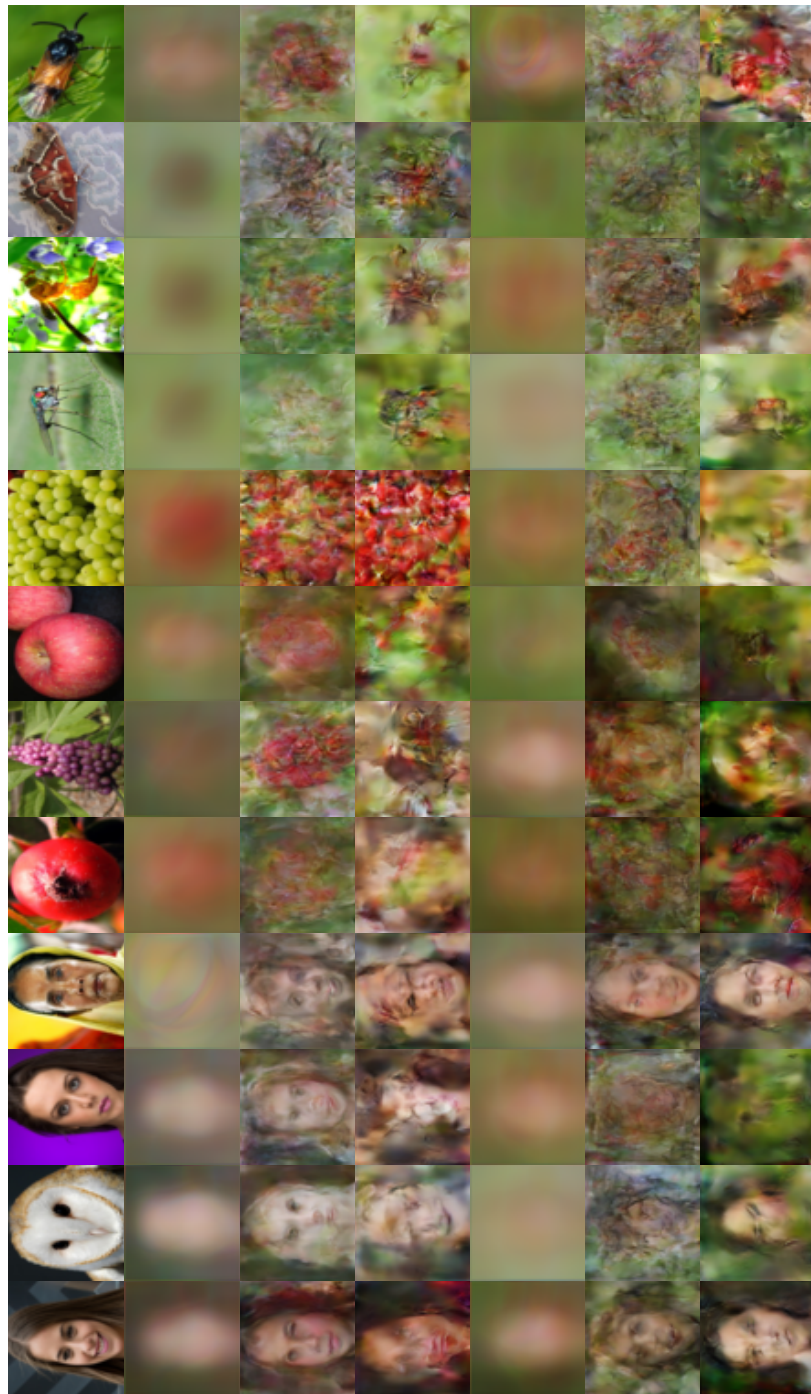
3.4 Experiments

3.4.1 Training

We used our ECoG dataset (see 2.4 for the details) to evaluate our image reconstruction models. For preprocessing ECoG signals, we first applied notch filtering (50 Hz) to remove line noise. Then, we normalized each trial’s data with the mean and standard deviation in the baseline period (from 500 to 200 ms before stimulus onset). For the input to reconstruction models, we used ECoG signals from 1 to 300 ms relative to the stimulus onset. We used images in three natural object classes (face, fruit, insect) for our experiments.

To test the ability to reconstruct novel natural images, we split prepared ECoG trials so that the validation and test sets do not include any image in the training set. In Subject 1’s dataset, the training, validation, and test set contains 6052, 2012, and 2023 trials, respectively. In Subject 2’s dataset, the training, validation, and test set contains 6110, 2035, 2035 trials, respectively.

To train our reconstruction models, we used downsampled 64×64 images as targets. Training was continued for 1,000 epochs with a batch size of 64. After every epoch of training, we measured the mean squared error (MSE) between reconstructions and presented images on the validation set. In the evaluation of each model on the test set, we used the best checkpoint with validation MSE over the training epochs. For optimization, we used the Adam optimizer [88] with a learning rate of



Ground truth

L1

L1-VGG-GAN

cGAN

S1

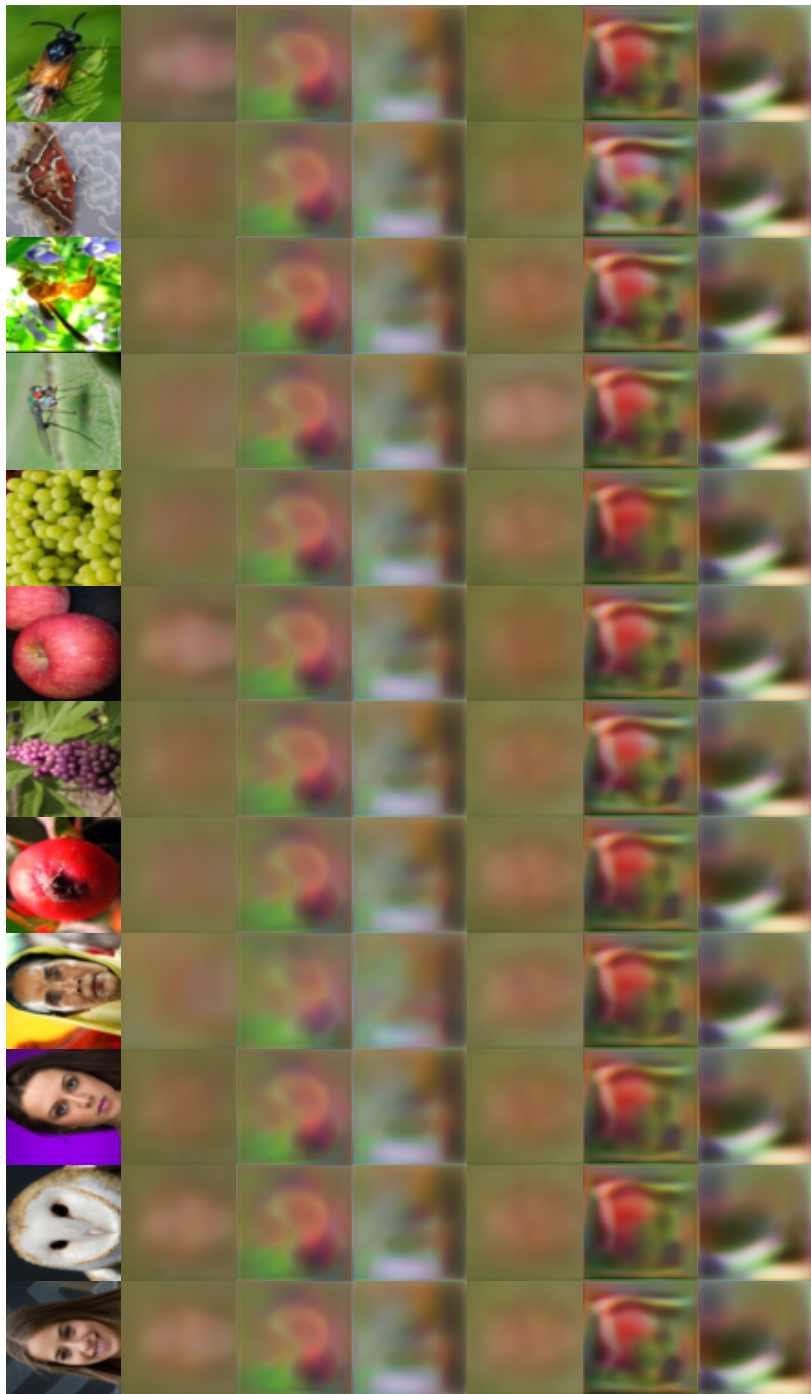
L1

L1-VGG-GAN

cGAN

S2

Figure 3-3: Example reconstructions for each subject and model. The first row shows ground truth images. The second to fourth rows show reconstruction results for Subject 1. The fifth to seventh rows show reconstruction results for Subject 2.



Ground truth

L1
 L1-VGG-GAN
 cGAN
 S1

L1
 L1-VGG-GAN
 cGAN
 S2

Figure 3-4: Example reconstructions with ECoG signals (downsampling width: 300 ms). The first row shows presented images (ground truth). The second to fourth rows show reconstruction results for Subject 1. The fifth to seventh rows show reconstruction results for Subject 2.

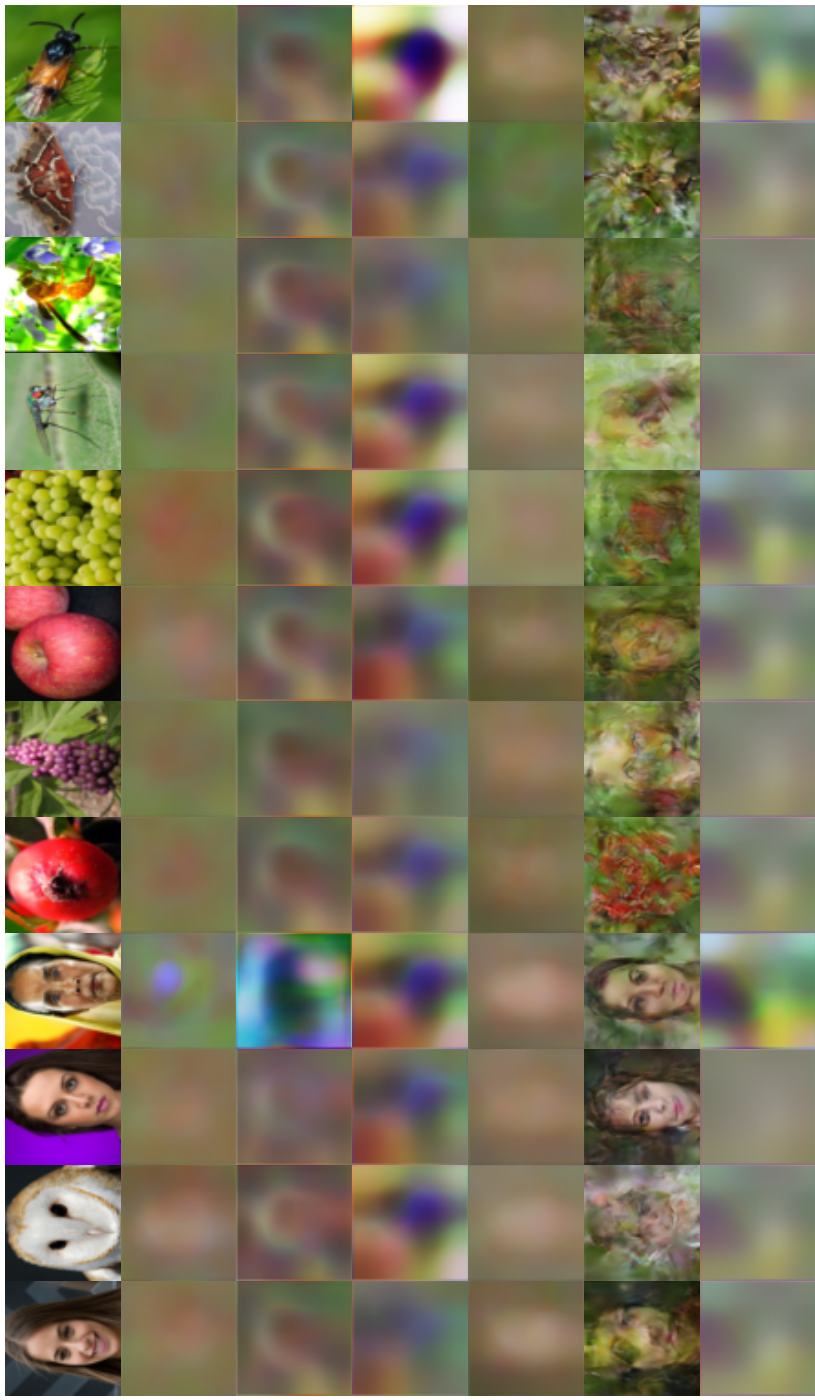


Figure 3-5: Example reconstructions with downsampled ECoG signals (downsampling width: 100 ms). The first row shows presented images (ground truth). The second to fourth rows show reconstruction results for Subject 1. The fifth to seventh rows show reconstruction results for Subject 2.

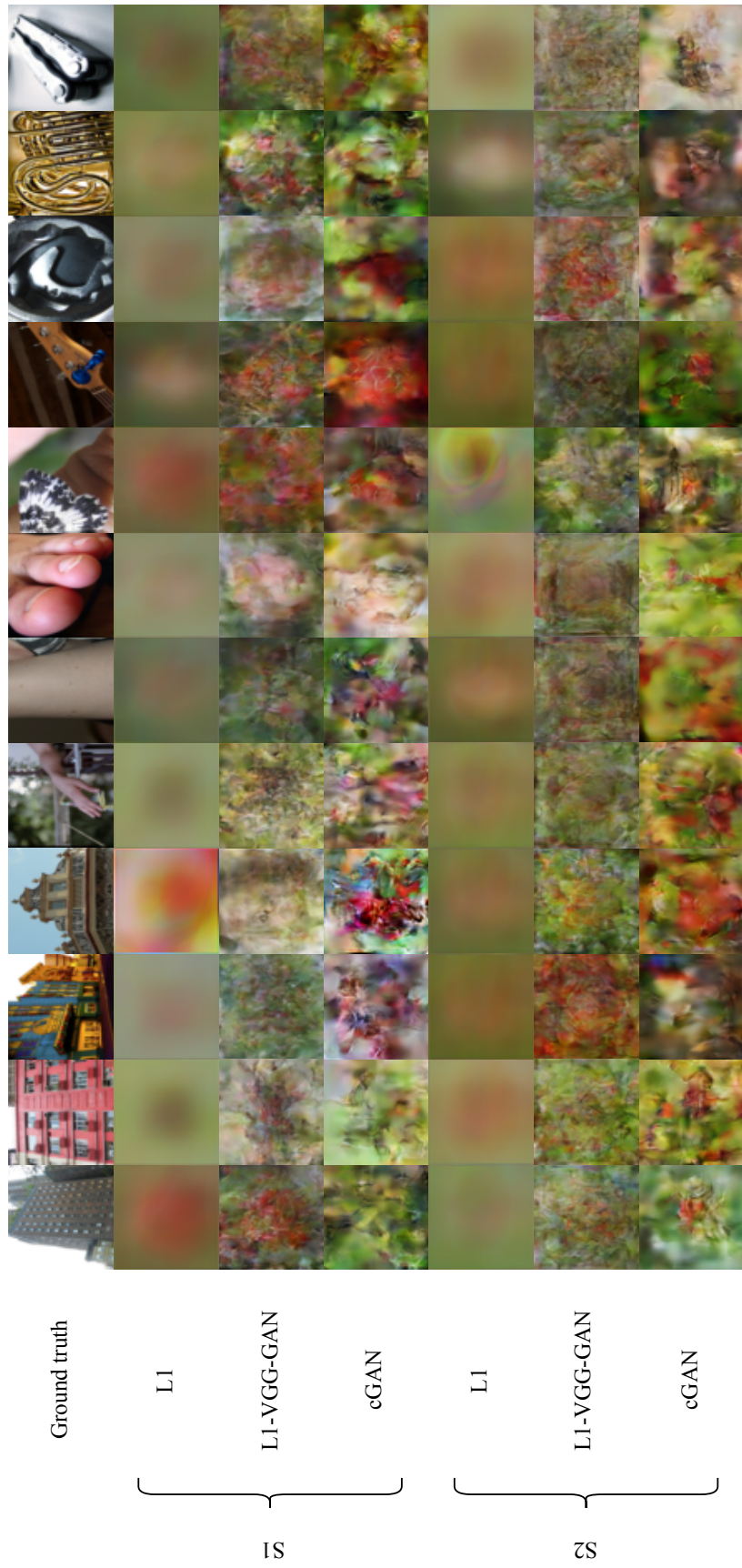


Figure 3-6: Example reconstructions for novel classes (building, body part, tool). The first row shows presented images (ground truth). The second to fourth rows show reconstruction results for Subject 1. The fifth to seventh rows show reconstruction results for Subject 2.

0.0001, $\beta_1 = 0$, and $\beta_2 = 0.9$. We used *PyTorch* [87] for implementing and conducting our experiments.

3.4.2 Reconstruction results

Figure 3-3 shows several example reconstructions and presented images for each subject and model. For both subjects, the reconstructions by the L1 models are significantly blurred and lack class- or object-specific attributes. On the other hand, the reconstructions by the L1-VGG-GAN and cGAN models contain more visible patterns that exist in the presented images. For successful examples, the L1-VGG-GAN and cGAN models produced class-specific patterns, such as human face (column 1-4 of Subject 1, column 1 and 4 of Subject 2), fruit color (column 6 and 7 of Subject 1), texture of small fruits (column 8 of Subject 1), blurred insect shape (column 9-12 of Subject 1 and 2).

To investigate the importance of rich temporal dynamics in ECoG signals for natural image reconstruction, we also trained reconstruction models with downsampled ECoG signals. Figure 3-4 shows several reconstruction results from ECoG signals with a downsampling width of 300 ms. Compared to results from original ECoG signals, these results lack class- or object-specific visual attributes compared to the original models, even for the L1-VGG-GAN or cGAN models. Similarly, Figure 3-5 shows several reconstruction results from ECoG signals with a downsampling width of 100 ms. In this case, the temporal resolution is slightly better, but each model failed to reconstruct perceptible visual patterns from ECoG signals. These results with downsampled ECoG signals indicate the importance of utilizing the rich temporal dynamics and high temporal resolution of ECoG signals for better natural image reconstruction.

We also tested the reconstruction models for novel classes that were not included in the training set (building, body part, tool) (Figure 3-6). Applicability to novel classes is a strength of regression-based decoding models over classification-based ones, which require re-training in this scenario. In our results, however, each model failed to successfully reconstruct visual information on novel classes compared to

reconstructions for trained classes.

3.4.3 Quantitative results

To quantitatively compare the reconstruction performance, we measured the peak signal-to-noise ratio (PSNR), the structural similarity index (SSIM) [122], and the Fréchet Inception Distance (FID) [123].

In computer vision, PSNR and SSIM are often used in studies on various image restoration tasks (e.g., single-image superresolution, image inpainting, deblurring, deraining) to measure the pixel-level similarity between ground truth and reconstructions. PSNR is defined as:

$$\text{PSNR} = 10 \cdot \log_{10} \left(\frac{\text{MAX}_I^2}{\text{MSE}} \right), \quad (3.11)$$

where MSE is the mean squared-error between reconstructions and ground truth:

$$\text{MSE} = \frac{1}{WH} \sum_{i,j} (x_{i,j} - \hat{x}_{i,j})^2. \quad (3.12)$$

SSIM incorporates structural information of local pixel groups (11×11 window) and the dynamic range of pixel values:

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}, \quad (3.13)$$

where μ_x and μ_y are the average of x and y , σ_x^2 and σ_y^2 are the variance of x and y , σ_{xy} is the covariance of x and y in the window. c_1 and c_2 are the variables to stabilize the division.

In addition to the above pixel-level distortion measures, we computed the Fréchet Inception Distance (FID) [123] to evaluate how photo-realistic reconstructions are. FID is a popular metric to evaluate the image generation quality on image synthesis tasks in computer vision. FID compares the statistics of two multivariate Gaussian

Table 3.3: Quantitative results of each subject and model. For PSNR and SSIM, higher values are better. For FID, lower values are better.

Loss	PSNR (\uparrow)		SSIM (\uparrow)		FID (\downarrow)	
	S1	S2	S1	S2	S1	S2
L1	11.93	11.96	0.2099	0.2047	330.8	331.2
L1-VGG-GAN	11.47	11.55	0.1292	0.1214	209.6	236.5
cGAN	10.92	11.04	0.1195	0.1174	196.4	192.5

distributions computed from ground truth and reconstructions:

$$d^2((\mu, \Sigma), (\hat{\mu}, \hat{\Sigma})) = \|\mu - \hat{\mu}\|_2^2 + \text{Tr}(\Sigma + \hat{\Sigma} - 2(\Sigma\hat{\Sigma})^{1/2}), \quad (3.14)$$

where (μ, Σ) and $(\hat{\mu}, \hat{\Sigma})$ are the mean and covariance of embedded samples from presented images and reconstructions, respectively. In this work, we used a pretrained Inception-v3 network for image embedding.

Table 3.3 shows quantitative results for each subject and model. For PSNR and SSIM, larger values are better. For FID, lower values are better. Comparing the models on the distortion-based evaluation metrics (PSNR, SSIM), the L1 models achieved better performance than the L1-VGG-GAN and cGAN models. However, in terms of the FID, which indicates the perceptual quality of reconstructions, the L1-VGG-GAN and cGAN models achieved better performance than the L1 models.

3.5 Discussion and Conclusion

In this work, we have conducted a large-scale experiment on natural image reconstruction from ECoG signals using deep learning. In successful cases, the L1-VGG-GAN and cGAN models produced reconstructions that contain various class- or object-specific visual attributes in presented images, suggesting that training reconstruction models with an adversarial loss is crucial to achieve better natural image reconstructions. Furthermore, our results with downsampled ECoG signals showed the impor-

tance of utilizing rich temporal dynamics in ECoG signals for better natural image reconstruction. In our experiments, we recorded ECoG signals from the macaque inferior temporal cortex (ITC). ITC is considered as the highest region in the ventral visual pathway. Although functional properties of neurons in the early visual cortex are relatively well investigated, those in the mid and higher visual cortex are still unclear. Therefore, it is notable that our results indicate the possibility of reconstructing diverse natural images from electrophysiological recordings of neuronal activities in ITC.

Although reconstructions by the L1-VGG-GAN and cGAN models were qualitatively better than the L1-based models, in our quantitative results, the L1-based models outperformed the other two models on the peak signal-to-noise ratio (PSNR) and the structural similarity index (SSIM), which are pixel-level distortion metrics. On the other hand, the L1-VGG-GAN and cGAN models outperformed the L1-based models on the Fréchet Inception Distance (FID), which compares the distribution of ground truth and reconstructions. Similar results have been observed in various image restoration tasks. For example, when evaluating with PSNR or SSIM, models trained only with a pixel-level loss (e.g., mean squared error: MSE) usually outperforms models trained with a combination of a pixel-level, perceptual, and adversarial losses [124, 125]. However, when evaluating with human opinion scores or the perceptual index [125], GAN-based models usually outperformed pixel-only models. We believe that the evaluation metric of image reconstruction models in brain decoding should be decided by the purpose of study. If models should reconstruct "accurate" images in terms of pixel values, they should be evaluated by pixel-level distortion metrics, such as PSNR and SSIM. On the other hand, if models should reconstruct more perceptible images, they should be evaluated by human opinion scores, the perceptual index, and image synthesis metrics (e.g., FID).

There are several limitation in our methods. First, in the training of our reconstruction models, it is assumed that brain signals of each trial reflect the presented image. Therefore, our methods are not directly applicable if models need to reconstruct imagined or perceived images, where explicit supervision is not easily available.

We believe that unsupervised or semi-supervised learning can be employed for more generic image reconstruction scenarios in brain decoding. Second, while our models reconstruct a single image from brain signals, each subject was continuously presented the image at every time step. Visual information in brain signals might depend on the time step. Therefore, to investigate what kind of visual information is contained at each time step, we need to train models that reconstruct a sequence of images from brain signals. This problem is also related to video reconstruction from brain signals [28].

Chapter 4

Deep Learning for Channel-Agnostic Brain Decoding across Multiple Subjects

4.1 Introduction

We can record complex spatiotemporal responses from the brain while the subject perceives or imagines something, using an electric or magnetic recording technique such as electroencephalography (EEG), magnetoencephalography (MEG), and electrocorticography (ECoG). The goal of brain decoding is to read out what was perceived or imagined from brain signals. Accurate decoding of motor states is crucial for creating practical BCI systems in the real-world. [126, 98, 127]. Furthermore, developing better brain decoding methods helps researchers investigate what kind of features are related to brain signals, by evaluating how well each decoding model predicts specific perceptual information from brain signals [128].

While a number of studies have proposed various decoding methods for brain signals [129, 130], most existing methods consider only static, single-subject cases, where a decoder is trained independently for each subject's dataset with the same recording equipment. In BCI applications, long calibration time and overly repeated

recording trials are painful for patients; thus, decoders are desired to be transferable to novel patients and conditions. For cognitive science, across-subject decoding analyses are useful when the number of trials for each subject is limited.

In the literature, various methods for across-subject decoding have been proposed, such as common spatial patterns (CSPs) [131, 132, 29, 133] and transfer learning [134, 31, 135]. However, few studies have investigated decoding methods that are robust to the shift of recording channels. In practice, it is tough to record brain signals with exactly the same equipment and conditions from a large number of subjects or from a subject over a long period. Moreover, even with the same equipment, channel locations or conditions can change in each session, especially when the recording requires breaks and/or repeated removals of the electrode. Also, if a decoder accepts only one fixed number of input channels, it is not applicable to novel subjects' dataset that have a different number of channels. Therefore, developing channel-agnostic decoding methods is crucial for creating scalable and transferable BCI systems.

In this work, we study brain decoding across multiple subjects with a different number of recording channels and channel location shifts. We consider channel-agnostic brain decoding as a multi-instance learning problem [136, 34]. In multi-instance learning, each input is considered as a set of independent instances (bag), and the task is considered as a weakly supervised learning problem where only one label is annotated for each entire bag. By using a multi-instance pooling operator, models can aggregate informative features over a variable number of input instances. This formulation naturally fits into channel-agnostic brain decoding, where the goal is to train better performing decoders that are robust to the change of the number and the location of recording channels.

Based on the multi-instance learning formulation, we propose a novel channel-agnostic decoder architecture with three building blocks. The first block, channel-wise feature extraction, uses a channel-wise version of temporal convolutional networks (TCNs) [117, 118], which applies shared 1D convolution kernels independently for each channel. The second block, across-channel transform, uses recently proposed multi-head self-attention [137] to model inter-channel interactions with channel

permutation invariance. The third block, multi-channel pooling aggregates features across a variable number of channels.

We conducted a thorough experiment to verify the design of our proposed decoder architecture in channel-agnostic brain decoding across multiple subjects. Our dataset has ECoG signals recorded from two subjects with a different number of channels and inconsistent channel locations. We trained our proposed models and baselines to predict six visual object classes from each subject’s single-trial data. Our results indicate the importance of using across-channel transforms with channel permutation invariance and inter-channel interactions for achieving better classification results in channel-agnostic brain decoding across multiple subjects.

4.2 Related work

4.2.1 Multi-subject decoding for fMRI data

Independent component analysis (ICA) is a popular method in fMRI analysis. For analysing common spatial properties of fMRI data across multiple subjects, *group ICA* [138] can be used. There are a variety of approaches for group ICA, such as aggregation of independent, single-subject ICA results, temporal concatenation, spatial concatenation, and tensor decomposition. Each approach has specific assumptions and drawbacks.

Processing raw fMRI data across multiple subjects requires either anatomical alignment or functional alignment. Anatomical alignment uses several anatomical landmarks to align each subject’s fMRI images, but its effectiveness is limited by the diversity of the size, shape and anatomical location of across subjects. *Hyperalignment* [139] is a functional alignment method that uses higher-order, orthogonal transformation, which can be formulated as a multi-set *canonical correlation analysis* (CCA). Hyperalignment and its extensions were used for multi-subject fMRI analyses (e.g., inter-subject classification).

Other studies proposed transfer learning [140] and domain adaptation [141] for

multi-subject fMRI decoding tasks.

4.2.2 Multi-subject decoding for EEG signals

Most existing methods for multi-subject EEG decoding employ common spatial patterns (CSPs) on top of the decoding model [131, 132, 29, 133]. Fazil *et al.* [29] proposed a method for subject-independent motor imagery classification by estimating subject-specific filters and subject-independent classifiers. CSP-based subject transfer is a popular method for avoiding calibration or fine-tuning for novel subjects [142]. However, CSP-based methods require the recording equipment use the same number of channels and similar channel locations, unless an channel interpolation method is used to impute missing channels.

Other scenarios on across-subject EEG decoding have been also studied, such as unsupervised domain adaptation [143], transfer learning [31, 135], and multi-subject decoding [30].

4.2.3 Open problems and the purpose of this work

Most existing methods for multi-subject brain decoding assume that the number of channels and channel locations are consistent across subjects; Few studies have investigated decoding methods that are robust to the shift of recording channels. Interpolation methods for imputing missing channels require the exact location and topography of channels, This limits the application of decoding methods for SUA/MUA/LFP, which records neuronal activities *in* the brain, and ECoG, which often involves channels on the gyral and sulcus. In this work, we propose a novel decoder architecture that can handle a variable number of channels, has permutation invariance to the order of input channels, and can capture inter-channel relationships. We prepared an ECoG dataset where the number and location of channels are not consistent across multiple subjects. To investigate the effectiveness of our proposed methods, we compared with other deep learning-based baselines.

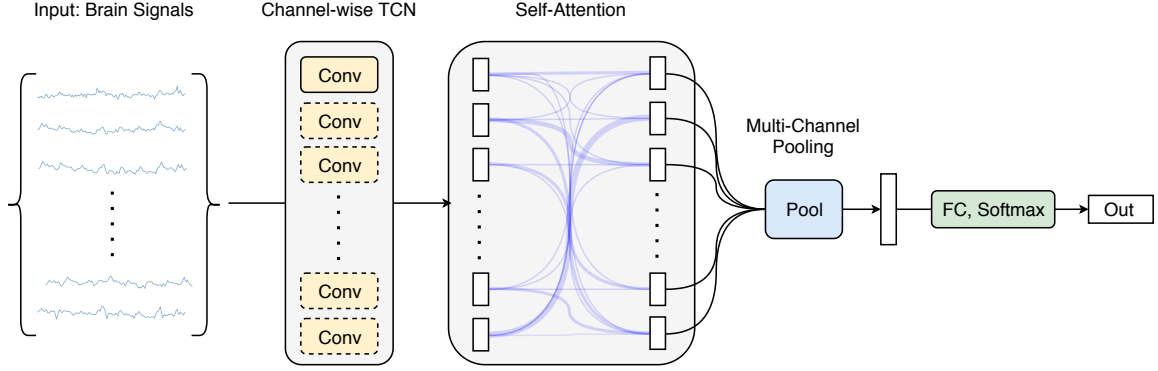


Figure 4-1: Proposed decoder architecture based on channel-wise temporal convolutional networks (TCNs), across-channel self-attention, and multi-channel pooling.

4.3 Methods

4.3.1 Channel-agnostic brain decoding as multi-instance learning

Suppose we have training examples from two subjects: $\mathcal{D}_A = \{(X_1^{(A)}, y_1), \dots, (X_{N_A}^{(A)}, y_{N_A})\}$ and $\mathcal{D}_B = \{(X_1^{(B)}, y_1), \dots, (X_{N_B}^{(B)}, y_{N_B})\}$. Here, $X^A \in \mathbb{R}^{C_A \times T}$ and $X^B \in \mathbb{R}^{C_B \times T}$ are preprocessed brain signals of each subject, where the number of channels (C_A, C_B) is not consistent between the subjects.

In this scenario, decoders need to be designed so that they can handle a variable number of channels. Multi-instance learning [136, 34] considers each input as a set (bag) which contains a variable number of instances; thus, naturally fits this situation. Most existing methods in multi-instance learning first independently extract instance-level features, and then pool these features to get bag-level features, followed by bag-level transformations:

$$S(X) = g(\text{pool}_{x \in X} f(x)), \quad (4.1)$$

where f is instance-level transformations; pool is multi-instance pooling; and g is bag-level transformations. Here, instance-level transformations and multi-instance

pooling should be invariant to the order of instances [144], because we need to optimize the order if they are not permutation-invariant.

To construct channel-agnostic brain decoders, we consider instances in multi-instance learning as channels in brain decoding. In the following, we will describe how we design channel-level transformations and across-channel pooling for effective decoding.

Because the number of channels for one subject the subject, the model needs to be designed so that it can handle a variable number of input channels. To tackle this, we propose a two-stage architecture (see Figure 4-1 for diagram). The first stage f is channel-wise transformation $\{x^{(1)}, \dots, x^{(C)}\}$, resulting in features for each channel $\{f(x^{(1)}; \theta_{(f)}), \dots, f(x^{(C)}; \theta_{(f)})\}$. Then, the second stage g aggregates features over the channels using a multi-instance pooling function:

$$S(X) = g\left(\sum_{x \in X} f(x)\right). \quad (4.2)$$

4.3.2 Channel-wise transform

In our proposed architecture, channel-wise transformations are grouped into two blocks. The first block is a series of channel-wise temporal convolutional networks (TCNs) [117, 118]. To extract informative features from input brain signals, this block applies shared 1D convolution kernels independently for each channel.

The channel-wise TCN has one 1D convolution layer and six residual blocks [117, 118]. The first convolution layer embeds each channel’s brain signals $x_c \in \mathbb{R}^T$ into hidden features $h_c \in \mathbb{R}^{d \times T}$, with the embedding size d of 16 in our experiments. Each residual block has two dilated convolution layers, each of which follows layer normalization [145] and ReLU activation (i.e. preactivation residual blocks [146]). The dilation factor and kernel size of each block’s convolution layers are 2 and 5, respectively. All the convolution layers and layer normalization are conducted independently for each channel. For regularization, we apply dropout [147] after each ReLU activation, with a probability of 0.3.

4.3.3 Across-channel transform

The second block is channel-level self-attention [137] for capturing inter-channel interactions. This deviates from traditional multi-instance learning methods, which assume instances are independent. When recording signals, nearby channels might be correlated each other. Moreover, recent studies suggest that various forms of interactions exist even between distant channels [148, 149, 150, 151]. Therefore, we hypothesize that considering inter-channel interactions is beneficial for effective brain decoding.

After embedding each channel’s brain signals with the channel-wise TCN, we use multi-head self-attention [137] for capturing inter-channel interactions. In multi-head self-attention, the outputs of each head’s self-attention are concatenated and transformed:

$$\text{Multihead}(Q, K, V) = \text{concat}(O_1, \dots, O_L) \mathbf{W}^O \quad (4.3)$$

$$O_l = \text{Attention}(Q \mathbf{W}_l^Q, K \mathbf{W}_l^K, V \mathbf{W}_l^V), \quad (4.4)$$

where $\mathbf{W}_l^O \in \mathbb{R}^{L d_{head} \times d}$, $\mathbf{W}_l^Q \in \mathbb{R}^{d \times d_{head}}$, $\mathbf{W}_l^K \in \mathbb{R}^{d \times d_{head}}$, and $\mathbf{W}_l^V \in \mathbb{R}^{d \times d_{head}}$ are respectively the matrices for output, query, key, and value projection. In our experiments, we employ scaled dot-product attention [137]:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_{head}}}\right)V, \quad (4.5)$$

where attention values of the softmax is computed over channels, resulting in a channel-by-channel attention matrix for feature aggregation.

4.3.4 Multi-channel pooling

After channel-wise feature extraction, we get a bag of channel-wise embeddings $H = \{h_1, \dots, h_C\}$, where $h_c = f(\mathbf{x}_c) \in \mathbb{R}^K$. In across-channel pooling, the bag needs to be aggregated to a fixed-size vector. We consider three pooling methods that are

proposed in the literature of multi-instance learning: mean, max, and attention-based pooling.

The mean operator have been popularly used in multi-instance learning [136, 34]. The mean operator simply takes the average over the channels:

$$z = \frac{1}{C} \sum_{c=1}^C f(x_c), \quad (4.6)$$

while the max operator takes the maximum over the channels for each element:

$$z^k = \max_{c=1, \dots, C} \{h_c^k\}, \quad k = 1, \dots, K. \quad (4.7)$$

The attention-based pooling is a recently-proposed method [152]. Attention-based methods have come into widespread use in natural language processing [153, 154] and computer vision [155]. The attention-based pooling aggregates channel-level embeddings by taking the weighted summation as:

$$z = \sum_{c=1}^C a_c f(x_c), \quad (4.8)$$

where a_c is the attention weight for c-th channel. Each attention weight is computed as:

$$a_c = \text{softmax}(\mathbf{w}^\top \tanh(\mathbf{V} f(x_c)^\top)). \quad (4.9)$$

In the attention-based pooling, attention weights are computed for each input, and the parameters , \mathbf{w} and \mathbf{V} , are trained with the dataset.

4.3.5 Baselines

To verify the effectiveness of self-attention in channel-agnostic brain decoding, we compare the performance against two alternative methods: channel-wise feed-forward and bi-directional long short-term memory networks (Bi-LSTM) [156, 157].

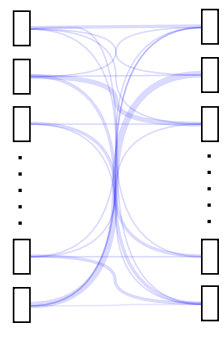
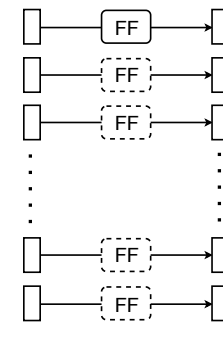
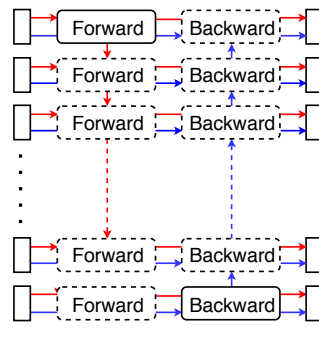
	Multi-Head Self-Attention	Channel-wise Feed-Forward	Bidirectional LSTM
Transform			
Permutation Invariance	✓	✓	✗
Inter-Channel Interactions	✓	✗	✓

Figure 4-2: Comparison of each across-channel transform modules.

In channel-wise feed-forward networks (FFN), two fully-connected layers are applied independently for each channel's features:

$$\text{FFN}(X) = \max(0, X\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2, \quad (4.10)$$

where $\mathbf{W}_1 \in \mathbb{R}^{d \times d_{\text{ff}}}$ and $\mathbf{W}_2 \in \mathbb{R}^{d_{\text{ff}} \times d}$. While FFN can handle a variable number of input channels and the output of this method does not depend on the order of channels, there is no interaction between channels.

Bi-LSTM is the bi-directional version of LSTM [158, 159]. The forward path of

Bi-LSTM at channel c is computed as follows:

$$\begin{aligned}
\mathbf{i}_{\text{fwd},c} &= \text{sigmoid}(\mathbf{W}_{\text{fwd}}^{(ii)}\mathbf{x}_c + \mathbf{b}_{\text{fwd}}^{(ii)} + \mathbf{W}_{\text{fwd}}^{(hi)}\mathbf{h}_{\text{fwd},c-1} + \mathbf{b}_{\text{fwd}}^{(hi)}) \\
\mathbf{f}_{\text{fwd},c} &= \text{sigmoid}(\mathbf{W}_{\text{fwd}}^{(if)}\mathbf{x}_c + \mathbf{b}_{\text{fwd}}^{(if)} + \mathbf{W}_{\text{fwd}}^{(hf)}\mathbf{h}_{\text{fwd},c-1} + \mathbf{b}_{\text{fwd}}^{(hf)}) \\
\mathbf{g}_{\text{fwd},c} &= \text{tanh}(\mathbf{W}_{\text{fwd}}^{(ig)}\mathbf{x}_c + \mathbf{b}_{\text{fwd}}^{(ig)} + \mathbf{W}_{\text{fwd}}^{(hg)}\mathbf{h}_{\text{fwd},c-1} + \mathbf{b}_{\text{fwd}}^{(hg)}) \\
\mathbf{o}_{\text{fwd},c} &= \text{sigmoid}(\mathbf{W}_{\text{fwd}}^{(io)}\mathbf{x}_c + \mathbf{b}_{\text{fwd}}^{(io)} + \mathbf{W}_{\text{fwd}}^{(ho)}\mathbf{h}_{\text{fwd},c-1} + \mathbf{b}_{\text{fwd}}^{(ho)}) \\
\mathbf{c}_{\text{fwd},c} &= \mathbf{f}_{\text{fwd},c} * \mathbf{c}_{\text{fwd},c-1} + \mathbf{i}_{\text{fwd},c} * \mathbf{g}_{\text{fwd},c} \\
\mathbf{h}_{\text{fwd},c} &= \mathbf{o}_{\text{fwd},c} * \tanh(\mathbf{c}_{\text{fwd},c}),
\end{aligned} \tag{4.11}$$

where $\mathbf{h}_{\text{fwd},c}$ and $\mathbf{c}_{\text{fwd},c}$ are respectively the hidden state and cell state at channel c , and $\mathbf{i}_{\text{fwd},c}$, $\mathbf{f}_{\text{fwd},c}$, $\mathbf{g}_{\text{fwd},c}$, and $\mathbf{o}_{\text{fwd},c}$ are respectively input, forget, cell, and output gates.

The backward path at time t is computed as follows:

$$\begin{aligned}
\mathbf{i}_{\text{bwd},c} &= \text{sigmoid}(\mathbf{W}_{\text{bwd}}^{(ii)}\mathbf{x}_c + \mathbf{b}_{\text{bwd}}^{(ii)} + \mathbf{W}_{\text{bwd}}^{(hi)}\mathbf{h}_{\text{bwd},c+1} + \mathbf{b}_{\text{bwd}}^{(hi)}) \\
\mathbf{f}_{\text{bwd},c} &= \text{sigmoid}(\mathbf{W}_{\text{bwd}}^{(if)}\mathbf{x}_c + \mathbf{b}_{\text{bwd}}^{(if)} + \mathbf{W}_{\text{bwd}}^{(hf)}\mathbf{h}_{\text{bwd},c+1} + \mathbf{b}_{\text{bwd}}^{(hf)}) \\
\mathbf{g}_{\text{bwd},c} &= \text{tanh}(\mathbf{W}_{\text{bwd}}^{(ig)}\mathbf{x}_c + \mathbf{b}_{\text{bwd}}^{(ig)} + \mathbf{W}_{\text{bwd}}^{(hg)}\mathbf{h}_{\text{bwd},c+1} + \mathbf{b}_{\text{bwd}}^{(hg)}) \\
\mathbf{o}_{\text{bwd},c} &= \text{sigmoid}(\mathbf{W}_{\text{bwd}}^{(io)}\mathbf{x}_c + \mathbf{b}_{\text{bwd}}^{(io)} + \mathbf{W}_{\text{bwd}}^{(ho)}\mathbf{h}_{\text{bwd},c+1} + \mathbf{b}_{\text{bwd}}^{(ho)}) \\
\mathbf{c}_{\text{bwd},c} &= \mathbf{f}_{\text{bwd},c} * \mathbf{c}_{\text{bwd},c+1} + \mathbf{i}_{\text{bwd},c} * \mathbf{g}_{\text{bwd},c} \\
\mathbf{h}_{\text{bwd},c} &= \mathbf{o}_{\text{bwd},c} * \tanh(\mathbf{c}_{\text{bwd},c}),
\end{aligned} \tag{4.12}$$

The output of Bi-LSTM is the concatenation of forward and backward paths:

$$\mathbf{h} = [\mathbf{h}_{\text{fwd},C}, \mathbf{h}_{\text{bwd},1}], \tag{4.13}$$

where C is the number of channels. Bi-LSTM can handle a variable number of input channels, and also capture inter-channel interactions in forward and backward computations. However, its results crucially depend on the order of inputs [160, 144]; thus not permutation-invariant.

4.4 Experiments

To evaluate the effectiveness of our proposed methods on channel-agnostic brain decoding, we conducted a thorough experiment on a multi-subject classification task. Models receive a subject’s trial data and predict one class from six visual object classes. Each subject’s dataset has a different number of recording channels, and their locations are also different.

We used our ECoG dataset (see 2.4 for the details) in our experiments. For Subject 2’s dataset, we used additional 64 channels on the prefrontal cortex (PFC), resulting in an inconsistent number of channels and channel locations between the subjects. For preprocessing ECoG signals, we first applied notch filtering (50 Hz) to remove line noise. Then, we normalized each trial’s data with the mean and standard deviation in the baseline period (from 500 to 200 ms before stimulus onset). For the input to reconstruction models, we used ECoG signals from 1 to 300 ms relative to the stimulus onset. We used all the six object classes (building, body part, face, fruit, insect, tool) for our experiments.

4.4.1 Training

In Subject 1’s dataset, the training, validation, and test set contains 12,149, 4,039, and 4,049 trials, respectively. In Subject 2’s dataset, the training, validation, and test set contains 12,204, 4,065, and 4,065 trials, respectively. To test the classification ability in novel situations, the validation and test sets do not include any image in the training set.

For a fair comparison, all models are implemented so that they have a similar parameter size. To adjust the parameter size over the multi-instance pooling methods, we add one additional fully-connected layer before mean or max pooling. We implemented and conducted our experiments with *PyTorch* [87].

We trained each model for 100 epochs with a batch size of 128. We used the Adam optimizer [88] for optimizing model parameters, with a learning rate of 0.0001,

Table 4.1: Comparison of classification accuracy between three across-channel transform modules. The results with the best multi-channel pooling function are shown for each transform module. For each model, we conducted eight runs with different weight initialization, and the average and standard error of classification accuracy over the eight runs are reported here.

Transform	Parameters	Accuracy		
	(10^6)	Mean	(S1)	(S2)
Self-Attention (Mean)	0.451	32.19 \pm 0.53	36.80 \pm 0.51	27.59 \pm 0.73
Feed-Forward (Max)	0.451	25.39 \pm 0.70	28.26 \pm 0.77	22.52 \pm 0.75
Bi-LSTM (Max)	0.485	28.98 \pm 0.63	32.44 \pm 0.58	25.52 \pm 0.77

a weight decay of 0.0001, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. After the training of 50-th epoch is finished, the learning rate is annealed with a factor of 0.5. We ran each model’s training for 8 times with different random seeds, and compare the models with the average and standard error over the 8 runs.

4.4.2 Classification results: Across-channel transform

First, we compare the difference of the classification accuracy between the three across-channel transform modules: multi-head self-attention, channel-wise feed-forward, and bidirectional LSTM.

Table 4.1 shows each model’s across-channel transformations, parameter size, and the classification accuracy (mean over the subjects, each subject). For each transformation, the results with the best multi-channel pooling function for each model are shown here. Although the parameter size of the Bi-LSTM-based models are slightly larger than the self-attention-based models, the self-attention-based module outperformed the other two modules on both the cross-subject mean and individual accuracy. These results indicate the importance of designing channel-level transformation so that they are permutation-invariant and can capture inter-channel relationships for effective channel-agnostic brain decoding. It is noteworthy that, even without any

Table 4.2: Comparison of classification accuracy between multi-channel pooling functions. The results for the multi-head self-attention module are shown here. For each model, we conducted eight runs with different weight initialization, and the average and standard error of classification accuracy over the eight runs are reported here.

Pooling	Parameters		Accuracy	
	(10^6)	Mean	(S1)	(S2)
Mean	0.451	32.19 \pm 0.53	36.80 \pm 0.51	27.59 \pm 0.73
Max	0.451	31.31 \pm 0.16	36.01 \pm 0.45	26.62 \pm 0.32
Attention	0.451	31.53 \pm 0.43	36.33 \pm 0.49	26.73 \pm 0.68

explicit spatial information of recording channels, self-attention-based models learned to capture inter-channel relationships and achieved far better performance than the chance level.

The comparison between the channel-wise feed-forward and Bi-LSTM also shows the importance of capturing inter-channel relationships in the across-channel transform. While the feed-forward module has the permutation invariance property, each channel-level features are transformed independently in the module. On the other hand, the Bi-LSTM module can capture inter-channel relationships thorough its recurrent processing in the forward and backward directions, although the output is not permutation invariant.

4.4.3 Classification results: Multi-channel pooling

Next, we compare the difference of the classification accuracy between the three multi-channel pooling functions: mean, max, and attention-based pooling.

Table 4.2 shows each self-attention-based model’s pooling function, parameter size, and the classification accuracy (mean over the subjects, each subject). In a recent study on multi-instance learning [152], the attention-based pooling was shown to be superior to other traditional pooling functions. However, in our results with

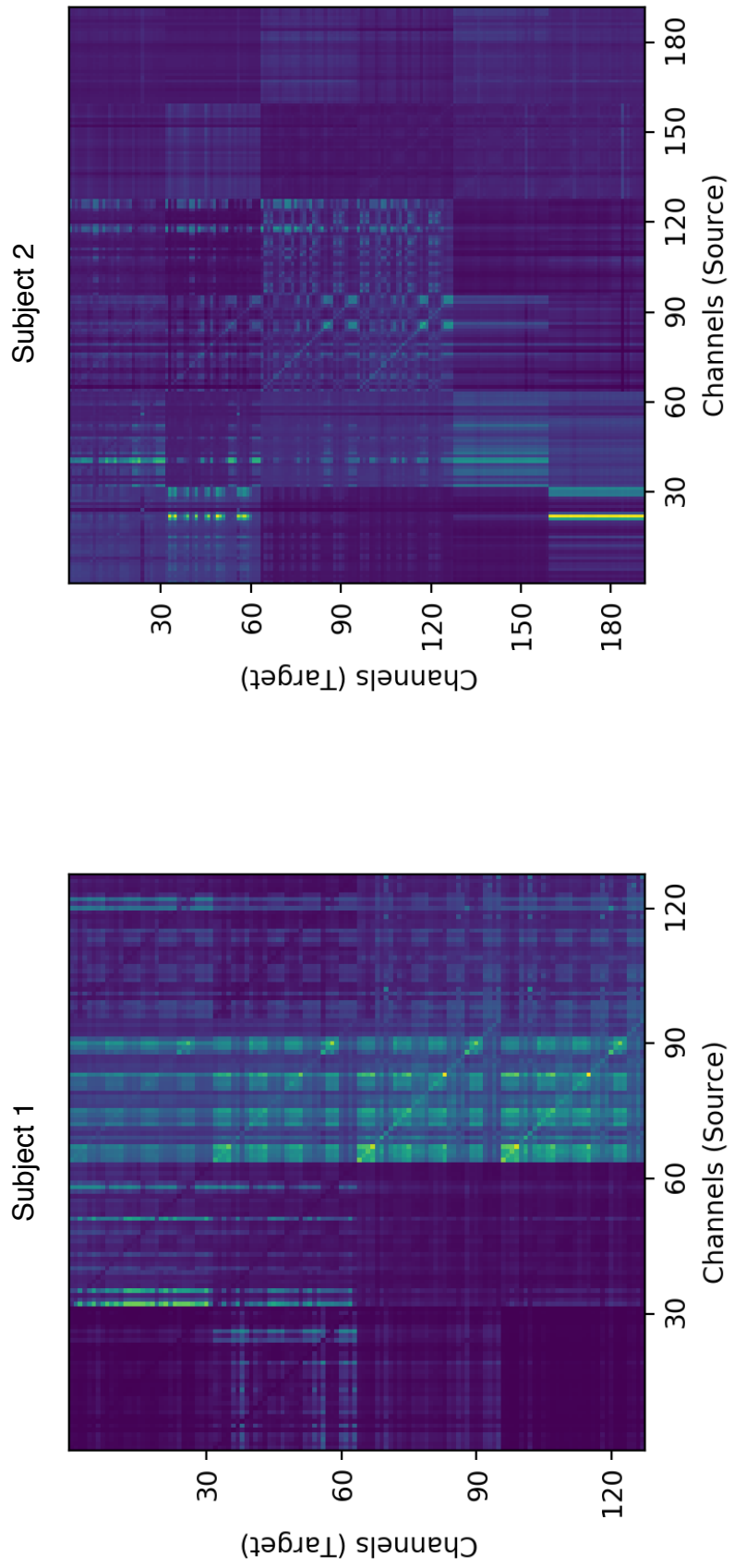


Figure 4-3: Visualized self-attention weights extracted from a trained model (transform: self-attention, pool: mean). The numbers on the horizontal and vertical axes indicate channel indices in each subject's ECoG electrode. For Subject 1, all 128 channels were implanted on the inferior temporal cortex. For Subject 2, channels 1-128 were implanted on the inferior temporal cortex, and channels 129-192 were implanted on the prefrontal cortex.

self-attention-based transform, the mean pooling is slightly better than max and attention-based pooling models.

4.4.4 Visualization of self-attention weights

To investigate whether our proposed architecture learned to capture inter-channel relationships, we visualized the attention matrix in self-attention. Figure 4-3 shows self-attention weights from the best performing model (transformation: self-attention, pooling: mean) for each subject. These self-attention weights are averaged over the test trials.

For Subject 1 (ITC: 1-128), most channels attended a portion of the posterior channels (channel ids around 60-90). On the other hand, a group of anterior channels (1-30) attended a few, more posterior channels (around 30). These results indicate that, even only in the inferior temporal cortex, the form of inter-channel relationships can be diverse.

For Subject 2 (ITC: 1-128, PFC: 129-192), several posterior-ITC channels (between 20 and 30) are strongly attended by groups of PFC channels; thus, indicating the importance of capturing inter-region relationships for effective brain decoding.

4.5 Discussion and Conclusion

In this work, we have studied brain decoding methods in a scenario where decoders need to adapt to a variable number of recording channels and channel location shifts across multiple subjects. Towards robust brain decoding in the scenario, we have proposed a novel decoder architecture based on the multi-instance learning framework. Our proposed architecture has three building blocks: channel-wise transform, across-channel transform, and multi-channel pooling. We proposed to use multi-head self-attention in the across-channel transform block to achieve permutation invariance and to capture inter-channel relationships for better decoding performance. We conducted a thorough experiment on the visual object classification from ECoG signals, where the number of channels and channel locations were not consistent across the

subjects. In our results, our self-attention-based across-channel transform outperformed the baselines, which lack either permutation invariance or the ability to capture inter-channel relationships. Our results suggest the importance of permutation invariance and inter-channel relationships for achieving better decoding performance in channel-agnostic tasks across multiple subjects. Furthermore, our visualization results of self-attention weights suggest intriguing properties about inter-channel relationships in visual perception.

Chapter 5

Conclusions

In this thesis, we have studied methods for encoding and decoding brain signals using deep learning, towards (1) understanding the relationship between complex brain activities and diverse visual features and (2) developing practical brain encoding and decoding methods for real-world brain-computer interface (BCI) tasks. We prepared a large-scale electrocorticography (ECoG) dataset by recording brain signals from macaque inferior temporal cortex while presenting visual stimuli to the subjects. We analyzed complex temporal properties of ECoG signals by developing an encoding analysis framework using optimized hierarchical visual features extracted from deep convolutional neural networks (CNNs). We also proposed advanced, flexible decoding methods based on state-of-the-art methods in deep learning.

In Chapter 2, we conducted an experiment on encoding frequency-specific ECoG signals using hierarchical visual features from pretrained convolutional neural networks (CNNs). We found that two different frequency bands, theta and gamma bands, are more related to visual features than the other bands. We also found that these two bands carry complementary features in terms of visual abstraction. While theta-band activities showed selectivity for higher-layers in CNNs, gamma-band activities showed selectivity for lower-layers. Our results suggest that neuronal oscillatory activities in theta and gamma bands carry distinct information in the hierarchy of visual features, and that distinct levels of visual information are multiplexed in frequency-specific brain signals. Furthermore, combining our results with previous

studies on frequency-specific roles in inter-areal communications, it could be possible that theta and gamma bands carry distinct visual information for different roles in inter-areal communications in the visual cortex.

In Chapter 3, we conducted an experiment on natural image reconstruction from ECoG signals using deep learning. To investigate what kind of models are effective for reconstructing photo-realistic natural images from brain signals, we considered three loss functions: L1 loss, VGG (a.k.a. perceptual) loss, and generative adversarial networks (GANs). To compare the impact of each loss function, we trained and evaluated three reconstruction models: L1, L1-VGG-GAN, and conditional GAN (cGAN). The L1 model was trained only with pixel-wise errors. The L1-VGG-GAN model was trained with a weighted combination of L1 loss, perceptual loss based a pretrained VGG network, and adversarial loss (generative adversarial network: GAN). The cGAN model was trained with the conditional version of GAN loss. In our results, while the L1-based models achieved better performance in terms of the pixel-level distortion metrics (peak signal-to-noise ratio: PSNR, structural similarity index: SSIM), the L1-VGG-GAN and cGAN models produced far better reconstructions in terms of perceptual quality (Fréchet Inception Distance: FID). In successful cases, the L1-VGG-GAN and cGAN models produced reconstructions that contain various class- or object-specific visual attributes in presented images, suggesting that training reconstruction models with an adversarial loss is crucial to achieve better natural image reconstructions. Furthermore, our results with downsampled ECoG signals showed the importance of utilizing rich temporal dynamics in ECoG signals for better natural image reconstruction. In our experiments, we recorded ECoG signals from the macaque inferior temporal cortex (ITC). ITC is considered as the highest region in the ventral visual pathway. Although functional properties of neurons in the early visual cortex are relatively well investigated, those in the mid and higher visual cortex are still unclear. Therefore, it is notable that our results indicate the possibility of reconstructing diverse natural images from electrophysiological recordings of neuronal activities in ITC.

In Chapter 4, we conducted an experiment on deep, multi-instance learning for

channel-agnostic brain decoding across multiple subjects. Towards robust brain decoding in the scenario, we consider the task from the view of multi-instance learning. We proposed a novel brain decoder architecture based on three building blocks: channel-wise transform, across-channel transform, and multi-channel pooling. Considering the physiological properties of multi-channel brain signals, we proposed to use multi-head self-attention in the across-channel transform block to achieve permutation invariance and to capture inter-channel relationships for better decoding performance. We conducted a thorough experiment on the visual object classification from ECoG signals, where the number of channels and channel locations were not consistent across the subjects. Our results showed that our self-attention-based across-channel transform outperformed the baselines, which lack either permutation invariance or the ability to capture inter-channel relationships, suggesting the effectiveness of our proposed architecture for achieving better decoding performance in channel-agnostic tasks across multiple subjects. Furthermore, our visualization results of self-attention weights suggest intriguing properties about inter-channel relationships in visual perception. We believe that our novel formulation of channel-agnostic brain decoding and proposed architecture lead to larger scale analyses using diverse brain recordings and more robust, useful decoding methods for real-world BCI applications.

Overall, our results indicate that state-of-the-art deep learning methods are invaluable tools for understanding neuronal representations in rich temporal dynamics of brain signals, decoding complex visual patterns from brain signals, and developing practical decoding methods applicable for multiple subjects. As the field of deep learning gets matured, it has been advocated that neuroscience should inspire deep learning again [161, 162]. It is noteworthy that, before the success of deep learning in artificial intelligence (AI), results of experimental and computational neuroscience inspired the development of deep learning. For example, basic building blocks in CNNs, convolution and pooling, are inspired from simple and complex cells in the primary visual cortex [41, 42]. Furthermore, the early motivation of studying neural networks is to develop computational models of neurons in the brain. Recently, as the effectiveness of deep learning in diverse AI tasks shown, deep learning has been popularly

used for analyzing brain activities. Thus, deep learning does have inspired cognitive neuroscience. The development of deep learning helps cognitive neuroscience in three aspects. First, as the performance of deep neural networks in diverse cognitive tasks is improved, researchers get an access to better computational models of cognition, because *good* models should achieve similar cognitive performance as the brain. Second, the capacity of learning rich representations from data helps researchers develop better-performing encoding and decoding models for complex spatiotemporal brain activities. Third, the flexibility of constructing deep neural networks helps researchers to study biologically-plausible computational models of the brain. Deep learning-based brain data analyses have been important for cognitive neuroscience. Cognitive neuroscience research in this direction could lead to more insights that cannot be achieved by traditional analysis methods. Furthermore, novel insights on the brain could lead to novel models and architectures for the progress of current deep learning methods.

Bibliography

- [1] Takeshi Matsuo, Keisuke Kawasaki, Takahiro Osada, Hirohito Sawahata, Takafumi Suzuki, Masahiro Shibata, Naohisa Miyakawa, Kiyoshi Nakahara, Atsuhiko Iijima, Noboru Sato, et al. Intracal electrocorticography in macaque monkeys with minimally invasive neurosurgical protocols. *Frontiers in Systems Neuroscience*, 5:34, 2011.
- [2] Lizette Heine, Andrea Soddu, Francisco Gómez, Audrey Vanhauzenhuyse, Lubaba Tshibanda, Marie Thonnard, Vanessa Charland-Verville, Murielle Kirsch, Steven Laureys, and Athena Demertzi. Resting state networks and consciousness. *Frontiers in psychology*, 3:295, 2012.
- [3] Thomas Naselaris, Kendrick N. Kay, Shinji Nishimoto, and Jack L. Gallant. Encoding and decoding in fmri. *NeuroImage*, 56(2):400–410, 2011.
- [4] Jean Rémi King, Laura Gwilliams, Chris Holdgraf, Jona Sassenhagen, Alexandre Barachant, Denis Engemann, Eric Larson, and Alexandre Gramfort. Encoding and decoding neuronal dynamics: Methodological framework to uncover the algorithms of cognition. 2018. hal-01848442.
- [5] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.
- [6] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61:85–117, 2015.
- [7] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.
- [8] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.
- [9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1026–1034, 2015.

- [10] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.
- [11] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164, 2015.
- [12] Daniel LK 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.
- [13] Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS Comput Biol*, 10(11):e1003915, 2014.
- [14] Umut Güçlü and Marcel AJ 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.
- [15] Tomoyasu Horikawa and Yukiyasu Kamitani. Generic decoding of seen and imagined objects using hierarchical visual features. *Nature Communications*, 8:15037, 2017.
- [16] Yağmur Güçlütürk, Umut Güçlü, Katja Seeliger, Sander Bosch, Rob van Lier, and Marcel AJ van Gerven. Reconstructing perceived faces from brain activations with deep adversarial neural decoding. In *Advances in Neural Information Processing Systems*, pages 4246–4257, 2017.
- [17] K Seeliger, U Güçlü, L Ambrogioni, Y Güçlütürk, and MAJ Van Gerven. Generative adversarial networks for reconstructing natural images from brain activity. *Neuroimage*, 181:775–785, 2018.
- [18] Guohua Shen, Kshitij Dwivedi, Kei Majima, Tomoyasu Horikawa, and Yukiyasu Kamitani. End-to-end deep image reconstruction from human brain activity. *Frontiers in Computational Neuroscience*, 13:21, 2019.
- [19] Gerwin Schalk and Eric C Leuthardt. Brain-computer interfaces using electrocorticographic signals. *IEEE Reviews in Biomedical Engineering*, 4:140–154, 2011.
- [20] Mark L Homer, Arto V Nurmikko, John P Donoghue, and Leigh R Hochberg. Sensors and decoding for intracortical brain computer interfaces. *Annual Review of Biomedical Engineering*, 15:383–405, 2013.
- [21] Stefano Panzeri, Jakob H Macke, Joachim Gross, and Christoph Kayser. Neural population coding: combining insights from microscopic and mass signals. *Trends in Cognitive Sciences*, 19(3):162–172, 2015.

- [22] James J DiCarlo, Davide Zoccolan, and Nicole C Rust. How does the brain solve visual object recognition? *Neuron*, 73(3):415–434, 2012.
- [23] Philippe G Schyns, Gregor Thut, and Joachim Gross. Cracking the code of oscillatory activity. *PLoS Biology*, 9(5):e1001064, 2011.
- [24] Yoichi Miyawaki, Hajime Uchida, Okito Yamashita, Masa-aki Sato, Yusuke Morito, Hiroki C Tanabe, Norihiro Sadato, and Yukiyasu Kamitani. Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. *Neuron*, 60(5):915–929, 2008.
- [25] Sanne Schoenmakers, Markus Barth, Tom Heskes, and Marcel van Gerven. Linear reconstruction of perceived images from human brain activity. *NeuroImage*, 83:951–961, 2013.
- [26] Gijs Joost Brouwer and David J Heeger. Decoding and reconstructing color from responses in human visual cortex. *Journal of Neuroscience*, 29(44):13992–14003, 2009.
- [27] Alan S Cowen, Marvin M Chun, and Brice A Kuhl. Neural portraits of perception: reconstructing face images from evoked brain activity. *Neuroimage*, 94:12–22, 2014.
- [28] Shinji Nishimoto, An T Vu, Thomas Naselaris, Yuval Benjamini, Bin Yu, and Jack L Gallant. Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology*, 21(19):1641–1646, 2011.
- [29] Siamac Fazli, Florin Popescu, Márton Danóczy, Benjamin Blankertz, Klaus-Robert Müller, and Cristian Grozea. Subject-independent mental state classification in single trials. *Neural networks*, 22(9):1305–1312, 2009.
- [30] Emanuele Olivetti, Seved Mostafa Kia, and Paolo Avesani. Meg decoding across subjects. In *2014 International Workshop on Pattern Recognition in Neuroimaging*, pages 1–4. IEEE, 2014.
- [31] Vinay Jayaram, Morteza Alamgir, Yasemin Altun, Bernhard Scholkopf, and Moritz Grosse-Wentrup. Transfer learning in brain-computer interfaces. *IEEE Computational Intelligence Magazine*, 11(1):20–31, 2016.
- [32] Li-Dan Kuang, Qiu-Hua Lin, Xiao-Feng Gong, Fengyu Cong, Jing Sui, and Vince D Calhoun. Multi-subject fmri analysis via combined independent component analysis and shift-invariant canonical polyadic decomposition. *Journal of Neuroscience Methods*, 256:127–140, 2015.
- [33] Yijun Wang and Tzyy-Ping Jung. A collaborative brain-computer interface for improving human performance. *PloS ONE*, 6(5):e20422, 2011.

- [34] Marc-André Carbonneau, Veronika Cheplygina, Eric Granger, and Ghyslain Gagnon. Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition*, 77:329–353, 2018.
- [35] Christopher R Holdgraf, Jochem W Rieger, Cristiano Micheli, Stephanie Martin, Robert T Knight, and Frederic E Theunissen. Encoding and decoding models in cognitive electrophysiology. *Frontiers in Systems Neuroscience*, 11:61, 2017.
- [36] David G Lowe. Object recognition from local scale-invariant features. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1150–1157, 1999.
- [37] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *European Conference on Computer Vision*, pages 404–417. Springer, 2006.
- [38] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, 1989.
- [39] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [40] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.
- [41] David H Hubel and Torsten N Wiesel. Receptive fields, binocular interaction and functional architecture in the cat’s visual cortex. *The Journal of Physiology*, 160(1):106–154, 1962.
- [42] David H Hubel and Torsten N Wiesel. Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology*, 195(1):215–243, 1968.
- [43] Kuniyuki Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36(4):193–202, 1980.
- [44] Andrew J Watrous, Juergen Fell, Arne D Ekstrom, and Nikolai Axmacher. More than spikes: common oscillatory mechanisms for content specific neural representations during perception and memory. *Current Opinion in Neurobiology*, 31:33–39, 2015.
- [45] Pascal Fries. Rhythms for cognition: communication through coherence. *Neuron*, 88(1):220–235, 2015.

- [46] György Buzsáki and Andreas Draguhn. Neuronal Oscillations in Cortical Networks. *Science*, 304, 2004.
- [47] Christoph Kayser and Peter König. Stimulus locking and feature selectivity prevail in complementary frequency ranges of v1 local field potentials. *European Journal of Neuroscience*, 19(2):485–489, 2004.
- [48] Philipp Berens, Georgios A Keliris, Alexander S Ecker, Nikos K Logothetis, and Andreas S Tolias. Feature selectivity of the gamma-band of the local field potential in primate primary visual cortex. *Frontiers in Neuroscience*, 2:37, 2008.
- [49] Joshua Jacobs and Michael J Kahana. Neural representations of individual stimuli in humans revealed by gamma-band electrocorticographic activity. *Journal of Neuroscience*, 29(33):10203–10214, 2009.
- [50] Christopher M Lewis, Conrado A Bosman, Nicolas M Brunet, Bruss Lima, Mark J Roberts, Thilo Womelsdorf, Peter de Weerd, Sergio Neuenschwander, Wolf Singer, and Pascal Fries. Two frequency bands contain the most stimulus-related information in visual cortex. *bioRxiv*, page 049718, 2016.
- [51] A. Belitski, A. Gretton, C. Magri, Y. Murayama, M. A. Montemurro, N. K. Logothetis, and S. Panzeri. Low-Frequency Local Field Potentials and Spikes in Primary Visual Cortex Convey Independent Visual Information. *Journal of Neuroscience*, 28(22):5696–5709, 2008.
- [52] Siddhesh Salelkar, Gowri Manohari Somasekhar, and Supratim Ray. Distinct frequency bands in the local field potential are differently tuned to stimulus drift rate. *Journal of Neurophysiology*, 120(2):681–692, 2018.
- [53] Astrid Von Stein, Carl Chiang, and Peter König. Top-down processing mediated by interareal synchronization. *Proceedings of the National Academy of Sciences*, 97(26):14748–14753, 2000.
- [54] Andre Moraes Bastos, Julien Vezoli, Conrado Arturo Bosman, Jan-Mathijs Schoffelen, Robert Oostenveld, Jarrod Robert Dowdall, Peter De Weerd, Henry Kennedy, and Pascal Fries. Visual areas exert feedforward and feedback influences through distinct frequency channels. *Neuron*, 85(2):390–401, 2015.
- [55] Georgios Michalareas, Julien Vezoli, Stan Van Pelt, Jan-Mathijs Schoffelen, Henry Kennedy, and Pascal Fries. Alpha-beta and gamma rhythms subserve feedback and feedforward influences among human visual cortical areas. *Neuron*, 89(2):384–397, 2016.
- [56] Craig G Richter, William H Thompson, Conrado A Bosman, and Pascal Fries. Top-down beta enhances bottom-up gamma. *Journal of Neuroscience*, 37(28):6698–6711, 2017.

- [57] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, pages 818–833. Springer, 2014.
- [58] Paul Dean. Effects of inferotemporal lesions on the behavior of monkeys. *Psychological Bulletin*, 83(1):41, 1976.
- [59] Charles G Gross. How inferior temporal cortex became a visual area. *Cerebral Cortex*, 4(5):455–469, 1994.
- [60] Antonio R Damasio, Daniel Tranel, and Hanna Damasio. Face agnosia and the neural substrates of memory. *Annual Review of Neuroscience*, 13(1):89–109, 1990.
- [61] Nancy Kanwisher and Morris Moscovitch. The cognitive neuroscience of face processing: An introduction. *Cognitive Neuropsychology*, 17(1-3):1–11, 2000.
- [62] Elizabeth K Warrington and Tim Shallice. Category specific semantic impairments. *Brain*, 107(3):829–853, 1984.
- [63] Rosaleen A McCarthy and Elizabeth K Warrington. Disorders of semantic memory. *Philosophical Transactions: Biological Sciences*, 346(1315):89–96, 1994.
- [64] Keiji Tanaka. Inferotemporal cortex and object vision. *Annual Review of Neuroscience*, 19(1):109–139, 1996.
- [65] Nikos K Logothetis and David L Sheinberg. Visual object recognition. *Annual Review of Neuroscience*, 19(1):577–621, 1996.
- [66] Robert Desimone, Thomas D Albright, Charles G Gross, and Charles Bruce. Stimulus-selective properties of inferior temporal neurons in the macaque. *Journal of Neuroscience*, 4(8):2051–2062, 1984.
- [67] Nikolaus Kriegeskorte, Marieke Mur, Douglas A Ruff, Roozbeh Kiani, Jerzy Bodurka, Hossein Esteky, Keiji Tanaka, and Peter A Bandettini. Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, 60(6):1126–1141, 2008.
- [68] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587, 2014.
- [69] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 779–788, 2016.
- [70] Matthew D Zeiler, Graham W Taylor, and Rob Fergus. Adaptive deconvolutional networks for mid and high level feature learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2018–2025, 2011.

- [71] Nicolas Pinto, David D Cox, and James J DiCarlo. Why is real-world visual object recognition hard? *PLoS Comput Biol*, 4(1):e27, 2008.
- [72] Jeremy Freeman and Eero P Simoncelli. Metamers of the ventral stream. *Nature neuroscience*, 14(9):1195–1201, 2011.
- [73] Thomas Serre, Aude Oliva, and Tomaso Poggio. A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences*, 104(15):6424–6429, 2007.
- [74] Jim Mutch and David G Lowe. Object class recognition and localization using sparse features with limited receptive fields. *International Journal of Computer Vision*, 80(1):45–57, 2008.
- [75] Radoslaw Martin Cichy, Aditya Khosla, Dimitrios Pantazis, Antonio Torralba, and Aude Oliva. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, 6, 2016.
- [76] Stefano Panzeri, Nicolas Brunel, Nikos K Logothetis, and Christoph Kayser. Sensory neural codes using multiplexed temporal scales. *Trends in Neurosciences*, 33(3):111–120, 2010.
- [77] Hanlin Tang, Calin Buia, Radhika Madhavan, Nathan E Crone, Joseph R Madsen, William S Anderson, and Gabriel Kreiman. Spatiotemporal dynamics underlying object completion in human ventral visual cortex. *Neuron*, 83(3):736–748, 2014.
- [78] Dean V Buonomano and Wolfgang Maass. State-dependent computations: spatiotemporal processing in cortical networks. *Nature Reviews Neuroscience*, 10(2):113–125, 2009.
- [79] John E. Lisman and Ole Jensen. The theta-gamma neural code. *Neuron*, 77(6):1002–1016, 2013.
- [80] a K Engel, P Fries, and W Singer. Dynamic predictions: oscillations and synchrony in top-down processing. *Nature reviews. Neuroscience*, 2(10):704–16, 2001.
- [81] Conrado A Bosman, Jan-Mathijs Schoffelen, Nicolas Brunet, Robert Oostenveld, Andre M Bastos, Thilo Womelsdorf, Birthe Rubehn, Thomas Stieglitz, Peter De Weerd, and Pascal Fries. Attentional stimulus selection through selective synchronization between monkey visual areas. *Neuron*, 75(5):875–888, 2012.
- [82] Timo Van Kerkoerle, Matthew W Self, Bruno Dagnino, Marie-Alice Gariel-Mathis, Jasper Poort, Chris Van Der Togt, and Pieter R Roelfsema. Alpha and gamma oscillations characterize feedback and feedforward processing in monkey

- visual cortex. *Proceedings of the National Academy of Sciences*, 111(40):14332–14341, 2014.
- [83] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems*, pages 3320–3328, 2014.
- [84] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 806–813, 2014.
- [85] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *International Conference on Machine Learning*, pages 647–655, 2014.
- [86] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.
- [87] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035, 2019.
- [88] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [89] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6541–6549, 2017.
- [90] Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. *University of Montreal*, 1341(3):1, 2009.
- [91] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2(11):e7, 2017.
- [92] Ilya Kuzovkin, Raul Vicente, Mathilde Petton, Jean-Philippe Lachaux, Monica Baciu, Philippe Kahane, Sylvain Rheims, Juan R Vidal, and Jaan Aru. Activations of deep convolutional neural networks are aligned with gamma band activity of human visual cortex. *Communications biology*, 1(1):107, 2018.

- [93] Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:4, 2008.
- [94] John-Dylan Haynes and Geraint Rees. Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience*, 7:523–534, July 2006.
- [95] Frank Tong and Michael S Pratte. Decoding patterns of human brain activity. *Annual Review of Psychology*, 63:483–509, 2012.
- [96] Brian N Pasley and Robert T Knight. Decoding speech for understanding and treating aphasia. In *Progress in Brain Research*, volume 207, pages 435–456. Elsevier, 2013.
- [97] James V Haxby, Andrew C Connolly, and J Swaroop Guntupalli. Decoding neural representational spaces using multivariate pattern analysis. *Annual Review of Neuroscience*, 37:435–456, 2014.
- [98] Neha Tiwari, Damodar Reddy Edla, Shubham Dodia, and Annushree Bablani. Brain computer interface: A comprehensive survey. *Biologically Inspired Cognitive Architectures*, 26:118–129, 2018.
- [99] Ghislain St-Yves and Thomas Naselaris. Generative adversarial networks conditioned on brain activity reconstruct seen images. In *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 1054–1061. IEEE, 2018.
- [100] Guohua Shen, Tomoyasu Horikawa, Kei Majima, and Yukiyasu Kamitani. Deep image reconstruction from human brain activity. *PLoS Computational Biology*, 15(1):e1006633, 2019.
- [101] Kendrick N Kay, Thomas Naselaris, Ryan J Prenger, and Jack L Gallant. Identifying natural images from human brain activity. *Nature*, 452(7185):352, 2008.
- [102] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.
- [103] Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579*, 2015.
- [104] Grigory V Rashkov, Anatoly S Bobe, Dmitry V Fastovets, and Maria V Komarova. Natural image reconstruction from brain waves: a novel visual bci system with native feedback. *bioRxiv*, page 787101, 2019.
- [105] Chi Zhang, Kai Qiao, Linyuan Wang, Li Tong, Ying Zeng, and Bin Yan. Constraint-free natural image reconstruction from fmri signals based on convolutional neural network. *Frontiers in Human Neuroscience*, 12:242, 2018.

- [106] Rufin VanRullen and Leila Reddy. Reconstructing faces from fmri patterns using deep generative neural networks. *arXiv preprint arXiv:1810.03856*, 2018.
- [107] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.
- [108] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.
- [109] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.
- [110] Ian Goodfellow. Nips 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*, 2016.
- [111] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.
- [112] Takeru Miyato and Masanori Koyama. cgans with projection discriminator. In *International Conference on Learning Representations*, 2018.
- [113] Scott Reed, Zeynep Akata, Xinchun Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. *arXiv preprint arXiv:1605.05396*, 2016.
- [114] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiao lei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5907–5915, 2017.
- [115] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.
- [116] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. *arXiv preprint arXiv:1808.06601*, 2018.
- [117] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alexander Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

- [118] Shaojie Bai, J Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.
- [119] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.
- [120] Kevin Roth, Aurelien Lucchi, Sebastian Nowozin, and Thomas Hofmann. Stabilizing training of generative adversarial networks through regularization. In *Advances in Neural Information Processing Systems*, pages 2018–2028, 2017.
- [121] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? *arXiv preprint arXiv:1801.04406*, 2018.
- [122] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.
- [123] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pages 6626–6637, 2017.
- [124] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4681–4690, 2017.
- [125] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *European Conference on Computer Vision Workshops (ECCVW)*, September 2018.
- [126] Fabien Lotte, Marco Congedo, Anatole Lécuyer, Fabrice Lamarche, and Bruno Arnaldi. A review of classification algorithms for eeg-based brain-computer interfaces. *Journal of Neural Engineering*, 4(2):R1, 2007.
- [127] Natasha Padfield, Jaime Zabalza, Huimin Zhao, Valentin Masero, and Jinchang Ren. Eeg-based brain-computer interfaces using motor-imagery: Techniques and challenges. *Sensors*, 19(6):1423, 2019.
- [128] Nikolaus Kriegeskorte and Pamela K Douglas. Interpreting encoding and decoding models. *arXiv preprint arXiv:1812.00278*, 2018.
- [129] Michel Besserve, Karim Jerbi, Francois Laurent, Sylvain Baillet, Jacques Martinerie, and Line Garnero. Classification methods for ongoing eeg and meg signals. *Biological Research*, 40(4):415–437, 2007.

- [130] Swati Aggarwal and Nupur Chugh. Signal processing techniques for motor imagery brain computer interface: A review. *Array*, 1:100003, 2019.
- [131] Zoltan Joseph Koles. The quantitative extraction and topographic mapping of the abnormal components in the clinical eeg. *Electroencephalography and Clinical Neurophysiology*, 79(6):440–447, 1991.
- [132] Yijun Wang, Shangkai Gao, and Xiaornog Gao. Common spatial pattern method for channel selection in motor imagery based brain-computer interface. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 5392–5395. IEEE, 2006.
- [133] Wei Wu, Zhe Chen, Xiaorong Gao, Yuanqing Li, Emery N Brown, and Shangkai Gao. Probabilistic common spatial patterns for multichannel eeg analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(3):639–653, 2014.
- [134] Morteza Alamgir, Moritz Grosse-Wentrup, and Yasemin Altun. Multitask learning for brain-computer interfaces. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 17–24, 2010.
- [135] Yuan Pin Lin and Tzyy Ping Jung. Improving eeg-based emotion classification using conditional transfer learning. *Frontiers in Human Neuroscience*, 11:334, 2017.
- [136] Veronika Cheplygina, David MJ Tax, and Marco Loog. On classification with bags, groups and sets. *Pattern Recognition Letters*, 59:11–17, 2015.
- [137] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.
- [138] Vince D Calhoun, Jingyu Liu, and Tülay Adalı. A review of group ica for fmri data and ica for joint inference of imaging, genetic, and erp data. *Neuroimage*, 45(1):S163–S172, 2009.
- [139] James V Haxby, J Swaroop Guntupalli, Andrew C Connolly, Yaroslav O Halchenko, Bryan R Conroy, M Ida Gobbini, Michael Hanke, and Peter J Ramadge. A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron*, 72(2):404–416, 2011.
- [140] Hejia Zhang, Po-Hsuan Chen, and Peter Ramadge. Transfer learning on fmri datasets. In *International Conference on Artificial Intelligence and Statistics*, pages 595–603, 2018.
- [141] Shuo Zhou, Christopher R Cox, and Haiping Lu. Improving whole-brain neural decoding of fmri with domain adaptation. In *International Workshop on Machine Learning in Medical Imaging*, pages 265–273. Springer, 2019.

- [142] Matthias Krauledat, Michael Tangermann, Benjamin Blankertz, and Klaus-Robert Müller. Towards zero training for brain-computer interfacing. *PLoS ONE*, 3(8):e2967, 2008.
- [143] Xin Chai, Qisong Wang, Yongping Zhao, Xin Liu, Ou Bai, and Yongqiang Li. Unsupervised domain adaptation techniques based on auto-encoder for non-stationary eeg-based emotion recognition. *Computers in Biology and Medicine*, 79:205–214, 2016.
- [144] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan R Salakhutdinov, and Alexander J Smola. Deep sets. In *Advances in Neural Information Processing Systems*, pages 3391–3401, 2017.
- [145] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.
- [146] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016.
- [147] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.
- [148] Luis Garcia Dominguez, Richard A Wennberg, William Gaetz, Douglas Cheyne, O Carter Snead, and Jose Luis Perez Velazquez. Enhanced synchrony in epileptiform activity? local versus distant phase synchronization in generalized seizures. *Journal of Neuroscience*, 25(35):8077–8084, 2005.
- [149] Qingguo Wei, Yijun Wang, Xiaorong Gao, and Shangkai Gao. Amplitude and phase coupling measures for feature extraction in an eeg-based brain-computer interface. *Journal of Neural Engineering*, 4(2):120, 2007.
- [150] Kei Majima, Takeshi Matsuo, Keisuke Kawasaki, Kensuke Kawai, Nobuhito Saito, Isao Hasegawa, and Yukiyasu Kamitani. Decoding visual object categories from temporal correlations of ecog signals. *Neuroimage*, 90:74–83, 2014.
- [151] Brendon O Watson, Mingxin Ding, and György Buzsáki. Temporal coupling of field potentials and action potentials in the neocortex. *European Journal of Neuroscience*, 48(7):2482–2497, 2018.
- [152] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International Conference on Machine Learning*, pages 2132–2141, 2018.
- [153] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

- [154] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015.
- [155] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pages 2048–2057, 2015.
- [156] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997.
- [157] Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks*, 18(5-6):602–610, 2005.
- [158] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.
- [159] Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. Learning to forget: continual prediction with lstm. In *International Conference on Artificial Neural Networks*, volume 2, pages 850–855, 1999.
- [160] Oriol Vinyals, Samy Bengio, and Manjunath Kudlur. Order matters: Sequence to sequence for sets. *arXiv preprint arXiv:1511.06391*, 2015.
- [161] Demis Hassabis, Dhharshan Kumaran, Christopher Summerfield, and Matthew Botvinick. Neuroscience-inspired artificial intelligence. *Neuron*, 95(2):245–258, 2017.
- [162] Adam H Marblestone, Greg Wayne, and Konrad P Kording. Toward an integration of deep learning and neuroscience. *Frontiers in Computational Neuroscience*, 10:94, 2016.

Acknowledgements

The work in this thesis would not have been accomplished without the help of a lot of people. First of all, I am grateful to my supervisor, professor Takayuki Okatani. When I applied to the graduate school, I was a completely newcomer to computer science because my major was Economics in my undergraduate. However, Prof. Okatani accepted me as a graduate student and have taught me a lot of things over the five years in my master's and Ph.D. courses. I have learned much from Prof. Okatani through many situations, such as discussion, meetings, and paper writing.

I would also thank all of my collaborators at Niigata University: professor Keisuke Kawasaki, professor Hasegawa, Takumi Hongo, and Harunori Miki. Especially, professor Kawasaki not only provided me their brain datasets for our experiments, but also have closely worked with me for the work in this thesis through meetings, discussion, and writing. Takumi and Harunori helped us with recording brain signals from macaque monkeys and discussing our research.

In writing and completing this thesis, professor Koichi Hashimoto and professor Shingo Kagami gave me a lot of valuable feedback for improving my thesis.

My research in Ph.D. was supported by the scholarship and research funding by the doctoral course scholarship of Division for Interdisciplinary Advanced Research and Education (DIARE), Tohoku University. In addition to financial support, DIARE gave me a lot of opportunities to present our work to other staffs and Ph.D. students at DIARE workshops, and to get valuable feedback from professor Yasuji Sawada in annual meetings.

I am also grateful to my colleagues and friends in our lab: professor Koya Yamaguchi for helping me set up my computational environment, managing our CPU/GPU servers, and always showing us his faithful attitude as a researcher; professor Mete Ozay for his continuous, passionate support through a lot of discussion; Akemi Sakane for always caring our lab; Masaki Saito, Makoto Ozeki, Koretka Ogata, Yan Zhang, Zhun Sun, Pongsate Tangseng, Xing Liu, Hu Junjie, Sumadianto Eka Putra, Takuya Yashima, Ryotaro Yada, Taro Hatsutani, Fazil Altinel, Duy Kien, Techapanurak En-

gkarat, Rito Murase, Yusuke Hosoya for a stimulating and wonderful research life in our lab.

Finally, I would like to thank my family for their continuous support and encouragement throughout all of my life.