Grounding Language Learning in Vision for Artificial Intelligence and Brain Research

by

Yizhen Zhang

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Electrical and Computer Engineering)
in the University of Michigan
2021

Doctoral Committee:

Associate Professor Zhongming Liu, Chair
Assistant Professor David Brang
Professor Jeffrey Fessler
Assistant Professor Andrew Owens

Yizhen Zhang
zhyz@umich.edu
ORCID iD: 0000-0002-2836-2666

DEDICATION

For my parents and husband.
For my dearest grandmother in Heaven.

ACKNOWLEDGMENTS

I have received a lot of support and assistance during my Ph.D. journey. I would like to express my deepest appreciation to my advisor, Prof. Zhongming Liu. The completion of my dissertation would not have been possible without his invaluable guidance and exceptional support. He has brought me into the fields of brain imaging, signal processing, machine learning, and neuroscience. He has mentored me on analytical skills, scientific thinking, and academic writing. More importantly, he has guided me to become an innovative and independent researcher. His broad knowledge, keen perspective, and endless passion for science have deeply infected me towards an academic career. I would also like to express my deepest gratitude to my committee members, Prof. David Brang, Prof. Jeffery Fessler, and Prof. Andrew Owens, for their kind encouragement and constructive feedback on my thesis work.

I have enjoyed working with all my current and former colleagues in the Laboratory of Integrated Brain Imaging (LIBI lab). Special thanks to Dr. Ulrich Scheven, Dr. Haiguang Wen, Dr. Lauren Lynch, Dr. Kun-Han Lu (Tom), Mr. Jun Young Jeong, Mr. Junxing Shi, Dr. Ranajay Mandal, Dr. Jiayue Cao (Cherry), Dr. Jung-Hoon Kim, Mr. Nishant Babaria, Mr. Steven Oleson, Mr. Kuan Han, Mr. Di Fu, Ms. Xiaokai Wang, Mr. Minkyu Choi, and Ms. Yuanhanqing Huang, for all the insightful discussion, valuable advice, and assistance on experiments. I would like to extend my sincere thanks to my collaborators, Dr. Robert Worth, Dr. Gang Chan, Dr. David Kemmerer, and Dr. Eugenio Culurciello, for their practical suggestions and helpful contribution to my research. I am also grateful to Dr. Kristin Mosier and Dr. Nicholas Barbaro for their help and support on the functional magnetic resonance imaging and electroencephalogram experiments at Indiana University School of Medicine.

Last but not least, I wholeheartedly thank my parents, Mr. Jiancheng Zhang and Mrs. Chunrong Zhang, who have always loved and supported me unconditionally. I also want to express my great appreciation to my husband, Mr. Ziyun Kong, who has accompanied and encouraged me throughout my graduate life. Thanks also to my friends who have always cheered me on and celebrated every accomplishment with me.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE

LIST OF TABLES

TABLE

ABSTRACT

Most models for natural language processing learn words merely from texts. However, humans learn language by referring to real-world experience and knowledge. My research aims to ground language learning in visual perception, taking one step closer to making machines learn language like humans. To achieve this goal, I have designed a two-stream model with deep neural networks. One stream extracts image features. The other stream extracts language features. The two streams merge to connect image and language features in a joint representation space. By contrastive learning, I have first trained the model to align images with their captions, and then refined the model to retrieve visual objects with language queries and infer their visual relations. After training, the model's language stream is a stand-alone system capable of embedding words in a visually grounded semantic space. This space manifests principal dimensions explainable with human intuition and neurobiological knowledge. The visually grounded language model also enables compositional language understanding based on visual knowledge and multimodal image search with queries based on image-text combination. This model can also explain human brain activity observed with functional magnetic resonance imaging during natural language comprehension. It sheds new light on how the brain stores concepts and organizes concepts by their semantic relations and attributes.

CHAPTER 1

Introduction

Language learning is a demanding function for both human and machine intelligence. Humans use language to record and express their experiences and knowledge. Language is also the key to establishing communication between one and another. Psycholinguists have been studying the mechanisms as to how humans acquire, store, and process language information in the brain [159, 16]. Researchers in artificial intelligence have also been building computational models to teach machines to understand natural language [183, 35, 39].

To date, most machine learning models designed for natural language processing [123, 162, 35, 21] are based on the distributional hypothesis [67, 42]. That is, words that occur in similar contexts carry similar meanings. These models, namely the *distributional semantic models*, show the most advanced performance on natural language understanding tasks [170] and support the basic learning principles behind this concise and elegant assumption. One of the most compelling and influential machine learning techniques inspired by this notion is word embedding [123, 137]. By characterizing the distributional characteristics (e.g., co-occurrence) of words or phrases in a large corpus, this technique represents each word as a vector in a continuous vector space with a dimension much lower than the total number of words. Since the representation is learned entirely from contextual relations, the words clustered in this space often exhibit similar syntactic functions and semantic meanings [149]. In addition, in some word embedding models, vector arithmetic captures semantic and syntactic relations and supports the simple composition of meanings [124]. For example, in word2vec space, *man* - *woman* + *queen* results in a vector close to the word *king*. For these properties, word embeddings have been widely used as input for more complicated language models and tasks. This technique has also been successfully applied to other sequential data that emphasize the co-occurrence of nearby elements, such as genes and proteins [5].

However, the distributional hypothesis is an imperfect principle for language learning. A brilliant thought experiment by Steven Harnad [66], which was originated from the famous "*Chinese Room Argument*"[152], demonstrated that humans could not learn a language only from textual information. In contrast, humans learn words by relating the meaning of the words to the referent and experience in the physical world. Suppose you have to learn Chinese as a second language,

and only a Chinese-Chinese dictionary is provided. Outstanding cryptographers may accomplish this mission by linking Chinese to another language that has already been grounded in real-world experience, just like the way they decipher ancient languages. However, learning Chinese as the first language with only a Chinese-Chinese dictionary seems to be an impossible task, because there is hardly any way to learn a language, which is a symbol system, if it is not explained by anything in the physical world but with merely meaningless symbols. [66].

In this sense, the predominant language models are all *ungrounded* since they learn words merely based on textual information. Arguably, these models cannot capture the semantics of words or concepts because they only learned "distributional properties" between words. This discrepancy retrospectively poses a question to researchers who develop brain-like computing models: As a more complex and intelligent computing machine, how does the human brain associate words with specific types of perceived objects or executable actions [139]?

The semantic grounding hypothesis from cognitive neuroscience assumes that concepts are grounded in action, perception, and emotion systems rather than represented by an isolated system that is detached from sensory and motor processes [8, 119]. This refers to the **grounded cognition theory**. It is supported by the notion that the brain is profoundly multimodal. Human sensory systems can educate each other without an external teacher [159]. The semantic memory is established by the bottom-up process from the low-level sensory system to the association areas and it is further retrieved by the top-down activation in the reverse direction as a simulation process [9]. Grounded cognition theory advocates that we learn words and concepts by grounding their meanings in perceptual systems [58, 2, 8, 15, 139, 119]. Consider how children learn the concept of *apple* - they first figure out that an apple is a round object, usually red or green, and can be held with one hand. It feels smooth to the touch, has a fruity taste, and is sweet and sour. Whenever their sensory systems catch such visual and gustatory features, they will recognize it as an *apple* In other words, the concept of *apple* in their mind is also intrinsically associated with those sensory features.

Inspired by the grounded cognitive theory, my research aims to establish a language learning model grounded in perceptual systems (e.g., visual perception) to bridge this fundamental gap between how humans and machines learn language. Briefly, the distributional semantic models only use *textual* contexts of a word to learn its vector representation. However, language models can learn concepts from text paired with sensory data, such as images. This concept of joint vision-language learning has been explored to perform cross-modal tasks, such as image captioning [106], visual qeustion answering [163], visual reasoning [28], and scene graph generation [164]. In line with these studies, we have developed a computational model with deep neural networks to ground language learning in vision. Specifically, we designed a two-stream model with a visual stream and a language stream connected at the top for multimodal learning. We first built a two-stream model to jointly learn visual and language representation from image-caption pairs. We then finetuned the

learned model by adding a cross-modal attention layer [114, 163] and bilinear operators [182] that were used to extract representations of the relations between visual objects. Both stages of learning utilized contrastive loss [68, 79].

After training, the visual and language streams can be separable and computable as stand-alone systems. We then extracted word representations from the language model and systematically evaluated the semantic space grounded in vision vs. the ungrounded semantic space learned only from the text. Our results suggest that after visual grounding, embeddings of concepts can be better organized and clustered by visual attributes, tend to be more predictive of human-defined norms of semantic features, and are useful for compositional language understanding and cross-modal image search.

To apply the word representation learned from the machine learning models to brain research, we further built a linear encoding model to predict functional magnetic resonance imaging (fMRI) responses and electrocorticography (ECoG) data collected from human subjects listening to stories. The encoding results indicate that the human semantic system involves a wide range of bilateral brain regions. We then used the trained encoding model as a "digital mirror" of the brain in terms of semantic processing and tested its utility for mapping the cortical representations of word categories, word relations, and principal axes of the grounded semantics. Our results collectively revealed the organization of conceptual and semantic features in the human brain. This is a potentially effective strategy that uses the computational language model as a tool for understanding the fundamental mechanisms of language processing and embodied cognition.

Chapter 2 defines the research problem and the specific research objectives. Chapter 3 introduces the background and related works in machine learning and neuroscience. Chapter 4 describes the detailed methods and results for visually grounded language learning. Chapter 5 elaborates how we evaluate and interpret the word embeddings learned from the language stream with human intuition and neurobiological knowledge. Chapter 6 explains the details about our experiments, methods, and findings by using the computational language model to explain brain data. Chapter 7 discusses some general future directions toward filling the gaps between how humans and machines learn language.

CHAPTER 2

Research Objectives

Computer vision (CV) and natural language processing (NLP) have become popular fields of artificial intelligence (AI) because of their wide applications. So far, machines have been trained to outperform humans on many visual tasks (such as image recognition [72]). Nevertheless, it is still challenging for a machine to learn, understand, and produce natural language in ways comparable to humans. The development of recent machine learning techniques, e.g., recurrent connection [122], word embedding [123], sequence to sequence learning [162], and attention mechanism [167], has significantly improved the performance on various NLP tasks. Besides, recent computational neuroscience studies suggest that such models are also helpful in studying language representation and semantic processing in the brain. [86, 77, 191]. However, even the most advanced NLP models are far from defeating humans on tasks such as question answering and text summarization [183]. In addition, the fundamental mechanisms underlying natural language processing in the human semantic system remain relatively unexplained by the current computational language models.

One of the main gaps between how machines and humans learn language is that the predominant language models are solely based on the distributional hypothesis [67], without grounding semantics in real-world experience. Such language models are different from the semantic system in our brain because humans learn language and concepts from multisensory and multimodal contexts. Distributional semantic models only access the word occurrence information in textual contexts during the learning process. No perceptual experience or knowledge is used to form the conceptual representation in such models. Given word co-occurrence, one might be able to guess that the words *monkey* and *banana* are closely related. It is still rather difficult to use word co-occurrence to infer the similarity in visual appearance between *monkey* and other primates or to know the visual features such as shape, color, or taste of a *banana* in its word embedding.

The discrepancy described above has motivated us to design a computational model that links language learning with physical experience. Inspired by the grounded cognition theory, we assume that humans learn, store, process, and produce language with a semantic system that is not isolated but grounded in action and perception systems [9, 8, 139, 119]. A semantic grounding scheme is needed to enhance the existing language learning models. Specifically, this study aims to establish a

brain-inspired model to ground language learning in visual perception. This model is expected to obtain more comprehensive concept representations through cross-modal contrastive learning with paired visual-language input. After training the language model with not only the textual context but also the visual context, we expect that the visually grounded word representation can be better interpreted according to human intuitions and can be further used to study semantic processing in the brain.

Towards these research objectives, this dissertation was based on the following three specific aims.

Aim 1: Develop a two-stream deep neural network for visually grounded language learning

We aim to develop a two-stream deep neural network for visually grounded language learning. One stream extracts hierarchical visual features from natural images, namely the visual stream. The other stream extracts contextual embeddings from natural language descriptions, namely the language stream. The two streams are connected and constrained by a cross-modal module to match the features of paired visual and language input at the top of both streams. During training, this model can simultaneously transfer the visual information to the language domain and vice versa and jointly learn multimodal representation of concepts. After training, each stream is reshaped by the cross-modal information and can be detached from one another to function as a stand-alone visual or language systems. To gradually ground language learning in vision, we break down the model training into three stages: the unimodal pretraining, visual grounding of natural language, and visual grounding of object relations.

Aim 2: Evaluate the word embedding space grounded in vision

We expect the language model can obtain more comprehensive and interpretable conceptual representations after being grounded with rich visual experience. To verify this hypothesis, we apply systematic intrinsic evaluations to the word embedding space from the language models. Specifically, we decompose the semantic space into orthogonal principal components and investigate the word distributions along the principal dimensions. We further assess word similarity, word categorization, and predictability of semantic feature norms with existing quantitative datasets based on human ratings. We also test the conceptual compositionality from visual reasoning and multimodal image search with a combination of image-text queries. The aim is to understand how visual grounding affects the language model in its learned semantic representations, by comparing the language models that share the same computational architecture but are trained with different contexts - unimodal context (distributional semantic models) or multimodal context (semantic grounding models).

Aim 3: Apply learned semantic representation to explain information processing in the brain

After training (Aim 1) and evaluation (Aim 2), we further apply the semantic representation from machine learning models to explain semantic processing in the brain. For this purpose, we first collect brain responses from human subjects under naturalistic paradigms, e.g., listening to stories, imagining a music piece with visual cues. We then build a linear voxel-wise encoding model to predict brain responses from vectorized word features, by delivering the same naturalistic stimuli to machines and human subjects. After that, the trained encoding model becomes a fully-computable model of the human brain in terms of natural language comprehension, which allows us to do high-throughput computational experiments for mapping a wide range of conceptual and semantic items to cortical representations, e.g., word categories, word relations, and principal axes in the visually grounded word embedding space. This approach and its resulting findings shed light on the cortical organization of high-dimensional semantic information.

CHAPTER 3

Background and Related Work

3.1 Language learning in machines

3.1.1 Learning principles

Distributional hypothesis. The learning principle in standard NLP models is mostly based on the distributional hypothesis [67, 14]. By characterizing the distributional properties of words or phrases in textual contexts, the word embedding technique represents each word as a point in a vector space, which has a much lower dimension than the total number of words [123, 137]. This vector space is capable of capturing some level of linguistic properties, such as word similarity (e.g., words with similar meanings tend to be clustered together) and word relationships (e.g., word pairs with the same relationship can be inferred by vector arithmetic) [124]. However, different from image features learned from convolutional neural networks [186], the vector representations of words are much harder to visualize, explain, and understand [149, 153]. The limitations and trade-offs between the explainability and learnability of the distributional semantic models have been recently discussed and highlighted [39, 95].

Symbol grounding hypothesis. As mentioned in the introduction, there is a fundamental gap in language learning between machines and humans. That is, the predominant NLP models remain *ungrounded* [55, 166, 17]. The process of connecting a *word* to its *meaning* by relating the word to its *referent* in the physical world is called *grounding* [168] (See illustration in Fig. 3.1). Early studies used binary perceptual features (e.g., "have 4 legs") labeled by human participants or image-based contexts as auxiliary information to build grounded semantic models by using similar techniques as used for learning word embeddings from textual contexts alone [156, 143, 23, 73, 185]. Findings from these studies demonstrate that perceptual features provide complementary semantic information to outperform word representations learned purely from texts. These studies also offer clues to implement the symbol grounding hypothesis as a language learning principal in deep learning models. However, it remains relatively unexplored how to ground different modalities into

the state-of-the-art contextual embedding models, e.g., Bert [35], and how grounding affects the word embedding space.

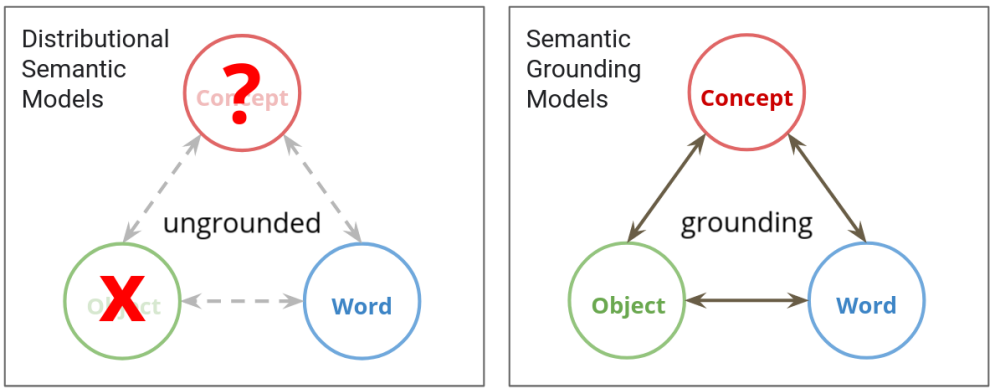


Figure 3.1: **Conceptual difference between distributional semantic models and semantic grounding models.** The distributional semantic models (left) are unable to connect the concept of a word with its real-world references.

3.1.2 Grounding language learning with multimodal data

The idea of grounding language learning in *physical* experience follows the distributional hypothesis, but extends the *textual* context to *multimodal* context [98]. For example, visual experience includes rich perceptual features (e.g., color, shape, texture, environment) of the words that describe various objects or scenes. Suppose an image and its caption are presented to the model simultaneously during training. The visual features will play an important role in altering the semantic features in the language learning model [7].

The computational structure for grounding language learning with multimodal processing can be summarized into three general methods: multimodal fusion, multimodal mapping, and multimodal alignment [11]. Multimodal fusion combines several unimodal representations into a single representation by concatenation or addition [22]. Multimodal mapping projects representations from one modality to another with simple computations, e.g., a linear mapping [43]. The approach of jointly learning for multimodal alignment first separates unimodal representation learning into independent streams and then optimizes the unimodal representations and cross-modal alignment simultaneously [68, 79].

Several prior works have been trying to integrate visual information into distributional semantic models to facilitate visually grounded language learning [89, 81]. Besides, some other modalities have also been investigated for language grounding, including auditory perception [88], olfactory perception [87], action [184, 25, 135, 115], and emotion [145].

3.2 Language learning in humans

3.2.1 Grounded cognition theory

How concepts are constructed and organized in the human brain remains an open question in cognitive science. Traditional theories assume that there exists an isolated semantic system detached from other modalities to represent and process concepts. However, the grounded cognition theory rejects this assumption and hypothesizes that human cognition (e.g., language processing, reasoning) is grounded in bodily, affective, perceptual, and motor processes [2, 8, 15, 139, 119].

For example, a classic model for grounded cognition is the Perceptual Symbol System proposed by Barsalou et al. [9]. This model hypothesizes that the perceptual experience activates the bottom-up process from sensorimotor areas to semantic integration and abstraction in the association areas. For the top-down semantic retrieval, the association areas reactivate the low-level sensorimotor areas to simulate the perceptual experience. The abstract representation of perceptual components (e.g., the semantic feature of *red*) are stored, processed, and retrieved by semantic memory. By simulating these perceptual components with attention-controlled top-down processing, concepts are connected back to the physical experience. This scheme establishes a fully functional conceptual system without the need for an amodal semantic system and proposes a plausible flow of neural information processing for grounded cognition [8].

3.2.2 Neuroimaging and behavioral evidence

Although it remains a theoretical hypothesis, the notion of grounded cognition theory is supported by neuroimaging studies, showing that both concrete concepts (e.g., *tools* [37]) and abstract concepts (e.g., *emotion* [38], *numbers* [101]) involve neural processing in sensorimotor areas. Specifically, recent lesion studies have shown that the sensorimotor cortex plays a more important role in the semantic processing of action-related words than regions (e.g., the frontoparietal cortex) in the language system defined by prior research [37].

Besides, behavioral studies on how humans learn novel concepts also advocate the theory that cognitive processing is grounded in perception and action systems [54, 61]. For example, mythical concepts (such as *Atlantis*) that cannot be directly grounded in perceptual experience are likely learned by connecting to descriptions with already grounded concrete concepts [61]. However, there is still a lack of a canonical basis in terms of computational modeling and systematic evaluations of brain data to establish that human cognition is inherently embodied and grounded [138].

3.3 Background and related works in machine learning.

3.3.1 Multimodal learning

Many research problems in machine learning and artificial intelligence are multimodal, such as image captioning [108], speech recognition[65], audio-visual alignment [68, 132], and visual question answering [75]. Besides, even for unimodal tasks, the performance could be improved by integrating information from another modality [43]. Among all multimodal learning problems, visual-language representation learning is especially of interest for its broad applications.

Most of the multimodal learning strategies directly or implicitly construct a shared embedding space for different modalities, either by concatenating or adding unimodal representations [22], or by using bilinear pooling with attention mechanism [44, 90], or by training a shared latent representation space through an encoder-decoder structure with paired image-text input [157], or by projecting and aligning representations from unimodal streams to a shared space for jointly learning the shared features [79]. These strategies are used to solve different types of multimodal tasks. The information fusion between the two modalities can happen at different levels in the model architecture. Using an early-stage fusion could potentially suppress within-modal interaction and make the processing of unimodal information less effective. Using a late-stage fusion would likely reduce cross-modal interaction that supports tasks requiring extensive information exchange across modalities, such as visual question answering [187].

3.3.2 Transformer encoder

The transformer encoder was first introduced in 2017 by Vaswani et al.[167]. The critical component in the transformer encoder is the multi-head self-attention mechanism. In each head, its output is a combination of features weighted by attention scores, which could be viewed as some sort of "pairwise relationship" between different feature locations (e.g., different words in natural language models [35] or different image patches in computer vision models [134]). Different heads are interpreted as capturing distinct features and relations [169]. Language models based on the transformer encoders pretrained for masked token prediction have shown outstanding performance in a range of natural language understanding tasks [35, 21].

The breakthrough success of the transformer encoder [167] in language learning has also inspired several works exploring multimodal representation learning by using self-attention to aggregate both within- and cross-modal information [163, 114, 161, 28]. Following Bert [35], these models are pre-trained for masked input prediction, where masking can be applied to either a word in the language input or an object in the image, showing supreme performance on cross-modal tasks such as visual question answering, text-image retrieval, and visual reasoning after transfer learning. But

since these models usually combine representations in language and visual domains at a very early stage and use the multimodal information throughout the processing hierarchy, it is not possible to separate the model into a stand-alone language model that could be used to evaluate the effect of visual grounding on the semantic representation of textual symbols [78].

3.3.3 Contrastive representation learning

The emergence of contrastive learning can be traced back to the early 1990s [10, 20]. The following intuition may help explain the idea. In the learning process, the model updates its parameters to drag similar representations (e.g., data from the same class) closer and push dissimilar representations (e.g., data from different classes) further away from each other. Commonly used contrastive learning losses include triplet loss [150], N-pair loss [160], and noise contrastive estimation (NCE) loss [62, 131].

The flexibility of contrastive learning makes it applicable to many circumstances where only weak supervision is needed, both for unimodal [26, 111] and multimodal [68, 79] representation learning. But it still suffers from the requirement of a larger dataset, a longer training period, and hard negative samplings [142]. The general framework and remaining challenges of contrastive representation learning has been discussed by a recent review paper [99].

3.3.4 Relational reasoning

Cognitive tasks involving relational reasoning, such as visual question answering, visual reasoning, and scene graph generation, are extremely difficult for machine learning. They require the model to infer generalizable and abstract knowledge embedded in images or sentences. Previous studies have attempted to model relations with multilayer perceptrons [113, 147] or graph neural networks [27, 104, 103]. Inspired by the vector arithmetic property in word2vec [124], another widely-used approach to model entities and relations in a large-scale knowledge base is to view relation embedding as a linear operator (e.g., translation [18, 173]) or bilinear operator [182, 128] over the low-dimensional embeddings of entities, which have shown promising progress on compositional reasoning [63]. Relational thinking is vital for human cognition and language learning because humans develop an understanding of a new concept by relating it to known concepts [36]. Furthermore, the relational reasoning task has led researchers to rethink how to reconcile symbolic AI with deep learning [13, 118], such that the advantage of data efficiency, generalizability, interpretability, and compositionality (or other formal logical processes) in symbolic AI could be accomplished in the deep learning framework to make up for the drawbacks of the latter one [47].

3.4 Bridging neural networks and biological brains

3.4.1 Brain signals

Functional MRI. Functional magnetic resonance imaging (fMRI) measures brain activity by detecting changes associated with blood flow. The blood flow and neuronal activation in the brain are coupled - when a brain region is activated, the blood flow to that region will also increase [112]. Although fMRI measures the BOLD (blood-oxygen-level dependent) signal instead of neural activity directly, it is widely used in brain mapping studies since it provides a non-invasive whole-brain measurement with relatively high spatial resolution compared to other brain signals [76]. But fMRI data has relatively low time resolution due to the sluggishness of neurovascular coupling, which delays and slows down the hemodynamic response.

Electrocorticography. Intracranial electroencephalography (iEEG), such as electrocorticography (ECoG), is a type of electrophysiological measurement that directly record electrical activity by placing electrodes on the exposed surface of the brain. Since a surgery is required to implant the electrode grid onto the cortical surface, ECoG is mostly applied to patients for clinically justifiable reasons. ECoG has relatively low spatial resolution due to the usually large spacing between electrodes and has relatively limited coverage because the electrode grid only covers a limited region on the cortex. Nevertheless, ECoG records neural signals with a superior signal-to-noise ratio and millisecond-level precision.

3.4.2 Naturalistic paradigm

The naturalistic paradigm uses diverse and dynamic stimuli similar to what we may encounter in the real world (such as movies, speeches, or music). Such stimuli are ecologically relevant to human perceptual, cognitive, and emotional experiences [192], in contrast to traditional paradigms that use artificial and controlled stimuli with strict experimental reductionism. Naturalistic stimuli tend to evoke consistent and reliable brain responses within and across subjects [70]. Thus, data collected from different sessions or different subjects under the same naturalistic stimuli can be directly integrated for group-level analysis given its high reproducibility [192]. The complex, dynamic, and multimodal naturalistic paradigm may involve widely distributed brain regions, providing unique opportunities to investigate the functional networks under real-life experience [102].

3.4.3 Neural encoding and decoding

Neural encoding and decoding involve a set of approaches that relate external stimuli to brain responses. Neural encoding predicts brain signals based on stimuli, while neural decoding predicts

the stimuli based on brain signals.

We usually use the input space, feature space, and activity space to describe various components in an encoding model. Input space stands for the space of stimuli. Feature space stands for the space of features encoded by the brain, which are related to specific inputs. Activity space stands for the space of brain activity. A widely accepted assumption is that the mapping between the input space and the feature space is nonlinear, since human brains process information in a complicated and non-linear fashion. While the mapping between the feature space and the activity space is linear since the relationship between the activation and the feature represented in the brain should be relatively simple and straightforward [127]. Thus, two steps are needed to build a neural encoding model. The first step is to find the non-linear mapping that extracts stimulus features from the input. The second step is to fit the stimulus features and brain activities with a linear regression model.

Linear encoding models for fMRI data have helped us understand the relationship between stimuli and brain responses in several prior studies with simple or human-defined features [126, 127]. Recent progress from deep representation learning has significantly increased the potential to extract a more high-dimensional and brain-like feature space for both visual [175, 129, 60, 155, 64] and language [77, 33, 191] systems. The reliability of using naturalistic stimuli [70] for neural encoding further allows the trained encoding models generalizable to different brains [174]. On the other hand, decoding with ECoG data on sensorimotor processing also provides a unique opportunity for developing groundbreaking brain-computer interfaces [3, 178].

CHAPTER 4

Grounding Language Learning to Vision

4.1 Rationale and Overview

[1] Grounding language in visual perception requires not only a *textual* context but also a *visual* context to be involved in the learning process [23]. Inspired by the strategy of extending the distributional semantic models to the multimodal context described in previous studies [156, 143, 23, 73, 185], we leverage the notion of visual grounding by using the state-of-the-art language learning models [35].

We first build a two-stream (i.e., visual and language streams) model to jointly learn visual and language representation from image-text pairs (Fig. 4.1). One stream takes the natural image as input and serves as a visual encoder to extract hierarchical visual features. The other stream takes the natural language description as input and serves as a language encoder to extract contextual embeddings of words. Finally, the two streams are connected at the top layer for image and language features to be projected into a shared representation space, constrained and trained with cross-modal contrastive learning.

We then finetune the learned model by using a multi-head cross-modal attention layer to extract visually grounded object representations and using bilinear operators to represent the relationships between visual objects. To learn visual relations between objects, we also apply contrastive learning with the similarity score defined by the learnable representations of **s**ubject-**p**redicate-**o**bject triplets to learn visual relations between objects.

After training, the visual and language streams are separable and computable as stand-alone systems after training. Thus, unlike prior works for multimodal representation learning [68, 114, 163, 79], this study focuses on the language model learned with visual data and characterizes the learned language model as a stand-alone system to extract visually grounded language representation as well as a system for performing cognitively demanding cross-modal tasks.

[1] This chapter is based on a conference paper [190] (under review).

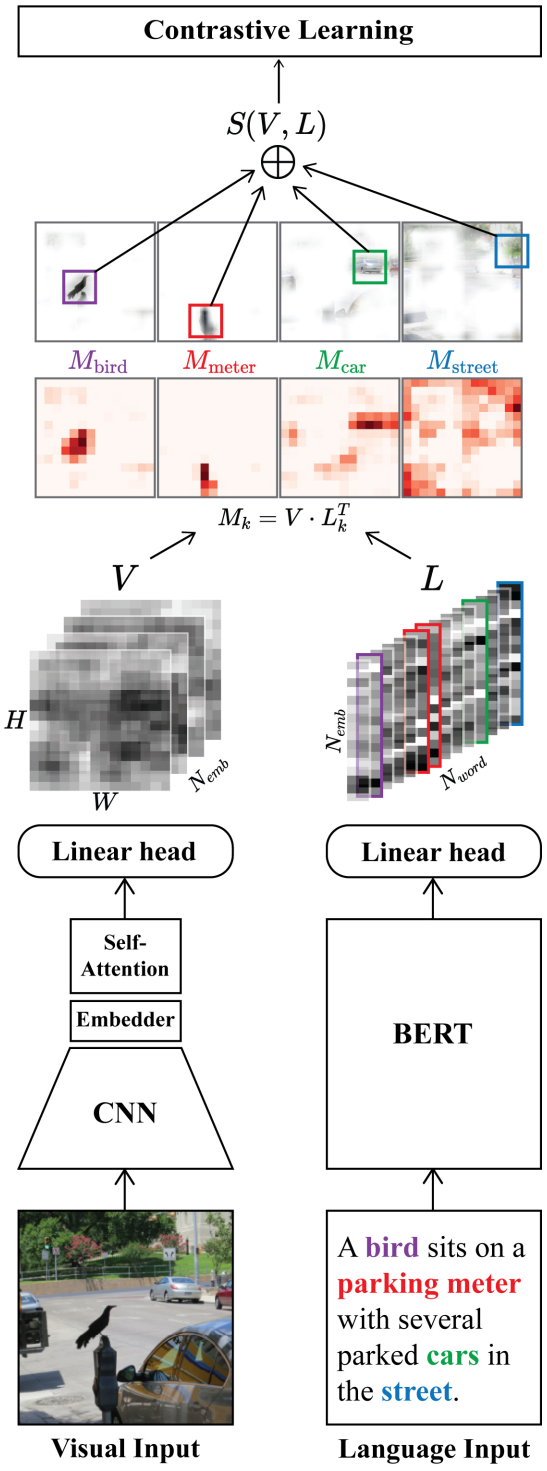


Figure 4.1: **The two-stream model for grounding natural language in vision.** The visual and language streams take an image and its caption as input respectively. The match-map is the inner-product between the visual feature maps and contextual word embeddings, forming a 3D tensor that highlights the matched visual and language content. The similarity score calculated from the match-map (Eq. 4.14) is used for cross-modal contrastive learning. See details in the following sections.

4.2 Approach

4.2.1 Visual stream

The visual stream (Fig. 4.1 bottom left) consists of a convolutional neural network, a linear transformation as an embedder layer to match the feature dimension with the language stream, and a multi-head spatial self-attention layer. The following sections describe these computational modules in detail.

4.2.1.1 Convolutional neural network as hierachical feature extractor

Convolutional neural network (CNN), initially inspired by the human visual system [74, 45, 141, 110], has shown a great success on various visual tasks, including image recognition [158], image segmentation [144], image denoising [188], image style transfer [85], and video prediction[130], etc. It applies translation-invariant filters to an input image by weight sharing of convolutional kernels. A typical CNN model learns features progressively emerging from simple edges, textures, and shapes, to complex and abstract semantics [100]. Such computational flow is similar to the feedforward processing in the brain's visual system [129, 175].

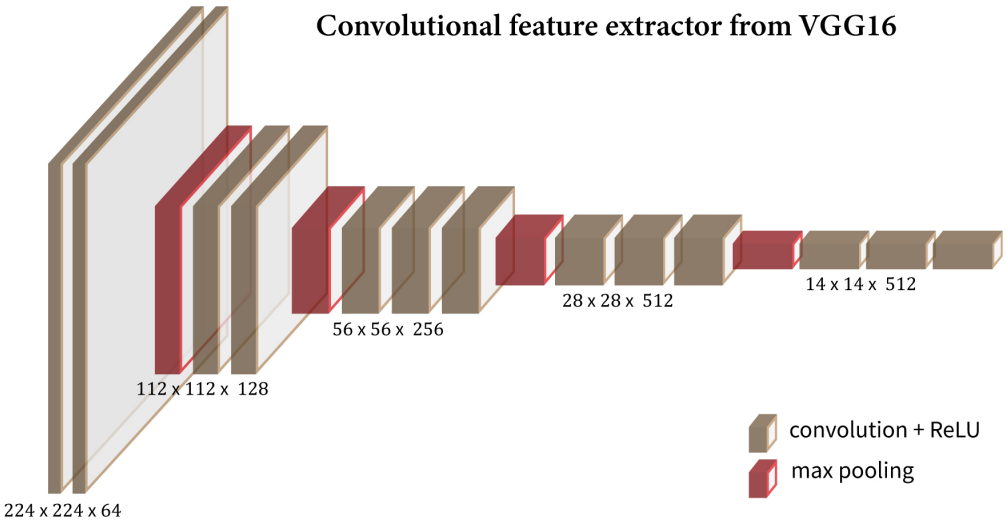


Figure 4.2: **An illustration of the VGG16 model structure.** Figure adapted from previous publications [158]. The number indicates the feature map size (width $\times$ height $\times$ channel).

In this study, we use a pre-established deep CNN structure - VGG16 [158], as the convolutional feature extractor in the visual stream [158]. VGG16 consists of sequential computational blocks. Each has stacked convolutional layers with nonlinear activation function (ReLU) followed by a max-pooling layer (Fig. 4.2). In VGG16, convolutions are all performed by 3 by 3 filters and max-pooling

is always over a 2 by 2 window with a stride equals to 2. Kernel size 3 by 3 is the smallest one that can aggregate neighbor information from surrounding pixels. Stacking convolutional layers with small kernels can reach a similar size for the effective receptive window but requiring significantly fewer parameters than a shallow convolutional layer with large kernels. The base CNN model can extract hierarchically organized spatial features by stacking sufficiently deep convolutional layers [158].

4.2.1.2 Multi-head spatial self-attention

In our design, the visual stream needs to perform multi-scale recognition to enable the language stream to match the detailed semantic knowledge of visual objects. Simonyan and colleagues [158] have shown that the performance of VGG models can be significantly improved by resizing the input images to multiple scales. This suggests that even deep CNN models may still focus too much on local details instead of aggregating enough global information [48, 12].

Inspired by the concept of *self-attention* for modeling long-term dependency in natural language [167, 35], we add a similar module (namely the **multi-head spatial self-attention**) into the visual stream. Integrating the attention mechanism into visual model may be a simple and efficient way to learn long-range association in the input image [172, 134].

The computations in spatial self-attention is conceptually similar to the transformer encoder [167], except that all the input features are from a 2D image instead of a 1D word sequence (Fig. 4.3).

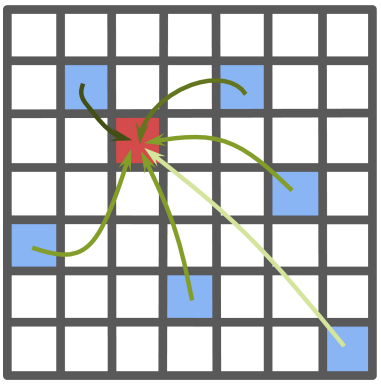


Figure 4.3: **A conceptual illustration of the spatial self-attention.** The red pixel is the query pixel, blue pixels refer to key pixels, and green arrows indicate the dynamic weights determined by the attention mechanism.

Suppose $\boldsymbol{I}$ is the input feature map. Here, $\boldsymbol{I} = \texttt{IE}+\texttt{PE}$, where IE is the image embedding from the embedder layer, PE is the location-wise positional encoding, "+" refers to the element-wise addition. The first step in spatial self-attention is converting the image feature into three sets *queries*, *keys*, and *values* though linear transformations $\boldsymbol{W}_{Q_i}, \boldsymbol{W}_{K_i}$, and $\boldsymbol{W}_{V_i}$ in the i-th attention head, respectively:

$$\boldsymbol{Q}_i = \boldsymbol{I}\boldsymbol{W}_{Q_i}, \quad \boldsymbol{K}_i = \boldsymbol{I}\boldsymbol{W}_{K_i}, \quad \boldsymbol{V}_i = \boldsymbol{I}\boldsymbol{W}_{V_i}. \tag{4.1}$$

After this step, *queries* $\boldsymbol{Q}_i$, *keys* $\boldsymbol{K}_i$, and *values* $\boldsymbol{V}_i$ are all 2D image feature maps with a feature dimension d lower than the input feature dimension in $\boldsymbol{I}$. The attention score is then calculated based on the inner-product between *queries* and *keys*, normalized by the square root of d,

$$\boldsymbol{o}_i = \text{softmax}(\frac{\boldsymbol{Q}_i \boldsymbol{K}_i^T}{\sqrt{d}}) \cdot \boldsymbol{V}_i, \tag{4.2}$$

$$\boldsymbol{y} = [\boldsymbol{o}_1, \boldsymbol{o}_2, \cdots, \boldsymbol{o}_h]\boldsymbol{W}_o. \tag{4.3}$$

Lastly, the output of this attention layer is the concatenation of outputs from all attention heads followed by a linear transformation with weight $\boldsymbol{W}_o$ (Eq. 4.3). More intuitive explanation about the multi-head self-attention mechanism is elaborated in Section 4.2.2.2: Transformer-based language modeling.

Besides, we also need to extend the positional encoding $\texttt{PE}$ from 1D version in the original transformer model to the 2D case. Prior works used either pre-defined positional encoding [134] or treat it as learnable parameters with a location-specific embedding layer. In this study, I also developed two options for positional encoding:

i. Pre-defined 2D positional encoding

In transformer [167], the 1D positional encoding is defined as a function of the position $x \in \mathbb{Z}$:

$$\texttt{PE}_{\text{1D}}^{2i}(x, d) = \sin\left(x \cdot C^{-\frac{2i}{d}}\right), \tag{4.4}$$

$$\texttt{PE}_{\text{1D}}^{2i+1}(x, d) = \cos\left(x \cdot C^{-\frac{2i}{d}}\right), \tag{4.5}$$

where d is the dimension of this positional encoding $\texttt{PE}_{\text{1D}}(x)$, which is the same as the dimension of 1D word embedding. The superscript refers to the $2i$-th or $(2i+1)$-th element in $\texttt{PE}_{\text{1D}}(x)$, $i \in \{0, 1, \ldots, d/2 - 1\}$. C is a pre-defined large constant (by default $C = 10,000$ in [167]).

This positional encoding has the following property,

$$\texttt{PE}_{\text{1D}}(x + k) = \boldsymbol{A}_{\boldsymbol{k}} \texttt{PE}_{\text{1D}}(x), \tag{4.6}$$

where $\boldsymbol{A}_{\boldsymbol{k}}$ is a matrix parameterized by k. This property is interpreted as the fact that if two pairs of positions have the same offset, they are related by the same linear transformation.

To generalize a similar idea from 1D to 2D, a simple way is to define $\texttt{PE}_{\text{2D}}$ as a function of the 2D location $(x, y) \in \mathbb{Z}^2$ by concatenating 1D positional encodings of x and y,

$$\texttt{PE}_{\text{2D}}((x, y), d) = \textbf{concat}[\texttt{PE}_{\text{1D}}(x, d/2), \texttt{PE}_{\text{1D}}(y, d/2)]. \tag{4.7}$$

Here, d is the dimension of $\texttt{PE}_{\text{2D}}$, which is the same as the feature dimension of the 2D image such that the positional encoding and the image feature can be added to form the attention input $\boldsymbol{I}$. Similar to the property of the 1D positional encoding, the 2D positional encoding defined in this

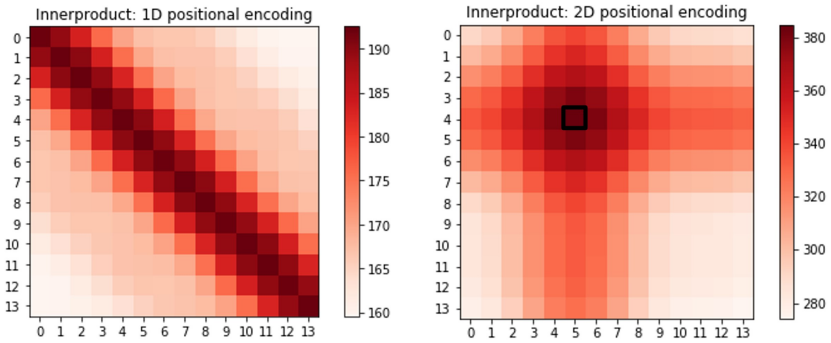


Figure 4.4: **The inner-product of positional encoding across location pairs.** Left: 1D positional encoding. Right: 2D positional encoding.

way has the following property,

$$\texttt{PE}_{\text{2D}}(x+k, y+l) = \boldsymbol{A}_{\boldsymbol{k},\boldsymbol{l}}\texttt{PE}_{\text{2D}}(x, y). \tag{4.8}$$

Besides, the "inner-product" based similarity between a pair of positional encoding decreases as their L1 distance increases (Fig. 4.4). This property adds a built-in spatial bias such that the model tends to attend to nearby image patches.

ii. Learnable 2D positional encoding

In the original transformer paper [167], the authors have claimed that the model performance is nearly identical if the positional encoding is either pre-defined or learned. Thus, in this study, we also keep an option of using a learnable 2D positional encoding:

$$\texttt{PE}_{\text{2D}}(x, y) = \mathbf{e}_{Nx+y}^{T}\boldsymbol{W}_{\texttt{PE}}, \tag{4.9}$$

where $\boldsymbol{W}_{\texttt{PE}}$ is a $N^2 \times d$ embedding matrix. N is the width and height of feature map, d is the size of feature dimension. $\mathbf{e}_{Nx+y}$ is a unit vector with the $(Nx+y)$-th element equals to 1 and all other elements equal to 0. $\boldsymbol{W}_{\texttt{PE}}$ is a set of learnable parameters.

4.2.2 Language stream

4.2.2.1 Word embedding

Word embedding refers to a process of mapping a large vocabulary to a dimension-reduced representational space. This mapping is usually first implemented by a linear layer with weight $\boldsymbol{W}_{\texttt{emb}}$ and then followed by non-linear transformations. Here, $\boldsymbol{W}_{\texttt{emb}}$ is a $N_{\texttt{voc}} \times N_{\texttt{emb}}$ matrix, where $N_{\texttt{voc}}$ is the total number of words in the vocabulary, $N_{\texttt{emb}}$ is the feature dimension of the representational

space, and $N_{\texttt{emb}} \ll N_{\texttt{voc}}$. Each row in the $\boldsymbol{W}_{\texttt{emb}}$ is a vector representation of a word in the vocabulary. Learning word embedding with a neural network has shown advantages in performing downstream natural language tasks [96, 84].

One of the most widely-used word embedding model is word2vec from Google [123]. It is trained with a shallow, two-layer neural network to reconstruct the local context of a given word. The word2vec preserves the semantic similarity of words in the low dimensional space as the cosine similarity between the corresponding vectors. It also preserves the semantic relations between words by vector arithmetic [124]. However, the drawbacks of word2vec (and other similar models of word embedding) are: 1) the learned word embedding does not change in different contexts, thus they are unable to account for the different meanings of the same word in different contexts (e.g., word *bank* in phrases *bank clerk* and *river bank* has entirely different meanings); 2) it only learns from texts and is not grounded in the physical world (e.g., visual experience).

4.2.2.2 Transformer-based language modeling

To mitigate the first drawback of word2vec, the language models based on bidirectional recurrent processing or transformer encoders aggregate contextual information into word representations [162, 167]. For instance, stacking self-attention layers in transformer encoders is able to capture long-range dependency between words in a sequence at a lower computational cost than recurrent neural networks.

In our model, the language stream is designed as a variation of Bert (i.e., Bidirectional Encoder Representations from Transformers) [35]. The language input (in most cases, a phrase or a sentence) is structured as a sequence of tokens (i.e., words or word pieces) [179]. The tokens are individually transformed to corresponding vector representations with a linear embedding layer. Each of the subsequent layers in the language stream consists of two sub-layers: one multi-head self-attention layer and one fully-connected layer.

In a multi-head self-attention layer, the input is first linearly transformed into three separate spaces with different functional roles: the **`query`** space, **`key`** space, and **`value`** space (Eq. 4.10). The inner-product between the *query* of a specific word and the *keys* of all words in the input sentence provides a set of attention scores, which is further divided by the square root of *query-key* feature dimension d for normalization (Eq. 4.11). Each attention score quantifies the association between a *key* word and a *query* word. After going through a soft-max function, the attention scores are converted to a probability distribution and are used as the weights for summing up the corresponding *value* vectors (Eq. 4.12). The attention-weighted sum of *value* vectors forms the output of the attention layer at the position of the given *query* word. This step aggregates the contextual information from every word in the input sequence according to its pair-wise relations with other words in the same sequence.

Such a process is repeated for each attention head. Different heads use different linear transformations for *query*, *key*, and *value*. As a result, different heads define different ways to evaluate the attention and generate different attention-weighted outputs. The output of the multi-head self-attention layer results from concatenating the output from every head, followed by a linear transformation (Eq. 4.13). The computational process in the multi-head self-attention layer is summarized as the following equations:

$$\boldsymbol{Q}_i = \boldsymbol{x}\boldsymbol{W}_{Q_i}, \quad \boldsymbol{K}_i = \boldsymbol{x}\boldsymbol{W}_{K_i}, \quad \boldsymbol{V}_i = \boldsymbol{x}\boldsymbol{W}_{V_i}, \tag{4.10}$$

$$\boldsymbol{A}_i = \frac{\boldsymbol{Q}_i \boldsymbol{K}_i^T}{\sqrt{d_k}}, \tag{4.11}$$

$$\boldsymbol{o}_i = \text{softmax}(\boldsymbol{A}_i) \cdot \boldsymbol{V}_i, \tag{4.12}$$

$$\boldsymbol{y} = [\boldsymbol{o}_1, \boldsymbol{o}_2, \cdots, \boldsymbol{o}_h]\boldsymbol{W}_o, \tag{4.13}$$

where $\boldsymbol{x}$ is the input with dimension $n \times d_e$, d_e is the embedding dimension and n is the number of input tokens. $\boldsymbol{W}_{Q_i}$ and $\boldsymbol{W}_{K_i}$ with dimension $d_e \times d_k$, $\boldsymbol{W}_{V_i}$ with dimension $d_e \times d_v$, are the weights that define the linear transformation from the input space to the `query`, `key`, and `value` spaces for the i-th head, respectively. $[\boldsymbol{o}_1, \boldsymbol{o}_2, \cdots, \boldsymbol{o}_h]$ is the concatenation of the output from h heads with dimension $n \times hd_v$. $\boldsymbol{W}_o$ is a matrix that defines the last linear transformation with dimension $hd_v \times d_e$. The motivation of using multiple attention heads is to allow the model to capture different types of syntactic and semantic relations between words (or tokens).

Following the above computational process, a fully-connected layer, which includes two linear transformations with a ReLU in between, is further applied. The dimensions of the input and the output are the same as the embedding dimension. The dimension of the hidden layer is denoted as d_h, which equals $4d_e$ in the default setting.

4.2.3 Cross-modal contrastive learning

We then connect the visual stream and the language stream at their top layers and train the two-stream model with cross-modal contrastive learning as illustrated in Fig. 4.1. First, the output features from both streams are projected to a common representational space through separate linear transformation heads [26]. In this common space, the inner-product between the visual representation $\boldsymbol{V}$ at every location and the language representation $\boldsymbol{L}$ of every word gives rise to a 3D match-map (Eq. 4.14), where each element indicates how a word in the text matches each region

in the image (See illustrations in Fig. 4.1). The similarity score $S(\boldsymbol{V}, \boldsymbol{L})$ between a pair of visual and language input is calculated by first taking the maximum value $S_k = \max_{i,j} \boldsymbol{M}_k(i,j)$ over the 2D match-map $\boldsymbol{M}_k$ of the k-th word, and then averaging the results across all words (Eq. 4.14):

$$\boldsymbol{M}_k(i,j) = \boldsymbol{V}_{i,j} \cdot \boldsymbol{L}_k^T, \quad S(\boldsymbol{V}, \boldsymbol{L}) = \frac{1}{K} \sum_{k=1}^{K} \max_{i,j} \boldsymbol{M}_k(i,j), \tag{4.14}$$

here i, j indicate the location in the 2D image feature map $\boldsymbol{V}$, and k indicates the k-th word in $\boldsymbol{L}$.

4.2.3.1 Triplet loss

Inspired by a prior work that uses a similar two-stream model for joint audio-visual learning [68], we construct a contrastive learning scheme by evaluating two triplet losses (Eq. 4.15). The triplet sample includes an anchor sample a, a positive sample p, and a negative sample n. The objective is to learn a representation with the distance $d(a,n)$ between the anchor sample and the negative sample larger than the distance $d(a,p)$ between the anchor sample and the positive sample with a constant margin m [176]. Here the anchor sample is defined as the representation of the input from one modality (e.g., vision). The positive sample and the negative sample are defined as the representation of the corresponding input and a random input from the other modality (e.g., language), respectively. The similarity metric between an anchor sample and a positive or negative sample is defined in Eq. 4.14, and the loss function is defined as

$$\text{Loss}_{\text{triplet}} = -\frac{1}{B}\Big(\sum_{i=1}^{B} \max(0, S(\boldsymbol{V}_i, \widetilde{\boldsymbol{L}_i}) - S(\boldsymbol{V}_i, \boldsymbol{L}_i) + m) + \max(0, S(\widetilde{\boldsymbol{V}_i}, \boldsymbol{L}_i) - S(\boldsymbol{V}_i, \boldsymbol{L_i}) + m) \Big), \tag{4.15}$$

where B is the batch size. For each sample in the batch, the triplet loss consists of two terms. The first one takes $\boldsymbol{V}_i$ as the anchor sample and the second one takes $\boldsymbol{L}_i$ as the anchor sample. $\widetilde{\boldsymbol{V}_i} = \boldsymbol{V}_j$ and $\widetilde{\boldsymbol{L}_i} = \boldsymbol{L}_k$ are negative samples randomly selected from the batch ($1 \leq j, k \leq B, j \neq i, k \neq i$). $m \in \mathbb{R}^+$ is the margin hyperparameter (by default $m = 1$).

4.2.3.2 NT-Xent loss

Another type of contrastive loss is based on noise contrastive estimation [62]. Inspired by prior studies that use the unimodal normalized temperature-scaled cross-entropy (NT-Xent) loss [131, 26, 79], we construct the cross-modal contrastive loss using the anchor sample from one modality and the positive sample and negative samples from the other modality. Similar to the triplet loss, we define two loss functions with the anchor sample from either images or texts and positive/negative samples from either texts or images, respectively:

$$\text{Loss}_l = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(S(\boldsymbol{V}_i, \boldsymbol{L}_i)/\tau)}{\sum_{j=1}^{B}\exp(S(\boldsymbol{V}_i, \boldsymbol{L}_j)/\tau)}, \tag{4.16}$$

$$\text{Loss}_v = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(S(\boldsymbol{V}_i, \boldsymbol{L}_i)/\tau)}{\sum_{j=1}^{B}\exp(S(\boldsymbol{V}_j, \boldsymbol{L}_i)/\tau)}. \tag{4.17}$$

For Loss_l in Eq. 4.16, the anchor sample $\boldsymbol{V}_i$ is an input image and the positive sample $\boldsymbol{L}_i$ is the corresponding image caption, whereas the negative samples $\boldsymbol{L}_j$ are unmatched textual descriptions included in the same batch (B is the batch size). Similarly, Loss_v in Eq. 4.17 is defined to contrast the positive and negative image samples against an anchor textual sample. For training the two-stream model with image-text pairs, we use the learning objective as the sum of Loss_l and Loss_v.

4.2.4 Relational grounding with cross attention and bilinear operator

After visually grounding the language model with image-text pairs (Section 4.2.3), we further finetune the model for visual relation prediction as illustrated in Fig. 4.5. In this stage, we remove the linear transformation heads in Fig. 4.1 and add a multi-head cross-modal attention module [114, 163]. Different from self-attention, this cross-attention module uses a query based on the embedding of an object description from the *language* stream (Query_L) and uses keys (Key_V) and values (Value_V) from local image features in the *visual* stream,

$$\text{Key}_V^i = \boldsymbol{V}\boldsymbol{W}_K^i, \quad \text{Value}_V^i = \boldsymbol{V}\boldsymbol{W}_V^i, \quad \text{Query}_L^i = \boldsymbol{L}\boldsymbol{W}_Q^i. \tag{4.18}$$

The attention score $\boldsymbol{A}_{L\to V}$ is the inner-product between Query_L and Key_V,

$$\boldsymbol{A}_{L\to V}^i = \text{softmax}\{\text{Query}_L^i(\text{Key}_V^i)^T\ /\sqrt{d}\}, \tag{4.19}$$

The attention-weighted sum of the Value_V from the visual stream is concatenated across attention heads to generate a visually grounded object representation,

$$\boldsymbol{O} = \text{concat}\{\boldsymbol{A}_{L\to V}^1\text{Value}_V^1, \cdots, \boldsymbol{A}_{L\to V}^h\text{Value}_V^h\} \tag{4.20}$$

where i in Eq. 4.18 and Eq. 4.19 refers to the i-th attention head. d in Eq. 4.19 refers to the query/key feature dimension in each attention head. h in Eq. 4.20 refers to the total number of attention heads in the cross-attention module (by default $h = 8$).

A bilinear operator is then applied to the grounded object representation in Eq. 4.20 for predicting the visual relation between two objects (linguistically a subject and an object). For both the subject and the object, their grounded representations are linearly transformed to a subspace $D = \mathbb{R}^d$ (by

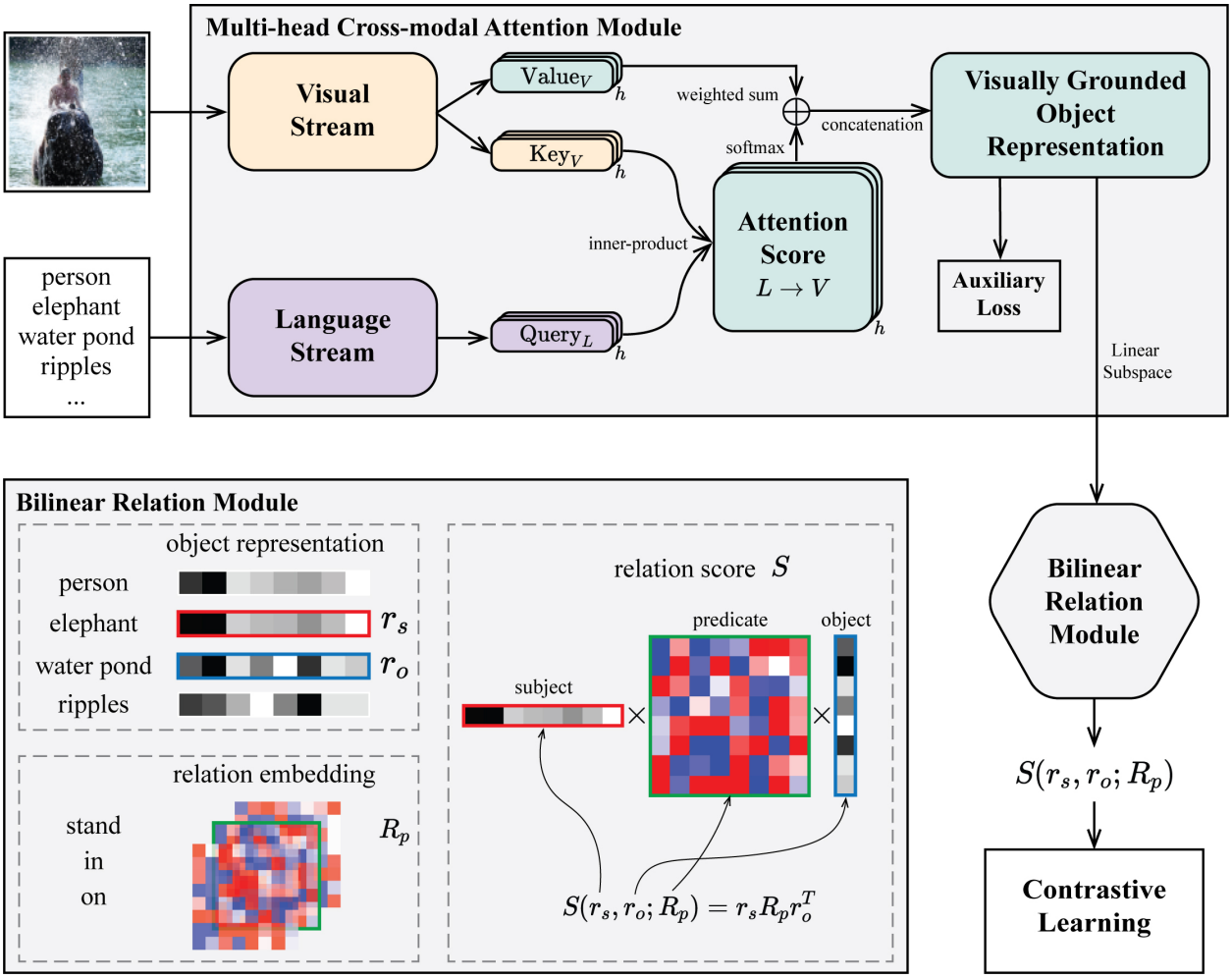


Figure 4.5: **Model architecture for visual grounding of object relations.** The language stream uses an object description as input (e.g., large black elephant; we only show the object name "elephant" in this illustration for simplicity). The multi-head cross-attention module outputs a set of visually grounded object representations (See detailed methods in Appendix A.3). The bilinear relation module (bottom left) further generates a relation score given representations of a (**s**ubject,**p**redicate,**o**bject) triplet (e.g., `(elephant,in,water pond)`) for contrastive learning.

default $d = 32$), denoted as $\boldsymbol{r}_s$ and $\boldsymbol{r}_o$, respectively. A predicate p is represented as a learnable bilinear operator $F_p :< D, D >\rightarrow \mathbb{R}$, which represents the relation embedding $\boldsymbol{R}_p$. Applying this bilinear operator to the subject vs. object representations measures their relation score S specific to the given predicate [182] expressed as

$$S(\boldsymbol{r}_s, \boldsymbol{r}_o; \boldsymbol{R}_p) = F_p(\boldsymbol{r}_s, \boldsymbol{r}_o) = \boldsymbol{r}_s \boldsymbol{R}_p \boldsymbol{r}_o^T. \tag{4.21}$$

Since the bilinear operation can be rewritten as the inner-product between the vectorization of $\boldsymbol{R}$ and the vectorization of the outer-product of $\boldsymbol{r}_1$ and $\boldsymbol{r}_2$,

$$S(\boldsymbol{r}_1, \boldsymbol{r}_2; \boldsymbol{R}) = \boldsymbol{r}_1 \boldsymbol{R} \boldsymbol{r}_2^T = \langle \text{vec}(\boldsymbol{r}_1^T \boldsymbol{r}_2), \text{vec}(\boldsymbol{R}) \rangle, \tag{4.22}$$

we further add the Frobenius norm constraint $\|\boldsymbol{R}\|_F = 1$ to each relational embedding matrix. Thus, theoretically, the optimal relation between a given object pairs $\boldsymbol{r}_1$ and $\boldsymbol{r}_2$ has an embedding matrix $\boldsymbol{R}^*_{(\boldsymbol{r}_1, \boldsymbol{r}_2)}$ with the following form

$$\boldsymbol{R}^*_{(\boldsymbol{r}_1, \boldsymbol{r}_2)} = \underset{\|\boldsymbol{R}\|_F = 1}{\arg\max}\, S(\boldsymbol{r}_1, \boldsymbol{r}_2; \boldsymbol{R}) = \frac{\boldsymbol{r}_1^T \boldsymbol{r}_2}{\|\boldsymbol{r}_1\|_2 \|\boldsymbol{r}_2\|_2}, \tag{4.23}$$

which implies that relation modeled in this way follows a *compositional property* achieved by matrix multiplication

$$\boldsymbol{R}^*_{(\boldsymbol{r}_1, \boldsymbol{r}_3)} = \boldsymbol{R}^*_{(\boldsymbol{r}_1, \boldsymbol{r}_2)} \boldsymbol{R}^*_{(\boldsymbol{r}_2, \boldsymbol{r}_3)}. \tag{4.24}$$

In algebraic logic [82, 148], this is known as the "composition of relations" [2]: for example, suppose we have three "objects" $\boldsymbol{r}_1, \boldsymbol{r}_2, \boldsymbol{r}_3$, which are members of a family. Then the relation *"is an uncle of"* between $\boldsymbol{r}_3$ and $\boldsymbol{r}_1$ is the composition of relations $\boldsymbol{r}_3$ *is a brother of* $\boldsymbol{r}_2$ and $\boldsymbol{r}_2$ *is a parent of* $\boldsymbol{r}_1$.

For training the proposed model for visual relation prediction, we also use contrastive learning with two loss functions by taking either relation embedding or subject/object representations as positive or negative samples,

$$\text{Loss}_{\text{rel}} = -\frac{1}{|\mathcal{B}|} \sum_{(\boldsymbol{r}_s, \boldsymbol{r}_o; \boldsymbol{R}_p) \in \mathcal{B}} \log \frac{\exp(S(\boldsymbol{r}_s, \boldsymbol{r}_o; \boldsymbol{R}_p)/\tau)}{\sum_{k \in \mathcal{K}_{\text{rel}}} \exp\big(S(\boldsymbol{r}_s, \boldsymbol{r}_o; \boldsymbol{R}_p^k)/\tau\big)}, \tag{4.25}$$

$$\text{Loss}_{\text{obj}} = -\frac{1}{|\mathcal{B}|} \sum_{(\boldsymbol{r}_s, \boldsymbol{r}_o; \boldsymbol{R}_p) \in \mathcal{B}} \log \frac{\exp(S(\boldsymbol{r}_s, \boldsymbol{r}_o; \boldsymbol{R}_p)/\tau)}{\sum_{k \in \mathcal{K}_{\text{obj}}} \exp(S(\boldsymbol{r}_s^k, \boldsymbol{r}_o^k; \boldsymbol{R}_p)/\tau)}. \tag{4.26}$$

In Loss_{rel}, $\mathcal{K}_{\text{rel}}$ is the set that contains all relations available. The anchor sample is a pair of

[2] Wikipedia: Composition of relations

subject and object in an image. The positive sample is the embedding of the ground-truth relation. The negative samples are the embeddings of all other relations. In Loss_{obj}, the anchor sample is a given relation. The positive sample is a subject-object pair that holds this relation. The negative samples are other subject-object pairs in a different relation. For both loss functions, the positive and negative samples are drawn from the same batch $\mathcal{B}$.

In addition, we also add a classification head (two fully connected layers with ReLU in between) and apply it to the grounded object representation $\boldsymbol{O}$ in Eq. 4.20. We use this object classification as an auxiliary objective (with a cross-entropy loss in Eq. 4.27) to constrain the grounded object representation captures visual and semantic information sufficient for separating different object classes:

$$\text{auxiliary loss}(\boldsymbol{x}, l) = -\log \frac{\exp\left(\boldsymbol{x}[l]\right)}{\sum_{i=1}^{N} \exp\left(\boldsymbol{x}[i]\right)} \tag{4.27}$$

where x is the output vector from the classifier with dimension N. N is the number of object classes. $l \in \{1, \dots N\}$ is the index of the ground-truth object class.

4.2.5 Training and testing

We progressively train the proposed model in three stages. In the *first* stage (Section 4.2.5.1), we pretrain each single stream with a large set of images or language corpus for training good unimodal encoders. For the language stream, we directly download the pretrained Bert[3] (embedding dimension$=768$; query/key dimension$=64$; number of self-attention layers$=12$; number of attention heads$=12$; trained on lower-cased English text) as our baseline model. For the visual stream, we pretrain the VGG16 with one self-attention layer on ImageNet for object classification. In the *second* stage (Section 4.2.5.2), we train the two-stream model as illustrated in Fig. 4.1 with the MS COCO dataset [108] which consists of image-caption pairs. In this stage, we freeze the CNN and the lower layers of Bert, and only allow the visual self-attention layer and the top k layers in Bert to be trainable (by default $k = 8$). In the *third* stage (Section 4.2.5.3), we transfer the model to perform visual relation prediction by adding one cross-modal attention layer and a bilinear relation module as illustrated in Fig. 4.5. We only finetune the visual self-attention layer and the higher l layers in Bert (by default $l = 2$). We train the model on the Visual Genome dataset [94], which contains images paired with scene graphs that are densely annotated with objects, attributes, and relationships. We clean the dataset and keep 114 relation labels and 55 object classes to balance training samples.

[3] bert-base-uncased

4.2.5.1 Unimodal pretraining

For training a deep convolutional neural network to perform visual tasks, a well-established strategy is to first pretrain the model using a large dataset for the object recognition task (e.g., ImageNet [32]) to learn a universal visual representation, and then to finetune the model with a specific downstream task which usually has fewer training data. This strategy has advantages of faster convergence and reasonable generalizability [51, 71].

In our model, we pretrain the visual stream on ImageNet for object classification and evaluate whether adding a self-attention layer can help learn a better representation of the image by efficiently aggregating the global information. The linear embedder transforms the image feature from 512 channels to 768 channels to match the word embedding dimensions in the Bert model. The visual self-attention layer has 12 heads. For the attention to account for spatial information, we use by default a learnable 2D positional encoding $\texttt{PE}_{\text{2D}}(x, y)$ as described in Section 4.2.1.2 and Eq. 4.9. We use the same hyper-parameter setting for training VGG16 and attention-enhanced visual model (batch size$= 200$, optimizer=SGD, learning rate$= 0.01$, momentum$= 0.9$, weight decay$= 1\text{e}{-4}$; learning rate decay by half for every 20 epochs).

4.2.5.2 Visual grounding of natural language

Then, we train the two-stream model as illustrated in Fig. 4.1 on the MS COCO dataset [108] with NT-Xent loss functions as defined in Eq. 4.16 and Eq. 4.17. The training data consists of 118287 images, each of them has 5 captions. For each iteration, we randomly choose 1 out of the 5 captions. We train the model with `Adam` optimizer (learning rate$= 5\text{e}{-5}$, weight decay$= 5\text{e}{-7}$, $\beta = (0.95, 0.999)$; dropout$= 0.3$; learning rate decay by half after every 15 epochs; batch size=180; total training epochs=100). The temperature parameter in the contrastive loss is always set as 0.1.

During training, we freeze the CNN and the lower layers of Bert and only allow the visual self-attention layer and the top k layers in Bert to be trainable (by default $k = 8$). we also freeze the query and key weights (both are linear transformation layers) in all Bert self-attention layers. This training constraint is given two considerations. First, we want to control the number of learnable parameters to avoid overfitting. Second, we want to separate the functional roles of query (Q), key (K), and value (V) in the self-attention mechanisms, such that Qs and Ks are trained to learn syntactic relation and contextual information between words in textual contexts, which have been optimized in the unimodal pretraining stage, whereas Vs are trained to learn an informative conceptual representation space being optimized with both *textual* contexts and *multimodal* contexts.

4.2.5.3 Visual grounding of visual object relation

In this training stage, instead of utilizing region annotations or a prior object detection module as in previous studies [163, 114], our model jointly learns the object representation and relation representation from raw image input to make the framework more flexible and generalizable [80]. Since the relation labels are relatively imbalanced in the original Visual Genome dataset [94], we filter the data to create a cleaner training and testing set for fine-tuning our model on the visual relation prediction task (as described in Section 4.2.4). To create object labels, we first extract the WordNet synset [125] of each object in the Visual Genome data annotation. We then investigate the distribution of the hypernyms of all object synsets and summarize them into 55 classes as shown in Table 4.1. To create relation labels, we first extract the "predicate" term from the data annotation and only preserve the ones with more than 250 instances in the training dataset (which remains 292 out of 37342 unique labels). We then manually merge equivalent predicates into a single relation label. For instance, we merge *near, next to, on side of, beside, standing next to, next, standing near, to right of, near a, close to, on side* to "**near**". After this, we end up having 114 unique relation labels as shown in Table 4.2. We further filter out image samples with fewer than 5 subject-predicate-object triplets (results in 98512 images). We then randomly split this cleaned dataset into training (93512 samples) and testing (5000 samples) set. At this training stage, we also freeze the CNN (VGG16 encoder) in the visual stream and lower layers in Bert. For each self-attention layer in the language stream, we freeze the weight in query and key transformation. We train the model with Adam optimizer (learning rate$= 1\mathrm{e}{-5}$, weight decay$= 5\mathrm{e}{-7}$, $\beta = (0.95, 0.999)$; dropout$= 0.1$; learning rate decay by half after every 15 epochs; batch size=180; total training epochs=150). The temperature parameter in the contrastive loss is always set as 1.0.

Table 4.1: Object classes defined from WordNet synsets for visual grounding of object relations.

feline.n.0	equine.n.01	mammal.n.01
bird.n.01	animal.n.01	body_part.n.01
bread.n.01	vegetable.n.01	fruit.n.01
meat.n.01	beverage.n.01	food.n.01
tree.n.01	herb.n.01	vessel.n.02
wheeled_vehicle.n.01	aircraft.n.01	vehicle.n.01
road.n.01	clothing.n.01	furniture.n.01
tableware.n.01	home_appliance.n.01	stairs.n.01
building_material.n.01	decoration.n.01	room.n.01
building.n.01	container.n.01	surface.n.01
machine.n.01	measuring_instrument.n.01	instrument.n.01
tool.n.01	device.n.01	paper.n.01
man.n.01	woman.n.01	person.n.01
equipment.n.01	sport.n.01	activity.n.01
symbol.n.01	sign.n.02	number.n.02
writing.n.02	body_of_water.n.01	facility.n.01
geological_formation.n.01	location.n.01	atmospheric_phenomenon.n.01
phenomenon.n.01	communication.n.02	structure.n.01
artifact.n.01		

Table 4.2: Relation labels after merging synonymous predicates for visual grounding of object relations.

on	have	in	of	wear
with	behind	hold	near	under
by	above	sit	in front of	to
at	over	for	around	ride
stand	hang	carry	eat	walk
cover	play	lay	along	among
and	watch	belong to	painted	against
from	parked	made of	say	covered
mounted	across	fly	lying	grow
use	outside	cross	worn	printed
full of	filled with	swing	built	pull
touch	adorn	a	hit	support
written	lean	drive	rest on	held
connected to	cut	throw	line	through
float	show	face	graze	cast
stick out of	catch	drink	reflected in	be
beyond	lead	read	swim	white
off	seen	push	shining on	ski
wait	surf	down	make	feed
run	take	enjoy	that	at end of
stuck	reflect	stacked	black	plugged
overlook	form	without	do	kick
visible on	brush	blue	work on	

4.3 Results

4.3.1 Image classification with occlusion experiments

We compared the model performance on ImageNet [146] classification between VGG16 and its variation with spatial attention enhancement. See results in Table 4.3. By adding a single self-attention layer to enforce global information aggregation for large objects and long-range dependency between distant objects, the top-1 accuracy on the ImageNet validation dataset has been improved from 71.6% to 74.3%.

Table 4.3: Object classification accuracy on ImageNet validation dataset.

	Object classification accuracy (%)		
Model	Top-1	Top-5	Top-10
VGG16	71.6	90.4	94.0
VGG16+attention	**74.3**	**91.8**	**95.1**

To evaluate how self-attention changes the feature representation, we have further performed an occlusion experiment [186]. For each image in a small validation dataset, a fixed-sized window (32×32) centered at a specific location is occluded with a grey square. The center of this occlusion is iterated throughout the whole image (stride = 8) for individual trials of the occlusion experiment. Each trial of occlusion outputs a probability of the correct class. It is expected that after occluding different portions of the input image, the model prediction (i.e., the probability of classifying the occluded input as the correct label) may result in different confidence levels. If the occluded region includes a key feature of the correct class, the probability may drop significantly. The effect of occlusion is evaluated and visualized as a heat-map, which shows the probability of correct classification as a function of the center of occlusion. For example, (see Fig. 4.6), VGG16 fails to classify the image of jay (a bird) as the correct label when any part of the bird is occluded. After adding the self-attention layer, the classification is compromised only when a very small part of the image is occluded. Similarly, in an image with a bridegroom, the classification performance drops only when a key feature (the Boutonnière) is occluded, whereas the performance of VGG16 is sensitive to occlusions at multiple regions. In another example image (Newfoundland dog), the attention-enhanced model is insensitive to the occlusion placed anywhere. In rarer cases, attention makes the model more sensitive to occlusion. See the last row of Fig. 4.6. Such cases usually involve a large-sized object in the image and the object identity is mostly defined by the local texture (e.g., the dishcloth). Overall, adding the self-attention helps aggregate information across the image and makes the model much less sensitive to image occlusion.

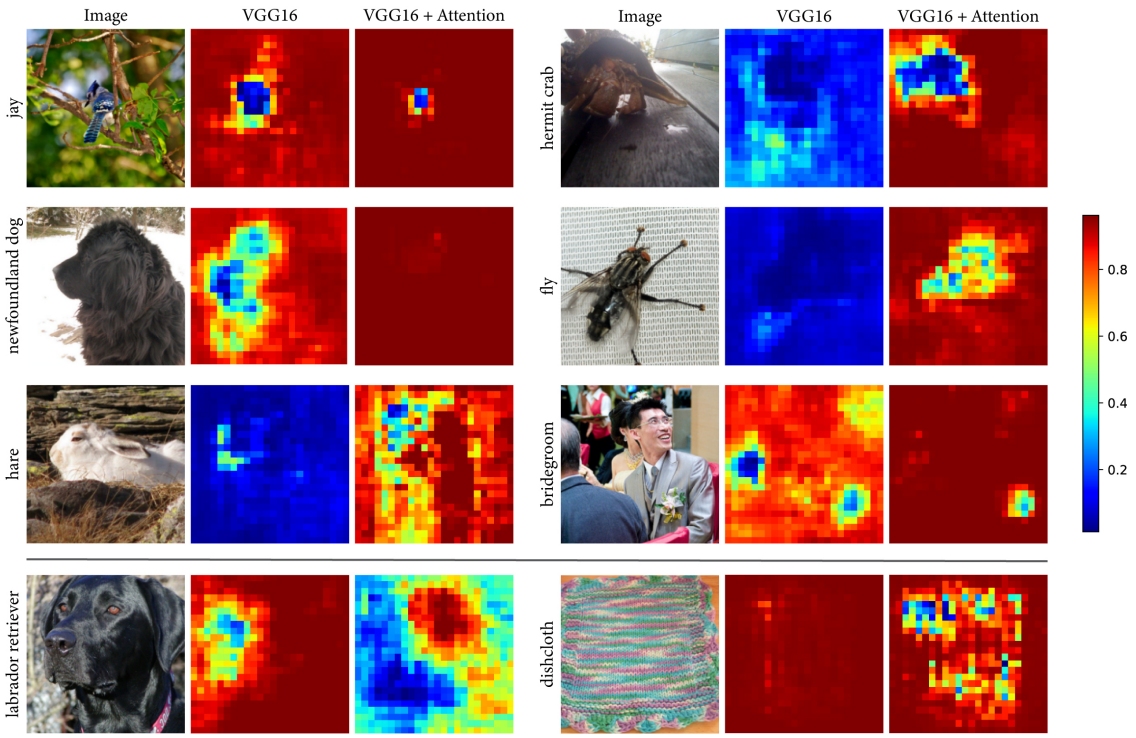


Figure 4.6: **Example results for occlusion experiment.** Each example contains three images (from left to right): the input image, the heatmap showing VGG16 prediction accuracy on occluded images, and the heatmap showing attention-enhanced VGG16 prediction accuracy on occluded images. The ImageNet class label is shown on the left. The first three rows show examples of when attention makes the model's performance less sensitive to occlusion. The last row shows examples of the opposite.

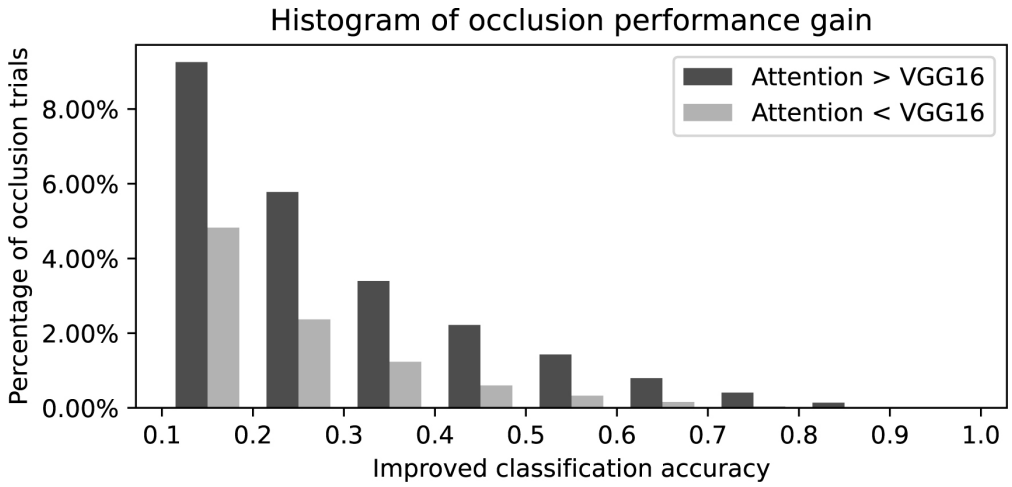


Figure 4.7: **Quantitative results for occlusion experiments.** The y axis shows the percentage of occlusion experiment trials. The x axis shows the absolute difference of the classification accuracy between the model w/wo attention.

Quantitatively, after adding a visual self-attention layer, the averaged top-1 classification accuracy on occluded images improved from $61.9 \pm 5.1\%$ to $66.6 \pm 5.6\%$. The standard deviation is taken from the 24 by 24 occluded trials in each image, and is then averaged across all validation samples. We further compare the probability of correct classification between VGG16 and its variation with attention for each trial of occlusion. We count the number of trials that the attention mechanism increases (or decreases) the probability of correct classification relative to VGG16, and evaluate the histogram by the size of increase (or decrease). As shown in Fig. 4.7, visual attention improves the classification of occluded images in many more trials than its baseline VGG16. Overall, the self-attention layer makes the model more robust against occlusion.

4.3.2 Visualizing the match-maps

To understand and validate how two streams are aligned in the joint metric space, we visualize the match-map given a pair of image and its caption, which follows the definition in Eq. 4.14 and the illustration in Fig. 4.1. We first threshold each match-map by zero to only preserve positive values, and then linearly scale it to a range from 0 to 1. For the visualization results as shown in the following figures, we overlay the match-map with its corresponding original image input after converting it into a mask by first taking square of the match-map values (still ranges 0 to 1) and then resizing the match-map to the input image size.

In a preliminary study, we have trained the two-stream model with triplet loss (See detailes in Section 4.2.3.1) and we visually inspect the effect of how the existence and different types of positional encoding changes the cross-modal alignment in the proposed model (Fig. 4.8). The results suggest that without attention the match-map learned from the two-stream model is unable to correctly match the whole object especially when the object size is relatively large (the first row). By adding an self-attention layer without positional encoding as a spatial constraint, the match-map becomes too inclusive, highlighting not only the object itself but also some background regions (the second row). The match-map appears the most reasonable only when the visual stream is enhanced by a self-attention layer with positional encoding (the last two rows), whereas the pre-defined and learnable `PE` do not show any significant difference.

After the two-stream model illustrated in Fig. 4.1 is trained on MS COCO dataset with NT-Xent loss as described in Section 4.2.5.2, we also visualize the match-map to ensure that the two stream alignment is correctly captured by the similarity score defined in Eq. 4.14. In the testing dataset, we first evaluate the examples with a *bird* object that occurs in both the image and its caption (Fig. 4.9). For the visualization purpose, we overlap each 2D math-map with the input image such that only the highlighted pixels are visible while other locations are masked out. In this specific example, we find that the 2D match-maps for word *bird* always catch the location of a bird in its corresponding image,

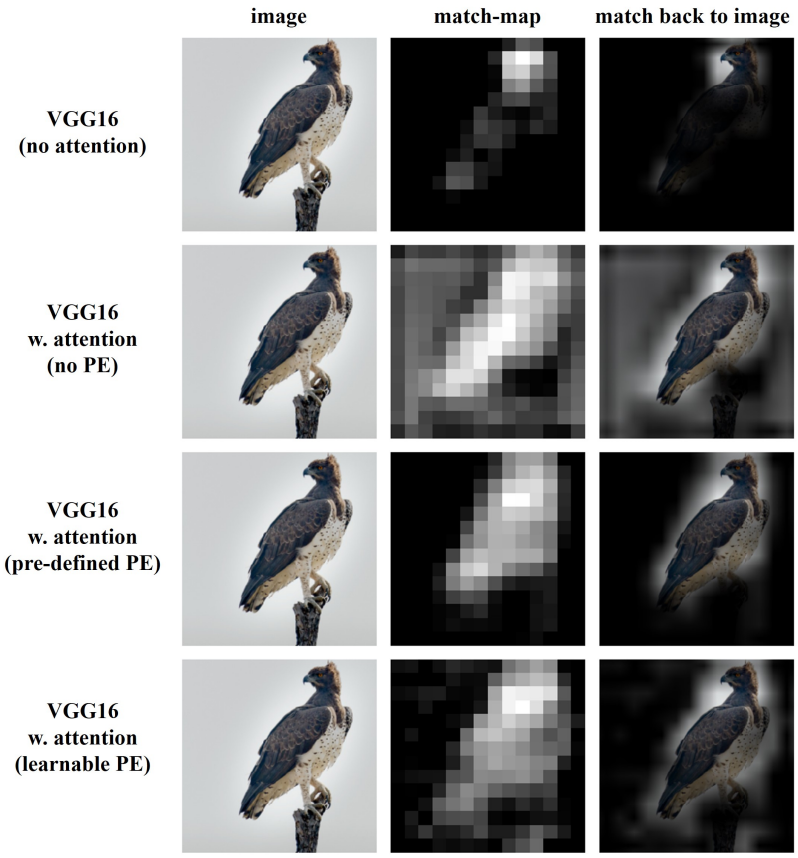


Figure 4.8: Match-map visualization for different visual models.

no matter the size of bird is small (upper rows) or large (lower rows). However, the match-map tends to be a bit over inclusive in some examples (the right column).

We further visualize how the match-map changes for different words in the same image caption (Fig. 4.10). The results in this example shows that the match-map highlights different parts of the image when the semantic content changes in the language input (e.g., "a bird" highlight the *bird*, "on a parking meter" highlight the *parking meter*, "several parked cars" highlight the *cars*, "in the street" highlight the *road* and *trees*). These visualization also implicitly validate the notion that the contextual embeddings of word from transformer encoders are grouped into meaningful segments.

4.3.3 Cross-modal retrieval

Fig. 4.11 shows the image-to-text and text-to-image retrieval performance on the validation set of MS COCO which contains 5000 images. The results suggest that allowing more layers in Bert learnable (i.e., earlier stages of visual grounding) results in better cross-modal retrieval accuracies, while freezing weights on query and key transformations (blue dots) reduces the number of learnable parameters without compromising the performance.

Besides, we have also done a separate experiment by always using one fixed caption for each image during training. The result suggests that although model architectures and learning objectives are the same, utilizing multiple synonymous sentences at different iterations instead of using a fixed sentence for all iterations significantly improves the accuracy for both image-to-text retrieval (top-1 accuracy: from 18.24% to 25.30%) and text-to-image retrieval (top-1 accuracy: from 17.22% to 23.90%).

4.3.4 Visual relation prediction

4.3.4.1 Performance with the testing dataset given different training settings

After further training the model as illustrated in Fig. 4.5 with our cleaned Visual Genome dataset, we test the model's performance on object classification and visual relation prediction with different hyper-parameter settings in the language stream (Fig. 4.12). The result suggests that the trained model can better classify objects from visually grounded representations (Fig. 4.12 left) when the query and key weights are learnable (orange dots) or when more layers in bert are learnable (star markers). But different levels of visual grounding on image captions as in the MS COCO pretraining stage (indicated by the x axis labels) has a minor effect on grounded object classification performance. In contrast, applying visual grounding to earlier stages of natural language processing by Bert results in better performance for relation prediction (Fig. 4.12 right). However, all these differences were relatively minor (Fig. 4.12).

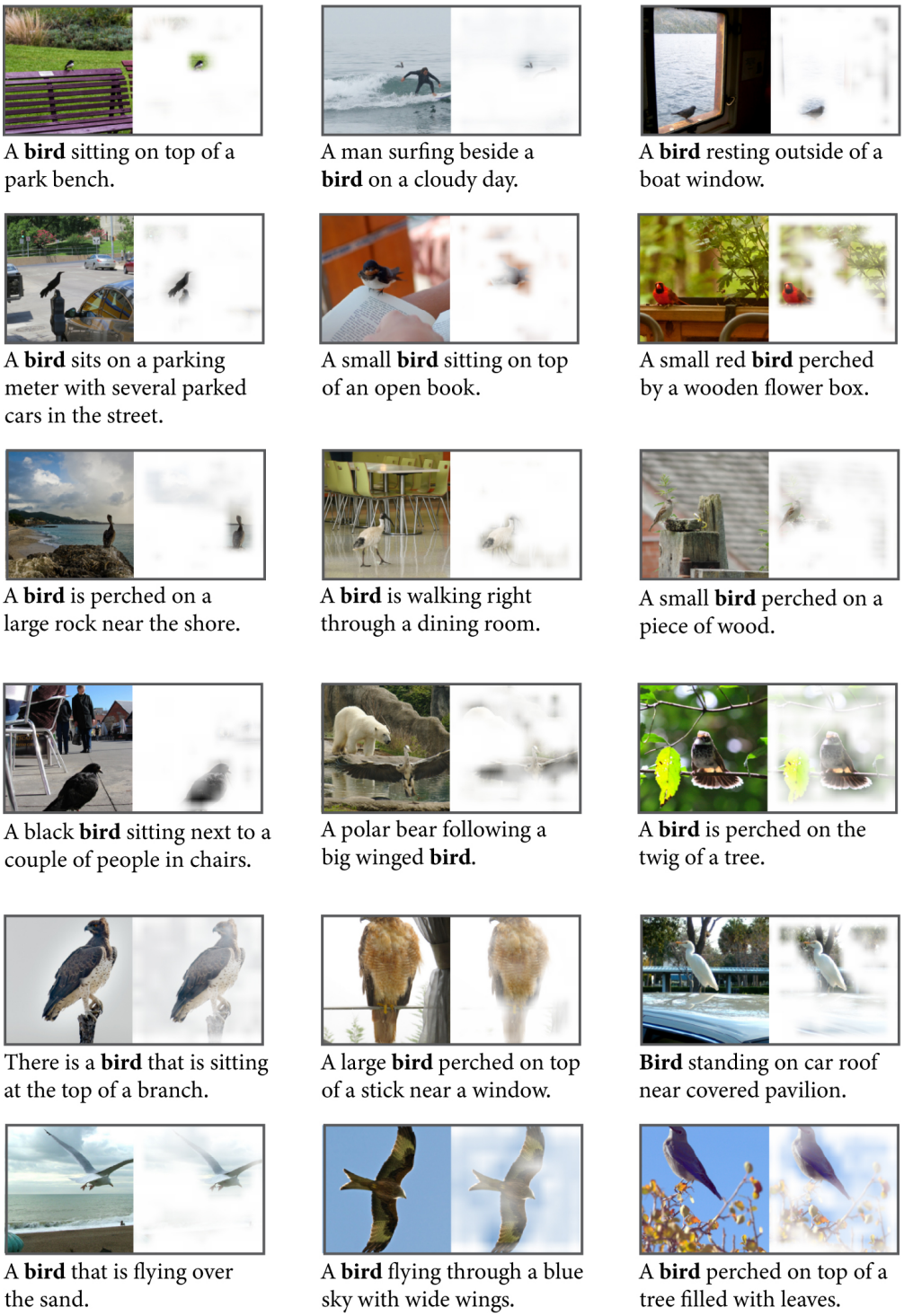


Figure 4.9: Match-map visualization for examples in the testing dataset that include **bird** objects.

Figure 4.10: Match-map visualization for all words in an example caption.

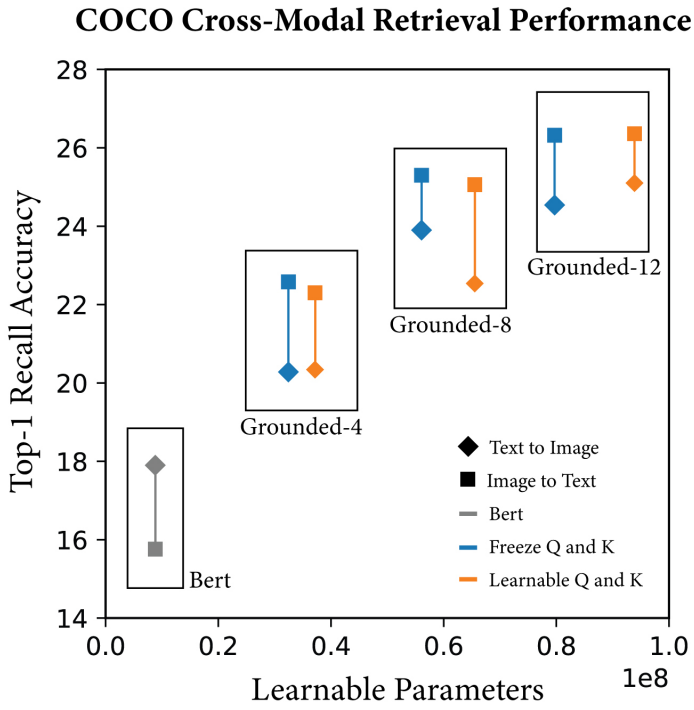


Figure 4.11: **Cross-modal retrieval performance on MS COCO.** The x axis shows the number of learnable parameters at the 2nd training stage (i.e., visual grounding of natural language). The y axis shows the top-1 recall accuracy. The label under each black box in the figure corresponds to a different setting of the learnable transformer layers in the language stream. `Bert`: the whole language stream is frozen. `Grounded-`k: the top k layers in Bert is learnable.

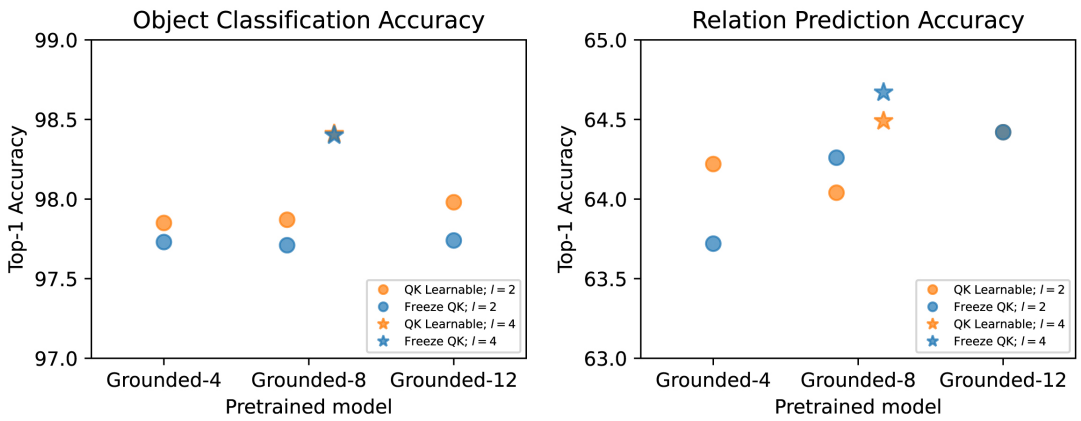


Figure 4.12: **Model performance on object classification and visual relation prediction with different pretrained models and training settings.** `Grounded-k`: the top k layers in Bert are learnable in the MS COCO pretraining stage. l in the figure legend refers the number of learnable layers in Bert at the stage for grounding visual object relations with Visual Genome dataset.

4.3.4.2 Ablation study for different training losses

Since the training objective is constructed with three different loss functions (Loss_{rel}, Loss_{obj}, the auxiliary object classification loss), we also check how each loss function contributes to the model performance, by excluding individual loss in a set of ablation experiments. The results suggest that the model shows the best performance on relation prediction by combining all three losses (Table 4.14, Fig. 4.13). If either the auxiliary object recognition loss or the contrastive loss Loss_{obj} is excluded, the model still tends to have comparable performance. But if Loss_{rel} is excluded, the model shows worse performance, which suggests Loss_{rel} is the key component that allows the model to learn visual relation. We also find that Loss_{obj} and Loss_{rel} are somewhat entangled during training (i.e., minimizing one loss would also decrease the other loss, see Fig. 4.14). The model converges faster (Fig. 4.13) when we combine both two contrastive losses.

Table 4.4: Effects of different loss functions evaluated with ablation experiments.

Model	Object Classification	Relation Prediction (Top-1)	Relation Prediction (Top-10)
Combined loss	97.71	**64.26**	**95.21**
No auxiliary loss	0.60	64.19	94.99
No Loss_{rel}	97.95	29.97	68.78
No Loss_{obj}	**98.39**	64.14	95.14

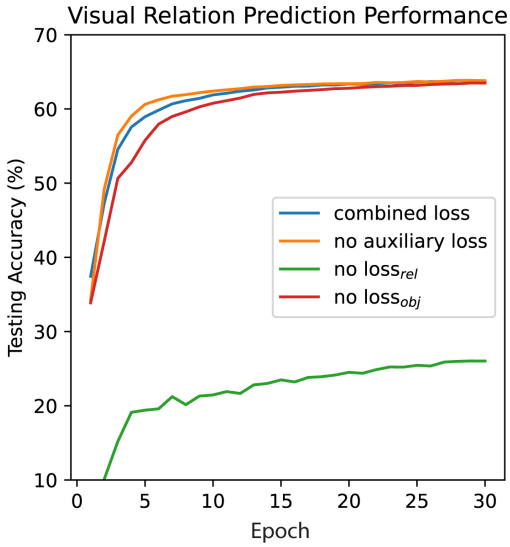


Figure 4.13: **Learning curve in terms of the accuracy of visual relation prediction with the testing dataset for the first** 30 **training epochs.** The blue curve shows the performance when the model is trained with the default loss that combines Loss_{rel}, Loss_{obj}, and the auxiliary loss for object classification. The other three curves show the performance when the model is trained when one of the three losses is excluded.

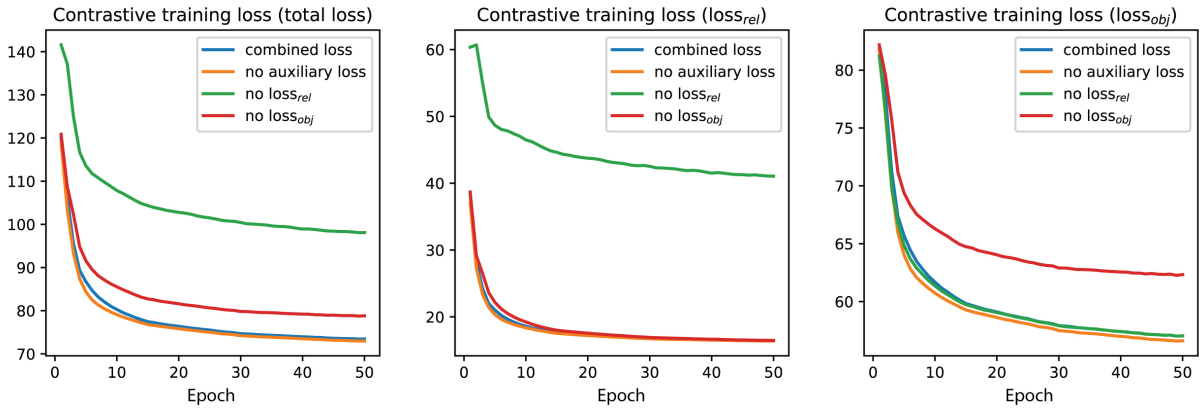


Figure 4.14: **Learning curve of the contrastive loss functions.** The *total loss* in the first figure refers to the summation $\text{Loss}_{\text{rel}} + \text{Loss}_{\text{obj}}$.

4.3.4.3 Visualizing test examples

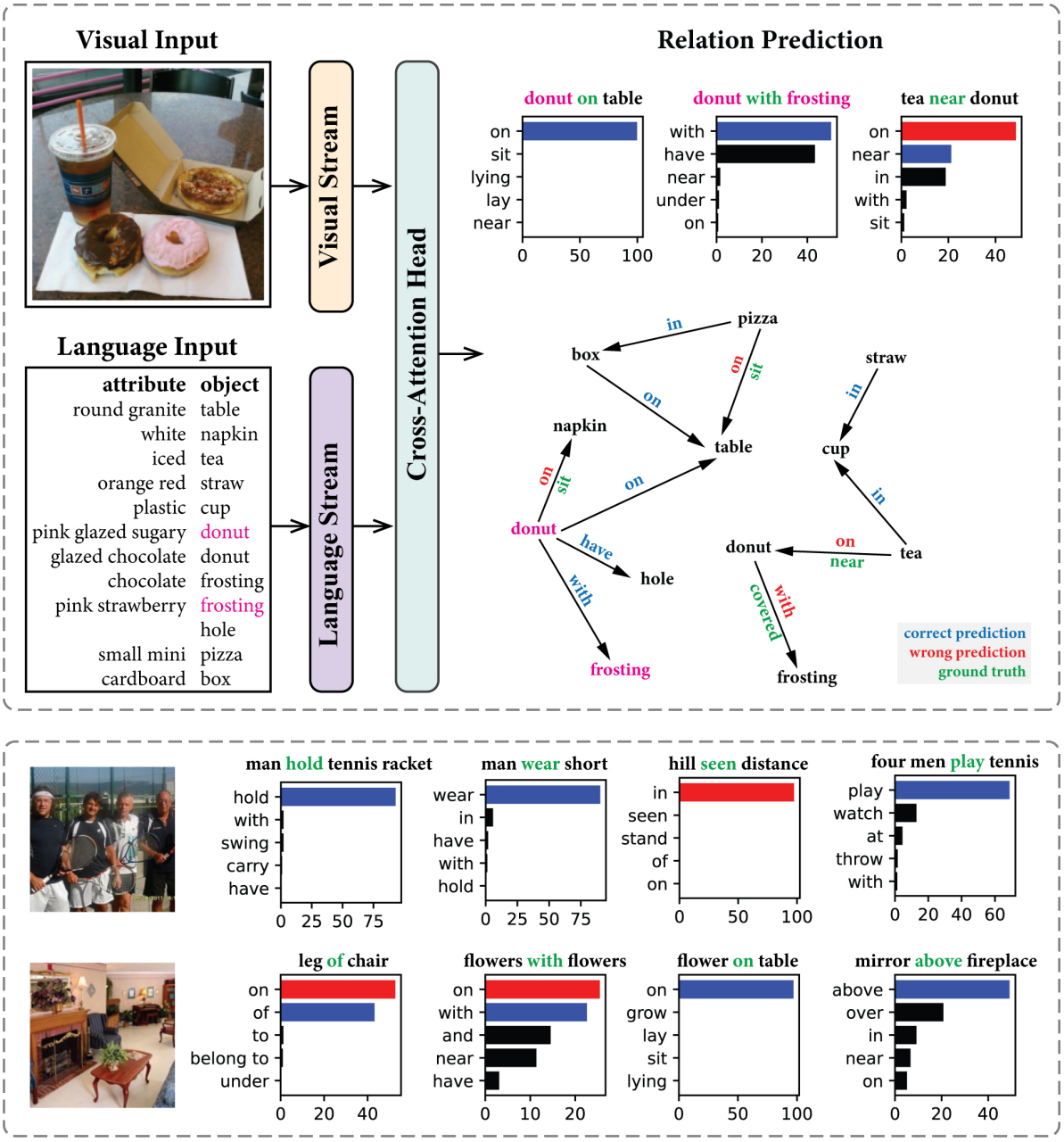


Figure 4.15: **Test examples of visual relation prediction from the testing dataset.** Top: For visual grounding of object relations, the visual input is a natural image and the language input is a set of object descriptions. The bar charts show the examples of top-5 predicted visual relations for pairs of objects (e.g., donut and table). The directed graph shows top-1 predicted visual relations on all object pairs in the given example. Each arrow points from a subject to its corresponding object. Along each arrow, the relation marked in blue indicates that the top-1 model prediction is the same as the ground truth. The relation marked in red indicates the top-1 model prediction is wrong, and the ground truth label below is marked in green. Bottom: Other examples.

4.4 Summary and Discussion

To summarize this chapter, we have built a two-stream deep neural network for visually grounded language learning. The visual stream and the language stream are independent feature extractors pretrained with a large unimodal dataset. Then the two streams are associated by grounding based on natural images paired with textual descriptions (Section 4.2.5.2) and object relations (Section 4.2.5.3). The model shows good performance on cross-modal alignment (Section 4.3.2), image-text retrieval (Section 4.3.3), and visual relation prediction (Section 4.3.4).

Prior work uses a triplet loss with paired visual-audio input to jointly learn cross-modal alignment and representations from two modalities [68]. We use normalized the temperature-scaled cross entropy (NT-Xent) loss [26]. The triplet loss contrasts the similarity between an anchor sample and a positive sample vs. the similarity between the anchor example and a negative sample. In contrast, the NT-Xent loss contrasts one positive sample with a noise distribution generated from many negative samples. The latter one tends to force the model to learn a better-structured representational space by leveraging contrastive learning with more negative samples.

Our model is different from recent works that use transformer encoders with multiple layers of cross-modal attention for visual-language learning [163, 114, 161, 28]. We intentionally delay the stage of multimodal fusion and keep the same architecture as Bert [35] in the language model. The language stream is a stand-alone system that can be detached from the visual stream after visual grounding. Although early fusion of multimodal information can support better performance in cross-modal tasks, it also loses the possibility to obtain a separable language system that may support the representational learning of grounded semantics. Since this work focuses on how the joint learning of multimodal information reshapes the language model, we are particularly interested in evaluating the intrinsic properties of the grounded word embedding space (See details in Chapter 5) and applying the grounded semantic representation to explain brain data (See details in Chapter 6).

Note that our model only requires paired visual-language input without further annotations (e.g., the object bounding box). Therefore, it can be easily generalized to large-scale data for multimodal representation pretraining, as explored in a recent study [79]. Our study utilizes two types of datasets: images with captions and images with scene graphs describing visual objects and their relations. The image captioning data is used to "pretrain" the two-stream model to align the visual output and the language output in a shared representational space. Images with scene graphs are used to model visual object relations, by adding two additional modules to the pretrained two-stream model - a cross-attention module to extract visually grounded object representations and a bilinear operator to encode relational embeddings. In future, we can use a similar strategy to finetune our model on other tasks and datasets, such as visual reasoning and visual question answering, to further

ground language representations to more demanding cognitive experience.

Again, the primary focus of this work is to fuse multimodal information for refining unimodal feature extractors by cross-modal learning. Especially we want to merge visual knowledge into the language stream to understand the effect of perceptual grounding on language learning inspired by the grounded cognition theory [8]. Such an architecture should be generalizable to multiple modalities (including but beyond vision) by including other independent streams (See detailed discussions about grounding language in other modalities in Section 7.1). Another standard visual-language learning model follows an encoder-decoder framework inspired by the language translation models [162], focusing on learning a joint latent representation space for generative purposes, such as generating image captions [181]. An ability to generate data is critical to machine and human intelligence, as explained by this famous quote from Richard Feynman: *"What I cannot create, I do not understand."* I would like to highlight that the two-stream (or dual-encoder) and encoder-decoder structures are not incompatible. Instead, after learning a more informative semantic representational space grounded in different modalities, we will be able to further train a decoder for performing generative tasks. Considering how humans learn to understand and produce language, these two abilities may be gradually and collaboratively acquired during the early stage of language development. Thus, it is worth exploring for future studies to add a generative component that mimics language production with the grounded language model.

CHAPTER 5

Visually Grounded Semantic Space

5.1 Rationale and Overview

[1] In Chapter 4, we have developed an approach to ground language learning in vision by training a two-stream model with three progressive stages - the unimodal pretraining, grounding natural language, and grounding visual object relations. To understand and assess how visual grounding affects the learned distribution of semantic representations, in this Chapter we systematically evaluate the semantic space grounded in vision vs. the ungrounded semantic space learned from pure texts.

After training the two-stream model, we treat its language stream as a stand-alone language model. We first extract the word embeddings of a large vocabulary set [153], which consists of $9,197$ commonly used English words. These words are further segregated into 100 word categories. We apply principal component analysis to the representations of all words in this vocabulary set to linearly decompose the semantic space into orthogonal bases. We then examine the principal axes of the semantic space, which are the top principal components that carry the largest variance in word representations. In addition, we evaluate whether the semantic norms of concepts defined by human understandings [34] are predictable by the word representations through logistic regression. We also assess word similarity [41] and word clustering [153] before and after visual grounding. We further test whether the visually grounded language model enables compositional language understanding based on visual knowledge and multimodal image search with queries based on images, texts, or their combinations.

5.2 Approach

We first extract the word embeddings from the language models studied herein, which share a Bert-based structure but has been trained with different levels of visual grounding (**Bert**: no

[1] This chapter is based on a conference paper [190] (under review).

visual grounding; **Grounded**: visual grounding of natural language (Section 4.2.5.2); **Relational Grounded**: visual grounding of object relations (Section 4.2.5.3)). To do this, we input every single word (or phrase) preceded with a special token `[CLS]` and followed by a special token `[SEP]` (according to the original Bert [35] paper) into the language stream, and extract the average pooled output at the last hidden layer as the output embedding of a given word or phrase. Since some feature channels have much greater values than other channels, we further use the mean and standard deviation of the output embeddings from the $30,522$-token vocabulary [179] to standardize the representation in each channel. The same process has been applied to both the Bert model and the visually grounded language models.

For each input word (or phrase), its output representation is a d-dimensional vector ($d = 768$). All embeddings of commonly used English words from the vocabulary set S [153] form a set of vector representations (as denoted by a 2D matrix $\boldsymbol{X} \in \mathbb{R}^{|S| \times d}$ in the following sections) in this high-dimensional semantic space (as denoted by V in the following sections), where each row in $\boldsymbol{X}$ is the vector representation of a single word $w \in S$.

5.2.1 Principal component analysis of word representations

Applying principal component analysis (PCA) to $\boldsymbol{X}$ defines a new coordinate system that spans the semantic space. This coordinate system uses orthogonal bases ordered by the corresponding portion by which each basis explains the variance of all word representations in $\boldsymbol{X}$. We then project $\boldsymbol{X}$ onto these orthogonal bases, where the top few dimensions, i.e., the top principal components from the principal component analysis, are viewed as the "principal axes" in the semantic space. By visual inspection of the word distribution projected onto each principal axis, we further evaluate the interpretability of each principal axis against human intuitions and neurobiological knowledge for the grounded vs. ungrounded models.

For PCA, we first apply the singular value decomposition (SVD) to $\boldsymbol{X}$ (Eq. 5.1) and then project $\boldsymbol{X}$ onto the orthogonal space spanned by the columns of $\boldsymbol{W} \in \mathbb{R}^{d \times d}$ (Eq. 5.2):

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{W}^T, \tag{5.1}$$

$$\boldsymbol{T} = \boldsymbol{X}\boldsymbol{W}, \tag{5.2}$$

where $\boldsymbol{T} \in \mathbb{R}^{|S| \times d}$ is the score matrix. Each row in $\boldsymbol{T}$ represents a word $w \in S$ in the new coordinate system defined by the columns in $\boldsymbol{W}$, which are the principal axes derived from $\boldsymbol{X}$.

5.2.2 Relating semantic representations to human understandings

After applying PCA, we calculate the Pearson's correlation between word representations projected onto the first principal axis (i.e., the first column in $\boldsymbol{T}$) and the human rating of word concreteness collected from a prior study [24], and compare the results across the **Bert** language model, the **Grounded** language model, and the **Relational Grounded** language model. The aim is to test whether the visually grounded language model is more capable of capturing the fundamental semantic features (specifically, the concrete-abstract attribute) without explicitly trained to do so.

We further investigate whether the visually grounded word embeddings are capable of predicting semantic features defined by humans, according to the standard evaluation methods established in prior studies [105, 185]. For this purpose, we adopt the concept property norm dataset collected from the Centre for Speech, Language and the Brain (CSLB) [34]. This dataset includes binary semantic features (e.g., `has_wheels`) labeled for 638 concepts by 123 human participants. We hypothesize that the word embeddings can be read out through a linear and sparse projection to readily support binary classification attainable by humans, especially after visual grounding. To test this hypothesis, we train a logistic regression model with L1 regularization to predict each binary semantic feature from the corresponding word representation, and repeat the same process for the **Bert**, the **Grounded**, and the **Relational Grounded** language models for comparison.

Since many binary semantic norms contain only a few positive samples in the CSLB dataset [34], we first filter out the feature norms with less than 5 positive samples and retain 390 out of 2725 feature norms that are assigned to 5 feature types according to the CSLB dataset: 156 "visual perceptual" features (e.g., `has_wheels`); 29 "other perceptual" features (e.g., `has_flavors`); 94 "functional" features (e.g., `does_cut`); 65 "encyclopaedic" features (e.g., `is_dangerous`); 46 "taxonomic" features (e.g., `is_clothing`).

For the i-th semantic norm, we build a binary classifier with a logistic regression model p^i as:

$$p^i(y_{ij} = 1|\boldsymbol{x_j}) = \sigma(\boldsymbol{w_i}^T \boldsymbol{x_j}), \tag{5.3}$$

here $\boldsymbol{x_j}$ is the word representation of the j-th word x_j in the CSLB dataset after transforming the representation to the orthogonal bases obtained with PCA. $\boldsymbol{w}_i$ is a linear weight specific to the i-th semantic norm. $y_{ij} \in \{0, 1\}$ is the binary label indicating whether the x_j holds the i-th semantic norm. We add an L1-norm of $\boldsymbol{w}_i$ to the loss function, as the sparsity constraint, to avoid overfitting when training the logistic regression model.

$$\boldsymbol{w_i}^* = \arg\min_{\boldsymbol{w_i}} \Big(-\sum_j \big[y_{ij} \log\big(\sigma(\boldsymbol{w_i}^T \boldsymbol{x_j})\big) + (1 - y_{ij}) \log\big(1 - \sigma(\boldsymbol{w_i}^T \boldsymbol{x_j})\big) \big] + \lambda_i \|\boldsymbol{w_i}\|_1 \Big), \tag{5.4}$$

where $\lambda_i \in \mathbb{R}^+$ is the regularization parameter, which is determined by a leave-one-out cross validation to minimize the following objective function

$$\mathcal{L}_i(\lambda_i) = \sum_j \mathcal{L}_{ij}(\lambda_i). \tag{5.5}$$

Suppose $P_i = \{k|y_{ik} = 1\}$ and $N_i = \{k|y_{ik} = 0\}$ are the sets of positive and negative word samples for the i-th semantic feature, respectively ($|P_i \cup N_i| = 638$, i.e., the total number of words in the CSLB dataset). Then $\mathcal{L}_{ij}(\lambda_i)$ is defined as

$$\mathcal{L}_{ij}(\lambda_i) = \frac{1}{|P_i|} \sum_{k \in P_i, k \neq j} (\log p^i_{\lambda_i,j}(y_{ik} = 1|\boldsymbol{x_k})) + \frac{1}{|N_i|} \sum_{k \in N_i, k \neq j} (\log p^i_{\lambda_i,j}(y_{ik} = 0|\boldsymbol{x_k})), \tag{5.6}$$

here j indicates the left-out word sample, and $p^i_{\lambda_i,j}$ is the corresponding regression model trained with regularization parameter λ_i. After λ_i is determined by minimizing Eq. 5.5, we train the L1-norm regularized logistic regression model p^i with all word samples and calculate the F_1-score for positive labels

$$\mathrm{F}_1 = \frac{\mathtt{tp}}{\mathtt{tp} + \frac{1}{2}(\mathtt{fp} + \mathtt{fn})}, \tag{5.7}$$

where `tp` is the number of true positive cases, `fp` is the number of false positive cases, `fn` is the number of false negative cases. For each of the five semantic norms, we then compare the F_1-score obtained with different language models and assess their pairwise differences for statistical significance with one-sided Wilcoxon Signed Rank Test [177].

5.2.3 Assessing word similarity

We also test whether visual grounding reshapes the semantic space to drag words with similar conceptual meanings and semantic features closer to each other. For different language models, we evaluate the similarity between word pairs. The evaluation is based on the WordSim-353 dataset [41], which contains 353 pairs of nouns with the similarity score rated by human judges in a numeric scale from 0 to 10. Unless stated otherwise, all similarity between a pair of word representation $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$ is evaluated by the pair-wise cosine similarity:

$$\mathtt{Similarity}(\boldsymbol{x_i}, \boldsymbol{x_j}) = \cos(\boldsymbol{x_i}, \boldsymbol{x_j}). \tag{5.8}$$

For simplicity, we directly assess the Pearson's correlation between the similarity metric captured by the language model (Eq. 5.8) and the human rated similarity score. The results are compared across different language models. We also calculate the similarity between all pairs of words in the

vocabulary set S [153] as a baseline distribution separately evaluated for **Bert**, the **Grounded**, and the **Relational Grounded** language models.

5.2.4 Clustering by word categories

After visual grounding, we hypothesize that the semantic representations group themselves based on perceptual similarity. To test this, we use the SemCat dataset ($9,197$ English words from $N = 100$ categories) [153] and calculate the Silhouette coefficient (ranges from -1 to 1) to measure the degree by which these words are clustered in the space into human-defined word categories. The distance metric d between word embeddings is measured as the cosine distance: $d(\boldsymbol{x_i}, \boldsymbol{x_j}) = 1 - \cos(\boldsymbol{x_i}, \boldsymbol{x_j})$. Let us denote N as the number of word categories, $\boldsymbol{x_i}$ as the embedding of word w_i which is assigned to the category C_i in the SemCat dataset, the degree to which word w_i falls in affinity with other words assigned to the same category is measured by the modified Silhouette coefficient $s(i)$ defined as

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, j \neq i} d(\boldsymbol{x_i}, \boldsymbol{x_j}), \tag{5.9}$$

$$b(i) = \frac{1}{N - 1} \sum_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} d(\boldsymbol{x_i}, \boldsymbol{x_j}), \tag{5.10}$$

$$s(i) = \frac{b(i) - a(i)}{\max\left(a(i), b(i)\right)}, \tag{5.11}$$

here we use the average instead of the minimum when calculating $b(i)$ in Eq. 5.10 since the 100 categories in the SemCat dataset are not mutually exclusive (e.g., in the SemCat dataset, category `mammal`, category `bird`, and category `fish` have overlapping word samples with category `animal`).

To assess a group-level metric for categorization, we simply average the Silhouette coefficient for words in each category C_i and obtain $N = 100$ values from -1 to 1. We then compare these category-level clustering metrics between different language models and evaluate their pairwise differences for statistical significance with one-sided Wilcoxon Signed Rank Test [177].

5.2.5 Semantic compositionality based on visual knowledge

As mentioned in the introduction, the ungrounded semantic distributional models have a critical drawback that the learned language model is unable to know the perceptual knowledge about a concept, thus lacking the ability to make visually informed compositional reasoning, e.g., *zebra is a horse with black and white stripes*. Such knowledge is rarely or never exposed to the language model during training with natural language corpa alone [66]. Since the visual implication about a concept of *zebra* is too obvious by human experiences for anyone to explicitly describe a zebra like

this in language. Similarly, we don't say *yellow banana* since *yellow* is an intrinsic feature for a *banana*.

However, it is more plausible for the grounded language models to capture conceptual composition from visual information, since the model has been already trained with paired image-text and thus the language model has been exposed to the visual perceptual features when learning the semantic representation of concepts. To test this hypothesis, we use a few example words of which the meanings can be inferred by a combination of concepts based on human intuition, especially based on visual perception. We examine how the embedding of a composed query phrase $\boldsymbol{q}$ (e.g., striped horse) matches with the common English words in S by measuring their cosine similarity in the semantic space $S_q(k) = \texttt{Similarity}(\boldsymbol{q}, \boldsymbol{x_k})$. We rank all words in the vocabulary by their similarity to the query phrase,

$$\texttt{Ranking}(\boldsymbol{q}) = \texttt{sort}[S_q(k)] = \texttt{sort}[\cos(\boldsymbol{q}, \boldsymbol{x_k})]. \tag{5.12}$$

We compare the so ranked words against human intuition when the word representations are based on the **Bert**, the **Grounded**, or the **Relational Grounded** language models.

5.2.6 Multimodal image search in the joint representational space

After the two-stream model is established and trained with visual grounding of natural language and object relations (See section 4.2.5.3), the cross-attention module establishes a joint representational space for both visual and textual data. We further explore whether this joint space can support cross-modal tasks, e.g., image search based on image, text, or their combinations [79]. For this specific task, two additional heads F_V and F_L are added to the model. Each head includes two linear layers with ReLU in between followed by average pooling:

$$\boldsymbol{Q}_I = F_V(\text{Key}_V) = \frac{1}{HW}\sum_{i,j}\Big(\big(\text{ReLU}(\text{Key}_V[i,j,:]\boldsymbol{W}_V^1 + \boldsymbol{b}_V^1)\big)\boldsymbol{W}_V^2 + \boldsymbol{b}_V^2\Big), \tag{5.13}$$

$$\boldsymbol{Q}_W = F_L(\text{Query}_L) = \frac{1}{K}\sum_{k}\Big(\big(\text{ReLU}(\text{Query}_L[k,:]\boldsymbol{W}_L^1 + \boldsymbol{b}_L^1)\big)\boldsymbol{W}_L^2 + \boldsymbol{b}_L^2\Big), \tag{5.14}$$

where Key_V or Query_L are from the cross-attention module described in Fig. 4.5 and Eq. 4.18, after being concatenated across all attention heads along the feature dimension. $\boldsymbol{W}$s and $\boldsymbol{b}$s are the weights and biases of the linear transformations in these two head functions, with size $d \times d$ and $d \times 1$ ($d = 768$). H and W are the height and width of the image feature output ($H = W = 14$). K

is the number of words in an image caption, which varies for different language inputs. F_V and F_L are applied to visual and textual representations (Key_V or Query_L) in the joint space respectively, and result in a single vector representation for either an image (denoted as $\boldsymbol{Q}_I \in \mathbb{R}^d$) or a text (denoted as $\boldsymbol{Q}_W \in \mathbb{R}^d$).

In this step of transfer learning, the trained model described in Section 4.2.5.3 is frozen. The two additional heads are trained with contrastive loss to match the average-pooled representations of paired images and texts in terms of their cosine similarity using the MS COCO dataset, similar to the method described in Section 4.2.3. The loss function is similar to the one describe in Eq. 4.16 and Eq. 4.17, except the similarity score becomes $S(I, W) = \cos(\boldsymbol{Q}_I, \boldsymbol{Q}_W)$.

To use the model for multimodal image search, we apply a weighted sum to the normalized representations (Eq. 5.15) of a query image and a query text. The linear weighting α (for text) and $(1 - \alpha)$ (for image) range from 0 to 1,

$$\boldsymbol{Q}_I \leftarrow \frac{\boldsymbol{Q}_I}{\|\boldsymbol{Q}_I\|_2}, \quad \boldsymbol{Q}_W \leftarrow \frac{\boldsymbol{Q}_W}{\|\boldsymbol{Q}_W\|_2}, \tag{5.15}$$

$$\boldsymbol{Q}_{\text{search}} = (1 - \alpha)\boldsymbol{Q}_I + \alpha\boldsymbol{Q}_W. \tag{5.16}$$

we then use the multimodal query defined in Eq. 5.16 to search a held-out database[2] for the matched images ranked in terms of cosine similarity.

As mentioned above, α controls the weighting between the textual and visual queries. We test how the image search returns different results as α increases from 0 (image only) to 1 (text only).

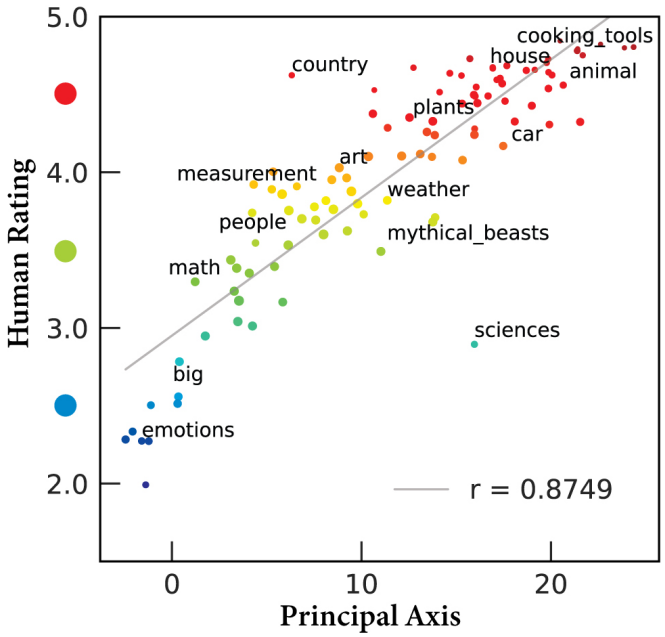
5.3 Results

5.3.1 Principal axes capture explainable semantic attributes

Interestingly, the first principal dimension in the visually grounded semantic space is readily interpretable as an abstract-to-concrete axis (Fig. 5.1). For example, words that end up with the highest values when they are projected on this axis are *ostrich, seagull, albatross, blender, pelican, broccoli, parakeet, lettuce, sailboat, vegetables*, whereas words with the lowest values are *displeasure, liking, to, outgoing, present, experienced, profitable, faithful, meaningful, multitude*. The representations of words along this axis is significantly correlated with human rating of their concreteness (ranging from 1 to 5) from a prior study [24] (Fig. 5.1). For the **Grounded** model, the Pearson correlation coefficient reaches 0.8749 or 0.6615 across word categories or words, respectively. For the **Relational Grounded** model, the Pearson correlation coefficient reaches

[2] 41,600 images from the validation dataset of Open Images Dataset $V6$.

0.8001 for word categories and 0.6948 for words.



Examples:

word	category	principal axis	human rating
oven	cooking tool	30.86	4.97
zebra	animal	24.74	4.86
car	car	22.06	4.89
furniture	house	20.83	4.89
defense	country	4.65	4.19
chemistry	sciences	18.67	3.64
wood	plants	16.22	4.85
cartoon	art	11.84	4.33
humid	weather	11.58	3.48
angel	mythical beasts	7.22	3.82
lover	people	6.01	3.68
thousand	math	3.65	3.07
huge	big	0.17	3.54
cheerful	emotions	- 2.54	2.34

Figure 5.1: **The first principal component in the grounded semantic space captures the concrete-abstract axis of semantics.** Left: Each dot represents a word category with the color indicative of the averaged human-rated concreteness (the y axis) and the size proportional to the standard deviation. The x axis indicates the corresponding value of the word representations projected onto the first principal axis. Right: Example words in labeled categories.

Table 5.1: Correlation between the 1st principal component and human rating of word concreteness

	Correlation (Pearson's r)		
Group	Bert	Grounded	Relational Grounded
word-level	0.1040	0.6615	**0.6948**
category-level	0.3538	**0.8749**	0.8001

In contrast, the first principal axis of the ungrounded semantic space from the baseline **Bert** model is not straightforward to interpret and shows a much weaker correlation with human ratings of concreteness ($r = 0.3538$ for categories, $r = 0.1040$ for words). The comparison between different language models is further summarized in Table 5.1 and visualized in Fig. 5.2.

Besides the first principal component being interpreted as abstract-concrete axis, other principal components are also intuitively explainable. For example, PC 2 captures the human vs. non-human axis, PC 3 captures the object vs. scene axis, PC 4 captures the artificial vs. natural axis, PC 5

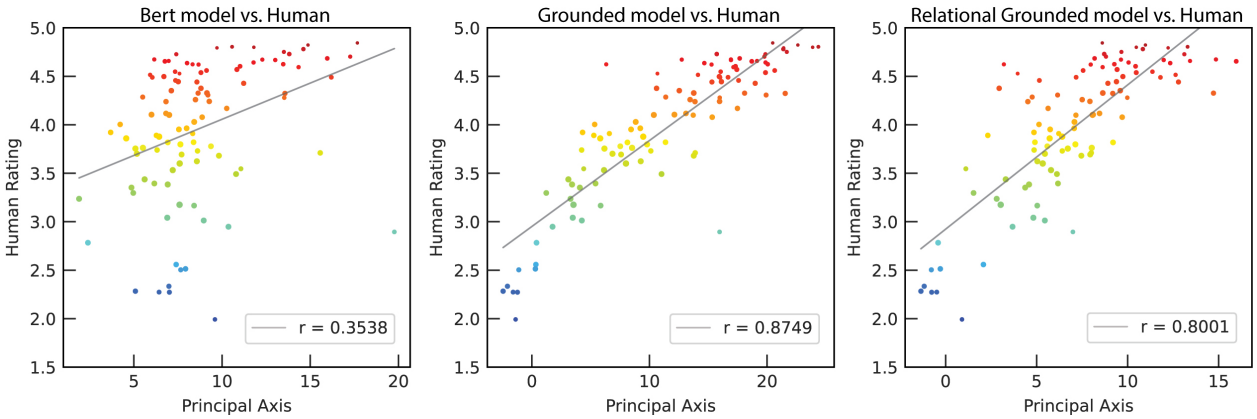


Figure 5.2: The first principal component in the word representation space captures the concrete-abstract axis only after visual grounding. The color indicates the concreteness rating of a category (blue: abstract; red: concrete).

captures the outdoor vs. indoor axis, PC 6 highlights words related to food. The distributions of all word categories after being projected onto the first 6 principal axes are visualized in Fig. 5.3.

5.3.1.1 2D visualization of the first three principal components

To visualize the grounded semantic representation in a subspace spanned by its first three principal axes, we first color-code each word category by the three dimensional RGB code according to the following associations: (PC1, red), (PC2, green), (PC3, blue). Only for the sake of visualization, the coefficient associated with each principal component have been linearly re-scaled into the range $[0, 1]$. We then project the three dimensional representations of the 100 word categories onto three 2D subspace, as shown in Fig. 5.4 (PC2 vs. PC3), Fig. 5.5 (PC1 vs. PC2), and Fig. 5.6 (PC1 vs. PC3).

The results suggest that the first quadrant of the PC2. vs. PC3 plane (as shown in Fig. 5.4) captures concepts describing natural scenes (e.g., biomes, rocks), the second quadrant captures concepts related to scenes with human activities (e.g., roadways, rooms), the third quadrant encodes human related non-scene concepts (e.g., jobs, musical instruments), the fourth quadrant encodes non-human objects (e.g., animal, foodweb). In addition, the abstract words (e.g., positive words, negative words, emotions) tend to be squeezed around the origin in this subspace.

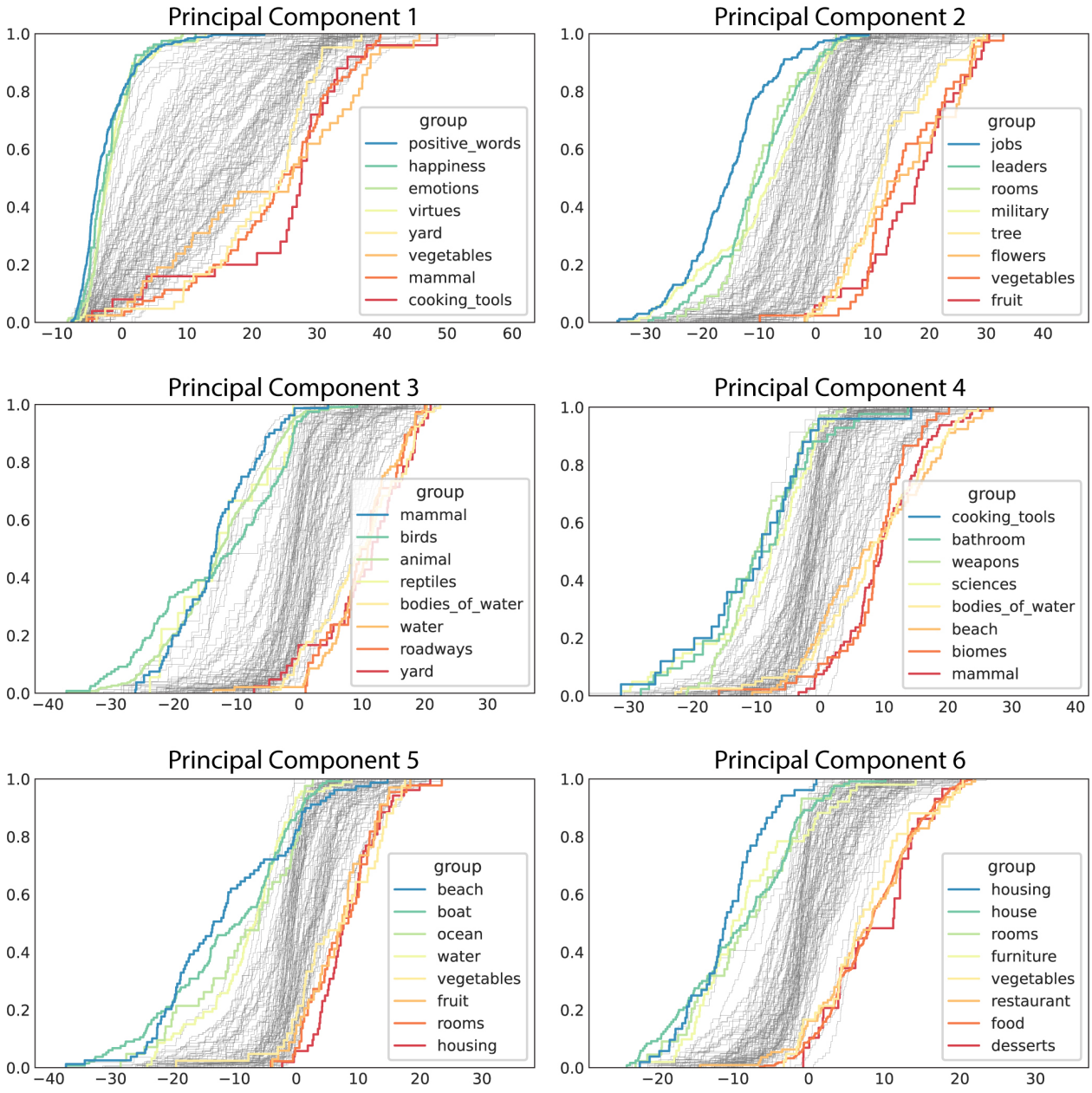


Figure 5.3: **Other principal components in the visually grounded word representation space.** Each plot shows a set of cumulative distribution functions (CDFs) for every word categories after being projected onto a principal axis. The principal dimensions capture the semantic attributes that can be interpreted by human intuition. PC1: abstract vs. concrete; PC2: human vs. non-human; PC3: object vs. scene; PC4: artificial vs. natural; PC5: outdoor vs. indoor; PC6: non-food vs. food.

Explainable Pricipal Components in the Visually Grounded Semantic Space

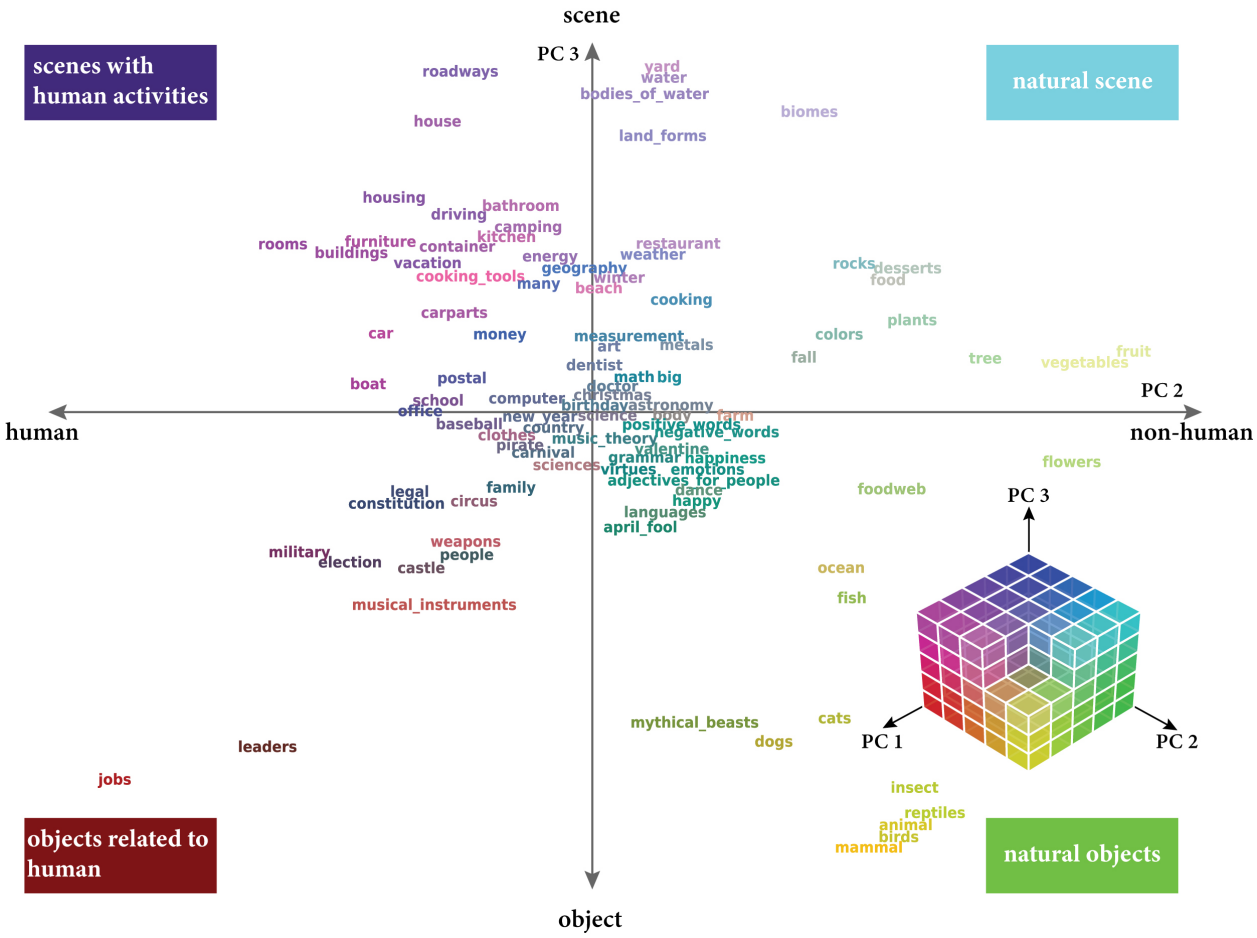


Figure 5.4: 2D visualization of PC2 and PC3.

Explainable Pricipal Components in the Visually Grounded Semantic Space

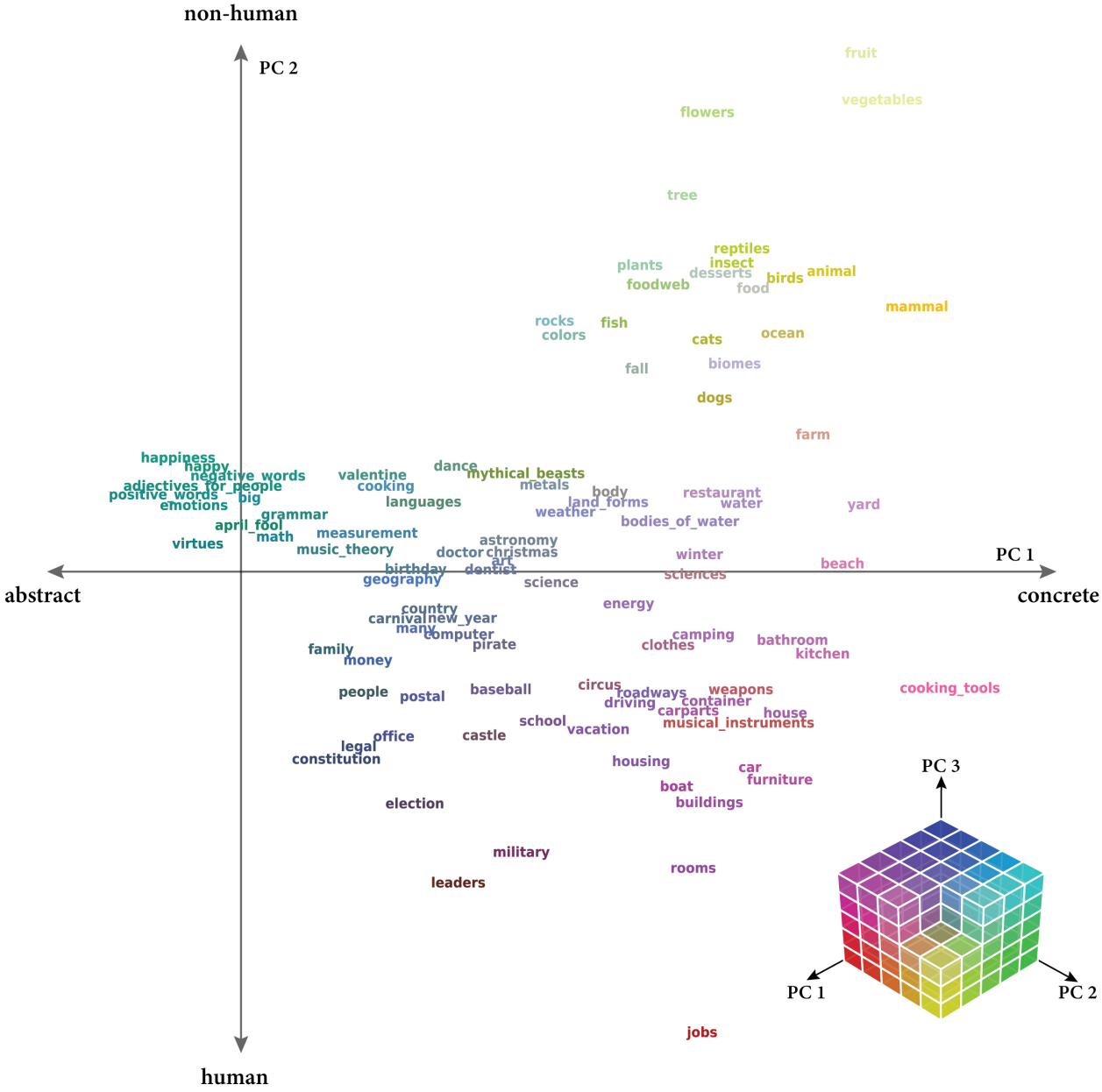


Figure 5.5: 2D visualization of PC1 and PC2.

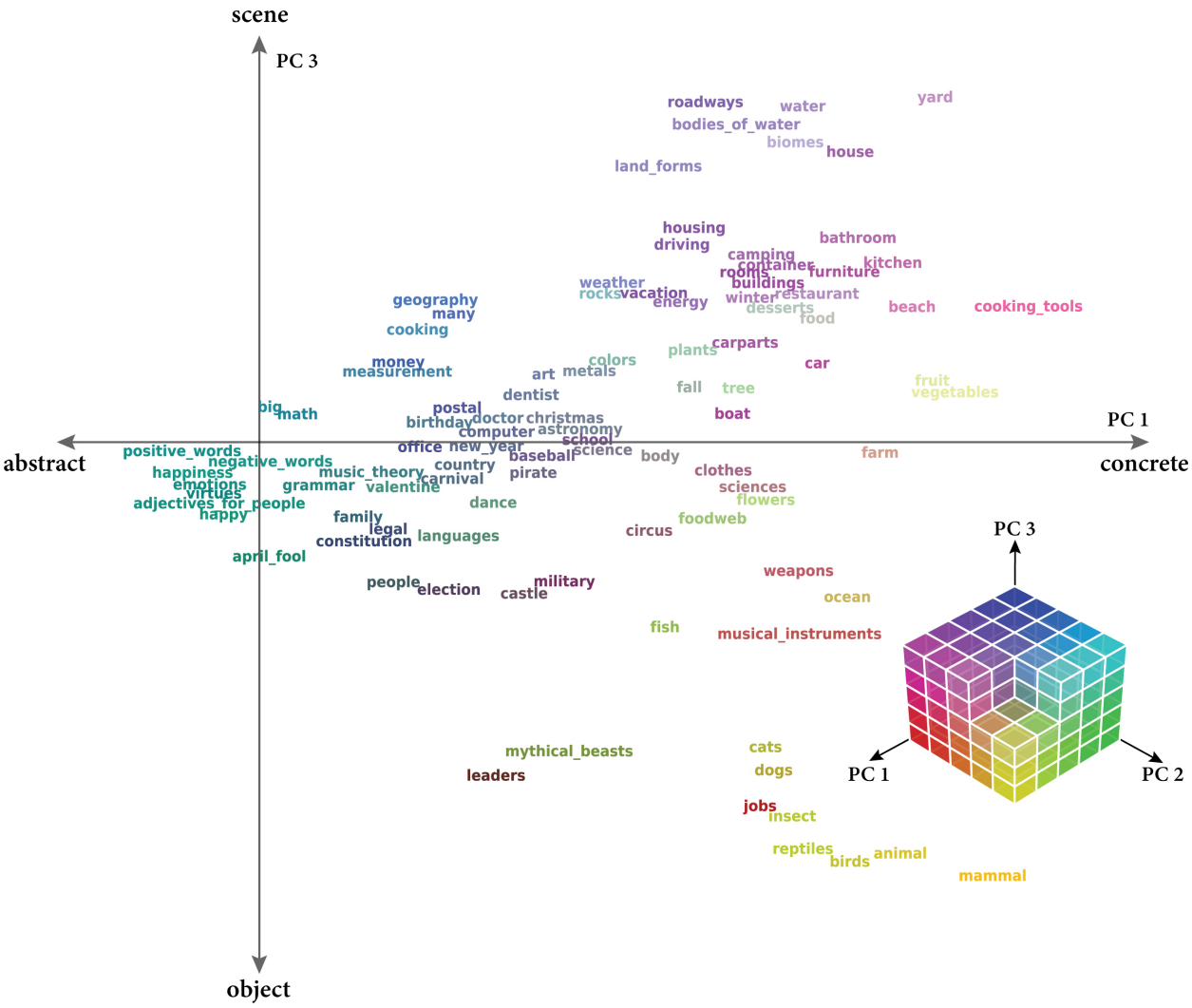


Figure 5.6: 2D visualization of PC1 and PC3.

In the subspace of PC1 vs. PC2 (Fig. 5.5) or PC1 and PC3 (Fig. 5.6), we also observe that although concrete concepts are distributed and scattered widely, abstract concepts are relatively squeezed around the origin and are not spread out to form separable distributions. This is perhaps due to the lack of information for the model to separate emotional words, which is not surprising since we only ground language learning in vision.

5.3.2 Predicting human-defined semantic features

Fig. 5.7 shows the F_1 scores indicative of how well word representations can predict human-defined binary semantic features through the logistic regression for the ungrounded **Bert** model, the

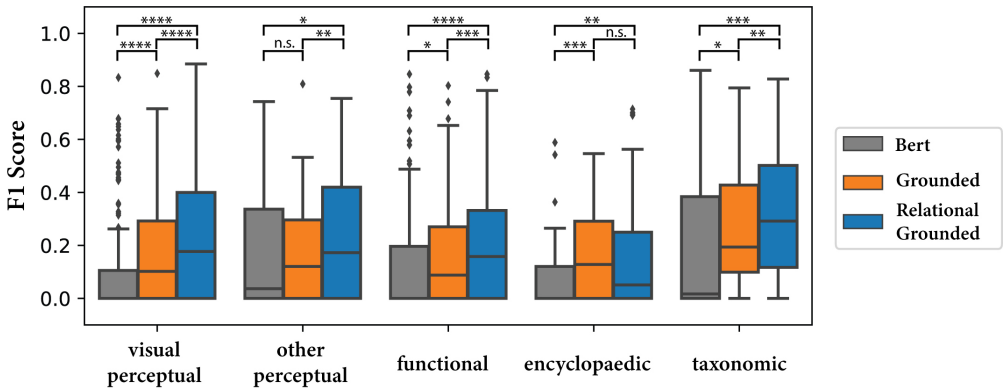


Figure 5.7: **The F1 score of predicting semantic feature norms from word representations before and after visual grounding.** Each box shows the lower (25%) percentile, the higher (75%) percentile, and the median of F1 scores within a feature type. Whisker$= 1.5$. Significant level: n.s.: not significant; *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$; ****, $p < 0.0001$.

Grounded model, and the **Relational Grounded** model.

The result (Fig. 5.7) suggests that the grounded word embeddings are significantly more predictive of binary features for "visually perceptual" than their ungrounded counterparts obtained by Bert (Wilcoxon Signed Rank Test; $p < 0.0001$). To a lesser but still significant level, this difference also applies to the "functional" features, "encyclopaedic" features, and "taxonomic" features. After visual grounding of object relations, the word embeddings tends to be more capable of correctly predicting the semantic norm features. However, the difference between the **Grounded** model and the **Bert** model on capturing "other perceptual" features is not significant, which is unsurprising since only vision is used to ground language learning.

Since we have added a strong L1-norm regularization term in the logistic regression model to avoid over-fitting (see Section 5.2.2), the results suggest 230 out of 390 semantic norms are not predictable by the ungrounded Bert model, while only 143 and 129 semantic norms are not predictable by the Grounded model and the Relational Grounded model respectively. We list the top-5 semantic norms that become more predictable after visual grounding based on the ranked difference in the F_1 score before and after visual grounding (Table 5.2).

Table 5.2: Top-5 semantic norms that are more predictable after visual grounding.

Feature type	**Grounded**	**Relational Grounded**
visual perceptual	has_wheels, has_a_handle_handles, has_skin_peel, has_pages, has_a_back	has_a_picture_pictures, has_pages, has_a_barrel, has_skin_peel, has_a_seat_seats
other perceptual	is_heavy, is_warm, does_smell_good_nice, is_juicy, has_flavours	is_warm, is_heavy, has_flavours, is_juicy, does_smell_good_nice
functional	does_fly, does_contain_hold, does_store, does_heat, is_used_to_see	does_fly, does_heat, does_cut, is_used_in_cooking, does_contain_hold
encyclopaedic	is_dangerous, is_found_in_seas, has_information, is_healthy, does_grow_on_trees	is_dangerous, has_information, does_grow_on_trees, is_found_in_seas, is_found_in_kitchens
taxonomic	is_clothing, is_a_weapon, is_a_vehicle, is_a_vegetable, is_transport	is_clothing, is_a_vehicle, is_a_vegetable, is_medicine, is_a_container

5.3.3 Visual grounding supports better and finer word categorization

5.3.3.1 Word similarity analysis on Wordsim-353 dataset

We test how well words are clustered by categories in the grounded vs. ungrounded semantic space, and further assess the Pearson's correlation between the model captured similarity metric (i.e., cosine similarity of paired word representations) and the human rated similarity score [41] (Fig. 5.8).

Table 5.3 shows the word-pair examples from the WordSim-353 dataset [41] with top-10 cosine similarity in the **Bert** model, the **Grounded** model, and the **Relational Grounded** model. The results suggest that most word pairs with significantly improved cosine similarity score after visual grounding have rich visual information in their meanings (e.g., (boy : lad), (car : automobile),

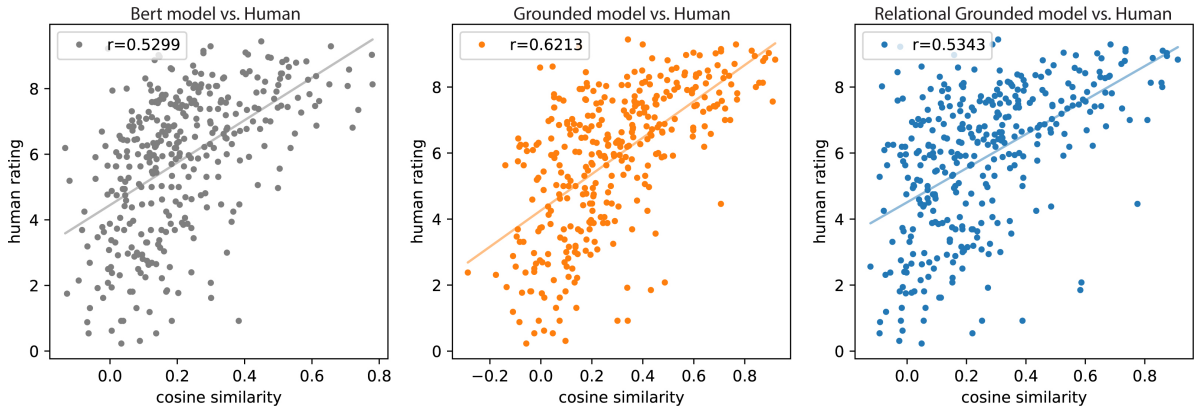


Figure 5.8: **The performance of language models on capturing word similarity.** r is the correlation between human-rated word similarity (y axis) and model-captured word similarity (x axis, measured by cosine similarity between two word embeddings). Each dot represents a word-pair in the Wordsim-353 dataset.

(furnace : stove), (street : avenue) etc.)

Besides, the averaged cosine similarity on the WordSim-353 dataset significantly increases ($p < 0.0001$; two-sided paired t-test after fisher z transformation) after visual grounding (See Fig. 5.9 and Table 5.4), while the baseline distribution of pairwise word similarity on common-word vocabulary S shows a similar pattern for all three language models (Fig. 5.10).

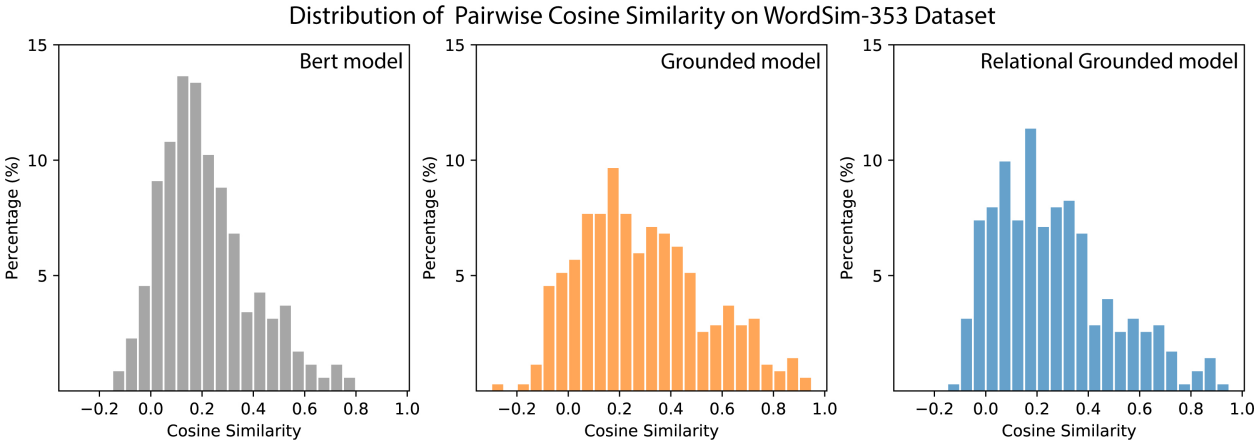


Figure 5.9: Distribution of word-pair cosine similarity on WordSim-353 dataset.

Table 5.3: Top-10 similar word pairs from Wordsim-353 assessed by the ungrounded and visually grounded language models. Two numbers are shown under each word pair. The first number indicates the model-captured word similarity (cosine; ranges from -1 to 1). The second number indicates the human-rated word similarity (ranging from 0 to 10).

Bert	**Grounded**	**Relational Grounded**
Harvard : Yale 0.78, 8.13	boy : lad 0.92, 8.83	boy : lad 0.91, 8.83
football : soccer 0.78, 9.03	tennis : racket 0.91, 7.56	car : automobile 0.87, 8.94
physics : chemistry 0.74, 7.35	football : soccer 0.89, 9.03	football : soccer 0.87, 9.03
football : basketball 0.72, 6.81	street : avenue 0.88, 8.88	coast : shore 0.86, 9.10
king : queen 0.71, 8.58	furnace : stove 0.87, 8.79	tiger : jaguar 0.86, 8.00
psychology : psychiatry 0.71, 8.08	Harvard : Yale 0.87, 8.13	vodka : brandy 0.86, 8.13
midday : noon 0.66, 9.29	vodka : brandy 0.86, 8.13	street : avenue 0.83, 8.88
vodka : brandy 0.65, 8.13	tiger : jaguar 0.84, 8.00	Harvard : Yale 0.82, 8.13
computer : keyboard 0.62, 7.62	car : automobile 0.84, 8.94	doctor : nurse 0.81, 7.00
drink : eat 0.61, 6.87	coast : shore 0.83, 9.10	lad : brother 0.78, 4.46

Table 5.4: The statistics (mean $\pm$ standard deviation) of the cosine similarity for the word pairs in the wordsim-353 dataset.

	Bert	**Grounded**	**Relational Grounded**
cosine similarity	0.22 ± 0.18	0.29 ± 0.24	0.26 ± 0.22

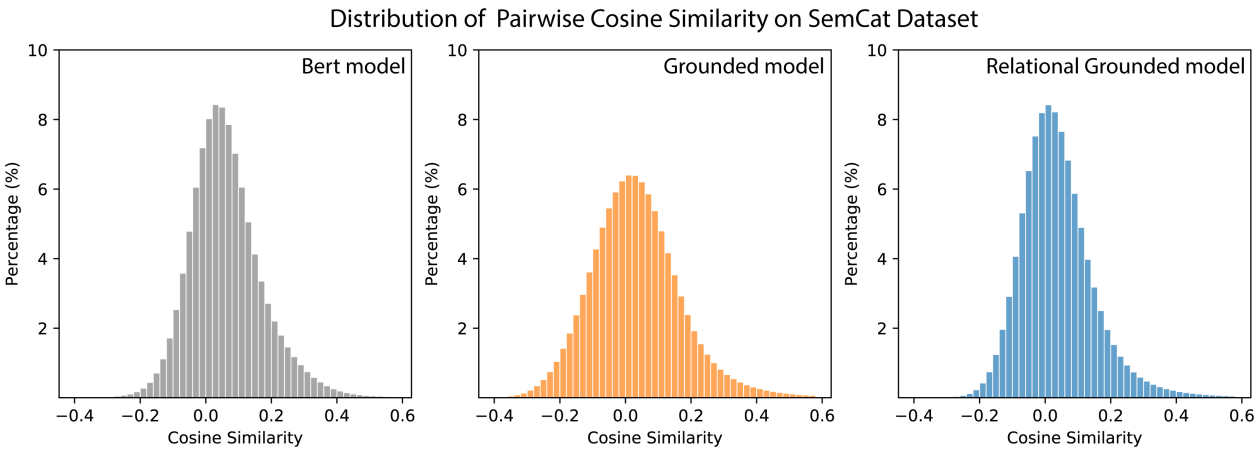


Figure 5.10: Distribution of word-pair cosine similarity on SemCat dataset.

5.3.3.2 Word categorization analysis on SemCat dataset

As shown in Fig. 5.11, after visual grounding, the Silhouette coefficients across 100 categories are significantly higher for the visually grounded semantics than ungrounded ones (Wilcoxon Signed Rank Test; $p < 0.0001$) (Fig. 5.11 left). The greatest gain in clustering is noticeable for categories that include concrete concepts (e.g., `car`, `housing`, `mammal`) with defining visual attributes (Fig. 5.11 right). For some abstract categories related to human emotion (e.g., `happy`), the grounded representations are also better clustered than the ungrounded ones.

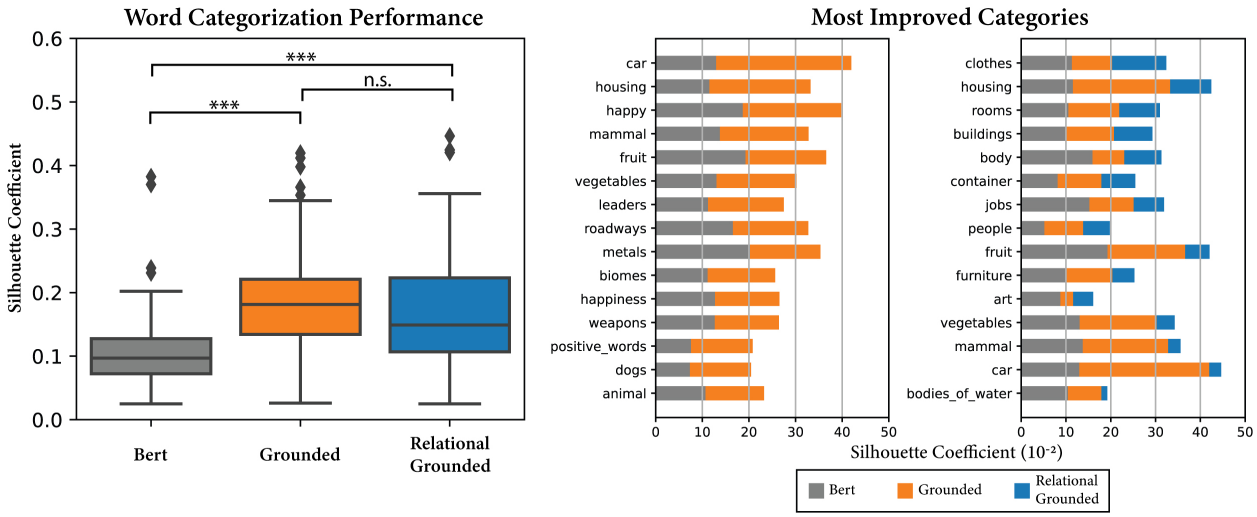


Figure 5.11: Left: A boxplot showing silhouette coefficients on word representations before and after visual grounding (whiskers=1.5). Right: Top-15 word categories that are better clustered after visual grounding of natural language and object relations.

We have also compared the word categorization performance after visual grounding of natural

language with different training settings (Fig. 5.12). The results suggest that earlier grounding (i.e., more learnable layers in Bert) tends to show better clustering by human-defined word categories. The models with frozen query and key transformations in Bert self-attention layers (as shown in blue bars) show similar categorization as the ones with learnable query and key weights for cross-modal training (as shown in orange bars). Models trained with larger dropout rate (0.3; as shown in opaque bars) show significantly higher performance on word categorization than its counterpart with smaller dropout rate (0.1; as shown in transparent bars).

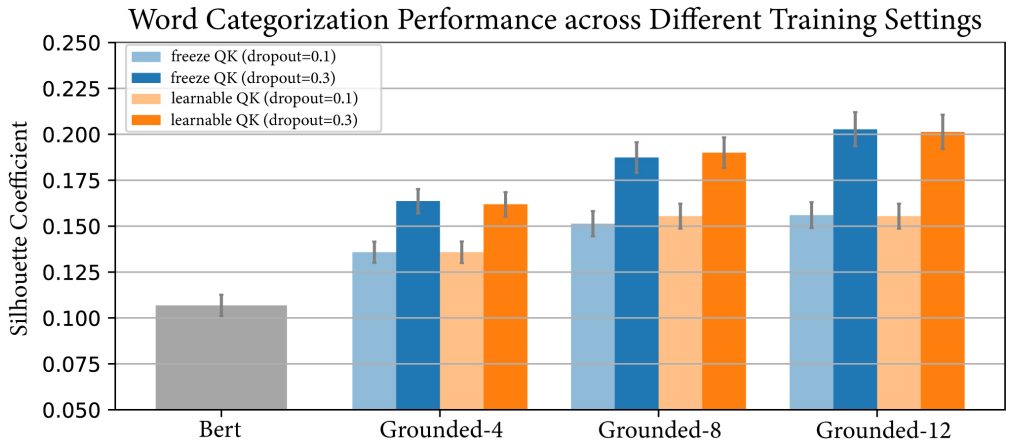


Figure 5.12: **Word clustering by category for comparative models with different training settings.** The y axis indicates the Silhouette Coefficient. Each bar shows the category-level clustering performance averaged across 100 categories. The error bar indicates the standard error.

To validate whether a better clustering performance on word w_i results from a higher sampling rate in the training dataset, we further calculate the correlation between the Silhouette coefficient $s(i)$ and the training occurrence rate of word w_i for all words in the vocabulary S. The result (Fig. 5.13) rejects this hypothesis by showing a non-positive correlation value between these two terms with both category-level ($r = -0.28$) and word-level ($r = -0.07$) analyses.

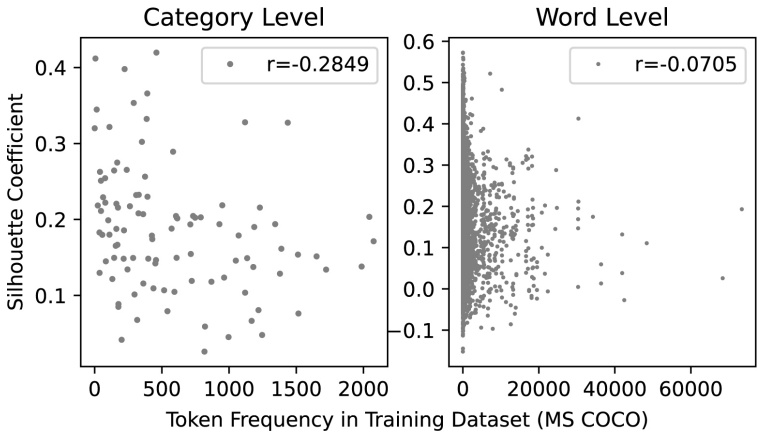


Figure 5.13: **Word clustering by category is uncorrelated with its occurrence rate during training.** Each dot represents a category (left figure) or a word (right figure). The y axis indicates the clustering performance measured by the Silhouette coefficient. The x axis indicates the occurrence of the corresponding word tokens in the training dataset. r is the Pearson's correlation between the clustering performance and training samples of all words or categories in the Semcat dataset.

Some general categories show further fine-grained clusters within themselves in the grounded semantic space. The following results are examples on word representation visualization for **vehicle** (Table 5.5, Fig. 5.14), **animal** (Table 5.6, Fig. 5.15), **food** (Table 5.7, Fig. 5.16), and **room** (Table 5.8, Fig. 5.17) subcategories. In each example, we first pick up three query words as the prototype for each subcategory (e.g., `boat`, `car`, `airplane` for **vehicle**). Then for each language model, we use the cosine similarity to sort out the top-15 closest words to each of these query words, as shown in the columns of these tables. We observe that only after visual grounding, the top similar words are well-aligned with the subcategory defined by the query word. For example, all words in the column under **Grounded** or **Relational Grounded** for the `boat` query belong to water transportation, but words `skeleton, nation, men, bike` found by Bert model are not water transportation. We further visualize the representation of the words in each subcategories from the **Grounded** model, by first calculating its cosine similarity to each of the query word (resulting in a three-dimensional cosine similarity representations), and then projecting these three-dimensional representations into the 2D plane spanned by a pair of query word. The visualization results shown in the following figures suggest that the representational distributions of word subcategories are separable only after visual grounding.

Table 5.5: Top-15 words for **vehicle** subcategories

query	**Bert**	**Grounded**	**Relational Grounded**
boat	canoe, sailboat, submarine, skeleton, nation, sailing, men, bike, ballast, boating, raft, motorboat, paddle, yacht, kayak	tugboat, yacht, riverboat, gunboat, canoe, boating, barge,dinghy, raft, sailboat, ship, steamboat, steamer, trawler, watercraft	riverboat, gunboat, canoe, sailboat, dinghy, tugboat, yacht, motorboat, ship, steamer, barge, steamboat, submarine, ferry, steamship
car	vehicle, jeep, sedan, truck, jaguar, auto, driver, motorcycle, chassis, motor, bike, boat, horse, speeding, automobile	sedan, suv, vehicle, automobile, limo, jeep, limousine, taxi, drive, traffic, auto, pickup, van, roadster, truck	sedan, limo, automobile, suv, jeep, van, limousine, taxi, truck, buggy, vehicle, cart, motorcycle, auto, convertible
airplane	plane, aircraft, propeller, automobile, airport, bird, flight, parrot, turbulence, fly, kite, butterfly, rocket, motorcycle, takeoff	plane, jet, flight, aircraft, takeoff, airport, corsair, pilot, hangar, undercarriage, missile, propeller, nuclear, nautical, flyby	plane, jet, aircraft, corsair, flight, balloon, missile, takeoff, automobile, cockpit, hangar, avian, rocket, freighter, nuclear

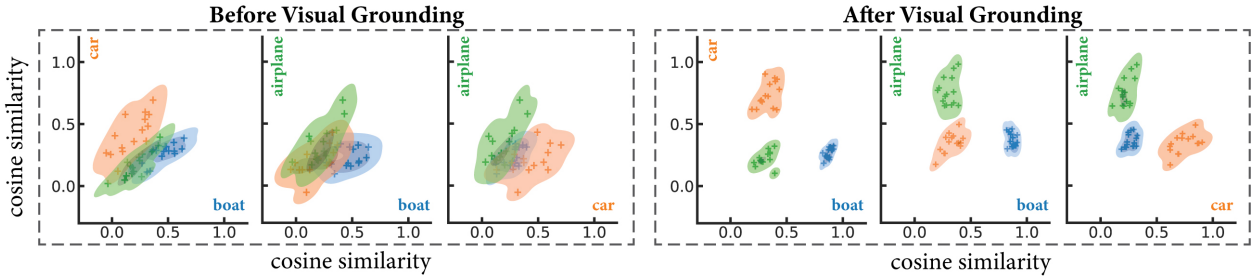


Figure 5.14: **Visualization of words related to vehicle subcategories.** Each plot shows a 2D plane expanded by the cosine similarity scores according to a pair of prototype words. Each dot represents a word color-coded by the prototype word (blue:boat; orange:car; green:airplane).

Table 5.6: Top-15 words for **animal** subcategories

query	**Bert**	**Grounded**	**Relational Grounded**
dog	pig, animal, bike, horse, mule, donkey, squirrel, bicycle, goat, monkey, motorcycle, moose, cat, gorilla, mouse	puppy, doge, terrier, canine, pug, hound, beagle, bulldog, dogwood, pup, mutt, spaniel, chihuahua, retriever, shepherd	puppy, doge, terrier, pug, canine, bulldog, beagle, hound, mutt, greyhound, bobcat, spaniel, hag, tomcat, donkey
goose	scare, geese, cow, calf, neighbor, puddle, flu, battleship, hog, displeasure, plank, herring, stir, scrambled, sock	geese, eagle, rooster, gull, pigeon, owl, duck, crow, parrot, partridge, falcon, harrier, sparrow, vulture, warbler	geese, eagle, pigeon, owl, parrot, duck, sparrow, rooster, falcon, crow, seagull, gull, warbler, partridge, harrier
horse	stallion, mule, dog, bike, trainer, boat, mare, car, animal, men, motorcycle, human, mountain, athlete, chestnut	stallion, mule, mare, donkey, seahorse, ox, steer, bull, unicorn, chestnut, foal, camel, oxbow, antelope, cow	stallion, mule, mare, seahorse, donkey, camel, bull, cow, cattle, ox, bison, deer, lassie, dog, animal

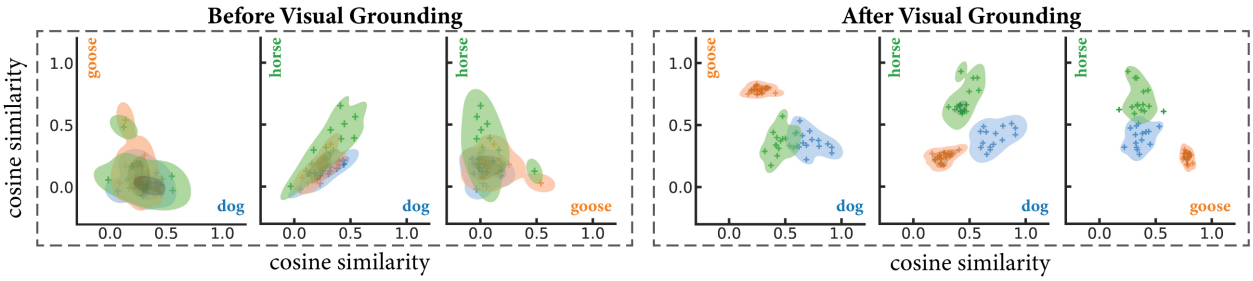


Figure 5.15: Visualization of words related to **animal** subcategories.

Table 5.7: Top-15 words for **food** subcategories

query	**Bert**	**Grounded**	**Relational Grounded**
drink	eat, feed, swallow, spend, study, spill, breathe, treat, bite, cough, drop, give, relax, wash, bleed	beverage, soda, cola, coke, juice, rum, bottle, champagne, drunk, flask, blender, mug, cup, blend, coffee	beverage, soda, cola, coke, juice, rum, flask, bottle, lemonade, coffee, cup, blend, mug, blender, brew
fruit	flower, foliage, citrus, flowers, plant, orchid, shrub, poisonous, inflorescence, eggs, omnivorous, seedling, nut, pineapple, snail	citrus, grape, strawberry, pear, pineapple, lemon, grapefruit, peach, mango, nuts, seeds, tomato,beets, tangerine, apple	citrus, strawberry, pineapple, grape, grapefruit, lemon, mango, peach, pear, apple, ripe, nuts, seeds, cherry, cranberry
vegetable	potato, beans, vegetables, cheese, mustard, beef, boiled, chicken, tomato, milk, corn, pig, bread, grape, grains	vegetables, salad, greens, botany, plantain,weeds, crops, algae, herb, legumes, perennial, lettuce, sprouts, vegetation, sprout	vegetables, botany, greens, legumes, salad, herb, plantain, herbs, lettuce, celery, crops, pomegranate, algae, asparagus, carrot

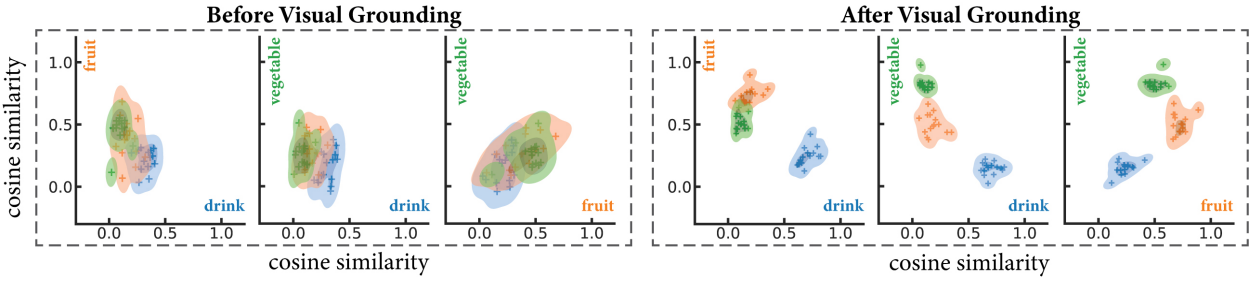


Figure 5.16: Visualization of words related to **food** subcategories.

Table 5.8: Top-15 words for **room** subcategories

query	**Bert**	**Grounded**	**Relational Grounded**
bathroom	restroom, bedroom, bath, kitchen, toilet, laundry, refrigerator, couch, bathtub, dresser, shower, hallway, mirror, towel, backyard	restroom, bath, shower, bathtub, toilet, sink, vanity, wash, tub, mirror, towel, soap, hygiene, hallway, shave	restroom, kitchen, cafeteria, bedroom, bath, room, shower, hospital, office, gym, classroom, pantry, hotel, gymnasium, motel
bedroom	bathroom, bed, room, dresser, downstairs, apartment, kitchen, condo, couch, upstairs, backyard, mattress, hallway, bath, attic	bed, mattress, pillow, room, closet, condominium, crib, dresser, apartment, motel, upstairs, cot, dorm, blanket, robe	room, closet, dorm, hotel, kitchen, apartment, motel, bathroom, household, dormitory, office, hostel, classroom, cafeteria, house
kitchen	refrigerator, bathroom, couch, fireplace, backyard, laundry, barn, furniture, sofa, basement, toilet, bedroom, cupboard, stairs, driveway	pantry, counter, cupboard, household, galley, stove, cook, microwave, oven, refrigerator, freezer, kettle, furnace, washer, chef	pantry, cafeteria, household, bathroom, bedroom, restroom, room, office, showroom, classroom, restaurant, garage, parlor, gym, dugout

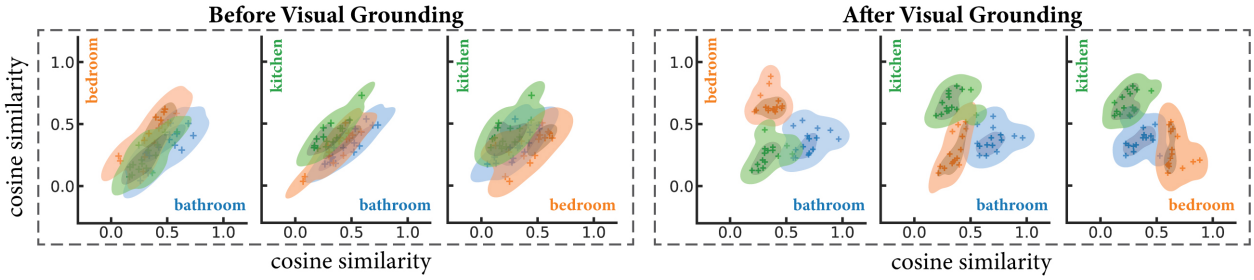


Figure 5.17: Visualization of words related to **room** subcategories.

5.3.4 Language composition based on visual knowledge

We also test whether the visually grounded semantics can perform compositional reasoning based on visual knowledge, without being explicitly trained to do so. For this purpose, we choose some words (Table 5.9) with meanings that can be intuitively inferred from the combination of other words (especially based on visual perceptual features). For example, we use a phrase "striped horse"

as a compositional query to search for the matched words ranked in terms of cosine similarity. With the grounded semantic representation, the phrase *striped horse* is highly similar to the word *zebra* (cosine similarity: 0.63), which is ranked as the 8-th among all common English words. Other words similar to *striped horse* all refer to animals within horse familiars (See visualization of the changes in ranking before and after visual grounding with a slope chart as shown in Fig. 5.18). In contrast, the ungrounded Bert model is not able to relate *striped horse* to *zebra* based on the similarity of their representations (cosine similarity: 0.12). See other examples in (Table 5.9 and Fig. 5.19).

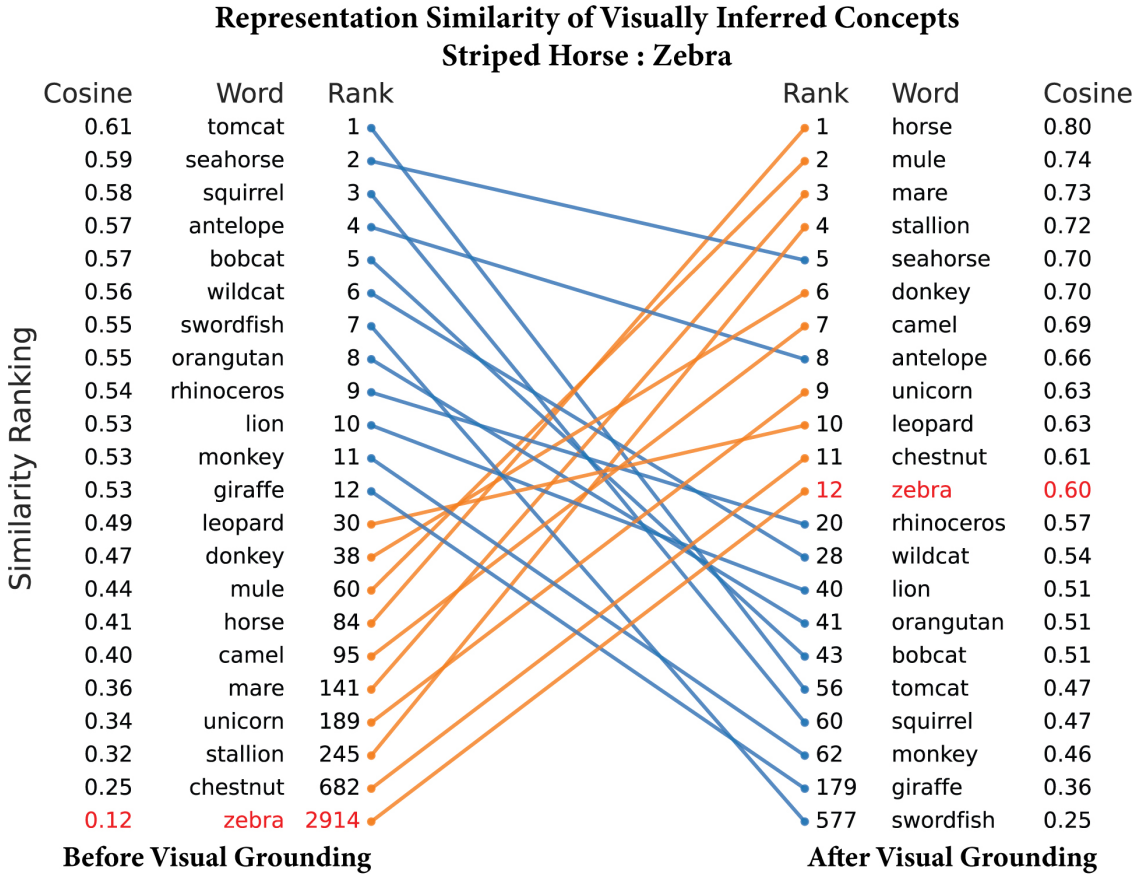


Figure 5.18: **Language composition based on visual knowledge (*striped horse*).** The left part shows the cosine similarity and ranking between the listed words and the query phrase (striped horse) before visual grounding. The right part shows the corresponding results after visual grounding of natural langauge. Orange curves indicate words with increased ranking after visual grounding (blue curves for the decreased cases). We highlight the target word "zebra" for this specific example, which shows a significant increase in cosine similarity value (from 0.12 to 0.60) and ranking (from 2914 to 12 out of 6238 unique words). Besides, the top words similar to "striped horse" are all horse-like animals after visual grounding, but this is not the case for ungrounded Bert model.

Table 5.9: Examples of vision based conceptual composition. Each row shows the cosine similarity and its ranking in the vocabulary (unique words in the Semcat dataset) between a pair of query phrase and target word. Except (`hot weather`, `summer`), all others are concepts supported by composition of *visual* knowledge in the query phrase.

Query Phrase	**Target Word**	**Similarity (cosine | rank)**					
		Bert		**Grounded**		**Relational**	
striped horse	zebra	0.12	2914	0.60	12	**0.63**	**8**
black and white bear	panda	0.13	2478	0.69	**2**	**0.81**	**2**
flying car	plane	0.36	167	**0.66**	**4**	0.61	11
round container	bowl	0.25	489	0.56	8	**0.67**	**2**
red fruit	strawberry	0.39	239	0.75	**3**	**0.85**	**3**
young dog	puppy	0.40	94	0.92	**2**	**0.93**	**2**
iced mountain	glacier	0.44	20	**0.86**	**1**	0.73	5
clear sky	sunny	0.27	631	0.31	184	**0.34**	**61**
hot weather	summer	0.27	903	0.52	14	**0.53**	**6**

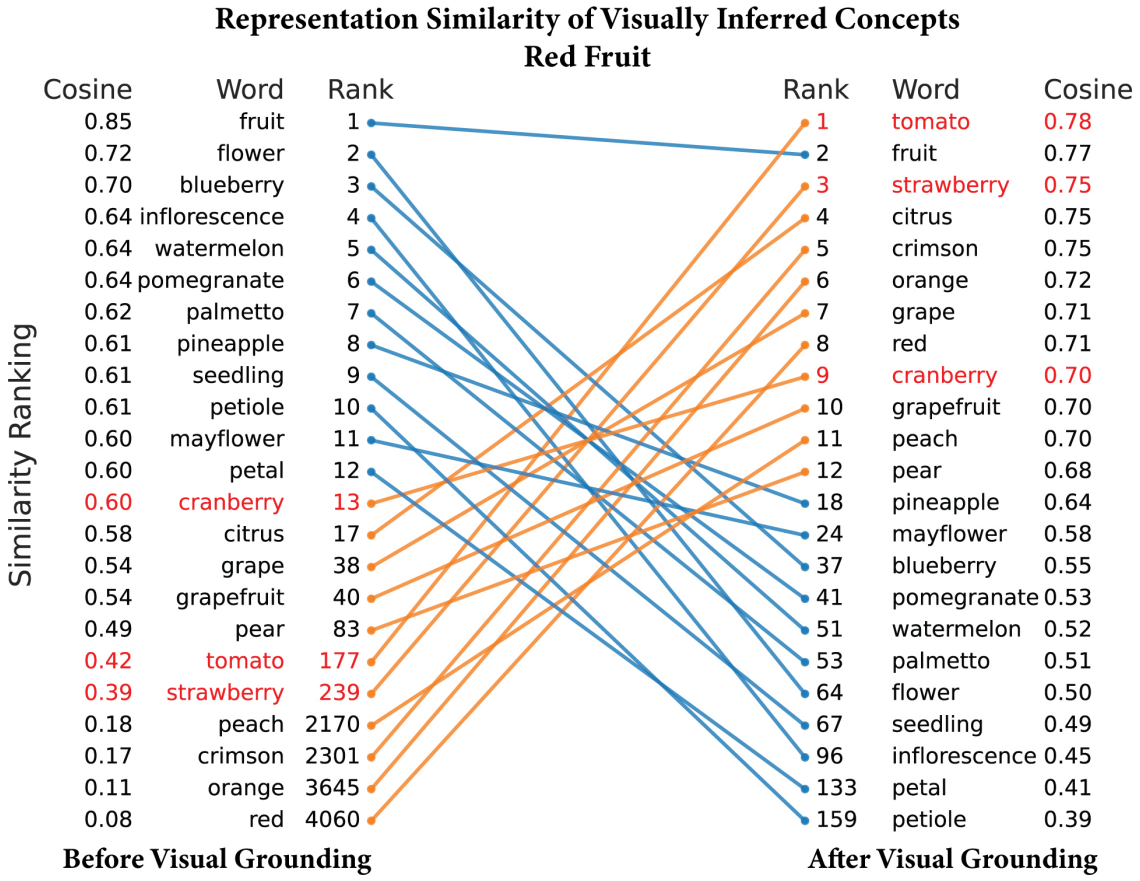


Figure 5.19: **Language composition based on visual knowledge (*red fruit*).**

5.3.5 A continuous semantic space shared across modalities

Figure 5.20: **Cross-modal image search.**. Top: Illustration of multimodal image search with a "zebra" example. The image query Q_I is the L2-normalized vector representation of a zebra's skin pattern. The word query Q_W is the L2-normalized vector representation of word *horse*. As the weight ratio α in the multimodal query Q_{search} increases from 0 to 1, the search results show grade change from zebra-like patterns, to a real zebra, and to a horse image. Bottom: Example multimodal image search results on image "cat" and word *sleep*, image "coffee" and word *milk*, image "car" and word *lego*. Increasing α from 0 to 1 gives the image search results from left to right.

As introduced in Section 5.2.6, we test the trained model for image search based on a multimodal image-text query, where the weighting between image and test information is controlled by a parameter α. For example, when we combine a word (*horse*) and an image (a stripped pattern) into a query, the multimodal search as described in Section 5.2.6 finds images similar to the zebra's skin pattern when α is close to 0, or finds images of typical horses when α is close to 1, but a zebra only when α is somewhere around 0.5 (Fig. 5.20, top).

This observation is generalizable to other examples (see Fig. 5.20 bottom). When the query image is a cat face and query word is *sleep*, the search results gradually change from an awake cat image, to a cat lying on a cushion, to a sleeping cat, and to sleeping (non-cat) animals when textual information continuously be added to the image query by increasing the weighting parameter α from 0 to 1. Similarly, when query image is a cup of coffee and query word is *milk*, the search results find coffee images when α is closed to 0 and find milk images when α is closed to 1, while images showing a cup of latte is found in between. In the last example, query image is a car and query word is *lego*, the multimodal search gradually changes its results from a real car to a toy car as α increased from 0 to 1. These results collectively suggest that the joint semantic space learned by grounding language in visual contexts is shared between visual and language domains and continuously captures conceptual representations.

5.4 Summary and Discussion

The current study demonstrates that after grounding language learning with visual experience, the word representations are reshaped in the semantic space to show interpretable semantic features more in line with human intuition and neurobiological knowledge. Specifically, the grounded semantic space can intrinsically capture the concrete-abstract axis in its first principal dimension. The visually grounded word embeddings can better predict human-defined binary features of concepts. Words are better clustered by categories and into fine-grained categories after visual grounding. The composition of concepts can be supported by visual knowledge (e.g., "zebra" = "striped horse"). Furthermore, the two-stream model can learn a semantic space to continuously represent concepts conveyed as a text or an image, or their combinations.

Methods of evaluating word embedding models can be summarized as two general schemes: *intrinsic* evaluation and *extrinsic* evaluation [171]. Intrinsic evaluations directly test the property of word vectors without performing external tasks. Extrinsic evaluations test the performance of word representations on downstream natural language tasks. The methods and results presented in this section focus on the *intrinsic* evaluation. We use data-driven analysis (e.g., PCA) with held-out datasets with human-defined or human-rated linguistic properties of concepts [153, 24, 34, 41]. The widely-used intrinsic evaluation methods include testing word similarity, word categorization, and

word analogy [124]. We focus on intrinsic evaluations in this research because we are interested in understanding and interpreting the organization of word distribution in the representational space before and after visual grounding. We conclude that the language model has learned a more explainable semantic space after learning with both language and visual input. Nevertheless, how the learned word embeddings can support downstream tasks is also useful and worth exploring in future studies.

It remains underexplored whether visual grounding should happen at an earlier stage or a later stage. Being earlier (or later) means more (or fewer) layers in the language model should be made learnable during cross-modal contrastive learning. Although early grounding seems to reshape the semantic space based on visual information with a more substantial effect, it may also cause potential harm to the textual processing capabilities of the language model. This assumption is supported by the following extrinsic evaluation results on some benchmark natural language understanding datasets.

Similar to the evaluation in [35], we test the language model before and after visual grounding on the General Language Understanding Evaluation (GLUE) benchmark [170]. For this purpose, the Bert encoder in the language models is fixed, while only a pooling layer (which is shared across all tasks in GLUE) and a linear classification layer (which is task-specific) were trainable. The testing results are submitted to and evaluated by the GLUE benchmark website [3].

Table 5.10: Model performance on GLUE benchmark.

Model	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-(m/mm)	QNLI	RTE	Average
Bert	40.7	92.6	87.8	81.8	71	83.3/82	89.4	73.8	78.04
Grounded-4	37.1	91.4	86	83.3	70.8	82.7/81.5	88.9	73	77.19
Grounded-8	38.6	91.5	86.3	83.4	70.8	82/81	87.9	72.2	77.08
Grounded-12	37.2	92.6	86.5	82.3	70.5	82.3/81.7	89.2	72	77.10
Relational-2	38	92.8	84.6	81.8	70.4	83.1/81.8	89.2	71.7	77.04

Table 5.10 summarizes the results. Grounded-k models have k learnable layers in Bert for visual grounding of natural language with MS COCO dataset. Relational-2 model is trained from Grounded-8 by finetuning 2 Bert layers for visual grounding of object relations. In general, the results suggest that the language model has a slightly decreased performance on GLUE after visual grounding, although the models tested here all share the same architecture as Bert. Allowing more

[3] https://gluebenchmark.com/

learnable parameters during visual grounding tends to result in worse performance for natural language understanding. This observation is unsurprising since we enforce the grounding process by matching short captions (MS COCO dataset) or phrases (Visual Genome dataset) to visual contents, which may not require extensive capacity for textual processing. This may be related to the notorious "catastrophic interference" problem in machine learning - the learning process of these new tasks in the language stream can interfere with the natural language understanding skills previously acquired from Bert pretraining. Future work on building grounded language learning models is needed to reconcile the trade-off between the performance for visual grounding vs. natural language understanding.

Another limitation in the current study is that we have not evaluated how the cross-modal training affects the visual stream. It is worth future exploration on whether the visual stream is able to learn more robust, abstract, and interpretable features by integrating high-level semantic information from the language model. Although the convolutional neural network (CNN) has reached great success in computer vision, several fundamental issues remain to be solved. Recent findings suggest that most CNNs are extremely vulnerable to even tiny perturbations of the input, e.g., adversarial examples [57]. Besides, all current ImageNet-trained CNNs are found to be severely biased towards local features, e.g., texture [48]. In addition, there is no well-established approach to easily scale up a trained model, for example, generalizing an image classification model to new labels [180]. One potential reason is that it is hard for CNNs to extract robust and abstract high-level features solely based on image inputs. Therefore, transferring the semantic knowledge from the language domain to visual models is a reasonable and likely promising solution to these problems [43, 91, 6, 29, 154]. Recent studies have used a similar strategy as in our research, i.e., training a two-stream model structure with contrastive learning on a large-scale noisy image-text dataset, showing the visual stream pretrained with matched semantic information can be transferred with a strong performance on multiple visual classification tasks [79].

Further, we have not thoroughly investigated the relational embeddings learned from the bilinear operator in the visual grounding stage with object relations. As the bilinear operator is carefully designed with some extra constraints, we expect the relational embedding matrices should reach some desired property, e.g., compositionality and transitivity (See details in Section 4.2.4). The preliminary results show interesting properties in some cases, but they are not significantly captured for all relational embeddings. For example, the embedding matrix of relation *wear* tends to be the transpose of the embedding matrix of relation *worn*, and the embedding matrix of relation *over* tends to be an idempotent matrix, supporting the intuition that $(A \textit{ over } B)$ `and` $(B \textit{ over } C) \implies (A \textit{ over } C)$. However, the current dataset does not contain enough well-defined relation labels, and it is still relatively imbalanced (top-10 occurred relation labels cover the majority of samples in the training and testing dataset). Thus, the relational embeddings might have not been fully trained to

form well-organized representations. And the training hyperparameters, including the number of heads in the cross-modal attention and the number of dimensions in the bilinear relational operator, remain to be carefully finetuned and optimized. To sum up, the bilinear operator, as initially inspired by prior studies on knowledge graph learning [182], shows promising results for modeling object relations with desired matrix computational properties that are worth in-depth explorations in future studies.

CHAPTER 6

Cortical Representations of Semantics

6.1 Rationale and Overview

[1] Inspired by biological neural networks, deep artificial neural networks have demonstrated near-human performance in some visual and language tasks [69]. Comparing artificial neural networks with biological brains has been increasingly used to investigate neural information processing in the brain [93]. This chapter summarizes how we applied the learned representations from language learning models to bridge artificial intelligence and neuroscience through neural encoding [127].

We first collected functional magnetic resonance imaging (fMRI) and electrocorticography (ECoG) data from humans under naturalistic stimulation [192], e.g., natural vision [175], natural language comprehension [193, 191], musical perception and musical imagery [189]. We trained an encoding model to relate the word representations learned from machine learning models to the cortical representation observed with fMRI or ECoG given the same set of stories delivered to computational models and human subjects. After establishing the encoding model, we used it as a high throughput computational analogy of the brain during semantic processing. We mapped word categories, word relations, and principal axes in the semantic space to their corresponding cortical representations. The results shed new light onto the cortical organization of not only concepts but also the relation between concepts, as well as the primary semantic features that define concepts for language and cognition.

For musical perception and imagery data, we time-locked and controlled the imagery process by using a visual cue [117] and correlated the brain responses when subjects were listening to and imagining the same music piece (Fig. 6.1). The corresponding results shed light on the shared cortical mechanism under high-level cognitive processing of information carried by different perceptual modalities.

[1] This chapter is based on the publications [189, 191] and a conference abstract [193].

6.2 Human Experiments

We collected three sets of data from our previous studies, including natural story comprehension with healthy subjects [191], natural story comprehension with epilepsy patients [193], and musical perception and visually-cued musical imagery with musicians [189]. All subjects provided informed written consent according to approved research protocols.

6.2.1 Natural story comprehension

For natural language comprehension, we collected both ECoG and fMRI data with epilepsy patents and healthy subjects, respectively [193, 191]. 6 epilepsy patients (3 females, age 38.0 ± 7.5) and 19 healthy human subjects (11 females, age 24.4 ± 4.8, all right-handed) participated in this study. Each subject was listening to several audio stories collected from The Moth Radio Hour [2]. ECoG was recorded at 2000Hz by using a grid of 64 electrodes on one hemisphere in epilepsy patients for whom the electrodes were implanted for surgical planning. T1 and T2-weighted MRI and fMRI data were acquired in a 3T MRI system (Siemens, Magnetom Prisma, Germany) with a 64-channel receive-only phased-array head/neck coil. The fMRI data were acquired with 2mm isotropic spatial resolution and 0.72s temporal resolution by using a gradient-recalled echo-planar imaging sequence (multiband$=8$, 72 interleaved axial slices, TR$=720$ms, TE$=31$ms, flip angle$=52°$, field of view$=21 \times 21$cm^2). During fMRI scanning, the stories were presented through binaural MR-compatible headphones (Silent Scan Audio Systems, Avotec, Stuart, FL). A single story was presented in each fMRI session (6 mins 48 secs $\pm$ 1 min 58 secs). For each story, two repeated sessions were performed for the same subject.

6.2.2 Musical imagery with visual cue

The fMRI data for musical perception and imagery with visual cue has been collected for this study [189]. Nine healthy volunteers (Age 19 to 27, 3 females, all right-handed with normal hearing and on average 10.9 years of musical training) participated in this study. The music stimulus was the first 8-minutes of the first movement of Beethoven's Symphony 9. This music piece was visualized as a movie through Stephen Malinowski's Music Animation Machine [117] (Fig. 6.1). In the musical perception sessions, each subject was instructed to listen to this music with his or her eyes closed while no movie was presented. During the musical imagery sessions, each subject was instructed to imagine the music piece while watching the silent music visualization.

Here, we focused on musicians because the task of imagining a long piece (8 min) of classic music was rather difficult for non-musicians. Otherwise, inclusion of data from any subjects who

[2] https://themoth.org/radio-hour

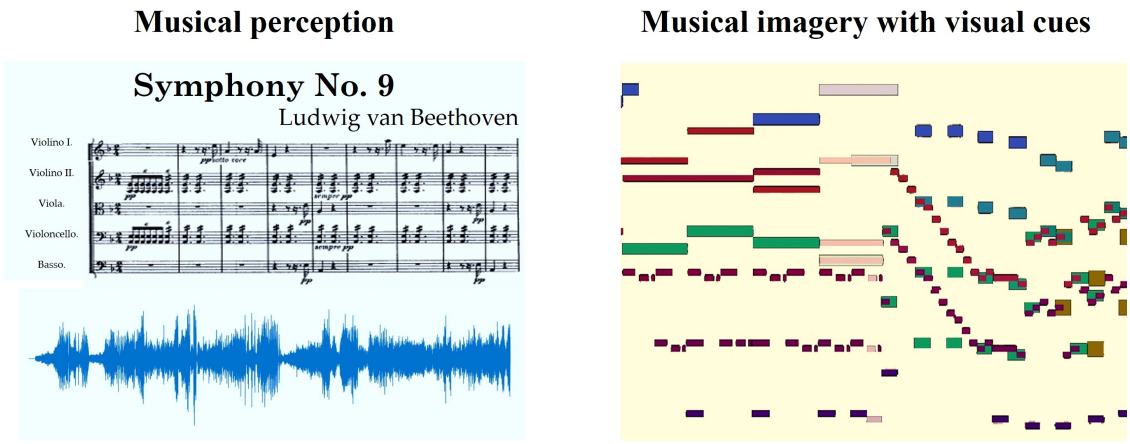


Figure 6.1: **Paradigm for musical perception (left) and imagery (right)**. The visualized music shown in is an animation with bars moving from right to left as the music flows. It includes all the musical information as in a standard music sheet: the length of the bars indicates the note length (rhythm and duration); the height of the bars indicates the keynote (pitch); the color of the bars indicates the instrument (timbre).

were unable to perform the musical-imagery task would complicate the study, of which the main goal was to map cortical activation and networks underlying musical imagery with visual cue.

6.2.3 Data preprocessing

We filtered ECoG data to extract high-gamma (70 to 150 Hz) activity from each channel and obtained the power envelope of the filtered signal, following the standard preprocessing method as described in prior studies [31, 19].

We preprocessed the MRI and fMRI data by using the minimal preprocessing pipeline established for the HCP (using software packages AFNI, FMRIB Software Library, and FreeSurfer pipeline). After preprocessing, the images from individual subjects were co-registered onto a common cortical surface template (see details in [53]). Then the fMRI data were spatially smoothed by using a gaussian surface smoothing kernel with a 2-mm standard deviation.

6.3 Computational Experiments

6.3.1 Correlational analysis

In our research, we use correlational analysis to evaluate 1) the reproducibility between repeated or related brain measurements; 2) the strength of linear relationship between stimulus features and brain responses. The former approach was used to map task-evoked cortical patterns within or

across subjects. We also used this strategy to map multimodal brain regions that were involved in a pair of related cognitive tasks, such as music perception and music imagery. The latter approach was used to assess the interpretability of human-defined or data-driven stimulus features for conveying information in the brain.

In our experiments, each stimulus was presented to the same subject twice in separate sessions. We calculated the voxel-wise Pearson correlation coefficient $r_i = \text{corr}(x_i^1(t), x_i^2(t))$ in the fMRI time series (or channel-wise correlation for ECoG data), where x_i^k is the cortical response at location i for $k-$th repeated session. The resulting cortical map could highlight brain locations with high correlation values. Since only the task-driven response would be consistent in the two repeated sessions, this map was expected to show the pattern of task-evoked cortical activation. Specifically, we used naturalistic stimuli due to its high reliability within and across subjects [70].

Besides, we also used a similar correlational analysis to evaluate the linear relation between a stimulus feature and the brain response. Suppose $f(t)$ is a 1D feature times series extracted from the stimulus. We first convolved $f(t)$ with the canonical hemodynamic response function (HRF) [109] and then correlated the resulting time series with brain responses at every location to create a cortical map showing the stimulus-response relation.

6.3.2 Training and testing the voxel-wise encoding model

The brain data collected during natural language comprehension provided a large set of diverse samples between stimulus and brain response. We applied the neural encoding method to test whether the word representation extracted from the machine learning models could predict the brain data through a linear and sparse transformation, as explored in previous studies [77, 127, 175]. Briefly, the machine learning models received the same input (i.e., the same natural stories) that the human subjects listened to during the fMRI scans (or ECoG recordings). The high dimensional representation extracted from the models (i.e., the word embeddings) was used as a set of independent regressors. For each location in the brain, its response was modeled as a linear combination of these regressors,

$$x_i = a_i + \boldsymbol{b}_i \boldsymbol{y} + \epsilon_i, \tag{6.1}$$

where x_i is the response at the i-th voxel, $\boldsymbol{y}$ is a column vector, which is the word feature extracted from the machine learning models. Each element in $\boldsymbol{y}$ corresponds to one axis (or feature) in the semantic space. $\boldsymbol{b}_i$ is a row vector of regression coefficients, a_i is the bias term, and ϵ_i is the error or noise.

We separated the collected data into a training dataset and a testing dataset. We used the (word, data) samples from the training stories to estimate the encoding model. As words occurred

sequentially in the audio story, each word was given a duration based on when it started and ended in the audio story. A story was represented by a time series of word embedding sampled every 0.1s. For each feature in the word embedding, its time-series signal was further convolved with a canonical hemodynamic response function (HRF) to account for the temporal delay and smoothing due to neurovascular coupling [109]. The HRF-convolved feature-wise representation was standardized and down-sampled to match the sampling rate of fMRI. It follows that the response of the i-th voxel at time t was expressed as

$$x_i(t) = a_i + \boldsymbol{b}_i \boldsymbol{y}(t) + \epsilon_i(t), \tag{6.2}$$

where $\boldsymbol{y}(t)$ is the HRF-convolved time series. An illustration of the voxel-wise encoding model trained with fMRI data and word2vec embedding features is shown in Fig. 6.2.

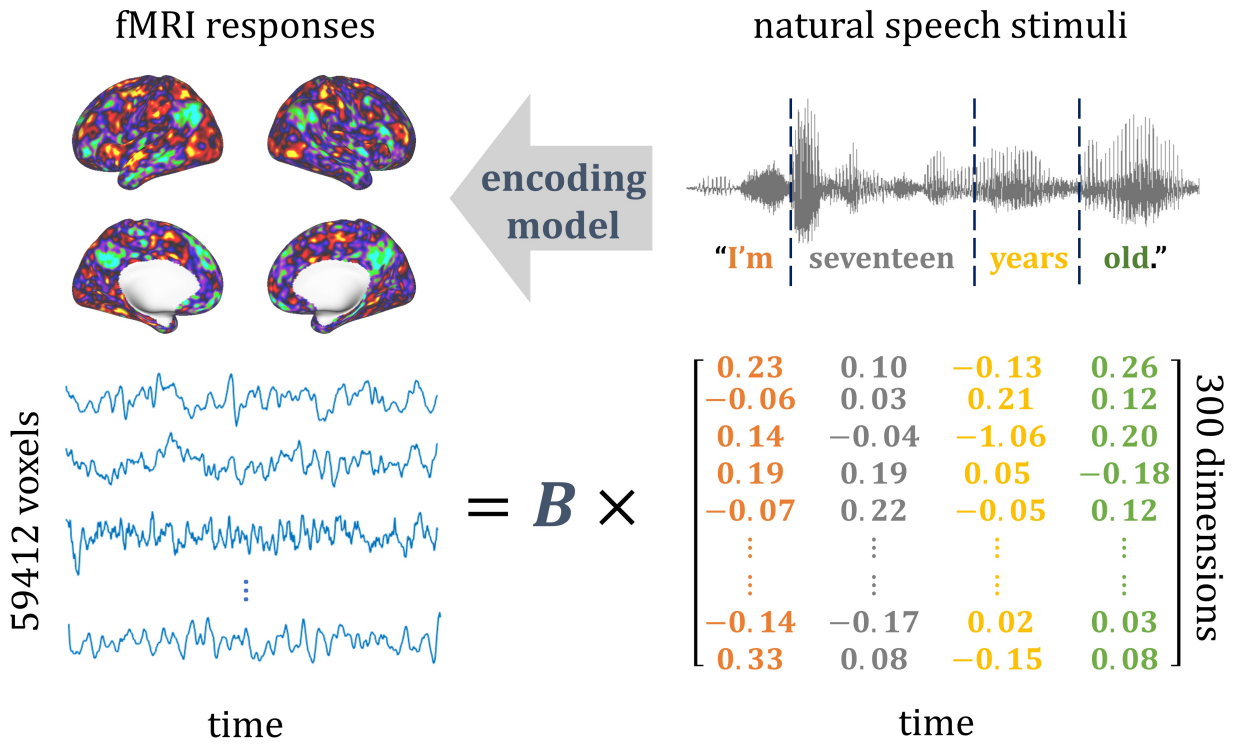


Figure 6.2: **The illustration of voxel-wise encoding model trained with fMRI responses and word2vec embedding features.** The encoding model was trained and tested for predicting the fMRI responses (top left) from a time series of words in audio-story stimuli (top right). Every word (as color coded) was converted to a 300-dimensional vector through word2vec. The encoding model was denoted as a $59,421$-by-300 matrix (B) to predict the voxel response to every word (bottom).

We estimated the coefficients $(a_i, \boldsymbol{b}_i)$ given time samples of $(x_i, \boldsymbol{y})$ by using least-squares estimation with L2-norm regularization. That is, to minimize the following loss function defined

separately for each voxel,

$$L_i = \frac{1}{T}\sum_{t=1}^{T}(x_i(t) - a_i - \boldsymbol{b}_i\boldsymbol{y}(t))^2 + \lambda_i\|\boldsymbol{b}_i\|_2^2, \tag{6.3}$$

where λ_i is the regularization parameter for the $i-$th cortical location, and T is the number of temporal samples.

We applied generalized cross-validation [56] to determine the regularization parameter λ_i. Specifically, the training data were divided evenly into ten subsets, of which nine were used for model estimation and one was used for model validation. The validation was repeated ten times such that each subset was used once for validation. In each time, the correlation between the predicted and measured brain responses was calculated and used to evaluate the validation accuracy. The averaged validation accuracy across all ten times was considered as the cross-validation accuracy. We chose the optimal regularization parameter that yields the highest cross-validation accuracy. Then we used the optimized regularization parameter and all training data for model estimation, ending up with the finalized model parameters denoted as $(\hat{a}_i, \hat{\boldsymbol{b}}_i)$ and the trained encoding model denoted as $\texttt{ENC}$.

We also tested how well the encoding model could be generalized to new data. For this purpose, the encoding model was applied to the testing dataset, generating a voxel-wise model prediction of the fMRI response to the testing story,

$$\hat{x}_i(t) = \hat{a}_i + \hat{\boldsymbol{b}}_i\boldsymbol{y}(t), \tag{6.4}$$

where $\boldsymbol{y}(t)$ is the HRF-convolved time series of word embedding extracted from the testing story. To evaluate the encoding performance, we calculated the correlation between the predicted brain response $\hat{x}_i$ and the actually measured brain response x_i. To evaluate the statistical significance, we used a block-wise permutation test [1] (20-s window size; $100,000$ permutations) with FDR $q < 0.05$.

In this research, we trained and tested two encoders: $\texttt{ENC}_{\text{word2vec}}$ was based on the existing word2vec model established by Google [123]. $\texttt{ENC}_{\text{grounded}}$ was based on our visually grounded language model established in Chapter 4, with its word embeddings evaluated in Chapter 5.

6.3.3 Mapping cortical representation through the encoding model

Through the encoding models, we synthesized cortical activation from a large datasets of words with high-throughput to map brain regions associated with different semantic properties, including categories of concepts, relationships between concepts, and semantic attributes of concepts.

6.3.3.1 Mapping word categories

In a prior study [191], we trained an encoding model with word2vec embedding and applied the encoding model to a large vocabulary set including about $40,000$ words [24]. We focused on the model prediction from nine categories: tool, human, plant, animal, place, communication, emotion, change, quantity (Table 6.1). For each word, we extracted its vector representation from the machine learning model, and then used the voxel-wise encoding model to map its cortical representation.

Semantic category	**Wordnet synsets**	**Semantic definition**	**# word samples**	**Example words**
tool	'tool.n.01'	an implement used in the practice of a vocation	200	axe, drill, fork, saw, wrench
animal	'animal.n.01'	a living organism characterized by voluntary movement	734	ant, cat, dog, fox, hen, owl
plant	'plant.n.02'	a living organism lacking the power of locomotion	387	aloe, beet, corn, lotus, rosewood
human	'adult.n.01'; 'worker.n.01'	a fully developed person; a person who works at a specific occupation	808	barber, clerk, groom, hunter, man
communication	'communication.n.02'	something that is communicated by or to or between people or groups	2,027	chat, discussion, gossip, idea, speaking
place	'location.n.01'; 'building.n.01'	a point or extent in space; a structure that has a roof and walls and stands more or less permanently in one place	814	arena, city, grave, inside, terminal
quantity	'definite_quantity.n.01'; 'measure.n.02'	a specific measure of amount; how much/many of something that you can quantify	958	billion, decade, single, gallon, megabyte
change	'change.v.01'; 'change.v.02'	cause to change, make difference; undergo a change	3,417	abort, fall, heal, reduce, thicken
emotion	'feeling.n.01'	the experiencing of affective and emotional states	504	anxiety, concern, daze, dream, mood

Table 6.1: Details of each semantic category.

As words were grouped by categories, we investigated the common cortical representation shared by those in the same category. For this purpose, we averaged the cortical representation of every word in each category, and threshold the average representation based on its statistical significance (one-sample t-test, FDR $q < 0.01$). We evaluated whether a given category was differentially represented by the left vs. right hemisphere. We also evaluated the semantic selectivity of each

voxel, i.e., how the voxel was more selective to one category than the others. For a coarse measure of categorical selectivity, we identified, separately for each voxel of significance, a single category that resulted in the strongest voxel response among all nine categories and associated that voxel with the identified category (or by "winners take all").

6.3.3.2 Mapping word relations

Semantic relation	**# paired samples**	**Example word pairs**
whole-part	178	hand-finger, zoo-animal, hour-second
class-inclusion	113	color-green, weapon-spear, tree-oak
object-attribute	63	fire-hot, child-innocent, heart-beat
case relations	106	coach-player, writer-story, barber-scissors
space-associated	58	library-book, mine-coal, mall-shopping
time-associated	44	morning-sunrise, winter-snow, christmas-presents
similar	160	house-home, kid-child, teach-instruct
contrast	162	hot-cold, rich-poor, top-bottom
object-nonattribute	69	fire-cold, slavery-freedom, optimist-despair
cause-effect	107	loss-grief, heat-sweat, study-learn

Table 6.2: Details of each semantic relation.

In word2vec embedding space, word relation is preserved as the differential vector of a word pair [123, 124]. Thus, applying the encoding model to the differential vector of a word pair could effectively generate the cortical representation of the corresponding word relation. With this notion, we used the encoding model to predict the cortical representations of semantic relations. Samples of word pairs were defined in the SemEval-2012 Task [83] dataset (Table 6.2). For each class of semantic relation, we calculated the relation vector of every word pair in that class, projected the relation vector onto the cortex using the encoding model, and averaged the projected patterns across word-pair samples in the class. For the averaged cortical projection, we tested the statistical significance for every voxel based on a paired permutation test. In this test, we flipped every word pair at random for $100,000$ trials. For every trial, we calculated the model-projected cortical pattern averaged across the randomly flipped word pairs, yielding a null distribution per voxel. Against this voxel-wised null distribution, we compared the average voxel value projected from non-flipped word pairs and calculated the two-sided p value with the significance level at FDR $q < 0.05$. The resulting pattern of significant voxels was expected to report the primary cortical representation of each semantic relation of interest.

6.3.3.3 Mapping principal axes in the grounded semantic space

In Chapter 5, we have found that the principal axes obtained by the singular value decomposition on visually grounded word representations manifest a set of explainable semantic attributes in line with human intuition. We further aim to understand how these principal axes of the visually grounded semantic space are represented by the human cortex. For this purpose, we trained a new voxel-wise encoding model $\texttt{ENC}_{\text{grounded}}$ for the visually grounded language model established in this study. We extracted the word embeddings from the `Grounded-4` model, which had the top 4 layers learnable in Bert and was trained with cross-modal contrastive learning on the MS COCO dataset for visual grounding of natural language, as illustrated in Fig. 4.1 and Section 4.2.5.2.

We decomposed this grounded semantic space into a set of orthogonal principal components (See detailed methods in Section 5.2.1). Each principal component defines an axis in the semantic space. The varying value on this axis quantified the representation of a specific semantic feature, such as **abstract** vs. **concrete** in PC 1, **non-human** vs. **human** in PC 2, and **object** vs. **scene** in PC 3. We then mapped each principle component, which also possessing a vector form in the semantic space, through the trained encoding model. We scaled each principal component with the corresponding singular value. Through the encoding model, we projected the scaled principal component in its positive and negative directions onto their corresponding cortical patterns. Their difference was interpreted as the cortical representation of the given principal dimension of the grounded semantic space,

$$\text{Pattern}_{\text{i}} = \texttt{ENC}_{\text{grounded}}(\boldsymbol{\Sigma}_{i,i}\boldsymbol{W}_{\cdot,i}) - \texttt{ENC}_{\text{grounded}}(-\boldsymbol{\Sigma}_{i,i}\boldsymbol{W}_{\cdot,i}), \tag{6.5}$$

where $\boldsymbol{W}_{\cdot,i}$ is the i-th column of $\boldsymbol{W}$ in the Eq. 5.1, which refers to the i-th principal component in the word embedding space. $\text{Pattern}_{\text{i}}$ is the resulting brain pattern of the i-th principal component mapped through the trained encoding model $\texttt{ENC}_{\text{grounded}}$. The cortical patterns of the first three principal components was expected to collectively revealed the cortical organization of the corresponding principal semantic components.

6.4 Results

All subsequent sections were based on $\texttt{ENC}_{\text{word2vec}}$ from our prior study using word2vec word embeddings [193, 191], except the results in Section 6.4.4, which was based on the encoding model $\texttt{ENC}_{\text{grounded}}$ trained from the visually grounded word embeddings.

6.4.1 Prediction accuracy for encoding models

We trained and examined the encoding model for natural language comprehension with both fMRI data and ECoG data. fMRI signal has a relatively higher spatial resolution but changes slow in time. In contrast, ECoG signal captures fast temporal dynamics of neural activity, but has limited spatial coverage. Combining results from these two types of brain signals would leverage their complementary advantages.

6.4.1.1 Encoding model trained with fMRI data for natural story comprehension

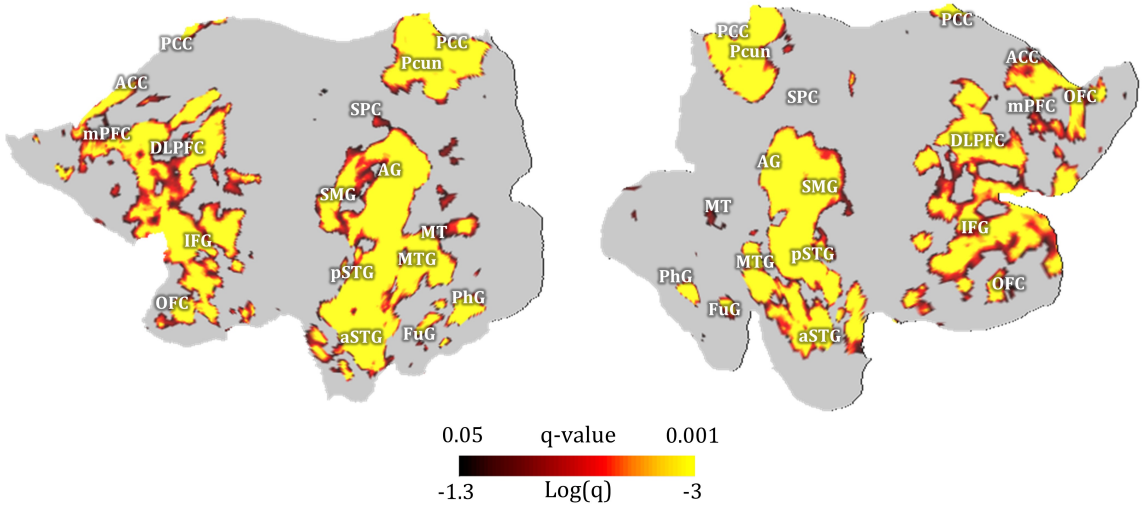


Figure 6.3: **A map of the semantic system obtained by 10-fold cross-validation of the encoding model.** The map is displayed on the flattened cortical surfaces for the left and right hemispheres. The color indiciates the FDR (or q value) in a logarithmic scale.

To estimate the voxel-wise encoding model, we acquired whole-brain fMRI data from 19 native English speakers listening to different audio stories. We used different stories for different subjects to sample more words collectively. In total, the story stimuli combined across subjects lasted 11h and included $47,356$ words (or $5,228$ words if duplicates were excluded).

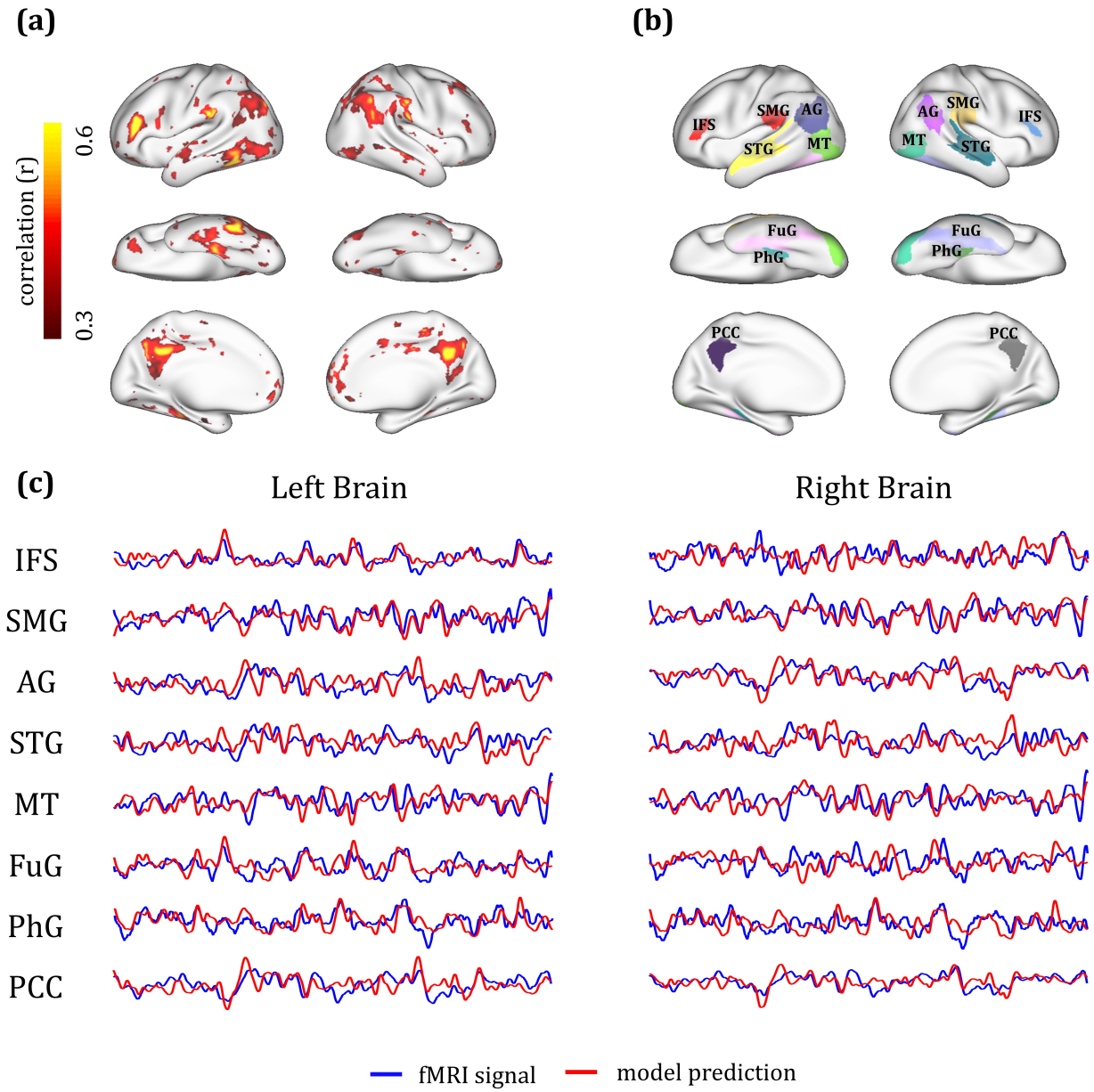


Figure 6.4: **Measured vs. model-predicted fMRI responses to a new (untrained) testing story.** (a). The voxel-wise correlation between fMRI responses and model predictions for the testing story. (b) Predefined ROIs (shown in different colors) are displayed on the cortical surfaces. (c) Response time series as measured (blue) or model-predicted (red) for each ROI, by averaging the time series across voxels within each ROI.

The voxel-wise encoding model was estimated based on the fMRI data concatenated across all stories and subjects. By 10-fold cross-validation[56], the model-predicted response was significantly correlated with the measured fMRI response (block-wise permutation test, false discovery rate or FDR $q < 0.05$) for voxels broadly distributed on the cortex (Fig. 6.3). This map, hereafter referred to as the semantic system, was widespread across regions from both hemispheres, as opposed to only the left hemisphere, which has conventionally been thought to dominate language processing and comprehension [92].

We also tested how well the trained encoding model could be generalized to a new story never used for model training and whether it could be used to account for the differential responses at individual regions. For this purpose, we acquired the voxel response to an independent testing story (6m 53s, 368 unique words) for every subject and averaged the response across subjects.

Color in Fig. 6.4(a) indicates the correlation coefficient between measured and model-predicted fMRI responses. The color-highlighted areas include the voxels of statistical significance (block-wise permutation test, one-sided, FDR $q < 0.05$). As shown in Fig. 6.4(a), we found that the encoding model was able to reliably predict the evoked responses in the inferior frontal sulcus (IFS), supramarginal gyrus (SMG), angular gyrus (AG), superior temporal gyrus (STG), middle temporal visual area (MT), left fusiform gyrus (FuG), left parahippocampal gyrus (PhG), and posterior cingulate cortex (PCC). These regions of interest (ROIs), as predefined in the human brainnetome atlas [40] (Fig. 6.4(b)), showed different response dynamics given the same story, suggesting their highly distinctive roles in semantic processing (Fig. 6.4(c)). Despite such differences across regions, the encoding model was found to successfully predict the response time series averaged within every ROI except the right FuG, suggesting its ability to explain the differential semantic coding (i.e., stimulus–response relationship) at different regions.

6.4.1.2 Encoding model trained with ECoG data for natural story comprehension

To estimate the encoding model with ECoG data, we used ECoG data from one epilepsy patient for whom the electrodes were implanted on left temporal and frontal cortices for surgical planning. During ECoG recording, the patient listened to 2h 19min of different audio stories.

The channel-wise encoding model was trained and tested by using leave-one-out cross-validation, while the encoding performance was evaluated as the correlation between the predicted and measured high-gamma activity. Fig. 6.5 shows the prediction accuracy of all channels in the ECoG data averaged across sessions. This result suggests the high-gamma power of ECoG in STG could be predicted by vectorized word representation during natural speech processing. To a lesser degree, somatosensory cortex and subtemporal area were also significantly predictable. The predicted high-gamma activity was a linear combination of the 300 elements in word2vec embedding space, showing similar temporal variation as the real high gamma activity (Fig. 6.6).

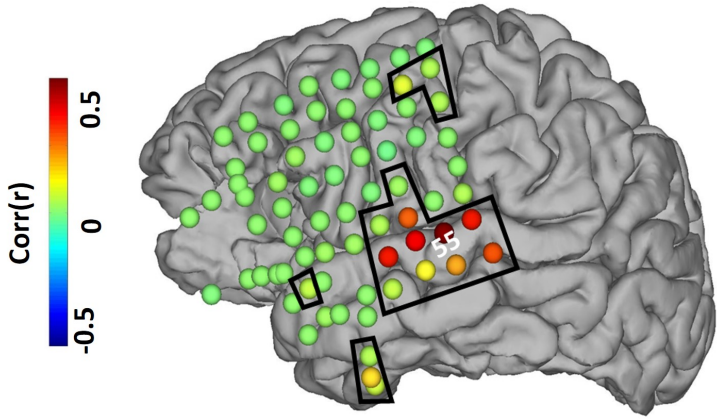


Figure 6.5: **Prediction accuracy of all channels in the ECoG data averaged across sessions.** The prediction accuracy was measured by channel-wise correlation between real and predicted high-gamma activity using leave-one-out cross-validation. 18 channels (inside black borders) showed significant positive correlation (FDR-corrected q-value< 0.01).

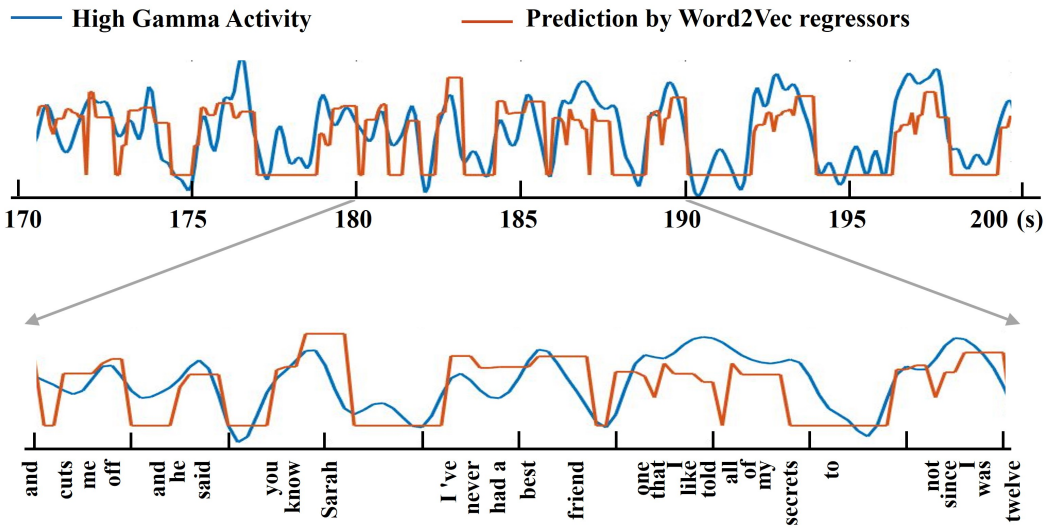


Figure 6.6: **Measured vs. model-predicted high gamma activity.** Examples of measured (blue) and predicted (red) high-gamma activity from one session in channel 55 (located in the STG; correlation coefficient $r = 0.51$. See Fig. 6.5 for reference.). Words aligned with high-gamma activity were demonstrated at the bottom.

6.4.2 Semantic categorization

Since the encoding model was generalizable to new words and sentences, we further used it to predict cortical responses to $> 9,000$ words from nine categories: tool, human, plant, animal, place, communication, emotion, change, quantity (Table 6.1), as defined in WordNet[125] and are representative of different conceptual domains. We confined the model prediction to the voxels in the semantic system for which the model fit was significant during cross-validation (Fig. 6.3).

Within each category, we averaged the model-predicted responses given every word and mapped the statistically significant voxels (Fig. 6.7; one sample t-test, FDR $q < 0.01$). This map represented each category being projected from the semantic space to the cortex, and thus was interpreted as the model-predicted cortical representation of each category. To each voxel in the semantic system, we assigned a single category that gave rise to the strongest voxel response, thus dividing the semantic system into category-labeled parcels (Fig. 6.8(a)). The resulting parcellation revealed how every category of interest was represented by a different set of regions, as opposed to any single region. In addition, the distinction in left/right lateralization was noticeable and likely attributable to the varying degree of concreteness for the words from individual categories. The concepts lateralized to the left hemisphere appeared relatively more concrete or exteroceptive, whereas those lateralized to the right hemisphere were more abstract or interoceptive (Fig. 6.8(b)). This intuitive interpretation was supported by human rating of concreteness (from 1 to 5) for every word in each category [24]. The maximum value of the concreteness rating is 5.00 and the minima value is 1.25. In the box plot Fig. 6.8(c), the central mark indicates the median, and the box edges indicate the 25th and 75th percentiles respectively. The concreteness rating was high (between 4 and 5) for the categories lateralized to the left hemisphere, whereas it tended to be lower yet more variable for those categories dominated by the right hemisphere (Fig. 6.8(c)).

6.4.3 Cortical representations of semantic relations

Through the word2vec model, we could also represent semantic relations as vectors in the semantic space [124]. Specifically, we represented the relationship between any pair of words based on their difference vector in word embedding.

For a given word-pair, their relation vector could be further projected onto the cortex through the encoding model. For an initial exploration, we applied this analysis to 178 word-pairs that all shared a same whole-part relationship. For example, in four word-pairs, (`hand`, `finger`), (`zoo`, `animal`), (`hour`, `second`), and (`bouquet`, `flower`), `finger` is part of `hand`; `animal` is part of `zoo`; `second` is part of `hour`; `flower` is part of `bouquet`. Individually, the words from different pairs had different meanings and belonged to different semantic categories, as `finger`, `animal`, `second`, and `flower` were semantically irrelevant to one another. Nevertheless, their

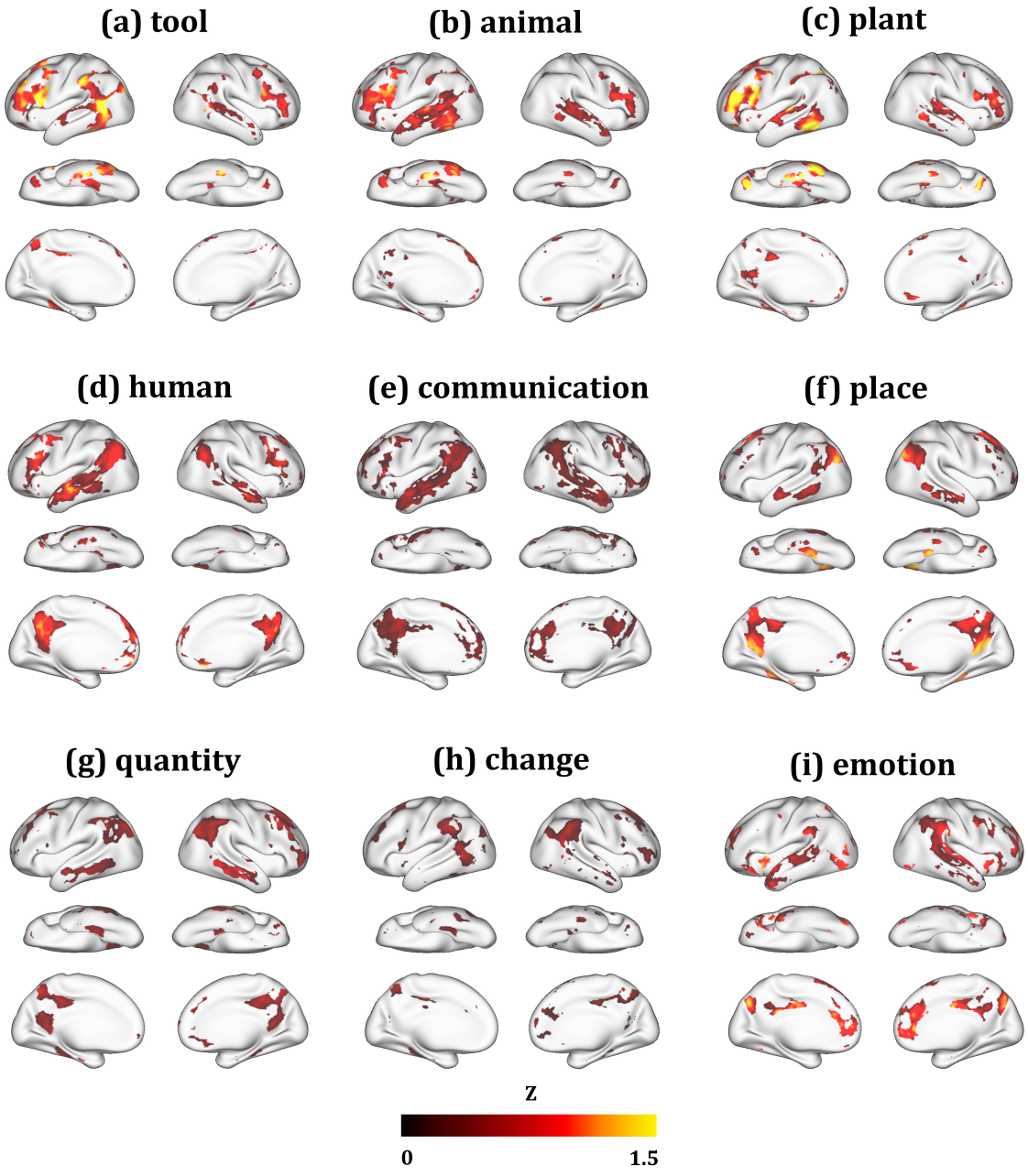


Figure 6.7: **Cortical representation of 9 word categories.**

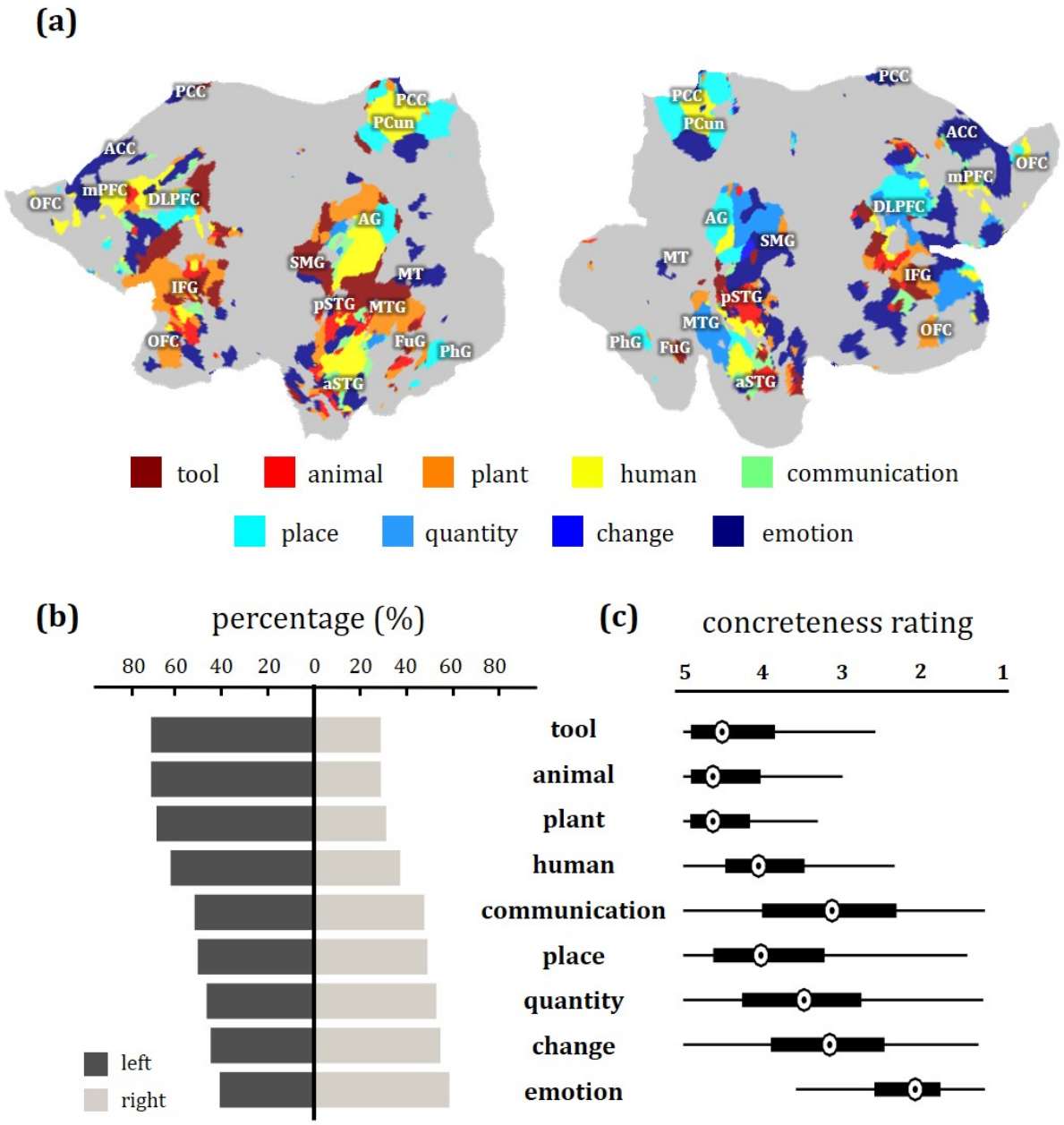


Figure 6.8: **Cortical organization of semantic category.** (a) Category-labeled parcellation based on voxel-wise selectivity using a "winners-take-all" strategy. (b) Cortical lateralization of categorical representations. For each category, the percentage value was calculated by counting the number of voxels on each hemisphere that represented the given category. (c) The concreteness rating of words in each category.

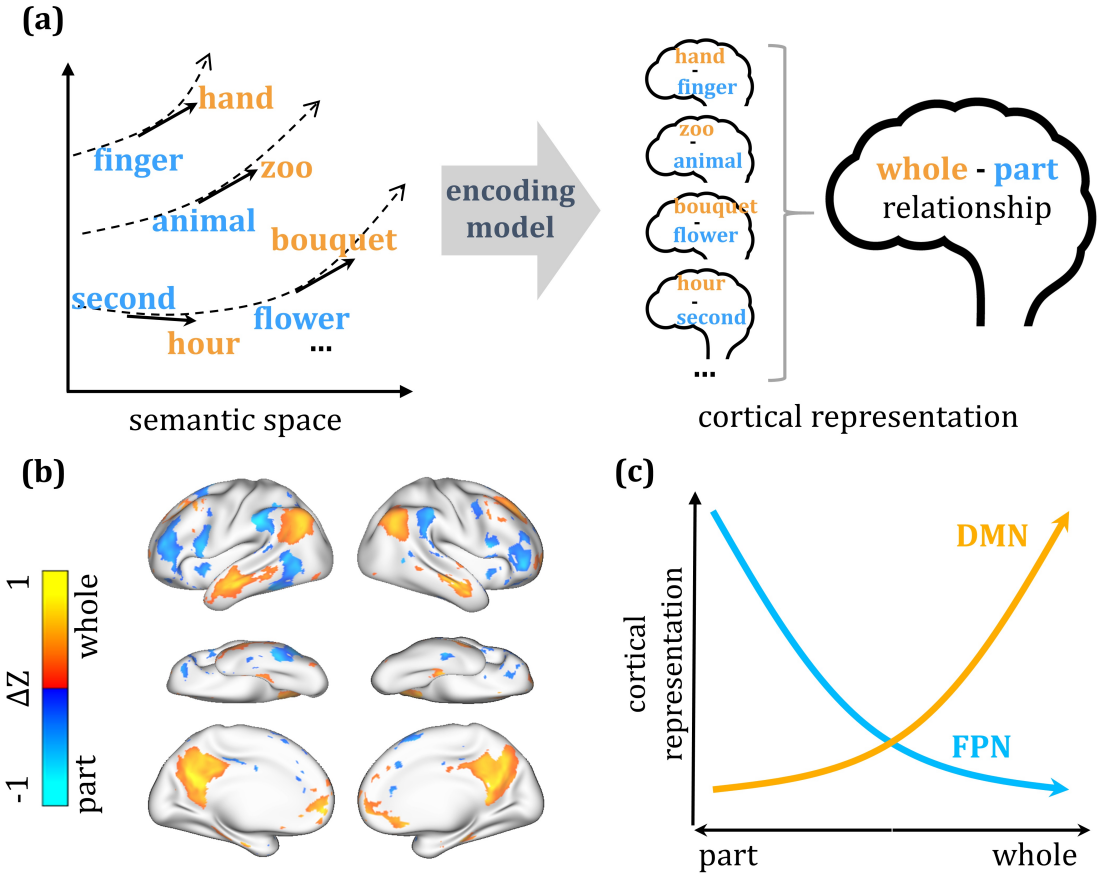


Figure 6.9: **Mapping cortical representation of the whole-part relation.** (a) The illustration of mapping the whole-part relation from the semantic space to the human brain through the voxel-wise encoding model. We viewed the whole-part relation as a vector field over the semantic space. This relation field was sampled by the difference vector of each word pair that held such a relation (left). The cortical representation of this difference vector was predicted by the voxel-wise encoding model. Cortical representation of the whole-part relation was then obtained by averaging representations of all word pairs (right). (b) Cortical representation of the whole-part relation. The statistical significance was assessed by a paired permutation test (178 word pairs, two-sided, FDR q<0.05). (c) The co-occurring activation of DMN and deactivation of FPN encodes the whole-part relation or the conceptual progression from part to whole.

pairwise relations all entailed the whole-part relation, as illustrated in Fig. 6.9(a). By using the encoding model, we mapped the pairwise word relationship onto voxels in the semantic system (as shown in Fig. 6.3), averaged the results across pairs, and highlighted the significant voxels (paired permutation test, FDR $q < 0.05$). The resulting cortical map represented each semantic relation being projected from the semantic space to the cortex, reporting the model-predicted cortical representation of the relation. We found that the whole-part relation was represented by a cortical pattern that manifested itself as the co-occurring activation of the default mode networks [140] (DMN, including AG, MTG, and PCC) and deactivation of the frontoparietal network [30, 151] (FPN, including LPFC, IPC and pMTG) (Fig. 6.9(b)). This cortical pattern encoded the whole-part relation independent of the cortical representations of the individual words in this relation. The co-activation and deactivation pattern indicated that conceptual progression from part to whole manifested itself as increasing deactivation of FPN alongside increasing activation of DMN, whereas progression from whole to part was shown as the reverse cortical pattern varying in the opposite direction, as illustrated in Fig. 6.9(c).

Similarly, we also mapped the cortical representations of several other semantic relations. Each relation was projected to a distinct cortical pattern (Fig. 6.10). For each of the ten relations in this figure, the number of significant voxels (paired permutation test, two-sided, FDR $q < 0.05$) is 10,607 (whole-part), 10,768 (class-inclusion), 9,037 (object-attribute), 9,550 (case relations), 11,496 (space-associated), 6,129 (time-associated), 0 (similar), 0 (contrast), 1,124 (object-nonattribute), 244 (cause-effect). Specifically, the class-inclusion relation, e.g., (`color`, `green`) where color includes green, was represented by the activation of AG and MTG and the deactivation of IFG and STG (Fig. 6.10(b)). The object-attribute relation, e.g., (`fire`, `hot`) where fire is hot, was represented by an asymmetric cortical pattern including activation primarily in the left hemisphere and deactivation primarily in the right hemisphere (Fig. 6.10(c)). The case relations, e.g., (`coach`, `player`) where a coach teaches a player, was represented by a cortical pattern similar to that of the whole-part relation (Fig. 6.10(d)), despite a lack of intuitive connection between the two relations. The space-associated relation, e.g., (`library`, `book`) where book is an associated item in a library, was represented by activation of AG and PCC and deactivation of STG (Fig. 6.10(e)). Lastly, the time-associated relation, e.g., (`morning`, `sunrise`) where sunrise is a phenomenon associated with morning, was also represented by a bilaterally asymmetric pattern (Fig. 6.10(f)). However, several nominal (human-defined) relations, e.g., similar, contrast, object-nonattribute and cause-effect, were projected onto either no or fewer voxels.

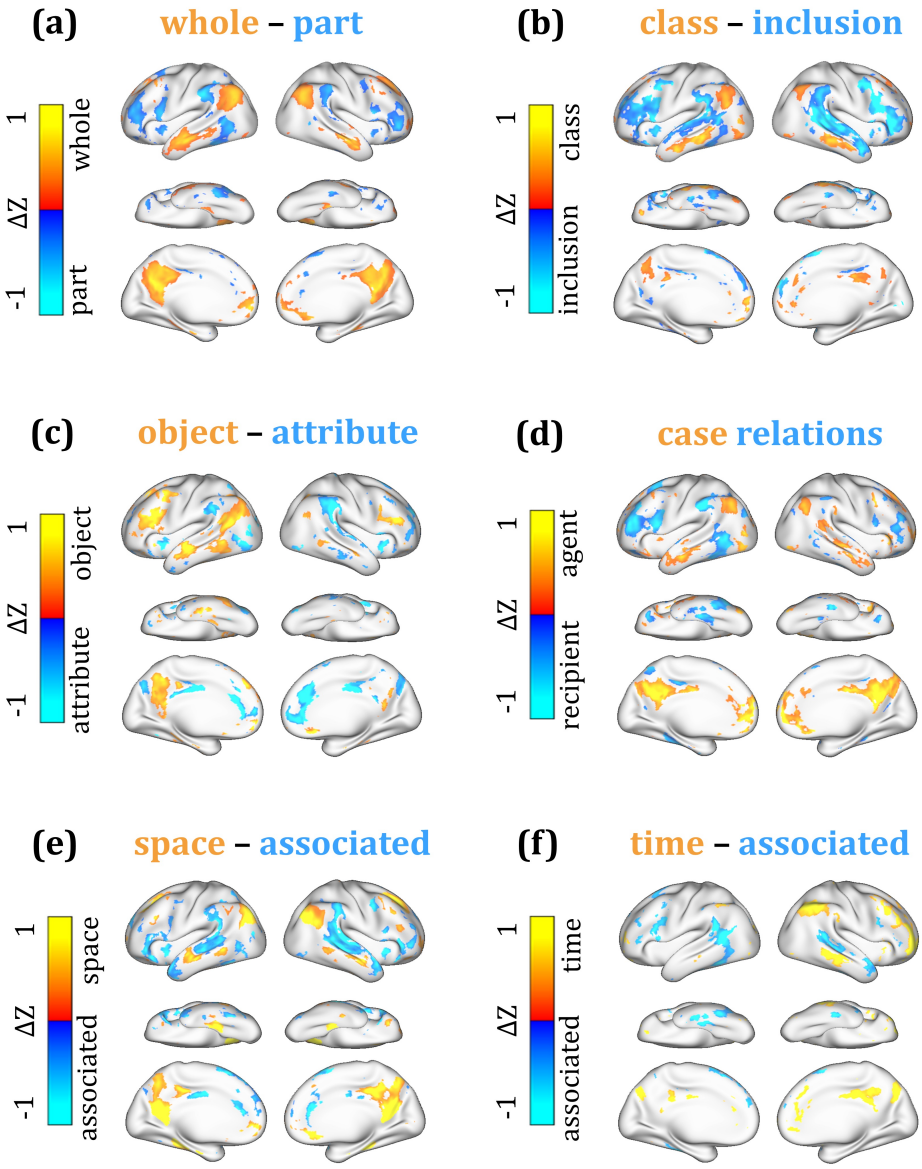


Figure 6.10: **Cortical representations of semantic relations.** The cortical pattern associated with each relation shows the average cortical projection of every word-pair sample in that relation and highlights only the voxels of statistical significance (paired permutation test, two-sided, FDR $q < 0.05$) based on voxel-wise univariate analysis.

6.4.4 Cortical organization of principal axes in the grounded semantic space

Results described in this section were based on the encoder $\mathrm{ENC}_{\mathrm{grounded}}$ trained with the visually grounded word embeddings. Since the visually grounded semantic space captured explainable principal axes (See details in Section 5.3.1), we mapped those principal axes onto the cortex through the encoding model with the method described in Section 6.3.3.3. The goal was to investigate the cortical organization of principal semantic attributes learned jointly from language and vision.

The results for all three principal components are shown in Fig. 6.11, Fig. 6.12, and Fig. 6.13. To visualize each map, we divided its positive and negative portions by the corresponding absolute maximum, resulting in a normalized cortical pattern with intensities ranging from -1 to 1. The collective findings from all three principal components were also summarized in a single RGB color-coded map to illustrate how the three principal semantic attributes were encoded by distributed networks in the human semantic system, as shown in Fig. 6.14.

Specifically, as shown in Fig. 6.11, the positive direction in PC1 reflected the **concrete** features. The word categories that show a pronounced representation in the positive direction of PC1 *cooking tools, mammal, ocean, fruit, reptiles, vegetables, desserts, animal* etc.). The cortical representation of this concrete semantics was almost exclusively lateralized to the left hemisphere. The most relevant areas included the left inferior frontal sulcus (IFS), left Brodmann area 44 (A44), left intraparietal sulcus (IPS), and left lateral temporal cortex (LTC).

To the opposite, the negative direction in PC1 reflected the **abstract** feature. The most pronounced examples were such word categories as *positive words, happiness, happy, adjectives for people, big, emotions, virtues, math*. The cortical representation of the abstract semantics involved both hemispheres. Despite an overall symmetry, the right hemisphere appeared to be more relevant to the abstract semantics than the left hemisphere. The most pronounced representations were found in the dorsomeidal parieto-occipital sulcus (dmPOS), and to a lesser degree in the superior temporal sulcus (STS) and right inferior parietal cortex (IPC). The differential degree of lateralization for the concrete and abstract attributes echoed the previous findings in word category representations (Section 6.4.2).

Fig. 6.12 suggests that the **non-human** feature (i.e., the positive direction in PC2: *fruit, flowers, reptiles, insect* etc.) was encoded by the left inferior frontal sulcus (IFS), left Brodmann area 44 (A44), left Brodmann area 45 (A45), and bilateral superior temporal sulcus (STS). The **human** feature (i.e., the negative direction in PC2: *rooms, boat, jobs, cooking tools* etc.) was slightly right-lateralized, mostly encoded by inferior parietal cortex (IPC) and posterior cingulate cortex (PCC).

As shown in Fig. 6.13, the representation of the **scene** feature (i.e., the positive direction in PC3: *water, biomes, land forms, bathroom* etc.) did not highlight a specific large region. Besides bilateral dorsomeidal parietooccipital sulcus (dmPOS), it was also coded by area PGs and area

PFm in the inferior parietal cortex (IPC) [52], right dorsolateral prefrontal cortex (dlPFC), and right anterior cingular cortex (ACC). In contrast, the representation of **object** feature (i.e., the negative direction in PC3: *jobs, mammal, musical instruments, birds* etc.) was coded mainly by the bilateral frontotemporal network.

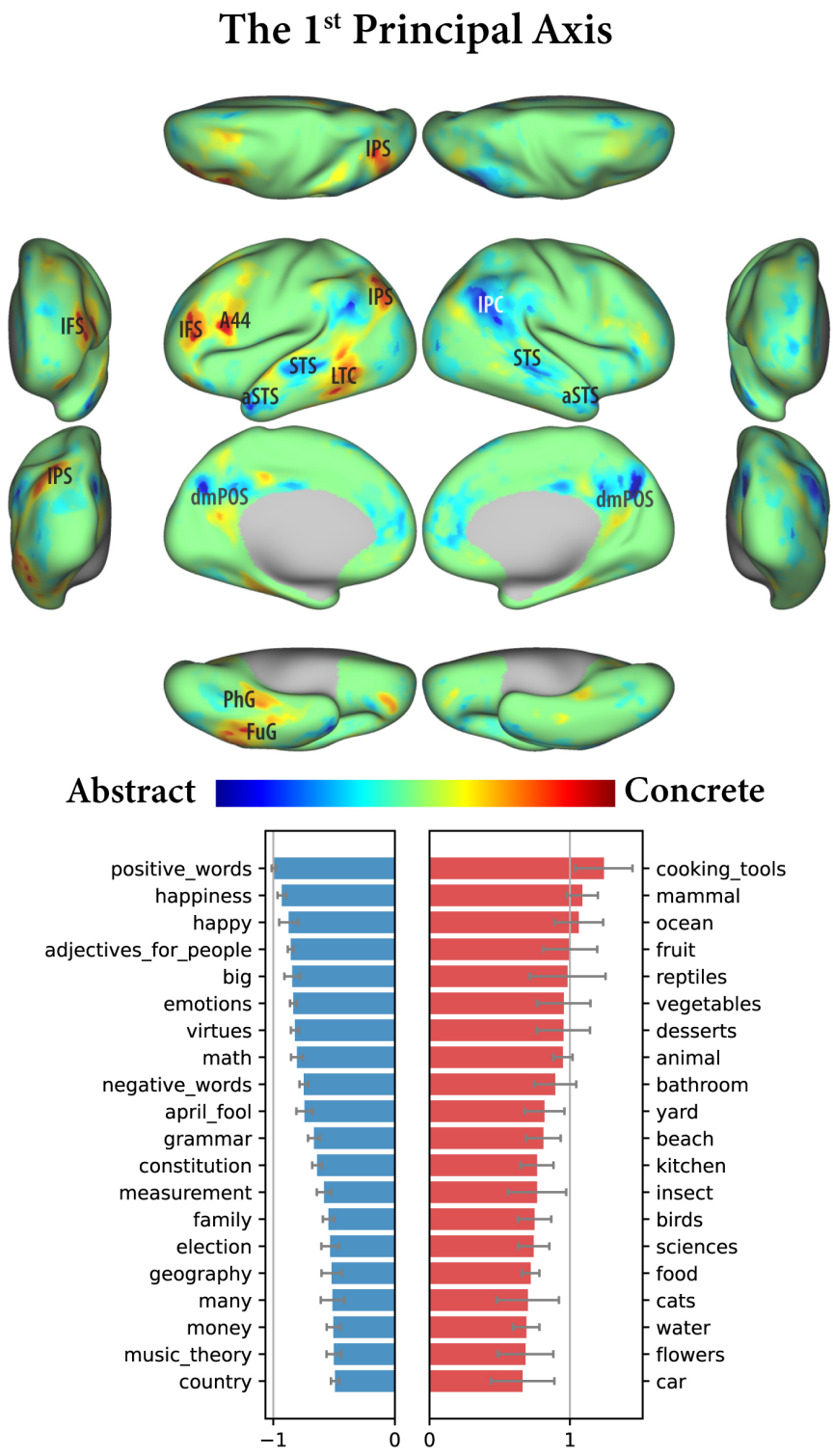


Figure 6.11: The cortical mapping and representative word categories for the first principal axis. The colorbar ranges from -1 to 1 after normalizing the resulting map. Blue: **Abstract**. Red: **Concrete**.

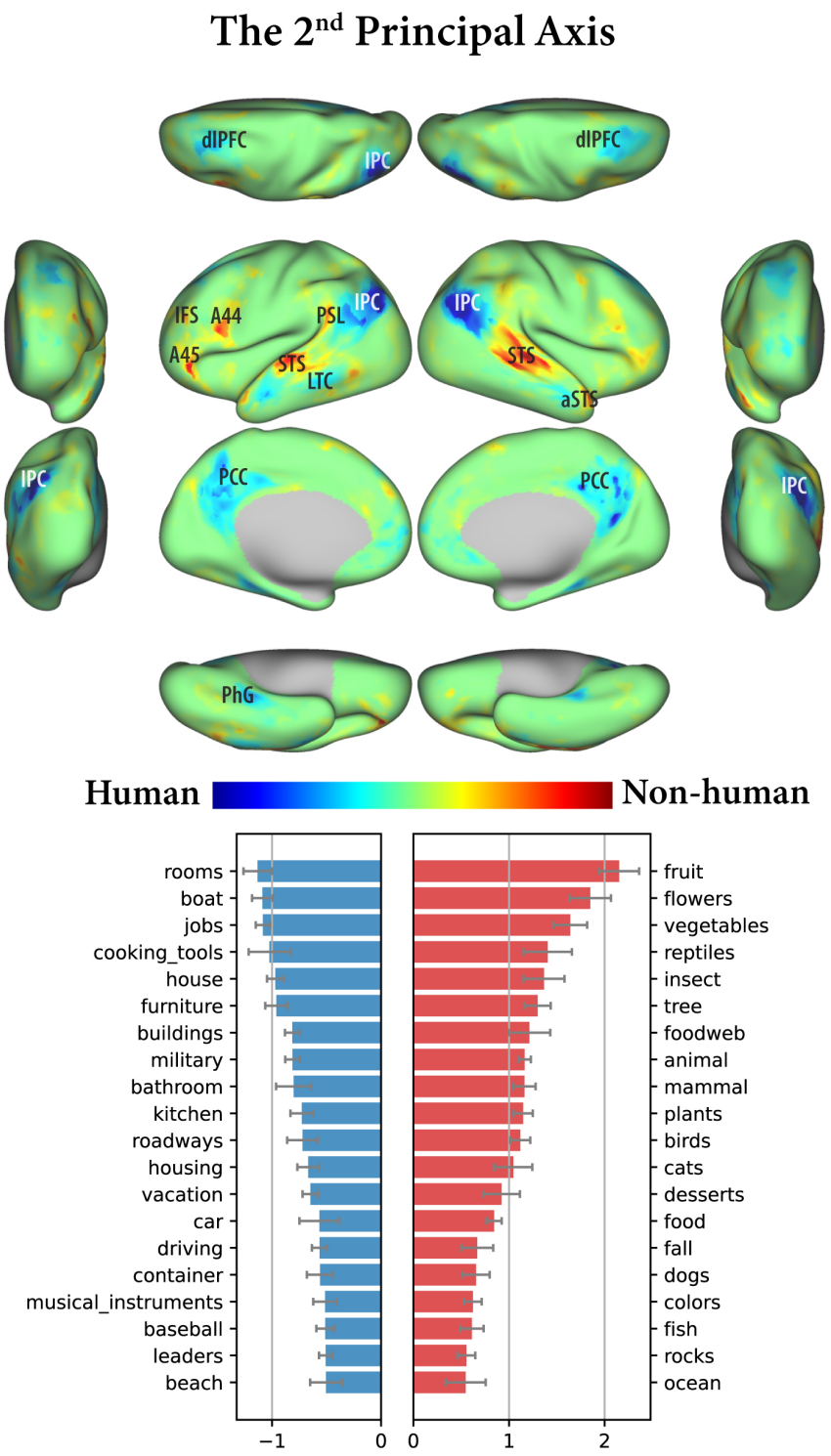


Figure 6.12: The cortical mapping and representative word categories for the second principal axis. The colorbar ranges from -1 to 1 after normalizing the resulting map. Blue: **Human**. Red: **Non-human**.

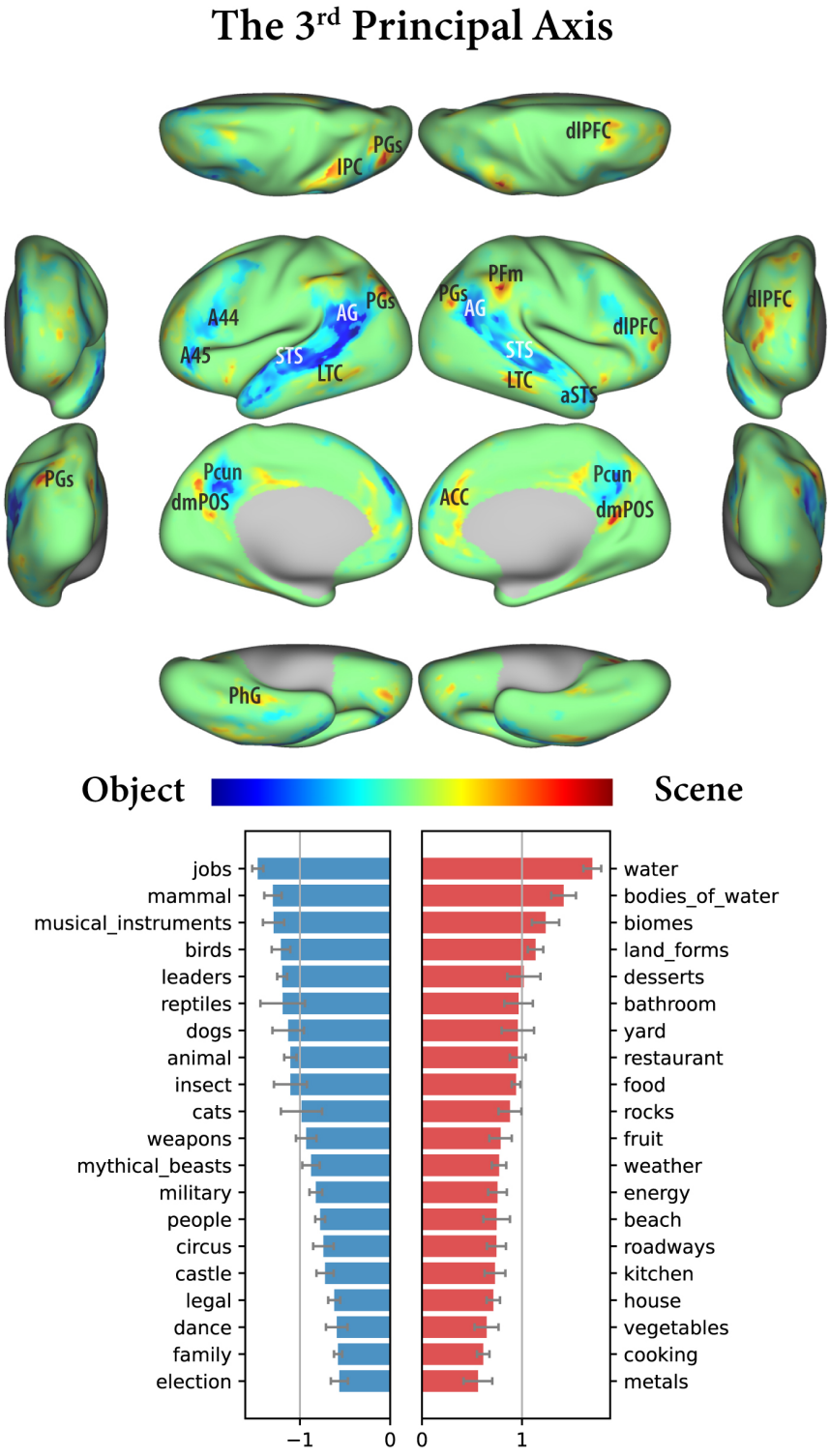


Figure 6.13: The cortical mapping and representative word categories for the third principal axis. The colorbar ranges from -1 to 1 after normalizing the resulting map. Blue: **Object**. Red: **Scene**.

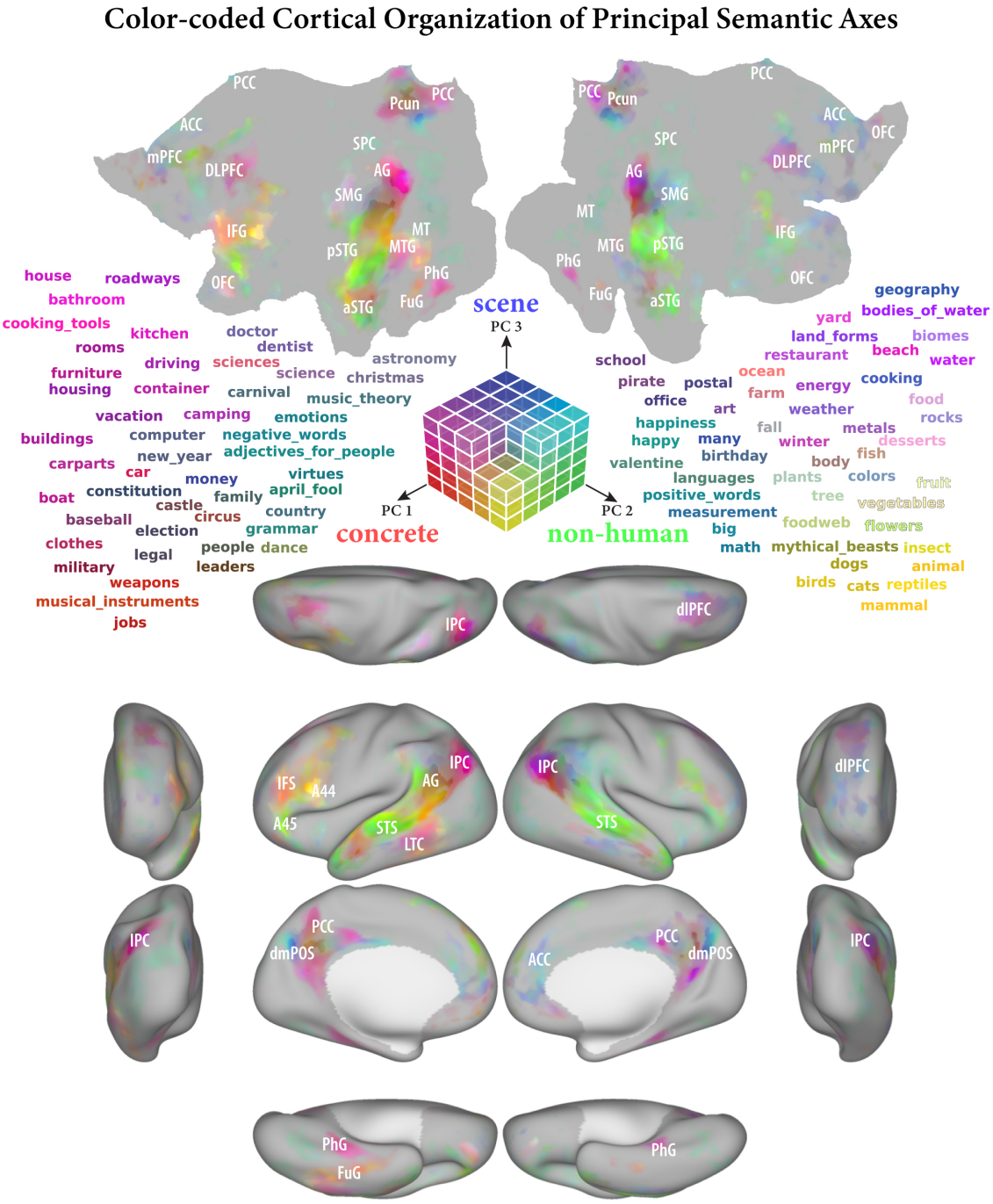


Figure 6.14: **Cortical representation of the first three principal axes in the grounded semantic space.** This map summarizes the cortical organization of all three principal components by color-coding each voxel with an RGB code, where reddish color towards **concrete**, greenish color towards **non-human**, bluish color towards **scene**.

6.4.5 Multimodal representation in the brain

In our prior study [189], we designed a fMRI experiment for musical imagery with visual cue to explore how visual and auditory information carrying the same (music) content are processed in the brain.

When a subject with musical training listened to an 8-min music piece eight times during the fMRI acquisition, activated cortical areas were mapped by identifying the voxels with reproducible fMRI signals across repetitions. Activated voxels were mostly confined to the auditory cortex, including the core and belt regions along the ventral auditory pathway, and Wernicke's area (Fig. 6.15(a)); the activations were slightly stronger in the right hemisphere than the left hemisphere.

In the second set of experiment, we visualized the music as a silent movie (Fig. 6.1, right). This movie provided real-time visual cues to inform subjects of the content of musical imagery and assisted in controlling the timing of the imagery process during the 8-min session. By watching this movie, each subject could consistently imagine the music piece for 12 repeated sessions of fMRI scans (All subjects reported experiencing vivid musical imagery during experiments, and could imagine the music with accurate tempo and pitch in mind). Similar to the activation analysis for musical perception, the cortical activation during the visually-cued musical imagery was mapped by assessing the intra-subject reproducibility of fMRI signals at the voxel level. The activated areas covered a large part of the cortex, including the primary visual cortex, dorsal visual areas, the parietal association cortex, the anterolateral belt, Wernicke's area, the frontal eye fields, the supplemental motor area and the premotor cortex (Fig. 6.15(b)). The responses at these areas could be attributable to either visual stimuli or musical imagery, since the task required the subject to process the visual cues as well as to imagine the music accordingly.

We further compared the task-evoked responses between the visually-cued musical imagery and musical perception conditions. This allowed us to localize the responses to musical imagery as opposed to the visual stimuli because no visual stimuli were given during the perception condition. We investigated the shared cortical substrates for both musical perception and imagery by assessing the fMRI signal correlation between each musical-imagery session and each music-perception session at each voxel. This analysis revealed the areas that showed consistent responses to both tasks (Fig. 6.15(c)). Such areas included Wernicke's area on the left hemisphere and its homologous area on the right hemisphere (herein we refer to them as the bilateral Wernicke's areas), and to a lesser degree the anterolateral belt, the supplementary motor areas, and the premotor cortex. Note that the perception task involved only auditory input but not visual input, whereas the imagery task involved only visual input but not auditory input. The voxel-wise correlation across these two task conditions was only attributable to the high-level music content in both conditions, regardless of its sensory modality. By visual inspection of the averaged fMRI signals, we found that both the imagery and perception tasks evoked complex but similar responses bilaterally in Wernicke's areas,

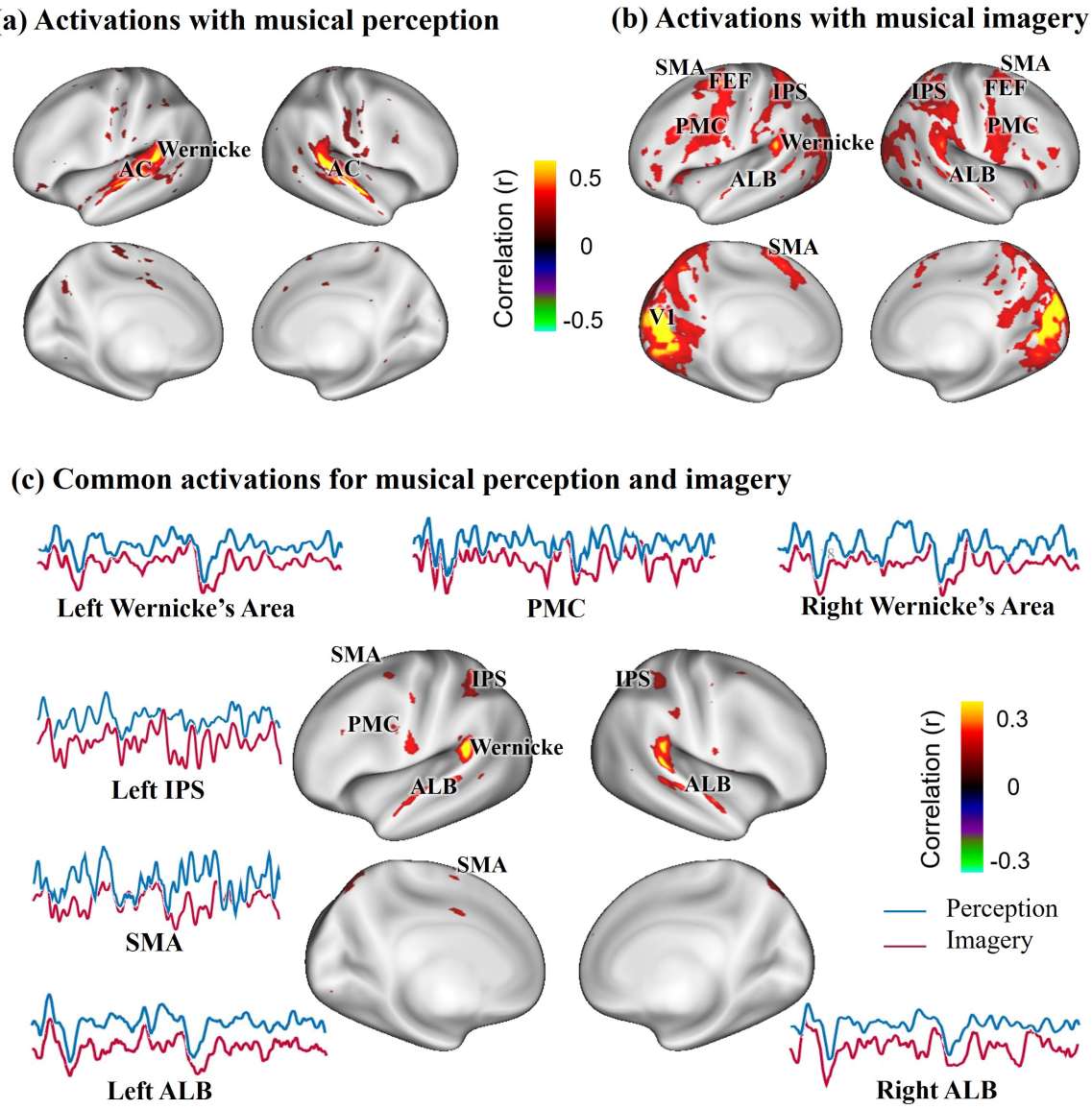


Figure 6.15: **Distinct and common cortical activations with musical perception and visually-cued imagery.** (a). Cortical activations for musical perception (two-tailed significance level $p < 0.01$). (b). Cortical activations for musical imagery (two-tailed significance level $p < 0.005$). (c). Shared cortical substrates between musical perception and musical imagery (two-tailed significance level $p < 0.01$). The time series was extracted from the fMRI signal averaged across the perception or imagery sessions from the labelled locations.

supplementary motor areas, and premotor cortex, over the entire duration of the music stimulus (Fig. 6.15(c))).

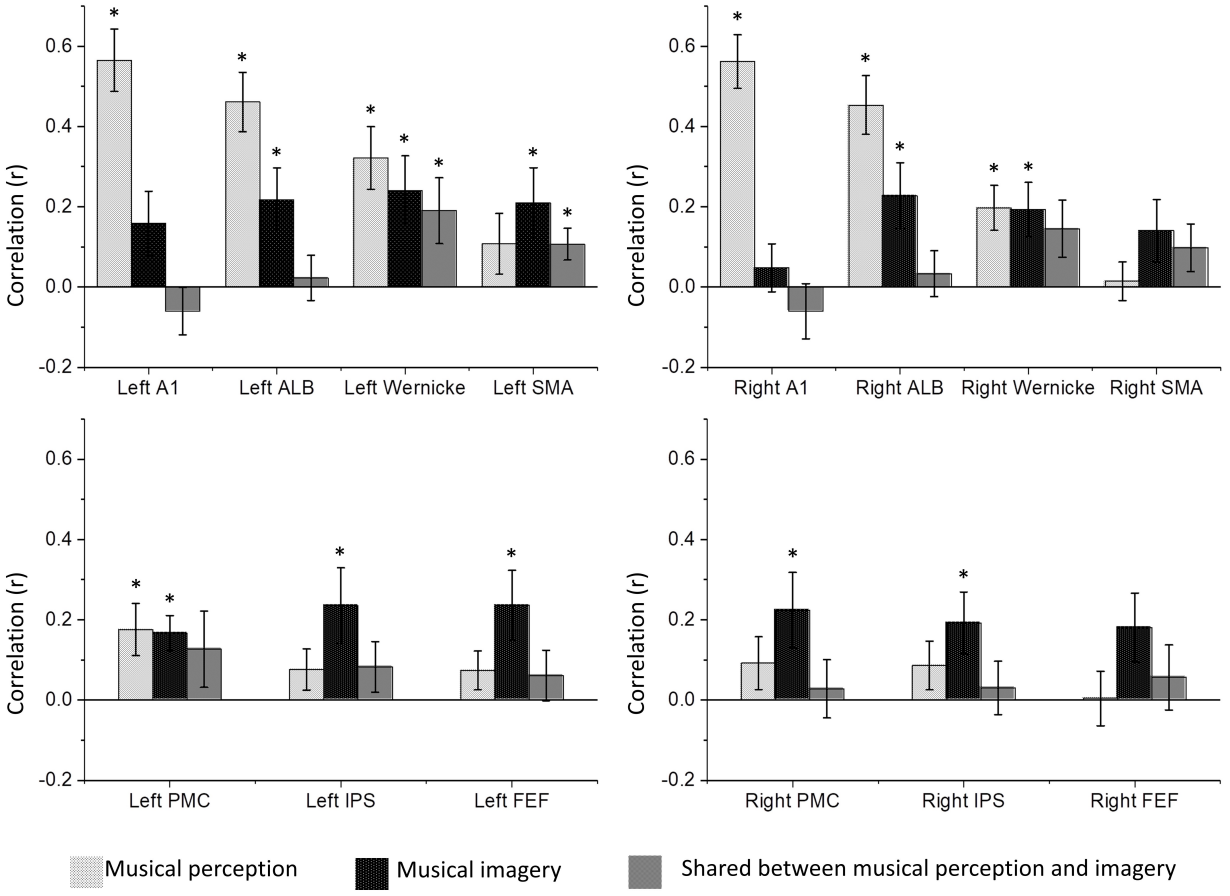


Figure 6.16: **Distinct and common cortical activations with musical perception and imagery from group-level analysis.** Each chart reflects the averaged correlation (r value) among all subjects in different regions of interest (ROI) compared across three conditions: reproducibility between musical perception sessions (light gray); reproducibility between musical imagery sessions (black); correlation between a musical perception session and a musical imagery session (dark gray).

Fig. 6.16 provided the results of group level analysis for distinct and common cortical activations with musical perception and imagery. The left two charts show the results for ROIs in the left hemisphere and the right two charts show the results for ROIs in the right hemisphere. The mark $*$ over a bar indicates that the specific ROI is consistently significantly activated by the musical perception or imagery task, or co-activated by both tasks among all subjects (two-tailed significance level $p < 0.05$). (A1: Primary auditory cortex; ALB: Auditory anterolateral belt; SMA: Supplementary motor area; PMC: Premotor cortex; FEF: Frontal eye field; IPS: Intraparietal sulcus).

We also explored how sound features were correlated with the cortical activity during musical

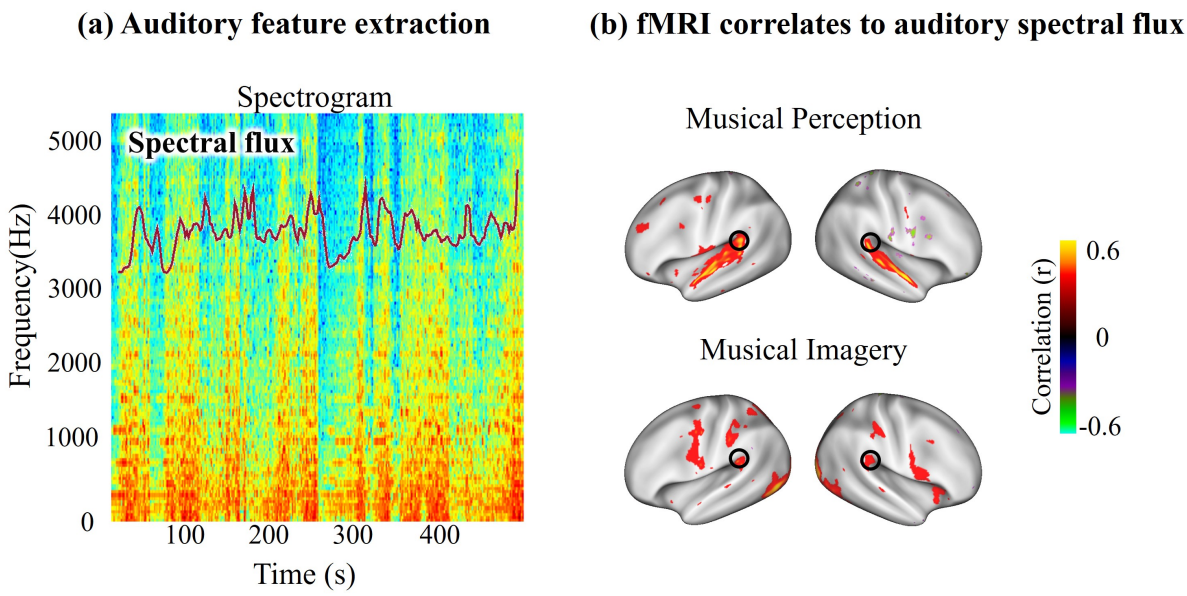


Figure 6.17: **Responses at Wernicke's areas coded musical features during imagery.** (a) The auditory spectral flux was extracted from the stimulus spectrogram as a feature showing how quickly the power spectrum of a sound wave changes over time. (b) Spectral flux was highly correlated with the fMRI signals (averaged across all subjects) in the common cortical regions shared between musical perception and imagery (corrected at false discovery rate (FDR) $q < 0.05$)

perception and musical imagery. As illustrated in Fig. 6.17(a), spectral flux, which measures the change in the power spectrum of a signal, was extracted by the MIRtoolbox in Matlab [97]. It is calculated as the 2-norm between the normalized spectra from adjacent frames [50]. Spectral flux is a sound feature related to music timbre. Fig. 6.17(b) shows that this feature was highly correlated with the fMRI signals (averaged across all subjects) in the common cortical regions shared between musical perception and imagery (corrected at false discovery rate (FDR) $q < 0.05$), especially in ventral visual areas and bilaterally in Wernicke's areas (as circled on the maps).

6.5 Summary and Discussion

With the existing word2vec embeddings, we built a neural encoding model for cortical response evoked by language comprehension using fMRI or ECoG signals from subjects listening to natural story stimuli. The encoding model could predict brain signals with high accuracy in widely distributed brain regions. We also mapped the cortical representation of semantic categories and semantic relations through the encoding model. After developing and evaluating a visually grounded language model, we further mapped the cortical organization of principal axes in the grounded semantic space by training a novel encoding model with the grounded word embeddings learned from Section 4.2.5.2. The results collectively suggest that the human semantic system represents conceptual features by distributed and overlapping cortical regions, including some multimodal association areas, supporting the grounded cognition theory. In another study, our preliminary results from the musical imagery experiments suggest that the representation of high-level multimodal information of the stimulus has invariant activities in some brain regions regardless of its input sensory modality, which sheds light on the cognitive processing mechanism of multimodal information in the brain.

Specifically, we found that semantic categories are represented not by segregated cortical regions, but by distributed and overlapping cortical patterns. Although the cortical representations of words collectively constitute a bilateral semantic system, the left hemisphere tends to be more selective to concrete concepts. More importantly, semantic relations reflecting conceptual progression from concreteness (i.e., *part*) to abstractness (i.e., *whole*) are represented by the co-occurrence of activation in the default-mode network and deactivation in the attention network. But not all differential vectors between word pairs with the same semantic relations were aligned well in the word2vec embedding space, especially for the relations that could not be consistently mapped to the brain with the current method (see details in Supplementary Information of [191]).

After grounding language in vision, the semantic space can be decomposed into a set of explainable principal axes (Section 5.3.1). These principal semantic axes are mapped onto distinctive patterns that involve highly selective cortical regions. It suggests that the principal semantic dimensions are encoded by different sub-systems in the human semantic system. Furthermore, we found early grounding (8 learnable layers in Bert; see results in Section 5.3.1) or late grounding (4 learnable layers in Bert; see results in Section 6.4.4) showed consistent semantic attributes for the first three principal axes in their semantic space, respectively: the 1st principal component always capture the contrast between abstract and concrete concepts, while the 2nd and the 2rd principal components always capture the semantic feature explaining concepts for non-human vs. human, and object vs. scene. This finding implies that these semantic attributes inherently and robustly span the dominant components in the grounded semantic space regardless of the different training

settings.

Overall, the results from the neural encoding models amount to a coherent conclusion that the human brain represents a continuous semantic space. Specifically, our findings go one step further than previous studies by showing the brain uses distributed cortical networks to encode not only concepts, but also the **relationships between concepts** and the **semantic attributes of concepts** to support conceptual inference and reasoning.

To assess the hierarchy of language processing system, we also did a latency analysis to estimate the delay of cortical activation responding to the natural story stimuli given the high temporal resolution in ECoG signals [193]. The latency analysis was explored separately for each channel, by allowing high gamma ECoG to lag behind speech stimuli by a variable delay (from 0 to 1500ms with 100ms increments) before calculating the correlation. The results suggest different ECoG channels had a varying latency for words encoded by the recorded high-gamma activity (Fig. 6.18). After subtracting the onset delay, the estimated latency in STG varied from 0 to 300ms, showing longer delays in higher cognitive regions. Future work connecting different layers in the grounded language model to the brain's hierarchical language processing is worth exploring with the ECoG data.

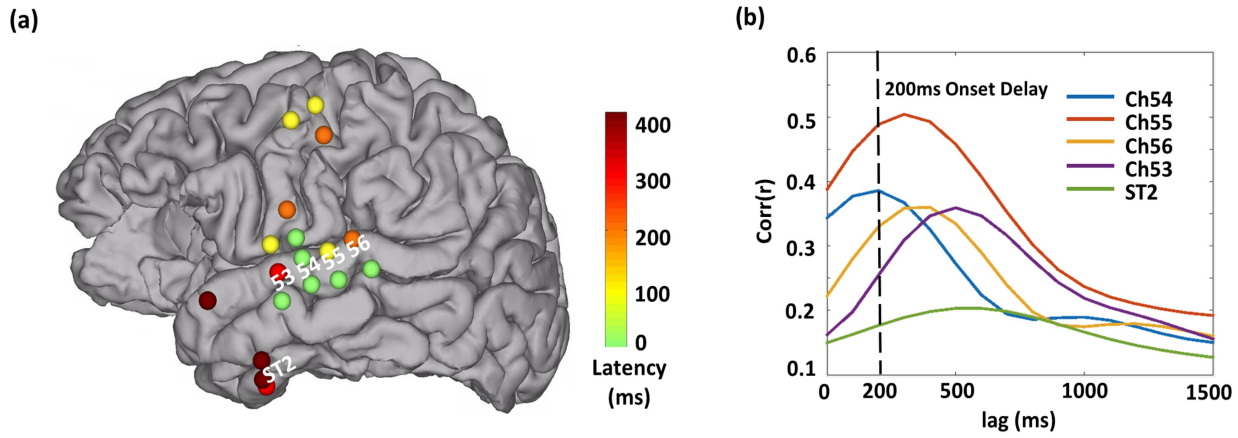


Figure 6.18: **Variant latency at different ECoG channels revealed by the encoding model** (a) The estimated latency in 18 significant channels. (b) Different brain regions have relatively different encoding latency for word-level information.

CHAPTER 7

Future Directions

This study is motivated by developing brain-inspired computational models to **bridge fundamental gaps between how humans and machines learn language**. On one hand, I hope we can incorporate the inspiration from human intelligence into machine learning to better support natural language processing. On the other hand, I hope a brain-like and interpretable language learning model can be used to help study the biological neural network to advance progress in neuroscience.

However, I have to admit that both deep learning models and human brains are like a "black box" system to some extent, which requires further explanation and understanding of their underlying computational mechanisms. Although studying language learning in either machines or humans remains a challenging problem, I believe that filling in the known gaps and constructing brain-like computational models is the right research direction, which can promote synergistic progress in both AI and neuroscience.

Specifically, I would like to discuss the potential future directions to further expand the current work. I emphasize two perspectives: 1. ground language learning in multiple modalities; 2. develop more biologically plausible learning schemes for language grounding models.

7.1 Grounding language learning in multiple modalities

Vision is just one of the neural systems for human perception. We not only perceive the world through our eyes but also through our ears (auditory), nose (olfactory), mouth (gustatory), and hands (tactile sensation). We also interact with real-world objects and environments through executable actions (sensorimotor). As human beings, emotions and feelings emerge in our minds through perception and interactions with the physical world. These experiences collectively embody our cognition. In this section, I will discuss the next steps toward developing grounded language learning in perception, action, and emotion. The ultimate goal is to build a comprehensive computational model to explain the grounded cognition theory.

7.1.1 Perception

Our brain is profoundly a multimodal system. The different sensory systems can educate each other without an external teacher [159]. Therefore, it is ideal for developing a unified multi-stream model with each stream as an analogy to the different sensory systems in the brain to ground language learning in various modalities. The cross-modal learning strategy demonstrated in this study can be naturally extended to include multiple modalities. For example, we can train a three-stream model given an audio-visual movie dataset, by inputting synchronized video frames, audio tracks, and text scripts (like descriptions of the movie contents). There are some handy datasets to train such models, such as documentary movies with narrations and movies with audio descriptions (e.g., The Audio Description Project).

Visual features and auditory features are relatively easy to model as real-valued tensor representations through convolution or spectral decomposition since the input data of these two modalities are already in well-structured digital forms. However, it is more difficult to model and extract data and features as computable items from olfactory, gustatory, and tactile systems. An indirect way to ground language learning in olfactory and gustatory systems is to use the information from chemical compounds of individual objects to build a representation of perceptual features based on the so-called "bag of chemical compounds" model [87]. How to computationally model smell, taste, or touch remains an open question, because even the most advanced electrochemical sensors cannot imitate and simulate these human sensations. Recent works using biological signals collected from receptors implanted on animals (e.g., bees) to build neuromorphic computing models of the olfaction system have shown promising progress in understanding the underlying computational mechanism [136]. In general, extending the current work to ground concept representations in multiple perceptual domains is worthy of future exploration.

7.1.2 Action

Grounding language learning in action is naturally related to training agents or robots in an environment with natural language instructions or communication [184, 25, 135, 115]. Action is an essential component for building semantic representations of verbs and prepositions. If a language model is only grounded in vision, concepts related to predicate words will be hard to learn since they require further information, such as spatial locations, spatial relations, the executable actions (e.g., we can *eat* or *cut* apples but not *drink* apples), and results of an action (e.g., we feel *full* by eating apples). Similarly, it may be necessary to pretrain a language model with a large textual corpus first and then transfer it for actional grounding. Because the sampling of words for instructions and communications with an robot agent in a virtual environment can hardly cover the semantic space without bias.

An intuitive way to ground language learning in action with a similar scheme as in the current study is to add an individual actional encoder to the visual-language two-stream model. The visual stream takes *what the agent sees* as input and output the visual embedding of the environment. The language stream takes *what the agent hears* (e.g., a natural language instruction "find an apple") as input and outputs the semantic representation of the target action. While the "action encoder" takes the state of the environment as input and outputs a feasible action. The whole model can be trained with reinforcement learning to optimize the reward of performing correct action given an instruction.

7.1.3 Emotion

Emotions play an essential role in human cognition. To ground word embeddings in emotions, traditional methods use tasks such as sthe entiment analysis to predict affective labels from text input [116]. Recent studies have also used text to pair with emojis collected from social media datasets to learn the affective information in natural language by predicting concurrent emojis [145]. Most words related to intense emotional characteristics are abstract concepts and cannot be directly captured by visual perception. Therefore, it is not surprising that in our results, all abstract concepts, no matter conveying positive or negative sentiment, are clustered together and cannot be separated into more exemplary sets in the visually grounded semantic space. The key limitation of emotional modeling or affective computing is the lack of high-dimensional space to represent emotional features. Although facial expressions, body gestures, and physiological signals (e.g., heart rate, body temperature, and galvanic skin response) all reflect rich emotional information, whether human emotions are collectively represented in a continuous space or as discrete states is still controversial [120]. Adding an emotional feature extractor in a language grounding model can not only lead to explainable word representations of abstract concepts but also shed light on the organization of emotional feature space.

7.2 Developing biologically plausible learning paradigms

Besides being "ungrounded" in multimodal contexts, current natural language learning models differ from human learners in many ways. Specifically, instead of passively perceiving multimodal information to learn concepts, humans are actively involved in interacting with the surrounding environment to learn concepts. In addition, our brain can quickly adapt to new knowledge in just a few attempts without forgetting previously learned skills. However, machine learning models usually require hundreds of repeated iterations to train on a single task. Furthermore, it is difficult for these models to maintain the performance of old tasks after retraining and finetuning on new

ones. Two brain-inspired learning schemes are likely helpful to bridge these gaps: interactive reinforcement learning and continual lifelong learning.

7.2.1 Interactive reinforcement learning

To imitate the way humans learn languages, it is desirable to place an autonomous agent in a virtual environment that exhibits rich multimodal information and allows executable operations with a high degree of freedom, such that the agent can learn concepts grounded in real-world experience. This objective fits well with the reinforcement learning scheme by treating an autonomous agent as an analogy of "language learner". Language learning is a demanding cognitive task that requires a large amount of data to learn from scratch. Thus, strategies like knowledge distillation and interactive learning that engage a more powerful "teacher" or "oracle" can likely boost the learning process. Specifically, interactive reinforcement learning involves a human-in-the-loop to provide feedback from external evaluations to tailor specific elements (e.g., policy reward function) in the algorithm modeling agent behaviors. Such a strategy shows a faster convergence rate and better performance by integrating prior knowledge from humans [4]. In addition, introducing multiple agents into the environment to promote interaction between subjects will further inject a social component into the computational modeling of grounded cognition. These ideas could be integrated into language grounding with an autonomous agent and a multimodal environment, as discussed in the Section 7.1.

7.2.2 Continual lifelong learning

One obvious limitation of machine learning models is that they are subject to catastrophic interference. That is, novel information interferes with the consolidated knowledge migrating a trained model for a new task [133]. This is also known as the stability-plasticity dilemma [121]. However, it is not a problem for humans to continuously and effectively acquire new skills and transfer knowledge across modalities while retaining what they have learned in their lifespan. Our brain's ability to simultaneously learning and protecting knowledge is achievable due to several key characteristics of the biological neural network. The neurosynaptic plasticity for developmental stages allows the flexibility to learn for new knowledge. At the same time, neurogenesis for memory formation enables the brain to remember previously learned skills, as suggested by cognitive science theories and neurophysiological studies [46].

Such biological foundations inspire some learning strategies toward continual lifelong learning in artificial intelligence models. For example, we can add a memory module to a computational system for storing learned information or allocating additional neural resources (e.g., more artificial neurons) to learn new tasks. However, these methods are not scalable to lifetime learning of an unlimited amount of tasks. According to how humans experience the world, another way is to

train the computational model with interleaved stimuli from different tasks instead of training it with sequential tasks. But this also suffers from generalizability on unpredefined tasks. Recently researchers have developed other strategies, including dual-memory learning systems to simultaneously reconcile short- and long-term memory [49], retraining neural networks with regularization on model plasticity [107], reusing and transferring information between tasks [165]. Inspired by the critical learning periods in humans and animals, using curriculum learning to incrementally start with simple tasks and gradually move onto more difficult tasks has also been explored for continual lifelong learning [59]. Conceptually, the three-stage training strategy (from unimodal to multimodal; from image-caption mapping to object-relations inference) in our approach follows the same intuition.

The continual lifelong learning paradigm can incorporate many intuitive learning principles imitating human learning behaviors, such as few-shot learning, unsupervised learning, and learning from intrinsic motivation (or curiosity). Thus, the long-term goal of bridging language learning in machines and humans is to understand and implement the brain's computational mechanism that makes us adaptive, efficient, robust, and lifelong learners.

BIBLIOGRAPHY

[1] D. Adolf, S. Weston, S. Baecke, M. Luchtmann, J. Bernarding, and S. Kropf. Increasing the reliability of data analysis of functional magnetic resonance imaging by applying a new blockwise permutation method. *Frontiers in neuroinformatics*, 8:72, 2014.

[2] M. L. Anderson. Embodied cognition: A field guide. *Artificial intelligence*, 149(1):91–130, 2003.

[3] G. K. Anumanchipalli, J. Chartier, and E. F. Chang. Speech synthesis from neural decoding of spoken sentences. *Nature*, 568(7753):493–498, 2019.

[4] C. Arzate Cruz and T. Igarashi. A survey on interactive reinforcement learning: Design principles and open challenges. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference*, pages 1195–1209, 2020.

[5] E. Asgari and M. R. Mofrad. Continuous distributed representation of biological sequences for deep proteomics and genomics. *PloS one*, 10(11):e0141287, 2015.

[6] Y. Aytar, C. Vondrick, and A. Torralba. See, hear, and read: Deep aligned representations. *arXiv preprint arXiv:1706.00932*, 2017.

[7] M. Baroni. Grounding distributional semantics in the visual world. *Language and Linguistics Compass*, 10(1):3–13, 2016.

[8] L. W. Barsalou. Grounded cognition. *Annu. Rev. Psychol.*, 59:617–645, 2008.

[9] L. W. Barsalou et al. Perceptual symbol systems. *Behavioral and brain sciences*, 22(4):577–660, 1999.

[10] S. Becker and G. E. Hinton. Self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature*, 355(6356):161–163, 1992.

[11] L. Beinborn, T. Botschen, I. Gurevych, et al. Multimodal grounding for language processing. In *COLING*, pages 2325–2339, 2018.

[12] I. Bello, B. Zoph, A. Vaswani, J. Shlens, and Q. V. Le. Attention augmented convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3286–3295, 2019.

[13] I. Beltagy, S. Roller, P. Cheng, K. Erk, and R. J. Mooney. Representing meaning with a combination of logical and distributional models. *Computational Linguistics*, 42(4):763–808, 2016.

[14] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin. A neural probabilistic language model. *The journal of machine learning research*, 3:1137–1155, 2003.

[15] J. R. Binder and R. H. Desai. The neurobiology of semantic memory. *Trends in cognitive sciences*, 15(11):527–536, 2011.

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

[17] Y. Bisk, A. Holtzman, J. Thomason, J. Andreas, Y. Bengio, J. Chai, M. Lapata, A. Lazaridou, J. May, A. Nisnevich, et al. Experience grounds language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735, 2020.

[18] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko. Translating embeddings for modeling multi-relational data. In *Neural Information Processing Systems (NIPS)*, pages 1–9, 2013.

[19] K. E. Bouchard, N. Mesgarani, K. Johnson, and E. F. Chang. Functional organization of human sensorimotor cortex for speech articulation. *Nature*, 495(7441):327–332, 2013.

[20] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah. Signature verification using a "siamese" time delay neural network. *Advances in neural information processing systems*, 6:737–744, 1993.

[21] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901, 2020.

[22] E. Bruni, G. Boleda, M. Baroni, and N.-K. Tran. Distributional semantics in technicolor. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 136–145, 2012.

[23] E. Bruni, N.-K. Tran, and M. Baroni. Multimodal distributional semantics. *Journal of artificial intelligence research*, 49:1–47, 2014.

[24] M. Brysbaert, A. B. Warriner, and V. Kuperman. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 46(3):904–911, 2014.

[25] J. Y. Chai, Q. Gao, L. She, S. Yang, S. Saba-Sadiya, and G. Xu. Language to action: Towards interactive task learning with physical agents. In *IJCAI*, pages 2–9, 2018.

[26] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[27] X. Chen, L.-J. Li, L. Fei-Fei, and A. Gupta. Iterative visual reasoning beyond convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7239–7248, 2018.

[28] Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu. UNITER: UNiversal Image-TExt Representation Learning. In *Computer Vision – ECCV 2020*, pages 104–120. Springer International Publishing, 2020.

[29] G. Collell, T. Zhang, and M.-F. Moens. Imagined visual representations as multimodal embeddings. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[30] M. Corbetta and G. L. Shulman. Control of goal-directed and stimulus-driven attention in the brain. *Nature reviews neuroscience*, 3(3):201–215, 2002.

[31] N. E. Crone, D. L. Miglioretti, B. Gordon, and R. P. Lesser. Functional mapping of human sensorimotor cortex with electrocorticographic spectral analysis. II. Event-related synchronization in the gamma band. *Brain: a journal of neurology*, 121(12):2301–2315, 1998.

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

[33] F. Deniz, A. O. Nunez-Elizalde, A. G. Huth, and J. L. Gallant. The representation of semantic information across human cerebral cortex during listening versus reading is invariant to stimulus modality. *Journal of Neuroscience*, 39(39):7722–7736, 2019.

[34] B. J. Devereux, L. K. Tyler, J. Geertzen, and B. Randall. The Centre for Speech, Language and the Brain (CSLB) concept property norms. *Behavior research methods*, 46(4):1119–1127, 2014.

[35] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[36] L. A. Doumas, J. E. Hummel, and C. M. Sandhofer. A theory of the discovery and predication of relational concepts. *Psychological review*, 115(1):1, 2008.

[37] F. R. Dreyer, T. Picht, D. Frey, P. Vajkoczy, and F. Pulvermueller. The functional relevance of dorsal motor systems for processing tool nouns–evidence from patients with focal lesions. *Neuropsychologia*, 141:107384, 2020.

[38] F. R. Dreyer and F. Pulvermüller. Abstract semantics in the motor system? - An event-related fMRI study on passive reading of semantic word categories carrying abstract emotional and mental meaning. *Cortex*, 100:52–70, 2018.

[39] G. Emerson. What are the goals of distributional semantics? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7436–7453, 2020.

[40] L. Fan, H. Li, J. Zhuo, Y. Zhang, J. Wang, L. Chen, Z. Yang, C. Chu, S. Xie, A. R. Laird, et al. The Human Brainnetome Atlas: A new brain atlas based on connectional architecture. *Cerebral cortex*, 26(8):3508–3526, 2016.

[41] L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin. Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web*, pages 406–414, 2001.

[42] J. R. Firth. A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*, 1957.

[43] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov. DeViSE: A deep visual-semantic embedding model. In *Proceedings of the 26th International Conference on Neural Information Processing Systems-Volume 2*, pages 2121–2129, 2013.

[44] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*, 2016.

[45] K. Fukushima and S. Miyake. Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In *Competition and cooperation in neural nets*, pages 267–285. Springer, 1982.

[46] A. Galván. Neural plasticity of development and learning. *Human brain mapping*, 31(6):879–890, 2010.

[47] M. Garnelo and M. Shanahan. Reconciling deep learning with symbolic artificial intelligence: representing objects and relations. *Current Opinion in Behavioral Sciences*, 29:17–23, 2019.

[48] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.

[49] A. Gepperth and C. Karaoguz. A bio-inspired incremental learning architecture for applied perceptual problems. *Cognitive Computation*, 8(5):924–934, 2016.

[50] D. Giannoulis, M. Massberg, and J. D. Reiss. Parameter automation in a dynamic range compressor. *Journal of the Audio Engineering Society*, 61(10):716–726, 2013.

[51] R. Girshick. Fast R-CNN. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[52] M. F. Glasser, T. S. Coalson, E. C. Robinson, C. D. Hacker, J. Harwell, E. Yacoub, K. Ugurbil, J. Andersson, C. F. Beckmann, M. Jenkinson, et al. A multi-modal parcellation of human cerebral cortex. *Nature*, 536(7615):171–178, 2016.

[53] M. F. Glasser, S. N. Sotiropoulos, J. A. Wilson, T. S. Coalson, B. Fischl, J. L. Andersson, J. Xu, S. Jbabdi, M. Webster, J. R. Polimeni, et al. The minimal preprocessing pipelines for the Human Connectome Project. *Neuroimage*, 80:105–124, 2013.

[54] A. M. Glenberg, T. Gutierrez, J. R. Levin, S. Japuntich, and M. P. Kaschak. Activity and imagined activity can enhance young children's reading comprehension. *Journal of educational psychology*, 96(3):424, 2004.

[55] A. M. Glenberg and D. A. Robertson. Symbol grounding and meaning: A comparison of high-dimensional and embodied theories of meaning. *Journal of memory and language*, 43(3):379–401, 2000.

[56] G. H. Golub, M. Heath, and G. Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223, 1979.

[57] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[58] A. Gopnik and A. N. Meltzoff. Semantic and cognitive development in 15-to 21-month-old children. *Journal of child language*, 11(3):495–513, 1984.

[59] A. Graves, M. G. Bellemare, J. Menick, R. Munos, and K. Kavukcuoglu. Automated curriculum learning for neural networks. In *international conference on machine learning*, pages 1311–1320. PMLR, 2017.

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

[61] F. Günther, T. Nguyen, L. Chen, C. Dudschig, B. Kaup, and A. M. Glenberg. Immediate sensorimotor grounding of novel concepts learned from language alone. *Journal of Memory and Language*, 115:104172, 2020.

[62] M. Gutmann and A. Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 297–304. JMLR Workshop and Conference Proceedings, 2010.

[63] H. Hakami, D. Bollegala, and H. Kohei. Why pairdiff works? - A mathematical analysis of bilinear relational compositional operators for analogy detection. *arXiv preprint arXiv:1709.06673*, 2017.

[64] K. Han, H. Wen, J. Shi, K.-H. Lu, Y. Zhang, D. Fu, and Z. Liu. Variational autoencoder: An unsupervised model for encoding and decoding fMRI activity in visual cortex. *NeuroImage*, 198:125–136, 2019.

[65] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, et al. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*, 2014.

[66] S. Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346, 1990.

[67] Z. S. Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.

[68] D. Harwath, A. Recasens, D. Surís, G. Chuang, A. Torralba, and J. Glass. Jointly discovering visual objects and spoken words from raw sensory input. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 649–665, 2018.

[69] D. Hassabis, D. Kumaran, C. Summerfield, and M. Botvinick. Neuroscience-inspired artificial intelligence. *Neuron*, 95(2):245–258, 2017.

[70] U. Hasson, R. Malach, and D. J. Heeger. Reliability of cortical activity during natural stimulation. *Trends in cognitive sciences*, 14(1):40–48, 2010.

[71] K. He, R. Girshick, and P. Dollár. Rethinking ImageNet pre-training. In *Proceedings of the IEEE international conference on computer vision*, pages 4918–4927, 2019.

[72] K. He, X. Zhang, S. Ren, and J. 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.

[73] F. Hill, R. Reichart, and A. Korhonen. Multi-modal models for concrete and abstract concept meaning. *Transactions of the Association for Computational Linguistics*, 2:285–296, 2014.

[74] D. H. Hubel and T. N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 160(1):106–154, 1962.

[75] D. A. Hudson and C. D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6700–6709, 2019.

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

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

[78] G. Ilharco, R. Zellers, A. Farhadi, and H. Hajishirzi. Probing text models for common ground with visual representations. *arXiv preprint arXiv:2005.00619*, 2020.

[79] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. V. Le, Y. Sung, Z. Li, and T. Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. *arXiv preprint arXiv:2102.05918*, 2021.

[80] H. Jiang, I. Misra, M. Rohrbach, E. Learned-Miller, and X. Chen. In defense of grid features for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10267–10276, 2020.

[81] X. Jin, J. Du, A. Sadhu, R. Nevatia, and X. Ren. Visually grounded continual learning of compositional phrases. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2018–2029, 2020.

[82] B. Jónsson. Maximal algebras of binary relations. *Contemporary Mathematics*, 33:299–307, 1984.

[83] D. Jurgens, S. Mohammad, P. Turney, and K. Holyoak. Semeval-2012 task 2: Measuring degrees of relational similarity. In ** SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 356–364, 2012.

[84] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.

[85] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.

[86] A. J. Kell, D. L. Yamins, E. N. Shook, S. V. Norman-Haignere, and J. H. McDermott. A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, 98(3):630–644, 2018.

[87] D. Kiela, L. Bulat, and S. Clark. Grounding semantics in olfactory perception. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 231–236, 2015.

[88] D. Kiela and S. Clark. Multi-and cross-modal semantics beyond vision: Grounding in auditory perception. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2461–2470, 2015.

[89] D. Kiela, A. Conneau, A. Jabri, and M. Nickel. Learning visually grounded sentence representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 408–418, 2018.

[90] J.-H. Kim, J. Jun, and B.-T. Zhang. Bilinear attention networks. In *Advances in Neural Information Processing Systems*, pages 1564–1574, 2018.

[91] R. Kiros, R. Salakhutdinov, and R. S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014.

[92] S. Knecht, M. Deppe, B. Dräger, L. Bobe, H. Lohmann, E.-B. Ringelstein, and H. Henningsen. Language lateralization in healthy right-handers. *Brain*, 123(1):74–81, 2000.

[93] N. Kriegeskorte and T. Golan. Neural network models and deep learning. *Current Biology*, 29(7):R231–R236, 2019.

[94] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.

[95] B. M. Lake and G. L. Murphy. Word meaning in minds and machines. *arXiv preprint arXiv:2008.01766*, 2020.

[96] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*, 2016.

[97] O. Lartillot and P. Toiviainen. A matlab toolbox for musical feature extraction from audio. In *International conference on digital audio effects*, pages 237–244. Bordeaux, 2007.

[98] A. Lazaridou, M. Baroni, et al. Combining language and vision with a multimodal skip-gram model. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 153–163, 2015.

[99] P. H. Le-Khac, G. Healy, and A. F. Smeaton. Contrastive representation learning: A framework and review. *IEEE Access*, 2020.

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

[101] T. Leibovich and D. Ansari. The symbol-grounding problem in numerical cognition: a review of theory, evidence, and outstanding questions. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, 70(1):12, 2016.

[102] D. A. Leopold and S. H. Park. Studying the visual brain in its natural rhythm. *Neuroimage*, 216:116790, 2020.

[103] K. Li, Y. Zhang, K. Li, Y. Li, and Y. Fu. Visual semantic reasoning for image-text matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4654–4662, 2019.

[104] L. Li, Z. Gan, Y. Cheng, and J. Liu. Relation-aware graph attention network for visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10313–10322, 2019.

[105] L. Li and J. Gauthier. Are distributional representations ready for the real world? evaluating word vectors for grounded perceptual meaning. In *Proceedings of the First Workshop on Language Grounding for Robotics*, pages 76–85, 2017.

[106] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020.

[107] Z. Li and D. Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017.

[108] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common Objects in Context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[109] M. A. Lindquist, J. M. Loh, L. Y. Atlas, and T. D. Wager. Modeling the hemodynamic response function in fMRI: efficiency, bias and mis-modeling. *Neuroimage*, 45(1):S187–S198, 2009.

[110] G. W. Lindsay. Convolutional neural networks as a model of the visual system: past, present, and future. *Journal of cognitive neuroscience*, pages 1–15, 2020.

[111] L. Logeswaran and H. Lee. An efficient framework for learning sentence representations. *arXiv preprint arXiv:1803.02893*, 2018.

[112] N. K. Logothetis, J. Pauls, M. Augath, T. Trinath, and A. Oeltermann. Neurophysiological investigation of the basis of the fMRI signal. *nature*, 412(6843):150–157, 2001.

[113] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei. Visual relationship detection with language priors. In *European conference on computer vision*, pages 852–869. Springer, 2016.

[114] J. Lu, D. Batra, D. Parikh, and S. Lee. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23, 2019.

[115] C. Lynch and P. Sermanet. Grounding language in play. *arXiv preprint arXiv:2005.07648*, 2020.

[116] A. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 142–150, 2011.

[117] S. Malinowski and L. Turetsky. Music animation machine. *Music Worth Watching"," http://www. musanim. com*, 2011.

[118] J. Mao, C. Gan, P. Kohli, J. B. Tenenbaum, and J. Wu. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. In *International Conference on Learning Representations*, 2018.

[119] A. Martin. GRAPES — Grounding representations in action, perception, and emotion systems: How object properties and categories are represented in the human brain. *Psychonomic bulletin & review*, 23(4):979–990, 2016.

[120] A. Martinez and S. Du. A model of the perception of facial expressions of emotion by humans: research overview and perspectives. *Journal of Machine Learning Research*, 13(5), 2012.

[121] M. Mermillod, A. Bugaiska, and P. Bonin. The stability-plasticity dilemma: Investigating the continuum from catastrophic forgetting to age-limited learning effects. *Frontiers in psychology*, 4:504, 2013.

[122] T. Mikolov, M. Karafiát, L. Burget, J. Černockỳ, and S. Khudanpur. Recurrent neural network based language model. In *Eleventh annual conference of the international speech communication association*, 2010.

[123] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[124] T. Mikolov, W.-t. Yih, and G. Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, 2013.

[125] G. A. Miller. *WordNet: An electronic lexical database*. MIT press, 1998.

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

[127] T. Naselaris, K. N. Kay, S. Nishimoto, and J. L. Gallant. Encoding and decoding in fMRI. *Neuroimage*, 56(2):400–410, 2011.

[128] M. Nickel, L. Rosasco, and T. Poggio. Holographic embeddings of knowledge graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

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

[130] J. Oh, X. Guo, H. Lee, R. Lewis, and S. Singh. Action-conditional video prediction using deep networks in Atari games. *arXiv preprint arXiv:1507.08750*, 2015.

[131] A. v. d. Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[132] A. Owens and A. A. Efros. Audio-visual scene analysis with self-supervised multisensory features. *European Conference on Computer Vision (ECCV)*, 2018.

[133] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019.

[134] N. Parmar, P. Ramachandran, A. Vaswani, I. Bello, A. Levskaya, and J. Shlens. Stand-alone self-attention in vision models. In *Advances in Neural Information Processing Systems*, pages 68–80, 2019.

[135] R. Paul, J. Arkin, N. Roy, and T. M Howard. Efficient grounding of abstract spatial concepts for natural language interaction with robot manipulators. *The International Journal of Robotics Research*, 37(10):1269–1299, 2018.

[136] F. Peng and L. Chittka. A simple computational model of the bee mushroom body can explain seemingly complex forms of olfactory learning and memory. *Current Biology*, 27(2):224–230, 2017.

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

[138] G. Pezzulo, L. W. Barsalou, A. Cangelosi, M. H. Fischer, K. McRae, and M. Spivey. Computational Grounded Cognition: a new alliance between grounded cognition and computational modeling. *Frontiers in psychology*, 3:612, 2013.

[139] F. Pulvermüller. How neurons make meaning: brain mechanisms for embodied and abstract-symbolic semantics. *Trends in cognitive sciences*, 17(9):458–470, 2013.

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

[141] M. Riesenhuber and T. Poggio. Computational models of object recognition in cortex: A review. 2000.

[142] J. Robinson, C.-Y. Chuang, S. Sra, and S. Jegelka. Contrastive learning with hard negative samples. *arXiv preprint arXiv:2010.04592*, 2020.

[143] S. Roller and S. S. Im Walde. A multimodal lda model integrating textual, cognitive and visual modalities. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1146–1157, 2013.

[144] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[145] A. S. Rotaru and G. Vigliocco. Constructing semantic models from words, images, and emojis. *Cognitive science*, 44(4):e12830, 2020.

[146] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. ImageNet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[147] A. Santoro, D. Raposo, D. G. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, and T. Lillicrap. A simple neural network module for relational reasoning. *Advances in Neural Information Processing Systems*, 30:4967–4976, 2017.

[148] G. Schmidt. *Relational mathematics*. Number 132. Cambridge University Press, 2011.

[149] T. Schnabel, I. Labutov, D. Mimno, and T. Joachims. Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 298–307, 2015.

[150] F. Schroff, D. Kalenichenko, and J. Philbin. FaceNet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[151] M. Scolari, K. N. Seidl-Rathkopf, and S. Kastner. Functions of the human frontoparietal attention network: Evidence from neuroimaging. *Current opinion in behavioral sciences*, 1:32–39, 2015.

[152] J. R. Searle. Minds, brains, and programs. *Behavioral and brain sciences*, 3(3):417–424, 1980.

[153] L. K. Şenel, I. Utlu, V. Yücesoy, A. Koc, and T. Cukur. Semantic structure and interpretability of word embeddings. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(10):1769–1779, 2018.

[154] H. Shi, J. Mao, T. Xiao, Y. Jiang, and J. Sun. Learning visually-grounded semantics from contrastive adversarial samples. *arXiv preprint arXiv:1806.10348*, 2018.

[155] J. Shi, H. Wen, Y. Zhang, K. Han, and Z. Liu. Deep recurrent neural network reveals a hierarchy of process memory during dynamic natural vision. *Human brain mapping*, 39(5):2269–2282, 2018.

[156] C. Silberer and M. Lapata. Grounded models of semantic representation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1423–1433, 2012.

[157] C. Silberer and M. Lapata. Learning grounded meaning representations with autoencoders. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 721–732, 2014.

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

[159] L. Smith and M. Gasser. The development of embodied cognition: Six lessons from babies. *Artificial life*, 11(1-2):13–29, 2005.

[160] K. Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 1857–1865, 2016.

[161] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*, 2019.

[162] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

[163] H. Tan and M. Bansal. LXMERT: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

[164] K. Tang, Y. Niu, J. Huang, J. Shi, and H. Zhang. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3716–3725, 2020.

[165] C. Tessler, S. Givony, T. Zahavy, D. Mankowitz, and S. Mannor. A deep hierarchical approach to lifelong learning in Minecraft. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[166] A. Utsumi. Exploring what is encoded in distributional word vectors: A neurobiologically motivated analysis. *Cognitive Science*, 44(6):e12844, 2020.

[167] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[168] P. Vogt. Language evolution and robotics: Issues on symbol grounding and language acquisition. In *Artificial cognition systems*, pages 176–209. IGI Global, 2007.

[169] E. Voita, D. Talbot, F. Moiseev, R. Sennrich, and I. Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. *arXiv preprint arXiv:1905.09418*, 2019.

[170] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018.

[171] B. Wang, A. Wang, F. Chen, Y. Wang, and C.-C. J. Kuo. Evaluating word embedding models: methods and experimental results. *APSIPA transactions on signal and information processing*, 8, 2019.

[172] X. Wang, R. Girshick, A. Gupta, and K. He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018.

[173] Z. Wang, J. Zhang, J. Feng, and Z. Chen. Knowledge graph embedding by translating on hyperplanes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 28, 2014.

[174] H. Wen, J. Shi, W. Chen, and Z. Liu. Transferring and generalizing deep-learning-based neural encoding models across subjects. *NeuroImage*, 176:152–163, 2018.

[175] H. Wen, J. Shi, Y. Zhang, K.-H. Lu, J. Cao, and Z. Liu. Neural encoding and decoding with deep learning for dynamic natural vision. *Cerebral Cortex*, 28(12):4136–4160, 2017.

[176] J. Weston, S. Bengio, and N. Usunier. Large scale image annotation: learning to rank with joint word-image embeddings. *Machine learning*, 81(1):21–35, 2010.

[177] F. Wilcoxon. Individual comparisons by ranking methods. In *Breakthroughs in statistics*, pages 196–202. Springer, 1992.

[178] F. R. Willett, D. T. Avansino, L. R. Hochberg, J. M. Henderson, and K. V. Shenoy. High-performance brain-to-text communication via handwriting. *Nature*, 593(7858):249–254, 2021.

[179] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

[180] Y. Xian, B. Schiele, and Z. Akata. Zero-Shot Learning – The Good, the Bad and the Ugly. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4582–4591, 2017.

[181] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015.

[182] B. Yang, W.-t. Yih, X. He, J. Gao, and L. Deng. Embedding entities and relations for learning and inference in knowledge bases. *arXiv preprint arXiv:1412.6575*, 2014.

[183] T. Young, D. Hazarika, S. Poria, and E. Cambria. Recent trends in deep learning based natural language processing. *ieee Computational intelligenCe magazine*, 13(3):55–75, 2018.

[184] H. Yu, H. Zhang, and W. Xu. Interactive grounded language acquisition and generalization in a 2d world. *arXiv preprint arXiv:1802.01433*, 2018.

[185] E. Zablocki, B. Piwowarski, L. Soulier, and P. Gallinari. Learning multi-modal word representation grounded in visual context. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[186] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

[187] C. Zhang, Z. Yang, X. He, and L. Deng. Multimodal intelligence: Representation learning, information fusion, and applications. *IEEE Journal of Selected Topics in Signal Processing*, 2020.

[188] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE transactions on image processing*, 26(7):3142–3155, 2017.

[189] Y. Zhang, G. Chen, H. Wen, K.-H. Lu, and Z. Liu. Musical imagery involves Wernicke's area in bilateral and anti-correlated network interactions in musicians. *Scientific reports*, 7(1):17066, 2017.

[190] Y. Zhang, M. Choi, K. Han, and Z. Liu. Explainable semantic space by grounding language to vision with cross-modal contrastive learning. In *Neural Information Processing Systems*. under review, 2021.

[191] Y. Zhang, K. Han, R. Worth, and Z. Liu. Connecting concepts in the brain by mapping cortical representations of semantic relations. *Nature communications*, 11(1):1–13, 2020.

[192] Y. Zhang, J.-H. Kim, D. Brang, and Z. Liu. Naturalistic stimuli: A paradigm for multi-scale functional characterization of the human brain. *Current Opinion in Biomedical Engineering*, page 100298, 2021.

[193] Y. Zhang, J.-H. Kim, H. Wen, and Z. Liu. High gamma electrocorticography in superior temporal gyrus represents words during natural speech. Suntec Singapore, June 2018. Organization for Human Brain Mapping. OHBM 2018 Merit Abstract.