Title
Modeling the Visual Word Form Area Using a Deep Convolutional Neural Network

Permalink
https://escholarship.org/uc/item/95t3w755

Author
Wiraatmadja, Sandy

Publication Date
2015

Peer reviewed|Thesis/dissertation
Modeling the Visual Word Form Area Using a Deep Convolutional Neural Network

A Thesis submitted in partial satisfaction of the requirements for the degree Master of Science

in

Computer Science

by

Sandy Wiraatmadja

Committee in charge:

Professor Garrison W. Cottrell, Chair
Professor Julian John McAuley
Professor Zhuowen Tu

2015
The Thesis of Sandy Wiraatmadja is approved and is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California, San Diego
2015
DEDICATION

I would like to dedicate this to my loving husband, Brian. Thank you for your encouragement, your love, and your support. You were always there when I needed you the most, lifted my spirit up when my sickness got the best of me, and accompanied me every step of the way through my treatments. I know that without you, I would not have made it this far. I love you.

I also want to dedicate this to my family. I become who I am today because of all of you. To my father and my mother, thank you for giving me the opportunity to pursue my education in the United States, and supporting me all the way. To my sisters, Sherley and Sheena, we might not always see eye to eye, but I know I can always depend on you, anytime anywhere anyplace.
# TABLE OF CONTENTS

Signature Page ................................................................. iii  
Dedication ................................................................. iv  
Table of Contents ............................................................... v  
List of Figures ................................................................. vi  
List of Tables ................................................................. viii  
Acknowledgements ........................................................... ix  
Abstract of the Thesis ..................................................... x  
Chapter 1  Introduction ...................................................... 1  
Chapter 2  Background ....................................................... 4  
  2.1 Visual Word Form Area (VWFA) ................................... 4  
  2.2 Convolutional Neural Network .................................... 12  
    2.2.1 The Winning Network of 2012 ILSVRC (AlexNet) .......... 13  
    2.2.2 LeNet-5 .......................................................... 14  
Chapter 3  Experimental Methods ....................................... 16  
  3.1 Visual Words Dataset .............................................. 16  
  3.2 Networks to Model the VWFA ...................................... 17  
    3.2.1 Modified AlexNet ............................................ 19  
    3.2.2 Modified LeNet-5 (VWFANet) .............................. 20  
  3.3 Training the Networks ............................................ 23  
    3.3.1 Training the Modified AlexNet ............................. 24  
    3.3.2 Training the VWFANet ...................................... 25  
  3.4 Selectivity to Real Words Experiments .......................... 25  
Chapter 4  Results and Discussion ..................................... 28  
  4.1 Network Performance .............................................. 28  
  4.2 Selectivity Analysis of VWFANet ................................. 30  
Chapter 5  Conclusion and Future Work ................................ 36  
Bibliography ...................................................................... 40
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Physical location of the VWFA in the left occipitotemporal sulcus bordering the fusiform gyrus (Dehaene and Cohen, 2011).</td>
</tr>
<tr>
<td>2.2</td>
<td>The VWFA activation of adult subjects with varying degrees of literacy when presented with different stimuli, plotted against the number of words that the subjects can read per minute (Dehaene and Cohen, 2011).</td>
</tr>
<tr>
<td>2.3</td>
<td>The anatomical and functional pathway model in visual word processing as proposed by McCandliss et al. (2003).</td>
</tr>
<tr>
<td>2.4</td>
<td>Plots of mean percent signal change in the fMRI-RA scans of the VWFA of participants in the experiments done by Glezer et al. (2009).</td>
</tr>
<tr>
<td>2.5</td>
<td>The AlexNet architecture, a Convolutional Neural Network that is trained to classify natural images from the ImageNet database Krizhevsky et al. (2012).</td>
</tr>
<tr>
<td>2.6</td>
<td>The LeNet-5 architecture, a Convolutional Neural Network that is trained for handwritten digits recognition LeCun et al. (1998).</td>
</tr>
<tr>
<td>3.1</td>
<td>Example of some visual words images in the dataset</td>
</tr>
<tr>
<td>3.2</td>
<td>Architecture of a modified AlexNet, with a replacement of the last layer into an 850-unit softmax layer, trained to classify visual words.</td>
</tr>
<tr>
<td>3.3</td>
<td>Architecture of a modified and shallower AlexNet, by using the features extracted in earlier convolutional layers and feeding them into an 850-unit softmax layer to classify visual words.</td>
</tr>
<tr>
<td>3.4</td>
<td>Architecture of the VWFANet, a modification of the LeNet-5 architecture, trained to classify visual words.</td>
</tr>
<tr>
<td>4.1</td>
<td>Validation set accuracy as the training progresses, up to the 100000-th iteration.</td>
</tr>
<tr>
<td>4.2</td>
<td>Test set accuracy of each word in the dataset, sorted by the accuracy.</td>
</tr>
<tr>
<td>4.3</td>
<td>Some examples of correct and incorrect predictions.</td>
</tr>
<tr>
<td>4.4</td>
<td>Euclidean distance of the activation in the output layer, which is a softmax with temperature parameter of 4.</td>
</tr>
</tbody>
</table>
Figure 4.5. The two activations in the output layer given a pair of real words: “arm”, “art”, and “key”. The prime word is “arm”, whose activation is shown in solid blue line. .............................. 33

Figure 4.6. The two activations in the output layer given a pair of pseudowords: “bot”, “sot”, and “lan”. The prime word is “bot”, whose activation is shown in solid blue line. .............................. 34
LIST OF TABLES

Table 2.1. Some examples of prime-target word pairs that were used in the experiments done by Glezer et al. (2009). ......................... 10

Table 3.1. A list of all font types that are used to generate the dataset. ....... 18

Table 4.1. Performance of the different networks. ......................... 29
ACKNOWLEDGEMENTS

I would like to acknowledge Professor Garrison W. Cottrell for his support as the chair of my committee. Through the ups and downs of my research, he always guides me and encourages me to go beyond what I imagine I could ever do.

I would also like to acknowledge members of the GURU lab, especially Yufei Wang and Sanjeev Jagannatha Rao, who have also provided help and suggestions along the way. Their supports have motivated me to keep trying, no matter how many failures I had to go through first.

Chapter 3, in part, is currently being prepared for submission for publication of the material. Wiraatmadja, Sandy; Cottrell, Garrison W.; Wang, Yufei. The thesis author was the primary investigator and author of this material.

Chapter 4, in part, is currently being prepared for submission for publication of the material. Wiraatmadja, Sandy; Cottrell, Garrison W.; Wang, Yufei. The thesis author was the primary investigator and author of this material.

Chapter 5, in part, is currently being prepared for submission for publication of the material. Wiraatmadja, Sandy; Cottrell, Garrison W.; Wang, Yufei. The thesis author was the primary investigator and author of this material.
ABSTRACT OF THE THESIS

Modeling the Visual Word Form Area Using a Deep Convolutional Neural Network

by

Sandy Wiraatmadja

Master of Science in Computer Science

University of California, San Diego, San Diego, 2015

Professor Garrison W. Cottrell, Chair

The visual word form area (VWFA) is a region of the cortex, located in the left fusiform gyrus, that appears to be a waystation in the reading pathway. The discovery of the VWFA occurred in the late twentieth century with the advancement in functional magnetic resonance imaging (fMRI). Since then, there has been an increased number of neuroimaging studies to understand the VWFA further for its properties. Because it is still relatively recent, there are disagreements in some properties of the VWFA. One such disagreement is regarding whether or not the VWFA is highly more selective for whole real words than pseudowords. A recent study provided evidences that the neurons
in the VWFA are tuned to be more selective to real words. This contradicts past studies which hypothesize that the VWFA is tuned to sublexical structure of visual words, and therefore has no preference for real words over pseudowords. The goal of this project is to develop a realistic model of the VWFA by training a deep convolutional neural network to map printed words to their labels. We then analyzed this network to see if we could observe the same selectivity the recent study found for whole real words. On the test set, the network that we trained from scratch is able to achieve an accuracy of 98.5%. Furthermore, we notice the same trends in our network, as in the results of the study, that show how the VWFA is highly selective for whole real words.
Chapter 1

Introduction

In recent years, with the advances in technologies, specifically the computing power of graphics processing units (GPUs), there has been an influx of deep Convolutional Neural Network (CNN) research. Deep neural networks have been shown to work really well in several image recognition tasks, such as object recognition (Krizhevsky et al., 2012), scene recognition (Zhou et al., 2014), and handwritten digits recognition (LeCun et al., 1998). Deep networks typically have many parameters which makes them hard to train, mostly due to the higher chance of overfitting, as well as the time it takes to fully train a network. However, there are techniques that can be done to reduce overfitting, such as using larger dataset by doing transformations on existing images to generate more data. Furthermore, with GPUs, training suddenly becomes much faster and more tractable.

Since the first success of GPU deep network training, many research projects attempt to implement novel ideas and algorithms to improve the performance of their architectures. The latest state of the art in image classification and detection is the GoogLeNet, a deep CNN which has 22 layers (Szegedy et al., 2014). They show that going deeper does not necessarily mean that the number of parameters has to increase as well, since they have 12x fewer parameters than the 8-layer network of Krizhevsky et al. (2012). This is important because more parameters can lead to overfitting problems,
while on the other hand, more layers are preferable because each layer can be thought of as representing a different set of features, with the features getting more abstract and higher level the deeper the layer. CNNs are chosen because they can model the human visual system quite well. Individual units are tiled together in such a way that they create overlapping regions that cover larger image field, similar to how the complex arrangements of cells work in the human visual cortex. Each cell is responsive to only a small region, but collectively they cover the entire visual field.

Computational models have become essential tools in Cognitive Science to provide ways to explain how humans or animals acquire and perform certain cognitive abilities, such as object categorization and reading, by simulating these cognitive processes and their behaviors under different stimuli. The model then provides a working hypothesis about how these processes are performed in the brain, especially if they are neurally plausible models. The model is explicit, compared to theories that are often just verbal descriptions of the process. Furthermore, computational models can also make novel predictions. Even though these models are designed to have certain properties that correspond to the hypothesis, there are times when they reveal other properties that can be investigated further to see if they are in accordance with the human cognitive process. This can lead to new hypotheses or offer new explanations to another previously observed phenomenon.

Cognitive modeling can also be used to model the different parts of the brain to deepen the understanding of how they work. This project aims to develop a deep CNN to realistically model the Visual Word Form Area (VWFA), a region of the cortex that appears to be a waystation in the reading pathway. We trained the network to map printed words to their labels. We then analyzed it to look for properties similar to the VWFA. We are especially interested to see if the model supports the neuroimaging evidence, as shown by Glezer et al. (2009), that suggests that the VWFA is highly sensitive to whole
real words. This is in contrast to previous studies which conclude that the VWFA is tuned to sublexical orthographic structure, and therefore has no preference for real words over pseudowords, which are legal letter strings that can be pronounced but they are not actual words with meanings.

This thesis is organized as follows. Chapter 2 provides some background information and previous studies that are related to the project. Chapter 3 describes the methods that were used in this project to train and analyze the networks that were used to model the VWFA. The results of the experiments are presented and discussed in Chapter 4. Then Chapter 5 gives a conclusion based on the results of this project, as well as any future work that can be explored following this project.
Chapter 2

Background

2.1 Visual Word Form Area (VWFA)

The VWFA is a region of the visual cortex that is activated during visual alphabetical word reading, similar to how the fusiform face area is responsive to faces. The idea of the existence of a specific region in the brain specialized for the reading process has been around since the nineteenth century, when a French neurologist, Joseph Jules Dejerine, reported a case of a patient with pure alexia due to a lesion in the brain in 1892. However, not until the late twentieth century, with advances in functional magnetic resonance imaging (fMRI), that the physical existence of the VWFA was discovered. Several brain imaging studies have been able to pinpoint this region to the same location within the left lateral occipitotemporal sulcus near the fusiform gyrus (Cohen et al., 2000; McCandliss et al., 2003; Vigneau et al., 2005; Dehaene and Cohen, 2011) which is shown in Figure 2.1. This area is found to be highly more responsive to visual words than any other similar stimuli, as demonstrated further through several lesion and interference studies. Lesions in the VWFA can cause pure alexia, where subjects experience severe visual reading impairment without any changes in ability to identify faces, objects, or even Arabic numerals, as well as ability to speak and understand words (McCandliss et al., 2003; Dehaene and Cohen, 2011). This is why several studies came to the same
conclusion, that the response of VWFA is strictly visual and prelexical, such that the words are recognized by VWFA visually, without giving them any meanings.

Considering the fact that written language was invented too recently to have an influence in the human evolution, one might wonder why the human brain is able to develop a specialized group of neurons that are activated during reading. Dehaene and Cohen (2007) suggested the neuronal recycling hypothesis, which states that some pre-existing cortical neurons are harnessed to be able to recognize visual words during learning that happens in the developmental stage. These neurons typically have similar functions and they are sufficiently plastic such that parts of them can be reoriented to the novel use of reading. This is suggested by the fMRI experiment done by Dehaene and Cohen (2011), studying the VWFA activation of subjects with different degrees of literacy (Figure 2.2). Figure 2.2a shows that the activation in VWFA increases proportionally to reading performance, when subjects are shown written sentences. The activation is also higher the more literate the subjects are. Furthermore, in Figure 2.2b, we can see how the activation for non-word stimuli decreases with reading performance. The VWFA is highly active on those stimuli in illiterate participants, but this activation decreases in

**Figure 2.1.** Physical location of the VWFA in the left occipitotemporal sulcus bordering the fusiform gyrus (Dehaene and Cohen, 2011).
(a) Written sentences.

(b) Other stimuli.

**Figure 2.2.** The VWFA activation of adult subjects with varying degrees of literacy when presented with different stimuli, plotted against the number of words that the subjects can read per minute (Dehaene and Cohen, 2011).

literate participants. These results support the neuronal recycling hypothesis since they suggest that the neurons in the VWFA become less responsive to non-word stimuli with proficiency in reading because they are recruited for reading.

Aside from the location and its prelexical nature that are largely accepted by now, there have been mixed results regarding other properties of the VWFA. One such property is the selectivity of the VWFA, whether it is tuned broadly to sublexical orthographic structure of the visual words, or whether it is actually tuned more tightly to whole real words. McCandliss et al. (2003) proposed a simple functional model of how words are processed in the visual pathway before it reaches the VWFA. This model can be seen in Figure 2.3. The word is first processed in ventral occipital regions V1 to V4, where the
neurons are tuned to features that are increasingly complex and abstract, running posterior to anterior along the visual pathway. The features progress from basic geometries like horizontal and vertical bars, to individual letters, to bigrams, and so on, until the sequence of the letters is identified. This hierarchical process in our visual system has been suggested by some neuroimaging studies (Dehaene et al., 2004; Vinckier et al., 2007), which lead to the findings that the VWFA has a role in whole word reading. However, experiments have failed to find any further evidence of this selectivity for whole words in the VWFA, leading to the hypothesis that the VWFA is tuned to sublexical structure of a word.

Studies supporting the VWFA broader tuning to sublexical structure

The following studies experimented on different properties of the VWFA, not specifically on the issue of whether or not the VWFA has preferences for real words over pseudowords. However, in the process, they failed to find evidence of a preference in the
VWFA for real words over pseudowords. This leads to them theorizing that the VWFA is broadly tuned to sublexical orthographic structure and that there is an insignificant difference between the activations under real words and pseudowords stimuli.

Dehaene et al. (2002) published his event-related fMRI study to test the hypotheses of prelexical representation and modality-specific character of the VWFA. The prelexical property refers to how the VWFA response to a visual word stimuli is purely visual without trying to give it any meanings. Meanwhile, the modality-specificity property refers to how the VWFA responds to a strictly visual process, rather than to auditory process. They measured the fMRI scans of participants who were given both visual and auditory stimuli. They were able to show that there is a significantly larger activation in the VWFA when subjects are presented written stimuli, whereas the auditory stimuli generate little to no activations. They also showed how the pseudowords stimuli, which have no meanings, cause a strong activation in the VWFA, which infers the prelexical nature of the VWFA. This hypothesis implies that the activity in the brain, to both words and pseudowords stimuli, should be similar, and that is exactly what their fMRI experiment results suggested.

An fMRI neuroimaging study was done by Vinckier et al. (2007) to test if a sublexical hierarchy is present in the VWFA region, with a posterior-to-anterior progression. Different words were presented to the participants under six different conditions: (1) false-font strings; (2) strings of infrequent letters; (3) strings of frequent letter but rare bigrams; (4) strings with frequent bigrams but rare quadrigrams; (5) strings with frequent quadrigrams; and (6) real words. They ran the experiment by taking the fMRI scans of the participants when given the different words, and calculating the percent activation to real words. They found that the neurons at the VWFA are more selective for higher-level stimuli, which matches the functional pathway shown in Figure 2.3. Furthermore, they observed that there is an insignificant difference between the activations in the VWFA
under real words and frequent quadrigrams stimuli, which mostly contain pseudowords. This supports the hypothesis that the VWFA is more tuned to orthographic regularity of letter strings, which makes it a stronger variable for this region than their lexical status.

**Studies supporting the VWFA high selectivity to real words**

Even though previous studies seem to show neuroimaging evidences against the VWFA neuronal tuning to whole words, some questions were raised regarding the accuracy of the methodology used to compare fMRI results in those studies. Most prior imaging studies, including Dehaene et al. (2002) and Vinckier et al. (2007), examined the average BOLD-contrast response in the VWFA when given a stimuli. However, it is harder to evaluate whether a large signal change is caused by a large number of neurons with small activity, or by a small number of neurons with large activity. This might cause the failure in obtaining evidence of VWFA preferences for real words. And therefore, inferring a certain property of the VWFA based on the average BOLD-contrast response becomes more complicated.

Glezer et al. (2009) conducted a neuroimaging study using a different technique called the fMRI rapid adaptation (fMRI-RA) (Grill-Spector et al., 2006). They were finally able to present evidences that the VWFA is more selective to whole real words. The fMRI-RA techniques measure the mean difference over a period of time between two BOLD-contrast responses when the subject is presented with a pair of stimuli that are shown sequentially in each trial. The mean percent signal change measurements reflect the similarity between the neuronal activation patterns that correspond to the two stimuli. If the mean percent signal change over time is low, that means there is a strong adaptation in the VWFA when given the pair of words, which means they activate the same groups of neurons. On the other hand, a high mean percent signal change indicate that the pair of words activate different groups of neurons instead.
Table 2.1. Some examples of prime-target word pairs that were used in the experiments done by Glezer et al. (2009).

<table>
<thead>
<tr>
<th>Prime Word</th>
<th>Target Word</th>
</tr>
</thead>
<tbody>
<tr>
<td>car</td>
<td>car</td>
</tr>
<tr>
<td>plane</td>
<td>plane</td>
</tr>
<tr>
<td>health</td>
<td>health</td>
</tr>
<tr>
<td>dut</td>
<td>dut</td>
</tr>
<tr>
<td>nount</td>
<td>nount</td>
</tr>
<tr>
<td>shatch</td>
<td>shatch</td>
</tr>
</tbody>
</table>

The subjects in the experiments were presented with prime/target pairs of real words and pseudowords, each with three different conditions: (1) “same”, where the prime and target words are the same, (2) “1L”, where the prime and target words differ by one letter, and (3) “different”, where the prime and target words share no common letters. Some examples of the pairs that were used in the experiments are shown in Table 2.1. They then divided the subjects into three experiment groups: the first group was shown only real word pairs, the second group was shown only pseudoword pairs, whereas the third group was shown both real word and pseudoword pairs. The results of the experiments, in terms of the mean percent signal change in the BOLD-contrast response in the VWFA, are shown in Figure 2.4. In the real word group, the difference in response for both “1L” and “different” conditions are insignificant, even though the pair of words in “1L” condition share some sublexical features. This suggests that the VWFA is tightly tuned to whole real words, supporting their hypothesis. On the other hand, in the pseudoword group, the result shows gradual increase in BOLD response with increasing dissimilarity. This suggests that the VWFA is not as tuned to pseudowords as it is to real words. Neurons in the VWFA are able to discriminate real words through substantial learning process, but they are not able to do the same thing with pseudowords.

Just recently, Glezer et al. (2015) went further to prove that learning new words
Figure 2.4. Plots of mean percent signal change in the fMRI-RA scans of the VWFA of participants in the experiments done by Glezer et al. (2009).
helps to tune the VWFA to be able to discriminate those new words better than when they are unlearned, which applies to pseudowords as well. This learning process effectively increases the size of the brain’s visual words dictionary. They implicitly trained some subjects on several pseudowords by exposing the words to them multiple times. In the end, they found similar results as in Figure 2.4a when the subjects were presented with prime/target pairs of previously trained pseudowords. This supports the hypothesis that the VWFA neurons are plastic enough that they can be refined through experience and learning.

2.2 Convolutional Neural Network

A Convolutional Neural Network (CNN) is a feed-forward artificial neural network. However, unlike the traditional neural network, a CNN uses convolution instead of general matrix multiplication. This means that it has sparse connectivity between the input and the output layer, resulting in many fewer parameters, making it easier to train, even though its performance might be slightly worse compared to standard feedforward neural networks with similarly-sized layers (Krizhevsky et al., 2012). CNNs are biologically-inspired variants of the Multilayer Perceptron (Bengio et al., 2015). There is a complex arrangement of cells in the visual cortex and they are sensitive to only a small region of the visual field, called a receptive field. These small regions are then tiled together to make up the entire visual field. This is exactly what a CNN model does, stacking the layers, leading to filters that are responsive to an increasingly larger region of pixel space. Therefore, it is only natural that we try to model the VWFA using a CNN. Specifically, we will use deep CNN that has been used successfully in several cognitive tasks. We discuss further the following two networks that are influential to this thesis.
2.2.1 The Winning Network of 2012 ILSVRC (AlexNet)

AlexNet is a large, deep CNN that was trained to classify 1.2 million high-resolution images in the ImageNet database into 1000 different classes (Krizhevsky et al., 2012). This network was the winner of the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) 2012 competition, achieving a top-5 test error rate of 15.3%. At the time, it was one of the first largest highly-optimized deep CNNs that was successfully trained within a reasonable time by using two GPUs, even though it still took 5-6 days to train. They were able to reduce the training time as well as the test error rate by introducing several novel and modified features, such as using Rectified Linear Units nonlinearity $f(x) = \max(0,x)$, instead of the usual hyperbolic nonlinearity function, $f(x) = \tanh(x)$. Since their implementation and their pre-trained weights are available publicly, it has since been used in many projects. The network consists of 5 convolutional layers and 3 fully-connected layers, as shown in Figure 2.5. The output of the second fully-connected layer is then fed into a 1000-way softmax to calculate the probability $P(class|input)$.

The input image is 224x224x3 pixels, extracted from the 256x256x3 images from ImageNet. It is filtered in the first convolutional layer with 96 kernels of size 11x11x3,
and with a stride of 4 pixels. Stride is the distance, in number of pixels, between each filter application. The second layer is a convolutional layer with 256 kernels, each of size 5x5x48. The outputs of the first two convolutional layers are normalized and pooled before being the input to the next layers. The third, fourth, and fifth convolutional layer consist of 384 kernels of size 3x3x256, 384 kernels of size 3x3x192, and 256 kernels of size 3x3x192, respectively. It is then followed by two fully-connected layers, which contain 4096 neurons each, and a softmax layer with 1000 units.

### 2.2.2 LeNet-5

LeNet-5 is a deep CNN that was developed by LeCun et al. (1998). The network was trained to classify handwritten digits recognition from the MNIST dataset, achieving a test-set accuracy of 99.05%. LeNet-5 comprises of 7 layers: 2 convolutional layers, each followed by a subsampling layer, 2 fully connected layers, and 1 output layer composed of Euclidean Radial Basis Function (RBF) units. The output layer has one unit for each of the 10 classes, corresponding to digits 0 to 9. The architecture is shown in Figure 2.6.

The input to the network is a 32x32 pixel image. The first convolutional layer C1 consists of 6 feature maps of size 28x28, where each unit is obtained from a 5x5
neighborhood in the input, with a stride of 1. The second layer is a subsampling layer \( S_2 \) with 6 feature maps of size 14x14. Each unit is connected to a 2x2 neighboring units from the C1 layer, where they are added and multiplied by a trainable coefficient, and then trainable bias is added to it in the end. This subsampling layer reduces the number of units by half from layer C1. The next convolutional layer C3 has 16 feature maps of size 10x10, by using a 5x5 neighborhood window from the feature maps of S2 with a stride of 1. The fourth layer is another subsampling layer \( S_4 \), with 16 feature maps of size 5x5. This is again a reduction by half in the number of units since each unit in S4 is connected to a 2x2 neighboring units from C3 layer. The next layer C5 is actually a convolutional layer with 120 feature maps where each unit is connected to a 5x5 neighboring units from the S4 layer. It just so happens that this results in the size of 1x1 for each feature map, which makes it the same as if it is a fully-connected layer instead. Layer F6 is a fully-connected layer of 84 units. The last layer is composed of 10 Euclidean RBF units, which can be interpreted as the unnormalized negative log-likelihood of a Gaussian distribution in the space of the 84-dimensional features of layer F6.

Unlike AlexNet, LeNet-5 is a much simpler network with only two convolutional layers. However, it is still able to generate features that can discriminate each digit almost perfectly, with test error rate of only 0.95%. On top of its simplicity and high accuracy, the inputs to LeNet-5 are images of handwritten digits, which consist of only black and white pixels, similar to visual word images. This is why we choose LeNet-5 as another inspiration for our model, as we believe that it will be able to generate features to discriminate visual words as well.
Chapter 3

Experimental Methods

As a first step in this research, we needed to develop a realistic model of the VWFA so that we could analyze it to look for properties that are similar to the VWFA. In order to do so, we built the dataset and trained a deep neural network that can recognize visual words by mapping them to their labels.

3.1 Visual Words Dataset

To minimize the scope of the project, we only use the list of 850 words of Basic English, that was created by Charles Kay Ogden. He claimed that these words are sufficient for ordinary communication in idiomatic English. The words are split into 3 different categories: (1) Operations, consisting of 100 words; (2) Things, consisting of 400 general words and 200 pictured words; and (3) Qualities, consisting of 100 general words and 50 opposite words. Using MATLAB, the words were printed in black onto a 227x227 blank white background, and saved in PNG format. MATLAB automatically saves each image with 3 channels, one for each colors: Red, Green, and Blue. Because the images are black and white, the three channels for each image are actually identical. However, since convolution is just a weighted sum of the input pixels, having three identical channels does not pose any issues, even though the two layers will not contribute any additional information. Therefore, for simplicity, we decided to leave
the three channels in the input images. To generate more images, we created the visual words with 75 different font types, as listed in Table 3.1, with three different sizes: 12, 15, and 18. We also rotated the text, with a rotation angle ranging from $-15^\circ$ to $15^\circ$, and translated the center of the text to be at least 75 and 100 pixels away from the top/bottom border and left/right border, respectively, maintaining enough space for longer words. We generated 1,296 images per word, totaling over 1.1 million images. Some sample images in the dataset can be found in Figure 3.1. The images were then divided randomly into 3 sets: the training set consisting of 850,000 images (1,000 images per word), the testing set consisting of 170,000 images (200 images per word), and the validation set consisting of 81,600 images (96 images per word).

3.2 Networks to Model the VWFA

In this section, we discuss the details of the networks that we trained to be used as models for the VWFA.
Table 3.1. A list of all font types that are used to generate the dataset.

<table>
<thead>
<tr>
<th>Aharoni</th>
<th>Gungsuh</th>
<th>MS Serif</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arial</td>
<td>Helvetica</td>
<td>Myanmar Text</td>
</tr>
<tr>
<td>Arial Black</td>
<td>Helvetica Narrow</td>
<td>New Century School Book</td>
</tr>
<tr>
<td>AR Julian</td>
<td>Impact</td>
<td>Nirmala UI</td>
</tr>
<tr>
<td>Avant Garde</td>
<td>Iskoola Pota</td>
<td>Palatino</td>
</tr>
<tr>
<td>Batang</td>
<td>Kartika</td>
<td>Roman</td>
</tr>
<tr>
<td>Batang Che</td>
<td>Khmer UI</td>
<td>Segoe UI</td>
</tr>
<tr>
<td>Bookman</td>
<td>Lao UI</td>
<td>Segoe UI Black</td>
</tr>
<tr>
<td>Calibri Light</td>
<td>Leelawadee</td>
<td>Segoe UI Emoji</td>
</tr>
<tr>
<td>Cambria</td>
<td>Leelawadee UI</td>
<td>Segoe UI Semibold</td>
</tr>
<tr>
<td>Cambria Math</td>
<td>Lucida Console</td>
<td>Segoe UI Symbol</td>
</tr>
<tr>
<td>Candara</td>
<td>Lucida Sans Unicode</td>
<td>Simplified Arabic</td>
</tr>
<tr>
<td>Comic Sans MS</td>
<td>Mangal</td>
<td>Sitka</td>
</tr>
<tr>
<td>Consolas</td>
<td>Meiryo</td>
<td>Tahoma</td>
</tr>
<tr>
<td>Constantia</td>
<td>Meiryo UI</td>
<td>Times</td>
</tr>
<tr>
<td>Courier</td>
<td>Microsoft Jheng Hei</td>
<td>Times New Roman</td>
</tr>
<tr>
<td>Courier New</td>
<td>Microsoft Jheng Hei UI</td>
<td>Traditional Arabic</td>
</tr>
<tr>
<td>Dok Champa</td>
<td>Microsoft New Tai Lue</td>
<td>Trebuchet MS</td>
</tr>
<tr>
<td>Dotum</td>
<td>Microsoft Sans Serif</td>
<td>Vani</td>
</tr>
<tr>
<td>Ebrima</td>
<td>Microsoft Tai Le</td>
<td>Verdana</td>
</tr>
<tr>
<td>Euphemia</td>
<td>Microsoft YaHei</td>
<td>Yu Gothic</td>
</tr>
<tr>
<td>Franklin Gothic Medium</td>
<td>Microsoft YaHei UI</td>
<td>Yumin</td>
</tr>
<tr>
<td>Gadugi</td>
<td>Mongolian Baiti</td>
<td>Yumin Demibold</td>
</tr>
<tr>
<td>Georgia</td>
<td>MSP Mincho</td>
<td>Yumin Light</td>
</tr>
<tr>
<td>Gisha</td>
<td>MS Sans Serif</td>
<td>Zapf Dingbats</td>
</tr>
</tbody>
</table>
3.2.1 Modified AlexNet

As a starting point, since AlexNet has been shown to generate generic features to represent visual images, we used it as a baseline network. Another motivation to use the AlexNet to see if it can accurately model the VWFA is related to the neuronal recycling hypothesis, as mentioned in Section 2.1. According to the hypothesis, the neurons in the VWFA are originally evolved to be selective for other stimuli, such as faces or tools, but they are harnessed for the task of reading. Therefore, we would like to see if the features generated by the AlexNet, trained for visual object recognition, are sufficiently discriminative for visual word images. To modify the network to solve the problem at hand, we changed the softmax layer to have 850 units, instead of 1000 units, since we classified the visual words into one of the 850 words in the dataset. The modified architecture can be seen in Figure 3.2. We used the pre-trained weights for the first 7 layers, without any fine-tuning, and we trained only the newly added layer.

As the layers go deeper, the higher-level features extracted tend to be more specialized to the training objective, which in the case of the AlexNet is to classify
Figure 3.3. Architecture of a modified and shallower AlexNet, by using the features extracted in earlier convolutional layers and feeding them into an 850-unit softmax layer to classify visual words.

objects in natural images. We wanted to see how well the full pre-trained network performs compared to the same pre-trained network but of a shallower depth. We believe that the low-level features generated by the earlier layers are more universal, similar to human early visual areas. An illustration of the shallower depth network is shown in Figure 3.3. The weights of the convolutional layers are pre-trained, without any fine-tuning. The output of the convolutional layer is then fed into an 850-unit softmax layer.

3.2.2 Modified LeNet-5 (VWFANet)

This network is inspired by the LeNet-5 architecture, as seen in Figure 2.6, with several refinements following ideas by Krizhevsky et al. (2012). We will refer to this network as the VWFANet for ease of reference. We chose LeNet-5 as the basis because both tasks involve visual character recognition. Furthermore, the network is much simpler compared to the AlexNet, and so it is more feasible to train from scratch, given our hardware limitations. The illustration of the VWFANet architecture can be found in Figure 3.4. The following subsections will describe each layer in more detail.
Figure 3.4. Architecture of the VWFANet, a modification of the LeNet-5 architecture, trained to classify visual words.

Image Layer

The input to the VWFANet is a 227x227x3 pixel image. Krizhevsky et al. (2012) utilized data augmentation to increase the number of training data, which is done by extracting patches out of the input image and mirroring it. Clearly, this is not a feasible option for our project. Therefore, in this image layer, we do not do any other transformation, besides scaling each pixel so that it is in the range [0,1].

Layer 1: Convolutional Layer

The first convolutional layer filters the input image with 20 kernels. Each kernel window is of size 5x5x3. Unlike LeNet-5, we take a stride of 2 to reduce the number of units in the output. This layer produces a feature map of dimension 112x112x20. As suggested in Krizhevsky et al. (2012), we used Rectified Linear Units (ReLU) as the activation function for this layer. This non-saturating nonlinearity $f(x) = \max(0, x)$, was demonstrated to learn faster than other saturating nonlinearities, such as the hyperbolic tangent function $f(x) = \tanh(x)$ or the sigmoid function $f(x) = (1 + e^{-x})^{-1}$. Using ReLU for this layer, which has a high dimension, produces sparse activations. These
sparse features have been shown to improve the network’s discriminative ability (Jarrett et al., 2009).

**Layer 2: Subsampling Layer**

We used max pooling on the first convolutional layer, reducing the dimensionality to 1/4 of its previous size. Each unit in the pooling layer is connected to a 2x2 neighboring units in the previous layer, and takes the largest value of the four. By choosing a stride of 2, adjacent pooling units do not overlap. This produces a feature map of dimension 56x56x20. We then applied Local Response Normalization to this output, which normalizes the activation over local regions. This scheme has heuristically been shown to aid generalization and make training faster. Each unit is divided by \((1 + \frac{\alpha}{n} \sum_i x_i^2)^\beta\), where \(x\) is the activation of the units, \(n\) is the size of each local region, and the sum is taken over the region that is centered at that unit (Jia et al., 2014). All constants, \(n, \alpha, \beta\), are hyper-parameters. We used the values \(n = 5, \alpha = 10^{-4}\), and \(\beta = 0.75\), which are the same values used by Krizhevsky et al. (2012).

**Layer 3: Convolutional Layer**

The third layer is another convolutional layer with 50 kernels, each of size 5x5x20, with a stride of 1. The output to this layer is 52x52x50. And similar to the first convolutional layer, we used ReLU as the activation function for this layer.

**Layer 4: Subsampling layer**

The fourth layer is another subsampling layer with the same parameters as the second layer. It involves a max-pooling with a 2x2 neighboring window with a stride of 2. Local Response Normalization is then applied to the output. This produces a feature map of dimension 26x26x50.
Layer 5: Fully Connected 1

The fifth layer is a fully-connected layer with 2048 units. The units in this layer compute a dot product between the weight vector and the input vector coming from the previous layer. It has a dense connection, which means that each unit is connected to all units in the previous layer. We also used ReLU as the activation function for this fully-connected layer to increase non-linearity.

Layer 6: Softmax Layer

We fed the output of the previous layer into an 850-way softmax. This produces a probability distribution over the 850 classes, $P(word\mid input)$.

3.3 Training the Networks

In order to train the networks as described in the previous section, we used Caffe, a framework designed for deep neural networks (Jia et al., 2014). Caffe provides the capability to easily train a deep network in either the CPU or the GPU. Caffe also provides some commonly used networks, including the pre-trained weights and configurations of AlexNet and LeNet-5. We trained the weights of the newly added layer in the modified AlexNet, as well as all the weights of the VWFANet from scratch, on a single NVIDIA GeForce GTX TITAN GPU which contains 2688 cores with 6GB of memory.

The networks were trained using the stochastic gradient descent method. At every iteration, the solver ran the network forward to compute the output and the loss, and then it did a back propagation to compute the gradients which would then be included in the weight updates. The objective function of this training is to minimize the cross-entropy classification loss of the softmax output. The update equation can then be summarized as
follows:

\[ V_{t+1} = \mu V_t - \alpha \nabla L(W_t) \]  \hspace{1cm} (3.1)

\[ W_{t+1} = W_t + V_{t+1} \]  \hspace{1cm} (3.2)

where \( V \) is the weight update value, \( W \) is the weights, and \( \nabla L(W) \) is the negative gradient (Jia et al., 2014). The subscript indicates the iteration number. The hyperparameters, \( \alpha \) and \( \mu \), required tuning to get the best results. The learning rate \( \alpha \) controls how much the negative gradient affects the new update value, while the momentum \( \mu \) is the weight of the previous update value. For this project, we decided to use the ideas that were presented by Krizhevsky et al. (2012) to choose the hyperparameter values. However, Bottou (2012) also had some good tricks that are useful in training with stochastic gradient descent method in his book. The following subsections give more training details for each network.

### 3.3.1 Training the Modified AlexNet

Since this network was previously trained by Krizhevsky et al. (2012), naturally we chose values that closely followed their strategy. The weights of the last layer were trained with the stochastic gradient descent method with a batch size of 256 examples, base learning rate of 0.001, momentum of 0.9, and weight decay of 0.0005. We also decreased the learning rate by a factor of 10 after every 2000 iterations. No fine-tuning was done in the training, therefore we had to make sure that the learning rate of all the other layers were set to 0, so that no weight update could happen. On the other hand, in the last layer, we started with a learning rate multiplier of 10, where this value was multiplied to the base learning rate. We started with a higher value because we wanted learning to happen in bigger steps in the beginning to reach convergence faster.
The weights were initialized from a Gaussian distribution with zero mean and standard deviation of 0.01, whereas the bias was initialized to 0. We trained the network for a maximum of 100,000 iterations, while checking the validation set accuracy after every 5,000 iterations so we could stop training once the accuracy started to go down.

### 3.3.2 Training the VWFANet

Similar to how the modified AlexNet was trained, all the weights for the VWFANet were also trained with base learning rate of 0.01, momentum of 0.9, and weight decay of 0.0005. We reduced the batch size to be 32 examples to reduce the training time. We found that increasing the batch size does not affect the accuracy significantly, but it takes a much longer time to run per iteration since it has to train all layers, unlike during the AlexNet training. The learning rate was adjusted using the inverse policy, where \( \lambda_n = \lambda_0 \times (1 + 0.0001 \times n)^{-3/4} \), where \( n \) is the current iteration number. The weights in each layer were initialized using the xavier algorithm, which uses the number of input and output neurons to determine the scale of initialization (Jia et al., 2014). The bias in each layer was initialized to 0. We trained the network for a maximum of 100,000 iterations, while checking the validation set accuracy after every 5,000 iterations so we could stop training once the accuracy started to go down.

### 3.4 Selectivity to Real Words Experiments

After we trained all the networks described above, we chose the one that was closest to modeling the VWFA by selecting the one with the highest classification accuracy. We used that model and analyzed it to see if it has a similar selectivity for whole real words as the VWFA, which was suggested by neuroimaging studies by Glezer et al. (2009). However, because we only trained one deep neural network model for the VWFA, we could only analyze the same trained model for both real words input pairs
and pseudoword input pairs. Therefore, we compared the analysis of the highest accuracy model to only the results of the third experiment group shown in Figure 2.4c, since it does a complete within-subject analysis when given both real words and pseudowords stimuli.

The main question now is how to measure the difference in activation so that we can compare it to the results in Figure 2.4c. We decided to measure the Euclidean distance of the activations in the last layer, which is the softmax layer. This is because each unit in this layer can be thought of as representing either a single neuron in the VWFA or a group of neurons that work in unison. In his book, Hebb (1949) proposed an idea that is now known as the Hebb’s rule, which states that: “When an axon of cell A is near enough to excite a cell B and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A’s efficiency, as one of the cells firing B, is increased.” This rule can be summarized as saying that cells that fire together, wire together. Therefore, our simplification of grouping neurons into one unit can be justified in the sense of treating each unit as a Hebbian cell assembly. Furthermore, we chose the last layer because by choosing this network to model the VWFA, we make a claim that the last layer of this model should reflect the VWFA, and that the units in the last layer represent the neurons in the VWFA. Meanwhile, the other layers correspond to regions in the visual pathway leading up to the VWFA.

The softmax layer imposes extreme values on the probabilities of outputs, that are likely to be more differentiated than actual neural activities. Hence we “soften” the output activity with a temperature parameter $T$ on the softmax, as follows:

$$P(word_i | input) = \frac{\exp^{x_i/T}}{\sum_{j=0}^{n-1} \exp^{x_j/T}}$$  \hspace{1cm} (3.3)
where $x$ is the input to the softmax layer. A high temperature parameter will distribute the probability more evenly, such that for $T \to \infty$, all words will have exactly the same probability. On the other hand, a low temperature will distribute the probability to only the highest value. We chose a temperature of $T = 4$, which creates a smoother probability distribution, such that there are a few labels that have non-zero probabilities, without losing the actual label information. Here, we chose this number arbitrarily, but it is a parameter that could be fit to the data.

We obtained the list of real words and pseudowords that Glezer et al. (2009) used in their study. We re-trained the network and added the study words to the original list of 850 Basic English words. We then created four visual images for each real word and pseudoword in the experiment. Similar to their experiments, we ran the network to classify a prime-target pair of visual words and measured the difference in activation between the two outputs. For each pair of words, we took the average of all possible input image pairs that we created out of the four generated images per word.

Chapter 3, in part, is currently being prepared for submission for publication of the material. Wiraatmadja, Sandy; Cottrell, Garrison W.; Wang, Yufei. The thesis author was the primary investigator and author of this material.
Chapter 4

Results and Discussion

In this chapter, we present the results of the different networks. After discussing these results, we analyze the highest accuracy network to look for properties similar to the VWFA.

4.1 Network Performance

The accuracy comparison between the networks can be seen in Table 4.1. Unlike the results from many other studies (Karayev et al., 2013; Zhou et al., 2014; Wang and Cottrell, 2015), we have found a problem for which AlexNet is not well-suited. Using the penultimate layer of features gives us very poor test set performance of 22.8%. While this is well above chance, it is not accurate enough for our purposes. We hypothesize that this is because the highest level features of AlexNet are highly specialized for object recognition. Objects have features that are very different from words. In particular, recognizing words is a fine-level discrimination task; similar-looking inputs must be discriminated from one another.

Hence we expect that the lower level-features in AlexNet, which resemble the Gabor-like receptive fields of V1, may perform better. This is indeed the case - there is a 5% gain using the lowest-level features, although this is still quite low. The remaining convolutional layers, however, give nearly equivalent performance of around 70%. The
best performance is obtained using the second level of convolutional features, which give about 72\% accuracy. Even though the features can identify low level representations of the visual words, the number of features in the second layer might raise some issues. The second convolutional layer consists of 256 feature maps, each of size 13x13, which sums up to 43264 features. Therefore, the chance of overfitting becomes higher. Furthermore, as mentioned in the previous chapter, these accuracy values are obtained from the pre-trained weight of the network, without any fine-tuning. While fine-tuning with heavy regularization may give better results, we chose to start fresh with a smaller network trained from scratch.

The accuracy obtained from the VWFA Net shows some promising results. The network is able to achieve test set accuracy percentage of $98.5\%$. Therefore, we use this model for further analysis. We trained the network for 100000 iterations, while checking for validation set accuracy after every 5,000 iterations, making sure we did not overfit the training data. Figure 4.1 shows the validation accuracy as the training progresses. We can see that the network is able to reach a high accuracy after 20,000 iterations, and climbs more slowly from there. Remarkably, the network never appears to overfit; accuracy on the validation set is monotonically increasing for another 80,000 iterations.

The accuracy of each word is shown in Figure 4.2, sorted by the accuracy values, and some samples of correct and incorrect predictions by the VWFA Net can be seen in

<table>
<thead>
<tr>
<th>Network</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>AlexNet (full network)</td>
<td>22.8%</td>
</tr>
<tr>
<td>AlexNet (up to first convolutional layer)</td>
<td>27.4%</td>
</tr>
<tr>
<td>AlexNet (up to second convolutional layer)</td>
<td>71.9%</td>
</tr>
<tr>
<td>AlexNet (up to third convolutional layer)</td>
<td>69.9%</td>
</tr>
<tr>
<td>AlexNet (up to fourth convolutional layer)</td>
<td>70.5%</td>
</tr>
<tr>
<td>AlexNet (up to fifth convolutional layer)</td>
<td>69.7%</td>
</tr>
<tr>
<td><strong>VWFANet</strong></td>
<td><strong>98.5%</strong></td>
</tr>
</tbody>
</table>
Figure 4.1. Validation set accuracy as the training progresses, up to the 100000-th iteration.

Figure 4.3. We can see that even though the VWFANet overall accuracy in visual words classification task is high, there are a few outliers with less than 60% accuracy: “tall” with an accuracy of 38%, and “hour” with an accuracy of 52.5%, respectively. However, as can be seen in Figure 4.3, the network makes reasonable mistakes. Out of the 200 images in the test set for the word “tall”, 124 of them are predicted incorrectly. And 113 out of those 124 images are predicted as “tail” which is very similar to “tall”. Similarly for “hour”, 89 out of the 95 incorrectly predicted images are predicted to be “how” which still has high resemblance to “hour”.

4.2 Selectivity Analysis of VWFANet

Figure 4.4 shows the result of calculating the euclidean distance between the activation in the last layer of the two input images. Even though we cannot compare it directly to the results of Glezer et al. (2009) study since the measurements are not the
Figure 4.2. Test set accuracy of each word in the dataset, sorted by the accuracy.

Figure 4.3. Some examples of correct and incorrect predictions.
**Figure 4.4.** Euclidean distance of the activation in the output layer, which is a softmax with temperature parameter of 4. The blue bar represents the distance in activation of two instances of the same word, while the orange bar represents the distance between the activations of two words differing by one letter. The gray bar represents the distance between two different words. Error bars represent the standard error of the mean. In each case, these are averages over the stimuli used in the human subject experiments.

same, we can still observe a similar trend to the findings shown in Figure 2.4c. This shows that the VWFANet supports the hypothesis made in the study, where the VWFA shows high selectivity for whole real words, compared to pseudowords.

For the real words experiment, there is a small Euclidean distance between the activations in the last layer of two different images from the “same” condition. This is expected because even though the stimuli are different versions of the same word, the activation of the one unit corresponding to the word is expected to be high for both stimuli, which in turn decreases the Eucliden distance. For the “1L” and “different” conditions, we observe a bigger Euclidean distance compared to the one in “same”, but the difference between the two conditions are not significant. This indicates that in each condition, the prime and target words activate disjoint groups of neurons, or different units in our model’s output layer. This shows that even when the two stimuli differ only
Figure 4.5. The two activations in the output layer given a pair of real words: “arm”, “art”, and “key”. The prime word is “arm”, whose activation is shown in solid blue line. By one letter, the model is still able to discriminate them and so they activate different neurons, similar to when the two inputs share no common letters. A sample of the activations in the output layer for a given pair of real words is shown in Figure 4.5.

For the pseudowords experiment, we again notice that the smallest Euclidean distance happens when the two stimuli are images of the same pseudoword, even though the model has not been trained on these. However, what is different from the real words experiment is that we observe a gradual increase in the Euclidean distance from the “same” condition to “1L” to “different”. This is also sensible because the network is trained for real words. Therefore, when the model is given a pseudoword, it activates partial representations of many different words. Figure 4.6 shows that the neurons from the softmax layer that get activated on pseudowords are more distributed, but there is more overlap when the stimuli are similar, compared to when the stimuli are completely different.

One might object that we are measuring very different things than the BOLD signal response change between two stimuli, the measurement made by Glezer et al. (2009). This is a time-series measurement that takes the mean of how much the activation signal changes over a period of time. On the other hand, the VWFANet does not provide
any temporal data. This is why deciding which measurement to use to compare our network to the neuroimaging results is not a trivial problem. All we have from our network is the activation data at each layer. However, we observe that the mean percent signal change is primarily used as an indicator of how much the two different stimuli activate overlapping regions of the VWFA. If the percent signal change is low, that means the BOLD-contrast signal change over the time period is small which suggests that the two individual stimuli activate similar groups of neurons. Meanwhile, a high percent signal change indicates that there is a big change in the BOLD-contrast response between the first and the second stimuli, which means that the two stimuli activate disjoint groups of neurons. Therefore, given the two outputs from the last layer of the VWFANet, we also want to see whether they correspond to similar or disjoint groups of neurons. Going back to our assumption, based on Hebb’s rule, that each unit in the softmax layer actually corresponds to a group of neurons, we can simplify this problem by checking for similarity between the two activation vectors. Euclidean distance is one way to measure similarity of two vectors. This is why we argue that it is reasonable to measure the Euclidean distance between the two activation vectors in the last layer for this experiment.
The decision to select the last layer to be measured is also not trivial. We choose to measure the activations of the last layer, because as mentioned in Section 3.4, we hypothesize that the units in the softmax layer of our model correspond to the neurons in the VWFA because these model neurons represent the recognition of each individual word, without any semantics, the hypothesized function of the VWFA. That is why it is reasonable to use the last layer to measure the similarity of the VWFA neuronal activation between two input stimuli. However, seeing that this is the classification layer, the distinct results for the real word experiment in “1L” and “different” condition are a completely expected result since we train the model with the objective function of minimizing the classification error. This is the main motivation for using the temperature parameter \( T = 4 \) to make sure that for each input word, there will be multiple units that get moderately activated in the softmax layer, instead of having only one highly activated unit. Therefore, other units, besides the unit corresponding to the true label of the input word, have a non-zero probability of being activated. If the VWFANet is tuned to sublexical orthographic structure, then the other units that get activated by the stimuli under “1L” condition should have more overlap and in turn, smaller Euclidean distance between the two softmax activation vectors. On the other hand, again assuming sublexical representations, the stimuli in the “different” condition should have less overlap and in turn, greater Euclidean distance. However, this is not what the data in Figure 4.4 shows us. For real words, the average Euclidean distance of the pair of activation vectors for both “1L” and “different” condition are almost identical. For further evidence, we can see from Figure 4.5b, the two words “arm” and “art” activate an almost completely disjoint set of units, even though their sublexical structures should be similar.

Chapter 4, in part, is currently being prepared for submission for publication of the material. Wiraatmadja, Sandy; Cottrell, Garrison W.; Wang, Yufei. The thesis author was the primary investigator and author of this material.
Chapter 5

Conclusion and Future Work

From the first part of our experiment, where we train several networks trying to find the one that can closely model the VWFA, we are able to learn several things. While the pre-trained full AlexNet is able to learn high-level features that are useful for many tasks, those tasks nearly universally involve objects, or people, or visual styles of images. However, in the case of visual words, the pre-trained AlexNet does not perform as well in identifying features that can differentiate one visual word to the other. On the other hand, the lower-level features generated from earlier layers of AlexNet are relatively universal, similar to the primate visual pathway, where the earlier visual areas compute basic features. This makes it clear that we need a separate network for words, and that is why we decided to train the VWFANet from scratch, by modifying the LeNet-5 architecture. This network is relatively shallow, compared to other state of the art object classifier networks. However, the two convolutional layers and two fully connected layers are able to extract important features from the visual words to be able to map them to the right labels. Even though there are some outliers, when we investigate further, we find out that their mistakes are reasonable, considering that different font types and sizes might make certain letters look more similar to others, such as i and l.

Now that we have a model for the VWFA, we can analyze it to test if it matches some hypotheses of the VWFA that have been previously suggested by neuroimaging
studies. One such hypothesis is the hierarchical sublexical structure of visual representation of the words in the visual pathway leading up to VWFA. This can be done by using a deconvolutional network to visualize the features at each layer. It will be interesting to see the progression of the features in later layers as well. However, we only have two convolutional layers in the VWFANet. Even though we already achieve such a high accuracy with this network, we should try adding more layers to see if the deeper model can generate better sets of features and possibly achieve a higher accuracy with fewer outliers. We have to make sure that the deeper network does not increase the number of parameters proportionally to avoid other problems, such as over-fitting. Once we have a deeper network, we should compare the features and see why our VWFANet performs well enough with only 2 convolutional layers. One possible reason why a shallower network works in this task is the simplicity of the input images. When we compare the visual word to natural images, such as the ImageNet images used in object classification, we realize that the visual words are much simpler than the natural images, since they have no colors and no background scenes. This is most likely why we can use fewer layers to represent the visual features.

With the VWFANet, we are also able to reproduce the same trends that Glezer et al. (2009) observed in their human neuroimaging study, which suggests that the VWFA is tuned to real words such that it has high selectivity for real words compared to pseudowords. Some questions might arise regarding which activation values we should measure the distance from. The output layer, which is the classification layer in our model, matches the real word data for unsurprising reasons, since the training was designed to classify the different inputs into different classes and minimize the classification error. One approach that we wish to explore is to use distributed representations, instead of the current localist representations where we dedicate one unit in the softmax layer to one word. With distributed representations, each neuron can be assigned to multiple words,
while similarly each word can activate multiple neurons. By doing this, we are no longer making this a classification task, and therefore, each unit in the output layer no longer corresponds to an individual word.

Taking the idea from a study done by Dehaene et al. (2004), we can further expand the experiments on this subject of the selectivity of the VWFA neuronal tuning to real words. In Glezer et al. (2009, 2015) studies, they compared three different conditions: “same”, “1L”, and “different”. Meanwhile, Dehaene et al. conducted a similar experiment of measuring priming to study if the VWFA has case-invariant and location-invariant representation of words. In their experiments, they also have 3 different conditions, where 2 of them are “same” and “different”. However, one condition that they use in place of the “1L” is “circular anagrams”, where pairs of words can transform into one another simply by moving a single letter from the front to the back, and vice versa. An example of a pair of words that are circularly anagrams is the French word “reflet” and “trefle”. If we can find some pairs of English words that are circularly anagrams, we can add this to our experiments. This goes a step further than “1L” because the target word is made up of exactly the same letters as the prime word, but with one location shift, instead of one letter change. If we still observe the same high percent signal change in the fMRI data, or high Euclidean distance in the VWFANet, then we provide an even stronger support to the hypothesis that the neurons in the VWFA are highly selective to whole words, instead of a broader tuning to sublexical orthographic structure of the words.

In conclusion, we have successfully trained a convolutional neural network to model the VWFA by trying to map the words to their labels, achieving an accuracy of 98.5% on the test set. This network might not be as deep as the latest state of the art for object recognition, but it seems to perform well for the task of modeling the VWFA. The VWFANet model is able to support the recent fMRI-RA evidence that there is a preference for real words over pseudowords in the VWFA. This VWFANet model may
open up more opportunities to study the VWFA and its properties further, especially with the recent increase in the neuroimaging studies for the VWFA.

Chapter 5, in part, is currently being prepared for submission for publication of the material. Wiraatmadja, Sandy; Cottrell, Garrison W.; Wang, Yufe. The thesis author was the primary investigator and author of this material.
Bibliography


