

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

Sandy Wiraatmadja (csandyw@gmail.com)

Department of Computer Science and Engineering, University of California San Diego
9500 Gilman Dr 0404, La Jolla, CA 92093 USA

Garrison W. Cottrell (gary@ucsd.edu)

Department of Computer Science and Engineering, University of California San Diego
9500 Gilman Dr 0404, La Jolla, CA 92093 USA

Abstract

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 of functional magnetic resonance imaging (fMRI). Since then, there has been an increasing number of neuroimaging studies to understand the VWFA, and there are disagreements as to its properties. One such disagreement is regarding whether or not the VWFA is more selective for real words over pseudowords¹. A recent study using fMRI adaptation (Glezer, et al., 2009) provided evidence that neurons in the VWFA are selectively tuned to real words. This contradicts the hypothesis that the VWFA is tuned to the sublexical structure of visual words, and therefore has no preference for real words over pseudowords. In this paper, we develop a realistic model of the VWFA by training a deep convolutional neural network to map printed words to their labels. The network is able to achieve an accuracy of 98.5% on the test set. We then analyze this network to see if it can account for the data Glezer et al. found for words and pseudowords, and find that it does.

Introduction

The VWFA is a region of the visual cortex that is activated during visual alphabetical word reading, similar to how the fusiform face area (FFA) 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, who in 1892 reported a case of a patient with pure alexia due to a brain lesion. However, it was 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, Cohen, & Dehaene, 2003; Vigneau, Jobard, Mazoyer, & Tzourio-Mazoyer, 2005; Dehaene & Cohen, 2011), shown in Figure 1. This area is found to be more responsive to visual words than any other similar stimuli, as corroborated by 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 (McCandliss et al., 2003; Dehaene & Cohen, 2011). The

¹Pseudowords are usually constructed from real words by changing a consonant, and hence follow the orthographic rules of English.

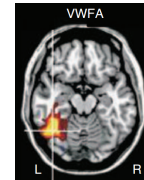


Figure 1: Activation of the VWFA in the left occipitotemporal sulcus near the fusiform gyrus (Dehaene & Cohen, 2011).

conclusion of this research is that the response of VWFA is strictly visual and prelexical, such that the words are recognized by VWFA visually, without giving them any meaning.

In this paper we develop a deep CNN model of the Visual Word Form Area (VWFA). 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 is consistent with the neuroimaging evidence reported by Glezer, Jiang, and Riesenhuber (2009), that suggests that the VWFA represents words discretely, and pseudowords in a more distributed manner. This is in contrast to previous studies which concluded that the VWFA is tuned to sublexical orthographic structure, and therefore has no preference for real words over pseudowords.

An influential descriptive model of the VWFA was proposed by McCandliss et al. (2003) (Figure 2). 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 horizontal and vertical bars, to individual letters, to bigrams, and so on, until the sequence of the letters is identified. This model is consistent with some neuroimaging studies (Dehaene et al., 2004; Vinckier et al., 2007). However, experiments have failed to find any further evidence of selectivity for whole words in the VWFA, leading to the hypothesis that the VWFA is tuned to the sublexical structure of a word.

We do not have space here to review the many studies of the VWFA. The reader is referred to (Dehaene & Cohen, 2011) for a relatively recent review. Here we focus on a study by Glezer et al. (2009), who used fMRI rapid adaptation (fMRI-RA) (Grill-Spector, Henson, & Martin, 2006) to investigate the nature of representations in the VWFA. fMRI-RA makes use of the fact that repeated stimuli lead to a reduction in the overall BOLD response. If a new stimulus leads to a resur-

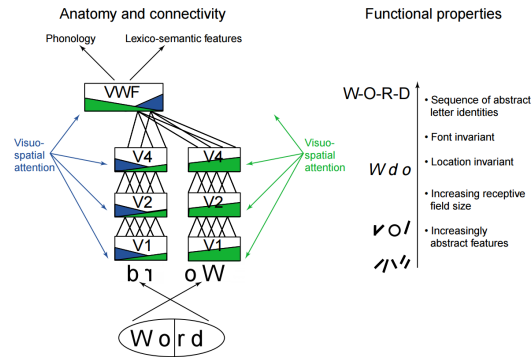


Figure 2: The anatomical and functional pathway model of visual word processing proposed by McCandliss et al. (2003).

Table 1: Some examples of prime-target word pairs that were used in the experiments done by Glezer et al. (2009).

Prime	Target Word			
	same	1L	different	
car	car	bar	lie	real words
plane	plane	place	cross	real words
health	health	wealth	ground	real words
dut	dut	dun	pim	pseudowords
nount	nount	naunt	spibe	pseudowords
shatch	shatch	chatch	joune	pseudowords

gence of activation, that is taken as evidence that different neurons represent that stimulus, i.e., the brain sees it as different from the prime stimulus.

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. Examples of pairs that were used in the experiments are shown in Table 1. Of greatest interest here is their third experiment, which presented these stimuli in a within-subject design. The results (consistent with the first two experiments) showed that single letter changes gave rise to recovery of the BOLD signal that was equal to the recovery for completely different words, suggesting that even words with one letter different did not share representations. On the other hand, responses to pseudoword changes were graded, suggesting that their representations were overlapping.

Experimental Methods

As a first step in this research, we needed to develop a realistic model of the VWFA that we could analyze to look for properties that are similar to the VWFA. In order to do so, we designed a convolutional neural network for this task, created a large dataset of word images, and trained the network to recognize visual words by mapping them to their labels. Here we describe the dataset, the network, our training procedures,

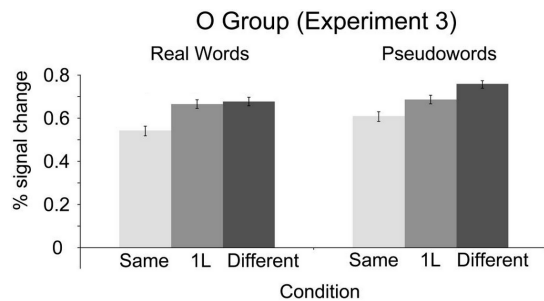


Figure 3: Mean percent signal change in the VWFA of participants in Glezer et al.’s third experiment.

and finally our experimental design to test the network on the stimuli used in Glezer et al. (2009).

Visual Words Dataset

To minimize the scope of the project, we used the 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. In order to model Glazer et al., we added 47 sets of three words each that they used for their stimuli. Since some of these words were already in the Basic English set, this resulted in 899 unique words.

Using MATLAB, the words were printed in black onto a 227x227 blank white background, and saved in PNG format. To generate an adequate variety of word images, we used 75 different font types, with three different sizes: 12, 15, and 18 pt. We also rotated the words slightly, with a rotation angle ranging from -15° to 15° , 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. In this manner, we generated 1,296 images per word, totaling over 1.1 million images. Some sample images in the dataset are shown in Figure 4. The images were then divided randomly into 3 sets: a training set of 899,000 images (1,000 images per word), a test set consisting of 170,000 images (200 images per word, plus four images each of the Glazer et al. words for simulating their experiments), and a validation set consisting of 86,304 images (96 images per word).

VWFA network model

Consistent with recent work in computer vision, we used a convolutional neural network (CNN) for the VWFA model. A CNN is a feed-forward neural network that uses the convolution of repeated features with small receptive fields across the image to create several feature maps. After a nonlinearity is applied, these maps are then downsampled by computing the max of a small patch of them, and the process is repeated. In this way, the receptive fields of the units get larger

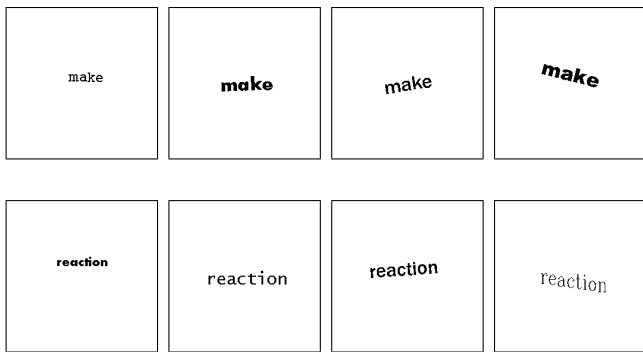


Figure 4: Example of some visual words images in the dataset

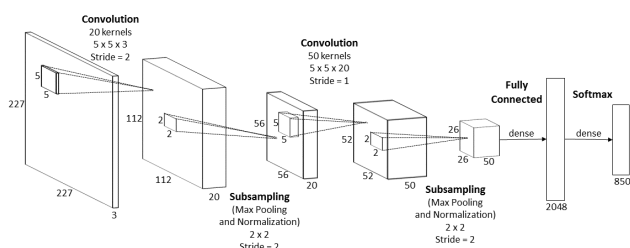


Figure 5: Architecture of the VWFANet, a modification of the LeNet-5 architecture, trained to classify visual words.

in deeper layers of the network (Krizhevsky, Sutskever, & Hinton, 2012), just as they do in visual cortex. CNNs are therefore biologically-inspired variants of multilayer perceptrons (Bengio, Goodfellow, & Courville, 2015).

A common approach in computer vision is to start with a pre-trained network, such as “Alexnet,” the network that won the 2012 Imagenet Large Scale Visual Recognition Challenge (ILSVRC) (Krizhevsky et al., 2012). However, for the task of recognizing images of words, Alexnet did not work well at all, presumably because it is tuned to objects, not words, which require fine-grained discrimination.

Hence we designed a modified version of LeCun’s zip code reading network, LeNet-5 (LeCun, Bottou, Bengio, & Haffner, 1998). We used several refinements following ideas by Krizhevsky et al. (2012). We chose LeNet-5 as the basis for our VWFA network because both tasks involve visual character recognition. An illustration of the VWFANet architecture is shown in Figure 5.

The input to the VWFANet is a 227x227 pixel image, scaled so each pixel is in the range [0,1]. The first convolutional layer filters the input image with 20 kernels of size 5x5, and uses a stride of 2. The stride is the offset of each kernel with respect to its neighbor, so here the kernels are overlapping. This results in a feature map of dimension 112x112x20. We used Rectified Linear Units (ReLU) as the nonlinear activation function for this layer. The ReLU function is defined as $f(x) = \max(0, x)$. It has been shown empirically to learn faster than 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, Kavukcuoglu, Ranzato, & LeCun, 2009).

The next step is pooling the responses of the first convolutional layer. We used max pooling on each 2x2 patch (i.e., the output of this operation is the maximum response of the four units), reducing the dimensionality to 1/4 of its previous size. Using 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 $(1 + (\alpha/n) \sum_i x_i^2)^\beta$, 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, α, β , 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).

The third layer is another convolutional layer with 50 kernels, each of size 5x5x20, with a stride of 1. The output of this layer is a 52x52x50 feature map, to which we apply the ReLU activation function. We then use max pooling on this layer with the same parameters as before, and again use response normalization. This produces a feature map of dimension 26x26x50.

At this point, we use a fully-connected (non-convolutional) hidden layer of 2048 units, each with 26x26x50 inputs from the previous layer. Again we used ReLU as the activation function. Finally, the output of this layer is connected to an 899-way softmax. This produces a probability distribution over the 899 classes (850 Basic English words plus 49 extra words used in Glezer et al.), $P(\text{word}|\text{input})$.

Training the Network

In order to train the network, we used Caffe, a framework designed for deep neural networks (Jia et al., 2014). We trained the weights of the VWFANet from scratch, on a single NVIDIA GeForce GTX TITAN GPU which contains 2688 cores with 6GB of memory.

The network was trained using cross-entropy loss, the minibatch stochastic gradient descent method, with momentum and weight decay. The minibatch method computes the error over a small “batch” of training examples, and then changes the weights. In our case, we used a minibatch size of 32. We used an initial learning rate of 0.01, momentum of 0.9, and weight decay of 0.0005. The learning rate was adjusted using the inverse policy, where $\lambda_n = \lambda_0 \times (1 + 0.0001 \times n)^{-3/4}$, where λ_0 is the initial learning rate and n is the current iteration number. The weights initialized to have 0 mean and standard deviation equal to $1/\sqrt{m}$, where m is the fan-in to the unit. 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. However, even after 100,000 iterations, the error was still dropping on the validation set.

Selectivity to Real Words Experiments

We analyzed the VWFANet to see if it has a similar selectivity for real words as the VWFA, as suggested by Glezer et al. (2009). We compared the analysis of the highest accuracy model to the results of the within-subject experimental results shown in Figure 3.

The main question now is how to model the release from adaptation effect. We assume that this effect is proportional to the difference in activation between stimuli. Hence we measure the Euclidean distance between stimuli as a proxy for signal change - a small distance means most neural activities are shared, and so there should be no or a small rise in the BOLD response, whereas a large distance means that the two stimuli do not share neural activations, so there should be a large change in the BOLD response. We measure this distance in the softmax layer. This is because each unit in this layer can be thought of as representing either a single neuron responding to a word, or a group of neurons (a Hebbian cell assembly) that work in unison.

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:

$$P(\text{word}_i | \text{input}) = \frac{\exp(x_i/T)}{\sum_{j=1}^n \exp(x_j/T)} \quad (1)$$

where x is the input to the softmax layer. A high temperature parameter will distribute the probability evenly, such that for $T \rightarrow \infty$, all words will have probability $1/n$. 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. We chose this number arbitrarily, but it is a parameter that could be fit to the data.

To model the Glezer et al. (2009) experiments, we ran each word in a pair through the network and measured the distance between their output activations. 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.

Results and Discussion

Network Performance

The VWFANet achieved 98.5% accuracy on the test set. We deemed this sufficient to proceed with testing the network in a paradigm similar to that used by Glezer et al. (2009).

Some samples of correct and incorrect predictions by the VWFANet can be seen in Figure 6. 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”

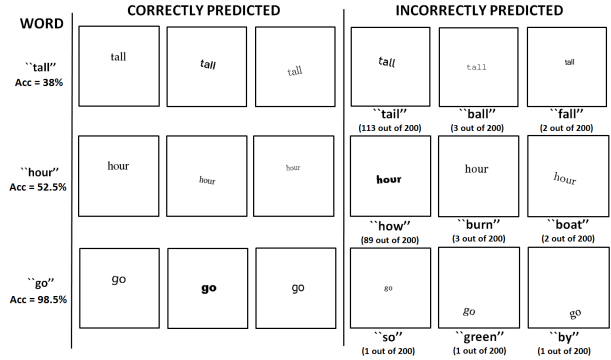


Figure 6: Some examples of correct and incorrect predictions. On the right, the incorrect predictions, and their frequency, are listed below the panels. Embiggen the pdf to see these better.

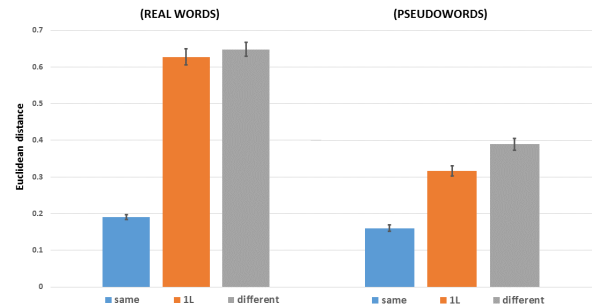


Figure 7: Euclidean distance between the activations in the output layer, with the softmax temperature set to 4. The blue bar represents the distance in activation of two instances of the same word (or pseudoword, on the right), while the orange bar represents the distance between the activations of two words or pseudowords differing by one letter. The gray bar represents the distance between two different words or pseudowords. Error bars represent the standard error of the mean. In each case, these are averages over the stimuli used in the human subject experiments.

with an accuracy of 52.5%, respectively. However, as can be seen in Figure 6, the network makes reasonable mistakes. Out of the 200 images in the test set for the word “tall”, 124 of them are predicted incorrectly; 113 of those 124 images are categorized 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 has high resemblance to “hour”.

Selectivity Analysis of VWFANet

Figure 7 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 directly to the results of the Glezer et al. (2009) study, since the measurements are not the same, we can still observe a trend similar to the findings shown in Figure 3.

It should be noted that there is a small distance between the activations in the last layer in the “same” condition. This

is because we use different images of the same word to compute the response. 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 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 8.

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 9 shows that the neurons from the softmax layer that get activated on pseudowords are more distributed (compare to Figure 8), but there is more overlap when the stimuli are similar, compared to when the stimuli are completely different.

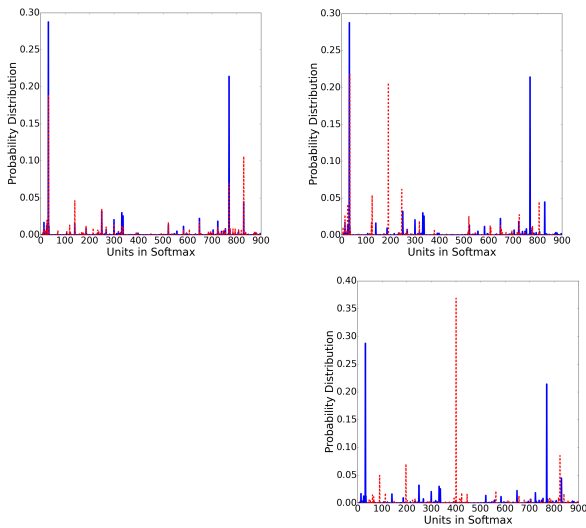


Figure 8: The activation patterns in the output layer for each pair of real words: two instances of **arm**, one of **arm** and **art**, and (bottom right) **arm** and **key**. **arm**’s activation is shown in blue. The words are sorted alphabetically, so in the upper right, the highest activation for **art** is right next to **arm**.

Conclusions

We developed a model (VWFANet) of the Visual Word Form Area using a deep convolutional neural network. Our model was trained on over a million images of almost 900 words, and achieved a very high accuracy of 98.5% on the test set.

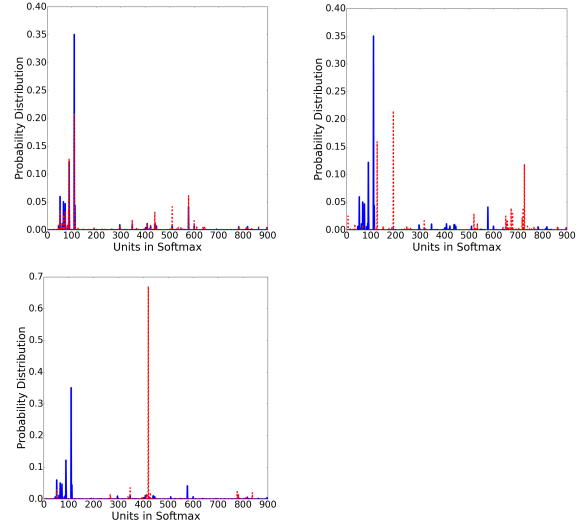


Figure 9: The activations patterns in the output layer for each pair of pseudowords: two instances of **bot**, **bot** and **sot**, and **bot** and **lan**. **bot**’s activation is shown in blue. Note that while **bot**, **sot** and **lan** are all real words to computer scientists, the network does not know that!

Most of the mistakes made by the network, while few, were very reasonable, such as confusing “tail” with “tail.”

Our model of the fMRI-RA results was predicated on the idea that the amount of release from adaptation upon the presentation of a new stimulus should be proportional to the difference in their representations. We measured this as the Euclidean distance between the representations of words at the output layer of the network. This is consistent with previous modeling work that measures the distance between stimulus representations as the distance between their output activations, and fits human judgments of similarity quite well (Dailey, Cottrell, Padgett, & Adolphs, 2002).

Using this interpretation of release from adaptation, VWFANet was able to qualitatively reproduce the data that Glezer et al. (2009) observed in their human neuroimaging study. That is, real words, no matter how similar, were equally distant from one another at the output layer of the network. The representations developed at the output layer were also relatively sparse (Figure 8). On the other hand, pseudowords displayed more graded distances, and less sparsity at the output layer (Figure 9). These results suggest that Glezer et al.’s interpretation of their fMRI-RA results are quite reasonable - real words are represented by different populations of neurons, while pseudowords are represented by distributed patterns of activation over the representations of real words.

It is of considerable note that the fMRI-RA results support the Interactive Activation Model’s representation of words versus pseudowords (McClelland & Rumelhart, 1981). In that model, words were represented by separate nodes in the network at a “word level.” The word superiority effect in letter perception was explained in the model as feedback to the letter level from the word level. On the other hand, the pseu-

doword superiority effect, for example in the pseudoword “mave,” was explained by the sum of feedback activation from many partially activated words at the word level (e.g., “save,” “have,” “wave,” etc.). The current results are consistent with that account.

The current model is just the first step in modeling the VWFA. There are many other experiments that could be modeled by this same architecture. For example, Dehaene et al. (2004) compared activation to the “same” words in different cases and positions, but they also include “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 circular anagrams are the French words “reflet” and “trefle.” If we can find some pairs of English words that are circular anagrams, we can add this to our experiments. This goes a step further than the “1L” condition 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 for 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. 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.

Acknowledgments

This work was supported by NSF grants IIS-1219252 and SMA 1041755 to GWC.

References

- Bengio, Y., Goodfellow, I. J., & Courville, A. (2015). *Deep learning*. (Book in preparation for MIT Press)
- Cohen, L., Dehaene, S., Naccache, L., Lehéricy, S., Dehaene-Lambertz, G., Hénaff, M.-A., & Michel, F. (2000). The visual word form area. *Brain*, *123*(2), 291–307.
- Dailey, M. N., Cottrell, G. W., Padgett, C., & Adolphs, R. (2002). Empath: A neural network that categorizes facial expressions. *Journal of Cognitive Neuroscience*, *14*(8), 1158–1173.
- Dehaene, S., & Cohen, L. (2011, Jun). The unique role of the visual word form area in reading. *Trends in Cognitive Sciences*, *15*(6), 254262.
- Dehaene, S., Jobert, A., Naccache, L., Ciuciu, P., Poline, J.-B., Le Bihan, D., & Cohen, L. (2004). Letter binding and invariant recognition of masked words: Behavioral and neuroimaging evidence. *Psychological Science*, *15*(5), 307–313.
- Glezer, L. S., Jiang, X., & Riesenhuber, M. (2009). Evidence for highly selective neuronal tuning to whole words in the “visual word form area”. *Neuron*, *62*(2), 199–204.
- Grill-Spector, K., Henson, R., & Martin, A. (2006, Jan). Repetition and the brain: neural models of stimulus-specific effects. *Trends in Cognitive Sciences*, *10*(1), 1423.
- Jarrett, K., Kavukcuoglu, K., Ranzato, M., & LeCun, Y. (2009, Sept). What is the best multi-stage architecture for object recognition? In *ICCV 2009* (p. 2146–2153).
- Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., ... Darrell, T. (2014). Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*.
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In F. Pereira, C. Burges, L. Bottou, & K. Weinberger (Eds.), *Advances in neural information processing systems 25* (pp. 1097–1105). Curran Associates, Inc.
- LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. In *Proceedings of the IEEE* (pp. 2278–2324).
- McCandliss, B. D., Cohen, L., & Dehaene, S. (2003). The visual word form area: expertise for reading in the fusiform gyrus. *Trends in Cognitive Sciences*, *7*, 293–299.
- McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part 1. an account of basic findings. *Psychological Review*, *88*(5), 375–407.
- Vigneau, M., Jobard, G., Mazoyer, B., & Tzourio-Mazoyer, N. (2005, September). Word and non-word reading: what role for the visual word form area. *NeuroImage*, *27*(3), 694–705.
- Vinckier, F., Dehaene, S., Jobert, A., Dubus, J. P., Sigman, M., & Cohen, L. (2007). Hierarchical coding of letter strings in the ventral stream: Dissecting the inner organization of the visual word-form system. *Neuron*, *55*(1), 143–156.