February 9, 2022
This is a preprint that has not undergone peer review yet.

A large-scale computational model to accurately predict early brain activity in response to written words

Marijn van Vliet[1*], Oona Rinkinen[1], Takao Shimizu[1], Barry Devereux[3], and Riitta Salmelin[1,2]

[1]Department of Neuroscience and Biomedical Engineering, Aalto University, Finland
[2]Aalto NeuroImaging, Aalto University, Finland
[3]School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, United Kingdom
[*]Corresponding author: marijn.vanvliet@aalto.fi

Abstract

To better understand the computational steps that the brain performs during reading, we used a convolutional neural network as a computational model of visual word recognition, the first stage of reading. In contrast to traditional models of reading, our model directly operates on the pixel values of an image containing text, and has a large vocabulary of 10k Finnish words. The same stimuli can thus be presented unmodified to both the model and human volunteers in a magnetoencephalography (MEG) scanner. In a direct comparison between model and brain activity, we show that the model accurately predicts the amplitude of three evoked MEG response components commonly observed during reading. We conclude that the deep learning techniques that revolutionized models of object recognition can also create models of reading that can be straightforwardly compared to neuroimaging data, which will greatly facilitate testing and refining theories on language processing in the brain.

1 Introduction

What computational operations is the brain performing when it recognizes some splotches of ink on a piece of paper as a meaningful word? This question has been the focus of a large number of neuroimaging studies that examine brain activity during reading. Noninvasive measurement techniques such as electroencephalography (EEG),[1] magnetoencephalography (MEG)[2] and functional magnetic resonance imaging (fMRI)[3] have provided a wealth of information about when and where changes in activity might be expected during various tasks involving orthographic processing.[4] However, these observations alone are unlikely to lead to a complete understanding of the computations performed by the brain during reading.[5] To develop such an understanding, we need to make these computations explicit, model them, and test and refine the model against the neural signatures observed during imaging studies.[6] In this study, we attempted to model the macro-level computations — that is, the net result of thousands of individual biological neurons working together — that manifest as evoked components in EEG/MEG data. We demonstrate a large-scale model of visual word recognition that enabled us, for the first time, to present to the model the same set of bitmap images that were presented to human volunteers in an earlier study,[7] and thereby directly compare the activation inside the model with MEG data.

[1] Grainger and Holcomb, 2009
[2] Salmelin, 2007
[3] Price, 2012
[4] Carreiras et al., 2014
[5] Poeppel, 2012
[6] Barber and Kutas, 2007; Price, 2018
[7] Vartiainen et al., 2011

In past decades, several computational models have been created that successfully capture some aspects of visual word recognition. Amongst the first such models, the interactive activation and competition (IAC) model of letter perception, by McClelland and Rumelhart,[8] showed how the interplay of bottom-up and top-down connections results in a system capable of "filling in the blanks" when faced with a partially obscured word. This model was later extended to model semantics as well, showing how the activation of some semantic features ("is alive", "is a bird") leads to the subsequent activation of more semantic features ("can fly", "lays eggs"),

[8] McClelland and Rumelhart, 1981; Rumelhart and McClelland, 1982

in what became known as the parallel distributed processing (PDP) framework.[9] Note that while these models consist of many interconnected units, those units and their connections do not aim to model a biological connectome, but are an abstract representation of macro-level computations performed by one. Coltheart et al. (2001) pointed out the benefits of explicitly defined connections in their dual-route cascaded (DRC) model of visual word recognition and reading out loud, as they grant the researcher exact control over the macro-level computations. However, as the scope of the models increases, "hand-wiring" the connections in the model becomes increasingly difficult. Therefore, most current models employ back-propagation to learn the connection weights between units based on a large training dataset. Together, PDP and DRC models have been shown to account for many behavioral findings, such as reaction times and recognition errors, in both healthy volunteers and patients.[10] However, in addition to verifying the eventual output of the model against observed behavior, it would seem informative and important to also verify the intermediate computations by comparing the internal state of the model against recordings of brain activity. Unfortunately, none of these models can produce detailed quantitative data that can be directly compared with neuroimaging data, leaving researchers to focus on finding indirect evidence that the same macro-level computations occur in both the model and the brain, with varying levels of success.[11]

More recently, progress has been made towards models whose internal state can be more easily compared with data obtained in neuroimaging studies. Laszlo and Plaut (2012) took an important step by creating a PDP-style model of visual word recognition that performs the computations required to translate a collection of letters (e.g., DOG) into an abstract semantic representation of a concept (e.g., a furry animal that barks), while producing an EEG-like signal.[12] By summing the activity of the computational units in specific layers of the model, the resulting time-varying signal resembles a well-known signal component, observed in EEG and MEG studies, referred to as the N400 potential.[13] However, while this model produces a neuroimaging-like signal, it still cannot be numerically compared with real neuroimaging data, since the simulated environment in the model is extremely simplified (e.g. it operates exclusively on 3-letter words), whereas the brain data will, by nature, reflect the reading process in more varied and realistic visual settings.

In addition to using a small and restricted vocabulary, another striking simplification made in both PDP-based and DRC models is in how the initial visual input is presented to the model. The models make use of "letter banks", where each letter of a written word is encoded in as a separate group of inputs. The letters themselves are encoded as either a collection of 16 predefined line segments[14] or a binary code indicating the letter.[15] This rather high level of visual representation sidesteps having to deal with issues such as visual noise, letters with different scales, rotations and fonts, segmentation of the individual letters, and so on. However, in studies of brain activity, these processes are thought to contribute a major part to the activity observed during the first 200 ms in EEG/MEG studies of visual word recognition,[16] and to the activity observed in the primary visual cortex and visual word form area (VWFA) through fMRI.[17]

Recent advances in deep learning and its software ecosystem are rapidly changing our notion of what is computationally tractable to model.[18] Convolutional neural networks (CNNs) have emerged that perform scale- and rotation-invariant visual object recognition at very high levels of accuracy.[19] Furthermore, they model some functions of the visual cortex well enough that a direct comparison between network state and neuroimaging data has become possible.[20] Consequently, our understanding of basic visual processing in the brain has increased tremendously.[21] Since visual word recognition, the first stage of reading, can be seen as a specialized

[9] McClelland and Rogers, 2003

[10] McClelland and Rogers, 2003; McLeod et al., 2000; Perry et al., 2007

[11] Barber and Kutas, 2007; Jobard et al., 2003; Protopapas et al., 2016

[12] Laszlo and Armstrong, 2014

[13] Kutas and Federmeier, 2011

[14] McClelland and Rumelhart, 1981

[15] Laszlo and Plaut, 2012

[16] Grainger and Holcomb, 2009; Pammer et al., 2004; Petit et al., 2006; Salmelin, 2007; Tarkiainen et al., 1999

[17] Cohen and Dehaene, 2004; Vinckier et al., 2007

[18] LeCun et al., 2015; Richards et al., 2019

[19] Dai et al., 2021; Krizhevsky et al., 2017; Simonyan and Zisserman, 2015

[20] Devereux et al., 2018; Schrimpf et al., 2018; Yamins and DiCarlo, 2016

[21] Lindsay, 2020

form of object recognition, CNNs may very well be a suitable tool for increasing the scale of traditional connectionist models of reading, allowing for a much more realistic modeling of the early orthographic processing steps than what has been possible so far. Indeed, earlier work has established that a CNN can form internal letter representations[22] and can be coerced into forming a VWFA-like region on top of an existing visual system.[23] In this study, we investigated whether a CNN can produce a signal that can be directly compared to neuroimaging data, in this case MEG, during a single-word reading task.

[22] Testolin et al., 2017

[23] Hannagan et al., 2021

We trained a VGG-11 network[24] to recognize 10 000 different Finnish words. Next, we ran through the model a set of stimuli that had been used in an earlier MEG study on human volunteers.[25] These stimuli consisted of valid Finnish words, pseudowords and consonant strings, and two stimulus types that traditional PDP and DRC models cannot handle, namely letter-like symbol strings, and words embedded in visual noise. We demonstrate that various layers in the model behave similarly to several well-studied evoked responses and evaluate this similarity both qualitatively and numerically.

[24] Simonyan and Zisserman, 2015

[25] Vartiainen et al., 2011

2 Results

2.1 The brain

For this study, we reused MEG data that was collected as part of a study by Vartiainen et al. (2011). During the recording session, 15 participants were presented with 560 orthographic stimuli in a silent reading task (Figure 1A). The stimuli were designed to form a series of experimental contrasts that highlight three processing stages during single word reading: early visual processing, letter shape detection and lexical access.

Source estimation of the sensor-level signals was performed using two methods. The first, minimum norm estimate, dynamic statistical parametric mapping (MNE-dSPM),[26] yields a comprehensive map of where activity can be found (Figure 1B), which is used in this study mostly to provide context for a deeper investigation into three peaks of activity that can be observed along the ventral stream. The second, guided equivalent current dipole (ECD) modeling,[27] summarizes the high-dimensional data as a sparse set of ECDs, each one capturing an isolated spatiotemporal component, allowing us to study selected components in more detail. The resulting participant-specific ECDs were grouped based on their location and the timing of peak signal strength. The current study re-uses three ECD groups defined in the original study[28] along the ventral stream (Figure 1B). Together, the grand-average time courses of signal strength at these ECD groups (Figure 1C) capture the different processing stages that the original study sought to highlight.

[26] Dale et al., 2000

[27] Hansen et al., 2010

[28] Vartiainen et al., 2011

The first group of ECDs is occipitally located and represents early onset activity in the visual cortex, peaking at 65 ms to 115 ms after the stimulus onset. The signal strength at these ECDs is driven by the visual complexity of the stimulus and is characterized in this study by a large response to noise-embedded words relative to all other stimulus types. Following Tarkiainen et al. (1999), we will refer to this effect as the Type I response.

The second group is located further along the fusiform gyrus, with the signal strength peaking slightly later at 140 ms to 200 ms. ECDs belonging to this group exhibit sensitivity to whether the stimulus contains letters that are part of the participant's native alphabet,[29] and is characterized in this study by a smaller response to stimuli containing symbol strings relative to stimuli that contain letters. Note that, in contrast to the Type I response, adding noise to the stimulus decreases the signal for these ECDs, further indicating that the underlying cognitive computations are more selective towards specific visual features. Following Tarkiainen et al.

[29] Tarkiainen et al., 1999

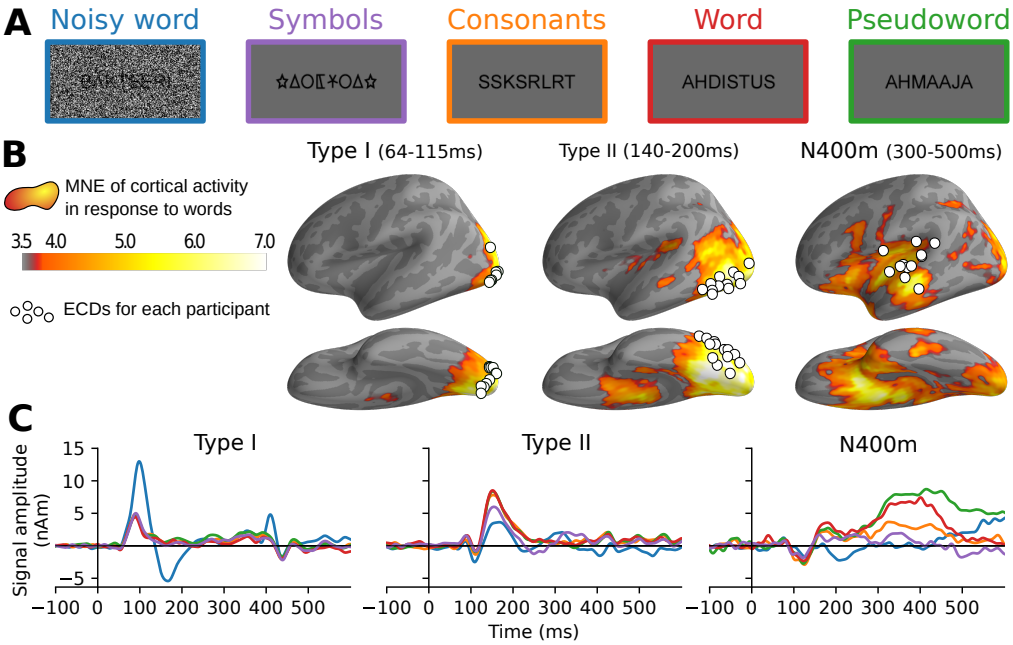

Figure 1: Summary of the MEG results obtained by Vartiainen et al. (2011).
A: Examples of stimuli used in the MEG experiment. Each stimulus contained 7–8 letters or symbols.
B: Source estimate of the evoked MEG activity, using MNE-dSPM. The grand-average activity to word stimuli, averaged for three time intervals, is shown in orange hues. For each time interval, white circles indicate the location of the most representative left-hemisphere ECD for each participant, as determined by Vartiainen et al. (2011).
C: Grand-average time course of signal strength for each group of ECDs in response to the different stimulus types. The traces are color-coded to indicate the stimulus type as shown in **A**.

(1999), we will refer to this as the Type II response. The Type II response is located in the same general cortical region as the VWFA reported using fMRI,[30] but they are not spatially or functionally identical.[31]

[30] Cohen and Dehaene, 2004

[31] Vartiainen et al., 2011

The third and final ECD group is located in the temporal cortex, peaking much later at 300 ms to 500 ms. Its signal strength is modulated by the lexical content of the stimulus, manifested in this study as a larger response to the word-like (i.e., words and pseudowords) versus the non-word-like stimuli. The location and response pattern of this ECD group corresponds to that of the N400m.[32]

[32] Halgren et al., 2002; Helenius et al., 1998; Salmelin, 2007; Service et al., 2007

2.2 The model

As a model of the computations underlying the brain activity observed during the MEG experiment, we used a VGG-11[33] network architecture, pretrained on ImageNet.[34] This architecture consists of five convolution layers (three of which perform convolution twice), followed by two densely connected layers, terminating in an output layer (Figure 2A). The model was trained to perform visual word recognition using a training set of 1 000 000 images, where each image depicted a word, rendered in varying fonts, sizes and rotations, with varying degrees of visual background noise (Figure 2B). The vocabulary size (10 000) was chosen to exceed the number of units in the densely connected layers (4 096), forcing the model to construct a sub-lexical representation. To add a detection element to what is otherwise a discrimination task, an additional 50 000 images containing only noise were added to the training set. The task for the model was to identify the correct word (regardless of the font, size and rotation used to render the text) by setting the corresponding unit in the output layer to a high value. If no word was present in the image, all output units should have a low value. This was achieved be replacing the output layer of the pre-trained model with one containing 10 001 randomly initialized units for word recognition training, where the final unit was trained to denote the "no word present" condition, and removing this final unit after training was complete. During training, the performance of the model was evaluated on an independent test set of 100 000 images that contained words and 5 000 that contained only noise. Training was stopped when the model's performance plateaued, at which point the accuracy on the test set was 99.3 %.

[33] Szegedy et al., 2015

[34] Russakovsky et al., 2015

After training, the model was used to classify the stimulus images presented during the MEG

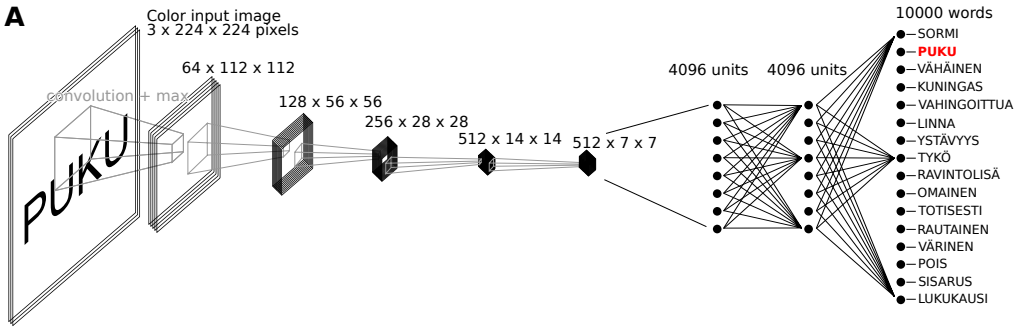

Figure 2: Overview of the proposed computational model of visual word recognition in the brain.
A: The VGG-11 model architecture, consisting of five convolution layers, two fully connected layers and one output layer.
B: Examples of the images used to train the model.

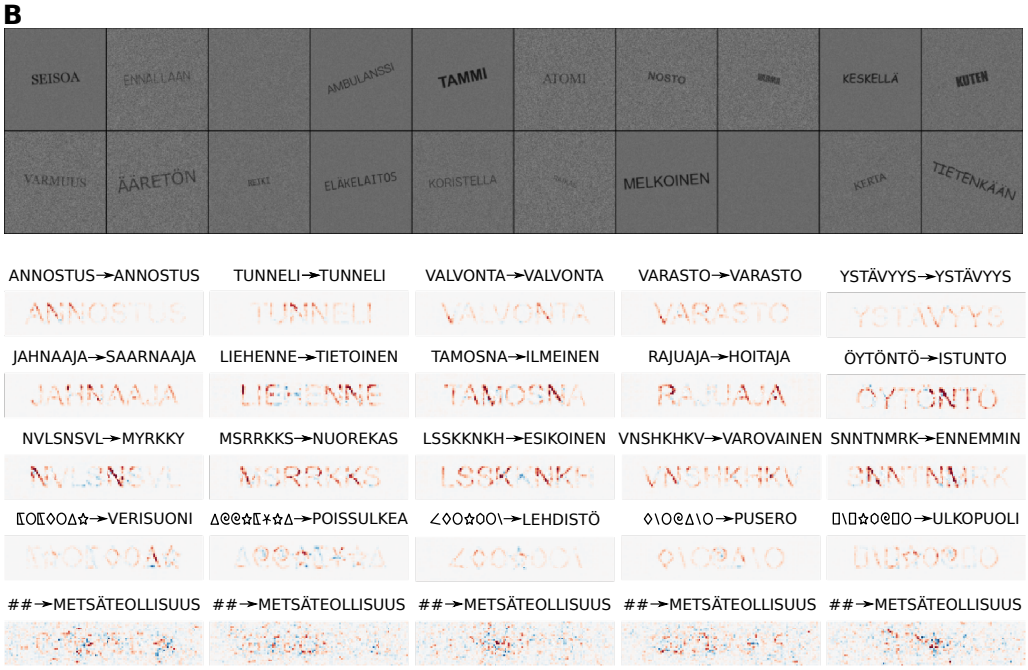

Figure 3: Contribution maps
For each type of stimulus (rows from top to bottom: words, pseudowords, consonants, symbols and noise embedded words), DeepLIFT contribution maps are shown for five exemplars. For each exemplar, the original word in the stimulus is given, followed by the word corresponding to the node in the output layer that was activated the strongest (i.e., the model prediction). The maps indicate the amount of influence each pixel has on this output node (red=positive, blue=negative, in arbitrary units).

experiment (see Supplementary Table 1). The model classified 114 of the 118 word stimuli correctly (accuracy 96.6 %). The 5 incorrectly classified stimuli were classified as close orthographic neighbors (e.g. LUOMINEN→TUOMINEN). All noise-embedded words were misclassified, since the amount of noise made them unrecognizable, with 99 out of 118 being classified as the word METSÄTEOLLISUUS, which is one of the words in the training set with the highest visual complexity. Since the model was only trained to detect and discriminate between words, all non-word stimuli (pseudowords, consonant strings, symbol strings) were misclassified as well. In this case, it was more difficult to find a pattern in the way they were classified. In many cases, similarity in letter shape seemed to play a role (e.g. SKKNTMT→SUKUNIMI, ÄHKÄÄJÄ→VARAAJA), but not always (e.g. INKRIHTI→EMOLEVY).

To better understand what kind of visual features the model used to determine the output word, we performed DeepLIFT[35] analysis to determine the contribution of each input pixel to the strongest activated output node (Figure 3). We observed similar contribution patterns for each individual letter, no matter its location in the string, indicating that the model learned to use individual letter shapes as building blocks.

[35] Shrikumar et al., 2019

2.3 Comparing model and brain

To directly compare layer activations in the model with the brain responses in word reading (Figure 4), we collected summary statistics, here referred to as "response patterns", for the ECD time courses, MNE-dSPM source estimates, and the rectified linear unit (ReLu) activations in the model. For each ECD, the response pattern was obtained by integrating the signal strength

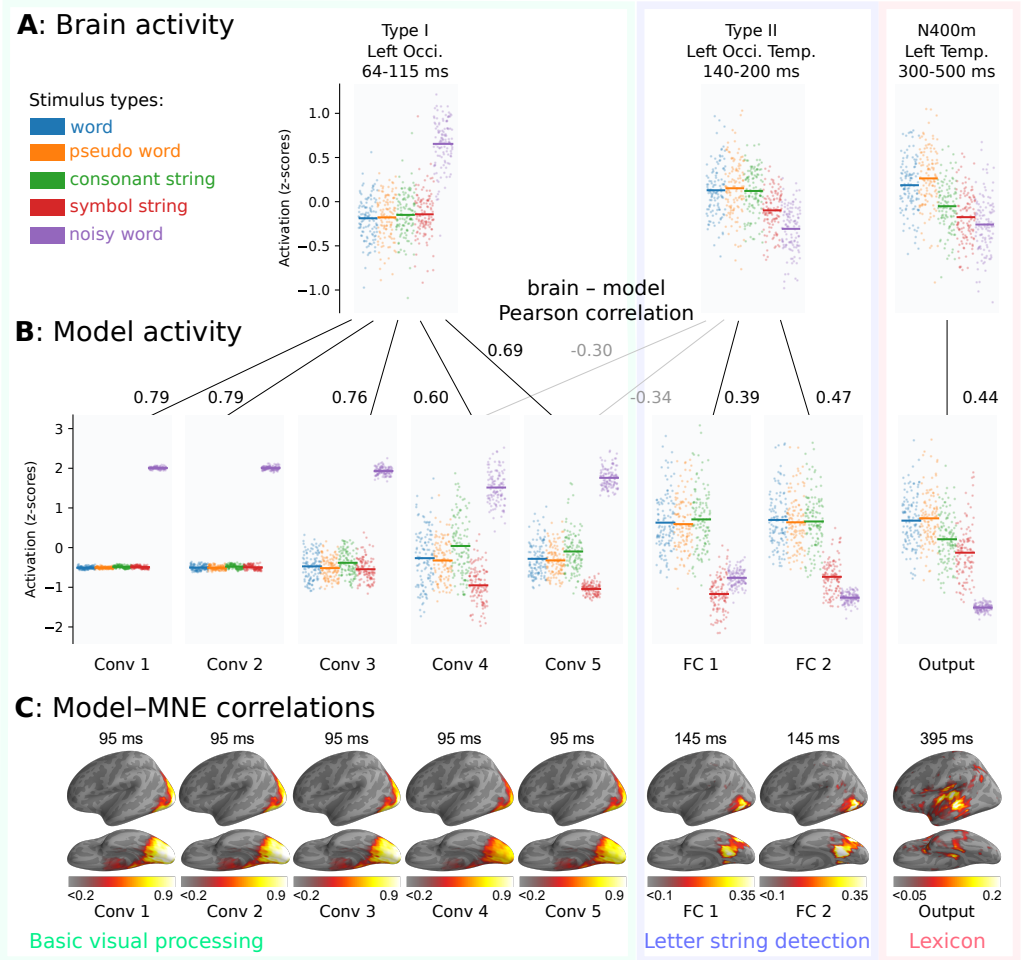

Figure 4: Grand-average comparison between brain and model. Based on the response patterns to each stimulus type, three processing stages were identified that correspond to different time windows in the MEG activity and different layer types in the model.
A: Grand-average response patterns obtained from evoked MEG activity, quantified using three ECD groups. For each stimulus, the ECD time courses were integrated over the indicated time interval.
B: Response patterns obtained from the model. For each layer, the mean ReLu activation in response to each stimulus. Matching response patterns between ECDs and model layers are shown using black lines, accompanied by the corresponding Pearson correlations. Two additional correlation values are shown in gray, for which the response patterns do not match, but which are discussed in the main text.
C: Pearson correlation between the model and the MNE-dSPM source estimate. Grand-average source estimates were obtained in response to each stimulus. The correlation map was obtained by correlating the activity at each source point with that for each layer of the model. The correlation map is shown at the time of peak correlation (using the same analysis window as in **A**). Only positive correlations are shown.

in response to each stimulus over the time window used in Vartiainen et al. (2011) for statistical analysis, followed by a z-transformation (Figure 4A). The single-participant response patterns were then averaged to obtain the grand-average response patterns per stimulus. For the model, response patterns were obtained for each layer, by computing the mean ReLu activation across all the units in the layer in response to each stimulus, followed by a z-transformation (Figure 4B). Finally, response patterns were obtained for each source/time point in the grand-average MNE-dSPM source estimates by computing the magnitude of the estimated signal at that point (Figure 4C).

Since the same individual stimuli were processed by both the model and the brain, we can measure the model-brain correspondence qualitatively by examining the overall shape of the response patterns, and quantitatively by computing the correlation between them (Figure 4, Table 1). Note that, in addition to real words with and without noise and pseudowords, the stimuli also included consonant and symbol strings, which the model was not trained to recognize.

In the first three convolution layers of the model, we see a much larger response to the noise-embedded word stimuli, relative to the other stimulus types. This is likely driven by the much higher visual complexity of the noisy stimuli. This corresponds well to the Type I response pattern, i.e. high response to noisy words and no distinction among the other stimulus types (correlations between grand-average patterns: 0.79/0.79/0.76, average single-participant correlations: 0.33/0.33/0.32 [$p < 0.05$]).

We see a change in the response pattern of the model in the next two convolution layers, layers 4 and 5. In these layers, the ReLu activation for symbol strings is now lower than that for the other stimulus types. This indicates the presence of convolution filters that are sensitive to the specific line arrangements that are found in letters of the Finnish alphabet, but not present in the symbol strings. In this regard, the response pattern of these layers is similar to the Type II response pattern. However, many of the filters remain sensitive to noise, which does not match the Type II response pattern, resulting in a negative correlation (correlations between grand-average patterns: -0.30/-0.34, average single-participant correlations: -0.10/-0.11 [n.s.]). Overall, the response patterns of layers 4 and 5 correlate best with the Type I response pattern, although they are not strictly consistent (correlations between grand-average patterns: 0.60/0.69, averaged single-trial correlations: 0.25/0.29 [$p < 0.05$]).

The selectivity for letter shapes becomes more pronounced in the two fully connected layers of the model. In these layers, we see an increase in ReLu activation for stimuli containing letter shapes, and a decrease in activation for noise-embedded words. However, no distinction is made between consonant strings and word-like stimuli, indicating that letters are detected in isolation. This means the response patterns of these layers are consistent with the Type II response pattern (correlations with grand-average: 0.39/0.47, averaged subject-subject correlations: 0.12 [$p < 0.01$], 0.15 [$p < 0.05$]), and is consistent with the Type II response.

Finally, at the output layer of the model, we observe that word-like stimuli produce more activity than consonant strings, indicating that the shapes of multiple letters are combined to produce the output. It is noteworthy that pseudowords, which were not present in the training set, produce a roughly equal amount of activation in the output layer as valid Finnish words. This means that the response pattern of the output layer corresponds best to that of the third ECD group in the temporal cortex (correlation with grand-average: 0.44, average subject-subject correlation: 0.16 [$p < 0.001$]), and is consistent with the N400m response.[36]

[36] Salmelin, 2007

To verify that the three ECD groups capture the most important correlations between the brain activity and the model, we compared the model to cortex-wide MNE-dSPM source estimates (Figure 4C). For each layer in the model, the areas with maximum correlation appear around the locations of the ECDs and at a time that is close to the peak activation of the ECDs.

3 Discussion

In this study, we built a model of visual word recognition that operates on the raw pixel values of a bitmap image, rather than the more abstract representations found in many earlier models.[37] Such a model can be tested with a wider variety of visual stimuli than traditional models, including a more naturalistic range of fonts, sizes, rotations, as well as non-letter stimuli such as visual noise and symbol strings. Furthermore, the deep-learning architecture scales to encompass a vocabulary with the same order of magnitude as a human's, enabling it to be

[37] Coltheart et al., 2001; Laszlo and Armstrong, 2014; McClelland and Rumelhart, 1981

Layer	Type	Type I	Type II	N400m
1	Convolution	0.33*	−0.15	−0.13***
2	Convolution	0.33*	−0.15	−0.13***
3	Convolution	0.32*	−0.15	−0.12***
4	Convolution	0.25*	−0.10	−0.08***
5	Convolution	0.29*	−0.11	−0.08***
6	Fully connected	−0.13*	0.12**	0.14***
7	Fully connected	−0.22*	0.15*	0.15***
8	Output	−0.26*	0.15*	0.16***

*$p < 0.05$ **$p < 0.01$ ***$p < 0.001$
linear mixed effects (LME) model, false discovery rate (FDR) corrected

Table 1: Single-participant comparison between brain and model. Shown are the average Pearson correlations between the single-participant ECD response patterns and the response patterns for each layer of the model.

tested meaningfully with pseudowords. The combination of all of this means that the model is able to operate on identical stimuli as those commonly used in neuroimaging studies of reading, and is able to perform the same task as the volunteer, with comparable accuracy and robustness against irrelevant visual detail. This bridges an important gap between modeling and experimental efforts in this domain, namely the ability to evaluate a model by its numerical fit to neuroimaging data.

Our model aims to provide a good fit to the brain activity observed during the first 400 ms after stimulus onset. Specifically, the mean ReLu activity in the layers of the model provides a good explanation of the responses of three important evoked components that are observed during MEG studies of word recognition.[38] Specifically, we observed that the convolution layers respond more to noise embedded words than the stimuli without noise, which mimics the Type I response that is observed around 64 ms to 115 ms in the visual cortex. Next, two fully connected layers responded more to valid Finnish letters than letter-like symbol strings, which mimics the Type II response that is observed around 140 ms to 200 ms occipital-temporally in the brain.[39] Finally, the output layer responded more to words and pseudo-words than consonant strings, which mimics the N400m response that is observed around 300 ms to 500 ms in the temporal cortex.[40] The progression of the signal through the layers in the model tracks the dynamics of the MEG data, with earlier layers matching with earlier time points.

[38] Salmelin, 2007

[39] Tarkiainen et al., 1999

[40] Halgren et al., 2002; Helenius et al., 1998; Salmelin, 2007; Service et al., 2007

The model uses a conventional CNN architecture that is similar to architectures that have previously been shown to describe the initial stages of visual object recognition.[41] No changes to the VGG-11 architecture[42] were needed to create a successful model. Using an approach that somewhat mimics how humans build an orthographic system on top of their existing visual system when they learn how to read,[43] we started with a model that was pre-trained to perform visual object recognition on ImageNet, and then further trained it using bitmap images of written words. The resulting model learned to recognize letter shapes and combine them into one of 10 000 possible words.

[41] Devereux et al., 2018; Schrimpf et al., 2018; Yamins and DiCarlo, 2016

[42] Simonyan and Zisserman, 2015

[43] Cohen and Dehaene, 2004; Vinckier et al., 2007

During training, the learning algorithm was required to adapt the existing visual network into a model of visual word recognition. This is in line with the observation that proficiency with written language has developed too recently in our history for any evolutionary changes in brain connectivity to support this skill to have occurred. Instead, acquisition of rapid visual word recognition likely relies on a reconfiguration of parts of an existing visual network.[44]

[44] Dehaene-Lambertz et al., 2018

Comparing unit activations in a CNN to MEG data means comparing two high-dimensional datasets, for which a mapping from one unto the other needs to be created. In other studies, this was often achieved through information-based techniques such as representational similarity analysis (RSA) or pattern-based techniques such as multivariate regression. Such methods allow for a lot of flexibility in how the model maps onto the brain signal. However, this flexibility is ultimately undesired when we strive for a more complete understanding of how computational processes in the brain manifest in the signal recorded by the scanner. In this study, we were able to make the mapping between the model and brain activity explicit enough to enable a rigid univariate comparison.

3.1 Limitations of the current model

The VGG-11 architecture was originally designed to achieve high image classification accuracy on the ImageNet challenge.[45] As a model of brain function to explain neuroimaging data during reading, it is still incomplete in many ways. As neuroimaging experiments of language started by mapping responses to basic visual and auditory stimuli, so does the modeling effort

[45] Simonyan and Zisserman, 2015

start with an implementation of only the most basic principles.

One important limitation of the current model is the lack of an explicit mapping from the units inside its layers to specific locations in the brain at specific times. Multiple layers have a maximum correlation with the source estimate at the exact same location and time (Figure 4C). Furthermore, there is no clear relationship between the number of layers the signal needs to traverse in the model to the processing time in the brain. The signal needs to propagate through five convolution layers to reach the point where it matches the Type II response at 140 ms to 200 ms, but only through two additional layers to reach the point where it matches the N400m response at 300 ms to 500 ms. There are network architectures like CORNet-S[46] that aim for an explicit mapping between layers in the model and the V1-V4 areas of the visual cortex, although the first layer of the model should probably be mapped onto the retina, rather than V1.[47]

[46] Kubilius et al., 2019

[47] Lindsey et al., 2019

As for all models that rely on machine learning to set millions of parameters, it's not always straightforward to understand what exact computations the model itself is performing. In this study, we used DeepLIFT contribution maps to gain some insight into this (Figure 3), and others have explored this issue more deeply in the specific context of CNNs that perform visual word recognition.[48]

[48] Hannagan et al., 2021; Testolin et al., 2017

Furthermore, the scope of the current model is limited to a bottom-up processing stream that starts from the initial pixel input up to recognition of a single orthographic word-form up until right before the formation of any abstract semantic representations. However, there is some basis for an extension of the model into the semantic domain, as there is a growing body of literature exploring how to learn suitable representations from text corpora[49] and how to combine multiple words into sentence representations.[50]

[49] Günther et al., 2019; Mikolov et al., 2013; Pennington et al., 2014

[50] Brown et al., 2020; Devlin et al., 2019

3.2 Conclusion

Despite its limitations, our model is an important milestone for computational implementations of cognitive theories of language on the level of the brain. The ability to directly compare a model with neuroimaging results greatly facilitates empirical verifiability of such theories. Our model is a powerful platform for future extensions that implement more accurate, detailed and complete theories, where the merit of such extensions can be evaluated based on the improvement of the fit to the data.

4 Methods

4.1 MEG study design and data analysis

The MEG data was collected by Vartiainen et al. (2011), and was re-analyzed for the current study using MNE-python[51] and FreeSurfer.[52] We refer to the original publication for additional details on the study protocol and data collection process.

[51] Gramfort et al., 2013

[52] Dale et al., 1999

Simultaneous MEG and EEG was recorded from 15 participants (who gave their informed consent, in agreement with the prior approval of the Ethics Committee of the Helsinki and Uusimaa Hospital District). The stimuli consisted of visually presented Finnish words, pseudowords (pronounceable but meaningless), consonant strings (random consonants), symbol strings (randomly selected from 10 possible shapes), and words embedded in high-frequency visual noise (Figure 1A). Each stimulus category contained 112 different stimuli and each stimulus contained 7–8 letters/symbols. The stimuli were presented sequentially in a block design. Each block started with a random period of rest (0 ms to 600 ms), followed by 7 stimuli of the same category. Each stimulus was shown for 300 ms with an inter-stimulus interval of 1500 ms of

gray background. In addition to the five stimulus conditions (16 blocks each), there were 16 blocks of rest (12 s). An additional 5 target blocks were added, in which one stimulus appeared twice in a row, that the participants were asked to detect and respond to with a button press. The target blocks were not included in the analysis. The same paradigm was applied in an fMRI-EEG recording of the same participants as well (data not used here).

The data were recorded with a Vector View device (Elekta Neuromag, now MEGIN Oy, Finland) in a magnetically shielded room. In addition, vertical and horizontal electro-oculography (EOG) were recorded. The signals were bandpass filtered at 0.03 Hz to 200 Hz and digitized at 600 Hz. For analysis, the data were further processed using the maxfilter software (MEGIN Oy, Finland) and bandpass filtered at 0.1 Hz to 40 Hz. Eye-movement and heart-beat artifacts were reduced using independent component analysis (ICA): the signal was high-passed at 1 Hz, and components with correlations larger than 5 standard deviations with any of the EOG channels or an estimated electro-cardiogram (based on the magnetometers without the maxfilter applied) were removed (3–7 out of 86–108 components). Epochs were cut −200 ms to 1000 ms relative to the onset of the stimulus and baseline-corrected using the pre-stimulus interval. Participant-specific noise covariance matrices of the MEG sensors were estimated using the pre-stimulus interval.

For ECD and MNE-dSPM source estimation, 3-layer boundary element method (BEM) forward models were used. These models were based on T1-weighted anatomical images that were acquired using a 3 T Signa EXCITE magnetic resonance imaging (MRI) scanner. FreeSurfer[53] was used to obtain the surface reconstructions required to compute the BEM models. Vartiainen et al. (2011) defines three groups of ECDs in the left-hemisphere, associated with the Type I, Type II and N400m responses. The locations and orientations of the ECDs in these groups were re-used in the current study. For each individual epoch, the signal at each of the three ECDs and a cortex-wide MNE-dSPM source estimate was computed using the noise covariance matrices and BEM forward models. Finally, the locations of the ECDs and MNE-dSPM source points were morphed to a template brain (FreeSurfer's fsaverage brain).

[53] Dale et al., 1999

4.2 Model

The computational model has a VGG-11 architecture[54] as defined in the TorchVision[55] module of the PyTorch[56] (Figure 2A). The architecture specifies 5 convolution layers, with the final 3 performing convolution twice. Each individual convolution step is followed by batch-normalization[57] and ReLu activation, and max-pooling is performed as final step of each layer. The 5 convolution layers are followed by 2 followed by 2 linear fully-connected layers, with ReLu activation, and employing 50 % dropout. The final output layer is also a linear, fully-connected layer with ReLu activation.

[54] Simonyan and Zisserman, 2015
[55] Marcel and Rodriguez, 2010
[56] Paszke et al., 2019
[57] Ioffe and Szegedy, 2015

The model parameters for all layers, except the output layer, were initialized to the ImageNet pre-trained model provided by TorchVision. Then, the model was trained using 1 000 000 bitmap images of rendered text on a background of visual noise, and 500 000 bitmap images containing only visual noise (Figure 2B). An independent test set containing 100 000 images of text and 50 000 images of only noise was used to track the performance of the model during training. None of the images in the training or test set were an exact duplicate of the images used as stimuli in the MEG experiment. Therefore, the stimulus set can be treated as an independent validation set.

The text on the images were one of 10 000 Finnish words. The list of words was compiled using the 112 words used in the MEG, extended with the most frequently used Finnish words, as

determined by Kanerva et al. (2014), excluding proper names and using their lemma form. A training image was assembled by overlaying text onto an image of visual noise. Visual noise was rendered by randomly selecting a gray-scale value (0 % to 100 %) for each pixel. Next, if the training image contained text, the word was rendered in uppercase letters, using a randomly chosen font (15 possible fonts), font size (14 pt to 32 pt), and rotation (-20° to 20°). The list of fonts consisted of: Ubuntu Mono, Courier, Luxi Mono Regular, Lucida Console, Lekton, Dejavu Sans Mono, Times New Roman, Arial, Arial Black, Verdana, Comic Sans MS, Georgia, Liberation Serif, Impact, and Roboto Condensed. The text and the visual noise were mixed at a ratio dictated by a randomly selected noise level (10 % to 30 %) by adjusting their respective opacities to add up to 100 %. For example, at a noise level of 20 %, first the visual noise was rendered at 20 % opacity, then the text was overlayed at $100 - 20 = 80$ % opacity. When the input image did not contain a word, all model outputs should be low, which was achieved by adding a unit to the output layer to indicate this condition. The training objective was to minimize the cross-entropy loss between the model output and a one-hot encoded vector of length 10 001, indicating the target word or the no-word condition. After training, the extra unit was removed from the model and a ReLu activation was placed on the remaining units. Training was performed using stochastic gradient descend with an initial learning rate of 0.001, which was reduced to 0.0001 after 10 epochs (an "epoch" in this context refers to one sequence of all 1 500 000 training images). In total, training was performed for 20 epochs, by which point the performance on the test set had plateaued at 99.6 %.

4.3 Comparison between model and MEG data

To compare the model and MEG, statistical summaries were made of the high-dimensional datasets, such that Pearson correlations could be computed between them. We will refer to these summaries as "activity scores".

After training, the model was applied to the bitmap images that were used during the MEG experiment (padding them to be 256×256 pixels with 50 % greyscale pixels). The activity at each layer of the model was quantified by recording the ReLu activation right before the maxpool operation, and computing the mean across the units. This yielded, for each stimulus image, a single number for each layer in the model representing the amount of activity within that layer. For each layer, the activity scores were z-scored (subtract the mean, divide by the standard deviation) across the stimulus images (Figure 4B).

The activity of each ECD was quantified by taking the mean of the ECD's timecourse (Figure 1C) during the time window used in Vartiainen et al. (2011). This yielded, for each stimulus image, a single number for each ECD for each participant, indicating the amount of signal at the ECD. These activity scores were first z-scored across the stimulus images independently for each participant. The ECDs are participant specific. Vartiainen et al. (2011) clustered the ECDs into groups using their location and timing, where the ECDs in each group were judged to encode the same evoked component in different participants. Three groups were selected as representing the Type I, Type II and N400m responses typically found in the literature.[58] The groups do not necessarily contain an ECD for each of the 15 participants, as some participants did not have an ECD to contribute that matched well enough in position and/or timing. The number of participants contributing an ECD to each group were: Type I-14, Type II-14, and N400m-15. During the grand-average analysis, the activity score of the ECDs was averaged within each group (Figure 4A).

[58] Salmelin, 2007

The MNE-dSPM source estimates used 4000 source points, distributed evenly across the cortex. The activity score at each source point was computed by taking the maximum signal value of

the source point's timecourse within the same time window as the ECD analysis. This yielded for each stimulus image, a single number for each source point for each participant, indicating the amount of signal at the source point (Figure 1C). These activity scores were first z-scored across the stimulus images independently for each participant. Then, during the grand-average analysis, the scores were averaged across participants.

In the grand-average comparison, Pearson correlations were computed between the activity scores of each layer in the model versus each group of ECDs (annotated in Figure 4B) and each MNE-dSPM source point (Figure 4C). In the single-participant comparison, a separate LME model was fitted for each layer in the model versus each group of ECDs, using participants as random effect (random intercept + random slope). The lme4[59] R-package implements the LME model and lmeTest[60] was used to estimate p-values.

[59] Bates et al., 2015

[60] Kuznetsova et al., 2015

4.4 Data availability

The trained model and MEG data necessary for performing the comparison is available at https://osf.io/nu2ep. As the consent obtained from the participants prohibits sharing of personal data, the MEG data that is made available has been processed to such a degree that individual participants cannot be identified. The code for training the model and reproducing the results of this paper is available at https://github.com/wmvanvliet/viswordrec-baseline.

5 Acknowledgements

We acknowledge the computational resources provided by the Aalto Science-IT project. This research was funded by the Academy of Finland (grants #310988 and #343385 to M.v.V, #315553 to R.S.) and the Sigrid Jesélius Foundation (to R.S.).

References

Barber, H. A., & Kutas, M. (2007). Interplay between computational models and cognitive electrophysiology in visual word recognition. *Brain Research Reviews*, *53*(1), 98–123. doi:10.1016/j.brainresrev.2006.07.002

Bates, D. M., Maechler, M., Bolker, B. M., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal Of Statistical Software*, *67*(1), 1–48. doi:10.18637/jss.v067.i01

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. *arXiv:2005.14165 [cs]*.

Carreiras, M., Armstrong, B. C., Perea, M., & Frost, R. (2014). The what, when, where, and how of visual word recognition. *Trends in Cognitive Sciences*, *18*(2), 90–98. doi:10.1016/j.tics.2013.11.005

Cohen, L., & Dehaene, S. (2004). Specialization within the ventral stream: The case for the visual word form area. *NeuroImage*, *22*(1), 466–476. doi:10.1016/j.neuroimage.2003.12.049

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological review*, *108*(1), 204–56. doi:10.1037/0033-295X.108.1.204

Dai, Z., Liu, H., Le, Q. V., & Tan, M. (2021). CoAtNet: Marrying convolution and attention for all data sizes. *arXiv:2106.04803 [cs]*.

Dale, A. M., Fischl, B. R., & Sereno, M. I. (1999). Cortical surface-based analysis: I. segmentation and surface reconstruction. *NeuroImage*, *9*(2), 179–194.

Dale, A. M., Liu, A. K., Fischl, B. R., Buckner, R. L., Belliveau, J. W., Lewine, J. D., & Halgren, E. (2000). Dynamic statistical parametric mapping: Combining fMRI and MEG for high-resolution imaging of cortical activity. *Neuron*, *26*(1), 55–67.

Dehaene-Lambertz, G., Monzalvo, K., & Dehaene, S. (2018). The emergence of the visual word form: Longitudinal evolution of category-specific ventral visual areas during reading acquisition. *PLOS Biology*, *16*(3), e2004103. doi:10.1371/journal.pbio.2004103

Devereux, B. J., Clarke, A., & Tyler, L. K. (2018). Integrated deep visual and semantic attractor neural networks predict fMRI pattern-information along the ventral object processing pathway. *Scientific Reports*, *8*(1). doi:10.1038/s41598-018-28865-1

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805 [cs]*.

Grainger, J., & Holcomb, P. J. (2009). Watching the word go by: On the time-course of component processes in visual word recognition. *Linguistics and Language Compass*, *3*(1), 128–156. doi:10.1111/j.1749-818X.2008.00121.x

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., & Hämäläinen, M. S. (2013). MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*, *7*(December), 1–13. doi:10.3389/fnins.2013.00267

Günther, F., Rinaldi, L., & Marelli, M. (2019). Vector-space models of semantic representation from a cognitive perspective: A discussion of common misconceptions. *Perspectives on Psychological Science*, *14*(6), 1006–1033. doi:10.1177/1745691619861372

Halgren, E., Dhond, R. P., Christensen, N., Van Petten, C., Marinkovic, K., Lewine, J. D., & Dale, A. M. (2002). N400-like magnetoencephalography responses modulated by semantic context, word frequency, and lexical class in sentences. *NeuroImage*, *17*(3), 1101–1116. doi:10.1006/nimg.2002.1268

Hannagan, T., Agrawal, A., Cohen, L., & Dehaene, S. (2021). Emergence of a compositional neural code for written words: Recycling of a convolutional neural network for reading. *Proceedings of the National Academy of Sciences*, *118*(46). doi:10.1073/pnas.2104779118

Hansen, P., Kringelbach, M., & Salmelin, R. (Eds.). (2010). *MEG: An introduction to methods*. Oxford University Press. doi:10.1093/acprof:oso/9780195307238.001.0001

Helenius, P., Salmelin, R., Service, E., & Connolly, J. F. (1998). Distinct time courses of word and context comprehension in the left temporal cortex. *Brain*, *121*(6), 1133–1142. doi:10.1093/brain/121.6.1133

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, 448–456.

Jobard, G., Crivello, F., & Tzourio-Mazoyer, N. (2003). Evaluation of the dual route theory of reading: A metanalysis of 35 neuroimaging studies. *NeuroImage*, *20*(2), 693–712. doi:10.1016/S1053-8119(03)00343-4

Kanerva, J., Luotolahti, J., Laippala, V., & Ginter, F. (2014). Syntactic N-gram collection from a large-scale corpus of internet Finnish. *Frontiers in Artificial Intelligence and Applications*, *268*, 184–191. doi:10.3233/978-1-61499-442-8-184

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84–90. doi:10.1145/3065386

Kubilius, J., Schrimpf, M., Kar, K., Rajalingham, R., Hong, H., Majaj, N., Issa, E., Bashivan, P., Prescott-Roy, J., Schmidt, K., Nayebi, A., Bear, D., Yamins, D. L., & DiCarlo, J. J. (2019). Brain-like object recognition with high-performing shallow recurrent ANNs. *Advances in Neural Information Processing Systems*, *32*.

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event related brain potential (erp). *Annu. Rev. Psychol.*, *62*, 621. doi:10.1146/annurev.psych.093008.131123

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2015). LmerTest: Tests for random and fixed effects for linear mixed effect models. R package, version 2.0-29.

Laszlo, S., & Armstrong, B. C. (2014). PSPs and ERPs: Applying the dynamics of post-synaptic potentials to individual units in simulation of temporally extended event-related potential reading data. *Brain and Language*, *132*, 22–27. doi:10.1016/j.bandl.2014.03.002

Laszlo, S., & Plaut, D. C. (2012). A neurally plausible parallel distributed processing model of event-related potential word reading data. *Brain and Language*, *120*(3), 271–281. doi:10.1016/j.bandl.2011.09.001

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. doi:10.1038/nature14539

Lindsay, G. W. (2020). Convolutional neural networks as a model of the visual system: Past, present, and future. *Journal of Cognitive Neuroscience*, *33*(10), 1–15. doi:10.1162/jocn_a_01544

Lindsey, J., Ocko, S. A., Ganguli, S., & Deny, S. (2019). A unified theory of early visual representations from retina to cortex through anatomically constrained deep CNNs. *International Conference on Learning Representations*. doi:10.1101/511535

Marcel, S., & Rodriguez, Y. (2010). TorchVision the machine-vision package of torch. *Proceedings of the 18th ACM International Conference on Multimedia*, 1485–1488. doi:10.1145/1873951.1874254

McClelland, J. L., & Rogers, T. T. (2003). The parallel distributed processing approach to semantic cognition. *Nature Reviews Neuroscience*, *4*(4), 310–322. doi:10.1038/nrn1076

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part I. an account of basic findings. *Psychological Review*, *88*(5), 580–596. doi:10.1016/B978-1-4832-1446-7.50048-0

McLeod, P., Shallice, T., & Plaut, D. C. (2000). Attractor dynamics in word recognition: Converging evidence from errors by normal subjects, dyslexic patients and a connectionist model. *Cognition*, *74*(1), 91–114. doi:10.1016/S0010-0277(99)00067-0

Mikolov, T., Corrado, G., Chen, K., & Dean, J. (2013). Efficient estimation of word representations in vector space. *Proceedings of the International Conference on Learning Representations (ICLR 2013)*, 1–12. doi:10.1162/153244303322533223

Pammer, K., Hansen, P. C., Kringelbach, M. L., Holliday, I., Barnes, G., Hillebrand, A., Singh, K. D., & Cornelissen, P. L. (2004). Visual word recognition: The first half second. *NeuroImage*, *22*(4), 1819–1825. doi:10.1016/j.neuroimage.2004.05.004

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., … Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems 32* (pp. 8024–8035). Curran Associates, Inc. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. doi:10.3115/v1/D14-1162

Perry, C., Ziegler, J. C., & Zorzi, M. (2007). Nested incremental modeling in the development of computational theories: The CDP+ model of reading aloud. *Psychological review*, *114*(2), 273–315. doi:10.1037/0033-295X.114.2.273

Petit, J.-P., Midgley, K. J., Holcomb, P. J., & Grainger, J. (2006). On the time course of letter perception: A masked priming ERP investigation. *Psychonomic Bulletin & Review*, *13*(4), 674–681. doi:10.3758/BF03193980

Poeppel, D. (2012). The maps problem and the mapping problem: two challenges for a cognitive neuroscience of speech and language. *Cognitive Neuropsycholy*, *29*(1-2), 34–55. doi:10.1080/02643294.2012.710600

Price, C. J. (2012). A review and synthesis of the first 20 years of PET and fMRI studies of heard speech, spoken language and reading. *NeuroImage*, *62*(2), 816–847. doi:10.1016/j.neuroimage.2012.04.062

Price, C. J. (2018). The evolution of cognitive models: From neuropsychology to neuroimaging and back. *Cortex*, *107*, 37–49. doi:10.1016/j.cortex.2017.12.020

Protopapas, A., Orfanidou, E., Taylor, J. S. H., Karavasilis, E., Kapnoula, E. C., Panagiotaropoulou, G., Velonakis, G., Poulou, L. S., Smyrnis, N., & Kelekis, D. (2016). Evaluating cognitive models of visual word recognition using fMRI: Effects of lexical and sublexical variables. *NeuroImage*, *128*, 328–341. doi:10.1016/j.neuroimage.2016.01.013

Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R. P., de Berker, A., Ganguli, S., Gillon, C. J., Hafner, D., Kepecs, A., Kriegeskorte, N., Latham, P., Lindsay, G. W., Miller, K. D., Naud, R., Pack, C. C., … Kording, K. P. (2019). A deep learning framework for neuroscience. *Nature Neuroscience*, *22*(11), 1761–1770. doi:10.1038/s41593-019-0520-2

Rumelhart, D. E., & McClelland, J. L. (1982). An interactive activation model of context effects in letter perception: Part II. the contextual enhancement effect and some tests and extensions of the model. *Psychological Review*, *89*(1), 60–94. doi:10.1037/0033-295X.89.1.60

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*(3), 211–252. doi:10.1007/s11263-015-0816-y

Salmelin, R. (2007). Clinical neurophysiology of language: The MEG approach. *Clinical Neurophysiology*, *118*(2), 237–254. doi:10.1016/j.clinph.2006.07.316

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., Kar, K., Bashivan, P., Prescott-Roy, J., Schmidt, K., Yamins, D. L. K., & DiCarlo, J. J. (2018). Brain-Score: Which artificial neural network for object recognition is most brain-like? *bioRxiv*, 407007. doi:10.1101/407007

Service, E., Helenius, P., Maury, S., & Salmelin, R. (2007). Localization of syntactic and semantic brain responses using magnetoencephalography. *Journal of Cognitive Neuroscience*, *19*(7), 1193–1205. doi:10.1162/jocn.2007.19.7.1193

Shrikumar, A., Greenside, P., & Kundaje, A. (2019). Learning important features through propagating activation differences. *arXiv:1704.02685 [cs]*.

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556 [cs]*.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *07-12-June*, 1–9. doi:10.1109/CVPR.2015.7298594

Tarkiainen, A., Helenius, P., Hansen, P. C., Cornelissen, P. L., & Salmelin, R. (1999). Dynamics of letter string perception in the human occipitotemporal cortex. *Brain*, *122*(11), 2119–2132. doi:10.1093/brain/122.11.2119

Testolin, A., Stoianov, I., & Zorzi, M. (2017). Letter perception emerges from unsupervised deep learning and recycling of natural image features. *Nature Human Behaviour*, *1*(9), 657–664. doi:10.1038/s41562-017-0186-2

Vartiainen, J., Liljeström, M., Koskinen, M., Renvall, H., & Salmelin, R. (2011). Functional magnetic resonance imaging blood oxygenation level-dependent signal and magnetoencephalography evoked responses yield different neural functionality in reading. *The Journal of Neuroscience*, *31*(3), 1048–1058. doi:10.1523/jneurosci.3113-10.2011

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. doi:10.1016/j.neuron.2007.05.031

Yamins, D. L. K., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19*(3), 356–365. doi:10.1038/nn.4244