Incorporating Context into Language Encoding Models for fMRI

Aayush Bhandari (2020101116), Amal Sunny (2021121011), Pratham Gupta (2020101080)

May 10, 2023

1 Introduction

We seek to implement and to some extent, extend the paper implemented by Jain et.al. [JH18] on the role that context plays in the representations of words in the brain.

The human brain needs context to understand language. Without context, we would fail to understand words with multiple meanings, organize phrases, or identify references. Knowing this, it is safe to assume that contextual information must be stored in some form in the human cortex and one of the tools we can use to figure out these mappings are encoding models. These models use features from the given stimuli to predict brain responses recorded with fMRI. Studies before this paper successfully mapped word-level semantic representations using embedding vectors, but this approach assumes that the response to each word is independent and neglects the effect of context. To overcome this limitation, language features that incorporate context must be extracted.

A major advancement in this domain (and to overcome the limitations of hand-designed feature spaces) was the use of representations discovered by unsupervised learning. These models were trained on tasks which the brain performs naturally: classifying objects in visual scenes [ASMG14], or words and musical genres in sounds [KYS+18]. The application of pre-trained neural networks in encoding models involves utilizing stimuli from fMRI experiments to feed the networks, and treating the activations obtained from each layer as distinct feature spaces. This approach has been observed to be highly effective in modeling brain responses in the visual and auditory cortex, indicating that the networks have the ability to identify representations that are similar to those present in the human brain. However, it should be noted that in both cases, the representations at the lowest-level (images, sound spectrograms) and highest-level (image categories, words, music genres) are already known, and the networks are utilized to identify intermediate representations.

On the other hand, the situation with language is different since the highest-level representation is not yet known, rendering the supervised approach unsuitable for this domain.

As an alternative for the domain of language, there exists self-supervised networks like neural language models(LMs) which are trained to merely predict the next word in a sequence based on the previous word provided as context[BDV00]. These models typically use long short-term memory (LSTM) networks[SSN12], which have been established to discover both semantic and contextual representations of the words and sequences provided - leading to them being used in many downstream natural language processing tasks. These LSTM LMs are also useful due to them generating different kinds of representations (based on varying context length and taking output of different internal layers in the network) which can be used to prove diverse representations in the brain. To begin with, it has been established that distinct regions of the brain possess varying ”temporal receptive fields,” which reflect their sensitivity to different contextual lengths [LHSH11]. This phenomenon can be simulated in LSTM network by adjusting the number of words or ”context length” utilized in generating representations. Furthermore, the various layers utilized would also generate distinctive representations, with each layer being tested separately.

Further improvements in the field of LMs have lead to the development of Transformer models [VSP+17] such as the Bidirectional Encoder Representations from Transformers (BERT) [DCLT18], offer a promising approach to extending the concept of language encoding models for fMRI. Unlike LSTMs, which can model contextual dependencies using a fixed-length context window, transformers employ self-attention mechanisms to encode sequences of words in a more flexible and adaptive

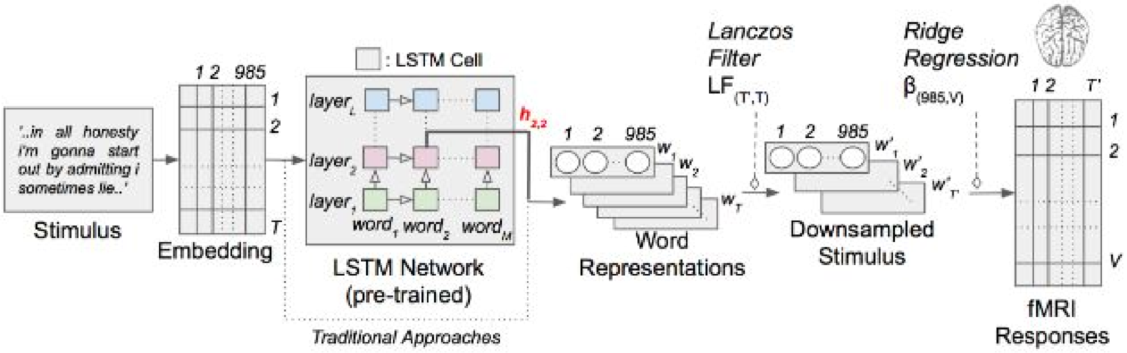

Figure 1: Contextual language encoding model with narrative stimuli. Each word in the story is first projected into a 985-dimensional embedding space. Sequences of word representations are then fed into an LSTM network that was pretrained as a language model. In this example, we extract representations for each word in the stimulus from layer 2 of the LSTM with context length 1 (i.e. considering only one word before the current word). A low-pass Lanczos filter is used to resample the contextual representations to the low temporal resolution of fMRI, and ridge regression is used to map downsampled contextual representations to fMRI responses. The dotted line indicates the path followed by the traditional, non-contextual approach that relies solely on individual word embeddings.

way. This allows transformers to capture long-range dependencies between words and generate highly-contextualized word representations. By leveraging these advantages, BERT-based language encoding models could potentially offer a more comprehensive and accurate account of the neural processes underlying language comprehension and production. Additionally, BERT's ability to learn from massive amounts of unlabeled text could enable the creation of richer and more diverse representations of language, which could facilitate the discovery of new insights in neuro-imaging research.

The authors of the original paper made two contributions. They incorporated context into encoding models that predict fMRI responses by using an LSTM LM and they compared the effectiveness of these models by varying the layer and context lengths used (Figure 1). They found their LSTM does significantly better than previously published word embedding models and revealed the effect of varying context length and layer representations on different brain regions. We extend their idea by replacing their use of an LSTM LM, with a Transformer based LM - BERT. We repeat the same analysis, but by using the embeddings generated by BERT with varying context lengths and using different layers in the network of BERT.

2 The Experiment

2.1 The Data

We use data from an fMRI experiment where the stimulus consisted of 11 naturally spoken narrative stories from The Moth Radio Hour, totalling over 2 hours or roughly 23,800 words [JH18]. Each story is transcribed and the transcripts are aligned to the audio, revealing the exact time when each word is spoken. These rich and complex stimuli are highly representative of the language that humans perceive on a daily basis, and understanding these stimuli relies heavily on contextual information. Measured brain responses consist of whole-brain blood-oxygen level dependent (BOLD) signals recorded from 6 subjects (2 female) using functional magnetic resonance imaging (fMRI), while they listened to stimuli. Images were obtained using gradient-echo EPI on a 3T Siemens TIM Trio scanner at the UC Berkeley Brain Imaging Center using a 32-channel volume coil, TR = 2.0045 seconds (yielding about 4,000 timepoints for each subject), TE = 31 ms, flip angle = 70 degrees, voxel size = 2.24 × 2.24 × 4.1 mm (slice thickness = 3.5 mm with 18% slice gap), matrix size = 100 × 100, and 30 axial slices.

2.2 Encoding Models

Encoding models aim to approximate the function f(S)-¿R that maps a stimulus S to observed brain responses R (16 ; 24 ; 10). For natural language, S is a continuous string of words $w_1, w_3, w_3 ... w_T$. n fMRI experiments, a separate encoding model $\hat{f_j}$ is typically estimated for the voxels that overlap with the *cerebral cortex* based on a training dataset, $S_t rn, R_t rn$. To evaluate model performance, the estimated model is used to predict responses in a separate testing dataset, $\hat{R_{test,j}} = \hat{f_j}(S_{test})$. Model performance for a single voxel is computed as the Pearson correlation coefficient between real and predicted responses, $r_j = corr(R_{test,j}, \hat{R_{test,j}})$.

In practice, the limited amount and quality of fMRI data cause generic nonlinear function approximators to perform poorly as encoding models. Instead, it is common to use linearized models, which assume that f is a nonlinear transformation of the stimuli into some feature space followed by a linear projection, $f(S) := g(S)_{(T,P)} \beta_{(P,V)}$, where g nonlinearly transforms S into a P -dimensional feature space, and β contains a separate set of P linear weights for each of the V voxels. The function g is typically chosen to extract stimulus features that are thought to be represented in the brain, and the weights β are learned using regularized linear regression. Here, we use ridge regression (L2-norm) to estimate weights for all encoding models.

BOLD responses are low-pass relative to the stimuli that elicited them. To account for this effect we resample and low-pass filter the stimulus representation $g(S)_{(T,P)}$ to the same rate as the fMRI acquisition with the use of a Lanczos filter, yielding $g'(S)_{(T,P)}$. Then, to account for hemodynamic delay we use a finite impulse response (FIR) model with 4 delays (2, 4, 6, and 8 seconds).

2.3 Word embeddings for language encoding

The representation of the stimuli is an extremely important question. In the study we follow, we make use of English1000, (a state of the art encoding model), to represent every word w_i as a 985-dimensional vector e_i, the representation being based on co-occurrence statistics from a large corpus of English text (similar to word2vec). This has been chosen as the baseline given the static nature of these embeddings which ignore context. Word embeddings capture the semantic similarity between words, such that words that would mean the same thing would have similar embedding vectors. The one major flaw with this model, is that it assumes each word has a brain response to it independent of other words in the stimulus. These models ignored temporal order and dependencies known to be significant for language processing in the brain [LHSH11].

3 Learning representations of context

BERT is a transformers model pretrained on a large corpus of English data in a self-supervised fashion. This means it was pretrained on the raw texts only, with no humans labeling them in any way (which is why it can use lots of publicly available data) with an automatic process to generate inputs and labels from those texts. More precisely, it was pretrained with two objectives:

- **Masked language modeling (MLM)**: taking a sentence, the model randomly masks 15% of the words in the input then run the entire masked sentence through the model and has to predict the masked words. This is different from traditional recurrent neural networks (RNNs) that usually see the words one after the other, or from autoregressive models like GPT which internally masks the future tokens. It allows the model to learn a bidirectional representation of the sentence.

- **Next sentence prediction (NSP)**: the models concatenates two masked sentences as inputs during pretraining. Sometimes they correspond to sentences that were next to each other in the original text, sometimes not. The model then has to predict if the two sentences were following each other or not.

Furthermore, the transformer encoder in BERT uses a self-attention mechanism to capture contextual information from both the left and right sides of the sequence. This allows BERT to model bidirectional context, unlike LSTMs which only capture information from the past. The self-attention mechanism computes a weighted sum of the embeddings of all the words in the sequence, with weights determined by the similarity between the word being predicted and all the other words in the sequence.

We have used the model at dimensionality 768 of the encoder layers and the pooler layer. Number of hidden layers in the Transformer encoder is equal to 12.

The model was trained on 4 cloud TPUs in Pod configuration (16 TPU chips total) for one million steps with a batch size of 256. The sequence length was limited to 128 tokens for 90% of the steps and 512 for the remaining 10%. The optimizer used is Adam with a learning rate of 1e-4, $\beta_1 = 0.9$ and $\beta_2 = 0.999$, a weight decay of 0.01, learning rate warmup for 10,000 steps and linear decay of the learning rate after.

3.1 Extracting contexual information

The BERT model uses information from up to $M = 20$ previous words to construct $L = 3$ separate context representations from three layers of the model viz, $4, 8, 12$ to abstract different levels of training information. By varying the number of words the BERT has seen before the representation is extracted, we can produce representations with different 'temporal receptive field' sizes (12). And by extracting representations from different layers we may be able to capture different types of high-level contextual representations (19). We extract $M \times L = 60$ separate representations of the stimuli, one for each context length and at each layer, by forward-propagating the stimuli through the pretrained BERT. To obtain context representations for a stimulus story $S = [w_1, ..., w_n]$ with a desired context length $m \in [0..M-1]$ and layer $l \in [1..L]$, we use the following procedure. First, for each word w_i we extract the previous m words to form a length $m+1$ sequence $[w_{i-m}, ..., w_i]$. The m context words and word w_i are sequentially fed into the pretrained BERT, and then the activations from layer 4 of BERT are extracted. The resulting 768-dimensional vector combines the information in w_i with the previous m words through recurrent connections, building an effective contextual representation for w_i. This procedure thus allows us to incorporate and experiment with varying temporal fields and contextual information in language encoding. $P(w_{i+1}|w_{i-m} \ldots w_i)$.

4 Results

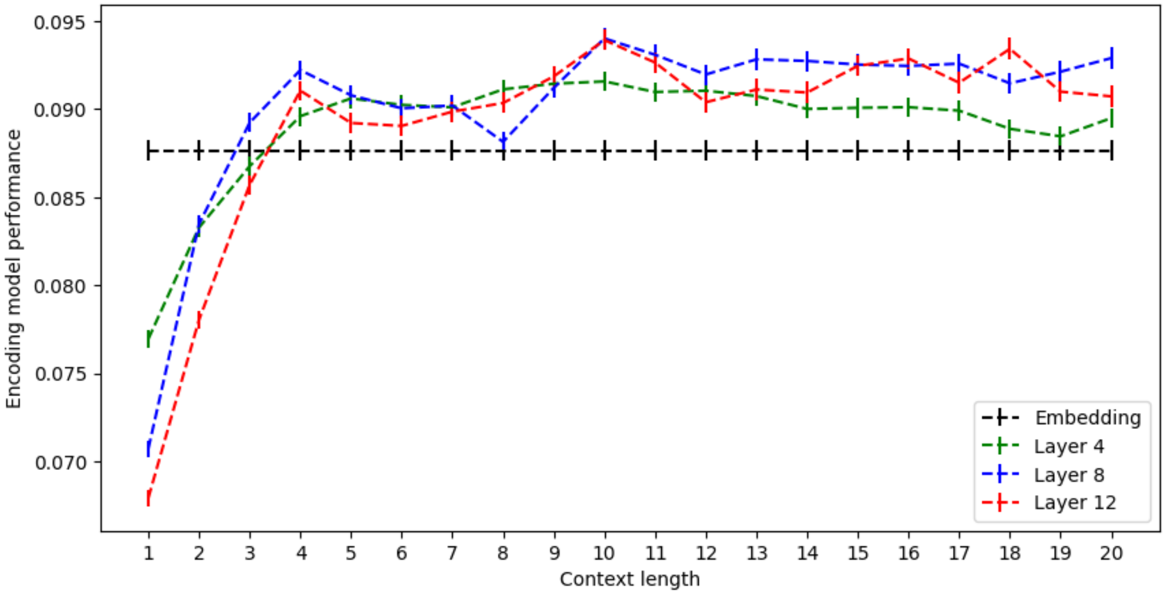

Figure 2: Contextual encoding model performance with different context lengths and layers vs. state-of-the-art embedding. Results are averaged across 6 subjects ± adjusted standard error of the mean. Contextual models from all layers outperform the embedding. Increasing context length uniformly improves performance in every layer, and different performance across layers suggests that each represents different information. Best performance is obtained using layer 2 with long context

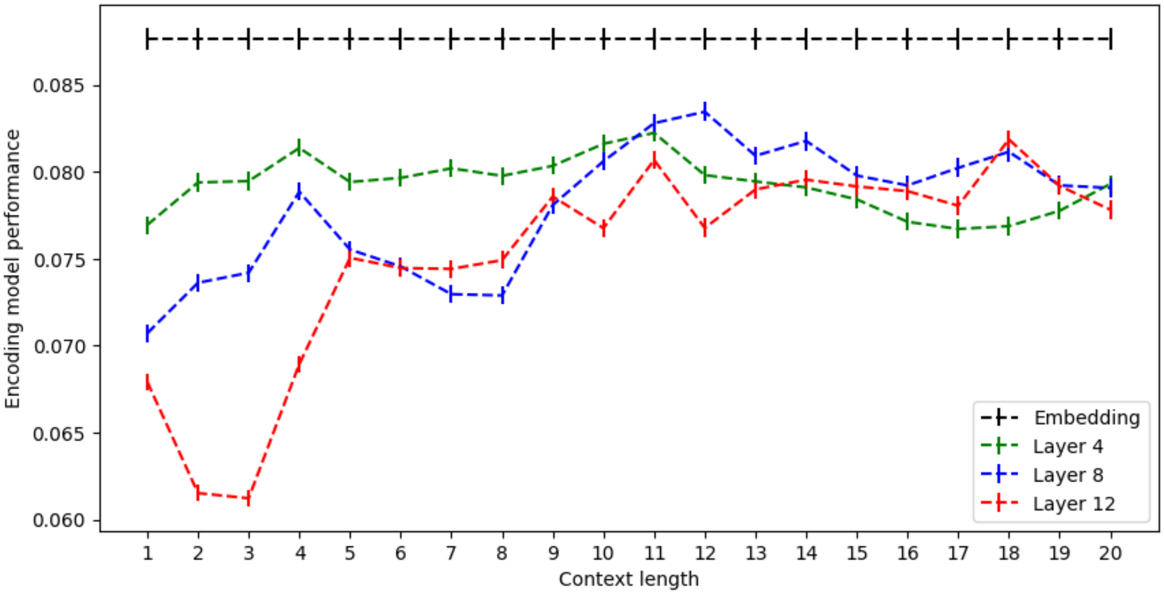

Figure 3: To show that contextual models do indeed rely on context, we swap context between different occurrences of each word and re-fit the encoding models. Swapped-context models also perform worse than the original, showing that linguistically plausible but wrong context also hurts the model.

4.1 Language encoding performance for context representations

After fitting the 60 contextual encoding models (with M = 20 different context lengths and L = 3 different layers) and a baseline embedding model from a worToVec like model (english1000), we evaluate model performance by predicting fMRI responses on a test dataset comprising one 10-minute story. Model performance in is summarized by computing mean correlations across all voxels i.e., $r_{lc} = \frac{\sum r_l c(v)}{number-of-voxels}$ for layer l and context length c Standard error of the mean (SEM) is used to compute error bars. Figure 2 shows encoding model performance for each model. models significantly beat the embedding in nearly every model variation. This supports our hypothesis that contextual representations are a better match to the human brain than static word embeddings.

4.1.1 Impact of context length

Figure 2 shows that model performance increases monotonically with context length, for all layers. This suggests the encoding models are successfully exploiting contextual information that BERT has learned to extract. However, model performance plateaus after 10-15 words, suggesting that BERT was unable to successfully incorporate contextual information beyond this timescale. This can be attributed to a higher level of overfitting of the regressor as a single word's meaning uses excessively long contexts that can be irrelevant stimulus to the brain.

The encoding model results in Figure 2 also suggest that the improvement of our model can be partially explained by BERT simply learning a better word embedding than the baseline. Models with zero context length use no contextual information and thus are simply nonlinear transformations of the input embedding. The model for layer 1 with zero context is substantially better than the embedding, suggesting that layer 1 learns to generate a new embedding that is a better match to the brain. However, the best model for each layer uses the longest context, suggesting that contextual information is important above and beyond the improved embedding.

4.1.2 Impact of LSTM layers.

Figure 2 also shows that there are differences between models that use different BERT layers. All the layers perform unsatisfactorily on short contexts but do reasonably well as compared to the baseline after the context length of about 4. Further, we observe that Layer 4 predicts better than other layers with short context, but only improves modestly with context length. Layer 8 and Layer 12 predict

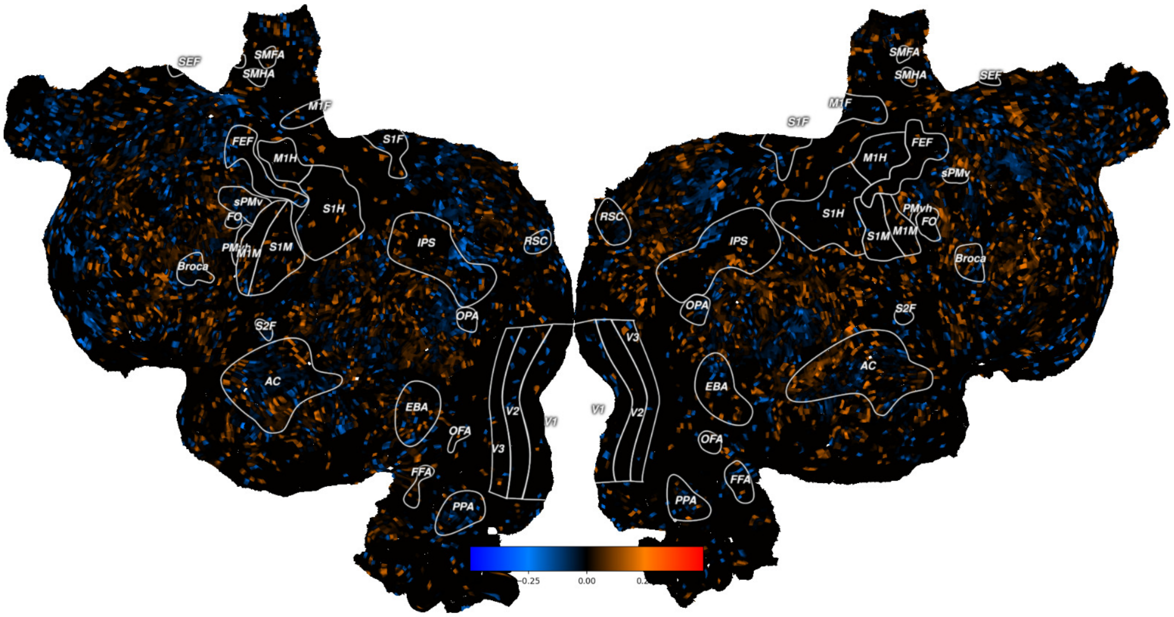

Figure 4: Difference between contextual and embedding model performance across cortex. The difference in r values is computed for each voxel in one subject (S1) and projected onto the subject's flattened cortical surface. White outlines show ROIs identified using other experiments. . Red voxels are best modeled by the contextual model and blue are best modeled by the embedding. Voxels poorly predicted by both models appear black. Overall, the contextual model performs better, although some voxels have approximately equal performance with both models

worse than the embedding with short context, but see large improvement with context length. This could be attributed to the model unable to provide proper representations for the word due to the small context space provided, leading to attention being misattributed due to lacking full context. We see it being rectified in larger contexts, providing more evidence for this. The layers lacking much difference can be attributed to the architecture of the transformer model and the relative good performance leaving little space for much difference, and the layers picking up useful representations early on and refining not leading to much benefit for encoding tasks. There is some benefit with respect to context between layers (layer 4 vs the rest), but still leaves layer 4 with good performance above the baseline.

4.2 Distorting context representations

Figure 2 shows that BERT produces useful representations for encoding. However, it also shows that part of the improvement is due to BERT learning better embeddings, as can be seen by the good performance of layer 4 with zero context. To directly test the importance of context in these models, we conduct an experiment where we degrade the quality of the context by 'swapping' the actual context with the context of another occurrence of the same word (e.g. "this is my dog" → "I saw the dog") for each stimulus word and context length. Swapping is done after inferring context vectors using BERT but before regression on the fMRI. For each context length, we pick the new context from that of the word that appears next in a simple search through story. Thus, the swapped contexts picked for each successive length are sometimes different, leading to the increased variance across context lengths. The results in Figure 3 show that linguistically plausible but wrong contexts also greatly impact encoding performance. The encoders built from BERT now perform worse than even the baseline model which demonstrates effectively that the context model is picking up information relevant for cortical representations.

4.3 Model preference across cortex

Here, we use the model we've trained above to investigate the representation of brain regions corresponding to the encoding of contextual information. We analyze across both context lengths and layer

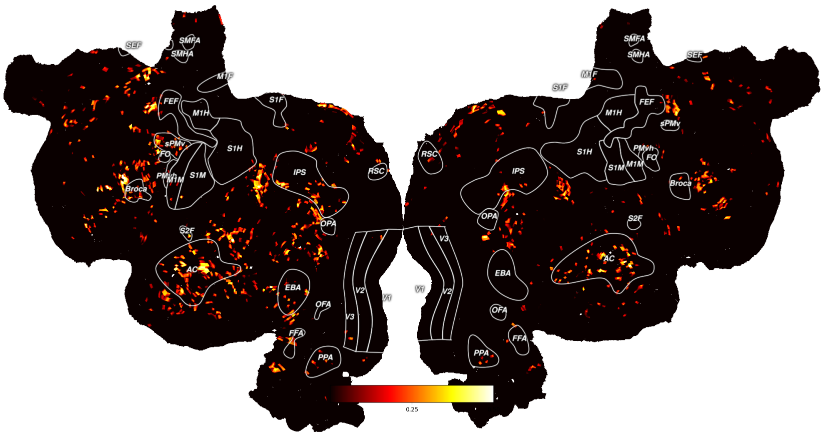

Figure 5: Projection of correlation scores on the brain from BERT encodings for context = 4

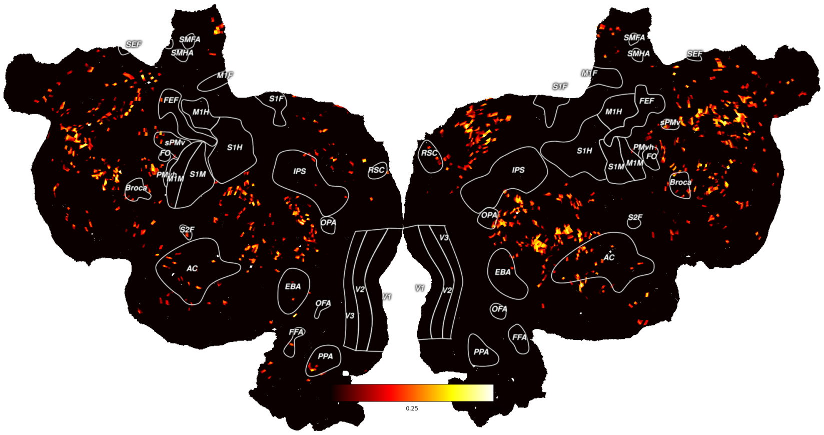

Figure 6: Projection of correlation scores on the brain from BERT encodings for context = 10

preferences across the cortex, along with a comparison to the baseline model (Figure 4.

4.3.1 Context length preference across cortex

While longer context is providing better encoding performance in the models we evaluated, the same cannot be said for every brain region. To investigate which regions correspond to it, we compute context length. To analyse the dependency on context length, we modelled the brain with the correlation values of the best performing layer, which in this case was layer 8, and found the maximum . Basically, we compute a context length preference index separately for each voxel. This index is defined as the projection of each voxel's 'context profile', a length-M vector containing prediction performance for each context length with the best in the layer. This index was selected because it accounts for the fact that performance plateaus for context lengths around range of 10 to 20 and it avoids computing a noisy argmax across context lengths.

We observe in our study that certain regions like the Auditory cortex and the Synaptosomal plasma membrane vesicles decrease in correlation with an increase in context length which could hint to a preference for smaller clusters of words rather than long sentences.

Figure 5, 6, 7 shows the context length preference for each voxel in one subject.

4.3.2 BERT layer preference across cortex

Earlier experiments found strong correspondence between layers of deep supervised networks and brain areas in the visual and auditory cortex. We modelled the encoding for the three layers to be compared and analysed the regions of correlation (Figure 8 , 9, 10). We observe that layer 4 seems to show the

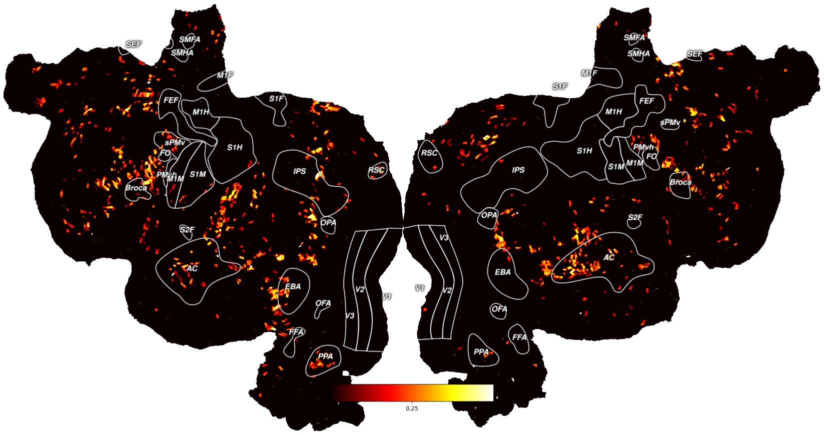

Figure 7: Projection of correlation scores on the brain from BERT encodings for context = 15

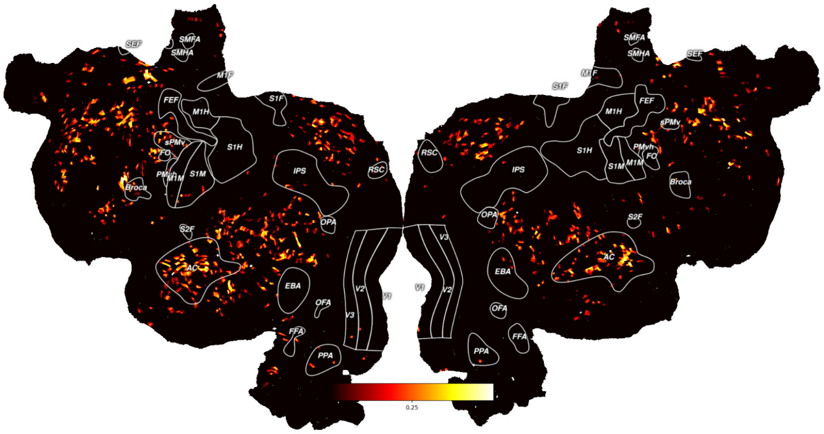

Figure 8: Layer 4 BERT projections of correlation values on the brain

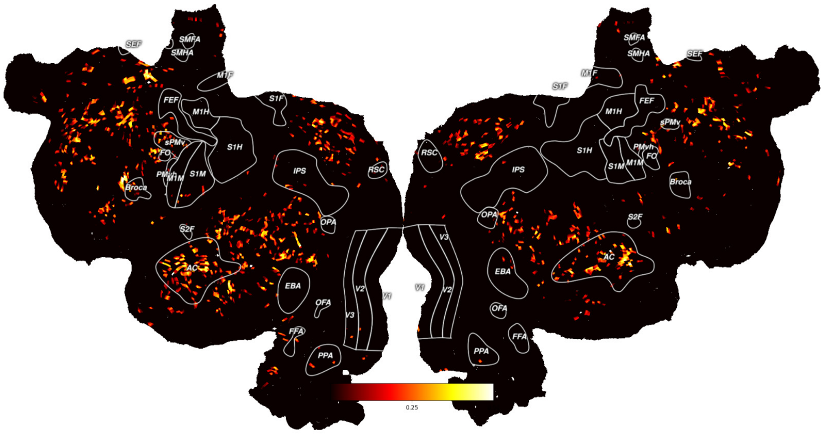

Figure 9: Layer 8 BERT projections of correlation values on the brain

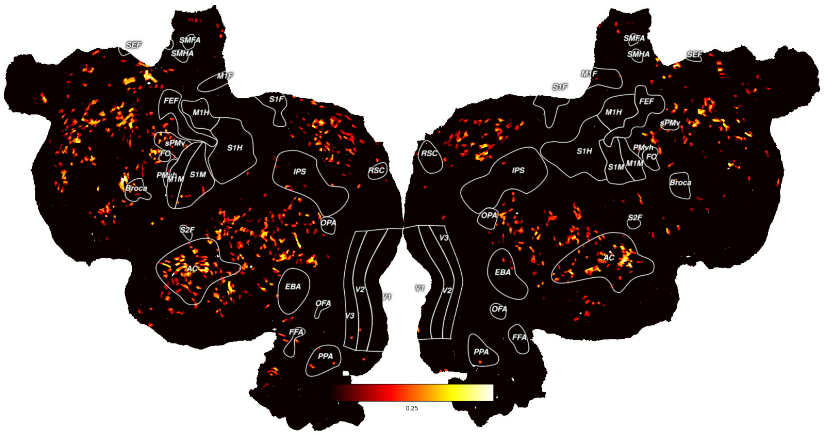

Figure 10: Layer 12 BERT projections of correlation values on the brain

minimum correlation with the brain regions (when compared to the other 2). Overall, the layers have similar correlation maps with only slight increases as you go higher up the layers. All three layers seem to show a clustering of correlation around the Auditory Cortex (AC) and the Synaptosomal plasma membrane vesicles (SPMV) with layer 12 and 4 showing the most clustered correlation in the areas, perhaps hinting at a preference for these two layers. The AC is also highly correlated with the lower level learning (as well as the semantic embeddings), which could hint at layers 4 and 12 being associated with learning lower level representations while layer 8 leans towards the higher level representations. There's also the fact that the Broca region, increasing correlation with increase in layer index.

5 Conclusions

This study demonstrates the successful integration of context representations into language using the BERT model. The results indicate that the BERT's representations outperform current embedding-based models and exhibit unique characteristics depending on the context length and layer. The findings suggest that these models can capture context and temporal order, although the mechanisms may differ across layers. Additionally, the study illustrates how these models can account for differences in language processing across various cortical regions, including both basic and advanced language areas.

References

[ASMG14] Pulkit Agrawal, Dustin Stansbury, Jitendra Malik, and Jack L Gallant. Pixels to voxels: modeling visual representation in the human brain. *arXiv preprint arXiv:1407.5104*, 2014.

[BDV00] Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. A neural probabilistic language model. *Advances in neural information processing systems*, 13, 2000.

[DCLT18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[JH18] Shailee Jain and Alexander Huth. Incorporating context into language encoding models for fmri. *Advances in neural information processing systems*, 31, 2018.

[KYS+18] Alexander JE Kell, Daniel LK Yamins, Erica N Shook, Sam V Norman-Haignere, and Josh H McDermott. A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, 98(3):630–644, 2018.

[LHSH11] Yulia Lerner, Christopher J Honey, Lauren J Silbert, and Uri Hasson. Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *Journal of Neuroscience*, 31(8):2906–2915, 2011.

[SSN12] Martin Sundermeyer, Ralf Schlüter, and Hermann Ney. Lstm neural networks for language modeling. In *Thirteenth annual conference of the international speech communication association*, 2012.

[VSP^{+}17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.