Long Short-Term Memory of Language Models for Predicting Brain Activation During Listening to Stories

Subba Reddy Oota, Frederic Alexandre, Xavier Hinaut

Long Short-Term Memory of Language Models for Predicting Brain Activation During Listening to Stories

Subba Reddy Oota
(subba-reddy.oota@inria.fr)
Inria Bordeaux, France

Frederic Alexandre
(frederic.alexandre@inria.fr)
Inria Bordeaux, France

Xavier Hinaut
(xavier.hinaut@inria.fr)
Inria Bordeaux, France

Abstract

Several popular sequence-based and pretrained language models have been found to be successful for text-driven prediction of brain activations. However, these models still lack long-term memory plausibility (i.e. how they deal with long-term dependencies and contextual information) as well as insights on the underlying neural substrate mechanisms. This paper studies the influence of context representations of different language models such as sequence-based models: Long Short-Term Memory networks (LSTMs) and ELMo, and a pretrained Transformer language model (Longformer). In particular, we study how the internal hidden representations align with the brain activity observed via fMRI when the subjects listen to several narrative stories. We use brain imaging recordings of subjects listening to narrative stories to interpret word and sequence embeddings. We further investigate how the representations of language models layers reveal better semantic context during listening. Experiments across all language model representations provide the following cognitive insights: (i) the representations of LSTM cell states are better aligned with brain recordings than LSTM hidden states, the cell state activity can represent more long-term information, (ii) the representations of ELMo and Longformer display a good predictive performance across brain regions for listening stimuli; (iii) Posterior Medial Cortex (PMC), Temporo-Parieto-Occipital junction (TPOJ), and Dorsal Frontal Lobe (DFL) have higher correlation versus Early Auditory (EAC) and Auditory Association Cortex (AAC).

Keywords: Brain Encoding; Linear Mapping; fMRI; LSTM; ELMo; Longformer; Transformer;

Introduction

In the past decade, artificial neural networks have witnessed a remarkable insights in the computational neuroscience community in understanding how the brain performs stimulus perception (1) given various forms of sensory inputs like visual processing in object recognition tasks (Yamins et al., 2014; Cadieu et al., 2014; Eickenberg, Gramfort, Varoquaux, & Thirion, 2017), or (ii) by studying higher-level cognition like language processing (Gauthier & Levy, 2019; Schrimpf et al., 2021; Schwartz, Toneva, & Wehbe, 2019). This line of work, namely brain encoding, aims at constructing neural brain activity given an input stimulus.

Sentence comprehension is studied using fMRI since a while (Constable et al., 2004), Some studies have looked into the modelling of language comprehension: e.g. how sequence-based language models such as echo-state networks (ESN) (Hinaut & Dominey, 2013) or long short-term memory networks (LSTM) (Jain & Huth, 2018; Variengien & Hinaut, 2020) encode syntactic structures and contextual information. Morevoer, (Jain & Huth, 2018) used LSTMs to get the context representation of sentences (with a next word prediction task) and then used this representation to predict fMRI data.

Some works studied the LSTM capacity of representing long-term information (Karpathy, Johnson, & Fei-Fei, 2015) and its ability to model working memory (O'Reilly & Frank, 2006), there still lacks investigation of the long-term memory cognitive plausibility of LSTM and its link to fMRI data. In this paper, we open the black box of LSTM to look at particular LSTM activations: the cell state and the hidden state. This can give more insights on longer-term and shorter-term information. Indeed, the *cell state* mechanism has been introduced in the original LSTM paper (Hochreiter & Schmidhuber, 1997) in order to keep the error gradient of backpropagation constant over long-time scales. Thus, its activity can represent more long-term information than the *hidden state* of the LSTM. We also investigate how the pretrained bi-directional sequence embedding language model ELMo (Peters et al., 2018) handles the longer context and interprets the LSTM layers representations that better predict brain activity.

Recently, the researchers studied how the representations from Transformer (Vaswani et al., 2017) based language models such as BERT (Devlin, Chang, Lee, & Toutanova, 2019) and RoBERTa (Liu et al., 2019) could directly predict fMRI data. Interestingly, such Transformer-based neural representations have been found to be very effective for brain encoding as well (Schrimpf et al., 2021). On the other hand, (Gauthier & Levy, 2019) fine-tune a pretrained BERT model on multiple natural language processing tasks to find tasks best correlated with high *decoding* performance. In recent works, (Caucheteux, Gramfort, & King, 2021a; Antonello, Turek, Vo, & Huth, 2021) interpret the representations of the Transformer model (GPT-2 (Radford et al., 2019)) by disentangling the high-dimensional Transformer representations of language models into four combinatorial classes: lexical, compositional, syntactic, and semantic representations to explore which class is highly associated with language cortical ROIs. However, these models are unable to handle the long-term dependencies (sequence length is fixed to 512 words) due to their self-attention operation. To overcome this limitation, recently, (Beltagy, Peters, & Cohan, 2020) introduced *Longformer* making it easy to process documents of thousands of tokens or longer and combining local windowed attention with global attention.

In this paper, we uncover insights about the association between fMRI voxel activations and representations of diverse language models: LSTM, ELMo, and Longformer. The predictive power of language model specific representations with brain activation is ascertained by (1) using ridge regression on such representations and predicting activations and (2) computing popular metrics like 2V2 accuracy and Pearson correlation between actual and predicted activations.

Specifically, we make the following contributions in this paper. (1) Given a language model pretrained on corpora by handling long-term dependencies, we propose the problem of finding which of these are the most predictive of fMRI brain activity for listening tasks. (2) The investigation of long-term context of language model results reveals that ELMo and Longformer representations display better correlation during narrative story listening. (3) We also investigate the internal memory representations of LSTM (cell state and hidden state) and derive interesting insights that the cell state representations yield better performance than hidden state representations.

Methodology

Brain Imaging Dataset

Narratives-Pieman (Listening to Stories) The "Narratives" collection aggregates a variety of fMRI datasets collected while human subjects listened to naturalistic spoken stories. The Narratives dataset that includes 345 subjects, 891 functional scans, and 27 diverse stories of varying duration totaling ∼4.6 hours of unique stimuli (∼43,000 words) was proposed in (Nastase et al., 2021). Similar to earlier works (Caucheteux, Gramfort, & King, 2021b), we analyze data from 82 subjects listening to the story titled 'PieMan' with 259 TRs (repetition time)[1]. A TR is the length of time between corresponding consecutive points in fMRI: here it is 1.5 sec. We list number of voxels per ROI (Region of Interest) in this dataset in Table 1. We use the multi-modal parcellation of the human cerebral cortex (Glasser Atlas: consists of 180 ROIs in each hemisphere) to display the brain maps (Glasser et al., 2016), since the Narratives dataset contains annotations tied to this atlas. The data covers ten brain ROIs, i.e., Left hemisphere (L), and Right hemisphere (R) for each of the following: (i) early auditory cortex (EAC: A1, LBelt, MBelt, PBelt, and R1) which plays a key role for sound perception since it represents one of the first cortical processing stations for sounds; (ii) auditory association cortex (AAC: A4, A5, STSdp, STSda, STSvp, STSva, STGa, and TA2) which is concerned with the memory and classification of sounds; (iii) posterior medial cortex (PMC: POS1, POS2, v23ab, d23ab, 31pv, 31pd, 7m) which has been implicated in tasks as diverse as attention, memory, spatial navigation, emotion, self-relevance detection, and reward evaluation; (iv) the temporo parieto occipital junction (TPOJ: TPOJ1, TPOJ2, TPOJ3, STV, PSL) which is a complex brain territory heavily involved in several high-level neurological functions, such as language, visuo-spatial recognition, writing, reading, symbol processing, calculation, self-processing, working memory, musical memory, and face and object recognition; and (v) the dorsal frontal lobe (DFL: L_55b, SFL, L_44, L_45, IFJA, IFSP) which covers the aspects of pragmatic processing such as discourse management, integration of prosody, interpretation of nonliteral meanings, inference making, ambiguity resolution, and error repair. These five brain ROIs (EAC, AAC, TPOJ, DFL, and PMC) span a cortical hierarchy supporting language and narrative comprehension (Huth, De Heer, Griffiths, Theunissen, & Gallant, 2016; Baldassano et al., 2017).

Table 1: # Voxels in each ROI in the Narratives Dataset. LH - Left Hemisphere. RH - Right Hemisphere. Pieman has 82 subjects.

ROIs→	EAC		AAC		PMC		TPOJ		DFL	
	LH	RH	LH	RH	LH	RH	LH	RH	LH	RH
# Voxels	808	638	1420	1493	1198	1204	847	1188	1061	875

Encoding Models

To explore how and where contextual language features are represented in the brain when listening to stories, we extract internal hidden representations from two sequence-based models: Random LSTM and LSTM, ELMo (obtaining context-dependent word embeddings), and popular pretrained Transformer language model (Longformer) used for describing each stimulus sentence and use them in an encoding model to predict brain responses. Our main objective is to compare the correlation between each model dense hidden representations and human cognitive process. In this paper, we train fMRI encoding models using Ridge regression on stimuli representations obtained using four models: Random LSTM, LSTM, ELMo, and Longformer. The main goal of each fMRI encoder model is to predict brain responses associated with each brain region given stimuli. In all cases, we train a ridge regression model per subject separately. Following the literature on brain encoding (Caucheteux et al., 2021b; Toneva, Stretcu, Póczos, Wehbe, & Mitchell, 2020), we choose to use a ridge regression model instead of more complex models. For instance, (Affolter, Egressy, Pascual, & Wattenhofer, 2020) used a neural network-based model that directly maps fMRI-to-word in a decoding setup. We plan to explore more such models as part of future work in brain encoding. Here, our main is objective to investigate the influence of context representations of different language models such as sequence-based models: LSTM and ELMo, and popular pretrained Transformer language model (Longformer).

LSTM

First, we train an LSTM (Hochreiter & Schmidhuber, 1997) network to predict the probability of the next word as a function of the history of previous words. The weights of LSTMs are learned using the error back-propagation through time

[1] 282 TRs (before preprocessing) and 259 TRs (after preprocessing).

(BPTT). To make the association between encoded stimuli from LSTM's internal components and fMRI brain activity, we do the folllowing: (i) At the time step t, we use vector a$_t$ to represent the internal neurons of encoded stimuli in LSTM. In this paper, a$_t$ may be hidden state vector (h$_t$) and cell state vector (c$_t$). (ii) In order to map the stimuli encoded vector a$_t$ of LSTM and brain activity at the t-th time step (Y_t), we define a simple linear model, ridge regression, to predict the brain activity ($\hat{Y}_t$) from a$_t$, as discussed in the ridge regression section.

Pretrained text Transformer: Longformer

Longformer (Beltagy et al., 2020) builds on BERT's language masking strategy and supports long document generative sequence-to-sequence tasks. We use the pretrained Longformer model with a local attention mechanism, where the default window size is set to 5. To obtain the stimuli representation, we use the last layer token representations where each token dimension is 768.

Linear Probing of Language Models

Here, we do not train Random LSTM, LSTM, ELMo, and Longformer networks to directly predict brain activities, for two reasons. First, the dimension of the fMRI voxels varies among different subjects and across different ROIs. Therefore, it is not convenient to design a universal neural network architecture for generating outputs of different dimensions. Second, the goal of this research is not to improve the performance of language models in predicting fMRI. We want to explore linear mappings between particular features of language models states and neural activities in the auditory and language brain ROIs. Namely, we look at (i) the characteristics of hidden state vectors (h$_t$) and artificial memory vectors (c$_t$) in both LSTMs and Random LSTMs, and (ii) local context vectors obtained from performance-optimized deep neural network models (ELMo and Longformer). Therefore, we avoid any possible supervision from the fMRI data when training LSTM and Random LSTM language models.

Ridge Regression

We trained a ridge regression based encoding model to predict the fMRI brain activity associated with the semantic vector representation obtained from each language model: randLSTM (hidden state, cell state), LSTM (hidden state, cell state), GloVe, ELMo, and Longformer. Each voxel value is predicted using a separate ridge regression model. Formally, at the time step (t), we encode the stimuli as $X_t \in \mathbb{R}^{N \times D}$ and brain region voxels $Y_t \in \mathbb{R}^{N \times V}$, where N denotes the number of training examples, D denotes the dimension of input stimuli representation, and V denotes the number of voxels in a particular region.

Hyper-parameter Setting: We used sklearn's ridge-regression with default parameters, 5-fold cross-validation, Stochastic-Average-Gradient Descent Optimizer, Huggingface for Longformer, MSE loss function, and L2-decay (λ):1.0. We used Word-Piece tokenizer for the Longformer model and Spacy-tokenizer for the GloVe and ELMo models.

Model Prediction Across Whole Brain

To determine the significant voxel predictions across the whole-brain, we ran the permutation tests where we shuffled the true responses 5000 times, computed the Pearson correlation scores, and finally obtained the FDR corrected p-values for the whole brain results using both Longformer and ELMo. We set the correlation score of voxels to zero if the p-value of the correlation obtained from the permutation test is above the significance threshold ($p > 0.05$, FDR corrected).

Evaluation Metrics

We evaluate our models using popular brain encoding evaluation metrics described in the following. Given a subject and a brain region, let N be the number of samples. Let $\{Y_i\}_{i=1}^{N}$ and $\{\hat{Y}_i\}_{i=1}^{N}$ denote the actual and predicted voxel value vectors for the i^{th} sample. Thus, $Y \in R^{N \times V}$ and $\hat{Y} \in R^{N \times V}$ where V is the number of voxels in that region.

2V2 Accuracy is computed as follows.

$$\begin{aligned} &2\text{V2Acc} = \\ &\frac{1}{N_{C_2}} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} I[\{cosD(Y_i,\hat{Y}_i) + cosD(Y_j,\hat{Y}_j)\} \\ &< \{cosD(Y_i,\hat{Y}_j) + cosD(Y_j,\hat{Y}_i)\}] \end{aligned}$$

where $cosD$ is the cosine distance function. $I[c]$ is an indicator function such that $I[c] = 1$ if c is true, else it is 0. The higher the 2V2 accuracy, the better.

Pearson Correlation (PC) is computed as PC=$\frac{1}{N}\sum_{i=1}^{n} corr[Y_i,\hat{Y}_i]$ where corr is the correlation function.

Comparison to Other Language Models

We compare the LSTM model with the Longformer and several other pretrained language models: Random LSTM, ELMo, and GloVe. We encoded the number of words same for both ELMo and Longformer to match the performance.

LSTM Training: We experimented with one layer of LSTM to perform the next word prediction. In our next word prediction, we first split each story in half (of 27 stories); we designate the first half as the training set and the second half as the test set. The model is implemented in Keras with TensorFlow backend (Abadi et al., 2016) with cross-entropy as loss, Adam optimizer (Kingma & Ba, 2014), the number epochs set to 100, the batch size is of 64, applied dropout with a keep-probability of 0.2, and tried LSTM with hidden state size is set to 100, the dimensionality of word embeddings is set to 100. The other hyper-parameters are learning rate (0.01), and maximum sequence length is set to 5.

Random LSTM: We use a random LSTM model where the LSTM neurons are randomly initialized and kept frozen. We use the output and cell state vectors at each time step to perform fMRI encoding. The configuration details of Random LSTM are the same as that original LSTM model.

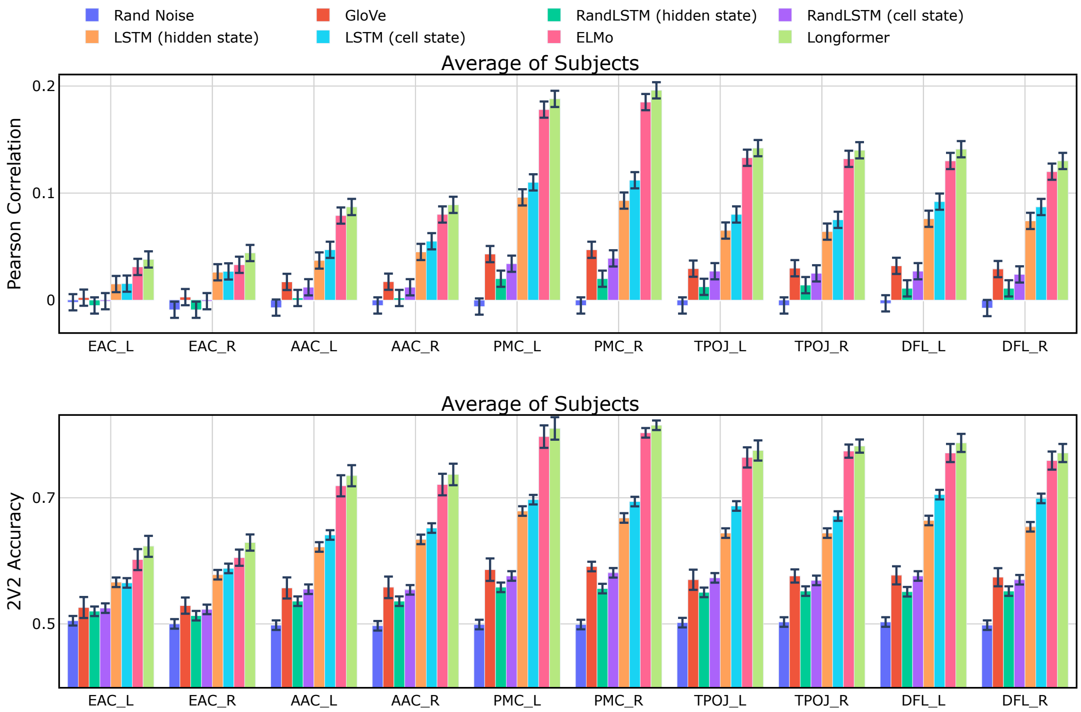

Figure 1: Pearson correlation coefficient (top figure) and 2V2 Accuracy (bottom figure) between predicted and true responses across different brain regions using a variety of language models (for Narratives-Pieman dataset). Results are averaged across all participants. ELMo and Longformer are the best. *Rand Noise* stands for a "Random noise vector". *(hidden state)* stands for the "short-term memory of internal state of the LSTM".

ELMo: ELMo (Embeddings from Language Models) is a successful NLP framework developed by the AllenNLP (Peters et al., 2018) group. Unlike earlier embeddings, ELMo embeddings represent words in a contextual fashion using a bidirectional LSTM model. We perform the average of word embeddings at each TR (1.5sec.) to obtain the contextual representation.

GloVe: based word vectors (each word is a 300-dimension vector) (Pennington, Socher, & Manning, 2014), and the average of word embeddings results in context vector in each time step.

Results

In order to assess the performance of the fMRI encoder models learned using the representations from a variety of language models, we computed the 2V2 accuracy and Pearson correlation coefficient between the predicted and true responses across various ROIs for the listening (Narratives-Pieman) dataset (Fig. 1).

Encoding performance of Language Models

From Fig. 1, we observe that the profiles of performance show low scores in the early auditory cortex (EAC) and auditory association cortex (AAC); average scores in TPOJ and DFL; and superior scores in PMC. This aligns with the known language hierarchy for spoken language understanding (Huth et al., 2016; Baldassano et al., 2017; Nastase, Liu, Hillman, Norman, & Hasson, 2020). Language models ELMo, and Longformer yield better performance in predicting the brain responses than the LSTM model across all the ROIs. These Pearson correlation (ρ) results are comparatively much higher compared to those obtained using the pretrained GPT2 model in (Caucheteux et al., 2021a) (ρ ranging from $0.02-0.06$). As shown in Fig. 1, our method obtains more than 3 times higher correlations (ρ ranging from $0.02-0.19$)[2]. The main reason is that the Longformer is designed to process documents of thousands of tokens or longer sequences while GPT-2 models are unable to handle the long-term dependencies (sequence length is fixed to 512 words). Also, the narrative dataset consists of longer documents (more than 2000 words in one story); the traditional transformer models consider the context up to 512 words, whereas Longformer handles even longer documents.

Further, from Fig. 1, we see that the bilateral posterior medial cortex (PMC) associated with higher language function exhibits a higher correlation among all the brain ROIs. ROIs, including bilateral TPOJ and bilateral DFL, yield higher correlations with the ELMo and Longformer, which is in line with the language processing hierarchy in the human brain. Finally, across all regions, Rand Noise vector and Rand LSTM models have worse correlation compared to LSTM and other language models. In summary, different and dis-

[2] we do not apply on the same number of subjects and/or same amount of stories than in (Caucheteux et al., 2021a). However, we tested with few other stories such as Lucy and Slumlord, and our results (higher correlations) show similar trends.

Table 2: p-values obtained using *post hoc* pairwise comparisons for the three best models (+ LSTM hidden state).

Models compared	EAC	AAC	PMC	TPOJ	DFL
Longformer vs ELMo	0.521	0.271	0.168	0.054	0.356
LSTM (Cell state vs hidden state)	0.991	0.177	0.0357*	0.0038*	0.158
Longformer vs LSTM (cell state)	0.048*	0.0008*	0.00002*	0.00003*	0.0015*
ELMo vs LSTM (cell state)	0.372	0.003*	0.00004*	0.00006*	0.0049*

tinct language model features seem to be related to the encoding performance in listening tasks.

In order to estimate the statistical significance of the performance differences, we performed one-way ANOVA on the mean correlation values for the subjects across the language models (GloVe, LSTM (cell state), LSTM (hidden state), ELMo, and Longformer) for the five brain ROIs. The main effect of the ANOVA test was significant for all the ROIs with p$\leq 10^{-2}$ with confidence 95%. Further, *post hoc* pairwise comparisons (Ruxton & Beauchamp, 2008) confirmed the visual observations that on both 2V2 accuracy and Pearson correlation measures, tasks such as ELMo and Longformer performed significantly better compared to other models, as shown in Table 2.

LSTM: Effects of Hidden State vs Cell State Vectors

In order to explore how LSTM hidden units learn to encode the long-term and short-term memory information and the interaction between the two types of working memories, we compare the encoding performance between representations of hidden state and cell state vectors. Fig. 1 showcases the fMRI encoding performance of both RandLSTM and LSTM models where the cell state representations (long term-memory vector) yield better performance than hidden state representations (short-term memory). This supports the cognitive plausibility of the LSTM cell architecture. Besides, the performance of GloVe and RandLSTM models have significantly equal performance, indicating that semantic context is missing in these models.

Which ELMo layers perform better encoding?

We investigate how the performance of ELMo changes at different layers (Embedding layer, LSTM layer-1, and LSTM layer-2), as they are provided in different contexts. The results are shown in Fig. 2. From Fig. 2, we observe that layer-2 displays better 2v2 accuracy and Pearson correlation score compared to other layers. We further observe that the layer 1 show a sharp increase in performance compared to embedding layer in the context of narrative story listening.

Which Longformer layers perform better encoding?

Given the hierarchical processing of language information across the Transformer layers, we further examine how these Transformer layers encode fMRI brain activity using encoder layers of Longformer. We present the layer-wise encoding performance results across brain ROIs in Fig. 3. We observe that in all the layers, intermediate layers (6 to 8) perform the best for narrative listening, followed by a decrease in performance.

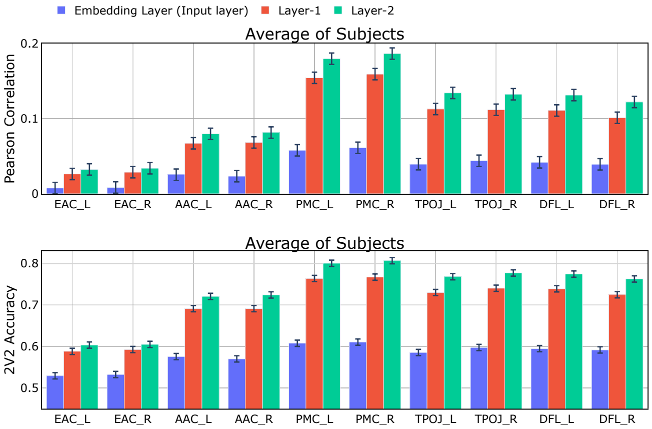

Figure 2: ELMo layers: Pearson correlation coefficient (top) and 2V2 Accuracy (bottom) between predicted and true responses across different brain regions using layers of ELMo model. Results are averaged across all participants.

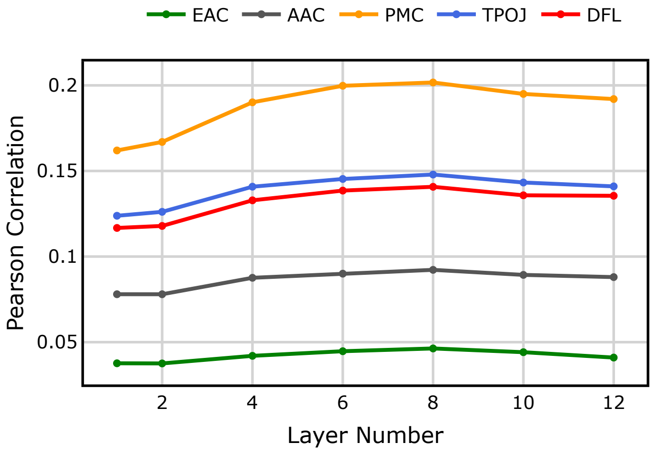

Figure 3: Longformer layers: Pearson correlation coefficient between predicted and true responses across different brain regions (different color lines) using Longformer. Results are averaged across all participants. Middle layers (6 and 8) show best correlation.

Cognitive Insights

We further analyse in more detail the prediction performance of the encoder model trained on sub ROIs for the ELMo and Longformer in Fig. 4. In the *EAC*, the sub ROI *pbelt (parabelt)* display higher Pearson correlation among other sub ROIs, and it is adjacent to the lateral belt on the exposed surface of the superior temporal gyrus (*STG*). Also, the *pbelt* area represent the next level in the auditory hierarchy and mainly concerned with memory or decision-making. Similarly in the *AAC*, the sub ROIs such as *A4*, *A5*, *stsvp*, and *stsda* yield better correlation, and these ROIs shares the primary medial and posterior borders with *TPOJ* (Trumpp, Kliese,

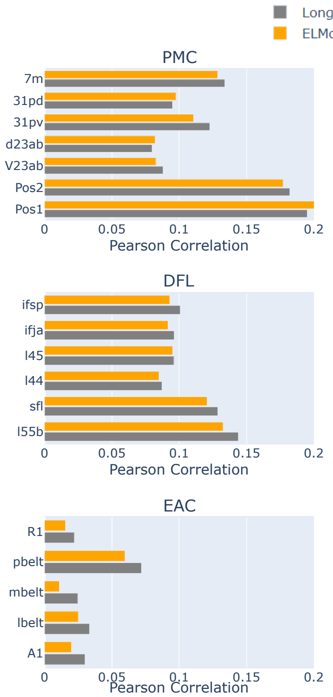

Figure 4: Pearson correlation coefficient between predicted and true responses across different sub ROIs of the Language Network using ELMo and Longformer. Results are averaged across all participants.

Hoenig, Haarmeier, & Kiefer, 2013). Further, there is evidence that these sub ROIs of the AAC process perceptual and conceptual acoustic sounds during auditory stories and social interaction tasks (Glasser et al., 2016). It can be observed that sub ROIs such as *Pos1* and *Pos2* have a higher Pearson correlation than other sub ROIs of the *PMC* region. Both *sfl* and *l55b* display a higher correlation among all the sub ROIs for the *DFL* ROI. However, all the sub ROIs in the *TPOJ* yield higher correlation, as shown in Fig. 4. The control and attention ROIs in the posterior cingulate cortex (for ex., *POS1* in *PMC*), together with the superior frontal language region (*sfl* in *DFL*) and *TPOJ*, are part of the language network associated with narrative comprehension (Nastase et al., 2020): it is encouraging to see that both ELMo and Longformer also relate to semantic analysis of the ongoing narrative because they obtain best performance, showing that capturing longer-term context is important.

Brain maps for whole brain predictions

The whole-brain prediction Pearson correlation for all the voxels using ELMo and Longformer is shown in Fig. 5. In the **listening task**, we observe from Fig. 5 that Longformer displays higher correlation values for many voxels than ELMo. From Fig. 5, we see that ROIs such as EAC and AAC have a lower percentage of voxels with a higher correlation compared to PMC and TPOJ brain ROIs (higher percentage of voxels with higher correlation).

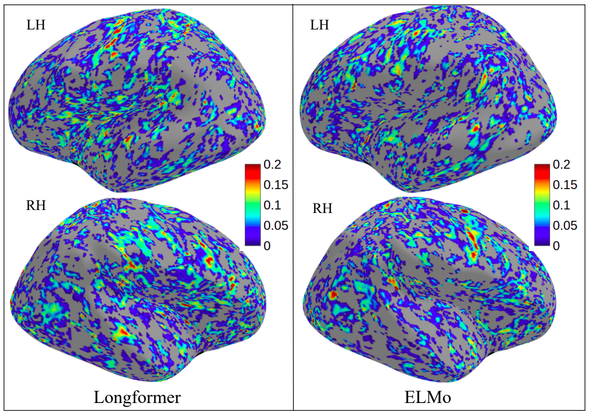

Figure 5: BrainMaps: Whole-brain prediction correlation using representations of Longformer (left) and ELMo (right) in one sample subject (subject 1) of Narratives-Pieman dataset.

Discussion

(1) We used a ridge regression model instead of more complicated models for encoding; (2) We experimented with several language models where the pretrained context representations (such as in ELMo and Longformer) are better predictors of voxel activations. (3) In LSTM, the cell state representations (long term memory vector) yield better encoding performance than hidden state representations; thus, internal dynamics of LSTMs seem to have more cognitively plausible activations than classically studied LSTM activations. (4) We used different layers of ELMo and Longformer, where higher layers display better correlation for ELMo while intermediate layers show superior performance for Longformer. (5) The control and attention ROIs in the posterior cingulate cortex, together with the superior frontal language region (sfl in DFL) and TPOJ, are part of the language network associated with narrative comprehension. (6) The posterior medial cortex (PMC), temporo-parieto-occipital junction (TPOJ), dorsal frontal lobe (DFL) have higher correlation versus early auditory and auditory association cortex.

Limitations

We believe that more complex models instead of simple ridge regression models can lead to further exciting insights: (i.e. we may be able to link more directly internal model mechanisms to brain activations). However, to achieve this, one would need more data (more subjects and longer stories) to train such complex models.

Conclusion

This paper studied the long-term memory plausibility of language models for brain encoding. We observe that building individual encoding models and interpreting the internal representations among models can provide a more in-depth understanding of the neural representation of language information. Our experiments on the Narrative listening stories dataset lead to interesting cognitive insights.

Acknowledgments

We thank the anonymous reviewers for their very helpful comments on previous version of this paper.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., … others (2016). Tensorflow: a system for large-scale machine learning. In *Osdi*.

Affolter, N., Egressy, B., Pascual, D., & Wattenhofer, R. (2020). Brain2word: Decoding brain activity for language generation. *arXiv preprint arXiv:2009.04765*.

Antonello, R., Turek, J., Vo, V., & Huth, A. (2021). Low-dimensional structure in the space of language representations is reflected in brain responses. *arXiv preprint arXiv:2106.05426*.

Baldassano, C., Chen, J., Zadbood, A., Pillow, J. W., Hasson, U., & Norman, K. A. (2017). Discovering event structure in continuous narrative perception and memory. *Neuron*, *95*(3), 709–721.

Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Cadieu, C. F., Hong, H., Yamins, D. L., Pinto, N., Ardila, D., Solomon, E. A., … DiCarlo, J. J. (2014). Deep neural networks rival the representation of primate it cortex for core visual object recognition. *PLoS Computational Biology*, *10*(12), e1003963.

Caucheteux, C., Gramfort, A., & King, J.-R. (2021a). Disentangling syntax and semantics in the brain with deep networks. In *International conference on machine learning* (pp. 1336–1348).

Caucheteux, C., Gramfort, A., & King, J.-R. (2021b). Model-based analysis of brain activity reveals the hierarchy of language in 305 subjects. *arXiv preprint arXiv:2110.06078*.

Constable, R. T., Pugh, K. R., Berroya, E., Mencl, W. E., Westerveld, M., Ni, W., & Shankweiler, D. (2004). Sentence complexity and input modality effects in sentence comprehension: an fmri study. *NeuroImage*, *22*(1), 11–21.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 4171–4186).

Eickenberg, M., Gramfort, A., Varoquaux, G., & Thirion, B. (2017). Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage*, *152*, 184–194.

Gauthier, J., & Levy, R. (2019). Linking artificial and human neural representations of language. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)* (pp. 529–539).

Glasser, M. F., Coalson, T. S., Robinson, E. C., Hacker, C. D., Harwell, J., Yacoub, E., … others (2016). A multi-modal parcellation of human cerebral cortex. *Nature*, *536*(7615), 171–178.

Hinaut, X., & Dominey, P. F. (2013). Real-time parallel processing of grammatical structure in the fronto-striatal system: A recurrent network simulation study using reservoir computing. *PloS one*, *8*(2), e52946.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Huth, A. G., De Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, *532*(7600), 453–458.

Jain, S., & Huth, A. G. (2018). Incorporating context into language encoding models for fmri. In *Proceedings of the 32nd international conference on neural information processing systems* (pp. 6629–6638).

Karpathy, A., Johnson, J., & Fei-Fei, L. (2015). Visualizing and understanding recurrent networks. *arXiv preprint arXiv:1506.02078*.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., … Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Nastase, S. A., Liu, Y.-F., Hillman, H., Norman, K. A., & Hasson, U. (2020). Leveraging shared connectivity to aggregate heterogeneous datasets into a common response space. *NeuroImage*, *217*, 116865.

Nastase, S. A., Liu, Y.-F., Hillman, H., Zadbood, A., Hasenfratz, L., Keshavarzian, N., … others (2021). Narratives: fmri data for evaluating models of naturalistic language comprehension. *bioRxiv*, 2020–12.

O'Reilly, R. C., & Frank, M. J. (2006). Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia. *Neural computation*, *18*(2), 283–328.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp)* (pp. 1532–1543).

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Ruxton, G. D., & Beauchamp, G. (2008). Time for some a priori thinking about post hoc testing. *Behavioral ecology*, *19*(3), 690–693.

Schrimpf, M., Blank, I., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., … Fedorenko, E. (2021). The neural architecture of language: Integrative reverse-engineering converges on a model for predictive processing. *PNAS*, *Vol.*

Schwartz, D., Toneva, M., & Wehbe, L. (2019). Inducing brain-relevant bias in natural language processing models.

Advances in Neural Information Processing Systems, *32*, 14123–14133.

Toneva, M., Stretcu, O., Póczos, B., Wehbe, L., & Mitchell, T. M. (2020). Modeling task effects on meaning representation in the brain via zero-shot meg prediction. *Advances in Neural Information Processing Systems*, *33*.

Trumpp, N. M., Kliese, D., Hoenig, K., Haarmeier, T., & Kiefer, M. (2013). Losing the sound of concepts: Damage to auditory association cortex impairs the processing of sound-related concepts. *Cortex*, *49*(2), 474–486.

Variengien, A., & Hinaut, X. (2020). A journey in esn and lstm visualisations on a language task. *arXiv preprint arXiv:2012.01748*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998–6008).

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624.