Brain Encoding using Randomized Recurrent Networks

Mallikarjuna Reddy Yeruva
(yeruvamalli4@gmail.com)
Tirumala Engineering College, India

Naga Yaswanth Reddy Jonnala
(yasreddi@gmail.com)
Mahindra University, India

Uday Kiran Reddy Atmakuri
(udaykiranreddyatmakuri@gmail.com)
Venkateswara University, India

Mounika Marreddy
(mounika.marreddy@research.iiit.ac.in)
IIIT-Hyderabad, India

Abstract

Seeking plausible models for brain computation has been a continuing effort in brain encoding and decoding. Most prior works have mapped the association between stimulus representation from language models and fMRI brain activity using ridge regression. However, these models are not biologically plausible from the perspective of representing neural dynamics of the brain underlying the fMRI recordings. In this work, our primary motivation is to challenge ridge regression models with simple neural architectures such as echo state network (ESNs) and long short-term memory (LSTMs) on the brain encoding task that requires full-sentence processing in the task of reading short sentences. We explore various pretrained Transformer language models for computing sentence representations and predict the fMRI brain activity from simple neural architectures that include initial layers with random parameterization and that do not require explicit training. Experiment results show that (i) ESNs with online learning can accurately predict the fMRI brain activity comparable to ridge regression models, (ii) Both cell-state (internal memory representation related to long term memory) and out-gate (related to short term memory) of LSTM display an equal level performance during short sentences in random LSTMs, (iii) left hemisphere language area has higher predictive brain activity versus right hemisphere language area, (iv) ESNs with online learning yield superior performance over offline learning, indicating the biological plausibility of ESNs and the cognitive process of sentence reading, and (v) among all the variants of transformer models, Longformer features facilitate better accuracy when utilized with both ridge regression and ESN online learning models. The proposed framework that combines input featurization, dynamic memory and learning modules offers a flexible, biologically plausible architecture for investigating brain encoding in neuroscience.

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

Introduction

In the past decade, biologically-inspired artificial neural networks have witnessed a resurgence in an application by the computational neuroscience community to gain the understanding of how the brain effortlessly performs perception and cognitive processing in a variety of tasks such as (i) visual processing in object recognition tasks (Yamins et al., 2014; Eickenberg, Gramfort, Varoquaux, & Thirion, 2017), and (ii) higher-level cognition in language processing (Gauthier & Levy, 2019; Schrimpf et al., 2021; Schwartz, Toneva, & Wehbe, 2019; Oota, Alexandre, & Hinaut, 2022b; Oota, Arora, Agarwal, et al., 2022; Toneva, Mitchell, & Wehbe, 2022; Aw & Toneva, 2022). This line of work, namely brain encoding, aims at constructing neural brain activity given an input stimulus.

Since the discovery of the relationship between language stimuli and functions of brain networks using fMRI (Constable et al., 2004), researchers have been interested in designing models that capture the mapping between linguistic stimuli and brain activity. Most of the existing brain encoding models use ridge regression to predict the brain activity from stimulus features (Schrimpf et al., 2021). On the other hand, some studies looked at how sequence-based language models such as echo-state networks (ESN) (Dominey, 2021; Oota, Alexandre, & Hinaut, 2022b) or long short-term memory networks (LSTM) (Jain & Huth, 2018; Oota, Alexandre, & Hinaut, 2022b) encode the stimulus information. For instance, (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.

Despite some efforts in understanding the internal memory mechanism of LSTM (Karpathy, Johnson, & Fei-Fei, 2015) and its architectural design (O'Reilly & Frank, 2006), the cognitive plausibility of sequence-based architectures (ESNs and LSTMs) as well as how their working mechanism relates to brain encoding and decoding remains largely unexplored. In this paper, we open the black box of both ESN and LSTMs to look at particular detailed reservoir states and LSTM activation (the cell state and the output gate state) and their relation to brain activation profile. This can give more insights on reservoir states in ESNs, representations of longer-term and shorter-term information in LSTMs. 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 *output gate state* of the LSTM which constitutes short-term information.

Recently, 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 (Toneva & Wehbe, 2019; Schrimpf et al., 2021; Caucheteux, Gramfort, & King, 2021; Oota, Alexandre, & Hinaut, 2022b). On the other hand, (Gauthier & Levy, 2019; Oota, Arora, Agarwal, et al., 2022) fine-tunes a pretrained

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

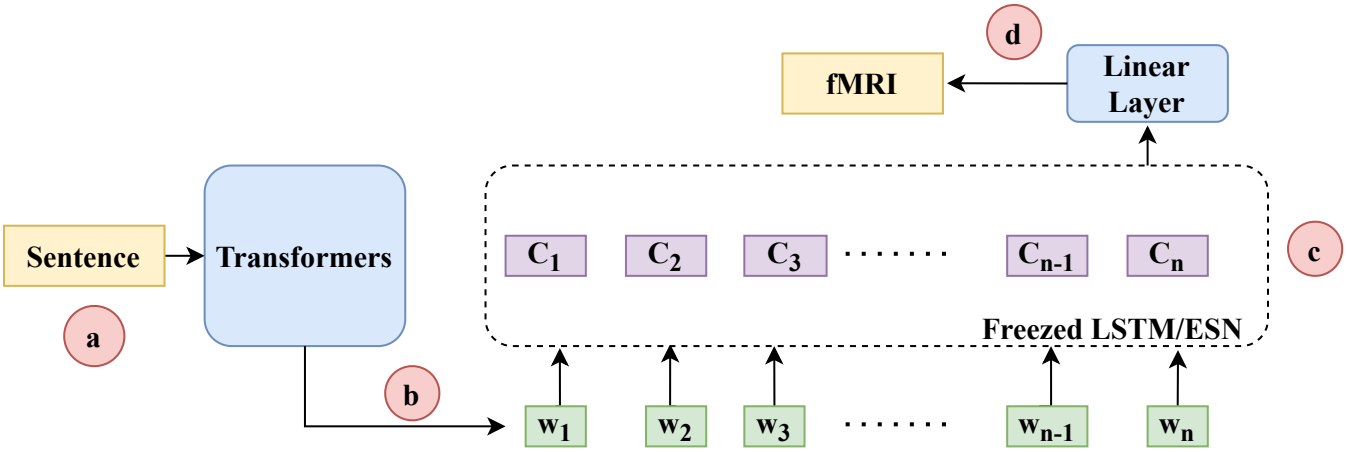

Figure 1: Workflow of our randomized recurrent brain encoder model. (a) denotes the extraction of word embeddings from variety of pretrained Transformers, (b) denotes the passing of word embeddings as input to LSTM/ESNs, (c) the recurrence mechanism (the input and LSTM layers are frozen and untrained), and (d) the prediction of fMRI brain activity by training the final linear layer.

BERT model on multiple natural language processing tasks to find tasks that best correlated with high *decoding* performance. In recent works, (Caucheteux et al., 2021; Antonello, Turek, Vo, & Huth, 2021; Oota, Gupta, & Toneva, 2022) 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-related 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.

Despite impressive performance with the ridge regression models, these models learn in a batch-mode from the whole training data and thus deviate from the typical sample-by-sample learning process adopted by humans. As such they are inappropriate for investigating hardest problems of language understanding. Our primary motivation is to challenge ridge regression models with simple neural architectures (like ESNs and LSTMs) on brain encoding task that requires full-sentence processing in a reading task. We aim to have models that could be easily grounded in cognitive modeling architectures while modeling language comprehension in the brain. Importantly, we do not want to focus on engineered neural architectures for biologically plausible purposes because we are also interested in exploring how relatively simple recurrent neural networks could generalize in such conditions while using incremental learning. In particular, one of the models we use, Echo State Networks (ESN) and, more generally, the Reservoir Computing paradigm, have already been used in several neuroscience applications (Maass, Natschläger, & Markram, 2002; Hinaut & Dominey, 2013) and are often referred to as a plausible computational principle for electrophysiological results (Rigotti et al., 2013; Enel, Procyk, Quilodran, & Dominey, 2016).

In this paper, we explore three architectures that predict fMRI brain activations from pre-trained word embeddings extracted from variety of Transformer language models as input to these architectures. Figure 1 depicts the workflow of our randomized recurrent brain encoder. In order to compare the ESNs with LSTMs, the input and LSTM layers are not trainable (frozen and no back-propagation). The sentence embeddings from both hidden and cell state are then used as output features for a fMRI prediction. The proposed framework that combines input featurization, dynamic memory and learning modules offers a flexible, biologically plausible architecture for investigating brain encoding in neuroscience (Oota, Alexandre, & Hinaut, 2022a). The predictive power of language model specific representations with brain activation is ascertained by (1) using ridge regression on such representations and predicting activations, (2) using biological plausible ESN and Random LSTM, and (3) computing popular metric like 2V2 accuracy (Toneva, Stretcu, Póczos, Wehbe, & Mitchell, 2020) between actual and predicted activations.

Specifically, we make the following contributions in this paper. (1) Given a pretrained Transformer language model, we propose the problem of finding which of these are the most predictive of fMRI brain activity for reading short sentences task. (2) Our cognitive plausibility of language model results reveals that ESNs with online learning can accurately predict the fMRI brain activity comparable to ridge regression models. (3) We also investigate the internal memory representations of LSTM (cell state and output gate), internal states of the reservoirs during short sentences reading task. (4) Our proposed framework that combines input featurization, dy-

Table 1: # Voxels in each ROI in the Pereira Dataset. LH - Left Hemisphere. RH - Right Hemisphere.

ROIs→	Language			DMN	Task Positive
↓Subj	LH	RH	Vision		
P01	5265	6172	12829	17190	35120
M02	4930	5861	11729	15070	30594
M04	5906	5401	12278	18011	34024
M07	5629	5001	12454	17020	30408
M15	5315	6141	12383	15995	31610

namic memory and learning modules offers a flexible, biologically plausible architecture for investigating brain encoding in neuroscience.

Overall, our goal is not to obtain a new state-of-the-art (SOTA), but to put current SOTA on more solid footing by 1) looking at how much they gain compared to biologically plausible ESNs and Random LSTMs; and 2) providing the field with more solid baselines, going forward.

Methodology

Brain Imaging Dataset: We work with Pereira dataset (Pereira et al., 2018). Similar to earlier work (Sun, Wang, Zhang, & Zong, 2019, 2020; Oota, Arora, Gupta, & Bapi, 2022), we combine the data from sentence-based experiments (experiments-2 and 3) from (Pereira et al., 2018). Five subjects were presented a total of 627 sentences from 48 broad topics, spanning over 168 passages, where each passage consists of 3-4 sentences. As in (Pereira et al., 2018), we focused on four brain ROIs (regions of interest) corresponding to four brain networks: (i) Default Mode Network (DMN) (linked to the functionality of semantic processing), (ii) Language Network (related to language processing, understanding, word meaning, and sentence comprehension), (iii) Task Positive Network (TP) (related to attention, salience information), and (iv) Visual Network (related to the processing of visual objects, object recognition). We briefly summarize the details of the dataset and the number of voxels corresponding to each ROI in Table 1. We use the AAL parcellation Atlas (116 brain ROIs) to present the brain map results, since Pereira dataset contains annotations tied to this atlas.

Encoding Models

In this section, we propose to employ brain encoding using three models, including simple ridge regression, ESN (i.e. Reservoir Computing), and LSTM. Here, we recall the definitions of Reservoir Computing and random features in ESN and LSTM and introduce the model architecture details.

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 pretrained Transformer language model. Formally, we encode the stimuli as $X \in \mathbb{R}^{N \times D}$ and brain region voxels $Y \in \mathbb{R}^{N \times V}$, 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.

LSTM: LSTM (Hochreiter & Schmidhuber, 1997) network has a memory cell and three gates: input gate, output gate and forget gate. The memory cell c_t keeps the useful history information which will be used for the next process. The weights of LSTMs are learned using the error back-propagation through time, BPTT, algorithm. In order to compare the performance of ESNs with LSTMs, we employ unidirectional LSTMs, but in our case without any training. The LSTM weight matrices and their corresponding biases are initialized uniformly at random and kept frozen (i.e both Input and LSTM layers are done with random initialization).

Echo State Networks (Reservoir Computing): Reservoir Computing techniques allow the use of a great variety of learning mechanisms to solve sequence prediction problems, where given a sequence X, we predict a label y for each step in the sequence. In ESNs, the learning rules are sorted in two categories: offline learning and online learning.

Offline Learning Offline learning rules are the most common learning rules in machine learning. Within the Reservoir Computing field, linear regression is probably the simplest and the more used way of training an artificial neural network. Linear regression is said to be an offline learning rule because parameters of the linear regression model are learned given all available samples of data and all available samples of target values. Once the model is learned, it cannot be updated without training the model on the whole dataset another time.

Online Learning As opposed to offline learning, we use the online FORCE learning algorithm (Sussillo & Abbott, 2009) which allows to update output weights $\mathbf{W}_{out}$ for each learning example, using such incremental learning is more biologically plausible than using the classical ESN offline learning approach. This method does not unfold time while training the network like back-propagation through time. As most deep learning algorithms cannot use such rules to update their parameters, as gradient descent algorithms requires several samples of data at a time to obtain convergence, Reservoir Computing algorithms can use these kind of rules. Indeed, only readout connections need to be trained.

Feature Spaces: To simultaneously test representations from multiple pretrained language models, we used the latent space features from each of the following eleven popular pretrained Transformer language models: BERT (Devlin et al., 2019), GPT-2 (Radford et al., 2019), XLNet (Yang et al., 2019), RoBERTa (Liu et al., 2019), ALBERT (Lan et al., 2019), DistilBERT (Sanh, Debut, Chaumond, & Wolf, 2019), ELECTRA (Clark, Luong, Le, & Manning, 2020), Transformer-XL (Dai et al., 2019), T5 (Raffel et al., 2020), Reformer (Kitaev, Kaiser, & Levskaya, 2019), and Longformer (Beltagy et al., 2020). Except Reformer and Longformer, remaining all the models have a fixed maximum sequence length (512) and do not handle longer sequences. Given an input sentence, each pretrained Transformer outputs token representations at the final layer. We use the #tokens $\times$ 768 dimension vector obtained from the last hidden

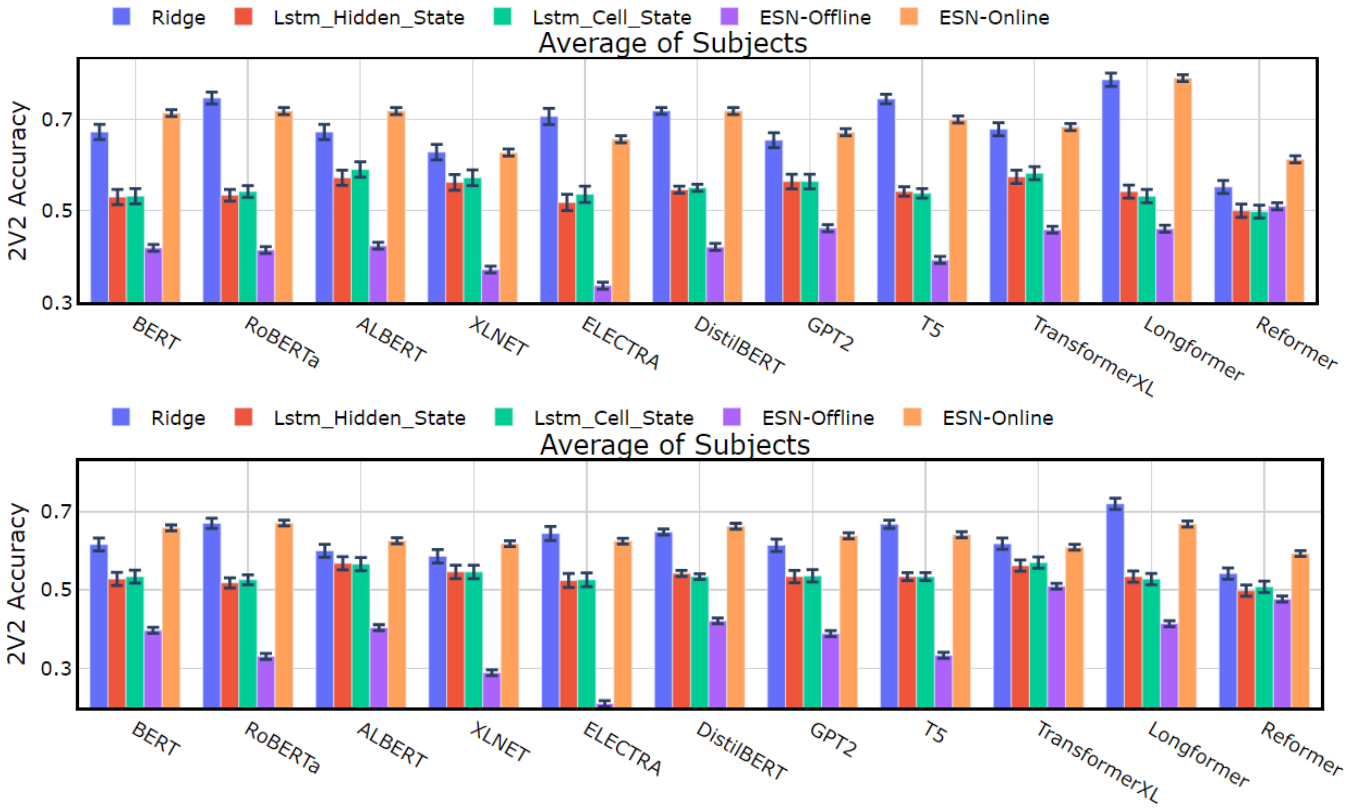

Figure 2: Language_LH (top), and Language_RH (bottom): 2V2 Accuracy between predicted and true responses using a variety of language models (for Pereira dataset). Results are averaged across all participants. Ridge and ESN Online are the best.

layer to obtain latent features for the stimuli. Since individual sentences were presented to the subjects while modeling, sentences were passed one by one to the pretrained Transformer model, and average-pooled representations were used to encode the sentence stimuli. We then build individual ridge regression models with the extracted latent features to predict brain responses and measure the accuracy between the prediction and the true response. For the sequence based models such as ESNs and LSTMs, we use the last layer token representations as input to predict the fMRI.

Cross-Validation: We follow K-fold (K=5) cross-validation. All the data samples from K-1 folds were used for training, and the model was tested on samples of the left-out fold.

Evaluation Metrics: We evaluate our models using popular brain encoding evaluation metric (2V2 Accuracy) (Toneva et al., 2020) 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.

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

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

Experimental Setup

We compare the ridge regression model with the ESNs and Random LSTMs.

Ridge Regression: We used sklearn's ridge-regression with default parameters, 5-fold cross-validation, Stochastic-Average-Gradient Descent Optimizer, Huggingface for Transformer models, MSE loss function, and L2-decay (λ) as 1.0.

ESN Training: We use the default parameters of ESN obtained from hyperopt library[1] for Pereira dataset as follows: {Size of the Reservoir = 500, Spectral Radius = 0.185, Leak Rate = 0.0097, Sparsity (on Reservoir Weight Matrix - $\mathbf{W}_{rec}$) = 0.5, Regularization coefficient = $1.3e^{-10}$, Input Scaling = 1.0}.

Random LSTM: We build a random LSTM model where the output layer is trained while the input and the LSTM layers are kept frozen. We use both the output and cell state vectors to perform fMRI encoding. The model is implemented in Keras with TensorFlow backend with mean suquared error (mse) as loss, Adam optimizer, the number epochs set to 20, the batch size is of 8, applied dropout with a keep-probability

[1] http://hyperopt.github.io/hyperopt/

of 0.2, learning rate (0.01), maximum sequence length is obtained from sentences of Pereira dataset, and tried LSTM with hidden state size set to 256.

Results and Discussion

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 between the predicted and true responses across various ROIs for the reading (Pereira) dataset (Fig. 2).

Encoding performance of Language Models in Language Region:

From Fig. 2, we observe that ESNs with online learning can accurately predict the fMRI brain activity comparable to ridge regression models. However, the performance of ESN offline learning and LSTM with both cell state and hidden state performance is low compared to ESN online learning. In order to estimate the statistical significance of the performance differences, we performed one-way ANOVA on the mean 2V2 scores for the subjects across the encoder architectures (Ridge, LSTM (output gate), LSTM (cell state), ESN offline, and ESN online) for the 11 pretrained Transformer models. The main effect of the ANOVA test was significant for all the Transformer models with $p \leq 10^{-2}$ with confidence 95%. Further, *post hoc* pairwise comparisons (Ruxton & Beauchamp, 2008) confirmed the visual observations that on 2V2 accuracy measures, tasks such as ridge and ESN-online learning performed significantly better compared to other models (indicated by *), as shown in Table 2. These results demonstrate that when reading short sentences, information processing of sentence in both cell state and hidden state constitute an equal level performance. Further, we compared the performance of representations of Transformer models and observe that Longformer featutes report higher 2v2 accuracy over other methods. The detailed p-values across Transformer models are reported in the supplementary. The 2v2 accuracy of other brain networks such as DMN and Task Positive are in the supplementary.

ESN: Effects of Offline vs Online Learning:

To explore the cognitive plausibility in terms of the internal representations in the hidden layer of ESN, we compare the encoding performance between both offline and online learning methods. Fig. 2 report the fMRI encoding performance of ESNs where the online learning method yields better performance than offline learning, indicating the biological plausibility of ESN and the cognitive process of sentence reading. To investigate the internal states of ESN during online learning, we report the absolute variation of the activation of reservoir neurons during the processing of the sentence in Figure 3. Since, we do not use any feedback in our reservoir, the states of the reservoir are fully determined by its initial random weights and the inputs received. In fact, the learning process happens by combining the useful activities given the random projections of the inputs done in the reservoir.

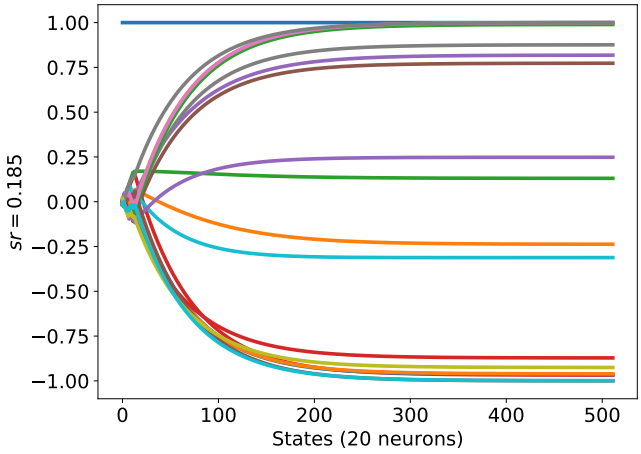

Figure 3: Absolute variation of the activation of reservoir neurons during the processing of the sentence.

RandLSTM: Effects of Output Gate vs Cell State Vectors:

In order to explore how RandLSTM 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 output gate and cell state vectors. Fig. 2 showcases the fMRI encoding performance of RandLSTM where both cell-state (internal memory representation related to long term memory) and out-gate (related to short term memory) of LSTM display an equal level performance during short sentences. Further, the pairwise comparisons using *post hoc* analysis confirmed the visual observations that on 2V2 accuracy, cell and hidden states do not differ (See in Table 2).

Effects of Language of sub ROIs:

To further investigate which sub ROIs (LPTG, LMTG, LATG, LFus, Lpar, Lang, LIFGorb, LIFG, LaMFG, LpMFG, and LmMFG) of the Language network are related to the predictive task features, we train encoding models for all the sub ROIs for the best performing models such as ESN online and ridge regression, as shown in Fig. 4. We notice that both LMTG (middle temporal gyrus) and LPTG (posterior temporal gyrus) are more accurately predicted than the other sub ROIs. On the other hand, LIFG-orb displays a lower Pearson correlation for both the encoder models. The presence of superior encoding information in the ROIs in the temporal gyrus as compared to those in the inferior frontal gyrus seems to mirror similar observations seen in decoder performance (Anderson et al., 2017).

Conclusion

In this work, we challenge the ridge regression models with simple neural architectures such as ESNs and LSTMs on the brain encoding task that requires full-sentence processing in the task of reading short sentences. Further, we explore vari-

Table 2: Language_LH: p-values obtained using *post hoc* pairwise comparisons for the three architectures across Pretrained Transformers.

Models compared	BERT	RoBERTa	ALBERT	XLNET	ELECTRA	DistilBERT	GPT-2	T5	Transformer-XL	Longformer	Reformer
Ridge vs Online	0.468	0.878	0.526	0.573	0.478	1	0.967	0.448	0.999	1	0.005*
Online vs Offline	0.000*	0.000*	0.000*	0.000*	0.000*	0.000*	0.000*	0.000*	0.000*	0.000*	0.000*
Online vs Cell	0.000*	0.000*	0.002*	0.006*	0.006*	0.000*	0.006*	0.000*	0.012*	0.000*	0.000*
Ridge vs Online	0.000*	0.000*	0.000*	0.000*	0.000*	0.000*	0.000*	0.000*	0.000*	0.000*	0.068
Cell vs Hidden	1	0.998	0.969	0.992	0.974	0.999	1	0.999	0.998	0.997	0.999

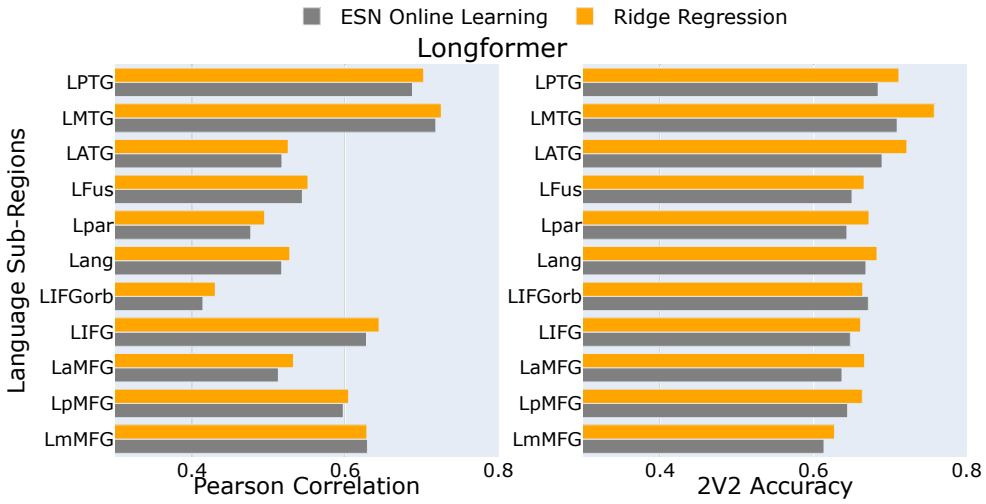

Figure 4: Pearson correlation coefficient and 2v2 accuracy measured between predicted and true responses across different sub ROIs of the Language Network using ESN online Learning and ridge regression. Results are averaged across all participants.

ous pre-trained Transformer language models for computing sentence representations and predict the fMRI brain activity from simple neural architectures that include initial layers with random parameterization and that do not require explicit training. Experiment results show that: ESNs with online learning can accurately predict the fMRI brain activity better than LSTM models. This is due to the fact that ESNs are more biologically plausible than LSTM and can learn incrementally by seeing each utterance only once, contrary to LSTMs that need to process the data for several epochs.

References

Anderson, A. J., Binder, J. R., Fernandino, L., Humphries, C. J., Conant, L. L., Aguilar, M., ... Raizada, R. D. (2017). Predicting neural activity patterns associated with sentences using a neurobiologically motivated model of semantic representation. *Cerebral Cortex*, *27*(9), 4379–4395.

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*.

Aw, K. L., & Toneva, M. (2022). Training language models for deeper understanding improves brain alignment. *arXiv preprint arXiv:2212.10898*.

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

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

Clark, K., Luong, M.-T., Le, Q. V., & Manning, C. D. (2020). Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

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.

Dai, Z., Yang, Z., Yang, Y., Carbonell, J. G., Le, Q., & Salakhutdinov, R. (2019). Transformer-xl: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th annual meeting of the acl* (pp. 2978–2988).

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 naacl: Human language technologies* (pp. 4171–4186).

Dominey, P. F. (2021). Narrative event segmentation in the cortical reservoir. *bioRxiv*.

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.

Enel, P., Procyk, E., Quilodran, R., & Dominey, P. F. (2016). Reservoir computing properties of neural dynamics

in prefrontal cortex. *PLoS computational biology*, *12*(6), e1004967.

Gauthier, J., & Levy, R. (2019). Linking artificial and human neural representations of language. In *Proceedings of the 2019 conference (emnlp-ijcnlp)* (pp. 529–539).

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.

Jain, S., & Huth, A. G. (2018). Incorporating context into language encoding models for fmri. In *Nips* (pp. 6629–6638).

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

Kitaev, N., Kaiser, L., & Levskaya, A. (2019). Reformer: The efficient transformer. In *Iclr*.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. In *Iclr*.

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*.

Maass, W., Natschläger, T., & Markram, H. (2002). Real-time computing without stable states: A new framework for neural computation based on perturbations. *Neural computation*, *14*(11), 2531–2560.

Oota, S. R., Alexandre, F., & Hinaut, X. (2022a). Cross-situational learning towards robot grounding.

Oota, S. R., Alexandre, F., & Hinaut, X. (2022b). Long-term plausibility of language models and neural dynamics during narrative listening. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 44).

Oota, S. R., Arora, J., Agarwal, V., Marreddy, M., Gupta, M., & Surampudi, B. R. (2022). Neural language taskonomy: Which nlp tasks are the most predictive of fmri brain activity? *arXiv preprint arXiv:2205.01404*.

Oota, S. R., Arora, J., Gupta, M., & Bapi, R. S. (2022). Multi-view and cross-view brain decoding. In *Proceedings of the 29th international conference on computational linguistics* (pp. 105–115).

Oota, S. R., Gupta, M., & Toneva, M. (2022). Joint processing of linguistic properties in brains and language models. *arXiv preprint arXiv:2212.08094*.

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.

Pereira, F., Lou, B., Pritchett, B., Ritter, S., Gershman, S. J., Kanwisher, N., ... Fedorenko, E. (2018). Toward a universal decoder of linguistic meaning from brain activation. *Nature communications*, *9*(1), 1–13.

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.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, *21*, 1–67.

Rigotti, M., Barak, O., Warden, M. R., Wang, X.-J., Daw, N. D., Miller, E. K., & Fusi, S. (2013). The importance of mixed selectivity in complex cognitive tasks. *Nature*, *497*(7451), 585–590.

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

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

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. *NeurIPS*, *32*, 14123–14133.

Sun, J., Wang, S., Zhang, J., & Zong, C. (2019). Towards sentence-level brain decoding with distributed representations. In *Proceedings of the aaai* (Vol. 33, pp. 7047–7054).

Sun, J., Wang, S., Zhang, J., & Zong, C. (2020). Neural encoding and decoding with distributed sentence representations. *IEEE TNNLS*, *32*(2), 589–603.

Sussillo, D., & Abbott, L. F. (2009). Generating coherent patterns of activity from chaotic neural networks. *Neuron*, *63*(4), 544–557.

Toneva, M., Mitchell, T. M., & Wehbe, L. (2022). Combining computational controls with natural text reveals aspects of meaning composition. *Nature Computational Science*, *2*(11), 745–757.

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. *NeurIPS*, *33*, 5284–5295.

Toneva, M., & Wehbe, L. (2019). Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). *arXiv preprint arXiv:1905.11833*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. In *Nips* (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. *PNAS*, *111*(23), 8619–8624.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *NeurIPS*, *32*.