Relating Simple Sentence Representations in Deep Neural Networks and the Brain

Sharmistha Jat[1]* **Hao Tang**[2] **Partha Talukdar**[1] **Tom Mitchell**[2]
[1]Indian Institute of Science, Bangalore
[2]School of Computer Science, Carnegie Mellon University
{sharmisthaj,ppt}@iisc.ac.in
htang1@alumni.cmu.edu, tom.mitchell@cs.cmu.edu

Abstract

What is the relationship between sentence representations learned by deep recurrent models against those encoded by the brain? Is there any correspondence between hidden layers of these recurrent models and brain regions when processing sentences? Can these deep models be used to synthesize brain data which can then be utilized in other extrinsic tasks? We investigate these questions using sentences with simple syntax and semantics (e.g., *The bone was eaten by the dog.*). We consider multiple neural network architectures, including recently proposed ELMo and BERT. We use magnetoencephalography (MEG) brain recording data collected from human subjects when they were reading these simple sentences.

Overall, we find that BERT's activations correlate the best with MEG brain data. We also find that the deep network representation can be used to generate brain data from new sentences to augment existing brain data. To the best of our knowledge, this is the first work showing that the MEG brain recording when reading a word in a sentence can be used to distinguish earlier words in the sentence. Our exploration is also the first to use deep neural network representations to generate synthetic brain data and to show that it helps in improving subsequent stimuli decoding task accuracy.

1 Introduction

Deep learning methods for natural language processing have been very successful in a variety of Natural Language Processing (NLP) tasks. However, the representation of language learned by such methods is still opaque. The human brain is an excellent language processing engine, and the brain representation of language is of course very effective. Even though both brain and deep learning methods are representing language, the relationships among these representations are not thoroughly studied. Wehbe et al. (2014b) and Hale et al. (2018) studied this question in some limited capacity. Wehbe et al. (2014b) studied the processing of a story context at a word level during language model computation. Hale et al. (2018) studied the syntactic composition in RNNG model (Dyer et al., 2016) with human encephalography (EEG) data.

We extend this line of research by investigating the following three questions: (1) what is the relationship between sentence representations learned by deep learning networks and those encoded by the brain; (2) is there any correspondence between hidden layer activations in these deep models and brain regions; and (3) is it possible for deep recurrent models to synthesize brain data so that they can effectively be used for brain data augmentation. In order to evaluate these questions, we focus on representations of simple sentences. We employ various deep network architectures, including recently proposed ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019) networks. We use MagnetoEncephaloGraphy (MEG) brain recording data of simple sentences as the target reference. We then correlate the representations learned by these various networks with the MEG recordings. Overall, we observe that BERT representations are the most predictive of MEG data. We also observe that the deep network models are effective at synthesizing brain data which are useful in overcoming data sparsity in stimuli decoding tasks involving brain data.

In summary, in this paper we make the following contributions.

- We initiate a study to relate representations of simple sentences learned by various deep networks with those encoded in the brain. We establish correspondences between activations in deep network layers with brain ar-

* This research was carried out during a research internship at the Carnegie Mellon University.

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 5137–5154
Florence, Italy, July 28 - August 2, 2019.

eas.

- We demonstrate that deep networks are capable of predicting change in brain activity due to differences in previously processed words in the sentence.

- We demonstrate effectiveness of using deep networks to synthesize brain data for downstream data augmentation.

We have made our code and data[1] publicly available to support further research in this area.

2 Datasets

In this section, we describe the MEG dataset and Simple Sentence Corpus used in the paper.

2.1 MEG Dataset

Magnetoencephalography (MEG) is a non-invasive functional brain imaging technique which records magnetic fields produced by electrical currents in the brain. Sensors in the MEG helmet allow for recording of magnetic fluctuations caused by changes in neural activity of the brain. For the experiments in this paper, we used three different MEG datasets collected when subjects were shown simple sentences as stimulus. These datasets are summarized in Table 1, please see (Rafidi, 2014) for more details. Additional dataset details are mentioned in appendix section A.1. In the MEG helmet, 306 sensors were distributed over 102 locations and sampled at 1kHz. Native English speaking subjects were asked to read simple sentences. Each word within a sentence was presented for 300ms with 200ms subsequent rest. To reduce noise in the brain recordings, we represent a word's brain activity by averaging 10 sentence repetitions (Sudre et al., 2012). Comprehension questions followed 10% of sentences, to ensure semantic engagement. MEG data was acquired using a 306 channel Elekta Neuromag device. Preprocessing included spatial filtering using temporal signal space separation (tSSS), low-pass filtering 150Hz with notch filters at 60 and 120Hz, and downsampling to 500Hz (Wehbe et al., 2014b). Artifacts from tSSS-filtered same-day empty room measurements, ocular and cardiac artifacts were removed via Signal Space Projection (SSP).

Dataset	#Sentences	Voice	Repetition
PassAct1	32	P+A	10
PassAct2	32	P+A	10
Act3	120	A	10

Table 1: MEG datasets used in this paper. Column 'Voice' refers to the sentence voice, 'P' is for passive sentences and 'A' is for active. Repetition is the number of times the human subject saw a sentence. For our experiments, we average MEG data corresponding to multiple repetitions of a single sentence.

2.2 Simple Sentence Corpus

In this paper, we aim to understand simple sentence processing in deep neural networks (DNN) and the brain. In order to train DNNs to represent simple sentences, we need a sizeable corpus of simple sentences. While the MEG datasets described in Section 2.1 contain a few simple sentences, that set is too small to train DNNs effectively. In order to address this, we created a new Simple Sentence Corpus (SSC), consisting of a mix of simple active and passive sentences of the form *"the woman encouraged the girl"* and *"the woman was encouraged by the boy"*, respectively. The SSC dataset consists of 256,145 sentences constructed using the following two sets.

- Wikipedia: We processed the 2009 Wikipedia dataset to get sentences matching the following patterns.
  "the [noun+] was [verb+] by the [noun+]"
  "the [noun+] [verb+] the [noun+]"
  If the last word in the pattern matched is not noun, then we retain the additional dependent clause in the sentence. We were able to extract 117,690 active, and 8210 passive sentences from wikipedia.

- NELL triples: In order to ensure broader coverage of Subject-Verb-Object (SVO) triples in our sentence corpus, we used the NELL SVO triples[2] (Talukdar et al., 2012). We subsample SVO triples based on their frequency (threshold = 6), a frequent verb list, and Freebase to get meaningful sentences. Any triple with subject or object or verb not in Freebase is discarded from the triple set.

  - Active sentence: Convert the verb to its past tense and concatenate the triple us-

[1] https://github.com/SharmisthaJat/ACL2019-SimpleSentenceRepr-DNN-Brain

[2] NELL SVO triples: http://rtw.ml.cmu.edu/resources/svo/

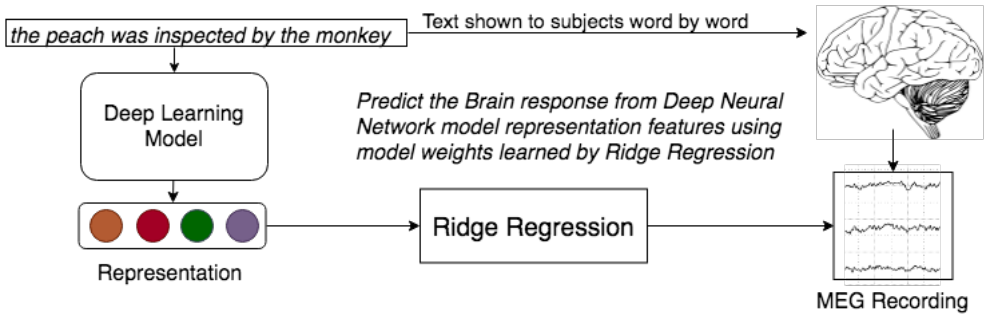


Figure 1: Encoding model for MEG data. 306 channel 500ms MEG signal for a single word was compressed to 306×5 by averaging 100ms data into a single column. This MEG brain recording data is then encoded from text representation vector to brain activity using ridge regression. The evaluation is done using 5 fold cross-validation. Please see Section 4 for more details.

ing the following pattern: *"the [subject] [verb-past-tense] the [object]"*.

– Passive sentence: Concatenate the triple using pattern: *"the [object] was [verb-past-tense] by the [subject]"*

We generate 86,452 active and 43,793 passive sentences in total from the NELL triples.

We train our deep neural network models with 90% of sentences in this dataset and test on the remaining 10%. We used the spaCy (Honnibal and Montani, 2017) library to predict POS tags for words in this dataset.

3 Methods

We test correlations between brain activity and deep learning model activations (LeCun et al., 2015) for a given sentence using a classification task, similar to previous works (Mitchell et al., 2008; Wehbe et al., 2014a,b). If we are able to predict brain activity from the neural network activation, then we hypothesize that there exists a relationship between the process captured by the neural network layer and the brain. The schematic of our encoding approach is shown in Figure 1.

We investigate various deep neural network models using context sensitivity tests to evaluate their performance in predicting brain activity. Working with these models and their respective training assumptions help us in understanding which assumption contributes to the correlations with the brain activity data. We process the sentences incrementally for each model to prevent information from future words from affecting the current representation, in line with how information is processed by the brain. For example, in the sentence *"the dog ate the biscuit"*, the representation of the word *"ate"* is calculated by processing sentence segment *"the dog ate"* and taking the last representation in each layer as the context for the word *"ate"*. The following embedding models are used to represent sentences.

- **Random Embedding Model**: In this model, we represent each word in a context by a randomly generated 300-dimensional vector. Each dimension is uniformly sampled between [0,1]. The results from this model help us establish the random baseline.

- **GloVe Additive Embedding Model**: This model represents a word context as the average of the current word's GloVe embedding (Pennington et al., 2014) and the previous word context. The first word in a sentence is initialized with its GloVe embedding as context.

- **Simple Bi-directional LSTM Language Model**: We build a language model following (Inan et al., 2016). Given a sequence of words $w_1 \dots w_t$, we predict the next word w_{t+1} using a two layer bidirectional-LSTM model (Hochreiter and Schmidhuber, 1997). The model is trained on the simple language corpus data as described in Section 2.1 with a cross-entropy loss. We evaluate our model on 10% held out text data. The perplexity for the Bi-directional Language model is 9.97 on test data (the low perplexity value is due to the simple train and test dataset).

- **Multi-task Model**: Motivated by the brain's multitask capability, we build a model to predict next word and POS tag information. The

multitask model is a simple two layer bidirectional LSTM model with separate linear layers predicting each of the tasks given the output of the last LSTM layer (Figure 2). The model is trained on the simple sentence corpus data as described in Section 2.1 with a cross-entropy loss. The model's accuracy is 96.9% on the POS-tag prediction task and has perplexity of 9.09 on the 10% test data. The high accuracy and low perplexity are due to the simple nature of our language dataset.

- **ELMO** (Peters et al., 2018): ELMo is a recent state-of-the-art deep contextualized word representation method which models a word's features as internal states of a deep bidirectional language model (biLM) pretrained on a large text corpus. The contextualized word vectors are able to capture interesting word characteristics like polysemy. ELMO has been shown to improve performance across multiple tasks, such sentiment analysis and question answering.

- **BERT** (Devlin et al., 2019): BERT uses a novel technique called Masked Language Model (MLM). MLM randomly masks some tokens inputs and then predicts them. Unlike previous models, this technique can use both left and right context to predict the masked token. The training also predicts the next sentence. The embedding in this model consists of 3 components: token embedding, sentence embedding and transformer positional embedding. Due to the presence of sentence embeddings, we observe an interesting performance of the embedding layer in our experiments.

4 Experiments and Results

With human brain as the reference language processing engine, we investigate the relationship between deep neural network representation and brain activity recorded while processing the same sentence. For this task, we perform experiments at both the macro and micro sentence context level. The **macro-context** experiments evaluate the overall performance of deep neural networks in predicting brain data for input words (all words, nouns, verbs etc.). The **micro-context** experiments, by contrast, focus on evaluating the performance of deep neural network representations in

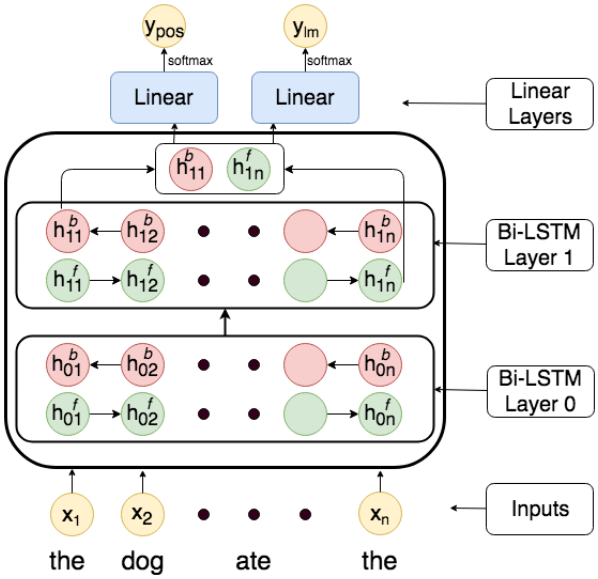


Figure 2: Architecture diagram for the simple multitask model. The second LSTM layer's output is processed by 2 linear layers each producing the next-word and the POS-tag prediction. We process each sentence incrementally to get the prediction for word at the nth position, this helps in removing forward bias from future words and therefore is consistent with the information our brain receives when processing the same sentence. Our Simple Bi-directional LSTM language model also has a similar architecture with just one output linear layer for next word prediction.

detecting minor changes in sentence context prior to the token being processed.

Regression task: Similar to previous research (Mitchell et al., 2008; Wehbe et al., 2014b), we use a classification task to align model representations with brain data. MEG data (Section 2.1) is used for these experiments. The task classifies between a candidate word and the true word a subject is reading at the time of brain activity recording. The classifier uses an intermediate regression step to predict the MEG activity from deep neural network representation for the true and the candidate word. The classifier then chooses the word with least Euclidean distance between the predicted and the true brain activity. A correct classification suggests that the deep neural network representation captures important information to differentiate between brain activity at words in different contexts. Detailed steps of this process are described as follows.

Regression training: We perform regression from the neural-network representation (for each layer) to the brain activity for the same input words in context. We normalized, preprocessed

and trained on the MEG data as described by (Wehbe et al., 2014b) (Section 2.3.2). We average the signal from every sensor (total 306) over 100ms non-overlapping windows, yielding a 306×5 sized MEG data for each word. To train the regression model, we take the training portion of the data in each fold, (X, Y), in the tuple (x_i, y_i), x_i is the layer representation for an input word i in a neural network model, and y_i is the corresponding MEG recording of size 1530 (flattened 306*5). The Ridge regression model (f) (Pedregosa et al., 2011) is learned with generalized cross-validation to select λ parameter (Golub et al., 1979). Ridge regression model's α parameter is selected from range $[0.1, \ldots, 100, 1000]$. The trained regression model is used to estimate MEG activity from the stimulus features, i.e., $\hat{y_i} = f(x_i)$.

Regression testing: The trained regression model is used to predict $\hat{y_i}$ for each word stimulus (x_i) in the test fold during cross-validation. We perform a pair-wise test for the classification accuracy (Acc) (Mitchell et al., 2008). The chance accuracy of this measure is 0.5. We use Euclidean distance (E_{dist}) as given in (1) for the measure.

$$\mathrm{Acc} = \begin{cases} 1, \text{ if } \mathrm{E}_{dist}(f(x_i), y_i) + \mathrm{E}_{dist}(f(x_j), y_j) \\ \quad \leq \mathrm{E}_{dist}(f(x_i), y_j) + \mathrm{E}_{dist}(f(x_j), y_i) \\ 0, \quad \text{otherwise} \end{cases} \tag{1}$$

4.1 Macro-context Experiments

The macro-context experiments aggregate classification performance of each model's layer on the entire stimuli set. We also evaluate on smaller sets such as only the nouns, verbs, passive sentence words, active sentence words, etc. The macro experiments help us to compare all the models on a large stimuli set. In summary, we observe the following: (1) the intermediate layers of state-of-the-art deep neural network models are most predictive of brain activity (Jain and Huth (2018) also observe this on a 3 layer LSTM language model), (2) in-context representations are better at predicting brain activity than out-of-context representations (embeddings), and (3) Temporal lobe is predicted with highest accuracy from deep neural network representations.

Detailed Observations: The results of pairwise classification tests for various models are presented in Figure 3. All the results reported in this section are for PassAct1 dataset. From the figure, we observe that BERT and ELMo outperform the simple models in predicting brain activity data. In the neural network language models, the middle layers perform better at predicting brain activity than the shallower or deeper layers. This could be due to the fact that the shallower layers represent low-level features and the deeper layers represent more task-oriented features. We tested this hypothesis by examining the performance scores at each lobe of the brain. For each area, we tested the left and right hemispheres independently and compared these performances with the bilateral frontal lobe as well as the activity across all regions. In particular, we examined the primary visual areas (left and right occipital lobe), speech and language processing areas (left temporal) and verbal memory (right temporal), sensory perception (left parietal) and integration (right parietal), language related movements (left frontal) and nonverbal functioning (right frontal). The frontal lobe was tested bilaterally as it is associated with higher level processing such as problem solving, language processing, memory, judgement, and social behavior.

From our results, we observe that lower layers such as BERT layer 5 have very high accuracy for right occipital and left occipital lobe associated with low-level visual processing task. In contrast, higher layers such as linear layers in the Multitask Model and in Language Model have the highest accuracy in the left temporal region of the brain. Figure 4 shows the pairwise classification accuracy for a given brain region for best layers from each model. The accuracy is highest in left temporal region, responsible for syntactic and semantic processing of language. These results establish correspondences between representations learned by deep neural methods and those in the brain. Further experiments are needed to improve our understanding of this relationship.

We performed additional experiments to predict on a restricted stimuli set. In each of these experiments, a subset of stimuli, for example active sentences, passive sentences, noun, and verb stimuli were used in classification training and testing. Detailed results for this experiment are documented in the appendix section (Figure 9). From the results, we observe that active sentences are predicted better (best accuracy = 0.93) than passive sentences (best accuracy = 0.87). This might be attributed to the nature of training datasets for

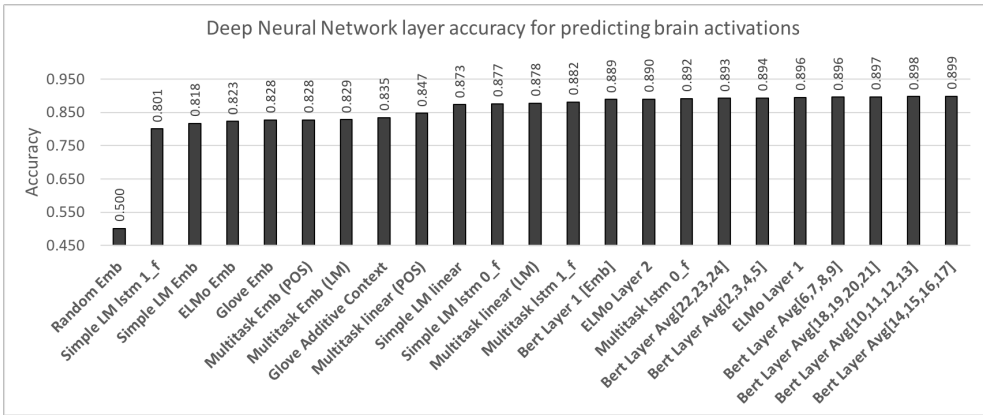


Figure 3: Pairwise classification accuracy of brain activity data predicted from various model layer representations. We average 4 consecutive layers of BERT into one value. We find that BERT and ELMO model layers perform the best. The middle layers of most models and BERT, in particular, are good at predicting brain activity. Read '_f' as forward layer and 'Emb' as the embedding layer.

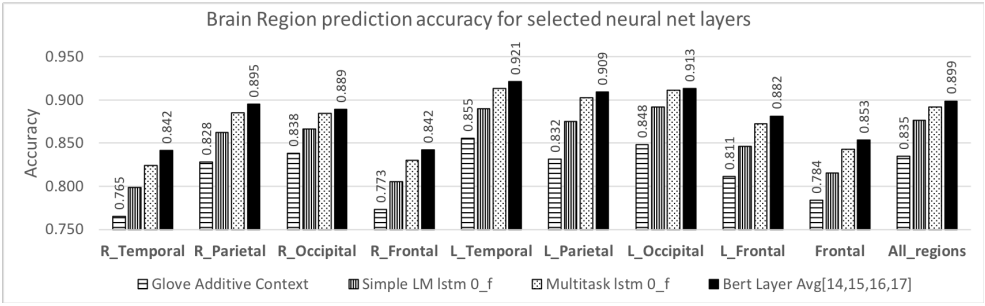


Figure 4: Pairwise accuracy of various brain regions from some selected deep neural network model layers. The left part of the brain which is considered central to language understanding is predicted with higher accuracy, especially left temporal region (L = left, R = right).

deep neural networks, as active sentences are dominant in the training data of most of the pre-trained models. We also observe that for passive sentences, our simple multitask model (trained using about 250K active and passive sentences) has a lower performance gap between active and passive sentence as compared to ELMO and BERT models. This may be due to a more balanced active and passive sentence used to train the multitask model. Noun stimuli are predicted with the highest accuracy of 0.81, while the accuracy for verbs is 0.65. Both Multitask and ELMo models dominate verb prediction results, while BERT lags in this category. Further experiments should be done to compare the ability of Transformer (Vaswani et al., 2017) versus Recurrent Neural Network based models to represent verbs.

4.2 Micro-context Experiments

In these micro-context experiments, we evaluate if our models are able to retain information from words in the sentence prior to the word being processed. For such context sensitivity tests, we only use the first repetition of the sentence shown to human subjects. This helps to ensure that the sentence has not been memorized by the subjects, which might affect the context sensitivity tests.

Training: The micro-context experiment setup is illustrated in Figure 5. To train the regression model, each training instance corresponding to a word has the form (x_i, y_i), where x_i is the layer representation for an input word i in a neural network model, and y_i is the corresponding MEG brain recording data of size 1530 (flattened 306×5). During testing, we restrict the pairwise tests to word pairs (x_i, x_j) which satisfy some conditions. For example in noun context sensitivity test, the pair of words should be such that, they appear in a sentence with the same words except the noun. We describe these candidate word test pairs, in detail, in the following sections.

In each of the following sensitivity tests, we perform a pair-wise accuracy test among the same candidate word (bold items) from sentences which

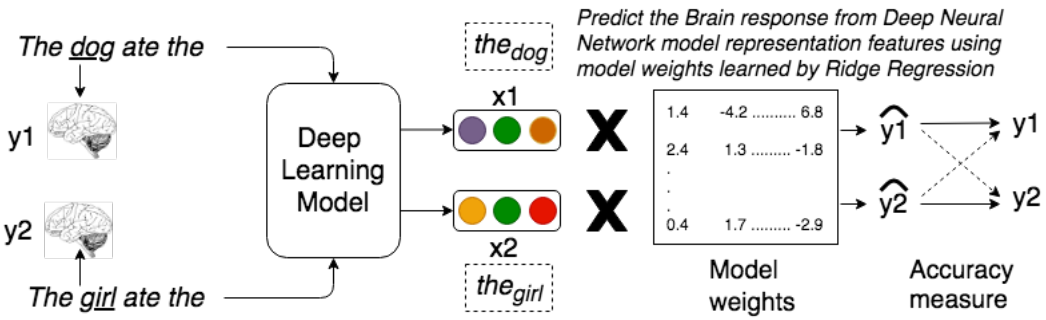


Figure 5: Experimental setup for micro-context tests. Given two sentences with similar words except one in the past (underlined), the test evaluates if the deep neural network model representation contains sufficient information to tell the two words apart. Please see Section 4.2 for more details.

are identical except for one word (underlined items). We vary the non-identical word type (noun, verb, adjective, determiner) among the two sentences to test the contribution of each of these word types to the context representation further in sentence. This test helps us understand what parts of the context are retained or forgotten by the neural network model representation. Detailed results of each test are included in the appendix section (Figure 10). Please note that the part of BERT word embedding is the sentence embedding, therefore the BERT embedding performs better than 0.5, unlike other embeddings.

4.2.1 Noun sensitivity

"*The dog ate* ***the***" vs. "*The girl ate* ***the***"

For the PassAct1 dataset, we observe that simple GloVe additive model (classification accuracy = 0.52) loses information about the noun while it is retained by most layers of other models like BERT (accuracy = 0.92), ELMo (accuracy = 0.91). Higher level layers, such as linear layer for POS-tag prediction (accuracy = 0.65), also perform poorly. This seems obvious due to the task it solves which focuses on POS-tag property at the word 'the' rather than the previous context. In summary, we observe that the language model context preserves noun information well.

4.2.2 Verb sensitivity

"*The dog saw* ***the***" vs. "*The dog ate* ***the***"

For the PassAct1 dataset, we observe that similar to noun sensitivity, most language model layers (accuracy = 0.92), except for simple GloVe Additive model, preserve the verb memory. By design, the GloVe Additive model retains little context from the past words, and therefore the result verifies the experiment setup.

4.2.3 First determiner sensitivity

"*A* ***dog***" vs. "*The* ***dog***"

For the PassAct2 dataset, we observe that determiner information is retained well by most layers. However, the shallow layers retain information better than the deeper layers. For example, BERT layer 3 (accuracy = 0.82), Multitask lstm 0_backward (accuracy = 0.82), BERT Layer 18/19 (accuracy 0.78). Since the earlier layers have a higher correlation with shallow feature processing, the determiner information may be useful for the early features in neural network representation.

4.2.4 Adjective sensitivity

"*The happy* ***child***" vs. "*The* ***child***"

For the Act3 dataset, we observe that middle layers of most models (BERT, Multitask) retain the adjective information well. However, surprisingly simple multitask model (lstm 1_forward layer accuracy = 0.89) retains adjective information better than BERT model (layer 7 accuracy = 0.84). This could be due to the importance of adjective in context for POS tag prediction. This result encourages the design of language models with diverse cost functions based on the kind of sentence context information that needs to be preserved in the final task.

4.2.5 Visualisation

We visualise the average agreement of model predicted brain activity (from BERT layer 18) and true brain activity for candidate stimuli in micro-sensitivity tests. Please note that the micro-sensitivity tests predict brain activity for stimuli with almost similar past context except one word, this makes the task harder. We preprocess the brain activity values to be +1 for all positive values and -1 for all negative values. The predicted brain

activity ($y^{'}$) and the true brain activity (y) are then compared to form an agreement activity ($y^{''}$), resulting in a zero value for all locations where the sign predicted was incorrect. We average these agreement activities ($y^{''}$) for all test examples in a cross-validation fold to form a single activity image ($Y^{''}$). Figure 8 shows $Y^{''}$ for the word '*the*' in noun-sensitivity tests Section 4.2.1 (additional results are in the appendix section). We observe that our model prediction direction agrees with brain prediction direction in most of the brain regions. This shows that our neural network layer representation can preserve information from earlier words in the sentence.

4.3 Semi-supervised training using synthesized brain activity

In this section, we consider the question of whether previously trained linear regression model (X1), which predicts brain activity for a given sentence, can be used to produce useful synthetic brain data (i.e., sentence-brain activity pairs). Constraints like high cost of MEG recording and physical limits on an individual subject during data collection, favor such synthetic data generation. We evaluate effectiveness of this synthetically generated brain data for data augmentation in the stimulus prediction task (Mitchell et al., 2008). Specifically, we train a decoding model (X2) to predict brain activity during a stimulus reading based on GloVe vectors for nouns. We consider two approaches. In the first approach, the same brain activity data as in previous sections was used. In the second approach, the real brain activity data is augmented with the synthetic activities generated by the regression model (X1).

In our experiment, we generate new sentences using the same vocabulary as the original sentences in the PassAct1 dataset. Details of the original 32 sentences (Section A.1.1) along with the 160 generated sentences (Section A.1.2) are given in the appendix section. We process the 160 generated sentences with BERT layer 18 to get word stimulus features in context. The encoding model (X1) was trained using the PassAct1 dataset. Please note that BERT layer 18 was chosen based on the high accuracy results on macro-context tests, therefore the layer aligned well with the whole brain activity. The choice of representation (deep neural network layer) to encode brain activity should be done carefully, as each representation may be good at encoding different parts of brain. A good criteria for representation selection requires further research.

To demonstrate the efficacy of the synthetic dataset, we present the accuracy in predicting noun (or verb) stimuli from observed MEG activity with and without the additional synthetic MEG data. With linear ridge regression model (X2), a GloVe (Pennington et al., 2014) feature to brain-activity prediction models were trained to predict the MEG activity when a word is observed . To test the model performance, we calculate the accuracy of the predicted brain activity given the true brain activity during a word processing (Equation 1). All the experiments use 4-fold cross-validation. Figure 7 shows the increase in the noun/verb prediction accuracy with additional synthetically generated data. The statistical significance is calculated over 400 random label permutation tests.

To summarize, these results show the utility of using previously trained regressor model to produce synthetic training data to improve accuracy on additional tasks. Given the high cost of collecting MEG recordings from human subjects and their individual capacity to complete the task, this data augmentation approach may provide an effective alternative in many settings.

5 Related Work

Usage of machine learning models in neuroscience has been gaining popularity. Methods in this field use features of words and contexts to predict brain activity using various techniques (Agrawal et al., 2014). Previous research have used functional magnetic resonance imaging (FMRI) (Glover, 2011) and Magnetoencephalography (MEG) (Hmlinen et al., 1993) to record brain activity. Prefrontal cortex in rhesus monkeys was studied in Mante et al. (2013). They showed that an appropriately trained recurrent neural network model reproduces key physiological observations and suggests a new mechanism of input selection and integration. Barak (2017) argues that RNNs with reverse engineering can provide a framework for modeling in neuroscience, potentially serving as a powerful hypothesis generation tool. Prior research by Mitchell et al. (2008), Wehbe et al. (2014b), Jain and Huth (2018), Hale et al. (2018), Pereira et al. (2018), and Sun et al. (2019) have established a general correspondence between a computational model and brain's re-

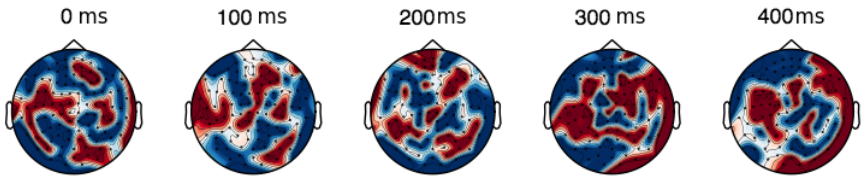
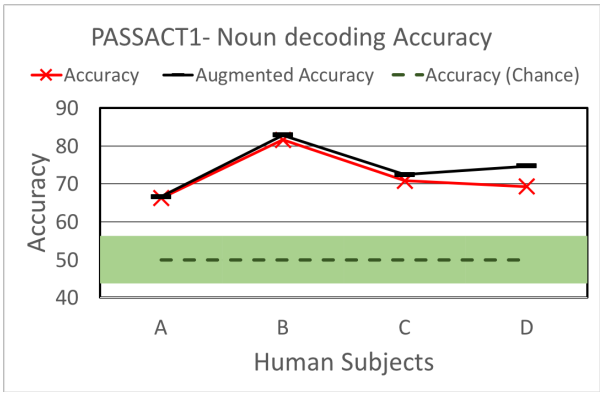
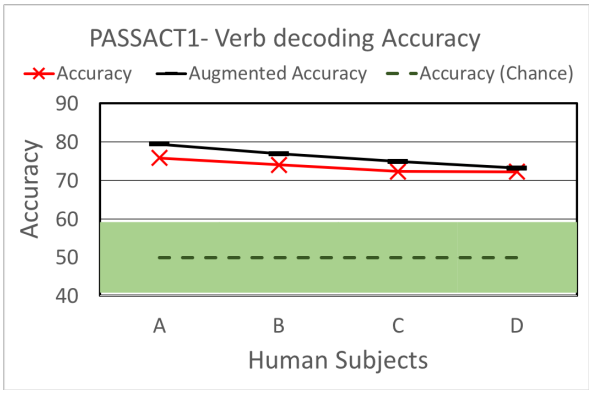


Figure 6: Average sign agreement activity for noun sensitivity stimuli '*the*'. The red and blue colored areas are the +ive and -ive signed brain region agreement respectively, while the white colored region displays brain regions with prediction error. We observe that in most regions of the brain, the predicted and true activity agree on the activity sign, thereby providing evidence that deep learning representations can capture useful information about language processing consistent with the brain recording.



(a) Noun prediction results



(b) Verb prediction results

Figure 7: Accuracy with and without synthetically generated MEG brain data on two stimuli prediction tasks: (a) Nouns (left) and (b) Verbs (right). We trained two models – one using true MEG brain recording and the other using both true and synthetically generated MEG brain data (Augmented data model). We observe that the augmented data model results in accuracy improvement on both tasks, on average **2.1%** per subject for noun prediction and **2.4%** for verb. Accuracy (chance) is the random permutation test accuracy, with the green shaded area representing standard deviation. Please see Section 4.3 for details.

sponse to naturalistic language. We follow these prior research in our analysis work and extend the results by doing a fine-grained analysis of the sentence context. Additionally, we also use deep neural network representations to generate synthetic brain data for extrinsic experiments.

6 Conclusion

In this paper, we study the relationship between sentence representations learned by deep neural network models and those encoded by the brain. We encode simple sentences using multiple deep networks, such as ELMo, BERT, etc. We make use of MEG brain imaging data as reference. Representations learned by BERT are the most effective in predicting brain activity. In particular, most models are able to predict activity in the left temporal region of the brain with high accuracy. This brain region is also known to be responsible for processing syntax and semantics for language understanding. To the best of our knowledge, this is the first work showing that the MEG data, when reading a word in a sentence, can be used to distinguish earlier words in the sentence. Encouraged by these findings, we use deep networks to generate synthetic brain data to show that it helps in improving accuracy in a subsequent stimulus decoding task. Such data augmentation approach is very promising as actual brain data collection in large quantities from human subjects is an expensive and labor-intensive process. We are hopeful that the ideas explored in the paper will promote further research in understanding relationships between representations learned by deep models and the brain during language processing tasks.

7 Acknowledgments

This work was supported by The Government of India (MHRD) scholarship and BrainHub CMU-IISc Fellowship awarded to Sharmistha Jat. We thank Dan Howarth and Erika Laing for help with MEG data preprocessing.

References

Pulkit Agrawal, Dustin Stansbury, Jitendra Malik, and Jack L. Gallant. 2014. Pixels to voxels: Modeling visual representation in the human brain. *CoRR*, abs/1407.5104.

Omri Barak. 2017. Recurrent neural networks as versatile tools of neuroscience research. *Current Opinion in Neurobiology*, 46:1 – 6. Computational Neuroscience.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL*.

Chris Dyer, Adhiguna Kuncoro, Miguel Ballesteros, and Noah A. Smith. 2016. Recurrent neural network grammars. In *Proc. of NAACL*.

Gary H Glover. 2011. Overview of functional magnetic resonance imaging. *Neurosurgery clinics of North America*, 22(2):133–vii.

Gene H. Golub, Michael Heath, and Grace Wahba. 1979. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223.

John Hale, Chris Dyer, Adhiguna Kuncoro, and Jonathan Brennan. 2018. Finding syntax in human encephalography with beam search. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2727–2736. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Matthew Honnibal and Ines Montani. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*.

Matti Hmlinen, Riitta Hari, Risto Ilmoniemi, Jukka Knuutila, and Olli V. Lounasmaa. 1993. Magnetoencephalography: Theory, instrumentation, and applications to noninvasive studies of the working human brain. *Rev. Mod. Phys.*, 65:413–.

Hakan Inan, Khashayar Khosravi, and Richard Socher. 2016. Tying word vectors and word classifiers: A loss framework for language modeling. *CoRR*, abs/1611.01462.

Shailee Jain and Alexander Huth. 2018. Incorporating context into language encoding models for fmri. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 6628–6637. Curran Associates, Inc.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature*, 521:436.

Valerio Mante, David Sussillo, Krishna V. Shenoy, and William T. Newsome. 2013. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature*, 503:78 EP –.

Tom M. Mitchell, Svetlana V. Shinkareva, Andrew Carlson, Kai-Min Chang, Vicente L. Malave, Robert A. Mason, and Marcel Adam Just. 2008. Predicting human brain activity associated with the meanings of nouns. *Science*, 320(5880):1191–1195.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python . *Journal of Machine Learning Research*, 12:2825–2830.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543.

Francisco Pereira, Bin Lou, Brianna Pritchett, Samuel Ritter, Samuel J. Gershman, Nancy Kanwisher, Matthew Botvinick, and Evelina Fedorenko. 2018. Toward a universal decoder of linguistic meaning from brain activation. *Nature Communications*, 9(1):963.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

Nicole Rafidi. 2014. The role of syntax in semantic processing: A study of active and passive sentences. [Online; accessed 2-March-2019].

Gustavo Sudre, Dean Pomerleau, Mark Palatucci, Leila Wehbe, Alona Fyshe, Riitta Salmelin, and Tom Mitchell. 2012. Tracking neural coding of perceptual and semantic features of concrete nouns. *NeuroImage*, 62:451–63.

Jingyuan Sun, Shaonan Wang, Jiajun Zhang, and Chengqing Zong. 2019. Towards sentence-level brain decoding with distributed representations. AAAI Press.

Partha Pratim Talukdar, Derry Wijaya, and Tom Mitchell. 2012. Acquiring temporal constraints between relations. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 992–1001, New York, NY, USA. ACM.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Leila Wehbe, Brian Murphy, Partha Talukdar, Alona Fyshe, Aaditya Ramdas, and Tom Mitchell. 2014a. Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PloS one*, 9:e112575.

Leila Wehbe, Ashish Vaswani, Kevin Knight, and Tom M. Mitchell. 2014b. Aligning context-based statistical models of language with brain activity during reading. In *EMNLP*, pages 233–243. ACL.

A Appendices

A.1 Dataset details

Following are the sentences used in the paper for experiments described in Section 4. We list down the sentences in PassAct1 dataset and the generated sentences in the sections Section A.1.1 and Section A.1.2 respectively. The two datasets are disjoint in terms of the sentences they contain, but are built using the same vocabulary. Datasets PassAct2 dataset and Act3 dataset are detailed in subsections A.1.3 and A.1.4 respectively.

A.1.1 PassAct1 dataset sentences

the boy was liked by the girl
the girl was watched by the man
the man was despised by the woman
the woman was encouraged by the boy
the girl was liked by the woman
the man was despised by the boy
the girl was liked by the boy
the boy was watched by the woman
the man was encouraged by the girl
the woman was despised by the man
the woman was watched by the boy
the girl was encouraged by the woman
the man was despised by the girl
the boy was liked by the man
the boy was watched by the girl
the woman was encouraged by the man
the man despised the woman
the girl encouraged the man
the man liked the boy
the girl despised the man
the woman encouraged the girl
the boy watched the woman
the man watched the girl
the girl liked the boy
the woman despised the man
the boy encouraged the woman
the woman liked the girl
the boy despised the man
the man encouraged the woman
the girl watched the boy
the woman watched the boy
the boy liked the girl

A.1.2 PassAct1 dataset artificially generated sentences

the girl was despised by the man
the man despised the girl
the man was liked by the girl
the girl was liked by the man
the girl liked the man
the man liked the girl
the girl was encouraged by the man
the man encouraged the girl
the man was watched by the girl
the girl watched the man
the boy was despised by the man
the man despised the boy
the man was liked by the boy
the boy liked the man
the man was encouraged by the boy
the boy was encouraged by the man
the boy encouraged the man
the man encouraged the boy
the man was watched by the boy
the boy was watched by the man
the boy watched the man
the man watched the boy
the man was despised by the women
the women was despised by the man
the women despised the man
the man despised the women
the man was liked by the women
the women was liked by the man
the women liked the man
the man liked the women
the man was encouraged by the women
the women was encouraged by the man
the women encouraged the man
the man encouraged the women
the man was watched by the women
the women was watched by the man
the women watched the man
the man watched the women
the girl was despised by the man
the man despised the girl
the girl was liked by the man
the man was liked by the girl
the man liked the girl
the girl liked the man
the girl was encouraged by the man
the man encouraged the girl
the man was watched by the girl
the girl watched the man
the girl was despised by the boy
the boy was despised by the girl
the boy despised the girl
the girl despised the boy
the girl was encouraged by the boy

the boy was encouraged by the girl
the boy encouraged the girl
the girl encouraged the boy
the girl was watched by the boy
the boy watched the girl
the girl was despised by the women
the women was despised by the girl
the women despised the girl
the girl despised the women
the girl was liked by the women
the women was liked by the girl
the women liked the girl
the girl liked the women
the girl was encouraged by the women
the women was encouraged by the girl
the women encouraged the girl
the girl encouraged the women
the girl was watched by the women
the women was watched by the girl
the women watched the girl
the girl watched the women
the boy was despised by the man
the man despised the boy
the man was liked by the boy
the boy liked the man
the boy was encouraged by the man
the man was encouraged by the boy
the man encouraged the boy
the boy encouraged the man
the boy was watched by the man
the man was watched by the boy
the man watched the boy
the boy watched the man
the boy was despised by the girl
the girl was despised by the boy
the girl despised the boy
the boy despised the girl
the boy was encouraged by the girl
the girl was encouraged by the boy
the girl encouraged the boy
the boy encouraged the girl
the girl was watched by the boy
the boy watched the girl
the boy was despised by the women
the women was despised by the boy
the women despised the boy
the boy despised the women
the boy was liked by the women
the women was liked by the boy
the women liked the boy
the boy liked the women
the boy was encouraged by the women
the women was encouraged by the boy
the women encouraged the boy
the boy encouraged the women
the boy was watched by the women
the women was watched by the boy
the women watched the boy
the boy watched the women
the women was despised by the man
the man was despised by the women
the man despised the women
the women despised the man
the women was liked by the man
the man was liked by the women
the man liked the women
the women liked the man
the women was encouraged by the man
the man was encouraged by the women
the man encouraged the women
the women encouraged the man
the women was watched by the man
the man was watched by the women
the man watched the women
the women watched the man
the women was despised by the girl
the girl was despised by the women
the girl despised the women
the women despised the girl
the women was liked by the girl
the girl was liked by the women
the girl liked the women
the women liked the girl
the women was encouraged by the girl
the girl was encouraged by the women
the girl encouraged the women
the women encouraged the girl
the women was watched by the girl
the girl was watched by the women
the girl watched the women
the women watched the girl
the women was despised by the boy
the boy was despised by the women
the boy despised the women
the women despised the boy
the women was liked by the boy
the boy was liked by the women
the boy liked the women
the women liked the boy
the women was encouraged by the boy
the boy was encouraged by the women
the boy encouraged the women

the women encouraged the boy
the women was watched by the boy
the boy was watched by the women
the boy watched the women
the women watched the boy

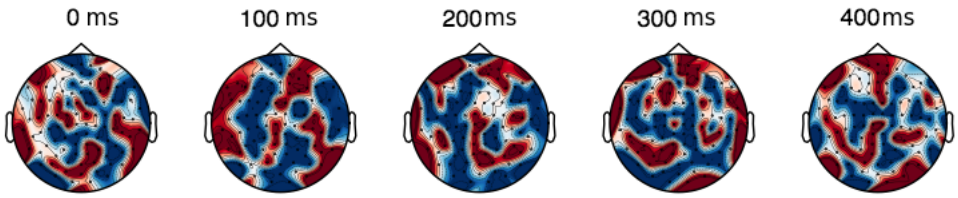
A.1.3 PassAct2 dataset sentences

the monkey inspected the peach
a monkey touched a school
the school was inspected by the student
a peach was touched by a student
the peach was inspected by the monkey
a school was touched by a monkey
a doctor inspected a door
the doctor touched the hammer
the student found a door
a student kicked the hammer
the student inspected the school
a student touched a peach
a monkey found the hammer
the monkey kicked a door
a dog inspected a hammer
the dog touched the door
a dog found the peach
the dog kicked a school
the doctor found a school
a doctor kicked the peach
a school was kicked by the dog
the peach was found by a dog
the door was touched by the dog
a hammer was inspected by a dog
the peach was kicked by a doctor
a school was found by the doctor
the hammer was touched by the doctor
a door was inspected by a doctor
the hammer was kicked by a student
a door was found by the student
the hammer was found by a monkey
a door was kicked by the monkey

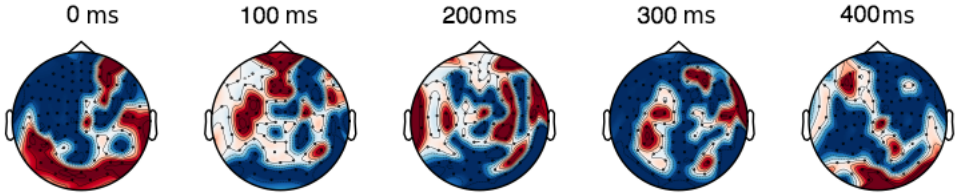
A.1.4 Act3 dataset sentences

the teacher broke the small camera
the student planned the protest
the student walked along the long hall
the summer was hot
the storm destroyed the theater
the storm ended during the morning
the duck flew
the duck lived at the lake
the activist dropped the new cellphone
the editor carried the magazine to the meeting
the boy threw the baseball over the fence
the bicycle blocked the green door
the boat crossed the small lake
the boy held the football
the bird landed on the bridge
the bird was red
the reporter wrote about the trial
the red plane flew through the cloud
the red pencil was on the desk
the reporter met the angry doctor
the reporter interviewed the politician during the debate
the tired lawyer visited the island
the tired jury left the court
the artist found the red ball
the artist hiked along the mountain
the angry lawyer left the office
the army built the small hospital
the army marched past the school
the artist drew the river
the actor gave the football to the team
the angry activist broke the chair
the cellphone was black
the company delivered the computer
the priest approached the lonely family
the patient put the medicine in the cabinet
the pilot was friendly
the policeman arrested the angry driver
the policeman read the newspaper
the politician celebrated at the hotel
the trial ended in spring
the tree grew in the park
the tourist hiked through the forest
the activist marched at the trial
the tourist ate bread on vacation
the vacation was peaceful
the dusty feather landed on the highway
the accident destroyed the empty lab
the horse kicked the fence
the happy girl played in the forest
the guard slept near the door
the guard opened the window
the glass was cold
the green car crossed the bridge
the voter read about the election
the wealthy farmer fed the horse
the wealthy family celebrated at the party
the window was dusty
the boy kicked the stone along the street
the old farmer ate at the expensive hotel

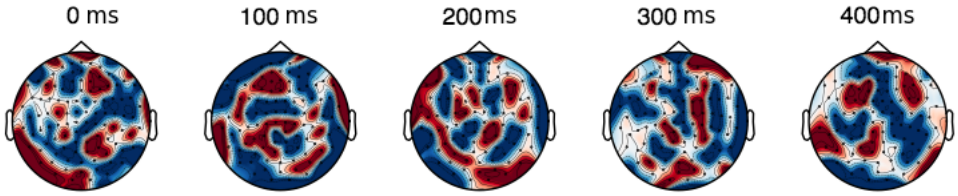
the man saw the fish in the river
the man saw the dead mouse
the man read the newspaper in church
the lonely patient listened to the loud television
the girl dropped the shiny dime
the couple laughed at dinner
the council read the agreement
the couple planned the vacation
the fish lived in the river
the flood damaged the hospital
the big horse drank from the lake
the corn grew in spring
the woman bought medicine at the store
the woman helped the sick tourist
the woman took the flower from the field
the worker fixed the door at the church
the businessman slept on the expensive bed
the businessman lost the computer at the airport
the businessman laughed in the theater
the chicken was expensive at the restaurant
the lawyer drank coffee
the judge met the mayor
the judge stayed at the hotel during the vacation
the jury listened to the famous businessman
the hurricane damaged the boat
the journalist interviewed the judge
the dog ate the egg
the doctor helped the injured policeman
the diplomat bought the aggressive dog
the council feared the protest
the park was empty in winter
the parent watched the sick child
the cloud blocked the sun
the coffee was hot
the commander ate chicken at dinner
the commander negotiated with the council
the commander opened the heavy door
the old judge saw the dark cloud
the young engineer worked in the office
the farmer liked soccer
the mob approached the embassy
the mob damaged the hotel
the minister spoke to the injured patient
the minister visited the prison
the minister found cash at the airport
the minister lost the spiritual magazine
the mouse ran into the forest
the parent took the cellphone
the soldier delivered the medicine during the flood
the soldier arrested the injured activist
the small boy feared the storm
the egg was blue
the editor gave cash to the driver
the editor damaged the bicycle
the expensive camera was in the lab
the engineer built the computer
the family survived the powerful hurricane
the child held the soft feather
the clever scientist worked at the lab
the author interviewed the scientist after the flood
the artist shouted in the hotel



(a) Verb sign agreement image between true and predicted brain activations



(b) Adjective sign agreement image between true and predicted brain activations



(c) Determiner sign agreement image between true and predicted brain activations

Figure 8: Sign agreement image for verb, determiner and adjective sensitivity test stimuli. The red and blue colored areas are the +ive and -ive signed brain region agreement. While, the white colored region displays brain regions with prediction error. We observe that in most regions of the brain the predicted and true image agree on the activity sign, thereby proving that deep learning representations can capture useful information about language processing.

Layers	Noun	Verb	Passive	Active
bert_layer_1_emb	0.795	0.62	0.87	0.926
bert_layer_10	0.795	0.618	0.868	0.931
bert_layer_11	0.8	0.636	0.869	0.931
bert_layer_12	0.802	0.648	0.868	0.93
bert_layer_13	0.802	0.625	0.864	0.933
bert_layer_14	0.796	0.636	0.865	0.931
bert_layer_15	0.803	0.633	0.866	0.928
bert_layer_16	0.801	0.609	0.873	0.93
bert_layer_17	0.816	0.608	0.874	0.927
bert_layer_18	0.81	0.597	0.872	0.925
bert_layer_19	0.807	0.603	0.872	0.925
bert_layer_2	0.79	0.621	0.867	0.926
bert_layer_20	0.807	0.605	0.87	0.927
bert_layer_21	0.807	0.605	0.871	0.924
bert_layer_22	0.805	0.605	0.871	0.924
bert_layer_23	0.808	0.597	0.872	0.923
bert_layer_24	0.803	0.583	0.87	0.918
bert_layer_3	0.788	0.614	0.866	0.925
bert_layer_4	0.789	0.63	0.864	0.924
bert_layer_5	0.791	0.629	0.866	0.923
bert_layer_6	0.79	0.639	0.866	0.922
bert_layer_7	0.792	0.631	0.864	0.925
bert_layer_8	0.789	0.633	0.866	0.926
bert_layer_9	0.792	0.631	0.868	0.931
elmo_emb_layer0	0.664	0.629	0.818	0.828
elmo_lstm_layer1	0.79	0.656	0.869	0.923
elmo_lstm_layer2	0.785	0.659	0.866	0.923
glove_additive_context_emb	0.78	0.643	0.836	0.847
glove_additive_word_emb	0.662	0.631	0.822	0.836
LM_bidirec_embedding	0.669	0.639	0.815	0.808
LM_bidirec_linear_nw	0.768	0.586	0.849	0.892
LM_bidirec_lstm_0_backward	0.668	0.532	0.723	0.764
LM_bidirec_lstm_0_forward	0.716	0.606	0.86	0.887
LM_bidirec_lstm_1_backward	0.745	0.581	0.771	0.828
LM_bidirec_lstm_1_forward	0.684	0.55	0.799	0.733
Multitask_bidirec_embedding_nw	0.664	0.629	0.823	0.828
Multitask_bidirec_embedding_pos	0.662	0.631	0.822	0.836
Multitask_bidirec_linear_nw	0.763	0.593	0.852	0.902
Multitask_bidirec_linear_pos	0.73	0.64	0.822	0.897
Multitask_bidirec_lstm_0_backward	0.742	0.651	0.834	0.906
Multitask_bidirec_lstm_0_forward	0.772	0.624	0.866	0.913
Multitask_bidirec_lstm_1_backward	0.772	0.645	0.836	0.914
Multitask_bidirec_lstm_1_forward	0.746	0.648	0.859	0.916
random_word_emb	0.525	0.502	0.521	0.463

Figure 9: Pairwise Accuracy of predicting brain encodings for noun, verb, passive & active sentences. For each of the category the Ridge regression model is learned and tested on the stimulus subset like only nouns or only passive sentences. The color of a cell represents the value within overall accuracy scale with red indicating small values, yellow intermediate and green high values. We observe that Nouns are predicted better than verbs. And active sentences are predicted better than passive sentences.

Layer Name	Noun	Verb	Determiner	Adjective
bert_layer_1_emb	0.927	0.922	0.794	0.754
bert_layer_10	0.927	0.927	0.813	0.744
bert_layer_11	0.927	0.927	0.798	0.846
bert_layer_12	0.927	0.927	0.806	0.769
bert_layer_13	0.927	0.927	0.821	0.846
bert_layer_14	0.922	0.927	0.806	0.795
bert_layer_15	0.922	0.927	0.798	0.821
bert_layer_16	0.922	0.927	0.794	0.846
bert_layer_17	0.922	0.927	0.79	0.846
bert_layer_18	0.922	0.927	0.786	0.821
bert_layer_19	0.927	0.927	0.786	0.718
bert_layer_2	0.927	0.927	0.802	0.795
bert_layer_20	0.927	0.927	0.806	0.821
bert_layer_21	0.927	0.927	0.794	0.846
bert_layer_22	0.927	0.927	0.802	0.769
bert_layer_23	0.927	0.927	0.813	0.769
bert_layer_24	0.927	0.927	0.806	0.769
bert_layer_3	0.927	0.927	0.821	0.744
bert_layer_4	0.927	0.927	0.794	0.744
bert_layer_5	0.927	0.927	0.798	0.769
bert_layer_6	0.927	0.927	0.81	0.821
bert_layer_7	0.927	0.927	0.802	0.846
bert_layer_8	0.927	0.927	0.794	0.795
bert_layer_9	0.927	0.927	0.786	0.744
elmo_emb_layer0	0.5	0.5	0.5	0.5
elmo_lstm_layer1	0.917	0.917	0.798	0.769
elmo_lstm_layer2	0.901	0.911	0.806	0.718
glove_additive_context_emb	0.526	0.661	0.544	0.492
glove_additive_word_emb	0.5	0.5	0.5	0.5
LM_bidirec_embedding	0.5	0.5	0.5	0.5
LM_bidirec_linear_nw	0.927	0.927	0.778	0.636
LM_bidirec_lstm_0_backward	0.927	0.927	0.782	0.518
LM_bidirec_lstm_0_forward	0.927	0.927	0.782	0.703
LM_bidirec_lstm_1_backward	0.927	0.927	0.8	0.626
LM_bidirec_lstm_1_forward	0.927	0.927	0.802	0.636
Multitask_bidirec_embedding_nw	0.5	0.5	0.5	0.5
Multitask_bidirec_embedding_pos	0.5	0.5	0.5	0.5
Multitask_bidirec_linear_nw	0.927	0.927	0.813	0.744
Multitask_bidirec_linear_pos	0.656	0.818	0.587	0.456
Multitask_bidirec_lstm_0_backward	0.927	0.927	0.821	0.795
Multitask_bidirec_lstm_0_forward	0.927	0.927	0.798	0.754
Multitask_bidirec_lstm_1_backward	0.927	0.927	0.778	0.703
Multitask_bidirec_lstm_1_forward	0.927	0.927	0.774	0.897

Figure 10: Micro-context sensitivity test results for all the layers. The color of a cell represents the value within overall accuracy scale with red indicating small values, yellow intermediate and green high values. We observe that noun and verbs are retained in the context with same accuracy followed by determiner and then adjective.