Relating Simple Sentence Representations in Deep Neural Networks and the Brain

Sharmistha Jat, Hao Tang, Partha Talukdar, Tom Mitchell
Association for Computational Linguistics (ACL) 2019
{sharmisthaj,ppt}@iisc.ac.in,htang1@alumni.cmu.edu, tom.mitchell@cs.cmu.edu

Contributions

Our main contributions are:

1. 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 areas.
2. We demonstrate that deep networks are capable of predicting change in brain activity due to differences in previously processed words in the sentence.
3. We demonstrate effectiveness of using deep networks to synthesize brain data for downstream data augmentation.

MEG Dataset

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

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.

Simple Sentence Corpus

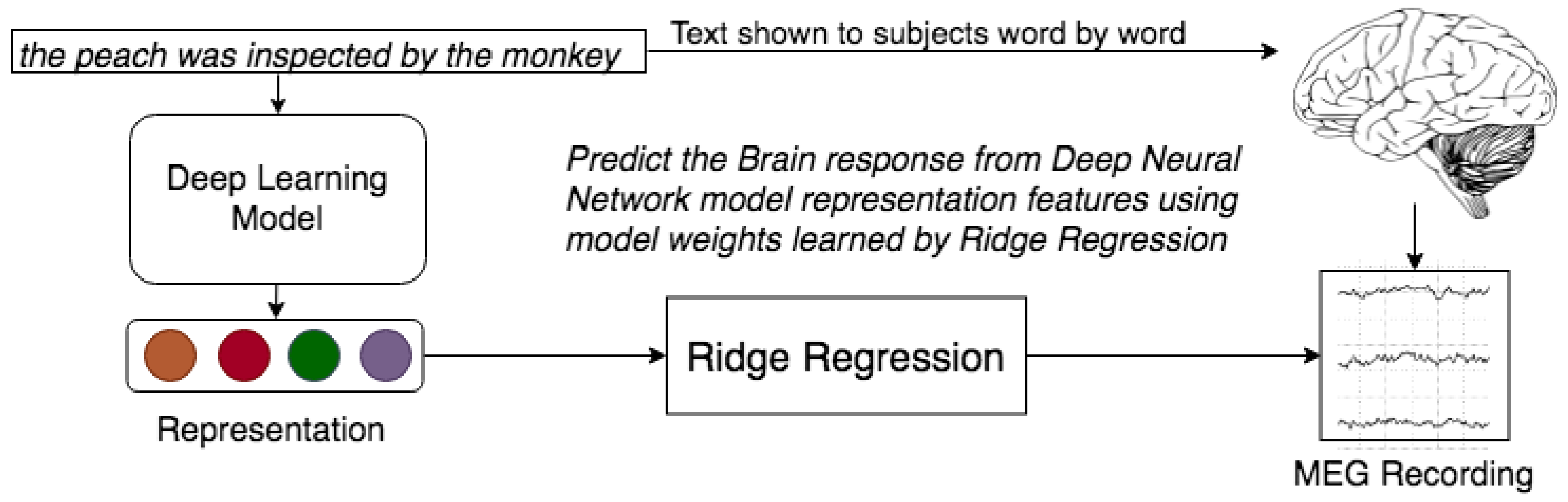
We created a new Simple Sentence Corpus (SSC), consisting of a mix of simple active and passive sentences of the form *"the woman encouraged the girl"* and *"the woman was encouraged by the boy"*, respectively, for training custom DNNs. The SSC dataset consists of 256,145 sentences constructed using the following two sets.

Source	#Sentences	Voice
Wikipedia	125,900	P+A
NELL Triples	130,245	P+A

Acknowledgement

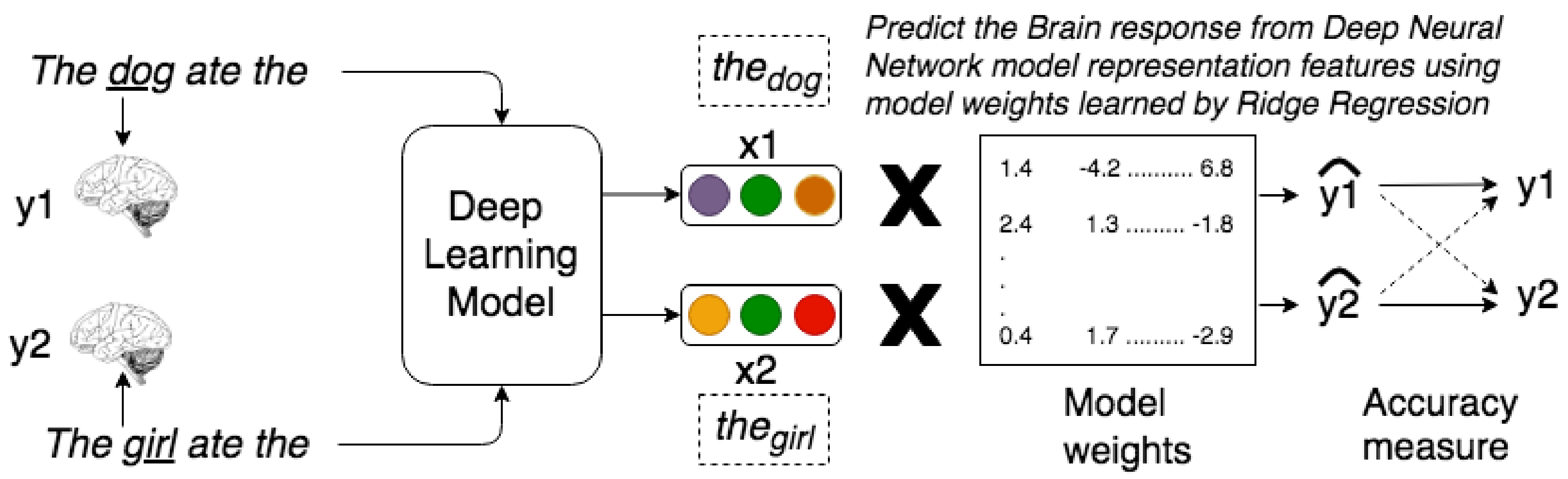
This work was supported by The Government of India (MHRD) scholarship and BrainHub CMU-IISc Fellowship awarded to Sharmistha Jat.

Macro-context tests


the peach was inspected by the monkey
Text shown to subjects word by word
Deep Learning Model
Representation
Ridge Regression
MEG Recording

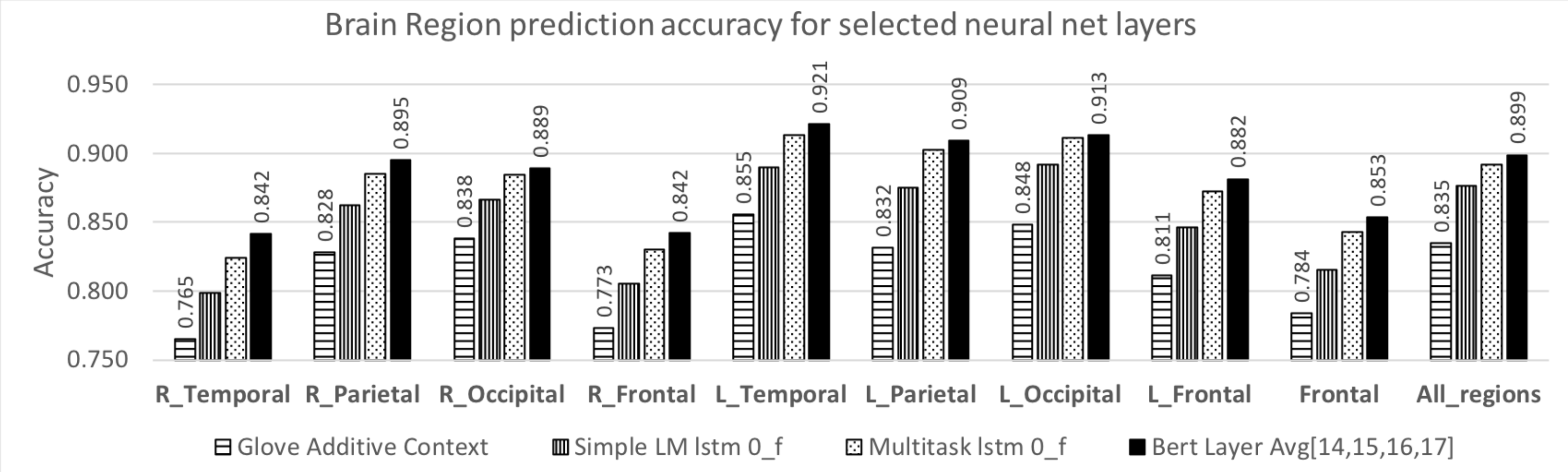
Predict the Brain response from Deep Neural Network model representation features using model weights learned by Ridge Regression

Micro-context tests



Predict the Brain response from Deep Neural Network model representation features using model weights learned by Ridge Regression

Macro-context results



Micro-context results

1. **Noun**:
   "*The* *dog* *ate* ***the***" vs. "*The* *girl* *ate* ***the***";
   Noun information is retained by most layers of other models like BERT (accuracy = 0.92), ELMo (accuracy = 0.91)
2. **Verb**:
   "*The dog* *saw* ***the***" vs. "*The dog* *ate* ***the***";
   Most language model layers retain verb information (accuracy = 0.92)
3. **Adjective**:
   "*The* *happy* ***child***" vs. "*The* ***child***";
   We observe that middle layers of most models (BERT, Multitask) retain the adjective information well
4. **First Determiner**:
   "*A* ***dog***" vs. "*The* ***dog***";
   The shallow layers retain determiner information better than the deeper layers. BERT layer 3 (accuracy = 0.82), Multitask lstm 0_backward (accuracy = 0.82), BERT Layer 18/19 (accuracy 0.78)

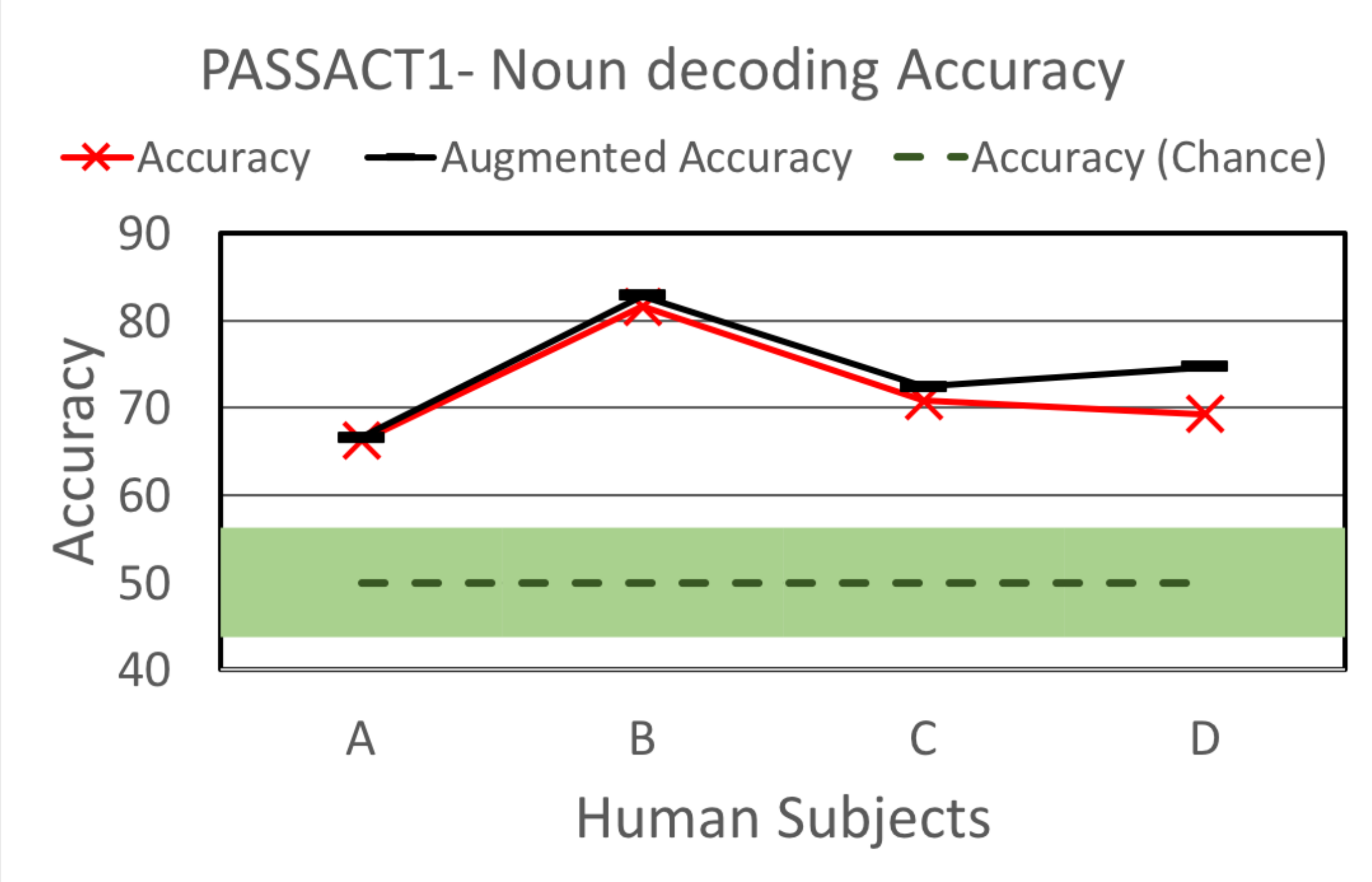
Sentence Representation:

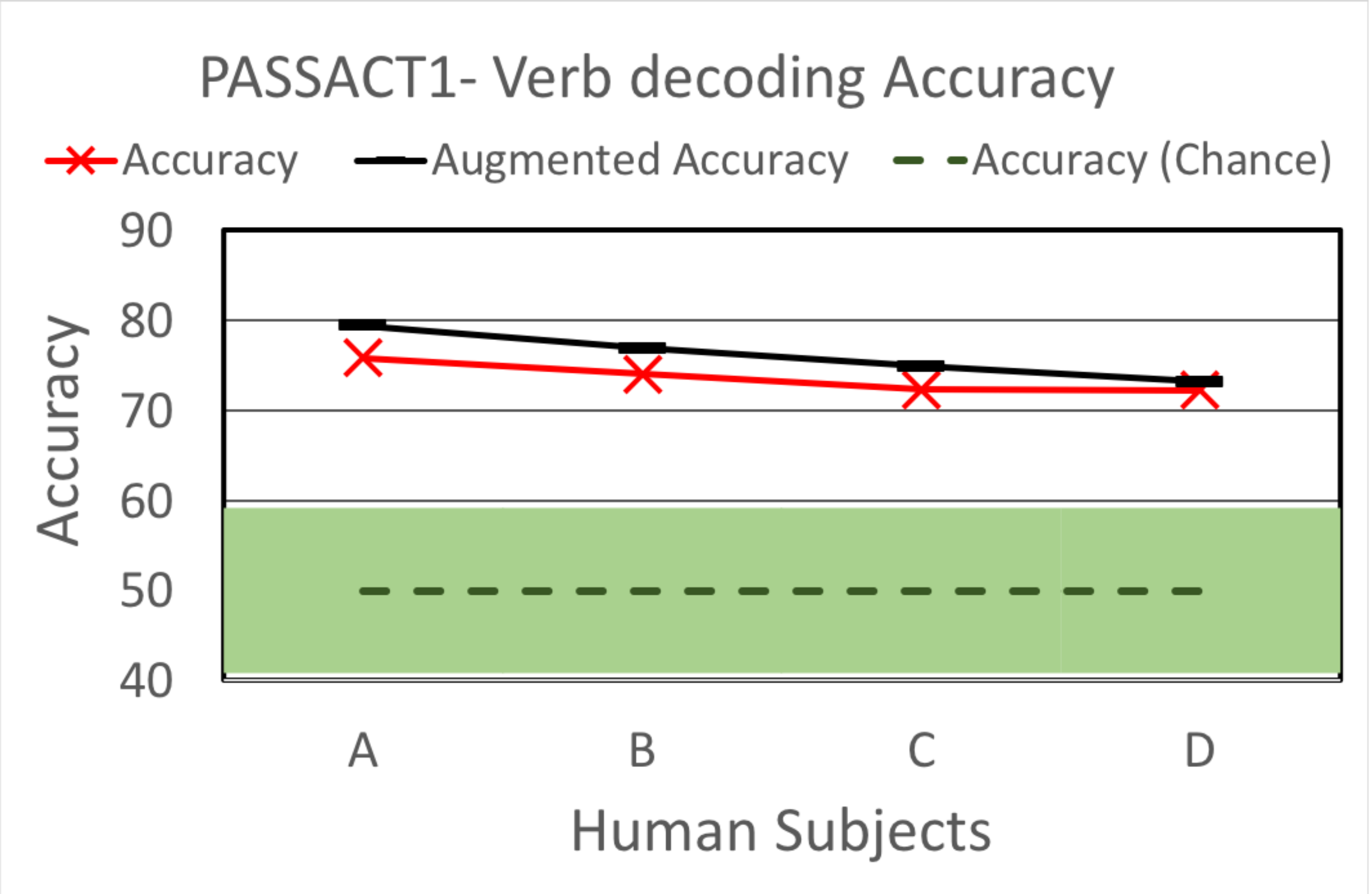
1. **Random Embedding Model**
2. **GloVe Additive Embedding Model**
3. **Simple Bi-directional LSTM Language Model**
4. **Multi-task Model**
5. **ELMO**
6. **BERT**

More details about each model in the paper.

Synthesized Brain Activity

Our 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.


PASSACT1- Noun decoding Accuracy
Accuracy
Augmented Accuracy
Accuracy (Chance)
Accuracy
A
B
C
D
Human Subjects


PASSACT1- Verb decoding Accuracy
Accuracy
Augmented Accuracy
Accuracy (Chance)
Accuracy
A
B
C
D
Human Subjects