IJCNN 2019. International Joint Conference on Neural Networks. Budapest, Hungary. 14-19 July 2019

StepEncog: A Convolutional LSTM Autoencoder for Near-Perfect fMRI Encoding

Subba Reddy Oota*[1], Vijay Rowtula*[1], Manish Gupta[1,3], Raju S. Bapi[1,2]
[1] IIIT Hyderabad, [2] University of Hyderabad, [3] Microsoft India
oota.subba@students.iiit.ac.in,vijay.rowtula@research.iiit.ac.in, and {manish.gupta,raju.bapi}@iiit.ac.in

***Abstract*—Learning a forward mapping that relates stimuli to the corresponding brain activation measured by functional magnetic resonance imaging (fMRI) is termed as estimating *encoding models*. Computational tractability usually forces current encoding as well as decoding solutions to typically consider only a small subset of voxels from the actual 3D volume of activation. Further, while reconstructing stimulus information from brain activation (brain decoding) has received wider attention, there have been only a few attempts at constructing encoding solutions in the extant neuro-imaging literature. In this paper, we present StepEncog, a convolutional LSTM autoencoder model trained on fMRI voxels. The model can predict the entire brain volume rather than a small subset of voxels, as presented in earlier research works. We argue that the resulting solution avoids the problem of devising encoding models based on a rule-based selection of informative voxels and the concomitant issue of wide spatial variability of such voxels across participants. The perturbation experiments indicate that the proposed deep encoder indeed learns to predict brain activations with high spatial accuracy. On challenging universal decoder imaging datasets, our model yielded encouraging results.**

***Index Terms*—fMRI, CNN-LSTM, encoding, functional MRI, Convolutional Neural Networks, Deep Learning**

I. Introduction

Apart from clinical use for diagnosing a variety of clinical conditions such as depression, Alzheimer's, dementia etc., functional magnetic resonance imaging (fMRI) studies are conducted extensively in neuroscience research to understand how knowledge is represented in the brain [1], [2]. This, in turn, enables the study of how the brain recovers partially from a stroke and to test how well a drug or behavioral therapy works [3], [4]. Since the work of Mitchell et al. [5], there has been an increasing interest in using computational models to interpret neural activity using either the decoding (learning brain activation to stimuli mapping) or encoding models (learning a mapping from stimuli to brain activation) [6]–[8]. An encoding model that predicts brain activity in response to stimuli is important for neuroscientists who can use the model predictions to investigate and test hypotheses about the transformation from stimulus to brain response in patients. In the context of fMRI, the voxel response is a proxy for neural activity and so an fMRI encoding model predicts voxel responses.

Recent approaches of modeling fMRI data use training data set to estimate a separate model for each recorded voxel. Encoding models have grown in popularity in fMRI [8], electro-corticography [7], and EEG/MEG [6]. Together, these models describe how information of the sensory stimulus or visual function is encoded in the measured brain activity [8]. The recent success of deep learning based word representations has raised the question of whether these models might be able to make association between brain activations and language [9]. Word embedding features were used to build encoding systems [10], [11] and comprise of semantic representation of brain from neurally-inspired models of vision [12] or audition [13]. Although voxel-specific encoding models provide the cortical representation of semantic content by visual and linguistic stimuli, they are typically measured for individual participant alone [14], [15]. Some methods rely on the parametric regression that assumes that the response is linearly related to stimulus features after fixed parametric nonlinear transformation(s) [5]. A recent work used encoding models with rich contextual representations derived from a long short-term memory (LSTM) networks based language model [16]. However, it is very difficult to estimate a model with minimal training data, especially when there are hundreds of stimulus features that need to be mapped to thousands of voxels.

In this paper, we present StepEncog: a deep autoencoding model that predicts the complete brain activity associated with multi-modal forms of concrete nouns, which include words and images. The theory underlying this computational model is that when a deep autoencoder is trained on large corpus, the model can transform the stimulus $\boldsymbol{S}$ which is either a word or image (or both) into corresponding 3D brain encoding $\boldsymbol{E}$. To meet the demand for larger training corpus for deep learning models, we split the 3D volume representation from the fMRI dataset into several 2D slices.

Our experiments include training separate models per subject, and training a single model across all subjects based on different modalities with the assumption regarding the underlying capability of the model to capture the complexities across subjects and stimuli. We present experimental evidence showing that the best encoding model is achieved when it is presented with multi-modal stimulus information rather than words or images alone. Our main contribution in this paper is a unified deep encoding model which can model all of the fMRI voxel activations for multiple users for given stimulus, and predict the activations on unseen stimulus. This can be seen as a departure from traditional method of using multi-regression methods on selected subset of voxels.

*The first two authors made equal contribution.

978-1-7281-2009-6/$31.00 paper N-19397.pdf

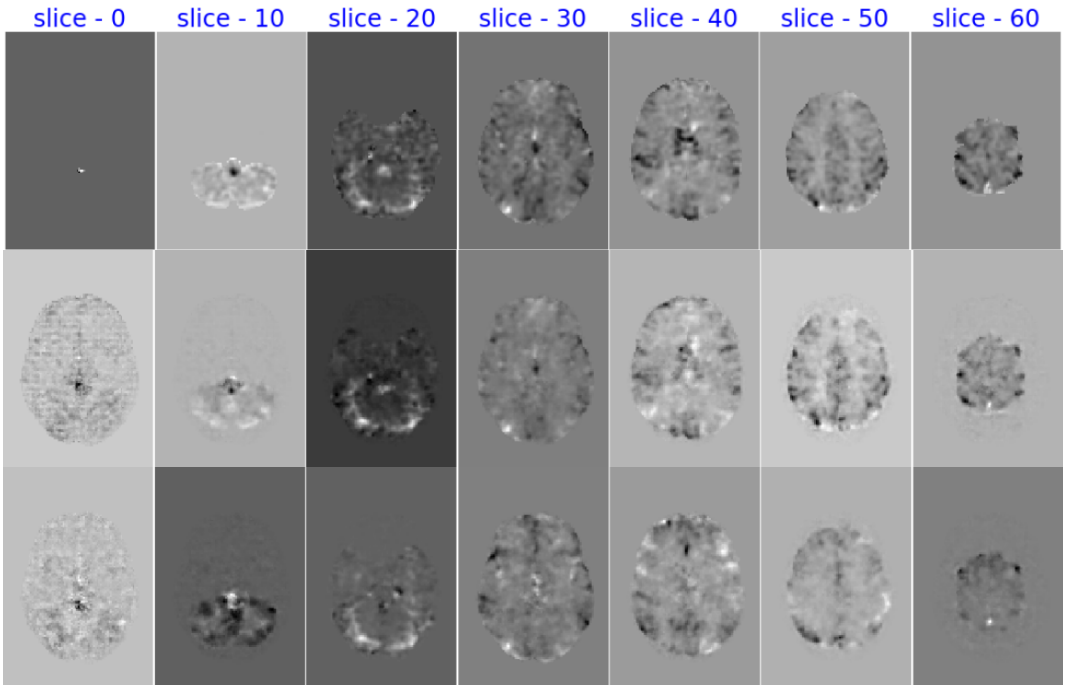

Fig. 1. The sequence of slices show (i) actual brain activation for the word "Apartment" after converting voxel activation per subject into 70 slices (top row), (ii) activation prediction by model trained on word "Apartment" + image (multi-modal) embeddings (middle row), and (iii) activation prediction by model trained on just GloVe embedding of "Apartment" (bottom row). Note that we show 7 of the 70 slices in each row.

II. Our Approach

Traditional methods either used a set of selective voxels from the dataset [17], [18] or handpicked region-based voxels to model brain encoding [11] and decoding. In the next sections, we discuss the disadvantages of such methods and our enhancements to overcome the issues.

A. Voxels and Semantic slices

A voxel is cuboid-shaped and smaller voxels contain fewer neurons on average and hence have less signal than larger voxels [19]. The scanner platform generates a 3D volume of the subject's head at the rate of one every repetition time (TR). The volume consists of an array of voxel intensity values, one value per voxel in the scan. The three-dimensional volume of the subject's head comprises several voxels arranged sequentially and can be unfolded into a single line (raster coding). Earlier studies used a subset of voxels for learning encoding models using multiple regression (ridge regression) to obtain maximum likelihood estimates of the voxel values. That is, obtain a set of voxel values that minimizes the sum of squared error in reconstructing the training fMRI images [5], [16].

Though earlier experiments were conducted with minimal subsets, behavioral and long-term studies of subjects may require the generation of the entire 3D fMRI volume when the subject is tested with various stimuli [20]. Also, voxel selection based on stimuli and subject is a difficult task which can result in vast output variation when a different set of stimuli are employed for analysis. This creates a necessity for encoding models that are capable of generating a complete 3D volume (all voxels) of the subject's brain based on past fMRI history. We attempted to perform such task by utilizing all voxels in the training data [18], by converting 3D fMRI volume to sequences of 2D fMRI slices. We argue that the slices provide enough semantic encoding information to train a spatio-temporal model, since we observed a gradual change in activation in regions across multiple slices, as seen in Figure I. We compare results from our models with the baseline results using multiple regression method to validate the advantages of our approach.

B. StepEncog Architecture

We used a CNN-LSTM (Convolutional Neural Network-Long Short Term Memory) based autoencoder model, whose architecture is inspired from [21], [22]. But, unlike regular autoencoders in literature [23]–[25] which use the same stimulus for both input (with noise) and output, we use output stimulus which is same as input but "one step ahead" of the input stimulus. Figure 2 describes a basic overview, where CNNs are used for fMRI slice encoding and decoding and LSTMs to learn temporal/semantic features across fMRI slices. We use convolution layers to learn feature representation from images. They have been widely used and studied for image tasks, and are currently state-of-the-art for object classification and detection [26]–[28]. Both the encoder and decoder have CNN layers with 64, 128 and 256 filters respectively, with ReLU activation, stride size 2 and "same" size padding. Our

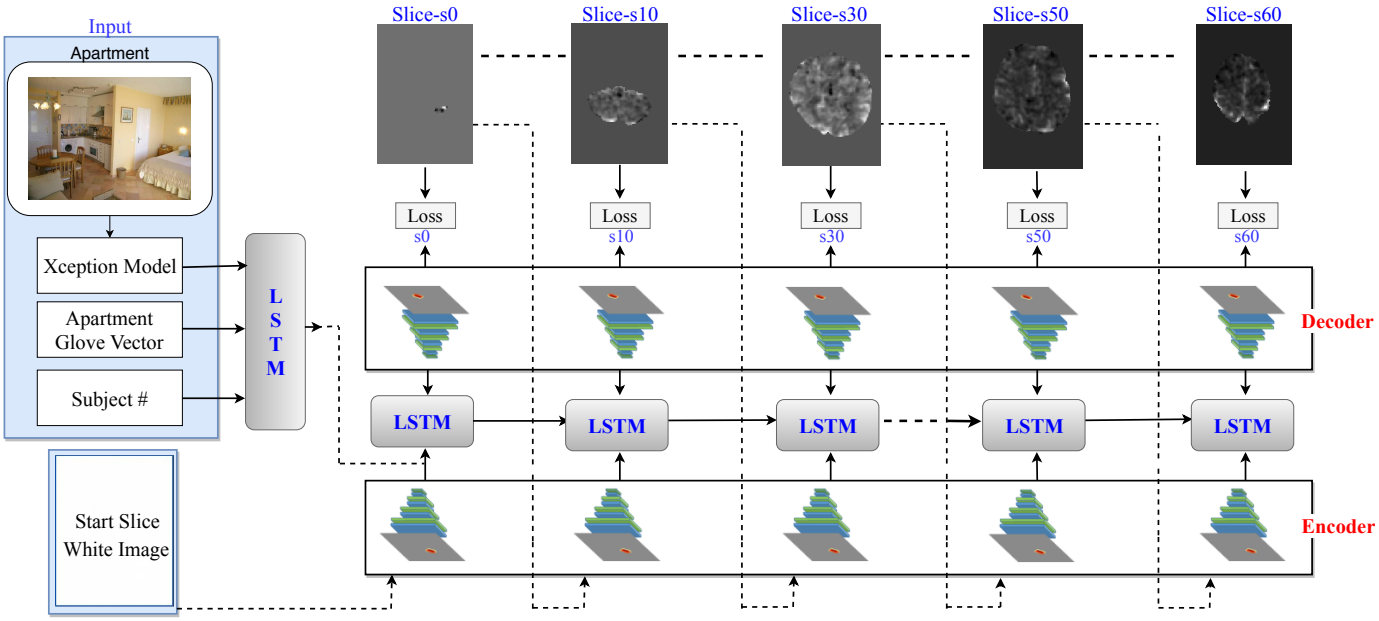

Fig. 2. Architecture of the StepEncog: the Convolutional LSTM autoencoder model used for our experiments. We used multi-modal embedding along with fMRI slices as input, and "step-ahead" fMRI slices as output.

particular choice of CNN layers and filters yields the best performance on our datasets.

LSTM networks are capable of modeling sequential/temporal aspects of data and hence are used widely for text, videos, and time-series analysis [29]. In our scenario, the model needs to capture the gradual variations that occur across fMRI 2D slices. We experimented with multiple layers of LSTMs to capture the semantic relation between slice-wise representations from CNN layers. Hence, multiple layers of LSTMs (128, 256, 512) were used as latent layers. We tried ConvLSTM, which has convolutional structures in both the input-to-state and state-to-state transitions as described in [30] and bidirectional LSTM instead of CNN and LSTM layers respectively. Both the models gave similar results when compared to the architecture described above, hence we report only CNN-LSTM based architecture results in the experiments section. The model is implemented in Keras with TensorFlow backend [31] with cross-entropy loss and Adam optimizer [32]. We used early-stopping method to stop model training when the loss starts to plateau.

Figure 2 shows an overview of our model where the input is multi-modal features of text, image, subject information along with a "start" slice, which is a blank image with white pixels to mark the beginning of 2D fMRI slices. Pre-trained model feature representations were used for both text, image stimulus and subject information was also added to the model as input. The multi-modal features representations, pass through two independent encoding layers of LSTM before concatenating to the output of CNN encoder to create a common input to the decoder. The model uses fMRI 2D slices as input and "one step ahead" slices as output during training. During testing, only the multi-modal input (image, word embedding, subject information, and start slice) is given to initiate the cascade of predictions. The model uses its own output at time step t as input in time step $t+1$, for 70 steps (for 70 fMRI slices as discussed in Section III-A), thus generating all the 2D slices required to regenerate complete 3D brain volume of voxels. A similar architecture is used to train both subject-specific model and multi-subject model. In the experiments and results section, we provide an in-depth analysis of model hyper-parameters and training.

C. Multi-modal Semantic models

For multi-modal learning [33], the model takes a corpus of images along with related and relevant word vectors as input and finds a correspondence between the two modalities. For the linguistic input, we use the popular context-predicting text-based semantic model GloVe [9] to obtain a 300-dimensional word embedding which represents the concept word. We also tested the model with word2vec embeddings [34], but we observed that our model performed better with GloVe word embeddings.

For image feature representation, we used the output from fully-connected layers of pre-trained neural network. The pre-trained models are usually trained on very large scale image classification problem datasets. The convolutional layers and fully connected layers act as feature extractors and they tend to learn very good discriminating features. Hence, image representation comprising 2048 features is obtained by using the output of the fully connected layer of pre-trained Xception model [35]. As described in Section II-B, we use word vector embeddings, the corresponding image features and fMRI slice as input to three independent encoders. The output of encoders is concatenated to provide input to the decoder. It should

TABLE I
LOSS OF THE PREDICTED BRAIN RESPONSES TO ACTUAL. PERFORMANCE RESULTS FOR INDIVIDUAL SUBJECTS ARE SHOWN SEPARATELY FOR SCENARIOS WHEN BASELINE MODEL (RIDGE REGRESSION), MULTI-MODAL (WITHOUT LSTM), AND MULTI-MODAL (CNN-LSTM) WAS UTILIZED.

Scenarios	Sub-0	Sub-1	Sub-2	Sub-3	Sub-4	Sub-5	Sub-6	Sub-7	Sub-8	sub-9	Sub-10	Sub-13	Sub-14	Sub-15	Sub-16	Sub-17
(1)	0.079	0.076	0.066	0.079	0.081	0.066	0.068	0.056	0.071	0.069	0.067	0.066	0.060	0.074	0.072	0.061
(2)	0.0015	0.0013	0.0014	0.0017	0.0015	0.0014	0.0018	0.0013	0.0012	0.0014	0.0014	0.0016	0.0022	0.0016	0.0017	0.0029
(3)	0.00079	0.00044	0.00063	0.00043	0.00058	0.00058	0.00059	0.00057	0.00044	0.00064	0.00054	0.00069	0.00047	0.00053	0.00089	0.00057
(1) Baseline (Ridge Regression), (2) Multi-Modal (Without LSTM), (3) Multi-modal (CNN-LSTM)																

be noted that the voxel information from 3D volume is converted into several 2D slices and the model needs to learn spatio-temporal features in these slices. Hence we maintain a batch size equal to the number of 2D brain slices, which represent the number of steps as required during LSTM model training. Since we use "step-ahead" slice as output for the corresponding input slice, a blank white image is prepended at the beginning of input slice sequence.

The StepEnCog autoencoder is expected to predict all 2D fMRI slices when stimuli are presented in the form of word, image, subject index and start slice. Hence during training, the actual multi-modal input is available only at the beginning of sequence. We provided zero value embeddings as text and image stimulus during the rest of the steps of a sequence. During testing of the model, the multi-modal embeddings along with "start" slice are given to predict the first actual fMRI slice. This output is reused as input in the next step along with zero-valued embeddings to generate second fMRI slice. This process continues till all 2D slices of a sequence (per subject, per word) are predicted by the model.

We initially trained separate models for 16 different subjects. The CNN-LSTM model described in Section II-B was used without the subject-index information. The trained model per user was expected to predict the brain responses using unknown stimuli for individual user (that is, *within-subject* prediction). The results are discussed in detail, in Section III-B. We later trained subject-independent model where information about the user (subject-index) is passed along the other multi-modal inputs. The trained model was able to predict the brain responses using unknown stimuli for any known user from training data. The experiments and results are discussed in the next sections.

III. EXPERIMENTS

A. Dataset

We used data from paradigm 1 of fMRI-Experiment 1 [18], where authors conducted experiments with multiple subjects by showing various forms of stimulus (sentence, picture, or both). Paradigm 1 contains three experiments. (i) In the first experiment, the target word was presented in the context of a sentence that made the relevant meaning salient. (ii) In the second, the target word was presented with a picture that depicted some aspect(s) of the relevant meaning. (iii) In the third, the target word was presented in a multi-modal form where both word and image were used. We retrieved 5 images per word from the image stimuli corpus for the 180 concepts (pictures) of Experiment 1 in Pereira et al.'s [18] dataset.

This fMRI dataset provided by the authors was collected from a total of 16 participants (subjects). For each participant in paradigm 1, a total set of 180 words (128 nouns, 22 verbs, 29 adjectives and adverbs, and 1 function word) were used as stimuli in multi-modal form (word, picture). The dataset contains fMRI captured as 128×88 voxel windows arranged as 85 slices, per subject per stimulus. Out of 85 slices, we ignored the first 9 slices and the last 7 slices since no activation is highlighted in any of the brain regions in these slices. Hence the number of slices is 69 plus one blank slice, making a total of 70 2D fMRI slices.

B. Results and Discussion

1) Baseline Experiments: We performed baseline experiments on the dataset using multi-regression to predict the voxel values first. As described in earlier research works [5], [16], a separate encoding model is used to estimate each voxel values in the training dataset, and the estimated model is used to predict voxel values on testing data. From Table I, we can observe the subject-wise testing loss using ridge regression. Before training the actual model described in Section II-B, we trained a similar model, without the LSTMs as latent layers. This experiment was conducted to understand the performance of StepEncog autoencoder on 2D fMRI slices, without temporal learning aspect. From Table I, we can clearly observe that the performance of our model without the LSTM layers is better than the multi-regression results, though it is not better than the final model. As observed in Figure I, the variation in brain response for the stimulus provided changes gradually across the 2D fMRI slices. A temporal learning element is required to capture this variation across slices and hence we added LSTM layers to our StepEncog model. We experimented with multiple layers (3 layers) of LSTM, inspired by the experiments from [16]. We observed that the model with 2 layers of LSTM gave us the best results and hence we report the results of models with 2 latent LSTM layers. The overall visual performance of our model is depicted using voxel activation locations in Figure 4. In the subsequent sections, we shall discuss the experiments of our StepEncog model with LSTM layers and analyze the results in detail.

2) Multi-modal Learning per Subject: Using the approach discussed in Section II, we trained separate encoding models per subject as the first experiment. The encoding performance was evaluated by testing models using subsets of 20 con-

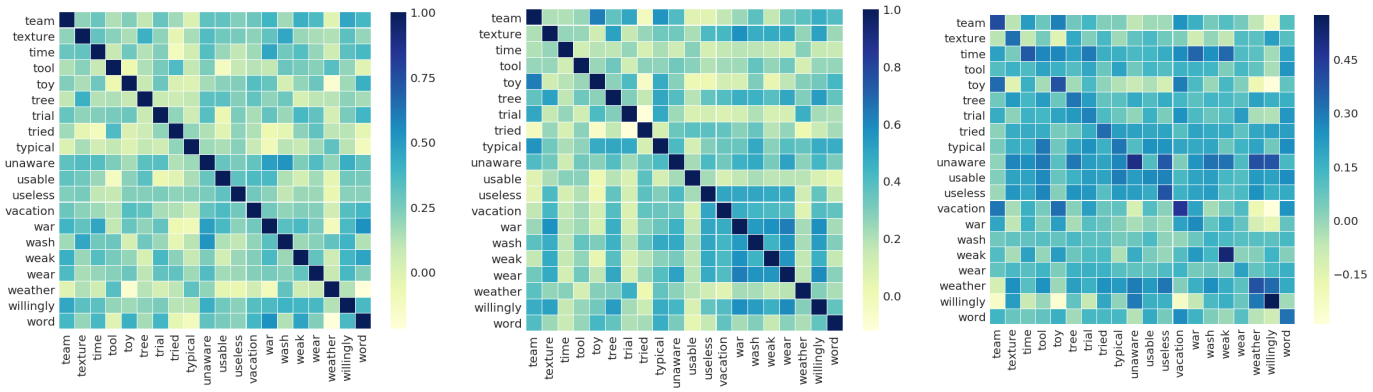

Fig. 3. Similarity structure between ground truth and predicted brain activations. (i) correlation between actual brain responses (left) (ii) correlation between predicted brain responses, to show that the prediction is unique (center) (iii) correlation between actual and predicted brain response with Multi-modal information (right)

TABLE II
PREDICTION ACCURACIES OF THE CORTICAL RESPONSES TO NOVEL CONCEPT WORDS (AVERAGED OVER 5-FOLD CROSS-VALIDATION). PERFORMANCE RESULTS FOR INDIVIDUAL SUBJECTS ARE SHOWN SEPARATELY FOR SCENARIOS WHEN MULTI-MODAL, GLOVE EMBEDDING, XCEPTION VECTOR (LAST FC LAYER), AND MULTI-SUBJECTS INFORMATION WAS UTILIZED.

	Subject-1			Subject-2			Subject-3			Subject-4			Subject-5			Average (all 16 subjects)		
Scenarios	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
(1)	0.83	0.98	0.86	0.78	0.99	0.85	0.86	0.99	0.90	0.81	0.96	0.85	0.82	0.97	0.86	0.832	0.981	0.860
(2)	0.83	0.97	0.86	0.75	0.99	0.83	0.86	0.98	0.90	0.81	0.95	0.85	0.81	0.97	0.86	0.817	0.968	0.853
(3)	0.81	0.98	0.85	0.72	0.97	0.82	0.85	0.98	0.89	0.82	0.96	0.86	0.81	0.96	0.85	0.806	0.954	0.845
(4)	0.87	0.99	0.90	0.77	0.99	0.84	0.86	0.99	0.90	0.82	0.96	0.86	0.82	0.98	0.88	0.831	0.984	0.882
(1) Multi-modal, (2) GloVe (Text), (3) Xception (Image), (4) Multi-subject																		

cepts of the 180 concepts per subject, in a 5-fold cross-validation scheme. The encoder models were trained until the epochs stopped due to early stopping method (around 65 epochs), or when validation loss did not change for few epochs. We observed an average validation loss of 0.0007 for word based models, 0.0006 validation loss for image-based models and 0.0004 validation loss for the multi-modal model (word + image) across all tested subjects. In order to assess the similarity between the actual and predicted brain slice, we compared the intensity of the voxels using slice-wise voxel coordinates. We measured the precision, recall, and F1-scores using voxel intensities and location of voxel coordinates between the predicted and actual slice data. Table II depicts the performance comparison between text alone model versus the model trained on multi-modal stimulus information. Although the precision, recall, F1-scores of two modalities are nearly similar, from Figure I, we observe that the similarities between ground truth and cortical brain responses from multi-modal based encoding model are better with a near-perfect recall. Some of the voxel intensity values predicted by the GloVe embedding model are very negligible in certain brain regions, which cause no activation.

Figure 3 shows the similarity (correlation) matrix between actual and predicted brain response with multi-modal stimuli. The average correlation values across all words in the dataset is 0.42. The correlation matrix is calculated by considering both the actual and predicted voxels in every brain slice. We considered voxels with high activations, that is, those with intensity values greater than a threshold (= mean + standard deviation) and discarded the remaining voxels with low activation values. Here, we found reliable correlations between fMRI responses from trained model and the actual brain responses for all the test words in the case of the model trained with multi-modal information as compared to the one with information from individual stimuli alone. Perturbation experiments (results not shown here) where random input is given as stimulus to the trained model yielded brain responses that had minimal correlation with any of the semantic encodings for the 180 concepts. These results verify the robustness of the learned encoding model.

3) Multi-Subject fMRI Brain Response Generation: In this experiment, we used model with the architecture described in Section II-B which includes the subject index as one of the multimodal inputs. We trained this model on all 16 subjects, where the dataset remained the same as described in the earlier section with 160 training stimuli and 20 test stimuli per subject, and instead of one model per subject, we trained a single model for all the subjects (for *across-subject* prediction). The model converged to the lowest validation loss of 0.0003 for multi-modal stimuli. For this setting, we did not attempt to train the model on individual modalities since earlier experiment inferred better performance with multi-modal stimuli rather than only word or image data alone. The model achieved an average precision of 0.83, recall of 0.98

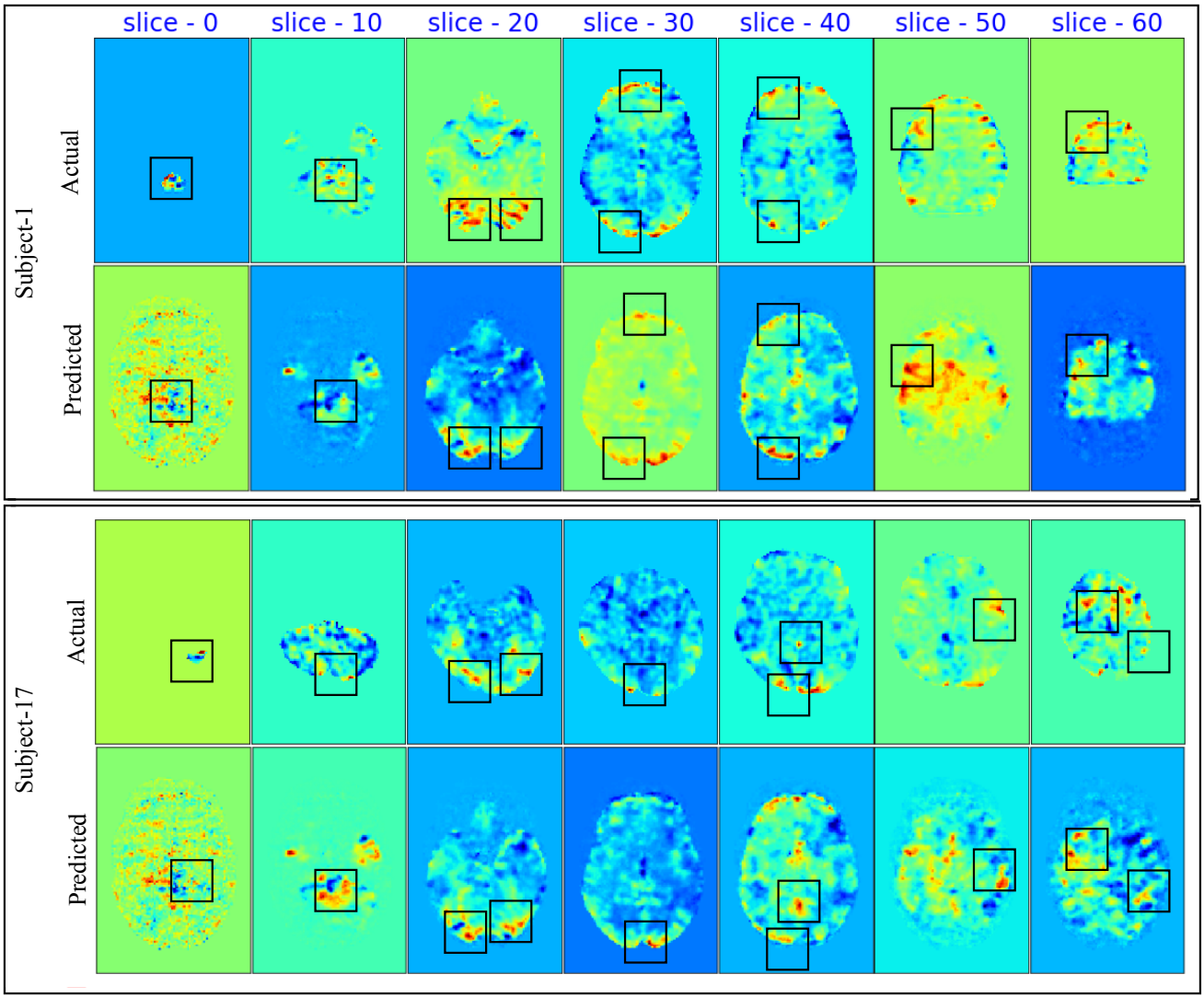

Fig. 4. The sequence of slices show the predictions from multi-subject trained model using multi-modal embeddings. (i) Subject-1 actual brain activation for the word "Team" after converting voxel activation per subject into 70 slices (top row) and brain activation prediction from the multi-subject model (second row), (ii) Subject-17 actual brain response for the word "Team" after converting voxel activation per subject into 70 slices (third row) and brain activation prediction from the multi-subject model (last row). Note: Box indicates the location where voxel activation between actual and predicted slices are compared.

with an average F1-score of 0.88 across all subjects when tested using subsets of 20 concepts of the 180 concepts per subject, in a 5-fold cross-validation scheme.

We also evaluated the performance of our multi-subject model empirically with the slice-wise region of activation in the brain which is shown in Figure 5. From the figure, we observe that regions of activation are approximately similar across subjects for the same word "Tool". We use the Automated Anatomical Labeling (AAL) atlas [36] with a parcellation of 116 brain regions, along with the nilearn [37] toolkit, to label the regions of activation in the slice-wise display of concurrence between actual and predicted responses. The regions which are highly activated slice-wise depict a sequence of "cerebellothalamocortical" ("cerebellum $\rightarrow$ thalamus $\rightarrow$ premotor cortex") and pass through the "peduncle." This pathway seems to emphasize the importance of the pre-motor cortex. Similarly, regions such as the "Fusiform" gyrus in the ventral occipital cortex, the "Temporal Lobe" and superior frontal gyrus depict slice-wise high activations across subjects.

IV. Conclusion

In this work, we proposed an encoder model which can generate a complete 3D model of the brain using multi-modal input, by training the model on subject's brain response for words in the training set. Different from previous work, our method predicts the complete set of voxels, as given in the dataset rather than selected few voxels per subject. The key distinction of our work is the utilization of machine translation inspired encoder-decoder model to generate complete brain image. In the future, we plan to experiment on all paradigms and experiments mentioned in the dataset, with a primary focus on attention-based models.

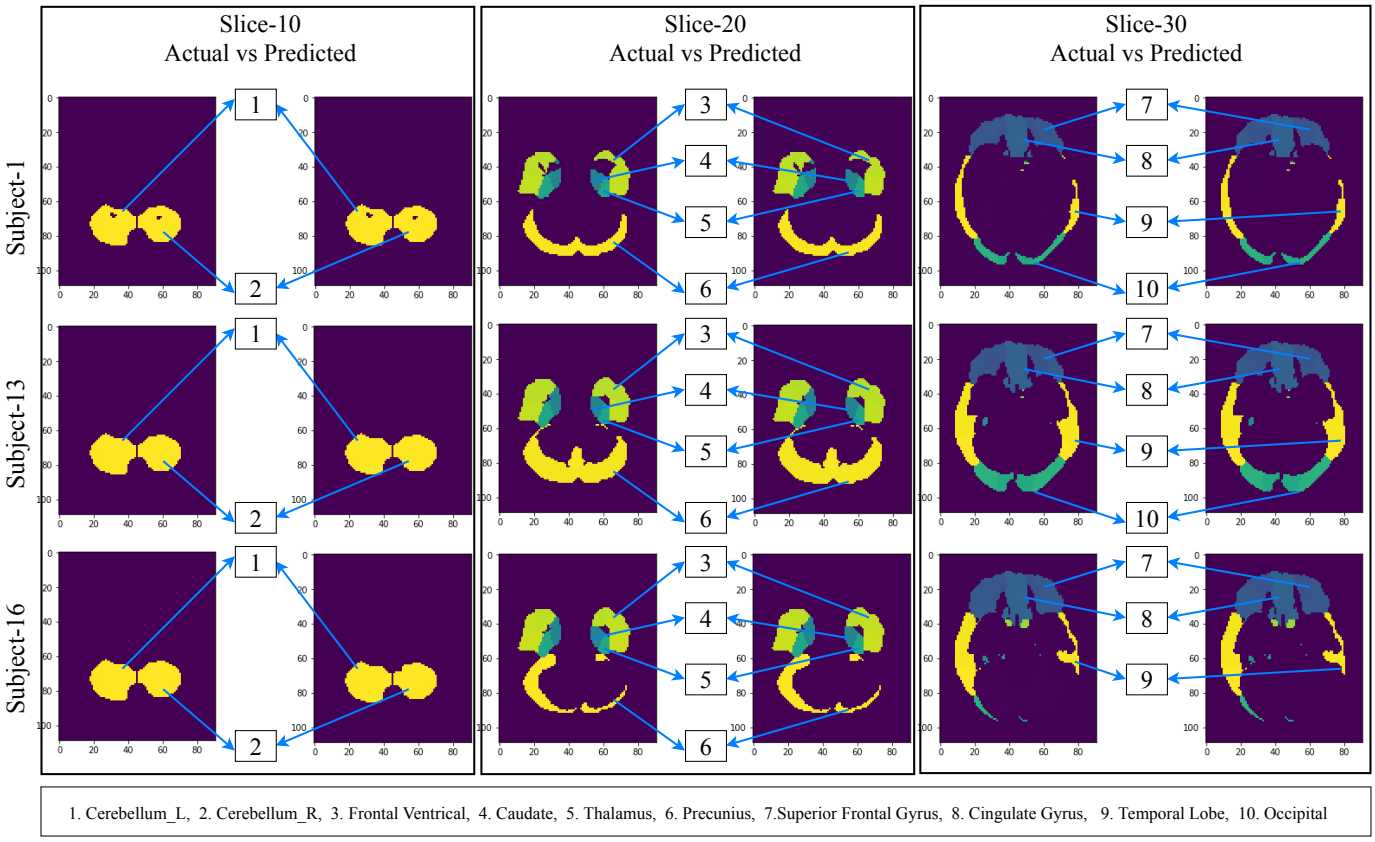

Fig. 5. Region of activation similarity between actual and predicted brain slices for the word "Tool". (i) slice-wise region of activation similarity between actual and predicted of subject1 (top row) (ii) slice-wise region of activation similarity between actual and predicted of subject13 (middle row) (iii) slice-wise region of activation similarity between actual and predicted of subject16 (bottom row).

REFERENCES

[1] S. A. Huettel, A. W. Song, G. McCarthy *et al.*, *Functional magnetic resonance imaging*. Sinauer Associates Sunderland, MA, 2004.

[2] M. D. Fox and M. E. Raichle, "Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging," *Nature Reviews Neuroscience*, 2007.

[3] S. C. Cramer, G. Nelles, R. R. Benson, J. D. Kaplan, R. A. Parker, K. K. Kwong, D. N. Kennedy, S. P. Finklestein, and B. R. Rosen, "A functional mri study of subjects recovered from hemiparetic stroke," *Stroke*, 1997.

[4] P. M. Rossini, C. Calautti, F. Pauri, and J.-C. Baron, "Post-stroke plastic reorganisation in the adult brain," *The Lancet Neurology*, 2003.

[5] T. M. Mitchell, S. V. Shinkareva, A. Carlson, K.-M. Chang, V. L. Malave, R. A. Mason, and M. A. Just, "Predicting human brain activity associated with the meanings of nouns," *science*, 2008.

[6] G. M. Di Liberto, J. A. OSullivan, and E. C. Lalor, "Low-frequency cortical entrainment to speech reflects phoneme-level processing," *Current Biology*, 2015.

[7] N. Mesgarani, C. Cheung, K. Johnson, and E. F. Chang, "Phonetic feature encoding in human superior temporal gyrus," *Science*, 2014.

[8] T. Naselaris, K. N. Kay, S. Nishimoto, and J. L. Gallant, "Encoding and decoding in fmri," *Neuroimage*, 2011.

[9] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

[10] S. Abnar, R. Ahmed, M. Mijnheer, and W. Zuidema, "Experiential, distributional and dependency-based word embeddings have complementary roles in decoding brain activity," in *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*, 2018.

[11] S. R. Oota, N. Manwani, and R. S. Bapi, "fmri semantic category decoding using linguistic encoding of word embeddings," in *International Conference on Neural Information Processing*. Springer, 2018, pp. 3–15.

[12] U. Güçlü and M. A. van Gerven, "Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream," *Journal of Neuroscience*, 2015.

[13] W. A. de Heer, A. G. Huth, T. L. Griffiths, J. L. Gallant, and F. E. Theunissen, "The hierarchical cortical organization of human speech processing." *Journal of Neuroscience*, 2017.

[14] A. G. Huth, W. A. de Heer, T. L. Griffiths, F. E. Theunissen, and J. L. Gallant, "Natural speech reveals the semantic maps that tile human cerebral cortex," *Nature*, 2016.

[15] A. G. Huth, S. Nishimoto, A. T. Vu, and J. L. Gallant, "A continuous semantic space describes the representation of thousands of object and action categories across the human brain," *Neuron*, 2012.

[16] S. Jain and A. Huth, "Incorporating context into language encoding models for fmri," *BioRxiv*, 2018.

[17] A. J. Anderson, D. Kiela, S. Clark, and M. Poesio, "Visually grounded and textual semantic models differentially decode brain activity associated with concrete and abstract nouns," *Transactions of the Association of Computational Linguistics*, 2017.

[18] F. Pereira, B. Lou, B. Pritchett, S. Ritter, S. J. Gershman, N. Kanwisher, M. Botvinick, and E. Fedorenko, "Toward a universal decoder of linguistic meaning from brain activation," *Nature Communications*, 2018.

[19] N. Kriegeskorte, R. Cusack, and P. Bandettini, "How does an fmri voxel sample the neuronal activity pattern: compact-kernel or complex spatiotemporal filter?" *Neuroimage*, 2010.

[20] D. Nie, H. Zhang, E. Adeli, L. Liu, and D. Shen, "3d deep learning for multi-modal imaging-guided survival time prediction of brain tumor patients," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2016.

[21] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A

neural image caption generator," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[22] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[23] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proceedings of the 25th International Conference on Machine Learning*, 2008.

[24] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[25] J. Masci, U. Meier, D. Cireşan, and J. Schmidhuber, "Stacked convolutional auto-encoders for hierarchical feature extraction," in *International Conference on Artificial Neural Networks*. Springer, 2011.

[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012.

[27] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[28] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[29] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, 1997.

[30] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," in *Advances in Neural Information Processing Systems*, 2015.

[31] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "Tensorflow: a system for large-scale machine learning." in *OSDI*, 2016.

[32] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[33] E. Bruni, N.-K. Tran, and M. Baroni, "Multimodal distributional semantics," *Journal of Artificial Intelligence Research*, 2014.

[34] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[35] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[36] R. C. Craddock, G. A. James, P. E. Holtzheimer III, X. P. Hu, and H. S. Mayberg, "A whole brain fmri atlas generated via spatially constrained spectral clustering," *Human Brain Mapping*, 2012.

[37] A. Abraham, F. Pedregosa, M. Eickenberg, P. Gervais, A. Mueller, J. Kossaifi, A. Gramfort, B. Thirion, and G. Varoquaux, "Machine learning for neuroimaging with scikit-learn," *Frontiers in Neuroinformatics*, 2014.