A Deep Autoencoder for Near-Perfect fMRI Encoding

Anonymous Author(s)
Affiliation
Address
email

Abstract

Encoding models of functional magnetic resonance imaging (fMRI) data attempt to learn a forward mapping that relates stimuli to the corresponding brain activation. 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 brain decoding (reconstructing stimulus information from the brain activation) has received wider attention, there have been only a few attempts at constructing encoding solutions in the extant neuroimaging literature. In this paper, we present a deep autoencoder consisting of convolutional neural networks in tandem with long short-term memory (CNN-LSTM) model. The model is trained on fMRI slice sequences and predicts the entire brain volume rather than a small subset of voxels from the information in stimuli (text and image). 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 (Pereira et al., 2018), our model yielded encouraging results.

1 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. Since the work of Mitchell et al. (2008), there has been an increasing interest in using computational models to interpret neural activity using either the decoding or encoding models (stimulus features are used to model brain activity) (Naselaris et al., 2011; Mesgarani et al., 2014; Di Liberto et al., 2015). 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 brain activity and so a 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. Together, these models describe how information of the sensory stimulus or visual function is encoded in the measured brain activity (Naselaris et al., 2011). Word embedding representations were used to build encoding systems (Oota et al., 2018; Abnar et al., 2018). Some methods rely on the parametric regression that assumes that the response is linearly related to stimulus features after fixed parametric nonlinear transformation(s) (Mitchell et al., 2008). 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 an autoencoding model that predicts the complete brain activity associated with multi-modal forms of concrete nouns, which include words and images. The theory underlying

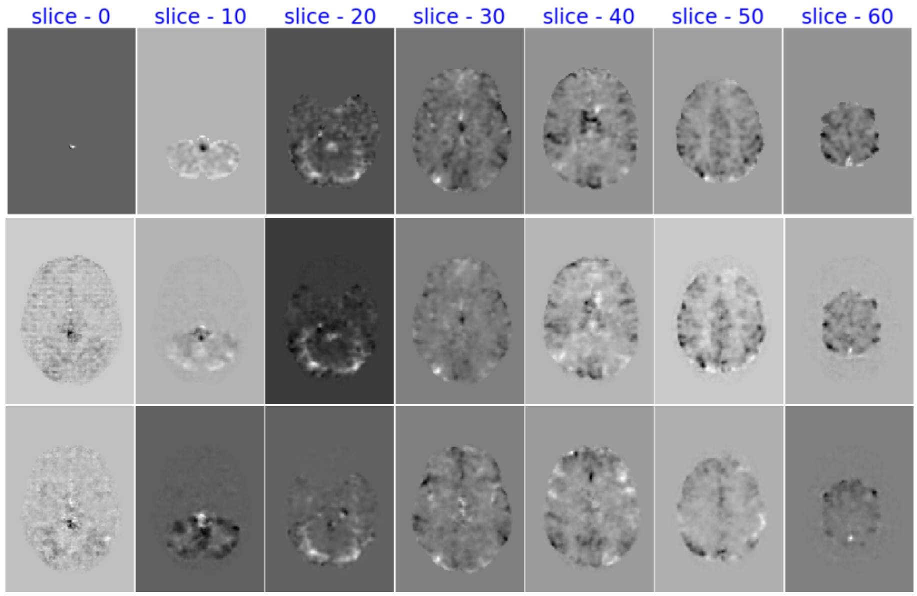

Figure 1: The sequence of slices show (i) actual brain activation for the word "Apartment" after converting voxel activation per subject into 70 slices (top row), (ii) activation prediction by model trained on multi-modal embeddings (middle row), and (iii) activation prediction by model trained on GloVe embedding (bottom row).

this computational model is that when the autoencoder is trained on sufficiently 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 into several 2D slices. 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.

2 fMRI Encoding: Our Approach

Traditional methods either used a set of selective voxels from the dataset (Anderson et al., 2017; Pereira et al., 2018) or handpicked region-based voxels to model brain encoding (Oota et al., 2018) and decoding analysis. In the next sections, we discuss the disadvantages of such methods and our enhancements to overcome these issues.

Voxels and Semantic slices: A voxel is a three-dimensional rectangular cuboid and smaller voxels contain fewer neurons on average and hence have less signal than larger voxels. 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 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 (Mitchell et al., 2008; Jain & Huth, 2018).

Though earlier experiments were conducted with minimal subsets, behavioral and long-term studies of subjects may require generation of the entire 3D volume when the subject is tested with various stimuli (Nie et al., 2016). This creates a necessity for encoding models that are capable of generating a complete 3D volume of the subject's brain based on past fMRI history. We attempted to perform the task of predicting complete 3D volume by utilizing all voxels in the training data (Pereira et al., 2018), converting them to sequences of 2D slices. We argue that the slices provide enough semantic encoding information to train a sequential spatial model, since we observed a gradual change in activation in regions across multiple slides, as seen in Figure 1.

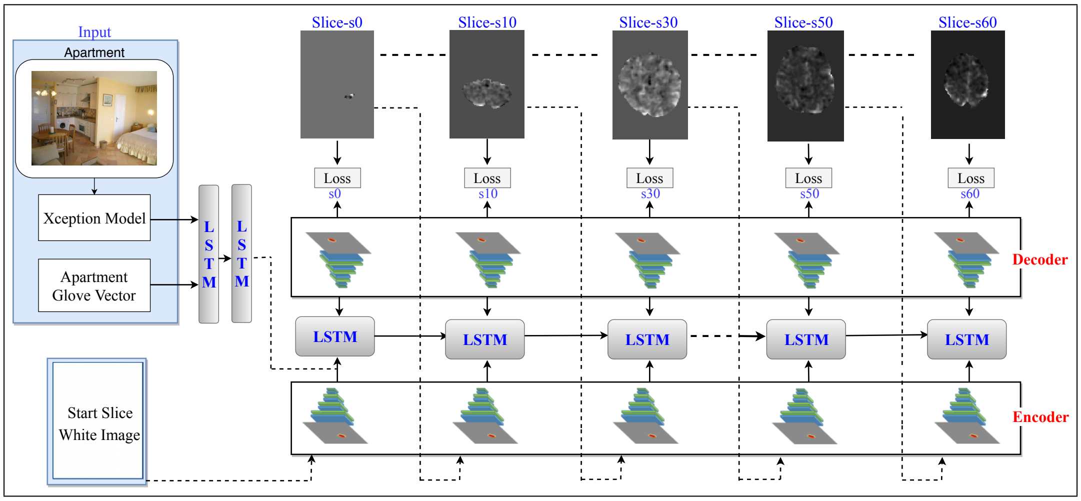

Figure 2: Proposed architecture of the CNN-LSTM autoencoder model used for our experiments.

Architecture: We used a CNN-LSTM based autoencoder model, whose architecture is inspired from Vinyals et al. (2015). Figure 2 describes a basic overview, where CNNs are used for fMRI slice encoding and decoding and LSTMs to learn temporal/semantic features across slides. Both the encoder and decoder have CNN layers with 64, 128 and 256 filters, respectively. Two layers of LSTMs (256, 128) were used as latent layers. The multi-modal features of text and image, pass through two independent layers of LSTM before concatenating to the outputs of CNN encoder. The model uses fMRI slice inputs and "one step ahead" slices as outputs during training. During testing, only the multi-modal input (image, word embedding, and start slice) is given to initiate the cascade of predictions. The modal uses its own output at time step $\boldsymbol{t}$ as input in time step $\boldsymbol{t+1}$.

Multi-modal Semantic models: In Multi-modal semantics (Bruni et al., 2014), a model takes a corpus of images with 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 (Pennington et al., 2014) to obtain a 300-dimensional word embedding which represents the concept word. Image representation comprising 2048 features is obtained by using the output of the fully connected layer of pre-trained Xception (Simonyan & Zisserman, 2014) model. We retrieved 5 images per word from the image stimuli corpus for the 180 concepts (pictures) of the experiment 1 in Pereira et al. (2018)'s dataset. We concatenate image features and the corresponding word vector to give as input to LSTM and a blank slice (start slice as in figure 2) as input to the CNN model.

3 Experiments

Dataset: We used data from paradigm 1 of fMRI experiment 1 (Pereira et al., 2018), where authors conducted experiments with multiple subjects by showing various forms of stimulus (sentence, word+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. This fMRI dataset was collected from a total of 16 participants. 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 initial 9 slices and the last 7 slices due to no activation present in any of the brain regions.

Results and Discussion: Using the approach discussed in Section 2, we trained separate encoding models per experiment for each subject. The encoding performance was evaluated by training and testing models using different subsets of the 180 concepts in a 5-fold cross-validation scheme.

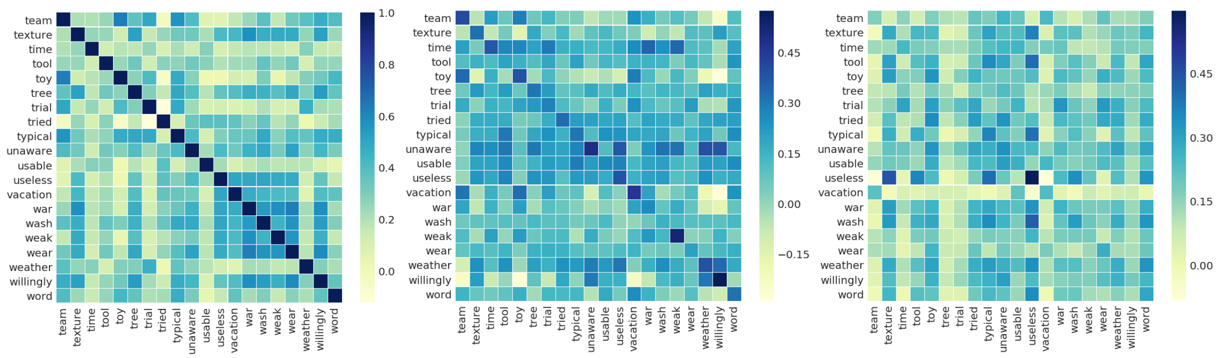

Figure 3: Similarity structure between ground truth and predicted brain activations. (a)correlation between predicted brain responses, to show that is prediction is unique (left) (b) correlation between actual and predicted brain response with Multi-modal (center), and (c) correlation between actual and predicted brain response with Glove embedding model alone (right)

The encoder models were trained until the epochs stopped due to early stopping method, when validation loss did not change for few epochs. We observed an average validation loss of 0.0007 for word based models and 0.0003 validation loss for multi-modal model across all tested subjects. In order to assess the similarity between the actual and predicted brain slice, we compared the slice-wise voxel coordinates and intensity of the voxels. We measured the precision, recall, and F1-scores using voxel intensities and location of voxel coordinates between the predicted and actual slice data. Table 1 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 1, 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 and word embedding stimulus. 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 word embedding information alone. Perturbation experiments (results not shown here) where the 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.

	Multi-modal			**GloVe (Text)**		
Subjects	Prec.	Rec.	F1	Prec.	Rec.	F1
(1)	0.83	0.98	0.86	0.83	0.97	0.86
(2)	0.78	0.99	0.85	0.75	0.99	0.83
(3)	0.86	0.99	0.90	0.86	0.98	0.90
(4)	0.81	0.96	0.85	0.81	0.95	0.85
(5)	0.82	0.97	0.86	0.81	0.97	0.86

Table 1: 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 cases when multi-modal and GloVe embedding information was utilized.

4 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 autoencoder.

References

Samira Abnar, Rasyan Ahmed, Max Mijnheer, and Willem 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)*, pp. 57–66, 2018.

Andrew J Anderson, Douwe Kiela, Stephen Clark, and Massimo Poesio. Visually grounded and textual semantic models differentially decode brain activity associated with concrete and abstract nouns. *Transactions of the Association of Computational Linguistics*, 5(1):17–30, 2017.

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49:1–47, 2014.

Giovanni M Di Liberto, James A O'Sullivan, and Edmund C Lalor. Low-frequency cortical entrainment to speech reflects phoneme-level processing. *Current Biology*, 25(19):2457–2465, 2015.

Shailee Jain and Alexander Huth. Incorporating context into language encoding models for fmri. *bioRxiv*, pp. 327601, 2018.

Nima Mesgarani, Connie Cheung, Keith Johnson, and Edward F Chang. Phonetic feature encoding in human superior temporal gyrus. *Science*, 343(6174):1006–1010, 2014.

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

Thomas Naselaris, Kendrick N Kay, Shinji Nishimoto, and Jack L Gallant. Encoding and decoding in fmri. *Neuroimage*, 56(2):400–410, 2011.

Dong Nie, Han Zhang, Ehsan Adeli, Luyan Liu, and Dinggang 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*, pp. 212–220. Springer, 2016.

Subba Reddy Oota, Naresh Manwani, et al. fmri semantic category decoding using linguistic encoding of word embeddings. *arXiv preprint arXiv:1806.05177*, 2018.

Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

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

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3156–3164, 2015.