Algonauts Project 2021

Challenge Model Building Process (Mini Track)

Aditya Lingam*, Aryan Kaul*

Under the Guidance of Prof. Bapi Raju, Aditya Jain Pansari, Madhuakr Dwivedi, and Chaitanya

*Both contributed equally

Introduction

The Algonauts Project aims to embark on understanding the nature of human intelligence and engineering more advanced forms of artificial intelligence together due to their increasingly intertwined nature. This year's challenge provides 1102 videos that are 3 seconds long with a frame rate of 30 fps and the recorded fMRI brain data of the first 1000 videos when displayed to 10 subjects. The objective of the challenge is to predict the recorded brain data, for each subject, for the remaining 102 videos. For the "Mini Track", a subset of the whole brain data is provided. Brain responses provided are from a set of specific regions of interest (ROIs) known to play a key role in visual perception. Since 3 trials are taken for each video, data in the format of "1000 X 3 X <number of voxels in ROI>" is provided for every subject and ROI combination. The predicted responses are scored on the basis of their correlation with the recorded responses. More details can be found on the challenge website here.

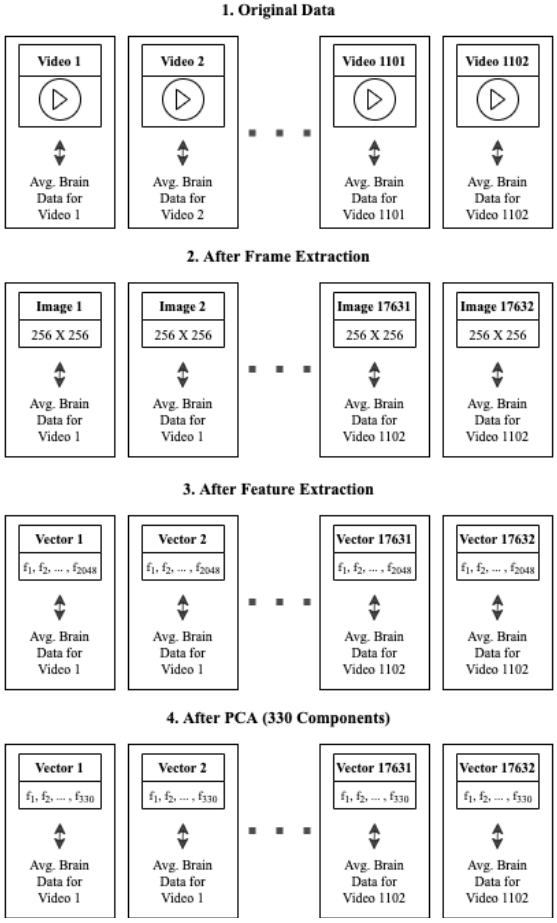
To predict the brain responses a feature vector, a representation of the video, is extracted for each video and sent to a model that predicts the brain data from the feature vector.

Feature Extraction

As the inputs are videos we initially tried using SlowFast, a state-of-the-art model in video analysis. However, after facing issues using features extracted from the whole video for feature extraction, we decided to extract features from an individual frames of the videos, hoping that the model would be able to predict the brain data using an individual frame of the video, as although individual frame doesn't contain the information that changes between frames, we were hoping a large part of the brain responses would be the processing of elements that remain constant through the

frames like the background and physical attributes of the subject.

To aid with data augmentation 16 uniformly spaced frames were taken from each video for feature extraction, each paired with the target brain data for the video. ResNet-152 is used for feature extraction with features extracted from the output of the last Average Pooling layer. The vector is then condensed using principal component analysis (PCA) from 2048 elements to 330 elements.



Brain Activity Prediction

The brain response was predicted using a different model for each subject as the number of voxels for a certain ROI is different for each subject. The model's input layer is shared by each ROI. Batch normalization and dropout layers are used with the dropout probability being 0.72. The ReLU activation function is used with the outputs of the layer being sent to individual output layers for each ROI. The loss is calculated by taking the average of the mean squared error losses of each output layer. This is done because our final score is based on the average of the scores for individual ROIs. The data is split with the first 14400 data points being used for the test set and the remaining 1600 being used for the validation set. The model is then trained for 80 epochs, with a large batch size of 512 and a learning rate of 0.01. The final predictions are generated using the parameters at the epoch with the lowest validation loss in the training process.

After training results can be generated for each subject and ROI combination with predicted activity for the 16 frames from the same video being averaged for the final values.

Results

The model was run with the seed value being set to 42. After the submission of the results the following scores were received:

Score Name	**Value**
LOC	0.5255 (12)
EBA	0.5442 (10)
FFA	0.6368 (11)

STS	0.4419 (11)
PPA	0.4518 (11)
V1	0.3421 (16)
V2	0.3424 (16)
V3	0.3618 (16)
V4	0.3957 (16)
Average	0.4491 (13)

The number in the brackets represents our position on the leaderboard.

Conclusion

From the results, we can see that we greatly underestimated the importance of movement and motion for visual processing. This can be seen in the low scores for the early and mid-level visual cortex ROIs (V1, V2, V3, V4). However we can see that information pertaining to the scene and other visual details, that do not change through the video, have a much greater effect on the responses of the ROIs in the higher-level cortex (EBA, FFA, STS, PPA) than the ones in the early and mid-level visual cortex. To increase the performance of the model we can try extracting a feature vector using a model architecture like the aforementioned SlowFast which uses frames from multiple points of the video. We can also try further experimentation with the structure of the prediction model described in the “Brain Activity Prediction” section.

Citations

Cichy, R. M., et al. “The Algonauts Project 2021 Challenge: How the Human Brain Makes Sense of a World in Motion.” ArXiv:2104.13714 [Cs, q-Bio], Apr. 2021. arXiv.org, http://arxiv.org/abs/2104.13714.