# Submission Report

Cong Hua, Bing Han, Ning Wang, Liang Zhang

School of Computer Science and Technology, Xidian University, Xi'an, China

{hc.zeppeli.fans, hanice511, wangning2049}@gmail.com, liangzhang@xidian.edu.cn

## 1.Introduction

*The 2021 Challenge: How the Human Brain Makes Sense of a World in Motion* organized within the Algonauts project, has as its aim increasing our understanding of what visual features are processed and encoded by the brain. The primary target of the 2021 challenge is to use computational models to predict brain responses recorded while participants viewed short video clips of everyday events. There are two challenge tracks: the Mini Track and the Full Track. The Mini Track focuses on pre-specified regions of the visual brain known to play a key role in visual processing. The Full Track considers responses across the whole brain.

The competition provides brain data measured with functional magnetic resonance imaging (fMRI), a widely used brain imaging technique with high spatial resolution that measures blood flow changes associated with neural response. fMRI is measured at 3T in 2.5×2.5×2.5mm cubes (called voxels) while 10 human participants viewed a set of 1102 3-s long video clips of everyday events(e.g., panda eating, fish swimming, a person paddling). 1000 of these videos are repeated 3times and form the training set. 102 videos were repeated 10 times and form the testing set.

At the deadline of the competition, we got 0.5917 and 0.3217 challenge score in mini track and full track respectively[1]. In the following, we will review the submission to the Algonauts 2021 Challenge in both the mini and full tracks and describe the building process for our best model.

---

[1] Our team is "Hc33" on the leaderboard.

# 2.Method

## 2.1 The overall Framework

In this study, we improved one common approach called a voxel-wise encoding model where the response of each voxel is predicted independently using the multiple features provided by a computational model. Fig 1 shows the proposed approach. Our method relies on two pretrained deep networks for image classification and behavior recognition. At the core of our method, we aggregate over the features produced at various stages along the networks. Then, we reduce the dimensionality of the high-dimensional features and map them to the response of each visual cortex area of the brain. The details of the proposed method will be elaborated in following subsections.

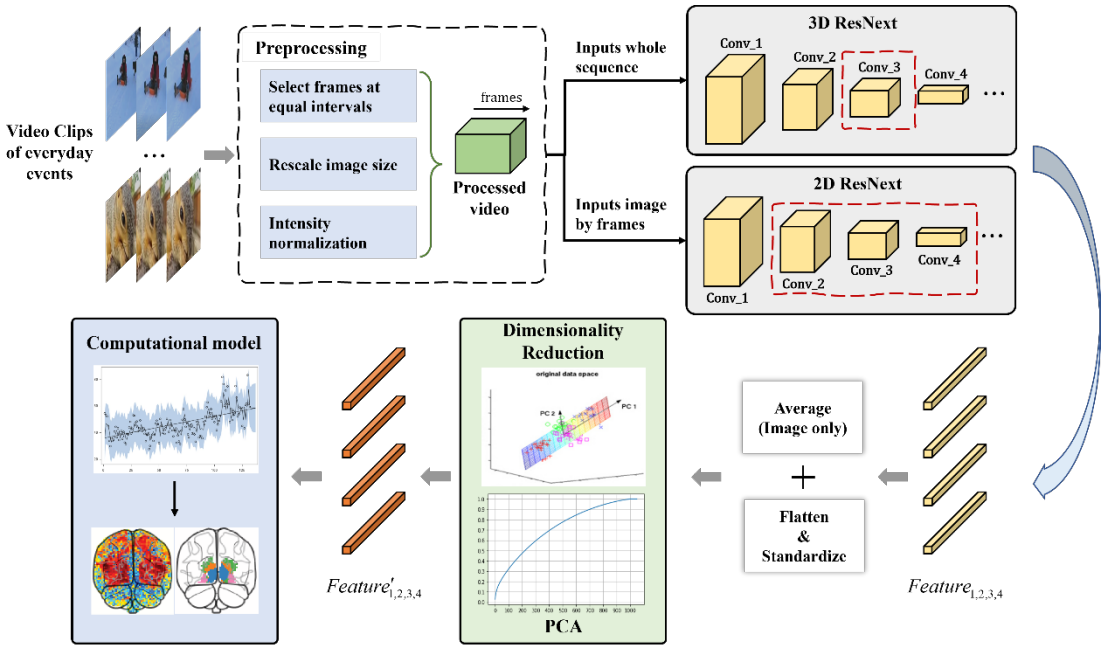


Fig 1. The overall framework

## 2.2 Feature Extraction and Compression

As mentioned above, getting features from video clips is first step in our framework. This changes the format of the data (from pixels to model features) and typically reduces the dimensionality of the data. The features of a given model are interpreted as a potential hypothesis about the features that a given brain area might be using to represent the stimulus.

**Feature Extraction.** Considering that the competition provides short video clips of such everyday events as inputs of the first step, we select 2D and 3D ResNext101 model as our feature extractors. 2D ResNext is pretrained on ImageNet and 3D ResNext is pretrained on Kinetics respectively.

As shown by the red dashed boxes in Fig.1, we get the activations at specific layers in the two models. According to our experimental results, the features obtained in the third layer (Conv_3) of the 3D RexNext model are most closely related to the subject`s brain response. So, we save the activation value of each video in this layer. In the 2D model, the second-, third- and fourth-layer features (Conv_2, Conv_3 and Conv_4).

### 2.3 Computational Model

The extracted features will be flattened and standardized before encoding. In particular, the activation values from the 2D model will be averaged by the number of frames. Then dimensionality reduction are performed due to a large number of high-dimensional features, which contain a lot of the redundant information. PCA (Principal Component Analysis) is a common data analysis method, often used for dimensionality reduction of high-dimensional data and can be used to extract the main feature components of the data. PCA is conducted here to reduce the useless information.

the processed features are mapped to the response value in visual cortex by encoding model. Here we select an implementation of computational model using Lasso regressor.

## 3.Result

### 3.1 Implementation Details

Before being fed into the feature extractor, video clips is preprocessed. The first step is to get a fixed number of frames from the video. For each video clip, 16 frames at equal intervals are selected as model inputs. After that, we scale each image to a

uniform size and then normalize it with corresponding RGB values. Next, these processed frames will be sent into the model and propagated forward, and we will get the activations of the model at a specific layer. It is worth noting that in the 3D model, these frames are fed as a whole sequence, but in the 2D model, the input is a single frame.

In the dimensionality reduction step, we set the number of principal components to be 25 or 50, which is a lower amount. Then we conduct separately for each-layer features instead of doing PCA for all features at the same time. This allows our computational model to obtain sufficient features and reduces the workload of data processing. These processed features are stacked together and used as input of computational model. When training the model, cross-validation is used to get the optimal hyperparameters.

### 3.2 Challenge Result

According to the results submitted to the Codalab platform, we have achieved a average score of 0.5917 in mini track with test brain data, the follow table shows the score for each region. In full track, our proposed method achieved a score of 0.3225.

Table 1. Our challenge score in mini track.

| **Brain Regions** | **V1** | **V2** | **V3** | **V4** | **LOC** | **EBA** | **FFA** | **STS** | **PPA** |
|---|---|---|---|---|---|---|---|---|---|
| **Challenge Score** | 0.5594 | 0.5525 | 0.5469 | 0.5901 | 0.6491 | 0.6649 | 0.6999 | 0.4971 | 0.5653 |

## Reference

[1] Naselaris, Thomas, et al. "Encoding and decoding in fMRI." Neuroimage 56.2 (2011): 400-410.

[2] Wu, Michael C-K., Stephen V. David, and Jack L. Gallant. "Complete functional characterization of sensory neurons by system identification." Annu. Rev. Neurosci. 29 (2006): 477-505.

[3] Cichy, Radoslaw Martin, et al. "The Algonauts Project 2021 Challenge: How the Human Brain Makes Sense of a World in Motion." arXiv preprint arXiv:2104.13714 (2021).