UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

DIPARTIMENTO DI INFORMATICA SISTEMISTICA E COMUNICAZIONE

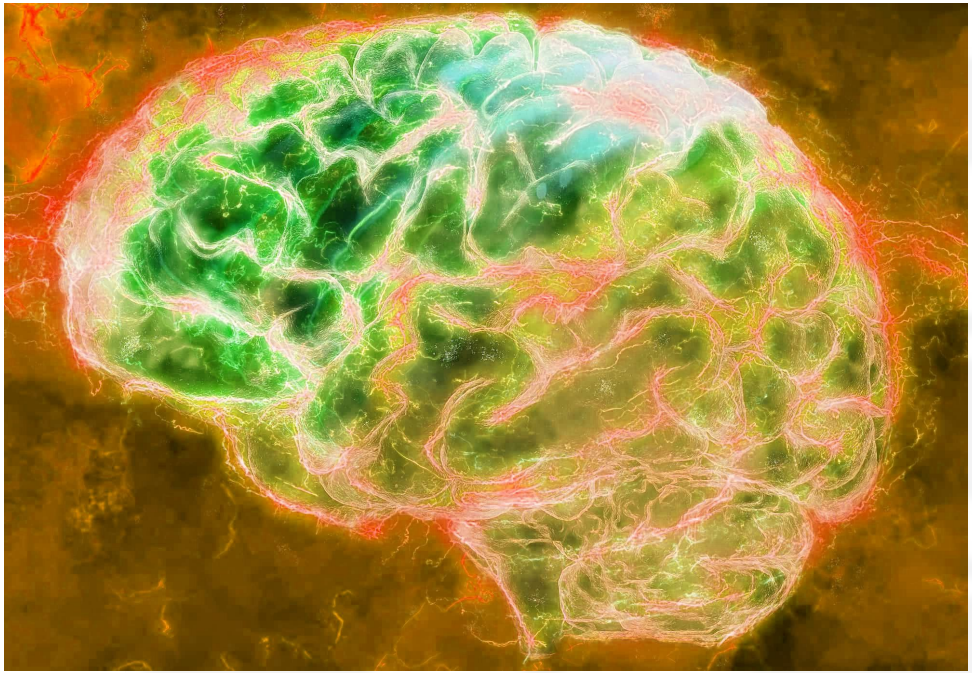
Anno Accademico 2022 - 2023

Exploration and comparison of deep learning architectures to predict brain response to realistic pictures

Relazione prova finale: Riccardo Chimisso, 866009

Relatore: Dimitri Ognibene

Co-relatore: Giuseppe Vizzari

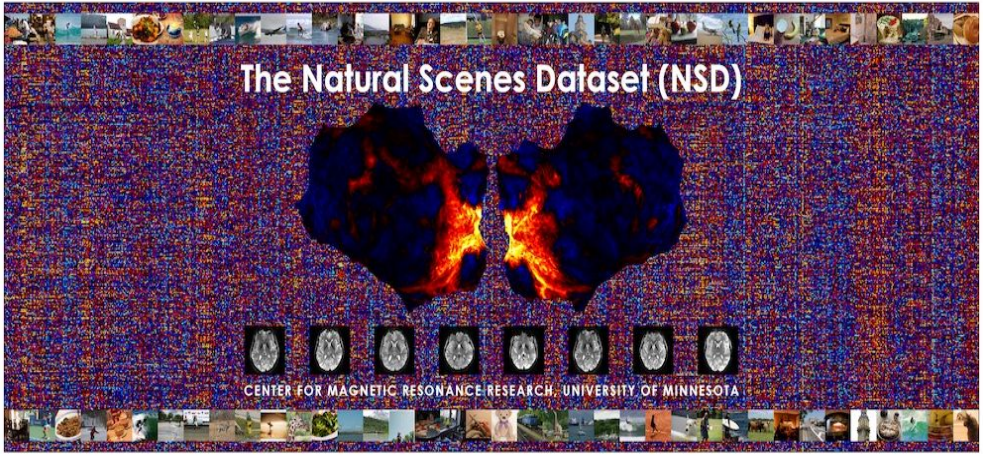


Algonauts Challenge 2023

How the Human Brain Makes Sense of Natural Scenes

AI and neuroscience challenge to predict fMRI brain activity in response to natural scenes.

Participants have to predict fMRI brain activity from natural scene images. Model performance is evaluated by mean noise-normalized squared correlation score.



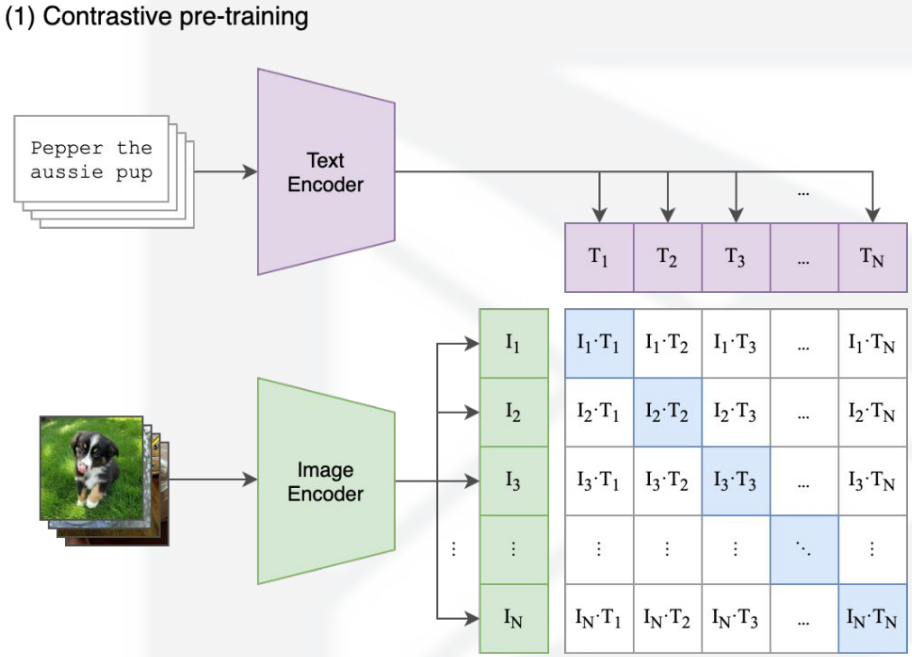
The Natural Scenes Dataset (NSD)

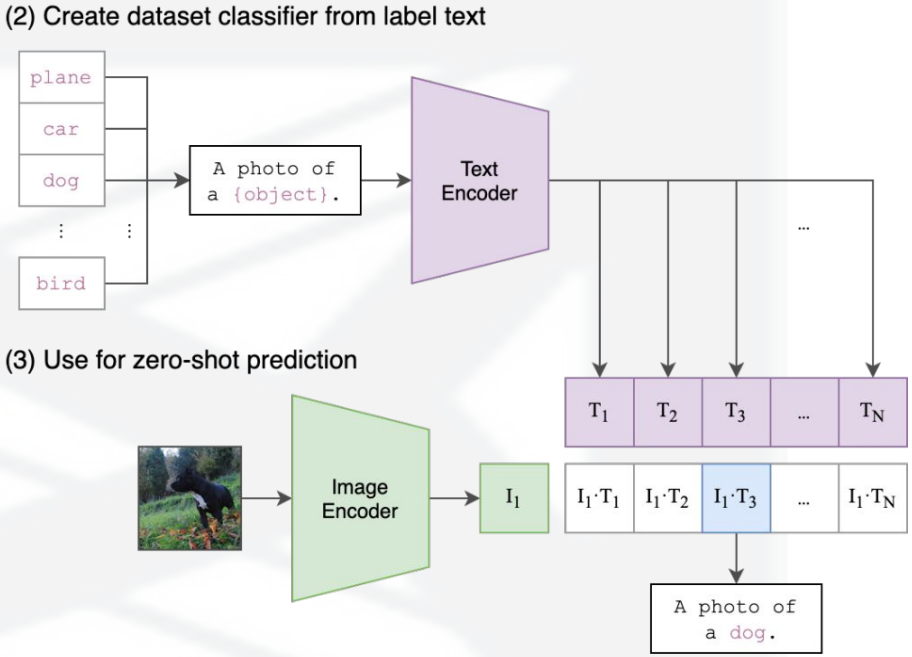
8 healthy adult subjects viewed thousands of color natural scenes while fMRI scans were taken, over the course of 30-40 scan sessions.

Each subject viewed 10,000 images, with 1000 shared across all subjects. In total 73,000 images (8 subjects times 9000 unique images plus 1000 shared ones). Each image was presented 3 times, meaning 30,000 image trials per subject.

The fMRI data is divided into left/right hemispheres, each with 19k/20k vertices. Subjects 6 and 8 have less due to missing data. Voxel activity is z-scored for each session, and then the fMRI responses are averaged across repeats of the same stimulus images.

CLIP (Contrastive Language-Image Pretraining)

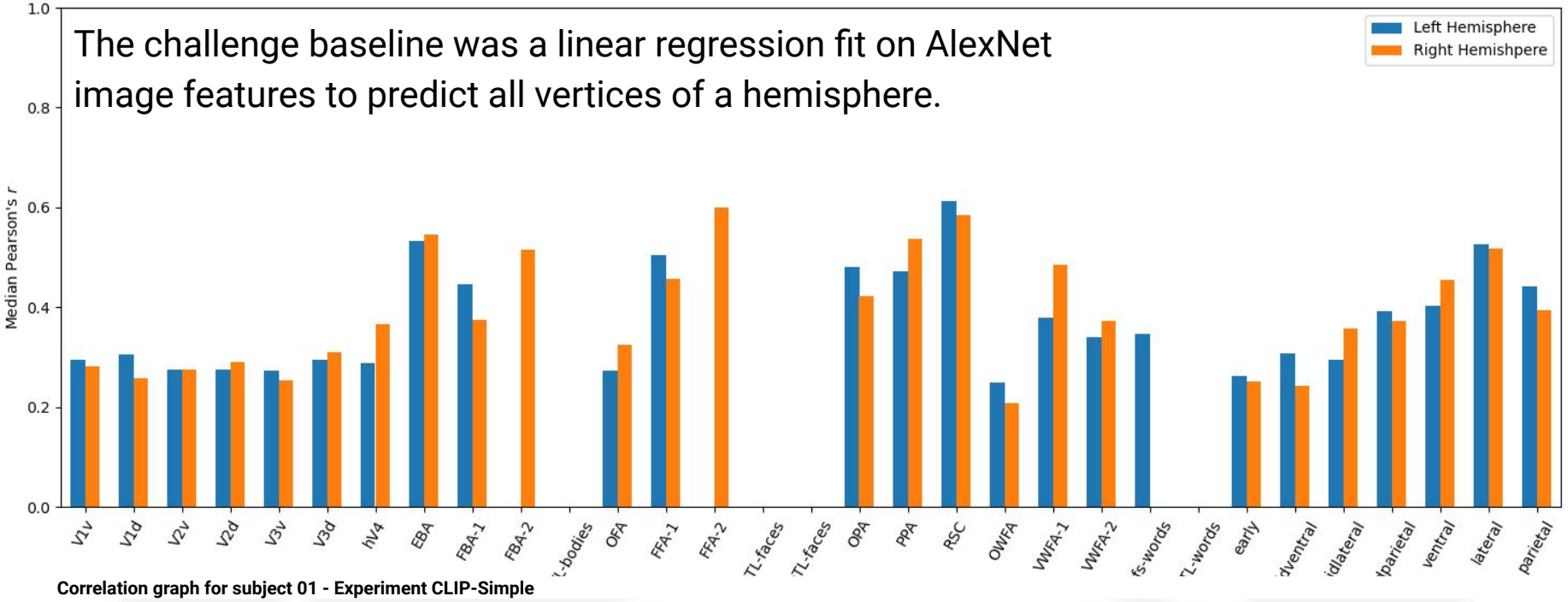




CLIP is a powerful model that can understand images and text. It can perform a variety of tasks without task-specific fine-tuning.

The main model we used for transfer learning as its embeddings are rich in semantics yet small enough to be processed quickly.

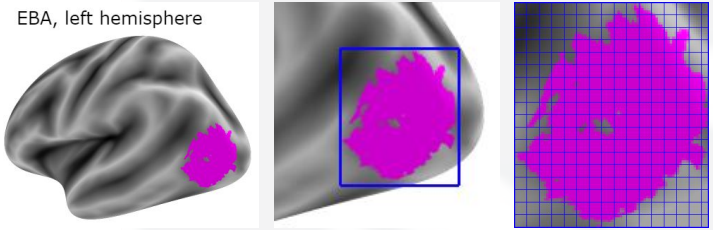
First attempts



Correlation graph for subject 01 - Experiment CLIP-Simple

Our initial models were simple and trained either single-subject or cross-subject.
Generalization was hard; single-subject yielded better results.
The best model was a single linear layer with CLIP embeddings as input.

Positional Encodings

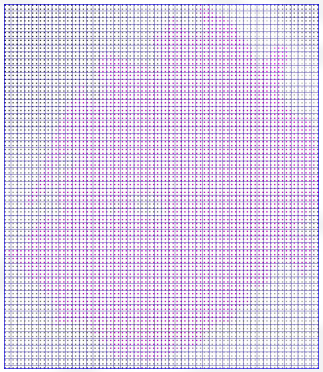


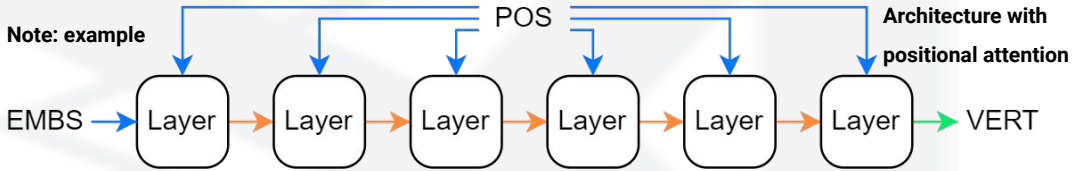
We thought the position of a vertex in its ROI could play a role in predicting its activation. We carefully studied the best way to encode positional information to benefit our models.

2D exemplification of the encoding process.

ROI -> BBox -> Grid -> Expanded Grid

The origin is the BBox center. Expanding helps differentiating vertices.
Position values are then encoded via sinusoids.

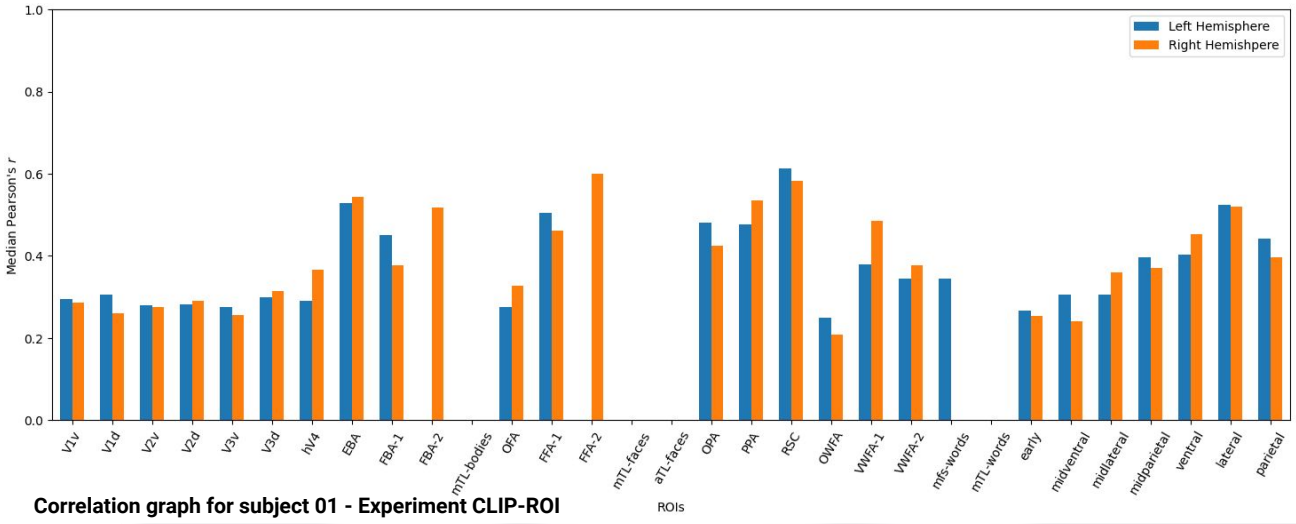




We tried combining these positional encodings with image embeddings, but CLIP-based models overfit and SAM-based models were generally unstable.

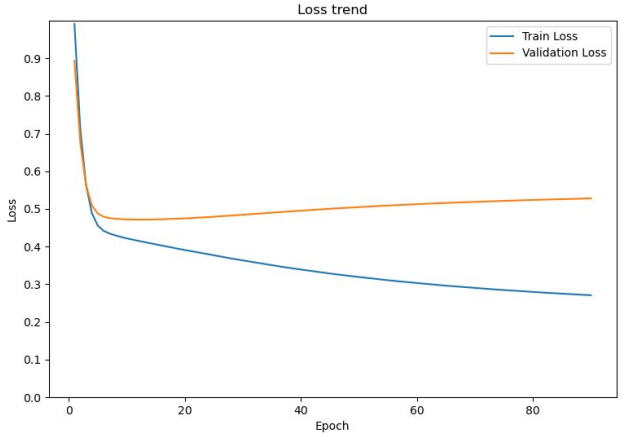
Applying dropout and resblocks improved performance, but never managed to outperform our best, not even when these techniques were applied to simpler models without positional information.

Last attempts



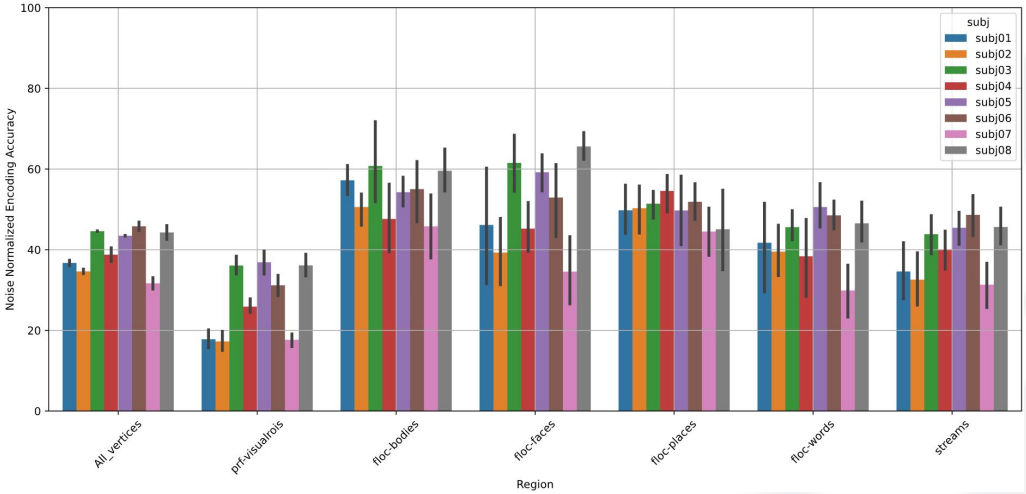
Correlation graph for subject 01 - Experiment CLIP-ROI

Only at last did we manage to get a slight improvement on the correlation scores compared to the best previous model.



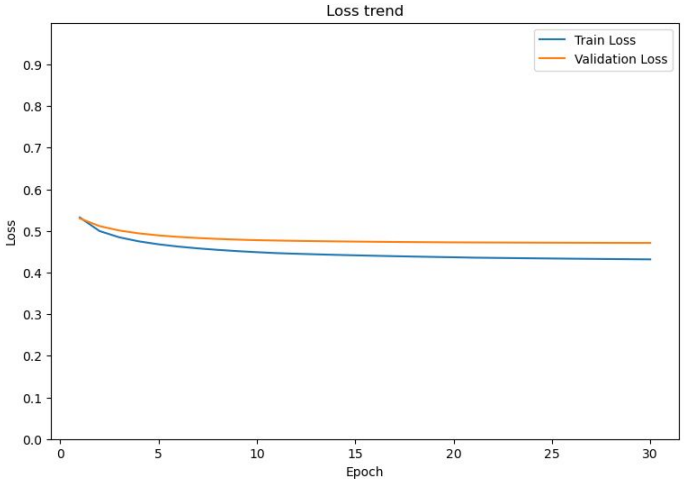
OFA, right hemisphere, subject 01 - Experiment CLIP-Conv1d

More complex models would overfit, so we took a step back and returned to a single linear layer.
This time, we trained a model ensemble: one for each ROI in each hemisphere of each subject.
Better correlation scores of about 0.01 across the board.
A long-awaited improvement, but not quite satisfying due to the simplicity of the architecture.



Best model and submission

We submitted our best models' predictions as time was running out.
We could compare our results with others' and our evaluation metrics with the challenge ones.



OFA, left hemisphere, subject 01 - Experiment CLIP-ROI

Our internal metrics resulted quite consistent with the challenge ones.
The subjects with missing data, 06 and 08, did not perform worse compared to other subjects.
We are currently placed 51st out of 100 with a score of 39.9926.

Possible improvements

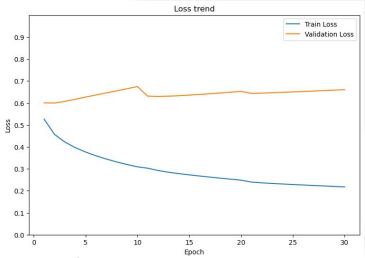
Increase bounding box expansion factor in positional encodings to make up for smaller ROIs.

Increasing the available training data to help preventing overfitting.
While it's true that the NSD dataset is a large one, it's small compared to other CV datasets.

Exploring different optimizers other than AdamW could lead to a better convergence.

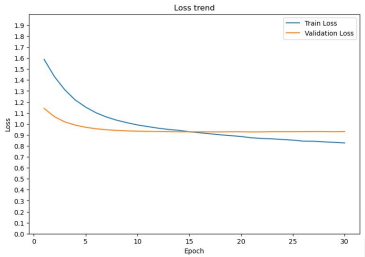
We used CLIP image embeddings, however correlation on some ROIs could improve by jointly utilizing CLIP text embeddings on COCO image captions.

Possible improvements



Left hemisphere, subject 01 - Experiment CLIP+Alex

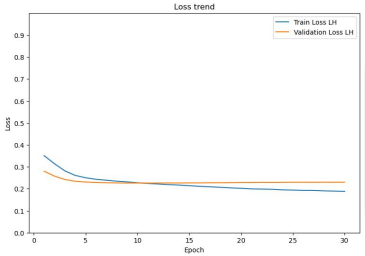
While CLIP embeddings brought an overall increase in correlation, AlexNet had the upper hand with ROIs in the “prf-visualrois” class. Unfortunately, combining these two embeddings only led to more overfitting.



OFA, left hemisphere, subject 01 - Experiment CLIP-Both w/MSE

ROIs in both hemispheres are correlated, so training a model to predict them in both at the same time could improve generalization. Train loss decreased considerably, but other than that, it did not improve much.

Along the same lines, another attempt could be to predict all ROIs of the same class in both hemispheres.



OFA, left hemisphere, subject 01 - Experiment CLIP-Both w/Huber

MSE loss is sensitive to outliers, so other loss functions could help. Huber and Pearson Correlation losses were explored, but correlation scores did not improve much.

Conclusion

We explored many architectures and approaches, from simple to complex. Complex models, alas, tended to overfit.

Our best architecture was an ensemble of models, one for each ROI in each hemisphere for each subject, scoring an average correlation of 39.99 and ranking 51st out of 100.

We will keep experimenting to improve our results. We won't stop trying new approaches and architectures to predict brain activity until the end of the competition.

Even after, we are eager to check out other participants methods and find out the differences between ours.