The role of gaze position in training visual brain encoders on free-viewing data

Marie St-Laurent1,2, Katja Seeliger2 & Martin Hebart2
1Centre de recherche de l’Institut universitaire de gériatrie de Montréal, 2Max Planck Institute for Human Cognitive and Brain Sciences

INTRODUCTION

- Brain-encoding models predict the brain’s response to (visual) input
- “End-to-end” brain encoders such as Neural Information Flow (NIF)3 are trained directly on human visual cortical activity
- To facilitate brain modelling, input images are commonly aligned with the participant’s fovea, which requires participants to fixate on visual stimuli
- While unconstrained viewing is more natural than fixation, the need for foveal alignment makes it challenging to train models on “free-viewing” data

Here, we tested whether NIF could be trained without foveal alignment using a large dataset of movie-viewing fMRI data

METHODS

- With video stimuli as input, NIF couples brain areas with CNN layers that encode spatiotemporal visual features, and learns to read out the activity tensors for mass univariate prediction of activity in individual voxels.

Neural Information Flow

- NIF was trained on fMRI data from a participant scanned while watching three seasons of the sitcom Friends for the Courtois Project on Neuronal Modelling (CNeuroMod)4
- It predicted responses from voxels in cortical areas V1, V2, V3, hV4, V3A and V3B delineated with population receptive fields4
- NIF predicted voxel activity at randomized target TRs based on 3*45 consecutive movie frames (~3 TRs) shown 3-5 TRs before.
- Frames were centered around a likely gaze position estimated with DeepGaze MR5, a static saliency model with a pre-trained VGG-19 backbone that predicts gaze distributions for movie frames

RESULTS

- We compared NIF performance across four conditions
  - Cropped images centered around:
    - Center: central gaze position estimated with eye-tracking
    - Peak DG: peak gaze likelihood from DeepGaze saliency map
    - Peak DG*: peak gaze likelihood IF one local maximum, else central gaze
    - Mean Peak DG: peak gaze likelihood averaged over input frames (per TR)

- DeepGaze predictions were validated against eye-tracking data acquired for 10 fMRI runs (5 Friends episodes)

CONCLUSION

- Deep net-based end-to-end model successfully trained on free viewing data despite lack of fixation or eye-tracking
- In higher level regions, performance benefited from input stability (mean centering) despite lack of foveation
- Recentering input around predicted gaze averaged per TR improved performance slightly in V1; recentering per frame did not help
- As a next step, contrast current performance with models trained on data acquired with fixation in ROIs sensitive vs. invariant to eye movement

References


ACKNOWLEDGEMENT: Support from the Courtois Fondation

www.cneuromod.ca  courtois-neuromod  @cneuromod