

Linking human and artificial neural representations underlying face recognition: Insights from MEG and CNNs

Hamza Abdelhedi (hamza.abdelhedi@umontreal.ca)

Cognitive & Computational Neuroscience Lab, Montreal, Canada
University of Montreal, Montreal, Canada

Karim Jerbi (karim.jerbi@umontreal.ca)

Cognitive & Computational Neuroscience Lab, Montreal, Canada
Psychology Department, University of Montreal, Montreal, Canada
Mila (Quebec AI Research Institute)
UNIQUE (Quebec Neuro-AI Research Center)



Abstract

Mounting evidence suggests that biological and artificial neural networks trained on similar tasks can exhibit remarkable functional similarities. In particular, Convolutional Neural Networks (CNNs) trained on object recognition have been shown to learn representations that model the processing hierarchy in the human visual system. But what about the specific case of face recognition? Would CNNs trained on face recognition learn representations that capture the neural dynamics of the neural circuits that mediate face recognition? Here, we investigated the putative similarities between the representations learned by three CNN architectures trained on face recognition (FaceNet, ResNet50 and CORnet-S) and brain patterns assessed with magnetoencephalography (MEG) recorded during a face recognition tasks. We found that the neuromagnetic brain signals (especially in visual areas) was correlated with activations in multiple CNN layers. However, these correlations were not as strong as those observed in the more general task of object recognition. This may suggest that CNNs trained on face recognition capture only a small portion of the complexity of the real brain patterns associated with face recognition. Our results contribute to an emerging stream of research that seeks to probe face recognition mechanisms by joint exploration of the associated neural representations in both brains and machines.

Keywords: Convolutional Neural Networks (CNN); Magnetoencephalography (MEG); Face Recognition; Similarity analysis

Intro

Artificial neural networks (ANN), originally inspired by fundamental properties of brain structure and function, have gone on to surpass human performance on many tasks and most of the latest developments in AI occur with little if any links to the field of neuroscience. This said, there is an increasing recognition of the added value of combining AI and neuroscience research to mutually reinforce on another. One promising approach within this subfield of Neuro-AI is to compare artificial and biological systems' internal representations in an array of cognitive tasks such as visual categorisation (Kubilius et al., 2019; Richards et al., 2019; Storrs, Kietzmann, Walther, Mehrer, & Kriegeskorte, 2020; Yamins et al., 2014). Here, we set out to investigate the similarity between neural representations in convolutional neural networks (CNNs) and electrophysiological human brain data during face recognition. To date, only a handful of studies have addressed this question. For instance, Chang, Egger, Vetter, and Tsao (2021) used several models to explain the facial stimuli representation in IT cortex of non-human primates. Jiahui et al. (2022) used fMRI recorded during processing of naturalistic face stimuli to compare responses in face processing regions in the brain to three different artificial networks. The present study is, to our knowledge, the first to probe

this question using magnetoencephalography (MEG). Previous work suggests that the high spatio-temporal resolution afforded by MEG could be crucial in investigating the links between CNN models and network dynamics of the human brain (e.g. (Kietzmann et al., 2019)).

Methods

MEG Data

We used publicly available MEG data acquired using a 306-channel system (Elekta Neuromag Vectorview) (Wakeman & Henson, 2015). Neuromagnetic signals were recorded from 16 subjects while they performed a face recognition task (150 Famous versus 150 Unfamiliar face stimuli and each picture was presented twice, i.e. 600 trials). However, three participants were excluded because of excessive artefacts. We replicated the preprocessing steps reported in the original study using the Biomag script (Jas et al., 2018). For each subject, we used 300 trials (i.e. single picture presentations) epoched into 800ms segments.

Network training and extraction of layer activations

We chose three networks: (i) The backbone of FaceNet (Schroff, Kalenichenko, & Philbin, 2015) which was specifically built for face recognition, (ii) ResNet50 (He, Zhang, Ren, & Sun, 2015) which was built for Object recognition, and (iii) CORnet-S (Kubilius et al., 2018), which was designed to model the visual cortex. To make sure the selected networks perform well on a face recognition task, we first trained them on the same classification tasks. An exhaustive search on hyperparameters led us to set a batch size of 32, a learning rate of 0.01 and a step size of 10. Finally, we used Cross-Entropy Loss for the all three networks.

We solved a closed-set face recognition task by training the networks on the VGGFace dataset (Cao, Shen, Xie, Parkhi, & Zisserman, 2017). We then fine-tuned the architectures on a distribution of the celebA dataset (Liu, Luo, Wang, & Tang, 2015) which is similar to the stimuli used in the MEG experiment. To extract the layer activations needed for the subsequent similarity analyses, we passed the face stimuli used in the MEG data to the networks, and saved the response of all layers (flattened in stored in 1D activation arrays). Depending on whether a given convolutional layer was followed by batch normalization and/or an activation function, we considered the layer activations to be the output of the layer response, of the batch normalization or of the convolutional layer. We also generated a single aggregate response for each network by concatenating all individual layer activations.

Representational Similarity Analysis (RSA)

Inline with previous work, we used Representational Similarity Analysis (RSA) (Kriegeskorte, Mur, & Bandettini, 2008) to assess the similarity between the activity patterns of the artificial and biological systems, in response to the same face stimuli. We computed representational dissimilarity matrices (RDMs) to quantify pairwise dissimilarities between the activations to

the stimuli. We computed one RDM per layer in each network, in addition to a general network-level RDM based on the whole-network activations. For the MEG data, we computed the RDMs at each sensor and then averaged them across all subjects. We computed the similarity between the RDMs in each sensor and the RDM of each layer of the three networks. We did the same for the whole network RDM.

All the correlations, in the RDM cells but also across RDMs (similarity scores) were calculated using Pearson correlation.

Results

Figure 1 C shows that the three networks provided similar global correlation levels with the MEG data when presented with the same face stimuli. Although the correlations levels are moderate, they are comparable to previous findings, including the study by Jiahui et al. (2022). We also found that MEG gradiometer channels yielded higher correlations than the magnetometers. In terms of classification accuracy, all three networks achieved reasonable performances on the CelebA stimuli, with the Resnet50 reaching the best decoding accuracy (Figure 1 B). In addition, the topographical maps of the MEG-Model similarity scores (Figure 1 C) suggest a prominent role of channels over occipital and, to some extent, central areas, with very similar topographies across the three architectures.

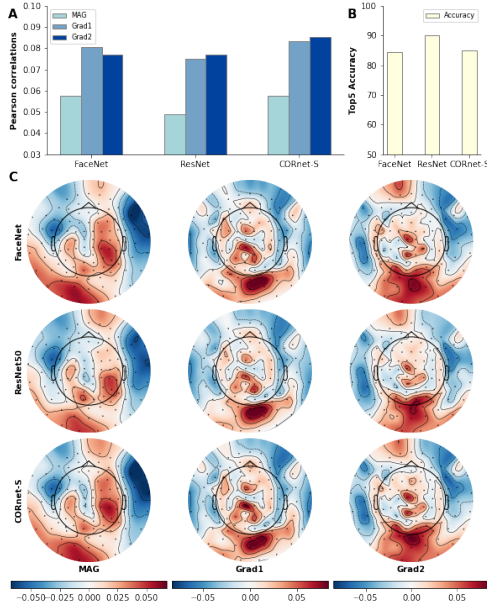


Figure 1: (A) Maximum similarity scores obtained across all MEG sensors for each architecture activations. (B) ANN performances on CelebA face stimuli. (C) Topographical maps of the MEG-Model similarity scores for each architecture and MEG sensor type (i.e. single RDM per architecture).

The topographies shown in Figure 1 C were obtained using a single RDM for the entire network. In contrast, Figure 2 depicts the max similarity scores of all layers of each network (panel A) and the topographical maps of correlations obtained

with the layer that yielded the highest correlation (panel B). Interestingly, we observed a drop in similarity between the MEG activity and the last few layers at the end of the three architectures, especially for ResNet50 and CORnet-S.

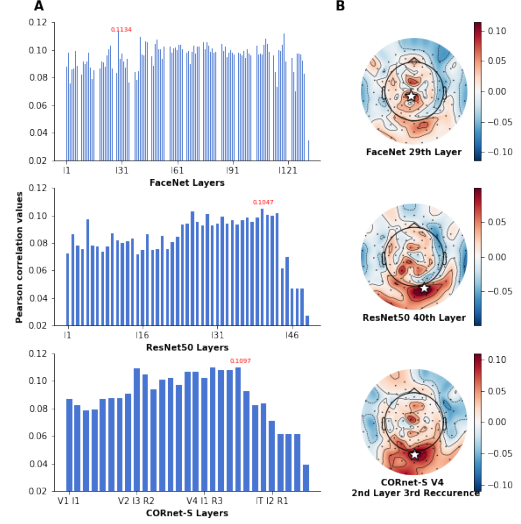


Figure 2: (A) Networks layers' maximum correlations with the MEG (planar gradiometer) sensors. (B) Topographical maps of the correlation values obtained with the specific layer that yields the highest similarity score. White stars indicate MEG sensors with the highest correlation values

Discussion

The three networks examined here succeeded to some extent in capturing neuromagnetic signatures of the brain dynamics involved in face recognition. However, this success is limited as the observed correlations are very moderate. This said, the levels of correlations are aligned with those reported in a similar study that used fMRI data (Jiahui et al., 2022). We should also keep in mind that the results presented here are first steps towards building new AI-driven models of the face recognition system in humans. We plan to extend these investigations in several directions. First we will further investigate the robustness of our observations including careful considerations of potential caveats and known limitations of the RSA approach (Dujmović, Bowers, Adolfi, & Malhotra, 2022). We also plan to explore the added value of source space MEG data analysis and examine distinct frequency bands of the MEG signal. Finally, even though the added neuroscientific value of the findings at this stage is arguably limited (i.e. what have we learned about the face recognition system that we did not already know?), we believe the current work to be part of a much-needed collective thrust in the field towards designing and validating new tools and methodological frameworks that will enable significant advances bridging neuroscience and AI research over years to come.

References

- Cao, Q., Shen, L., Xie, W., Parkhi, O. M., & Zisserman, A. (2017). *Vggface2: A dataset for recognising faces across pose and age*. doi: 10.48550/ARXIV.1710.08092
- Chang, L., Egger, B., Vetter, T., & Tsao, D. Y. (2021). Explaining face representation in the primate brain using different computational models. *Current Biology*, 31(13), 2785-2795.e4. doi: <https://doi.org/10.1016/j.cub.2021.04.014>
- Dujmović, M., Bowers, J. S., Adolphi, F., & Malhotra, G. (2022). The pitfalls of measuring representational similarity using representational similarity analysis. *bioRxiv*. doi: 10.1101/2022.04.05.487135
- He, K., Zhang, X., Ren, S., & Sun, J. (2015). *Deep residual learning for image recognition*. doi: 10.48550/ARXIV.1512.03385
- Jas, M., Larson, E., Engemann, D. A., Leppäkangas, J., Taulu, S., Hämäläinen, M., & Gramfort, A. (2018). A reproducible meg/eeg group study with the mne software: Recommendations, quality assessments, and good practices. *Frontiers in Neuroscience*, 12. doi: 10.3389/fnins.2018.00530
- Jiahui, G., Feilong, M., Oleggio Castello, M. V. d., Nastase, S. A., Haxby, J. V., & Gobbini, M. I. (2022). Modeling naturalistic face processing in humans with deep convolutional neural networks. *bioRxiv*. doi: 10.1101/2021.11.17.469009
- Kietzmann, T. C., Spoerer, C. J., Sörensen, L. K. A., Cichy, R. M., Hauk, O., & Kriegeskorte, N. (2019). Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences*. doi: 10.1073/pnas.1905544116
- Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2. doi: 10.3389/neuro.06.004.2008
- Kubilius, J., Schrimpf, M., Kar, K., Rajalingham, R., Hong, H., Majaj, N., ... DiCarlo, J. J. (2019). Brain-like object recognition with high-performing shallow recurrent anns. In (Vol. 32).
- Kubilius, J., Schrimpf, M., Nayebi, A., Bear, D., Yamins, D. L. K., & DiCarlo, J. J. (2018). Cornet: Modeling the neural mechanisms of core object recognition. doi: 10.1101/408385
- Liu, Z., Luo, P., Wang, X., & Tang, X. (2015, December). Deep learning face attributes in the wild. In *Proceedings of international conference on computer vision (iccv)*.
- Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A. J., ... Kording, K. P. (2019). A deep learning framework for neuroscience. *Nature Neuroscience*. doi: <https://doi.org/10.1038/s41593-019-0520-2>
- Schroff, F., Kalenichenko, D., & Philbin, J. (2015, jun). FaceNet: A unified embedding for face recognition and clustering. In *2015 IEEE conference on computer vision and pattern recognition (CVPR)*. doi: 10.1109/cvpr.2015.7298682
- Storrs, K. R., Kietzmann, T. C., Walther, A., Mehrer, J., & Kriegeskorte, N. (2020). Diverse deep neural networks all predict human it well, after training and fitting. *bioRxiv*. doi: 10.1101/2020.05.07.082743
- Wakeman, D., & Henson, R. (2015, 01). A multi-subject, multi-modal human neuroimaging dataset. *Scientific Data*, 2. doi: 10.1038/sdata.2015.1
- Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111. doi: 10.1073/pnas.1403112111