Neural Representations of face recognition in biological and artificial systems: Insights from MEG and CNNs

Summary: Artificial neural networks, inspired by brain structure and function, have surpassed human performance in various tasks, but the link between Artificial Intelligence and neuroscience is still underexplored. Combining these fields has offered mutual reinforcement, especially in the field of Neuro-AI, where comparing artificial and biological systems in cognitive tasks, such as visual categorization, has yielded promising insights (Kubilius et al., 2019; Yamins et al., 2014). Face recognition, however, is less explored in this context. Do Convolutional Neural Networks (CNNs) trained for face recognition mimic neural dynamics of face recognition in brain circuits? A question addressed only by a handful of studies, which in non-human primate mainly focus on the IT cortex, and in humans, largely rely on fMRI or behavioral data (Chang et al., 2021; Jiahui et al., 2023; Kathrina Dobes et al., 2023). Here we compare human brain activity collected using Magnetoencephalography (MEG) during a face recognition task to activations across seven CNNs trained on the same task. Compared to previous work, we leverage the high temporal resolution of MEG and source reconstruction techniques to compare these models to the brain across time, frequency, and space. Out of the tested models, FaceNet emerged as the most brain-like model during face recognition. Crucially, training on face recognition, rather than on object recognition or both simultaneously, was necessary and sufficient for high model-brain similarity. In terms of temporal alignment, peak similarities were observed around 170ms which corresponds to the M170-component linked with face perception. Examining the Fusiform Face Area (FFA), we observed that, compared to an untrained model, the similarity to FaceNet trained on face recognition significantly increased, from 0.02 to 0.08 in certain FFA regions. Our study provides novel insights into the spatio-temporal similarity patterns between artificial and biological neural responses associated with face recognition.

Methods

MEG Data: We utilized 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 using 3 types of stimuli: 150 Familiar (Famous), 150 Unfamiliar, and 150 scrambled faces. Familiar Stimuli are pictures of famous celebrities and unfamiliar are unknown pictures to subjects. We replicated the preprocessing steps reported in the original study using the Biomag script (Jas et al., 2018). For each subject, we had 450 trial of 1000ms: a 200 ms baseline and 800ms for the actual stimuli presentation. We performed source reconstruction to estimate the time courses of 8200 voxels on the brain surface. In the end, we obtained 450 trials for every subject, with a 1000ms signal for every voxel.

Network training and extraction of activity: We selected seven networks for our study: (i) The FaceNet backbone (Schroff, Kalenichenko, & Philbin, 2015), (ii) The SphereFace backbone, both specifically designed for face recognition, (iii) ResNet50 (He et al., 2015), (iv) MobileNet (Mark et al., 2019), (v) VGG16 (Simonyan & Zisserman, 2014), (vi) Inception (Szegedy et al., 2015) designed for object recognition, and finally (vii) CORnet-S (Kubilius et al., 2018), built to model the visual cortex. For the seven networks we performed the same classification tasks. First, we trained the networks from scratch on the VGGFace dataset (Cao et al., 2017). Subsequently, we fine-tuned on a distribution of the celebA dataset (Liu et al., 2015), similar to the stimuli used in the MEG experiment. Additionally, we trained all the seven architectures on ImageNet (Deng et al., 2009), which lacks a human faces category. Finally, we trained these networks on a dual task of face and object recognition by augmenting the ImageNet dataset with a face category, by selecting random faces from celebA. To extract the layer activations necessary for subsequent similarity analyses, we fed the three types of stimuli used in the MEG experiment to the networks and saved the responses of all layers.

Representational Similarity Analysis (RSA): In line with prior research, we employed Representational Similarity Analysis (RSA) (Kriegeskorte et al., 2008) to evaluate the similarity between the activity patterns of artificial and biological systems in response to identical face stimuli. For each network and experiment type (face recogni-

tion, object recognition, and dual task), we generated one Representational Dissimilarity Matrix (RDM) per layer, as well as for randomly initialized models. For the MEG data, we segmented the 8200 voxels into 450 regions of interest (ROI) based on the "aparc_sub" brain atlas. RDMs were computed using two methods: (i) We took the mean signal of each region using the time segment from 100ms to 600ms. (ii) For every region we took the voxels response for each time point to build an RDM. To determine similarity scores, we calculated correlations between the RDMs of each layer across the networks for the three experiments and the randomly initialized network, and the brain RDMs obtained from both methods (i) and (ii). All correlations, within the RDM cells and across RDMs (similarity scores), were computed using Pearson correlation.

Results We began by a cross-model comparison. In each model, we identified the layer that exhibited the highest similarity to the brain regions and used it as a model-score. FaceNet emerged as the model with the highest similarity to the brain, with a similarity score of 0.125 (while the other models hover around 0.1, and for a random network it was at 0.06). The results below primarily feature results obtained with FaceNet. We then explored whether a distinct training paradigm could yield higher similarities. However, for the FaceNet trained on object recognition or the dual task the similarities did not reach the same level as when training on face recognition. This suggests the necessity of having a model specifically trained on face recognition to achieve high similarities with the brain. To understand the evolution of similarities across layers, we presented three different types of stimuli (familiar, unfamiliar, and scrambled) to both trained and untrained FaceNet (specifically trained on face recognition). Our findings revealed that only the trained FaceNet, when presented with familiar stimuli, exhibited high similarity. Conversely, the other two types of stimuli, when presented to either trained or untrained models, yielded low similarities. The same was true for familiar stimuli presented to an untrained model. Additionally, we observed a decline in similarities in the trained FaceNet at the deeper layers. This observation suggests that these layers may not be engaging in face recognition in a manner analogous to the brain. To understand how the similarity scores were distributed across brain regions, each brain region was assigned its highest similarity score when compared to a trained FaceNet. These values were then visualized on a 3D brain map (Fig. 1). This analysis revealed distinct clusters with high similarities primarily in occipital areas, and some less prominent clusters in later visual areas. However, when plotting the similarities obtained after presenting scrambled and unfamiliar stimuli, these clusters disappeared. In addition, when considering the Fusiform Face Area (FFA) (Fig. 2), we observed that training on the face recognition task increased its similarities from 0.02 to 0.08, a significant improvmeent that was not observed in other brain regions. This effect remained consistent specifically when presenting familiar stimuli. Finally, we identified a peak of similarities at around 170 ms, which closely aligns with M170—an Event-Related Potential (ERP) associated with face perception in the brain.

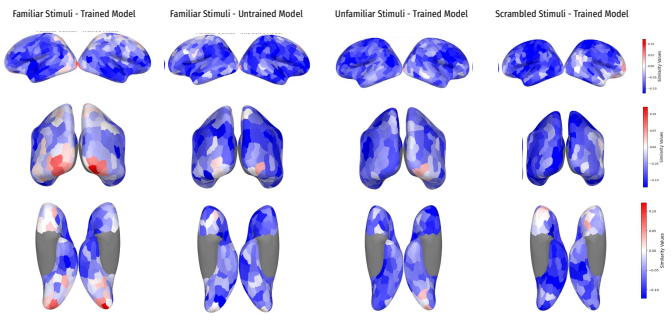

Figure 1: FaceNet-regions similarity values

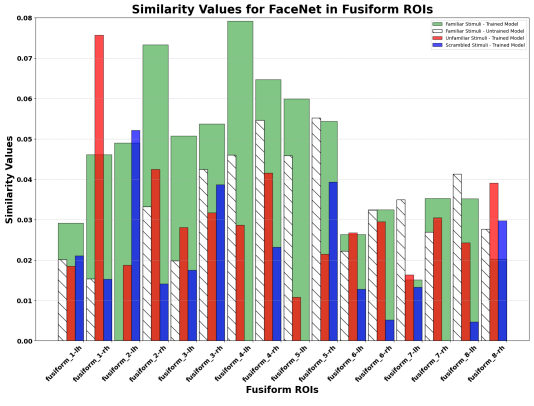

Figure 2: FaceNet-FFA similarity values

Conclusion Our work suggests that training ANNs on the same face recognition task conducted by humans, increases the similarity between ANN activations (esp. FaceNet) and neuromagnetic brain responses. Our findings also highlight the importance of considering the temporal dynamics and spatial distribution of the similarities across layers and brain regions.