Convolutional neural networks align early in training with neural representations

H. Steven Scholte (h.s.scholte@uva.nl)
Psychology Research Institute, University of Amsterdam, The Netherlands

Julio Smidi (julio.smidi@student.uva.nl)
Psychology Research Institute, University of Amsterdam, The Netherlands

Jessica Loke (j.loke@uva.nl)
Psychology Research Institute, University of Amsterdam, The Netherlands

Niklas Müller (n.muller@uva.nl)
Psychology Research Institute, University of Amsterdam, The Netherlands

Iris I. A. Groen (i.i.a.groen@uva.nl)
Informatics Institute, University of Amsterdam, The Netherlands
Psychology Research Institute, University of Amsterdam, The Netherlands

Marcel A. J. van Gerven (marcel.vangerven@donders.ru.nl)
Donders Institute for Brain Cognition and Behaviour, Radboud University Nijmegen, The Netherlands

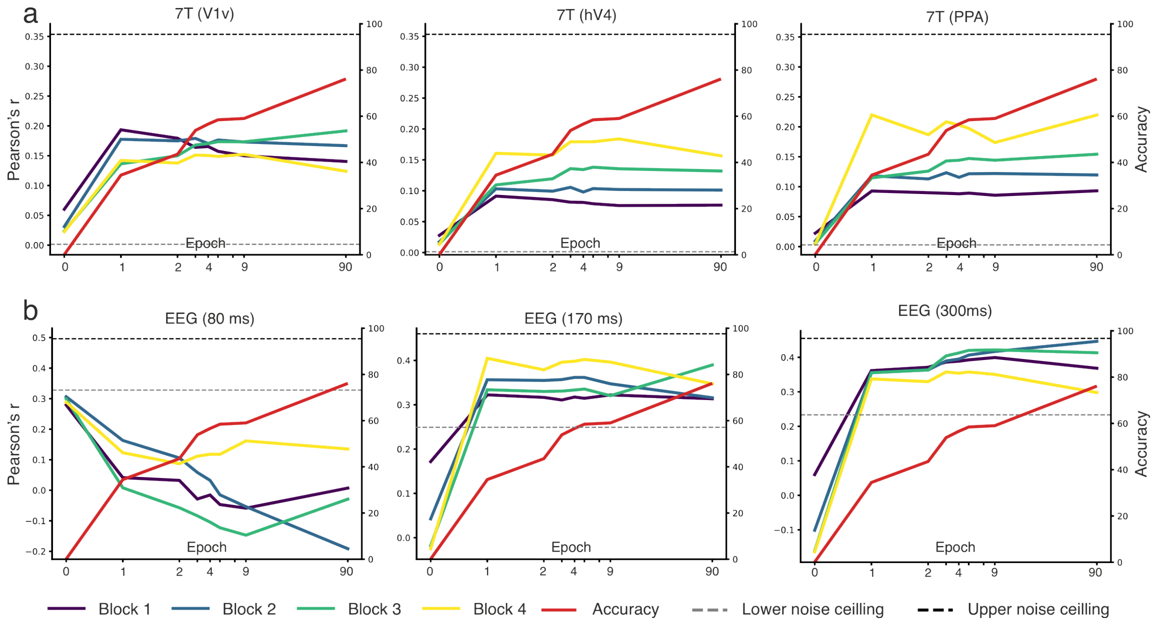

Figure 1: **a)** Alignment (RSA) of EEG (posterior electrodes) with output of same 4 ResNet-50 blocks for 80, 170 and 300 ms. **b)** Alignment (RSA) of BOLD-MRI with output of the 4 ResNet-50 blocks for V1v, hV4 and PPA. For both **(a)** and **(b)** epochs are plotted on a log-scale.

Abstract

Task-optimized deep convolutional neural networks (DCNNs) achieve human-level performance in object recognition and are leading in explaining neural activity across various brain measurement modalities. DCNNs are trained over numerous iterations to improve performance on a task, typically object recognition, whereby the underlying assumption is that optimizing network performance translates to better explanatory power for brain activity. Contrary to this assumption, our analysis of two published datasets (fMRI, EEG) reveals that the optimal alignment between brain activity and DCNNs already occurs after the first or one of the earliest iterations, and that changes in the brain-alignment are unrelated to changes in task-performance. This implies that extensive training on one task does not result in optimal brain alignment with visual cortex. It further suggests that much could be gained by aligning the training over epochs of a DCNN with learning in biological organisms.

Keywords: Alignment; Deep neural network; Development; Encoding models; EEG; BOLD-fMRI

Introduction

Deep convolutional neural networks (DCNNs) have emerged as state-of-the-art models of primate visual processing, in particular object recognition, rivaling human performance in specific tasks (Kell & McDermott, 2019). Representational similarity analysis, in which pairwise comparisons of stimulus responses are correlated between brain responses and DCNNs (Kriegeskorte et al., 2008) or linear encoding models that regress convolutional features of task-optimized DNNs onto neural data achieve high performance across multiple modalities, including EEG (Gifford et al., 2022), MEG (Seeliger et al., 2018), fMRI (Storrs et al., 2021), and electrophysiology (Yamins et al., 2014). The effects of training strategy, training dataset, and/or model architecture on DNN-brain alignment have been studied extensively (e.g., Conwell et al., 2022) . The majority of brain alignment studies have used DCNNs optimized for specific tasks, often object recognition, and the implicit assumption has been that the better model is at the task, the higher the brain alignment should be. However, it is possible that the alignment of neural activity with DCNNs is, for the vast majority, not based on the optimized tasks but rather more basic visual processes. For instance, Seijdel et al. (2020) and Loke et al. (2024) have shown that, for the datasets analysed in those studies, most of the brain alignment resulted from processing of scene segmentation and local features, not object processing. While untrained networks are often included as control comparisons to assess the influence of network training on brain alignment (Xu & Vaziri-Pashkam, 2021), the question how representations align as network training evolves has received much less attention. Here, we evaluate, for one DCNN architecture (ResNet-50) (He et al., 2016) that aligns well with neural activity, to what degree neural alignment changes as a function of training. If alignment is based on task specialisation we would expect to see an increase of alignment with an increase in task performance. However, if alignment is mainly based on more fundamental low- and mid-level statistical differences between images, it is possible that peak alignment occurs after only a limited amount of epochs and that changes in task performance and brain alignment are unrelated.

Methods

For evaluating brain alignment of DCNNs we used the Natural Scenes Dataset (NSD; Allen et al., 2022), an extensive fMRI dataset including neural responses of eight participants to 73,000 images (30,000 images per participant) from the COCO dataset (Lin et al., 2014). For the current analysis we used a subset of the 1,000 (872 after removing missing trials) COCO stimuli that were seen by all participants and the accessory fMRI per-trial responses as preprocessed in the development kit of Gifford et al. (2023). Specifically, we analyzed responses from the V1 ventral, V4 and parahippocampal place area (PPA) ROIs. We obtained EEG data from Gifford et al. (2022). The dataset contains EEG responses of 10 subjects over 82,160 trials each viewing THINGS (Hebart et al., 2019) images, consisting from 1854 concepts with 10 stimuli per concept. We used EEG measurements taken from all 17 channels overlying occipital and parietal cortex, averaging over the concept categories and stimuli repetitions. We train a set of five ResNet-50 (He et al., 2016) models on ImageNet (Russakovsky et al., 2015) for 90 epochs, saving model checkpoints at each epoch. We use different seeds for random weight initialization for each model instance to account for individual differences (Mehrer et al., 2020) in feature representations. For each of the datasets, we extract the features of all ResNet-50 instances after each identity block, for multiple epochs of training duration. Using representational similarity analysis (RSA; Kriegeskorte et al. (2008), we transform these layer features into representational dissimilarity matrices (RDMs), one for each chosen layer, epoch and seed. For the EEG data we obtained RDMs for each timepoint of interest (80, 170, and 300ms) and for the fMRI for each region of interest (V1v, V4, PPA). We compared the DCNN RDMs with the neural RDMs using Pearson correlation. We report the average over all subjects and ResNet-50 instances within each modality.

Results

Figure 1a shows the development of DCNN alignment with fMRI responses (Pearson correlation) as a function of training epochs (average of 5 ResNet-50s) on a logarithmic scale for ROIs V1v, V4, and PPA of the NSD dataset (Allen et al., 2022). For block 1 and 2 brain alignment peaks at epoch 1 for all three ROIs and drops or remains approximately constant afterwards. The highest alignment for V1v is observed for block 1 in epoch 1. Block 4 has the best alignment for hV4 and PPA at epoch 9 (hV4) and epoch 1 (PPA). Block 3 does show a consistent increase of alignment with training albeit only being the best performing block for V1v in the final epoch.

Correlations with EEG recordings **(Fig. 1b**) at 80 ms are maximal when the models are untrained, becoming worse for all four blocks during training. For later time points (170 and 300 ms) alignment is substantially increased after epoch 1 but is constant during further training. For these time points there is either a drop or no change in alignment for blocks 1, 2 and 4, while for block 3 (and block 2 for 300 ms) alignment keep increasing until the final epoch.

Together the NSD and EEG results are very similar, the bulk of the alignment has taken place after 1 to 3 epochs at which time model performance is still far from optimal. Also, apart from a change in alignment between epoch 0 and 1 we observe no clear relationship between changes in accuracy and changes in brain alignment in either data-set.

Finally we wish to note that a mapping of hierarchy of DCNNs over time and space with the brain (V1v and 80 ms EEG early layers, 300 ms and PPA later layers) appears to be more apparent when taking alignment over epochs into account than with only the fully trained networks.

Discussion

There is a weak and a strong version of the idea of aligning task-optimized DCNNs with the brain. In the strong version a network that is more optimized for a task should have a better alignment with the brain. The weak version of this idea presumes that a specific part of the brain, involved in specific tasks, should align better to DCNNs whose performance is optimized for those tasks (Dwivedi et al., 2021). No decrease in alignment with training is explicitly expected in either of these scenarios. The data presented in this paper is in direct contradiction with the strong version and potentially problematic for the weak version of the task-optimized approach since we find, in part of the data, decreases in brain alignment with an increase in training and only two data points (out of 144) for which a layer is optimally aligned after full training. At least two scenarios are possible to explain the observed patterns. First, the weak, but not the strong version of the task-optimized DCNNs approach applies. That is, the DCNN is a model of some part of brain processing, which is potentially not, or only to a limited degree, covered in the current data (although the drop in alignment remains an issue). Second, the task used for network optimisation (in this case object detection with ImageNet) aligns only to a limited degree with what humans do. Potentially, both the DCNN and the human brain learn the basic statistical structure of the world in their development / training but solve the task of object recognition ultimately in a different way, resulting in an earlier optimal alignment. This connects well with data indicating that DCNNs and the brain align in the partition of variance related to mid-level processing (Seijdel et al., 2020; Loke et al., 2024).

Conclusion

Our data shows that optimal alignment between the brain and DCNNs might not be found in task-optimized networks and that there is no correlation between increase in task performance and brain alignment, at least for the data under consideration. Further, alignment often peaks early in training suggesting that the basis of alignment is (partly) unrelated to a specific task. In general our results show that it is paramount to consider the entire training trajectory when considering brain/DCNN alignment.

Acknowledgments

This work is supported by the Interdisciplinary PhD Programme of University of Amsterdam Data Science Center as well as the Dutch Research Council (NWO) via the Dutch Brain Interface Initiative (DBI2) with project number 024.005.022.

References

Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Prince, J. S., Dowdle, L. T., . . . others (2022). A massive 7t fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, *25*(1), 116–126.

Conwell, C., Prince, J. S., Kay, K. N., Alvarez, G. A., & Konkle, T. (2022). What can 1.8 billion regressions tell us about the pressures shaping high-level visual representation in brains and machines? *BioRxiv*, 2022–03.

Dwivedi, K., Bonner, M. F., Cichy, R. M., & Roig, G. (2021). Unveiling functions of the visual cortex using task-specific deep neural networks. *PLoS computational biology*, *17*(8), e1009267.

Gifford, A. T., Dwivedi, K., Roig, G., & Cichy, R. M. (2022). A large and rich eeg dataset for modeling human visual object recognition. *NeuroImage*, *264*, 119754.

Gifford, A. T., Lahner, B., Saba-Sadiya, S., Vilas, M. G., Lascelles, A., Oliva, A., . . . Cichy, R. M. (2023). The algonauts project 2023 challenge: How the human brain makes sense of natural scenes. *ArXiv preprint arXiv:2301.03198*.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Hebart, M. N., Dickter, A. H., Kidder, A., Kwok, W. Y., Corriveau, A., Van Wicklin, C., & Baker, C. I. (2019). Things: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PloS One*, *14*(10), e0223792.

Kell, A. J., & McDermott, J. H. (2019). Deep neural network models of sensory systems: windows onto the role of task constraints. *Current Opinion in Neurobiology*, *55*, 121–132.

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*, 249.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., . . . Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *Computer vision–eccv 2014: 13th european conference, zurich, switzerland, september 6-12, 2014, proceedings, part v 13* (pp. 740–755).

Loke, J., Seijdel, N., Snoek, L., Sörensen, L. K., van de Klundert, R., van der Meer, M., . . . Scholte, H. S. (2024). Human visual cortex and deep convolutional neural network care deeply about object background. *Journal of Cognitive Neuroscience*, *36*(3), 551–566.

Mehrer, J., Spoerer, C. J., Kriegeskorte, N., & Kietzmann, T. C. (2020). Individual differences among deep neural network models. *Nature Communications*, *11*(1), 5725.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., . . . others (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*, 211–252.

Seeliger, K., Fritsche, M., Güçlü, U., Schoenmakers, S., Schoffelen, J.-M., Bosch, S. E., & Van Gerven, M. (2018). Convolutional neural network-based encoding and decoding of visual object recognition in space and time. *NeuroImage*, *180*, 253–266.

Seijdel, N., Tsakmakidis, N., De Haan, E. H., Bohte, S. M., & Scholte, H. S. (2020). Depth in convolutional neural networks solves scene segmentation. *PLoS Computational Biology*, *16*(7), e1008022.

Storrs, K. R., Kietzmann, T. C., Walther, A., Mehrer, J., & Kriegeskorte, N. (2021). Diverse deep neural networks all predict human inferior temporal cortex well, after training and fitting. *Journal of Cognitive Neuroscience*, *33*(10), 2044–2064.

Xu, Y., & Vaziri-Pashkam, M. (2021). Limits to visual representational correspondence between convolutional neural networks and the human brain. *Nature communications*, *12*(1), 2065.

Yamins, D. L., 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*(23), 8619–8624.