Implementing ANNs with TensorFlow

Project Report - Algonauts Challenge 2023

Introduction

This project is inspired by the "Algonauts Project 2023 Challenge: How the Human Brain Makes Sense of Natural Scenes" (Gifford et al., 2023). Our goal is to use artificial neural networks (ANNs) to predict the functional magnetic resonance imaging (fMRI) response of humans in the visual cortex to complex natural visual scenes, as recorded in the "Natural Scenes Dataset" (NSD) (Allen et al., 2022), which is in turn based on the "Common Objects in Context" dataset (COCO) (Lin et al., 2014). Being able to accurately predict fMRI data may help us to both better understand the complex processing of information in the brain, as well as inspire new approaches in computer vision.

The relationship between information processing in the brain and in ANNs has been subject of a large number of studies in recent years. This is especially true for the visual cortex and ANNs designed to process image data, such as convolutional neural networks (CNNs). Just as ANNs were originally inspired by the human brain (Agatonovic-Kustrin & Beresford, 2000), CNNs were inspired by the primary visual cortex (V1) (Kim, Kim & Lee, 2016).

Further, designing ANNs closely related to processing in the brain may yield significant advantages. For example it has been shown that CNN models incorporating a layer that better resembles V1 may improve robustness against adversarial attacks, which involve the manipulation of input data with the aim of inducing misclassification (*34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, 2020). Similarly, (Kim, Kim & Lee, 2016) proposed the implementation of a so called "ON/OFF ReLU" to improve CNN performance, which is inspired the retinal structure of the early visual system.

Conversely, ANNs present a powerful tool to analyze and explain processing of visual stimuli in the (human) brain (Cichy & Kaiser, 2019) (Yamins & DiCarlo, 2016). A multitude of previous studies have worked on finding and developing adequate ANN models to predict and represent neural correlates of visual processing, which has led to tremendously insight into brain functionality (Cichy, Khosla, Pantazis, Torralba & Oliva, 2016) (Dupre la Tour, Lu, Eickenberg & Gallant, 2021) (Dwivedi, Bonner, Cichy & Roig, 2021) (Han et al., 2019) (St-Yves & Naselaris, 2018) (Svanera, Morgan, Petro & Muckli, 2021) (Wang et al., 2021) (Yamins et al., 2014) (Zeman, Ritchie ,Bracci & Beeck, 2020) (Zhang et al., 2019) (Zhuang et al., 2021).

A variety of different ANNs have been applied in the literature. Most commonly used are CNN models, which have shown great success at predicting neural activity in the visual system (Dwivedi et al., 2021) (Yamins & DiCarlo, 2016) (Zeman et al., 2020). Used architectures include

state-of-the-art CNNs such as (a modified version of) ResNet-50 (Dwivedi et al., 2021), VGG-19, GoogLeNet and AlexNet (Zeman et al., 2020).

Recently there have also been advancements in the use of self-supervised models for this application, partly outperforming classical supervised networks (Svanera et al., 2021). Variational auto-encoders have been shown to match accuracy of CNNs in predicting brain fMRI activity in early visual areas, with lower accuracy in later areas (Han et al., 2019). Generally, unsupervised and supervised learning models seem to predict visual cortex activity roughly equally well (Zhuang et al., 2021). Other approaches focused on building networks closely representing available anatomical and neurophysiological information about brain areas, for example in a model for V1 in mice (Chen, Scherr & Maass, 2022), or the whole visual cortex of mice ("MouseNet") (Shi, Tripp, Shea-Brown, Mihalas & A Buice, 2022), taking knowledge about the number of neurons in each area and interconnectivity into account.

Strikingly, information processing in the visual cortex and DNNs seem to share a hierarchy of earlier and lower to later and higher level processing in distinct areas/layers (Saxe et al., 2021) (Dwivedi et al., 2021) (Cichy et al., 2016) (Zeman et al., 2020). Early deep neural network (DNN) layers show more similarities to low- and mid-level visual areas, while higher DNN layers are associated with activations in areas of later visual processing (Cichy et al., 2016). It has further been demonstrated that both early visual cortex areas such as V1 and early layers of CNNs encode shape information (Cichy et al., 2016) (Zeman et al., 2020), while higher-level visual areas such as the anterior ventral temporal cortex as well as final CNN layers encode category information (Zeman et al., 2020). Fittingly, DNNs trained for low-level visual tasks predict activity in early brain regions accurately, while DNNs trained for 3-dimensional perception tasks and semantic tasks perform well at predicting dorsal and ventral stream activity, respectively (Dwivedi et al., 2021).

Generally, ANNs performing better in visual tasks are found to be better at predicting visual cortex activity (Cadieu et al., 2014) (Saxe et al., 2021) (Yamins et al., 2014). However, this may depend on the specific brain area, with earlier visual areas being overall easier to predict (Cichy et al., 2016). Activations in V1 have even been reported to correlate better with untrained DNN models (Cichy et al., 2016). Overall, it further seems that optimizing performance of DNN models for computer vision task does not heighten correlation to brain activity beyond a certain point (Shi et al., 2022).

As we unfortunately do not have the time or resources available to build a model closely representing the human visual cortex or evaluate a variety of state-of-the-art computer vision models for their prediction of brain activity, we choose to work with the comparatively simple object detection model RetinaNet and plan to use feature maps of layers of differing depth to predict visual cortex activity in different areas, according to the hierarchical structure that DNNs and visual cortex share mentioned above.

RetinaNet follows a one-stage approach, which broadly means it uses only one network to make predictions regarding object localization and categorization, as opposed to a two-sage

approach, which first proposes regions of interest (ROIs), which are then fine-tuned in a second network (Lin et al., 2017). One-stage approaches are generally faster and involve models of lower complexity, making them suitable for our limited computational resources, but are considered less accurate (Lin et al., 2017). Importantly, the authors of the RetinaNet model proposed the new “Focal Loss”, which is designed to address a common problem of one-stage models called class imbalance (Lin et al., 2017). Class imbalance refers to the fact that most candidate locations for objects evaluated by the model during training do not contain any objects, which makes training inefficient, as these numerous easy negatives do not contribute strongly to useful training (Lin et al., 2017). Further, this imbalance constitutes a bias in training which can harm model performance (Lin et al., 2017). Focal loss deals with this by laying a heavier focus on badly classified examples during training (Lin et al., 2017). This seems to allow one-stage models such as the RetinaNet to match or even surpass accuracy of state-of-the-art two-stage models (Lin et al., 2017). Since performance on computer vision tasks is, as previously elaborated, of high importance for accurate prediction of visual cortex activity, RetinaNet seems to be an adequate choice for our model architecture.

.

Methods/Implementation

Our idea was to take the official algonauts colab tutorial (https://colab.research.google.com/drive/1bLJGP3bAo_hAOwZPHpiSHKlt97X9xsUw#scrollTo=gjQrI9AlzDqG) and change its implementation to work with Tensorflow instead of PyTorch, make it modular instead of a python notebook and use an object detection model instead of an object classification model. First of all, the data preprocessing was changed to work with Tensorflow by loading the images from the directory and differs from the colab tutorial in the way we map the corresponding fMRI data to the images. Further, we use a pretrained RetinaNet keras model that, in its original form (as .pb file), can take an image and predict object locations and classes but cannot output intermediate layer outputs. Therefore, it is necessary to know and have access to the exact architecture of the model (for a visualization see Fig. 1) which, in the case of the model we use, is known due to the fact that the pretrained model is built for keras.io that publishes the implementation used for training. With that we built a non-trained RetinaNet model, loaded the pretrained, same model and transferred the weights. Thus, we have a pretrained model with known layers and have access to it with a simple in-build function that outputs all intermediate features in a call. The algonauts colab tutorial follows mainly three steps to build a linearized encoding model for predicting brain data. These steps are first loading and downsampling features with principal component analysis (pca), training and predicting with linear regression and lastly evaluating the predictions with pearson's correlation. One can either do the whole process or single steps, but the second option requires to have the outputs of the preceding steps, therefore we save all intermediate variables. We chose to save the extracted and downsampled features and the predictions for the test and validation data since the challenge dataset is very large and therefore requires many computations and time.

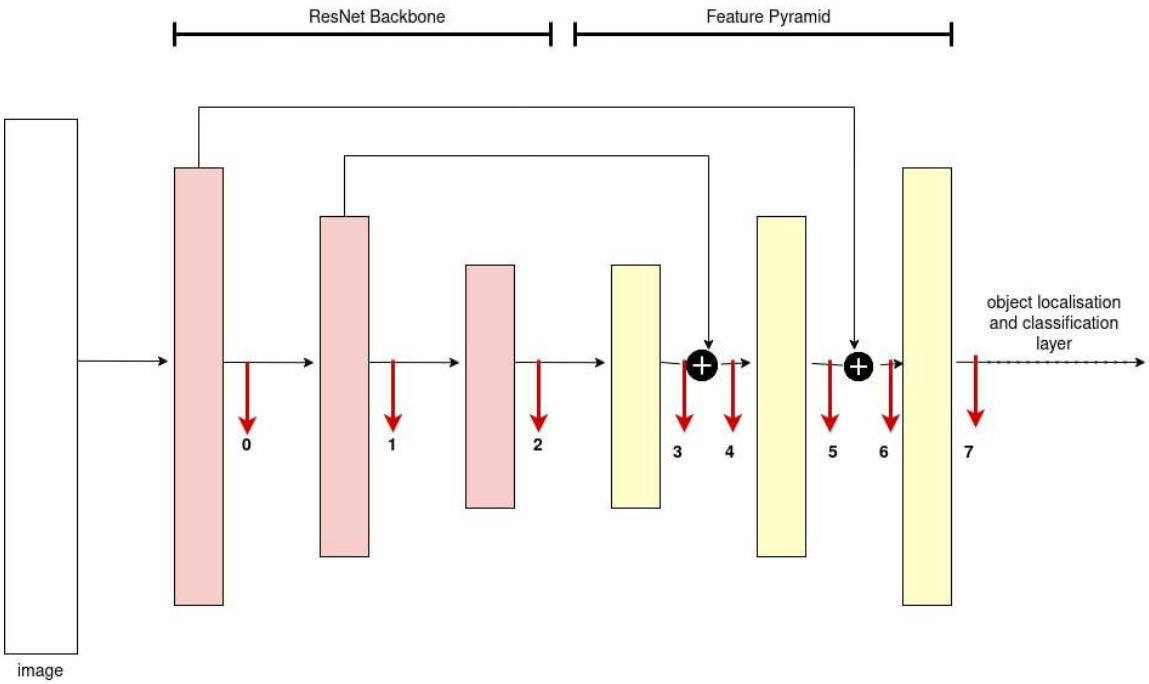

Figure 1 An abstract representation of the RetinaNet model architecture. The left white block represents the input image, the red blocks represent the layer blocks of the ResNet backbone and similarly the yellow blocks represent the layer of the feature pyramid. Each block or layer handles one size of feature maps and the intermediate outputs at the end of each block that we output with our extract function in the RetinaNet model is represented by red arrows with a number indicating which index should be set to get features of this layer.

Results

We tested the whole process for the first subject on the third and fifth feature map (as described in Fig.1), which took over an hour with the third layer as feature map and approximately 45 minutes with the fifth layer as feature layer. The accuracy was calculated with the pearson's correlation that results in values between –1 and 1 with 0 representing no correlation. The pearson's correlation averaged over the voxels is approximately 0, with a maximum value of 0.097 for the left hemisphere and 0.094 for the right hemisphere for both runs (for the third and the fifth layer as feature maps). Due to the computational and time limitations we were not able to run a lot of (not error throwing) full processes and in the first trials could only output an image like in the notebook. Therefore, we cannot provide more numbers. The displayed graphics show an overview over the correlation of the predicted and the target voxels based on feature maps from layer 3 (Fig.2) and layer 5 (Fig.3).

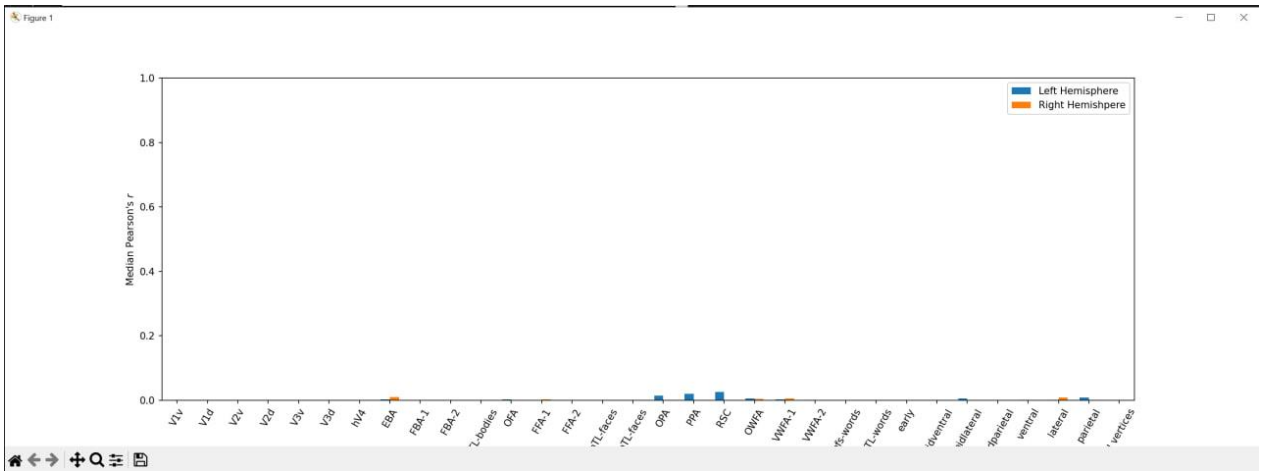

Figure 2 Pearson's correlation between fMRI activity predicted for each visual cortex area based on feature maps extracted from layer 3 and actual recorded fMRI activity

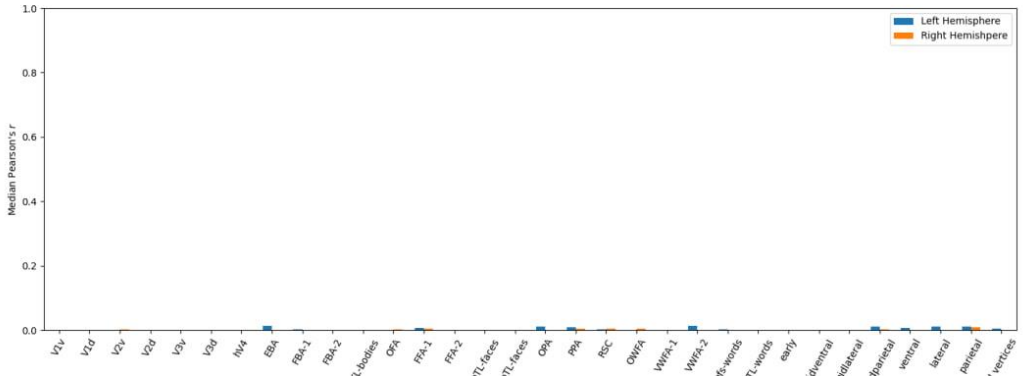

Figure 3 Pearson's correlation between fMRI activity predicted for each visual cortex area based on feature maps extracted from layer 5 and actual recorded fMRI activity

Discussion

In this project, we aimed at processing the data provided by the Algonauts Challenge 2023 to make it available in tensorflow, preprocess this data in tensorflow, extract feature maps from a pretrained model after applying the image data, downsampling the feature maps using pca and training a linearized encoding model to predict the fMRI reponse in the visual cortex of humans to the respective images. While we were able to perform all of these steps, our prediction accuracy is unfortunately very low.

One possible reason for our bad results is that the feature maps we extracted are non-informative, thus making it impossible for the linear regression to find sensible weights connecting the features to the fMRI data. This might be caused by an erroneous transfer of the weights of the pretrained RetinaNet onto our self-build RetinaNet model, however this is pure speculation as we could not identify a mistake in our code in this regard.

Additionally, our chosen model architecture may have been suboptimal, because RetinaNet is a feedforward model without recurrent connections. Such models have been found to perform worse at predicting fMRI responses to visual stimuli, which could be partly explained by missing recurrent connections constituting a deviation from the structure of the human visual system, where feedback connections are common (Kietzmann et al., 2019). However, this alone cannot explain our results, as feedforward models are still capable of achieving significantly higher accuracy than we attained. For example, the challenge tutorial uses a feedforward AlexNet and achieves a much higher accuracy.

Moreover, more sophisticated measures of evaluating adequate hyperparameters such as *x*-fold-cross validtation, in which training data is systematically split into different training and testing subsets, should be employed (BURMAN, 1989).

Future investigations might also make more elaborate distinctions between areas of the visual cortex which are to be predicted and investigate systematically which layers predict which areas best. For example predicting V1, V2, V3 and V4 separately instead of grouping them together as early visual regions, and especially separating predictions for higher areas of visual processing depending on their assumed function (for example grouping areas which are thought to be involved in facial recognition together).

Another approach to increase accuracy is the adoption of novel approaches of fitting the feature maps to the fMRI data. Dupre la Tour et al. (2021) criticized our approach of trying to fit feature maps from a given layer to certain visual cortex areas, instead arguing for learning a fitting of every layer to all regions, which we avoided in order to have fewer parameters in our model and thus make overfitting less likely. By learning a different regularization hyperparameter for each layer, their approach leads to non-predictive or redundant information being removed from the predictions while inducing finer mapping between layers and the visual cortex, resulting in higher prediction accuracy (Dupre la Tour et al., 2021).

Predicting a momentary fMRI response to images also does not take temporal dynamics into account, which might be crucial to advance our understanding of visual processing in humans (Dupre la Tour et al., 2021). For this purpose, additionally predicting other neural correlates such as magnetoencephalography (MEG) responses to videos seems very interesting. The advantage of MEG measurements in this regard is the considerably higher temporal resolution (Dale et al., 2000).

References

Dupre la Tour, T., Lu, M., Eickenberg, M., Gallant, J.L. (2021)

A finer mapping of convolutional neural network layers to the visual cortex. (2021). https://openreview.net/forum?id=EcoKpq43Ul8

Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017, August 7). *Focal Loss for Dense Object Detection*. http://arxiv.org/pdf/1708.02002v2

34th conference on neural information processing systems (NeurIPS 2020): Online, 6-12 December 2020. (2020). *Advances in neural information processing systems: Vol. 33*. Curran Associates, Inc.

Zhang, C., Qiao, K., Wang, L., Tong, L., Hu, G., Zhang, R.-Y., & Yan, B. (2019). A visual encoding model based on deep neural networks and transfer learning for brain activity measured by functional magnetic resonance imaging. *Journal of Neuroscience Methods*, *325*, 108318. https://doi.org/10.1016/j.jneumeth.2019.108318

St-Yves, G., & Naselaris, T. (2018). The feature-weighted receptive field: An interpretable encoding model for complex feature spaces. *NeuroImage*, *180*(Pt A), 188–202. https://doi.org/10.1016/j.neuroimage.2017.06.035

Han, K., Wen, H., Shi, J [Junxing], Lu, K.-H., Zhang, Y., Di Fu, & Liu, Z. (2019). Variational autoencoder: An unsupervised model for encoding and decoding fMRI activity in visual cortex. *NeuroImage*, *198*, 125–136. https://doi.org/10.1016/j.neuroimage.2019.05.039

Kim, J., Sangjun, O., Kim, Y., & Lee, M. (2016). Convolutional Neural Network with Biologically Inspired Retinal Structure. *Procedia Computer Science*, *88*, 145–154. https://doi.org/10.1016/j.procs.2016.07.418

Cichy, R. M [Radoslaw M.], & Kaiser, D. (2019). Deep Neural Networks as Scientific Models. *Trends in Cognitive Sciences*, *23*(4), 305–317. https://doi.org/10.1016/j.tics.2019.01.009

Agatonovic-Kustrin, S., & Beresford, R. (2000). Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. *Journal of Pharmaceutical and Biomedical Analysis*, *22*(5), 717–727. https://doi.org/10.1016/S0731-7085(99)00272-1

Dale, A. M., Liu, A. K., Fischl, B. R., Buckner, R. L., Belliveau, J. W., Lewine, J. D., & Halgren, E. (2000). Dynamic statistical parametric mapping: Combining fMRI and MEG for high-resolution imaging of cortical activity. *Neuron*, *26*(1), 55–67. https://doi.org/10.1016/S0896-6273(00)81138-1

Yamins, D. L. K., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19*(3), 356–365. https://doi.org/10.1038/nn.4244

Saxe, A., Nelli, S., & Summerfield, C. (2021). If deep learning is the answer, what is the question? *Nature Reviews. Neuroscience*, *22*(1), 55–67. https://doi.org/10.1038/s41583-020-00395-8

Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Prince, J. S., Dowdle, L. T., Nau, M., Caron, B., Pestilli, F., Charest, I., Hutchinson, J. B., Naselaris, T., & Kay, K [Kendrick] (2022). A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, *25*(1), 116–126. https://doi.org/10.1038/s41593-021-00962-x

Zeman, A. A., Ritchie, J. B., Bracci, S., & Beeck, H. op de (2020). Orthogonal Representations of Object Shape and Category in Deep Convolutional Neural Networks and Human Visual Cortex. *Scientific Reports*, *10*(1), 2453. https://doi.org/10.1038/s41598-020-59175-0

Cichy, R. M [Radoslaw Martin], Khosla, A., Pantazis, D., Torralba, A., & Oliva, A [Aude] (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, *6*, 27755. https://doi.org/10.1038/srep27755

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 of the United States of America*, *111*(23), 8619–8624. https://doi.org/10.1073/pnas.1403112111

Kietzmann, T. C., Spoerer, C. J., Sörensen, L. K. A., Cichy, R. M [Radoslaw 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 of the United States of America*, *116*(43), 21854–21863. https://doi.org/10.1073/pnas.1905544116

Zhuang, C., Yan, S., Nayebi, A., Schrimpf, M., Frank, M. C., DiCarlo, J. J., & Yamins, D. L. K. (2021). Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(3). https://doi.org/10.1073/pnas.2014196118

BURMAN, P. (1989). A comparative study of ordinary cross-validation, v -fold cross-validation and the repeated learning-testing methods. *Biometrika*, *76*(3), 503–514. https://doi.org/10.1093/biomet/76.3.503

Wang, H., Huang, L., Du, C., Li, D., Wang, B., & He, H. (2021). Neural Encoding for Human Visual Cortex With Deep Neural Networks Learning “What” and “Where”. *IEEE Transactions on Cognitive and Developmental Systems*, *13*(4), 827–840. https://doi.org/10.1109/TCDS.2020.3007761

Chen, G., Scherr, F., & Maass, W. (2022). A data-based large-scale model for primary visual cortex enables brain-like robust and versatile visual processing. *Science Advances*, *8*(44), eabq7592. https://doi.org/10.1126/sciadv.abq7592

Svanera, M., Morgan, A. T., Petro, L. S., & Muckli, L. (2021). A self-supervised deep neural network for image completion resembles early visual cortex fMRI activity patterns for occluded scenes. *Journal of Vision*, *21*(7), 5. https://doi.org/10.1167/jov.21.7.5

Cadieu, C. F., Hong, H., Yamins, D. L. K., Pinto, N., Ardila, D., Solomon, E. A., Majaj, N. J., & DiCarlo, J. J. (2014). Deep neural networks rival the representation of primate IT cortex for core visual object recognition. *PLoS Computational Biology*, *10*(12), e1003963. https://doi.org/10.1371/journal.pcbi.1003963

Dwivedi, K., Bonner, M. F., Cichy, R. M [Radoslaw Martin], & Roig, G [Gemma] (2021). Unveiling functions of the visual cortex using task-specific deep neural networks. *PLoS Computational Biology*, *17*(8), e1009267. https://doi.org/10.1371/journal.pcbi.1009267

Shi, J [Jianghong], Tripp, B., Shea-Brown, E., Mihalas, S., & A Buice, M. (2022). Mousenet: A biologically constrained convolutional neural network model for the mouse visual cortex.

PLoS Computational Biology, *18*(9), e1010427. https://doi.org/10.1371/journal.pcbi.1010427

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., & Dollár, P. (2014). *Microsoft COCO: Common Objects in Context.* https://doi.org/10.48550/arXiv.1405.0312

Gifford, A. T., Lahner, B., Saba-Sadiya, S., Vilas, M. G., Lascelles, A., Oliva, A [A.], Kay, K [K.], Roig, G [G.], & Cichy, R. M [R. M.]. (2023). *The Algonauts Project 2023 Challenge: How the Human Brain Makes Sense of Natural Scenes.* https://doi.org/10.48550/arXiv.2301.03198