



UvA-DARE (Digital Academic Repository)

A shallow residual neural network to predict the visual cortex response

Meijer, A.-R.J.; Visser, A.

Publication date

2019

Document Version

Submitted manuscript

[Link to publication](#)

Citation for published version (APA):

Meijer, A.-R.J., & Visser, A. (2019). *A shallow residual neural network to predict the visual cortex response*. Universiteit van Amsterdam. <https://arxiv.org/abs/1906.11578>

General rights

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

A shallow residual neural network to predict the visual cortex response

Anne-Ruth José Meijer

Universiteit van Amsterdam, The Netherlands

Arnoud Visser

Universiteit van Amsterdam, The Netherlands

Abstract—Understanding how the visual cortex of the human brain really works is still an open problem for science today. A better understanding of natural intelligence could also benefit object-recognition algorithms based on convolutional neural networks. In this paper we demonstrate the asset of using a shallow residual neural network for this task. The benefit of this approach is that earlier stages of the network can be accurately trained, which allows us to add more layers at the earlier stage. With this additional layer the prediction of the visual brain activity improves from 10.4% (block 1) to 15.53% (last fully connected layer). By training the network for more than 10 epochs this improvement can become even larger.

I. INTRODUCTION

Current state-of-the-art convolutional neural networks (CNNs) incorporate operations like maximum-based pooling of inputs [1], which were directly inspired by single-cell recordings from the V1 region of the mammalian visual cortex [2]. This inspiration is also beneficial for object recognition, because there is a correlation [3] observed between a CNNs ImageNet’s performance [4] and its Brain-Score [5]. However, this correlation can no longer be found for the most advanced CNNs [3]. Based on the hypothesis that by simplifying CNNs one could improve the understanding of ventral stream in the visual cortex, this year’s challenge of the Algonauts project is introduced [6]. The target of the 2019 challenge is to predict the activity in the human visual brain responsible for object recognition in two regions; the early visual cortex (EVC) and the inferior temporal (IT) cortex. Our study seems to confirm that simpler CNNs (more shallow) can have competitive prediction power when compared to “deeper” CNNs, because the earlier layers can be trained faster and more accurately.

II. EXPLAINING THE HUMAN VISUAL BRAIN CHALLENGE

The goal of the 2019 challenge is predict the response of two parts of the human brain responsible for object recognition. Two datasets are provided of brain recordings of 15 human subjects looking at pictures with objects from the ImageNet dataset [7]. The brain recordings are respectively functional magnetic resonance imaging (fMRI) for Track 1 and Magnetoencephalography (MEG) for Track 2 (see Fig. 1). fMRI is a technique which detect changes in blood flow with spatial resolution of millimeters; MEG is a technique that changes in magnetic fields with a temporal resolution of milliseconds.

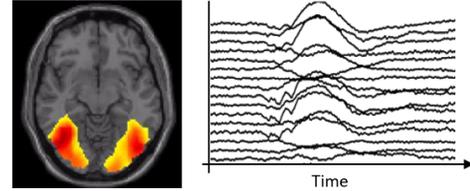


Fig. 1. Brain activity recorded with respectively the fMRI and MEG technique (Courtesy Algonauts project²)

The datasets are recorded in such a way that the observations correspond with activity in the early visual cortex (EVC) and the inferior temporal (IT) cortex, with respect to space (fMRI-data of Track 1) and time (MEG-data of Track 2). The response of the human brain and the CNN models is made possible by converting the incommensurate signals into the same similarity space, which are defined by representational dissimilarity matrices (RDMs) [8]. The recorded and predicted RDM are then compared based on the Spearman correlation, which is defined as the Pearson correlation between rank variables [9]. The result is normalized against the correlation an ideal model could give, taking into account the variation and noise in the dataset. As baseline, the prediction of the classic CNN AlexNet is given [10]. The overall evaluation procedure is illustrated in Fig. 2.

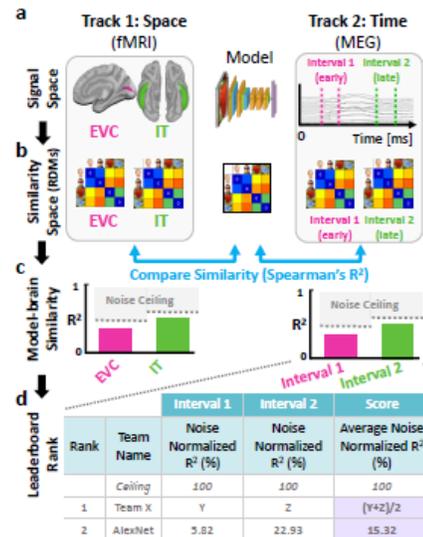


Fig. 2. The evaluation procedure of the 2019 Challenge (Courtesy [6])

²http://algonauts.csail.mit.edu/fmri_and_meg.html

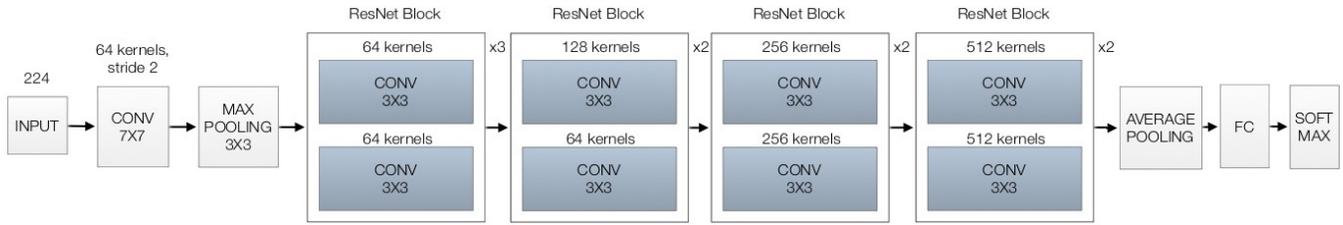


Fig. 3. ResNet20 architecture

III. RELATED WORK

The challenge is inspired by the initiative to find a Brain-Score [5], which found a correlation between the ImageNet performance and the Brain-Score. Yet, for the CNNs with the highest performance this correlation becomes less strong. The conclusion of the study was that DenseNet169, CORnet-S and ResNet-101 were the most brain-like CNNs. Yet, a number of smaller (i.e. more shallow) networks performed quite competitive, leaving the road open to better understand the ventral stream with simpler CNNs.

Interesting here is the good performance of CORnet-S [11], which is inspired by the relatively shallow cortical hierarchy (4-8 levels) which is observed in human brains. CORnet-S contains elements of ResNet and DenseNet, in the sense that it allows the backpropagation to reach the earlier layers by skipping layers, yet it does this with a biological type of unrolling. Unfortunately we could not reproduce the good performance of CORnet-S for this challenge (see for more details [12]), so this biologically interesting approach is no further studied in this paper.

Another interesting approach are the deep CNNs proposed by Kar *et al* [13]. Here they observed that deeper CNNs predicted neural responses better than shallower models, if they have unrolling mechanisms present. The best predictions were from ResNet-50 and ResNet-101. Yet the prediction were for the MEG-data of the IT region, while in this study we want to concentrate on the fMRI-data of both regions.

IV. REPLICATION STUDY

To get a feeling on the challenge, a number of networks which scored well at the Brain-Score leaderboard³ were tested on the challenge dataset. This data and metrics are not completely the same, so the results can not 1-to-1 be translated. The tested convolutional neural networks are AlexNet, VGG, ResNet18, ResNet-34, ResNet-50, ResNet-101, ResNet-152, DenseNet121, DenseNet161, DenseNet169, DenseNet201, CorNet-S, and CorNet-Z. AlexNet is used as baseline. All networks were already pre-trained on the ImageNet dataset.

Each network was tested on the provided 92 and 118 datasets. The best layer of a network is the layer that scored average the best on both training datasets. The best layers of the pre-trained networks were submitted to the Algonauts challenge. In figure 4 the score of the submitted models at the Algonauts challenge is visualized, more details can be found

³<http://www.brain-score.org/>

in [12]. The score is the average noise normalized squared correlation score of the EVC and IT region for each network. The average highest scoring network is the DenseNet201, with second place ResNet-18 and third ResNet-34. After this top 3, only ResNet-101, DenseNet121 and DenseNet161 score higher than the challenge baseline.

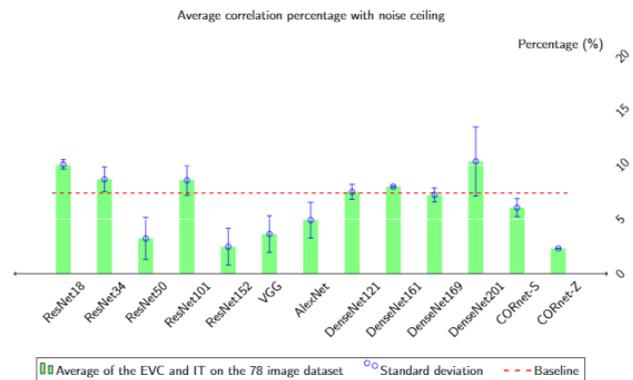


Fig. 4. The noise normalized squared Spearman correlation percentage of the Algonauts test set for different CNN's, along with the standard deviation.

Based on the good performance of the 'shallow' ResNet-18 and ResNet-34, we decided to design an own 'shallow' network (ResNet-20) and train this for the competition.

V. OUR METHOD

For this study a variation of a ResNet with 20 layers is designed. Because our hypothesis is that the earliest layers are important, this is in principal the same architecture as ResNet-18, but the first ResNet block occurring 3 times (instead of 2). This block has the same number of layers as the ResNet blocks in the design of ResNet-18 and ResNet-34. The architecture of the ResNet-20 is visualized in figure 3. As can be seen, the main building blocks are so-called ResBlocks. These ResBlocks contain two convolution layers, each with a kernel of 3x3. Each of these ResBlocks can be skipped to bring the feedback signal fast back to the earlier layers.. This is only possible because the spatial size of the input of every ResBlock does not differ from the output of the block. The ResNets includes a max pooling, an average pooling, a fully connected, and a softmax layer.

This ResNet-20 is trained on a GPU containing the AMD Radeon RX VEGA 64 Chip. The network is trained for a maximum of 120 epochs, although the first and every 5 epochs a model is saved. This ensures that interim models

can be tested. Similar to the testing phase, the images used while training are normalized and re-sized. The optimizer starts with a learning rate of 0.1, a momentum of 0.9, and a weight decay of 0.0001. The learning rate affects the speed at which a network trains [14]. A learning rate of 0.1 indicates that each time the weights are updated, they are updates with 10 percent of the estimated weight error. Momentum is a method used for pushing the gradient in the right direction [14]. During training an optimization function can get stuck in a local minimum, while the algorithm thinks the global minimum is reached. The momentum tries to avoid this problem by increasing the steps taken towards the minimum. Within a local minimum the increase of steps could cause the function to jump out of a local minimum. The value of the momentum indicates with what size the steps increase. The weight decay is the value with which the weights are multiplied after each update [14], preventing the weights from growing too large.

VI. RESULTS

At the time of writing the maximum of 120 epochs is not reached, models were saved after training for 1 epoch, 5 epochs and 10 epochs. All the models produce results above the baseline. In all cases the fully connected layer (FC), the last layer before the SOFTMAX, gave the best prediction.

The response of the FC layers were all submitted to the Algonauts challenge. Figure 5 shows these results, along with the standard deviation. The model trained with 10 epochs preforms the best, which looks promising for further training.

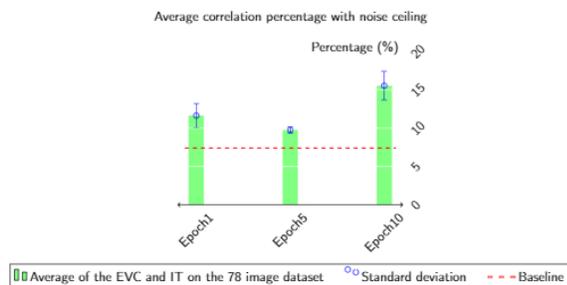


Fig. 5. The average noise normalized squared Spearman correlation percentage of the EVC and IT of the test set for different trained ResNet-20 models, along with the standard deviation

The result is not (yet) good enough to outperform other participants, but good enough for a top-10 ranking at the Algonauts leaderboard⁴.

VII. DISCUSSION & FUTURE WORK

This is a promising result, with a 'shallow' network as ResNet-20 outperforming deeper, more complex networks as DenseNet201 and ResNet-101. The Algonaut score of ResNet-20 is 15.53%, which should be compared with 10.31% of the best performing architecture (DenseNet201) in the replication scenario. Without further experiments it is difficult to conclude if this should be contributed to the

'shallow' design with extra convolutional layer in the earliest ResBlock, or to the training. Yet, DenseNet201 is already outperformed after the first epoch of training, so we should not overestimate the training contribution.

VIII. CONCLUSION

In this paper we demonstrate the asset of using a shallow residual neural network with 20 layers. The benefit of this approach is that earlier stages of the network can be accurately trained, which allows us to add more layers at the earlier stage. With this additional layer the prediction of the visual brain activity improves to 15.53% (last fully connected layer) outperforming deep neural networks with more than 200 layers.

REFERENCES

- [1] D. L. Yamins and J. J. DiCarlo, "Using goal-driven deep learning models to understand sensory cortex," *Nature neuroscience*, vol. 19, no. 3, p. 356, 2016.
- [2] D. H. Hubel and T. N. Wiesel, "Receptive fields of single neurones in the cat's striate cortex," *The Journal of physiology*, vol. 148, no. 3, pp. 574–591, 1959.
- [3] D. L. Yamins, H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, and J. J. DiCarlo, "Performance-optimized hierarchical models predict neural responses in higher visual cortex," *Proceedings of the National Academy of Sciences*, vol. 111, no. 23, pp. 8619–8624, 2014.
- [4] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85 – 117, 2015. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0893608014002135>
- [5] M. Schrimpf, J. Kubilius, H. Hong, N. J. Majaj, R. Rajalingham, E. B. Issa, K. Kar, P. Bashivan, J. Prescott-Roy, K. Schmidt, D. L. K. Yamins, and J. J. DiCarlo, "Brain-score: Which artificial neural network for object recognition is most brain-like?" 2018. [Online]. Available: <https://www.biorxiv.org/content/early/2018/09/05/407007>
- [6] R. M. Cichy, G. Roig, A. Andonian, K. Dwivedi, B. Lahner, A. Lascelles, Y. Mohsenzadeh, K. Ramakrishnan, and A. Oliva, "The Algonauts Project: A Platform for Communication between the Sciences of Biological and Artificial Intelligence," *arXiv e-prints*, p. arXiv:1905.05675, May 2019.
- [7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.
- [8] N. Kriegeskorte and R. A. Kievit, "Representational geometry: integrating cognition, computation, and the brain," *Trends in cognitive sciences*, vol. 17, no. 8, pp. 401–412, 2013.
- [9] J. Hauke and T. Kossowski, "Comparison of values of pearson's and spearman's correlation coefficients on the same sets of data," *Quaestiones geographicae*, vol. 30, no. 2, pp. 87–93, 2011.
- [10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105. [Online]. Available: <http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf>
- [11] J. Kubilius, M. Schrimpf, A. Nayebi, D. Bear, D. L. K. Yamins, and J. J. DiCarlo, "Cornet: Modeling the neural mechanisms of core object recognition," *bioRxiv*, 2018. [Online]. Available: <https://www.biorxiv.org/content/early/2018/09/04/408385>
- [12] A.-R. Meijer, "Explaining the human visual brain using a deep neural network," Bachelor thesis, Universiteit van Amsterdam, June 2019.
- [13] K. Kar, J. Kubilius, K. Schmidt, E. B. Issa, and J. J. DiCarlo, "Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior," *Nature neuroscience*, vol. 22, no. 6, p. 974, 2019.
- [14] Stanford-University, "Cs231n: Convolutional neural networks for visual recognition," 2019, [Online; accessed 11-june-2019]. [Online]. Available: <http://cs231n.github.io>

⁴<http://algonauts.csail.mit.edu/challenge.html>