A residual neural-network model to predict visual cortex measurements

Anne-Ruth José Meijer and 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 residual neural network of only 20 layers for this task. The advantage of this limited number is that earlier stages of the network can be more easily 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% to 15.53%.

1 Introduction

Current state-of-the-art convolutional neural networks (CNNs) incorporate operations like maximum-based pooling of inputs [18], which were directly inspired by single-cell recordings from the V1 region of the mammalian visual cortex [8]. This inspiration is also beneficial for object recognition, because there is a correlation [19] observed between a CNNs ImageNet's performance [15] and its Brain-Score [16]. However, this correlation can no longer be found for the most advanced CNNs [19]. 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 [3].

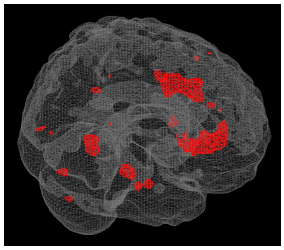

Fig. 1. A visualization of the prediction of the UvABrain team on the activity of the human brain for the Algonauts project: Explaining the Human Visual Brain challenge.

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. Earlier studies [16] showed that CNNs with up to 169 layers (i.e. DenseNet169) showed the best performance in predicting the activity in the brain. Yet, a number of smaller CNNs performed quite competitive, which is the rational for the UvABrain team to design a network with 20 layers. Our study seems to confirm the observation 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.

The novelty of this modification (a "shallow" residual neural-network with one additional earlier layer) is modest, but for the 2019 challenge this is best way forward. It should be clear from our replication study (Section 4), related work (Section 3) and the work of Yamin *et al* [19] that it is hard to see in advance which designs could lead to further advances on the Brain-Score. Advances on this Algonauts project should be made in modest steps with conscientious experimental work.

2 Explaining the Human Visual Brain Challenge

The goal of the 2019 challenge is to 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 [5]. The brain recordings are respectively functional Magnetic Resonance Imaging (fMRI) for Track 1 and Magnetoencephalography (MEG) for Track 2 (see Fig. 2 and Cichy *et al* [2]). 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.

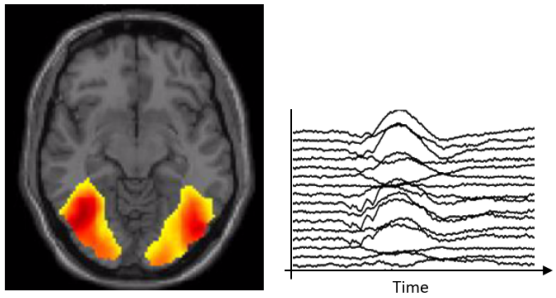

Fig. 2. Brain activity recorded with respectively the fMRI and MEG technique (Courtesy Algonauts project[2]).

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

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). This challenge is unique in the sense that not the final output of the CNNs (the classification of the image) of importance, but that the weights of the intermediate layers are interpreted as activation signals. 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) [10]. The recorded and predicted RDM are than compared based on the Spearman correlation, which is defined as the Pearson correlation between rank variables [6]. 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 [11]. The overall evaluation procedure is illustrated in Fig. 3.

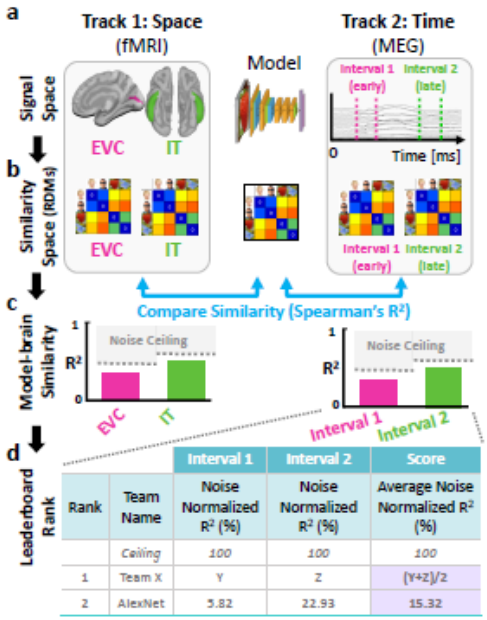

Fig. 3. The evaluation procedure of the 2019 Challenge (Courtesy [3]).

3 Related Work

The challenge was inspired by the initiative to find a Brain-Score [16], 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-liked 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 [12], 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 [14]), nor any other of the competitors.

Winners of the competition were Janik [9], Agrawal [1], and Lage-Castellanos and De Martino [13]. Although not the final winner, Romuald Janik very extensively studied the effects of different feature selection approaches on the performance boost on both the fMRI-data and the MEG-data [9]. According to Janik, the key difficulty of this challenge is the relative small number of training images in comparison with the enormous number of parameters of advanced CNNs, which the risk of overfitting. So, Janik concentrated on a number of relatively simple modifications on well-known CNNs, instead of more advanced approaches like a Siamese network.

The use of a Siamese network was precisely the key to success for Aakash Agrawal on the MEG-data [1]. Agrawal used the baseline AlexNet [11] for the early RDMs and a pretrained VGG-16 [17] for the late RDMs, where both networks were fine-tuned by minimizing the Pearson correlation distance between the predicted and observed RDMs. Because it is known that the early stage is sensitive to small features like texture and orientation [4], AlexNet was parameterized to use the largest filter size in the first layer. AlexNet was less successful for the later RDMs, where the representations of the IT areas of the cortex are known to be more categorical, and VGG-16 was able to predict the late RDMs of MEG-data well (when trained for at least 13 epochs).

Augustin Lage-Castellanos *et al.* used the small and the categorical aspect by extending the training sets. To perform well in the early stage, they extracted edges and applied Gaussian smoothing to create perceptual RDMs. To perform well in the second stage, they manually labelled the training dataset and learn to distinguish eight categories (e.g. human-faces, monkey-faces and animal-faces) [13]. They achieved 90.22% prediction accuracy by training a Gaussian Naïve Bayes classifier after the fully connected layer (1000 features) of a pretrained VGG network. They created a "categorical" 8x8 RDM by averaging the RDM values belonging to images of the same category, as illustrated in Fig. 4. At the left side of Fig. 4 you see a typical RDM, with on each column / row the activation of one of the 92 images. The images are already sorted by category, which is visible in the vague checkerboard pattern. The activation of the ± 10 images belonging to the same categories are then averaged by a *mean* operation. The result is displayed at the right side of Fig. 4. Note that the inferior temporal cortex is by the authors Lage-Castellanos *et al*[13] abbreviated to ITC.

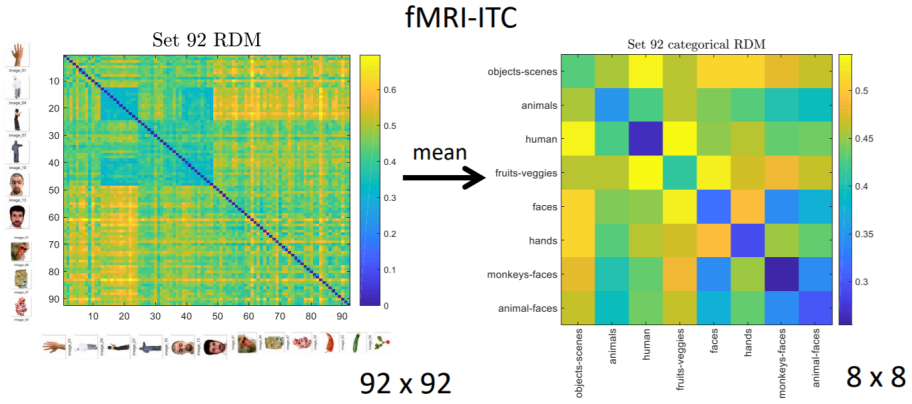

Fig. 4. Creating a categorical RDM from the training data (Courtesy Agustin Lage-Castellanos and Federico De Martino[4]).

Both the perceptual RDM and categorical RDM were mixed with predictions of AlexNet, VGG and ResNet [7]. Of those CNNs ResNet showed the best performance (mixed with a 60% contribution of the combined perceptual and categorical RDM prediction). With this method Augustin Lage-Castellanos *et al.* was able to predict the fMRI-data the best of all competitors.

4 Replication study

To get a feeling on the challenge, a number of networks which scored well at the Brain-Score leaderboard[5] 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 Fig. 5 the score of the submitted models at the Algonauts challenge is visualized, more details can be found in [14]. 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.

[4] 'RDM mixtures for predicting visual cortices responses', presentation at Explaining the Human Visual Brain Workshop, MIT, Cambridge MA, July 20, 2019

[5] http://www.brain-score.org/

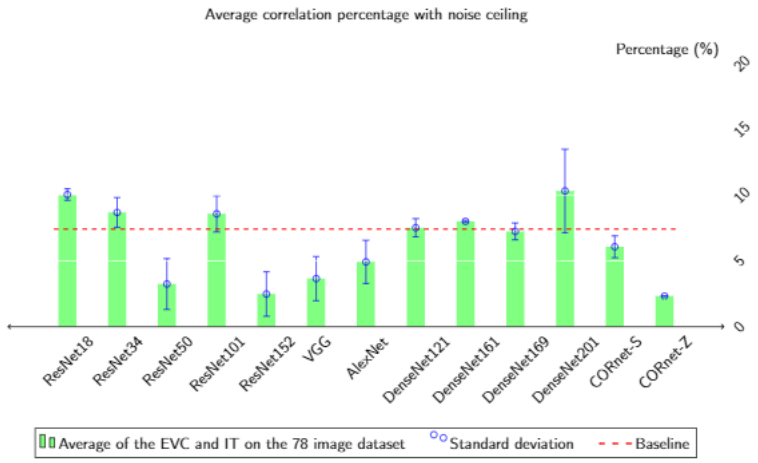

Fig. 5. The noise normalized squared Spearman correlation percentage of the Algonauts test set for different CNN's, along with the standard deviation.

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.

5 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 principle 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 Fig. 6. 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.

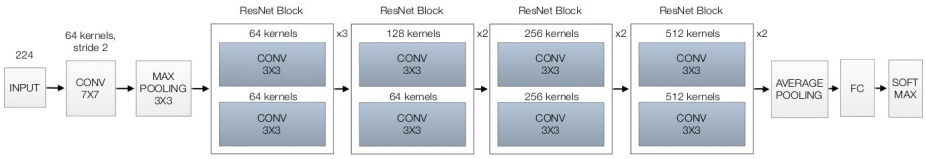

Fig. 6. ResNet20 architecture of the UvABrain team.

This ResNet-20 is trained on a GPU containing the AMD Radeon RX VEGA 64 Chip. The network is trained for 10 epochs, with a model saved the first and every 5 epochs. This ensures that interim models can be tested. Similar to the testing phase, the images used while training are normalized and resized. The optimizer starts with a learning rate of 0.1, a momentum of 0.9, and a weight decay of 0.0001. This values are textbook suggestions[6] and not extensively validated in a systematic parameter search. For a nice example of such a systematic search for this challenge, check the work of Janik [9].

6 Results

The model was trained, with intermediate models saved at 1 epoch, 5 epochs and 10 epochs. Training was considered, when more resources would have been available. Yet, the result is already promising. 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.

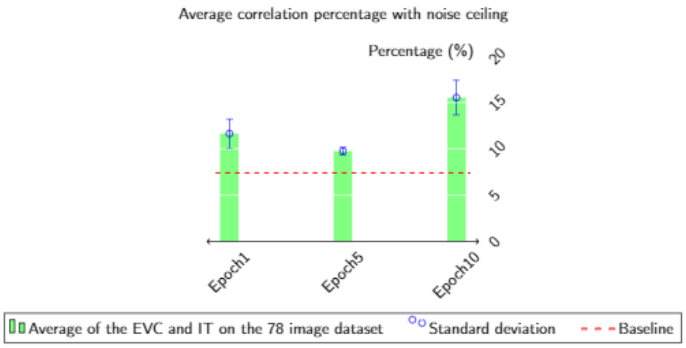

Fig. 7. The average noise normalized squared Spearman correlation percentage of the EVC and IT of the test set for different trained ResNet-20 models.

The responses of the FC layers were all submitted to the Algonauts challenge[7]. Fig. 7 shows these results, along with the standard deviation. The model trained with 10 epochs performs the best, which looks promising for further training.

The Algonaut score of our ResNet-20 model is 15.53%, which is not (yet) good enough to outperform other participants, but it is a clear improvement from the baseline and an improvement to Algonaut score of ResNet-18 of 10.04% (the architecture where our model is a modification from). The result of the UvABrain team was also good enough for a top-10 ranking at the Algonauts leaderboard.

[6] http://cs231n.github.io
[7] http://algonauts.csail.mit.edu/challenge.html

7 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.

Our choice for ResNet for the fMRI dataset is confirmed by Augustin Lage-Castellanos *et al.*, although they received a boost of at least 20.99 percentage point from their perceptual and categorical method. Aakash Agrawal selected as model VGG-16, which was trained for 160 epochs[8]. Training over 160 epochs improved the performance with a factor two from $\sim 14\%$ to $\sim 26\%$, which gives an indication of the performance gain which could be made with our ResNet-20 which was trained for more than 10 epochs. Also Janik experimented had good initial results with ResNet-18 for the fMRI EVC-dataset, but got a final boost by combining those features with the best features from maxpool2 of VGG-19. Yet, the biggest performance boost (18 percentage point) was made by simply dropping the corners[9]. For the fMRI IT-dataset a ResNet-50 was used, which was combined with a Multidimensional Scaling to create an embedding of 300 features. The success of this approach that on this level structure can be found, as also demonstrated by the combined perceptual/categorical approach of Augustin Lage-Castellanos *et al.*

The value of this ResNet-20 can be demonstrated in future work, when in a replication study the simple modifications which gave large performance boosts found by the other competitors [9,1,13] will be applied to our architecture. This will also give us time to see the effect of training our network for more than 160 epochs.

8 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 the baseline and several other neural networks designs with more complexity and layers.

[8] 'Dissimilarity learning via Siamese network predicts brain imaging data', presentation at Explaining the Human Visual Brain Workshop, MIT, Cambridge MA, July 20, 2019

[9] 'RDM peculiarities, effective receptive fields and surrogate features', presentation at Explaining the Human Visual Brain Workshop, MIT, Cambridge MA, July 20, 2019

References

1. Agrawal, A.: Dissimilarity learning via Siamese network predicts brain imaging data. arXiv:1907.02591 (Jul 2019)
2. Cichy, R.M., Pantazis, D., Oliva, A.: Similarity-Based Fusion of MEG and fMRI Reveals Spatio-Temporal Dynamics in Human Cortex During Visual Object Recognition. Cerebral Cortex **26**(8), 3563–3579 (07 2016)
3. Cichy, R.M., Roig, G., Andonian, A., Dwivedi, K., Lahner, B., Lascelles, A., Mohsenzadeh, Y., Ramakrishnan, K., Oliva, A.: The Algonauts Project: A Platform for Communication between the Sciences of Biological and Artificial Intelligence. arXiv:1905.05675 (May 2019)
4. De Valois, R.L., De Valois, K.K.: Spatial vision. Annual review of psychology **31**(1), 309–341 (1980)
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: IEEE conference on computer vision and pattern recognition. pp. 248–255. IEEE Computer Society (June 2009)
6. Hauke, J., Kossowski, T.: Comparison of values of pearson's and spearman's correlation coefficients on the same sets of data. Quaestiones geographicae **30**(2), 87–93 (2011)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. arXiv:1512.03385 (2015), http://arxiv.org/abs/1512.03385
8. Hubel, D.H., Wiesel, T.N.: Receptive fields of single neurones in the cat's striate cortex. The Journal of physiology **148**(3), 574–591 (1959)
9. Janik, R.A.: Explaining the Human Visual Brain Challenge 2019 – receptive fields and surrogate features. arXiv:1907.00950 (Jul 2019)
10. Kriegeskorte, N., Kievit, R.A.: Representational geometry: integrating cognition, computation, and the brain. Trends in cognitive sciences **17**(8), 401–412 (2013)
11. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, vol. 25, pp. 1097–1105. Curran Associates, Inc. (2012)
12. Kubilius, J., Schrimpf, M., Nayebi, A., Bear, D., Yamins, D.L.K., DiCarlo, J.J.: Cornet: Modeling the neural mechanisms of core object recognition. bioRxiv:408385 (September 2018)
13. Lage-Castellanos, A., De Martino, F.: Predicting stimulus representations in the visual cortex using computational principles. bioRxiv:687731 (July 2019)
14. Meijer, A.R.: Explaining the human visual brain using a deep neural network. Bachelor thesis, Universiteit van Amsterdam (June 2019)
15. Schmidhuber, J.: Deep learning in neural networks: An overview. Neural Networks **61**, 85 – 117 (2015)
16. Schrimpf, M., Kubilius, J., Hong, H., Majaj, N.J., Rajalingham, R., Issa, E.B., Kar, K., Bashivan, P., Prescott-Roy, J., Schmidt, K., Yamins, D.L.K., DiCarlo, J.J.: Brain-score: Which artificial neural network for object recognition is most brain-like? bioRxiv:407007 (September 2018)
17. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 (September 2014)
18. Yamins, D.L., DiCarlo, J.J.: Using goal-driven deep learning models to understand sensory cortex. Nature neuroscience **19**(3), 356 (2016)
19. Yamins, D.L., Hong, H., Cadieu, C.F., Solomon, E.A., Seibert, D., DiCarlo, J.J.: Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proceedings of the National Academy of Sciences **111**(23), 8619–8624 (2014)