Ablation Studies in Artificial Neural Networks

Richard Meyes, Melanie Lu, Constantin Waubert, Tobias Meisen
Institute of Information Management in Mechanical Engineering,
RWTH Aachen University, Dennewartstr. 27, 52064 Aachen, Germany
{richard.meyes, melanie.lu, constantin.waubert,
tobias.meisen}@ima-ifu.rwth-aachen.de

Abstract

1 Introduction

Recent research on deep learning (DL) has brought fourth a number of remarkable applications for different problems in a variety of domains, such as visual object recognition, object detection and semantic segmentation in field of computer vision (CV) [1–5], speech recognition and speech separation in the field of natural language processing (NLP) [6–10] or self-learning agents based on deep reinforcement learning (DRL) for video games [11–14], classic board games [15–17] as well as locomotion and robotic control [18–23]. During the last few years of research on DL, the strong increase in availability of computational power combined with the facilitation of new computing paradigms such as GPU programming [1] and asynchronous methods for training deep neural networks (DNNs) [24, 25] resulted in an increase of the average size, i.e. the number of trainable weights, of state of the art DNNs. Despite the growth in size and complexity of those DNNs, the main research focus was placed on increasing the performance and speed of those networks solving specific benchmark tasks rather than on the development of new methods and perspectives to understand how knowledge, that is acquired during training, is represented in these networks. Considering that the research on DNNs has been confronted only recently with larger networks (networks exhibiting some kind of holistic behaviour that is not trivially explained by just considering the functional mechanism of key components such as single units, their activation functions, regularization mechanisms and so on), methods and perspectives from the field of neuroscience, a research field, which dealt with large and complex neural systems for decades, may prove useful to investigate the structure of knowledge representation in state of the art DNNs.

In this paper, we follow a neuroscience inspired approach to analyze the structure of represented knowledge within DNNs. Our approach is inspired by the principle of ablation studies, which are based on carefully damaging neural tissue in a controlled manner while investigating how the inflicted damage influences the brain's capabilities to perform a specific task. This way, insights about the functional role of the damaged brain regions can be gained as well as insights about how the processing of external stimuli is structured, organized and mapped in the brain. One of the most prominent examples for such an organized mapping is the cortical humonculus found in the primary motor cortex and the primary sensory cortex of primates and humans. It is a distorted representation of the human body mapped onto specific regions of the neocortex responsible for processing motor functions or sensory functions for different parts of the body. In the past, ablation studies were used to uncover structure and organization in other parts of the brain. For instance, neonatal cochlear ablations in cats revealed that binaural interactions, i.e. the perception of sound via intensity differences arriving at the two ears, are exhibited early in postnatal life, well before structural maturation of the auditory pathways from the ear to the cortex is complete [26]. In another exemplary study, the ablation of subplate neurons in the visual cortex of adult cats revealed their role for the functional development of ocular dominance [27]. Considering that ablation studies proved to be be a valuable method to investigate large, complex neural systems, like the brain of vertebrates

and primates, it seems reasonable to investigate their potential for tackling state-of-the-art artificial neural systems.

In our work, we aim to transfer the principle of ablation studies to artificial neural networks (ANNs) to open up a new perspective to understand knowledge representation in these ANNs. We investigated correlates between the spatial as well as structural characteristics of single units within a small shallow multi layer perceptron (MLP), i.e. their location within the network and the distribution of their weights, and their contribution to the overall accuracy as well as the class specific accuracy of the network by means of single unit ablations. We found, that some single units are universally important for the classification task, while other single units are only selectively important for the classification of a specific class. Furthermore, we found that the importance of a single unit for the classification task correlates with the extent to which the distribution of weights of incoming connections of that single unit after training differs from the initial random weight distribution. We further investigated the robustness of the network's classification capabilities by looking for redundant knowledge representations of specific classes in different areas of the network by pairwise ablations of single units. We found that the pairwise ablation of single units has a stronger effect on the network's classification accuracy than the summed effects of single ablations of the same single units. Second, we investigated a larger state-of-the-art CNN for correlates between the size as well as the depth resolved spatial characteristics of the ablated portions of the network and the overall accuracy as well as the class specific accuracy of the network by means of ablating groups of filters of the convolutional layers in different depths of the network, aiming to examine the network for a similar hierarchical organization as it is found in the primary visual cortex [28, 29].

We found that, in general, the larger the ablated network portion, the stronger the effect on the network's classification accuracy, however, this effect greatly varies across different depths of the network. We further found that some layers are universally more important for the classification task than other layers, however, this effect shows some variation across specific classes. We further investigated the possibilities to repair the inflicted damage by further training the damaged network, aiming to recover the networks original classification accuracy. We found that most of the negative effect of ablations on the network's classification accuracy could be recovered within a single episode of recovery training, even in cases of severe structural damage (up to 80% of ablated filters within a single convolutional layer).

Interestingly, for both networks, we found that ablations, despite having a general negative effect on the universal classification accuracy of the networks, consistently show positive effects on the classification accuracy for specific classes. After an ablation, the network's classification accuracy for specific classes increased rather than decreased, giving raise to the notion that a trained network's structure may be purposefully manipulated to increase its classification capabilities beyond the local optimum that was reached during training by means of fitting the network's weights via back-propagation.

Conclusion and Outlook?

2 Related Work

The principle of an ablation, i.e. removing trainable parameters from a trained DNN, is an idea also followed when pruning networks in order to reduce their size and computational cost, thus speeding up training and inference, while retaining as much of their original performance as possible. The idea is that some parameters of a trained network contribute very little or not at all to the output of the network and are therefore negligible and can be removed [30]. Recent research on pruning state of the art convolutional neural networks (CNNs) like the VGG-16 oder the ResNet-110 focused on the optimization of a network's structure by removing kernels and entire filters [31, 32] and methods to find an appropriate ranking of units to tackle the simple but challenging combinatorial optimization problem of how to chose what combination of units to be removed for best results [33–35]. While pruning is mainly conducted as a measure to fine-tune a network's architecture, we aim to utilize the approach of ablations not merely to optimize the size and the speed of DNNs, but to gain insight about how the represented knowledge is structured and organized within the network, offering transparency and interpretability of the network's behaviour. This objective is closely related to the question of how a network reaches its decisions and what are the most important underlying factors for this decision making process. Some recent work on this matter demonstrated how to

explain the contribution of a network's input elements to its decision by means of Deep Taylor Decomposition [36] or Gradient-weighted Class Activation Mapping (Grad-CAM) [37, 38]. Another recent example focusing on the processes within a network rather than on the input showed how latent representations within CNNs are stored in individual hidden units that align with a set of humanly interpretable semantic concepts [39]. One of the most recent neuroscience inspired contributions utilizing ablations demonstrated that a network's capability to generalize a classification task is related to its reliance on class selective single units within the network. Specifically, networks that generalize well contain less class selective units than networks that merely memorize the data set presented during training [40].

3 Methods

In this study, we investigated two neural network architectures trained on different data sets.

First, we trained a small and shallow MLP to recognize hand written digits using the MNIST data set [41]. The network's input layer is comprised of 784 units corresponding to the 28x28 pixels of the input images. The network has two hidden layers with 20 and 10 hidden units respectively, whereas ReLU activation was chosen for all hidden units. The network's output layer contains 10 units, corresponding to the ten classes of the data set, with softmax activation. The network was trained for 100 epochs on 60.000 images of the training set and reached an accuracy of 94.64% validated on 10.000 images of the test set. After training, ablations of single units were performed by manually setting the weights of all incoming connections to zero, essentially preventing any kind of information flow through this unit. As we trained the network without biases, zeroing a unit's incoming weights is equivalent to removing this unit from the network alltogether. In order to investigate the effect of the ablation, we evaluated the network's performance on the test set and compared its accuracy with the original accuracy of the undamaged network. We used t-SNE on the complete 10.000 images of the test set to visualize the effects of the ablations.

Second, we investigated the VGG19 network with batch normalization pre-trained on the ImageNet data set [42] as a representative of today's state-of-the-art CNNs for object recognition tasks. The VGG19 features 19 layers with learnable weights, 16 convolutional and 3 linear layers. Additionally, the network has ReLU, maxpooling and batch normalization inbetween. The advantage this particular network offers is its sufficiently large size for investigating depth dependent effects of the ablations. Details about the data set, the network's architecture and the training process can be found in [2]. The ImageNet dataset used for the study consists of 1000 categories with a total of 1.2 million images in the training set and 50 images per category in the validation set. The study of the VGG19 was made up of two parts. In the first part, differently sized groups of filters were ablated in each of the convolutional layers of the network. For each layer groups of 1%, 5%, 10% and 25% of the number of filters were ablated. The groups were chosen based on a measure of the similarity of the filter weights. The similarities of the filters of each layer was determined based on the absolute Euclidean distance of the normalized filter weights. The ablations were achieved by setting the weights and biases to 0, which is equivalent to setting the activation of a filter to 0. The effect of the ablations was evaluated with the performance of the network on the validaton dataset, more precisely, the classification accuracy was tested. The usual performance metric for the ImageNet dataset is the top-5 error, which measures how often the correct class is among the first five predictions. In this study, the top-5 as well as the top-1 error were calculated after each ablation. Two aspects were regarded for the analysis of the effects, the first one being the amount of the filters ablated and the second one being the depth of the ablation. It should be noted that the number of the ablated filters varied depending on the layer due to the different number of filters in the layers. In the second part of the study, two layers were ablated and subsequently the ablated network was retrained on the ImageNet dataset. The goal was to assess the ability of the network to recover after several ablations.

- pre-trained on ImageNet, num classes? size of train set? size of val set?
- ablation of groups of filters with different group sizes, 1%, 5%, 10%, 25%.
- ablation was achieved by zeroing the weights of the kernel that corresponds to the filters.
- ablation was done only in conv layers in different depths
- effect of ablation was tested evaluating the damaged network's performance on the validation data set of ImageNet

- evaluation of the effect of an ablation on the networks performance tested by first ablating a part of a layer and second test the network with the test set and calculate the loss of accuracy
- two aspects investigated. One: effect of the amount of ablated filters (1, 5 10, 25 percent) Two: effect of where the filters were ablated, i.e. in what layer

4 Results

4.1 Single Unit Ablations in a shallow MLP

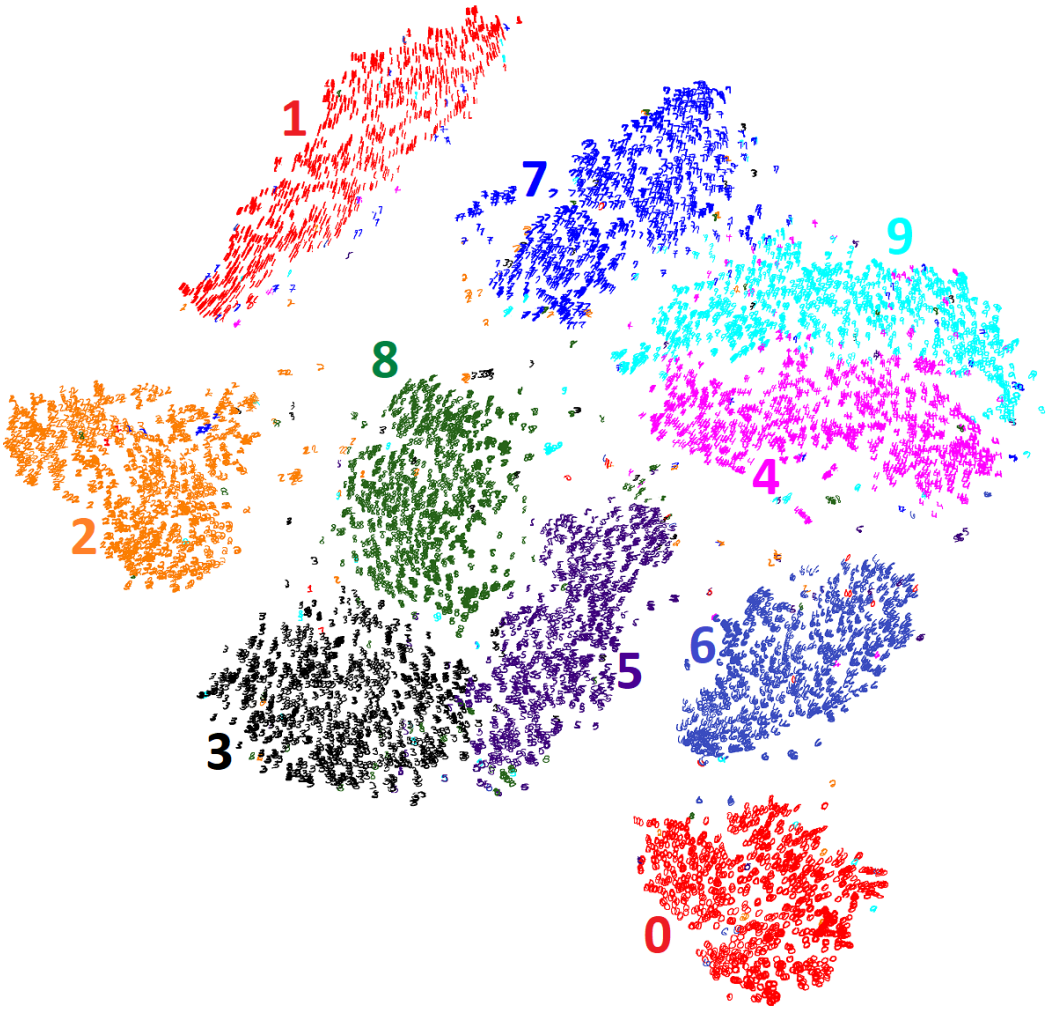
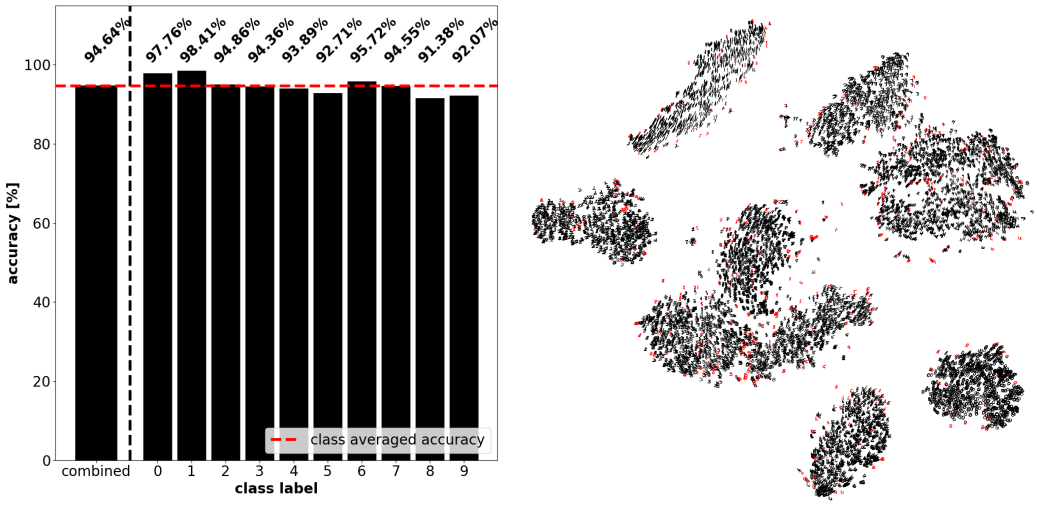


Figure 1: Text

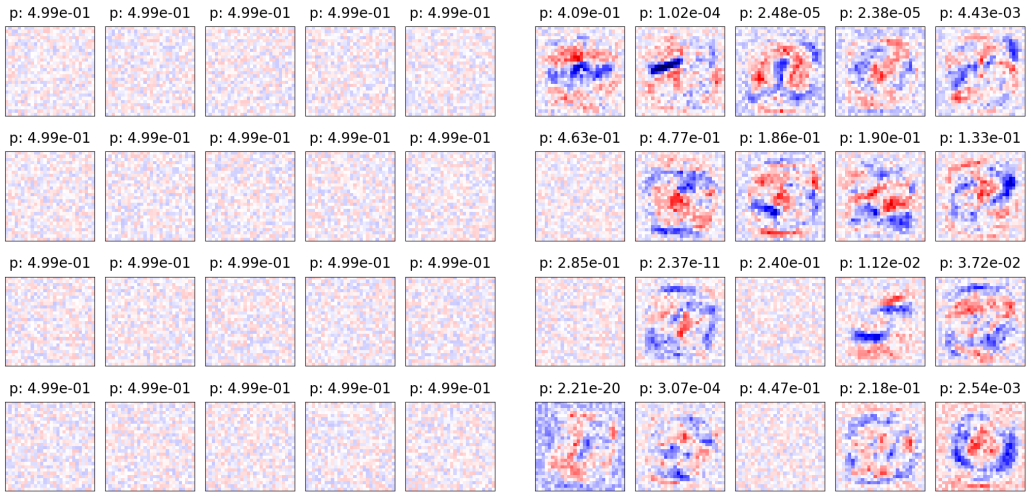
We found that degree of difference of the distribution of the incoming weights of a unit before and after training correlates positively with the unit's contribution to the network's overall accuracy. We further found, that

- characteristics of the unit which determines its importance WITHOUT a functional test?
- similar to estimate whether a banana is gonna taste well without actually eating it
- characteristics such a color, firmness, etc...
- correlates of unit importance to some characteristics
- MNIST - MLP



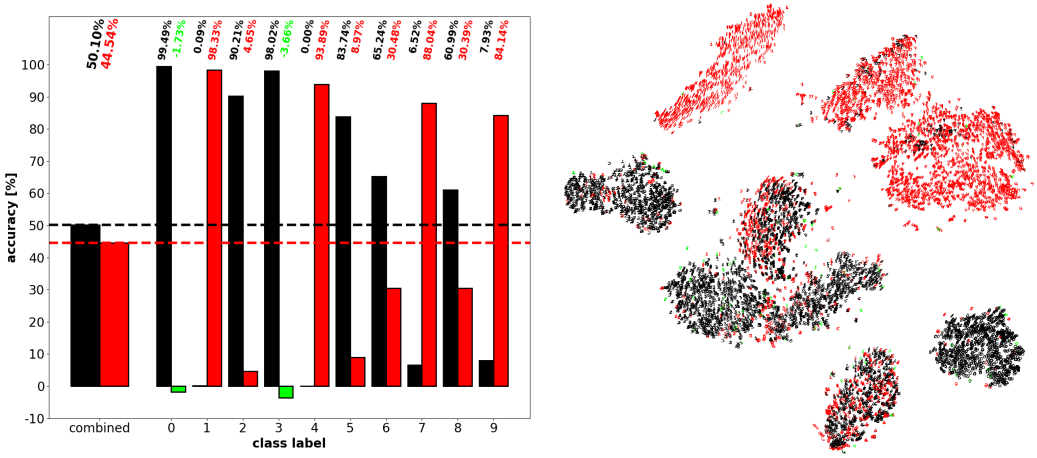
(a) accuracy (b) tSNE

Figure 2: acc and tSNE



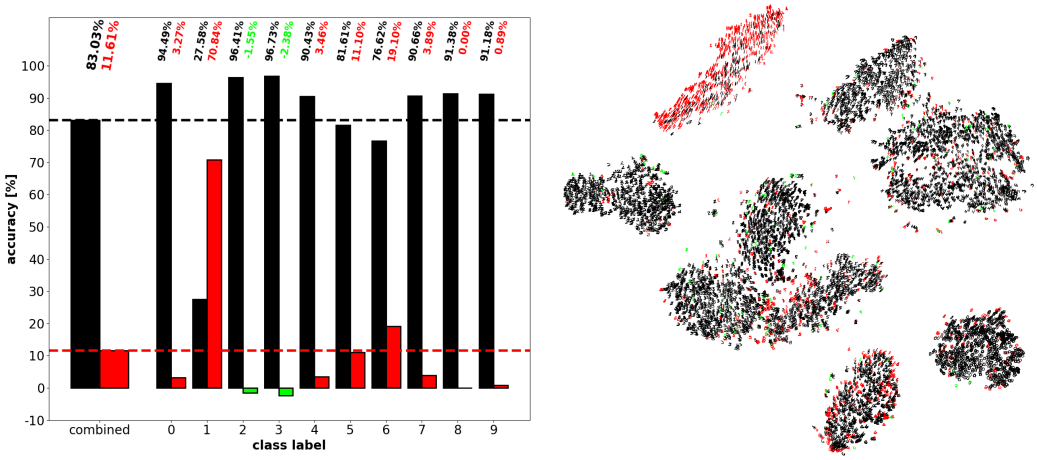
(a) untrained weight distrib (b) ttrained weight distrib

Figure 3: comparision of weight distrib before and after training. 4 untrained units!. Difference quanitified by mann whitney U p value



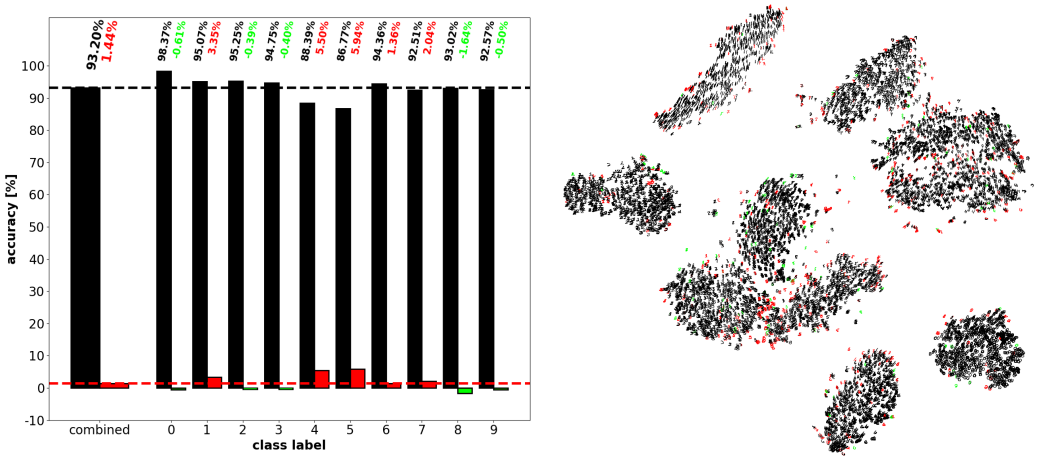
(a) accuracy (b) tSNE

Figure 4: acc and tSNE - Strongest effect, important for many classes



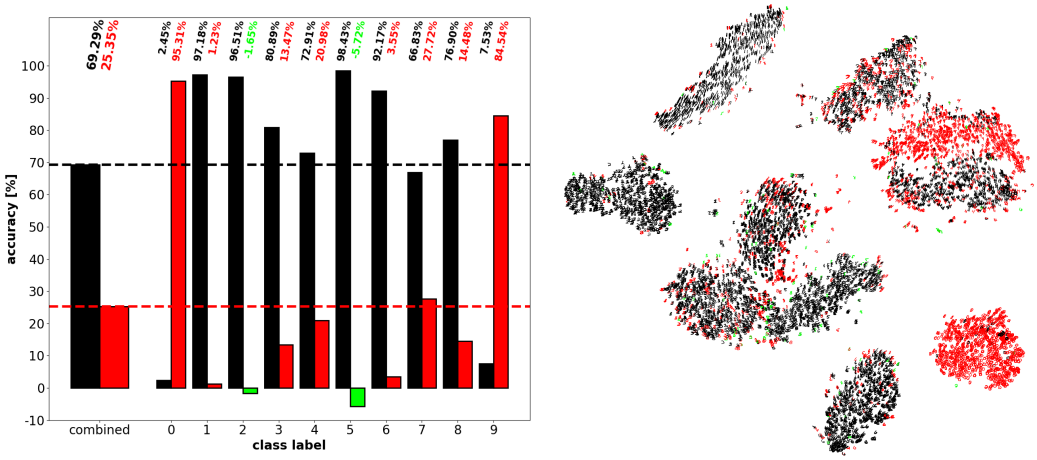
(a) accuracy (b) tSNE

Figure 5: acc and tSNE - very class selective, only 1, a little bit 6 (consistent with previous work on selectivity, [40])



(a) accuracy (b) tSNE

Figure 6: weakest effect! important for nothing. units could be dropped! pruning apporach!

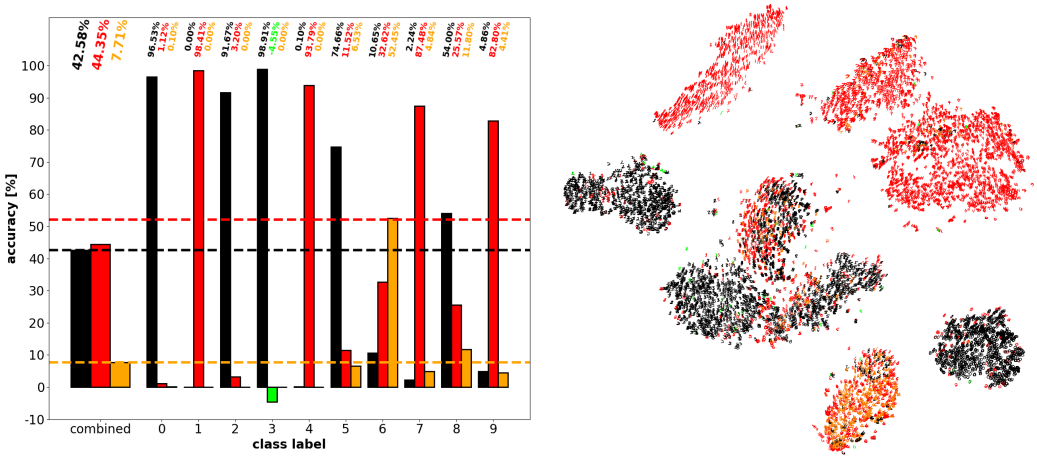


(a) accuracy (b) tSNE

Figure 7: tSNE ZOOM IN showing positive gain of accuracy! CONSTI!

- Ablation in first layer, different kinds of effects on functionality
- Class selective representation
- question: what makes a neuron important?
- answer: weights distribution difference between untrained and trained state? Correlation evidence
- usual effect: negative on performance, however ALSO POSITIVE EFFECTS but smaller magnitude
- question: can the positive effect be achieved without the negative?

4.2 Pairwise Unit Ablations in a shallow MLP

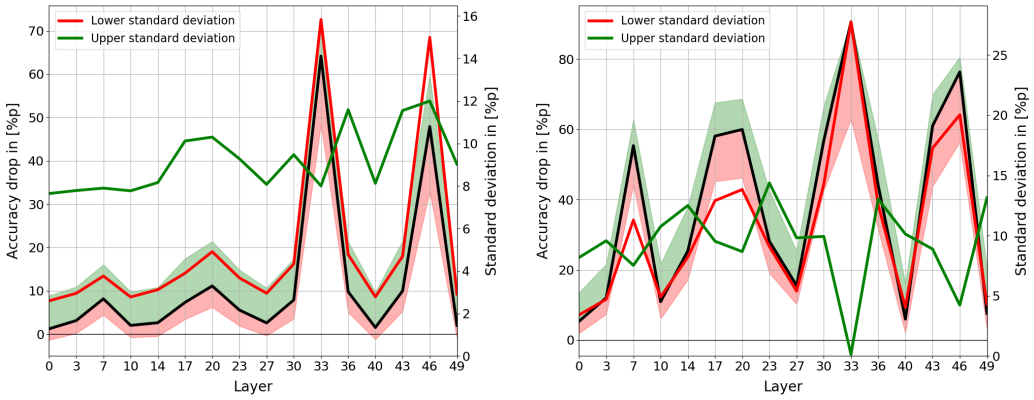


(a) accuracy (b) tSNE

Figure 8: positive effect for class 6 INCREASED!, negative effects also increased beyond the sum of single ablations!

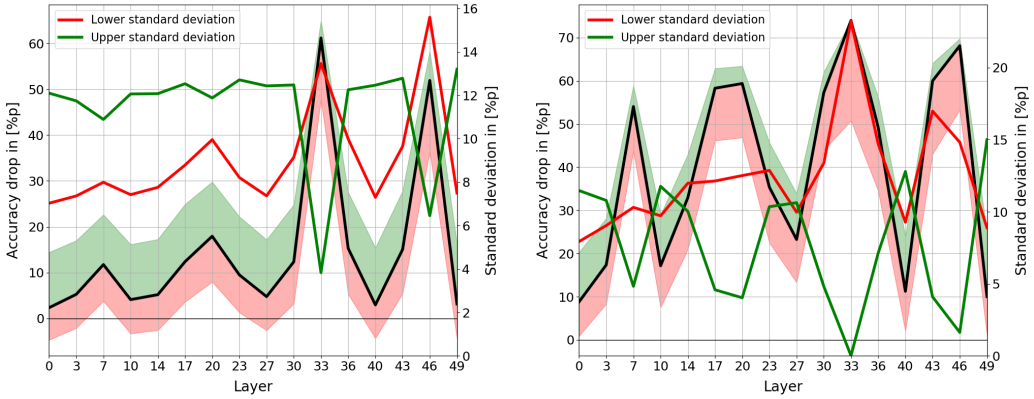
4.3 Grouped Unit Ablations in a Deep CNN

The first finding was as expected, that the higher the amount of ablated filters, the more severe the effect of the ablation on the performance. The second finding was that some layers seem to be more



(a) after 10% ablation (b) after 25% ablation

Figure 9: top-5 accuracy drop



(a) after 10% ablation (b) after 25% ablation

Figure 10: top-1 accuracy drop

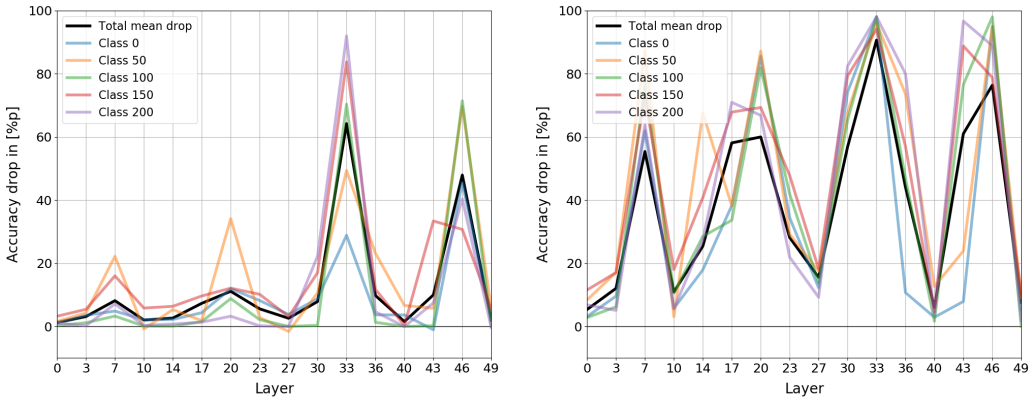
important than other layers for the network's performance, with regard to the VGG19 layer 33 and 46 showed particular importance. 9 and 10 show the average top-5 and top-1 accuracy drop for ablations ratios of 10% and 25% over all convolutional layers. Additionally, the standard deviation is plotted for the values above and below the mean.

Following these observations, the next question was if the drop in accuracy was the same for all classes or if the changes were class specific. For this purpose, the accuracy drop per class for different ablations was calculated. ... and ... show that the drop differs greatly dependent on the class. Furthermore, the standard deviation of the accuracy drops varies for different layers.

figure with class specific accuracy drop

We further found that the classes were effected differently depending on the layer. This means that classes effected strongly in one layer were barely effected in other layers, as shown in ....

- first finding is expected: the higher the amount of ablated filters, the larger the effect on the performance
- second finding is, that some layers seem to be more important than other layers for the networks performance, specifically 33 and 46
- Question: are those effects of general nature or are they class specific?
- Looking at the class specific accuracy, we found that some classes are strongly influenced by the ablation, other classes not so much.
- interestingly, the class specific accuracy for some classes IMPROVED after an ablation



(a) after 10% ablation

(b) after 25% ablation

Figure 11: top-5 accuracy drop

- Also, the depth dependence is class specific, i.e. the ablation of filters has a different class specific effect depending on where the ablation was performed
- at last, we wondered whether it was possible to recover the caused damage and retrained the network with frozen weights too speed up computation.
- we found that the damage could be recovered almost completely, even when we ablated nearly 80 percent in the most important layers (33 and 46)

5 Conclusions and Future Work

...

Acknowledgments

...

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[2] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[3] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[5] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 580–587, 2014.

[6] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal processing magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[7] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*, pp. 6645–6649, IEEE, 2013.

[8] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing," *ieee Computational intelligenCe magazine*, vol. 13, no. 3, pp. 55–75, 2018.

[9] J. Chen and D. Wang, "Dnn based mask estimation for supervised speech separation," in *Audio source separation*, pp. 207–235, Springer, 2018.

[10] Y. Luo, Z. Chen, and N. Mesgarani, "Speaker-independent speech separation with deep attractor network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 4, pp. 787–796, 2018.

[11] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," *arXiv preprint arXiv:1312.5602*, 2013.

[12] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, p. 529, 2015.

[13] M. Hessel, J. Modayil, H. Van Hasselt, T. Schaul, G. Ostrovski, W. Dabney, D. Horgan, B. Piot, M. Azar, and D. Silver, "Rainbow: Combining improvements in deep reinforcement learning," *arXiv preprint arXiv:1710.02298*, 2017.

[14] T. Pohlen, B. Piot, T. Hester, M. G. Azar, D. Horgan, D. Budden, G. Barth-Maron, H. van Hasselt, J. Quan, M. Večerík, *et al.*, "Observe and look further: Achieving consistent performance on atari," *arXiv preprint arXiv:1805.11593*, 2018.

[15] G. Tesauro, "Temporal difference learning and td-gammon," *Communications of the ACM*, vol. 38, no. 3, pp. 58–68, 1995.

[16] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, *et al.*, "Mastering the game of go with deep neural networks and tree search," *nature*, vol. 529, no. 7587, pp. 484–489, 2016.

[17] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, *et al.*, "Mastering chess and shogi by self-play with a general reinforcement learning algorithm," *arXiv preprint arXiv:1712.01815*, 2017.

[18] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971*, 2015.

[19] S. Levine, C. Finn, T. Darrell, and P. Abbeel, "End-to-end training of deep visuomotor policies," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 1334–1373, 2016.

[20] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[21] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in *Intelligent Robots and Systems (IROS), 2017 IEEE/RSJ International Conference on*, pp. 23–30, IEEE, 2017.

[22] M. Andrychowicz, F. Wolski, A. Ray, J. Schneider, R. Fong, P. Welinder, B. McGrew, J. Tobin, O. P. Abbeel, and W. Zaremba, "Hindsight experience replay," in *Advances in Neural Information Processing Systems*, pp. 5048–5058, 2017.

[23] M. Andrychowicz, B. Baker, M. Chociej, R. Jozefowicz, B. McGrew, J. Pachocki, A. Petron, M. Plappert, G. Powell, A. Ray, *et al.*, "Learning dexterous in-hand manipulation," *arXiv preprint arXiv:1808.00177*, 2018.

[24] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, "Asynchronous methods for deep reinforcement learning," in *International Conference on Machine Learning*, pp. 1928–1937, 2016.

[25] L. Espeholt, H. Soyer, R. Munos, K. Simonyan, V. Mnih, T. Ward, Y. Doron, V. Firoiu, T. Harley, I. Dunning, *et al.*, "Impala: Scalable distributed deep-rl with importance weighted actor-learner architectures," *arXiv preprint arXiv:1802.01561*, 2018.

[26] R. A. Reale, J. F. Brugge, and J. C. Chan, "Maps of auditory cortex in cats reared after unilateral cochlear ablation in the neonatal period," *Developmental Brain Research*, vol. 34, no. 2, pp. 281–290, 1987.

[27] P. O. Kanold, P. Kara, R. C. Reid, and C. J. Shatz, "Role of subplate neurons in functional maturation of visual cortical columns," *Science*, vol. 301, no. 5632, pp. 521–525, 2003.

[28] D. C. Van Essen and J. H. Maunsell, "Hierarchical organization and functional streams in the visual cortex," *Trends in neurosciences*, vol. 6, pp. 370–375, 1983.

[29] D. J. Felleman and D. E. Van, "Distributed hierarchical processing in the primate cerebral cortex.," *Cerebral cortex (New York, NY: 1991)*, vol. 1, no. 1, pp. 1–47, 1991.

[30] Y. LeCun, J. S. Denker, and S. A. Solla, "Optimal brain damage," in *Advances in neural information processing systems*, pp. 598–605, 1990.

[31] P. Molchanov, S. Tyree, T. Karras, T. Aila, and J. Kautz, "Pruning convolutional neural networks for resource efficient inference," *arXiv preprint arXiv:1611.06440*, 2016.

[32] H. Li, A. Kadav, I. Durdanovic, H. Samet, and H. P. Graf, "Pruning filters for efficient convnets," *arXiv preprint arXiv:1608.08710*, 2016.

[33] S. Anwar, K. Hwang, and W. Sung, "Structured pruning of deep convolutional neural networks," *ACM Journal on Emerging Technologies in Computing Systems (JETC)*, vol. 13, no. 3, p. 32, 2017.

[34] C. Louizos, M. Welling, and D. P. Kingma, "Learning sparse neural networks through l_0 regularization," *arXiv preprint arXiv:1712.01312*, 2017.

[35] L. Theis, I. Korshunova, A. Tejani, and F. Huszár, "Faster gaze prediction with dense networks and fisher pruning," *arXiv preprint arXiv:1801.05787*, 2018.

[36] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K.-R. Müller, "Explaining nonlinear classification decisions with deep taylor decomposition," *Pattern Recognition*, vol. 65, pp. 211–222, 2017.

[37] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, *et al.*, "Grad-cam: Visual explanations from deep networks via gradient-based localization.," in *ICCV*, pp. 618–626, 2017.

[38] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 839–847, IEEE, 2018.

[39] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba, "Network dissection: Quantifying interpretability of deep visual representations," *arXiv preprint arXiv:1704.05796*, 2017.

[40] A. S. Morcos, D. G. Barrett, N. C. Rabinowitz, and M. Botvinick, "On the importance of single directions for generalization," *arXiv preprint arXiv:1803.06959*, 2018.

[41] Y. LeCun and C. Cortes, "MNIST handwritten digit database," 2010.

[42] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR09*, 2009.