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 filters 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 altogether. 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 [42] 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 [43] 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. 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 1,000 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, we ablated filters in each of the convolutional layers of the network. Consecutively, groups of 1%, 5%, 10% and 25% of the total number of filters in each layer were ablated. The groups were chosen based on a measure of similarity of the filter weights. We determined the similarities of the filters of each layer 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 the 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 after each ablation. 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, we calculated the top-5 as well as the top-1 error 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 second part of the study, we ablated two layers and subsequently retained the ablated network on the ImageNet dataset. The goal was to assess the ability of the network to recover after ablations. Based on the results of the ablations, the layers with the highest impact on the classification performance were chosen for the recovery training experiments. The training was split into two different parts. For the first part, the network was retrained for several epochs with randomly chosen set of filters ablated from the two layers multiple times. We set the ablation ratio at 25%. The motivation for this experiment was to test to what extent the network would be able to recover its original functionality with ablated filters. Each pass consisted of an ablation followed by several epochs of recovery training. Between the passes, the network was reset to the original pre-trained VGG19, before the ablations were performed. The weights of the early layers were frozen up to the convolutional layer before the layer, in which the filters were ablated. This way, the adaptation of weights to recover the original functionality was

restricted to those layers that lay behind the damaged layer in order to allow the weights to adapt to the change of information flow that is caused by the ablation. Additionally, the ablated filters were also not retrained, as if the filters did not exist anymore leaving only 75% of the original amount of filters. For the second experiment, we ablated the same network randomly and retrained it for several passes without resetting the ablated filters. The ablation ratio was again set at 25%, although since the 25% were chosen from all filters including the already ablated one, the number of ablated filters increased non-linearly. In this case, the number of epochs was limited by the condition that at least five training epochs had to be completed and after that the improvement in accuracy had to be increased by at least 0.05% over two epochs. The goal of this experiment was to test how well the network could recover with an increasing number of ablated filters and if functional recovery would be prevented entirely by a too large amount of ablated filters. We tested the classification accuracy after each ablation and after each training epoch.

4 Results

4.1 Single Unit Ablations in a shallow MLP

Figure 1 shows a t-SNE visualization of the 10,000 digits in the test set and serves as a basis for the visual evaluation of the effects of ablations. As t-SNE tries to preserve the global and local structure of the data when embedding the original 784-dimensional data set into a 2-dimensional space, it allows to investigate whether this structure is represented in an organized manner in the network. The overall accuracy of the trained MLP evaluated on the test set was 94.64% with a slight variation across the specific classes ranging from 91.38% for the digit eight to 98.41% for the digit one. Figure 2 shows the overall accuracy, its class specific variation and the corresponding t-SNE plot. The black and red digits correspond to the correctly and incorrectly classified input images, respectively.

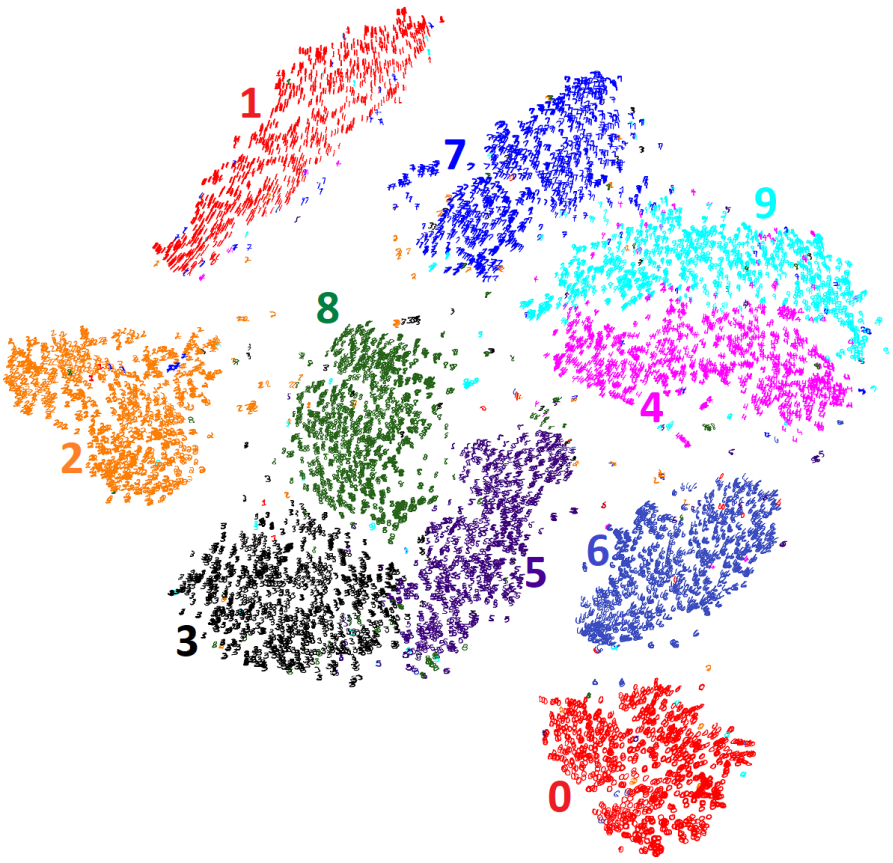


Figure 1: t-SNE visualization of the complete 10,000 digits of the MNIST test set.

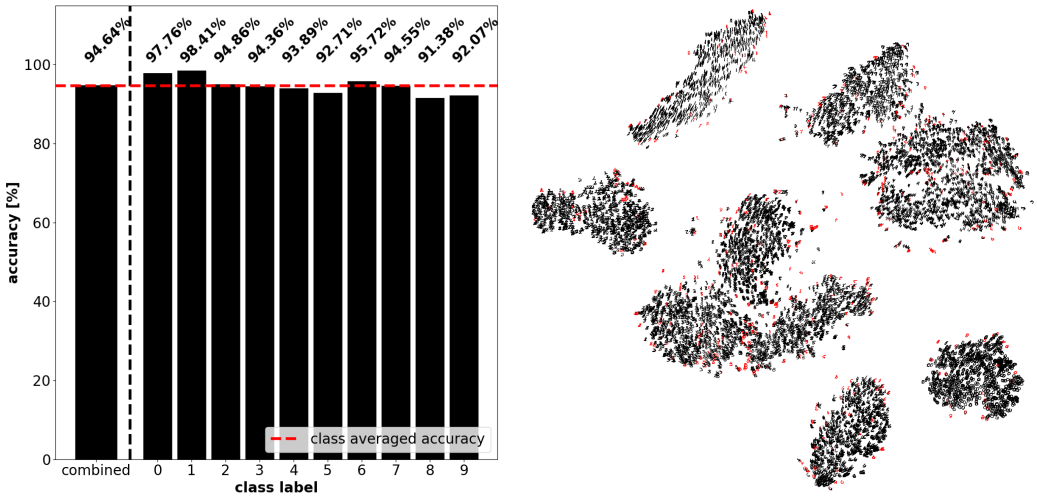


Figure 2: Overall accuracy, class specific accuracy and t-SNE visualization of the trained MLP.

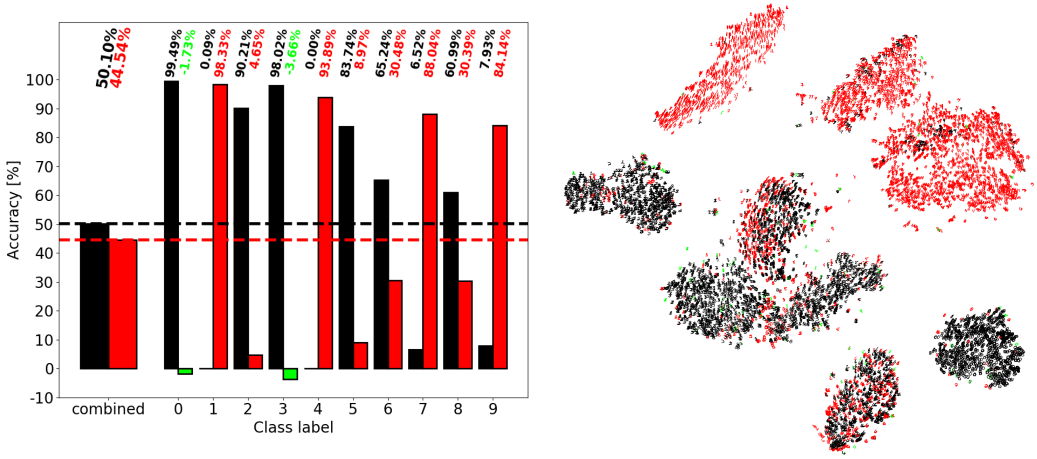


Figure 3: Overall accuracy, class specific accuracy and t-SNE visualization of the damaged MLP after the ablation of unit 12 in the first hidden layer. This unit is an example for the representation of features corresponding to many different classes

We found that the ablations of single units affected the network's accuracy in different ways. In general, the network's overall accuracy decreased, whereas the effect on single classes differed for different ablations. Figure 3 shows the effects of the ablation of Unit 12 in the first layer of the MLP, which resulted in the highest drop of overall accuracy of 44.54% for a single ablated unit. The black and red bars correspond to amount of correctly and incorrectly classified digits, respectively. Note that the red colored digits in the corresponding t-SNE plots do not contain the digits that were incorrectly classified by the undamaged network and only display the change of the classification as a result of the ablation of a single unit. The network lost its ability to correctly classify most digits of the classes one, four, seven and nine with a drop in class specific accuracy of more than 80%. The effects on the classes six and eight are less severe with a drop in class specific accuracy of around 30% while the effect on all other classes is smaller than 10%. The t-SNE plot suggests that this unit represents certain features in the data that are similar to each other across classes, as the majority of incorrectly classified digits are found close to each other in the upper part of the plot. Figure A.1 shows another example of such a representation, where most of the incorrectly classified units are found in the bottom right part of the t-SNE plot.

Figure 4 shows the effects of the ablation of Unit 19 in the first layer of the MLP, which resulted in drop of overall accuracy of 11.61%. In contrast to unit 12, this unit seems to represent features distinct to a single class, as the effect on the class specific accuracy for the class one is much stronger

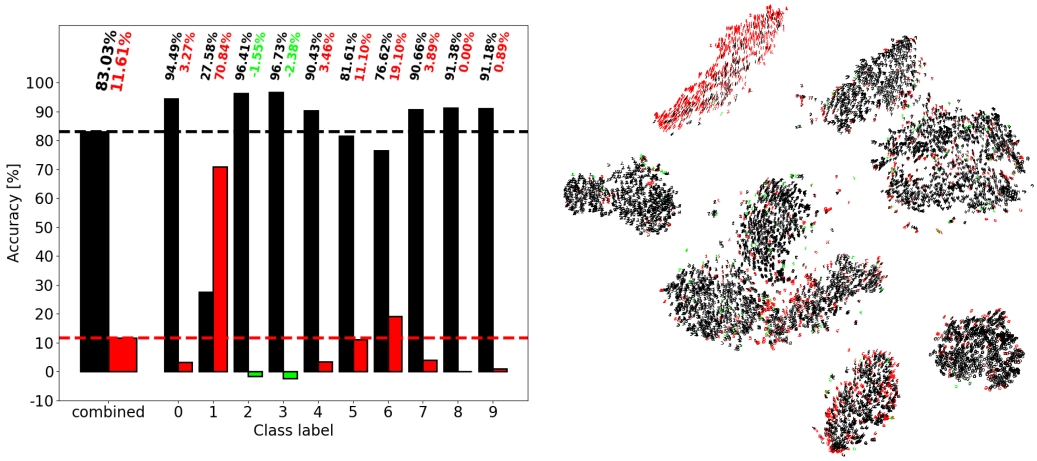


Figure 4: Overall accuracy, class specific accuracy and t-SNE visualization of the damaged MLP after the ablation of unit 19 in the first hidden layer. This unit is an example for the selective representation features distinct to a single class.

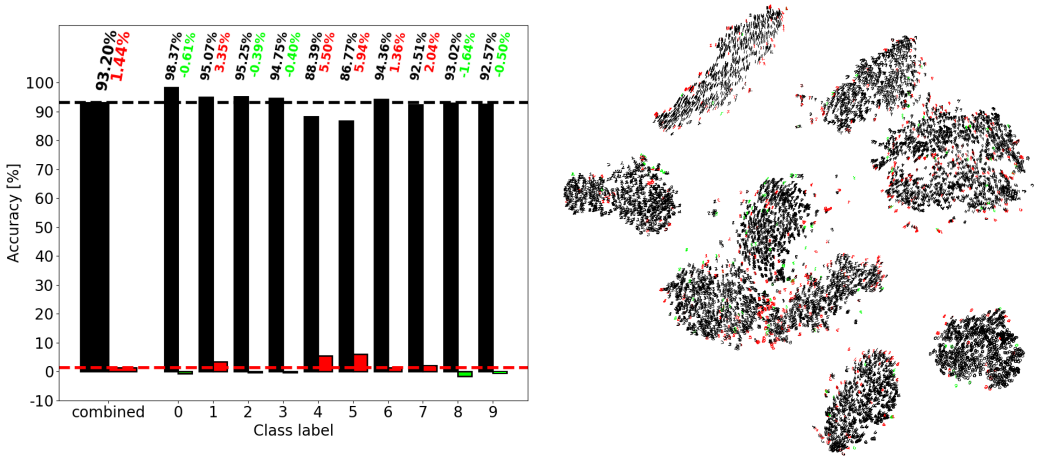


Figure 5: Overall accuracy, class specific accuracy and t-SNE visualization of the damaged MLP after the ablation of unit 6 in the first hidden layer. This unit is an example for a negligible contribution to the classification task and could be pruned to optimize network size.

than for all other classes. Although this unit is easy to interpret as it seems to represent a single class quite distinctively, it is not more important for the classification task than other units, e.g. unit 12, in terms of how strongly its ablation affects the network's classification performance. This result is consistent with previous investigations on the interpretability and importance of single units of an MLP classifier [40].

Figure 5 shows the effects of the ablation of Unit 6 in the first layer of the MLP, which resulted in drop of overall accuracy of only 1.44%. This unit seems to play no major role in the classification task as the effect of its ablation on the networks accuracy is small. We found four out of the 20 units in the first hidden layer, unit 6, 11, 13 and 18, showed this kind of behaviour and would be top candidates for pruning, if one would want to optimize the size of the network (c.f. Figure A.2).

Figure 6 shows the effects of the ablation of Unit 20 in the first layer of the MLP, which resulted in a drop of overall accuracy of 14.61%. This unit seems to represent features corresponding to subtle and smoothly changing characteristics distinct to the classes one, six and nine. The t-SNE visualization reveals that most of the incorrectly classified digits within a class can be found close to each other rather than evenly distributed across the whole class.

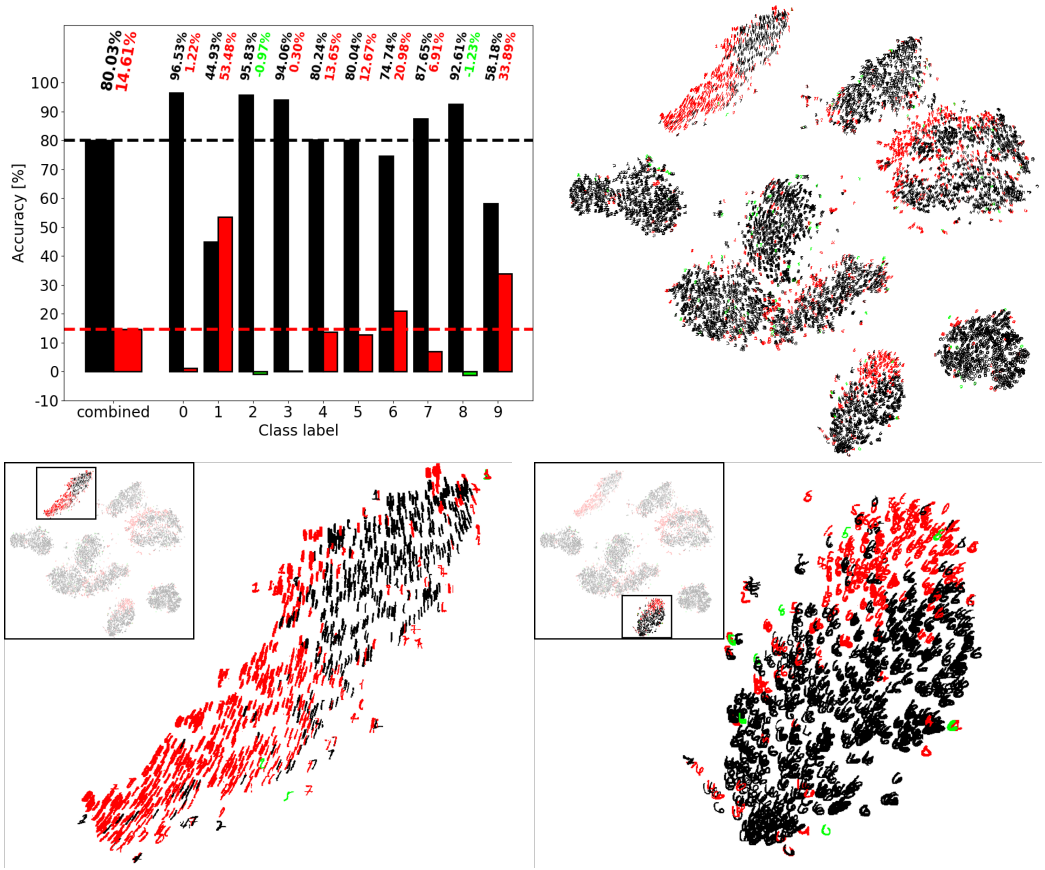


Figure 6: Overall accuracy, class specific accuracy and t-SNE visualization of the damaged MLP after the ablation of unit 20 in the first hidden layer. This unit is an example for the representation of features that are distinct to a subset of digits within different classes.

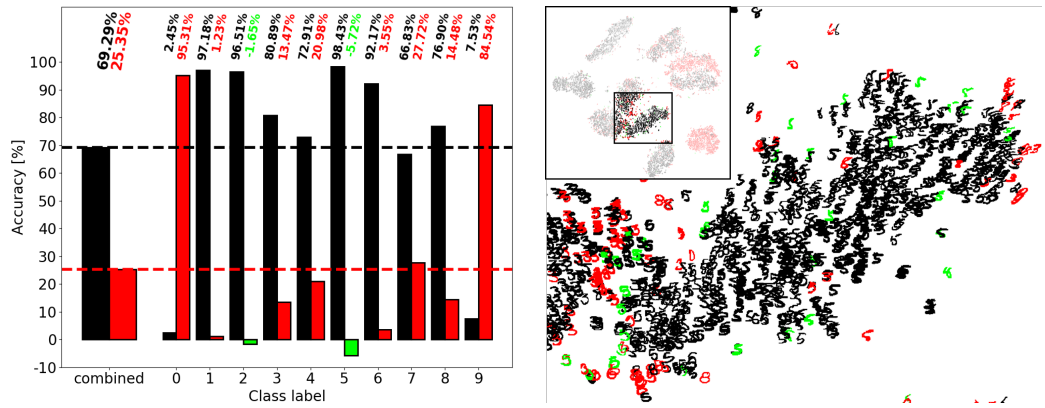


Figure 7: Overall accuracy, class specific accuracy and t-SNE visualization of the damaged MLP after the ablation of unit 3 in the first hidden layer. This unit shows the strongest positive effect of an ablation, i.e. the increase of the class specific accuracy of class five.

Figure 7 shows the effects of the ablation of Unit 3 in the first layer of the MLP, which resulted in a drop of overall accuracy of 25.35% but showed an increase of the class specific accuracy of 5.72% for class five, which is the strongest effect of all units in the first hidden layer. In general, we found that the damaged network would correctly classify some digits that were incorrectly classified by the undamaged network. This observation is consistent across different ablation localizations and in some cases showed a small increase of the class specific accuracies. This raises the question whether the

classification performance of a network can be increased beyond its trained capabilities by selectively ablating single connections to achieve the desired increase in accuracy without suffering from the negative effects.

Following the presented observations of the ablations, we aimed to find characteristics of single units which correlate with the drop in the network's overall accuracy after ablation of these units in order to be able to describe the importance of these units for the classification task. We found that the degree to which the distribution of the incoming weights of a unit after training differs from the randomly initialized normal distribution of weights before training is a good indication for the unit's importance for the classification task. We quantified this difference by the p-value of the Mann-Whitney U test, a non-parametric statistical test, which determines whether two independent observations were sampled from the same distribution. The p-value indicates the likelihood of both distributions being the same ($p = 1$) or being different from each other ($p \to 0$). Figure 8 shows a comparison of the network's first hidden layer's single unit weight distributions before and after training, whereas each distribution is visualized as a 28x28 pixel image. Note that the distributions of unit 6, 11, 13, and 18 did not change significantly during training (c.f. Figure A.2).

Figure 9 shows the pearson and spearman correlation of the Mann-Whitney U's p-value and the drop in accuracy after ablation. The left hand side shows 20 samples corresponding to the 20 units in the first hidden layer of the network from which the previous results were generated. In order to verify that the observed correlation is not a result of the random initialization of the network, we trained 20 more networks with different initializations and calculated the correlation coefficients for

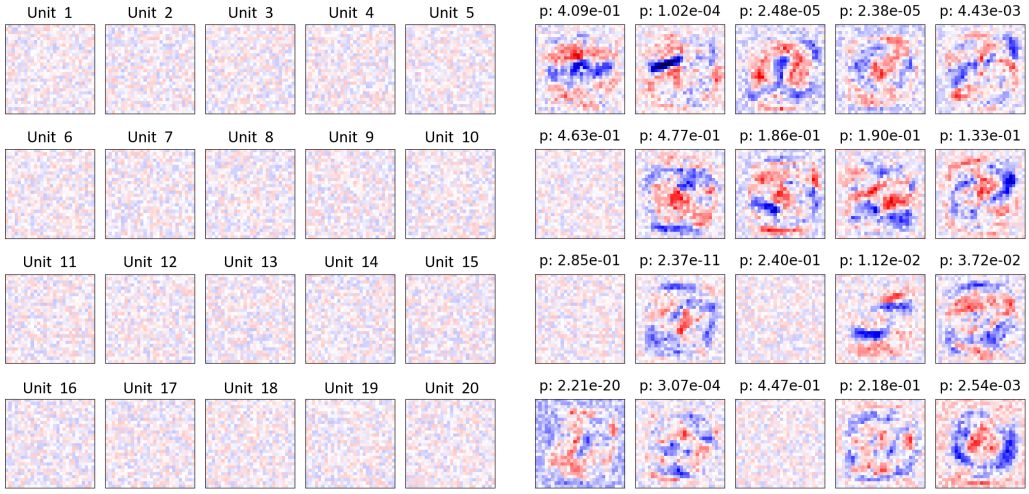


Figure 8: Comparison of the distributions of the incoming weights for the 20 single units in the first hidden layer before training (left) and after training (right).

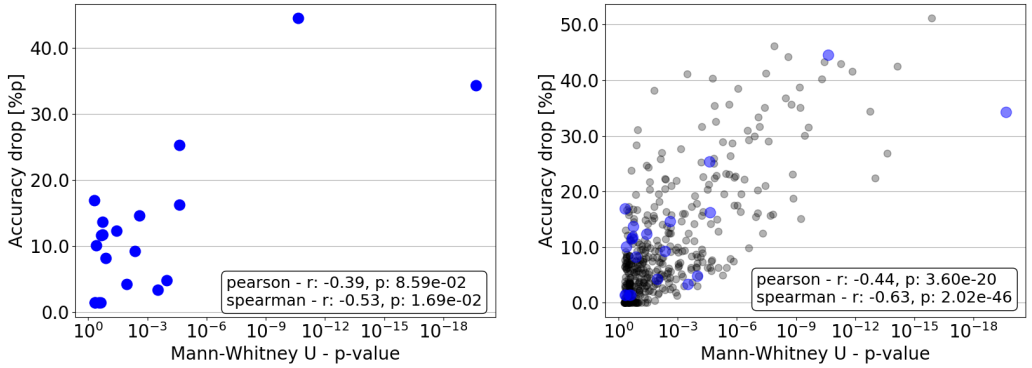


Figure 9: Correlation of the Mann-Whitney U's p-value with the drop in accuracy after ablation of a single unit in the first hidden layer.

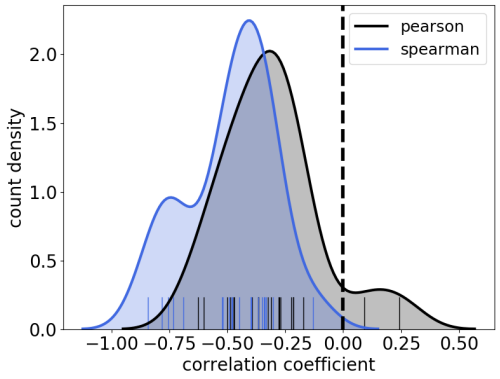


Figure 10: Distributions of the calculated pearson and spearman correlation coefficients for the 20 networks.

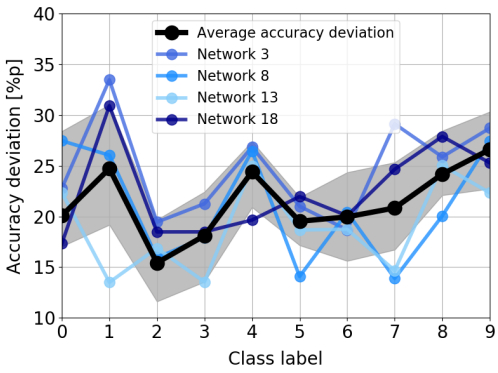


Figure 11: Class specific averaged deviation across the 20 networks of the dropped accuracy after ablations.

all 400 units within the first hidden layers of the 20 networks (c.f. Figure 9, right hand side). The results suggest that, in general, the more a single unit's distribution of incoming weights changes during training, the more important this unit is for the overall classification task. Figure 10 shows a kernel density estimated distribution of the calculated pearson and spearman correlations from all 20 networks and, except for two pearson coefficients, supports the average trend shown in Figure 9.

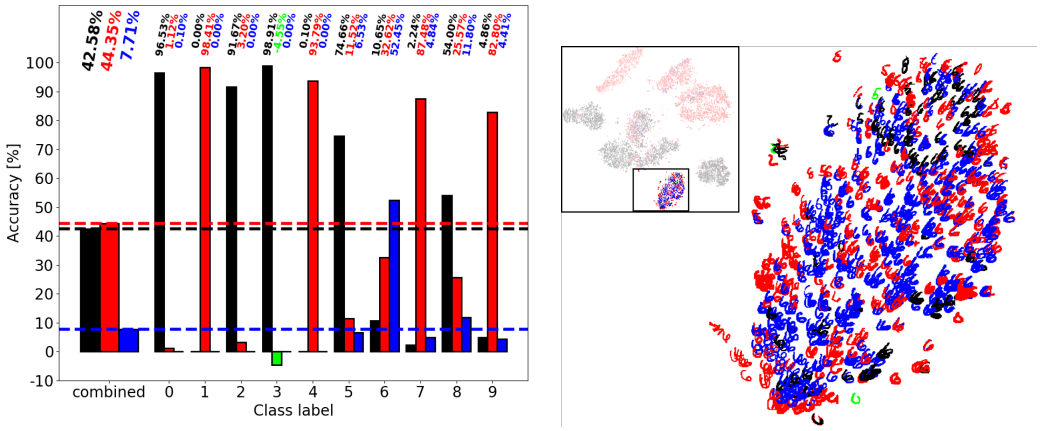
We wondered whether the representation of some classes within the networks is more selective than for other classes, i.e. whether the drop of the class specific accuracy after an ablation is similar for all units within a network or whether it shows a strong deviation. A high deviation would mean that some units within the network strongly represent a class while other units don't, suggesting that this class representation is somewhat localized in the network rather than evenly distributed across all units. Therefore, for each of the 20 networks, we computed the class specific drop in accuracy for all 20 single unit ablations in the first hidden layer and calculated the standard deviation. We further calculated the mean of this class specific accuracy deviation averaged across all 20 networks in order to compare the deviations of the single networks to the population mean. Figure 11 shows the population averaged accuracy deviation and four examples of a single network accuracy deviation. The black line corresponding to the population averaged accuracy deviation shows that some classes are represented more selectively than other classes. For instance, the classes one and four have a much higher deviation than class two, suggesting that, in general, class two is much more evenly represented across the first hidden layer than the classes one and four. However, this trend is not universal for all 20 networks indicated by the single networks' accuracy deviations. The fact that the blue lines cross the population average suggests that, despite the general trend, the selectivity of the representation of the 10 classes is somewhat unique to each network. This means that some networks develop a more selective representation for some classes than others.

4.2 Pairwise Unit Ablations in a shallow MLP

4.3 Grouped Unit Ablations in a Deep CNN

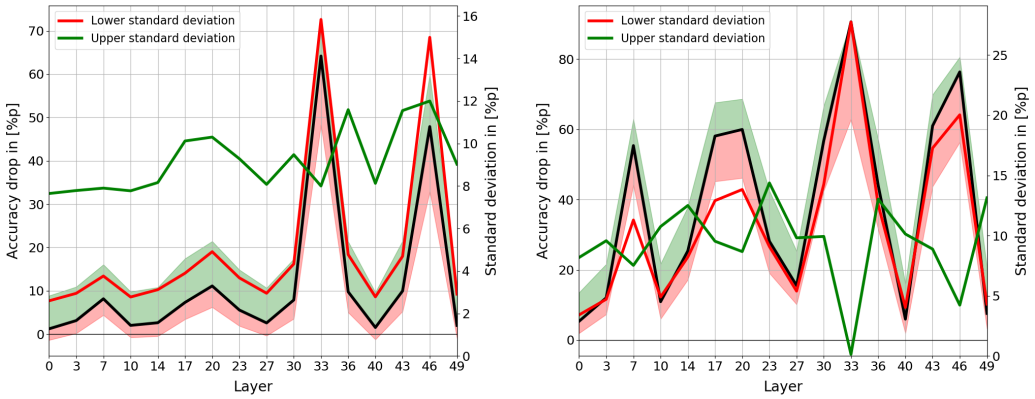
The first finding of the ablations was as expected. Generally, the higher the amount of ablated filters was, the more severe was the effect on the classification performance. The second finding was that some layers seemed to be more important than other layers for the network's performance. With regard to the VGG19 layer 33 and 46 showed particular importance, as can be seen in Figure 13 and Figure 14, which show the average top-5 and top-1 accuracy drop for ablation ratios of 10% and 25% over all convolutional layers.

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, we calculated the accuracy drop per class for different ablations. Figure 15 and Figure 16 show that the drop differs greatly dependent on the class. This suggests, that either some classes are generally harder to predict than others or that in each layer there is a certain class-selectivity in the stored information. Furthermore, the standard deviation of the accuracy drops varies for different layers. This could imply, that the way information



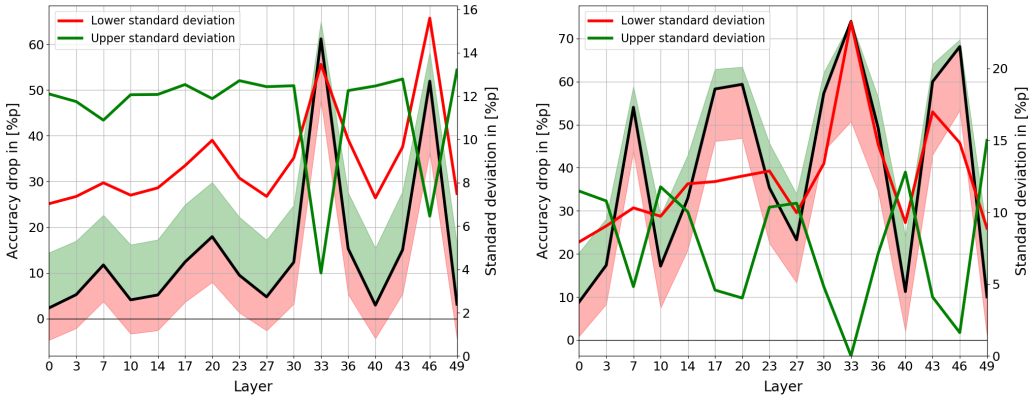
(a) accuracy (b) tSNE

Figure 12: positive effect for class 6 INCREASED!, negative effects also increased beyond the sum of single ablations!



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

Figure 13: top-5 accuracy drop

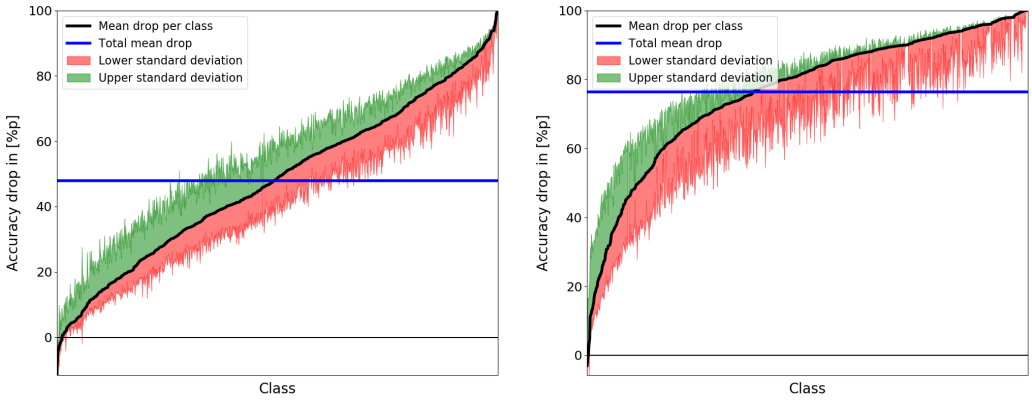


(a) after 10% ablation (b) after 25% ablation

Figure 14: top-1 accuracy drop

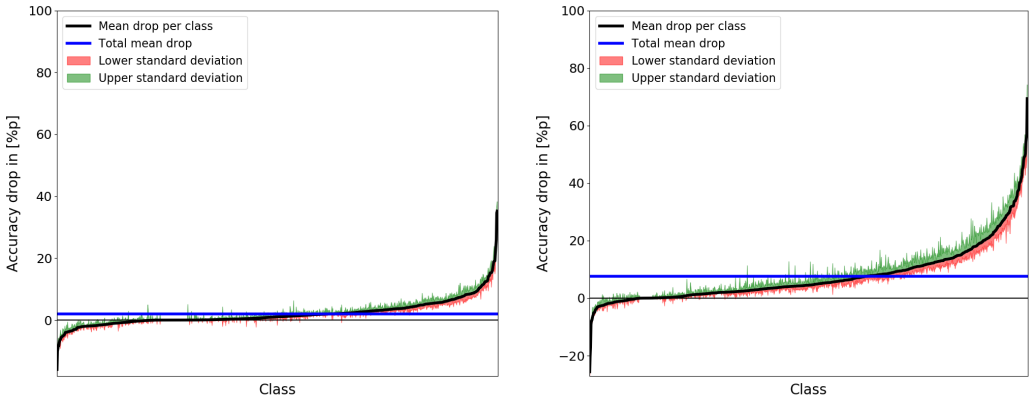
is distributed and stored differs from layer to layer leading to dissimilar effects on the classification

performance. Another interesting finding was that while for most classes the accuracy dropped after the ablation, for some classes the class-specific accuracy actually improved.



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

Figure 15: top-5 class-specific accuracy drop after ablation of layer 46

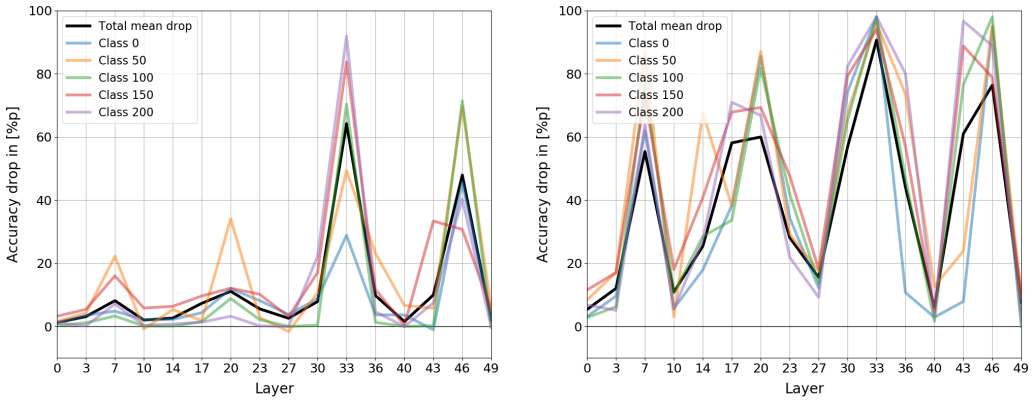


(a) after 10% ablation (b) after 25% ablation

Figure 16: top-5 class-specific accuracy drop after ablation of layer 49

As shown in Figure 17a and Figure 17b, the idea, that some classes are harder to predict than others is not clearly supported, since the lines for the accuracies of the different classes intersect. This means that classes effected strongly relative to other classes in one layer were effected comparatively weak in other layers, as shown in Figure 17. This could imply, that the information for correctly predicting particular classes is distributed differently in the layers of the network depending on the class. This means that there is a certain degree of class-selectivity in the layers and therefore the layers have a differing relative importance depending on the class.

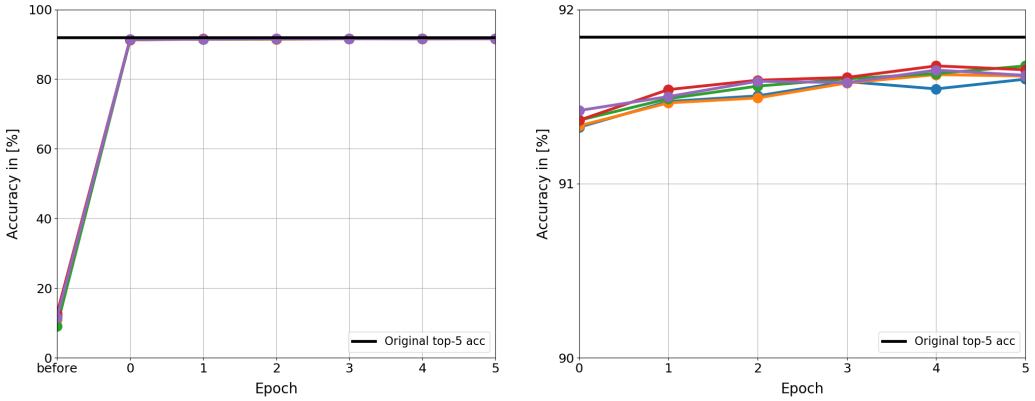
For the first experiment, Figure 18 and Figure 19 show the top-5 classification performance of the network for each training epoch. Each colored line represents one pass of the training. The VGG19 was able to recover performance-wise almost completely from the ablations, even though the ablated part was left out of the recovery. The difference between the recovered and the original accuracy is less than 1%p. The classification performance improves mostly within the first epoch. After the first epoch, the accuracy only increases marginally, as can be seen in Figure 18b and Figure 19b, which show the performance after the first training epoch. Generally, the original accuracy does not seem to be easily exceedable by recovery training. However, we always stopped the recovery training after 6 epochs due to limitation of computational training time, at which point the accuracy still seemed to be increasing. To be able to make a definite conclusion about the ability to recover, the recovery training needs to be carried out for even more epochs and possibly with an adaptive learning rate, which is able to account for small gradients towards the original accuracy. Furthermore, the set of ablated



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

Figure 17: top-5 accuracy drop for specific classes

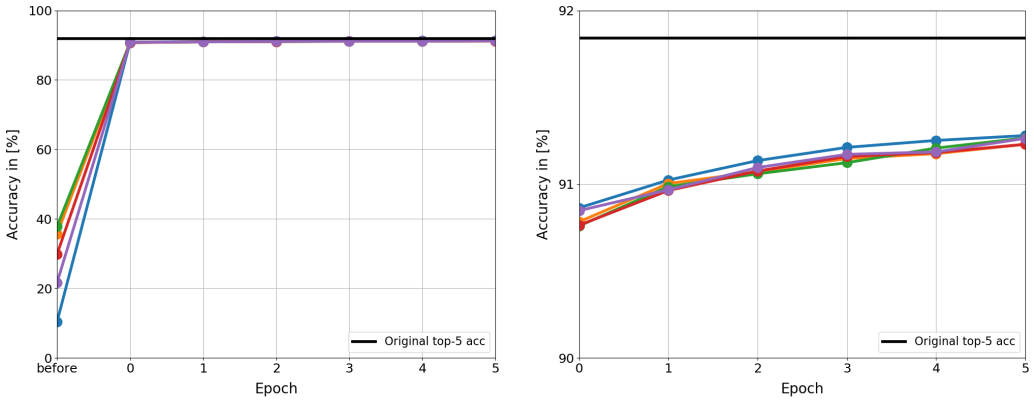
filters did not seem to make a difference in the recovery process. While for the iterations in layer 33, the accuracies after the ablation are rather close to each other, the accuracies for the iterations in layer 46 differed by around 30%p. However, during the training, the stronger drop did not have a negative impact on the recovery of the performance. We made the same observations for the top-1 classification performance.



(a) all epochs (b) after first epoch

Figure 18: top-5 accuracy for recovery training of layer 33 - experiment 1

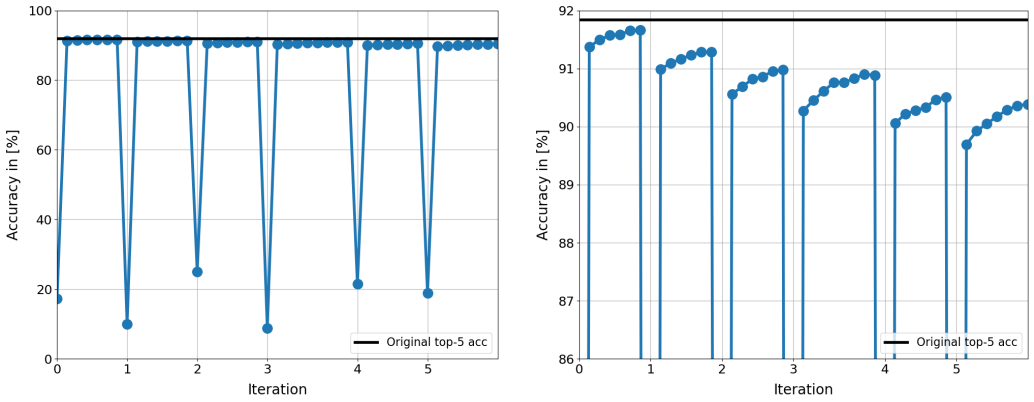
For the second experiment, we ablated the retrained network for 6 iterations without resetting the ablated filters. The number of ablated filters does not increase linearly, since the filters were chosen randomly from all filters, so that part of the chosen filters were already ablated. In the last iteration about 80% of the filters are ablated. Remarkably, the network, as before, was able to recover almost completely from the ablations, despite the increasing number of ablated filters. Figure 20 and Figure 21 show the classification performance of the network for each epoch of every iteration for the second experiment of the recovery training. Similarly, as in the first experiment, the performance rapidly increased during the first training epoch for each iteration and only improved slowly after that. The difference between the recovered and the original accuracy grew only slightly larger from iteration to iteration. The same observations were made for the top-1 classification performance. This means that with only about 20% of the filters in the ablated layer left, the network was still able to relearn and represent most of the necessary information. Additionally, the accuracy drop did not seem to increase, despite the accumulating number of ablated filters but varied depending on the set of ablated filters. For further analysis, it would be interesting to test how the accuracy develops for even higher ablation ratios and how the information content evolves throughout the network due to the recovery training, more precisely how the impact of an additional ablation study of the already manipulated



(a) all epochs (b) after first epoch

Figure 19: top-5 accuracy for recovery training of layer 46 - experiment 1

network would vary from the previously obtained results. As a side note, the accuracy drops of the ablations of layer 33 and layer 46 for the recovery training experiments seem to be weaker compared to the drop in the ablation study. This could suggest that the similarity by which the set of the filters for the ablations were chosen has a stronger impact on the classification performance than random ablations, which would imply that the similarity metric holds information in some form. However, to verify this result, further ablation experiments need to be conducted.



(a) all epochs (b) after first epoch

Figure 20: top-5 accuracy for recovery training of layer 33 - experiment 2

5 Conclusions and Future Work

...

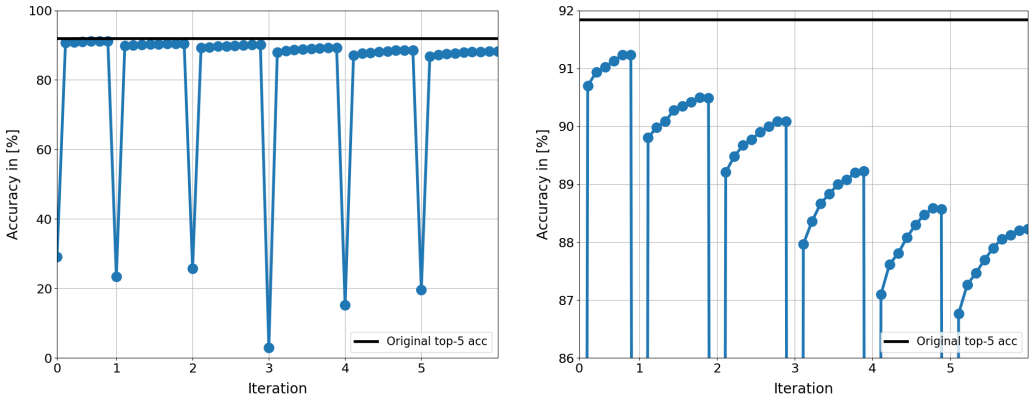
Acknowledgments

...

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.



(a) all epochs

(b) after first epoch

Figure 21: top-5 accuracy for recovery training of layer 46 - experiment 2

[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] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[43] 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.

A Appendix

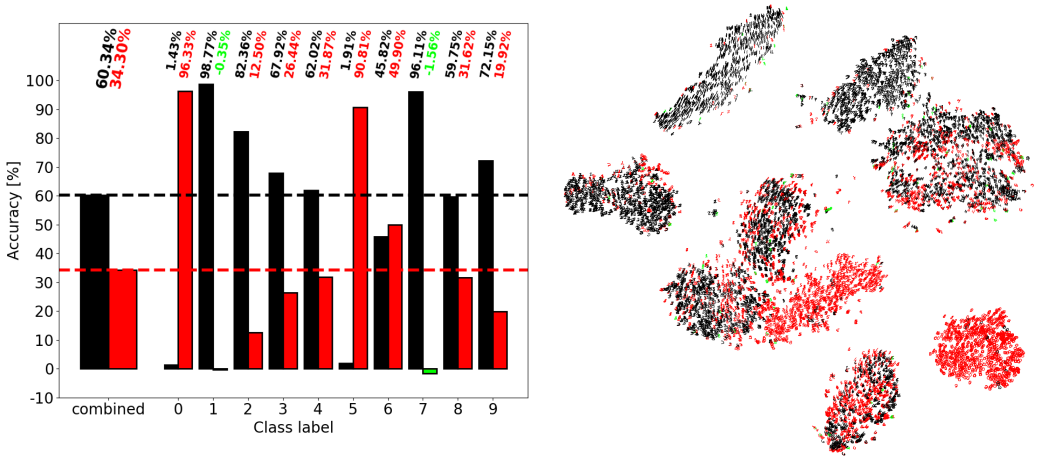


Figure A.1: Overall accuracy, class specific accuracy and t-SNE visualization of the damaged MLP after the ablation of unit 16 in the first hidden layer. This unit is an example for the representation of features corresponding to many different classes

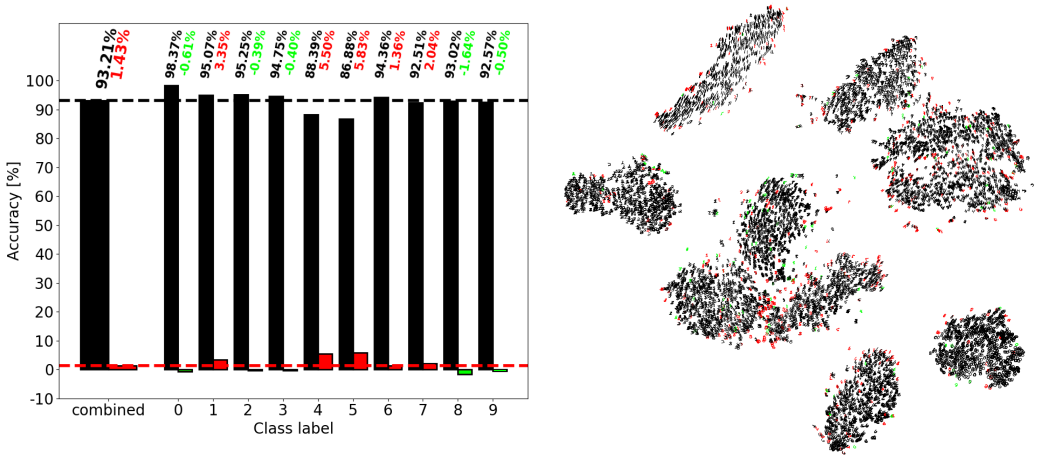


Figure A.2: Overall accuracy, class specific accuracy and t-SNE visualization of the damaged MLP after the ablation of units 6, 11, 13 and 18 in the first hidden layer. These units do not play a major role in the classification task and would be top candidates for pruning.