Ablation Studies to Uncover Structure of Learned Representations in Artificial Neural Networks

R. Meyes¹, M. Lu², C. Waubert de Puiseau² and T. Meisen²
¹ Chair of Technologies and Management of Digital Transformation, Bergische Universität Wuppertal, 42119 Wuppertal, Germany
² Institute of Information Management in Mechanical Engineering, RWTH Aachen University, 52068 Aachen, Germany

Abstract - Ablation studies have been widely used in the field of neuroscience to uncover structure and organization in complex biological systems. In this paper, we transfer the principle of ablation studies to two types of artificial neural networks trained to investigate the structure of their learned representations. We found that features distinct to the local and global structure of the training data are selectively and sometimes redundantly represented in specific parts of the network. Further, we found that the importance of these specific parts for the learning task can be determined based on the distribution of incoming weights of single units. Finally, we found that the network’s class-specific accuracy can be partly increased after training via ablations.

Keywords: Ablations, Artificial Neural Networks, Learning Representations, AI Transparency, Explainable AI

1 Introduction

Recent research on deep learning has brought forth a number of remarkable applications for different problems in a variety of domains. Prominent examples are visual object recognition, object detection and semantic segmentation in the field of computer vision [1-5], speech recognition and speech separation in the field of natural language processing [6-10] or self-learning agents based on deep reinforcement learning for video games [11-14], classic board games [15-17] as well as locomotion and robotic control [18-23]. During the last few years, the strong increase in availability of computational resources 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 parameters, of state-of-the-art DNNs. Due to this development, enabling the use of more complex algorithms and brute force methods, the main research focus in the past has been placed on increasing performance and speed of the trained networks to solve specific benchmark tests. Meanwhile, the development of new methods and perspectives for a deeper understanding of the structure of the learned representations in these complex networks was largely neglected.

In this paper, we follow a neuroscience-inspired approach based on the idea of ablation studies to analyze the structure of learned representations in DNNs. In these studies, neural tissue is damaged in a controlled manner while investigating how the inflicted damage influences the brain’s capabilities to perform a specific task. In the past, ablation studies proved to be a valuable method to investigate large, complex neural systems [26, 27], so it seems reasonable to investigate their potential for tackling state-of-the-art artificial neural systems.

Specifically, we performed ablation studies in two different network architectures trained on two different datasets. Conducting single unit ablations in a small, shallow multi-layer perceptron (MLP), we investigated correlations between the spatial as well as structural characteristics of the units and their contribution to the overall performance and to the class-specific performance of the network. We found, that some single units are important for the overall classification performance, while other single units are only selectively important for a specific class. Furthermore, the importance of a single unit for the classification performance correlates with the extent to which the unit’s weight distribution of incoming connections after training differs from the initial randomly initialized distribution. We further investigated the robustness of the classification performance by checking for redundant knowledge representations of specific classes in different areas of the network utilizing pairwise unit ablations. The results showed that pairwise unit ablations have a stronger effect on the classification performance than the summed effects of single ablations of the same units.

Second, we investigated a larger state-of-the-art convolutional neural network (CNN) for correlations between the size as well as the depth of the ablated portions of the network and the overall performance as well as the class-specific performance. We found that, in general, the larger the ablated network portion, the stronger the effect on the classification performance. However, this effect greatly varies across different depths of the network. The results show that some layers are universally more important for the classification task than other layers. However, this effect varies across specific classes.

Interestingly, for both networks, we found that ablations, despite having a general negative effect on the overall classification performance of the networks, consistently showed positive effects on the classification performance for specific classes. This raises the notion that the structure of a trained network may be purposefully manipulated to increase
its classification performance beyond the local optimum that was reached during training. Our code is publicly available\(^1\) to reproduce our results.

2 Related work

The basic idea of an ablation, i.e. removing trainable weights from a trained network, is also used when networks are pruned to reduce their size and computational cost, speeding up training and inference, while as much of the original performance as possible is retained. 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 [28]. Recent research on pruning state-of-the-art CNNs, like the VGG-16 or the ResNet-110, focused on the optimization of a network’s topology by removing kernels and entire filters [29, 30] and methods to find an appropriate ranking of units to tackle the seemingly simple but in fact quite challenging combinatorial optimization problem of how to choose the combination of units to be removed for the best results [31-33].

We aim to utilize the approach of ablations not merely to optimize the size and the speed of a network, but to gain insights about the structure and organization of the learned representation within the network, paving the way towards better transparency and interpretability of the network’s learned behavior. This objective is closely related to the question of how a network reaches its decisions and what the most important factors for this decision making process are. Some recent work on this matter demonstrated how the contribution of a network’s input elements to its decision could be explained by means of Deep Taylor Decomposition [34] or Gradient-weighted Class Activation Mapping [35, 36]. Another recent example, which focused 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 [37]. Complementary, some effort has been made to visualize the activity of hidden units of trained networks uncovering activity patterns propagating through the networks [38]. One of the most recent neuroscience-inspired contributions utilized ablations to demonstrate the relation between a network’s capability to generalize a classification task and its reliance on class-selective single units within the network. Specifically, networks which generalize well, contain less class-selective units than networks that merely memorize the dataset presented during training [39].

3 Study methods and design

3.1 Single/pairwise unit ablations in a shallow MLP

In the first part of the study, we trained a small, shallow MLP containing two hidden layers with 20 and 10 units (ReLU activation), respectively, on the MNIST dataset [40] to achieve a classification accuracy of 94.64%. After training, ablations of single units were performed by manually setting the weights of all incoming connections and the bias to zero, essentially preventing any kind of information flow through these units. All ablations were performed in the first hidden layer of the MLP. Effects of the ablations were evaluated visually based on t-SNE [41] projections of the complete test set (10,000 images) and quantitatively by comparing the performance of the network before and after ablations.

3.2 Ablations of filters in the VGG-19

In the second part of the study, we investigated the VGG-19 network pre-trained on the ImageNet dataset [42] (with batch normalization) as a representative of today’s state-of-the-art CNNs for object recognition tasks. Details about the architecture of the network and the training process can be found in [2]. Because of its size, the VGG-19 allows for depth resolved investigations of the effects of ablations. We performed ablations of groups of similar filters with increasing proportions relative to the total number of filters in each of the convolutional layers of the network. The similarity between filters within a group was calculated based on the absolute Euclidean distance of the normalized filter weights. Each of the filters in a layer was used once as a reference filter for the similarity-based choice of the remaining filters that were ablated as a group. Similar to ablations of single units in the MLP, ablations were performed by manually setting the weights and biases of all incoming connections of a filter to zero, effectively eliminating any activation of that filter. The effect of ablations was evaluated by comparing the network’s classification performance on the validation dataset using the top-5 accuracy before and after ablations.

4 Results

4.1 Single unit ablations in a shallow MLP

Figure 1 shows a t-SNE projection 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 dataset into the 2-dimensional space, it allows us to investigate whether this structure is represented in an organized manner in the network. The overall accuracy of the trained MLP on the test set was 94.6% with a slight variation across classes ranging from 91.4% for class 8 to 98.4% for class 1.

We found that the ablations of single units affected the accuracy in different ways. In general, the overall accuracy decreased, whereas the effect on single classes differed for specific ablations. Figure 2 shows the effects of the ablation of unit 12 in the first hidden layer of the MLP, which resulted in the highest drop of overall accuracy of 44.5% for a single ablated unit. The heights of the black and red/green bars correspond to the amount of correctly and incorrectly classified digits after the ablation, respectively. However, the red colored digits do not contain the digits that were incorrectly classified by the undamaged network and only display the change of the

\(^1\) https://github.com/richardmeyes/ablationstudies
almost exclusively, it is not more important for the classification task than other units, in terms of how strongly its ablation affects the overall classification performance. This result is consistent with previous investigations on the interpretability and importance of single units of an MLP classifier \[40\].

Contrary to the strong effects of the first two ablations, Figure 4 shows the effects of the ablation of unit 6 in the first hidden layer of the MLP, which resulted in a small drop of overall accuracy of only 1.4\(\%\). This unit seems to play only a minor role in the classification task as the effect of its ablation on the network’s accuracy is small. We found that 4 out of the 20 units in the first hidden layer, unit 6, 11, 13 and 18, showed similar effects which makes them top candidates for pruning, if one would want to optimize the size of the network.

In addition to the negative effects of ablations, Figure 6 shows an example of a positive effect of the ablation of unit 3 in the first hidden layer of the MLP, which resulted in a drop of overall accuracy of 25.4\(\%\). This unit seems to represent features corresponding to subtle and smoothly changing characteristics distinct to the classes 1, 6 and 9. The t-SNE projection reveals that most of the incorrectly classified digits within a class can be found close to each other rather than being evenly distributed across the whole class.

In addition to the negative effects of ablations, Figure 6 shows an example of a positive effect of the ablation of unit 3 in the first hidden layer of the MLP, which resulted in a drop of overall accuracy of 25.4\(\%\) but showed an increase of the class-specific accuracy of 5.7\(\%\) for class 5. Such positive effects are consistently observed across different ablated single units. After ablations, damaged network would correctly
classify some digits that were incorrectly classified by the undamaged network. These observations hint to a trade-off made while fitting the weights of the network via backpropagation, in which the recognition of a small number of digits is sacrificed for a much larger number of other digits. However, 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 observations of the ablations, we aimed to find characteristics of the single units which correlate with the drop in the overall accuracy after ablation of these units. Such characteristics would allow to describe the importance of single units for the classification task without the necessity to perform a functional test. We found that the degree to which the distribution of the incoming weights of a particular unit after training differs from the randomly initialized normal distribution of weights before training is a good indication of 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 to be the same \( p = 1 \) or to be different from each other \( p \to 0 \). Figure 7 shows a comparison of the weight distributions of the single units in the first hidden layer before and after training. Each distribution is visualized as a normalized 28x28 matrix, the same dimensions as the input images, with red and blue entries indicating high positive and negative weight values, respectively. The p-values indicate the difference of the distributions on the right-hand side of Figure 7 compared to the left-hand side of Figure 7. Note that the distributions of unit 6, 11, 13, and 18 did not change significantly during training.

Figure 5: Overall accuracy, class-specific accuracy and t-SNE projection 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 6: Overall accuracy, class-specific accuracy and t-SNE projection 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 5.

Figure 7: 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 8, left-hand side shows the Pearson and Spearman correlation of the Mann-Whitney U’s p-value and the drop in accuracy after ablations in 21 instances of the trained MLP. In order to verify that the observed correlation is not a result of the random initialization of the original network (blue dots), we trained 20 more networks with different random initializations and calculated the correlation coefficients for all 400 units within the first hidden layers of the 20 networks (black dots). 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 performance. Figure 8, right-hand side 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, left-hand side. This observation may prove useful for pruning neural networks. Units may be pruned based on the distributions of their incoming weights, thus, reducing the computational cost of repeatedly testing a pruned network on a large dataset.

Figure 8: Correlation of the Mann-Whitney U’s p-value with the drop in accuracy after ablation of single units in the first hidden layer (left). Distributions of the calculated Pearson and Spearman correlation coefficients for the 20 networks (right).
4.2 Pairwise unit ablations in a shallow MLP

In addition to single unit ablations, we performed pairwise unit ablations in the first hidden layer of the MLP to investigate the feature representations for redundancies, i.e. whether the effects of pairwise unit ablations are stronger than the sum of the corresponding single unit ablations. In this case, the network retains its capability to correctly classify some specific classes after a single unit ablation as another unit still represents the corresponding features sufficiently well. However, the pairwise ablation of both of these units causes the network to incorrectly classify those classes that were correctly classified in the case of the single unit ablations, as there are no more units left that redundantly represent the necessary features.

Figure 9 shows the effects of the pairwise unit ablation of units 4 and 16 in the first hidden layer of the MLP, causing the strongest observed effect to exceed the sum of the corresponding single unit ablations. The height of the black, red/green and blue bars correspond to the number of objects correctly classified after the pairwise unit ablation, the number of objects incorrectly classified after the corresponding single unit ablation and the number of objects incorrectly classified only after the pairwise unit ablation. The digits in the t-SNE plot are colored accordingly.

Figure 10: Overall accuracy, class-specific accuracy and t-SNE projection of the damaged MLP after the pairwise unit ablation of units 5 and 10 in the first hidden layer. The positive effect on class five (5.2%p) is stronger after the pairwise unit ablation than the summed effects after the corresponding single unit ablations (-3.1%p and -0.4%p).

4.3 Ablations of Filters in the VGG-19

Complementary to the investigation of a shallow MLP, we investigated the VGG-19 as a more complex representative of state-of-the-art CNNs for image classification tasks. Similar to the importance of single units in the MLP, we found that some layers are more important for the classification task than other layers. Figure 11 shows the drop in top-5 accuracy, for the ablation of 10% (left-hand side) and 25% (right-hand side) in all convolutional layers of the VGG-19. The black curve shows the accuracy drop in each layer, averaged over all ablations performed in this layer. The number of ablations is equal to the number of filters in each particular layer, since each filter was chosen once as a reference for the choice of the 10% and 25% ablations based on filter similarity (cf. section III.B). The red and green curves and shaded areas correspond to the lower and upper standard deviation from the average accuracy drop. Layer 33 and 46 showed a significantly higher drop in the top-5 accuracy compared to other layers. This effect is more distinct for the smaller number of ablated filters (10%) and becomes less pronounced for the larger number (25%). Concurrently, the effect of a larger number of ablated filters...
has a stronger impact on some layers than on others. For instance, layers 7, 17, and 20 show a significantly stronger drop in the top-5 accuracy for 25% of ablated filters compared to 10% of ablated filters, while layer 40 is almost not affected at all. This observation is somewhat surprising as we expect the upper layers to be the most important layers, as they are supposed to represent more general features common to many classes, whereas lower layers are supposed to represent more class-specific features [43]. Additionally, the fact that some layers, e.g., layer 40, are largely unaffected by the increase of the proportion of ablated filters from 10% to 25% suggests that the representation of features in this layer may be redundantly represented in other layers or in other filters in the same layer, rendering ablations mostly harmless for the overall performance. Consistent with the observations of positive effects of ablations in the MLP study, the ablation of some filters in some layers of the VGG-19 showed an increase in top-5 accuracy indicated by the crossing of the zero-line of the red shaded area in Figure 11.

Similar to the MLP study, we checked whether the importance of the layers for the overall classification performance shows class-specific variations. We found that, despite the general class-average trend (c.f. Figure 11, black line), some layers are much more important for specific classes than for others. Figure 12 shows the class-specific drop in top-5 accuracy averaged over all ablations for five example classes in addition to the average drop in top-5 accuracy as Figure 11.

In the case of the 10% ablations, class 50 shows a much higher drop in accuracy relative to the other classes after ablations in layer 7 and 20 and at the same time a lower drop in accuracy relative to the other classes after ablations in layer 33. Additionally, the drop in accuracy after the 25% ablation in layer 14 and 17 is much stronger and much weaker, respectively, than the average drop. This observation suggests that layers exhibit a certain degree of class-selectivity and therefore have different relative importance for the overall performance depending on the class. Based on this finding we further investigated how this selectivity is distributed across classes, i.e. to what extent a layer represents specific classes more than others.

We further investigated whether the different layers exhibit some class-specific selectivity in their representations, i.e. to what extent a layer represents specific classes more than others. Figure 13 shows two extreme examples for the class-specific drop in top-5 accuracy after ablations of 10% in layer 46 and 49. Layer 46 shows a much less pronounced class selectivity, as most classes are strongly affected by the ablations of filters in this layer. Contrary, a much smaller number of classes is affected strongly by ablations in layer 49, suggesting that this layer exhibits a stronger class selective representation. Consistent with the observations of the MLP study, ablations had a negative effect on the classification performance for most classes. For some classes, however, the class-specific top-5 accuracy improved after the ablations. In general, this effect was stronger for smaller ablations and in layers with a comparably small impact on the overall performance, such as layer 49.

5 Conclusion and future work

We investigated the effects of single and pairwise unit ablations on the classification performance of a shallow MLP trained on the MNIST dataset. As expected, we found that as a result of ablations, the overall classification performance generally decreased. However, in some cases the class-specific performance for some classes increased despite the overall impairing effect. Furthermore, the ablation of single units revealed their different contributions to the overall classification performance. Some units are universally important for the task, representing features distinct to a majority of the classes in the test set while other units are selectively important for specific classes and some units turned out to be not important at all. In our future work, we aim to investigate the origin of the positive effects in detail. We will investigate whether ablations can be used in a controlled manner to attain a performance which exceeds the network’s backpropagation-trained local optimum.

Furthermore, we found that the distribution of incoming weights as a structural characteristic of a single unit indicates its importance for the classification task. Specifically, the more a weight distribution changed during training, the more important this unit is for the classification task. This result may prove useful for pruning experiments, as the importance of a unit can be estimated with significantly reduced computational cost as compared to full functional tests of the network. Interestingly, the t-SNE projections of the ablation effects revealed that the features represented by single units mostly correspond to the global and local structure inherent to the data set. This suggests that information inherent to the stimuli, with which the network is trained, is mapped and locally represented in specific areas of the network. In a future study, we aim to investigate an underlying organization of this mapping.
We further found that the network exhibited robustness against the structural damage caused by ablations, as some features are represented redundantly in different single units. Specifically, pairwise unit ablations revealed effects on the classification performance that exceed the combined effects of the corresponding single unit ablations. Remarkably enough, this observation is true for the negative as well as for the positive effects on the performance.

Transferring the principle of ablations to the VGG-19, consistent with the observations for the MLP, we found that ablations had negative as well as positive effects on the classification performance. Specifically, the higher the number of ablated filters was, the stronger the effect. We further found that the different layers are not equally important for the classification performance. More precisely, the effect of ablations turned out to be significantly stronger for two deeper layers (33 and 46) than for all other layers. At first, this may seem surprising as we expected the upper layers to be the most important layers as they are supposed to represent more general features common to many classes, whereas lower layers are supposed to represent more class-specific features [43]. However, a possible explanation may be that the features represented in these two layers are not redundantly represented in other layers, while the features in the upper layers are.

Finally, we conclude that ablation studies performed in artificial neural networks are a feasible method to investigate the structure of their learned representations. It allows to determine the importance of specific areas of the network for the learning task and sheds light on how of these specific areas contribute to the network’s overall ability to solve the learning task.

6 References


