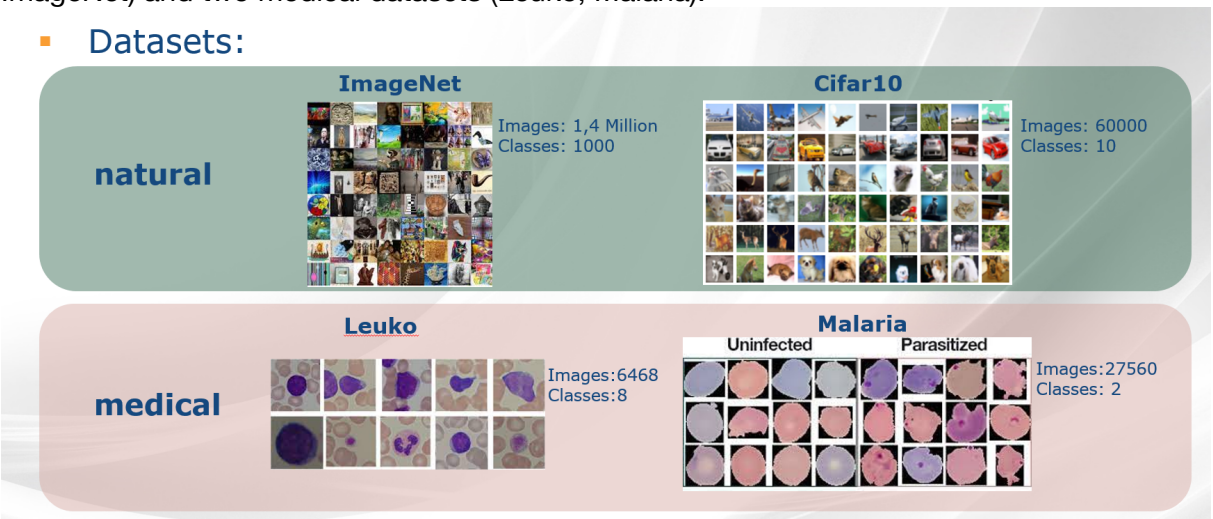
Project Description: Self-Supervised Learning

Background:
Deep learning is capable of achieving very good results in the field of computer vision. Furthermore, it enables the solution of new applications that would not be conceivable using classical image processing methods.
This makes deep learning an attractive tool for companies in the vision sector to offer to customers or for customer solutions.
Currently, the state of the art uses supervised learning methods to train these deep learning approaches. This learning method leads to very good results for many applications. The disadvantage is that a large amount of labeled data (several thousand to ten thousand images) is required, which usually must be provided by the customer.

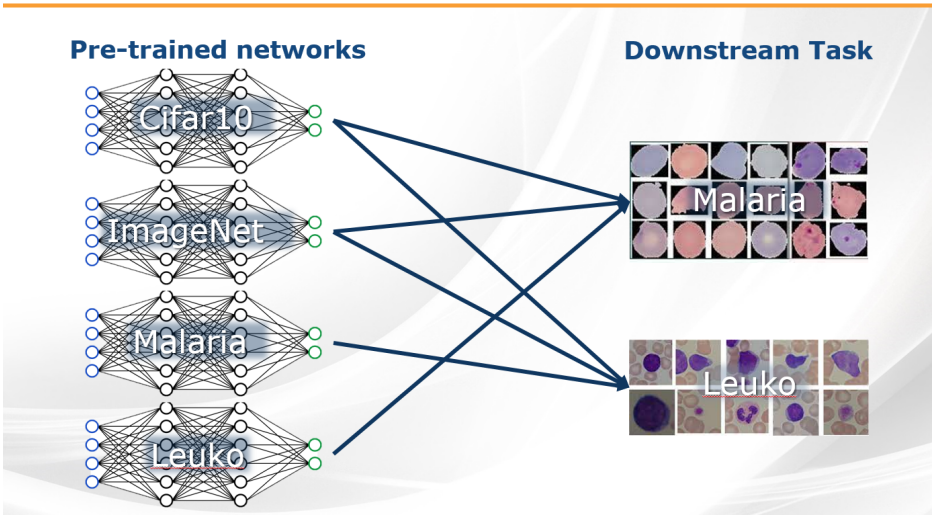
Objective:
The goal of this project is to reduce the effort and costs on the customer side by using algorithms and alternative learning methods that require fewer labeled data.

What was done?
As part of a 4-month internship, two different alternative learning methods were implemented and investigated.
Through transfer learning or the use of pre-trained networks, significantly fewer labeled data are required (state of the art). Google, Facebook, etc. provide open-source networks that have been trained on billions of data. This leads to the same pre-trained networks being used for many completely different applications and being adapted to the final application using "fine-tuning."
The available networks were mostly trained on natural images, e.g., dogs, cats, cars, etc.
It is now to be investigated to what extent these pre-trained networks are helpful for medical or industrial applications, or whether, for example, special pre-trained medical networks should be created and used.
For this purpose, four datasets were selected: two datasets with natural images (Cifar10, ImageNet) and two medical datasets (Leuko, Malaria).



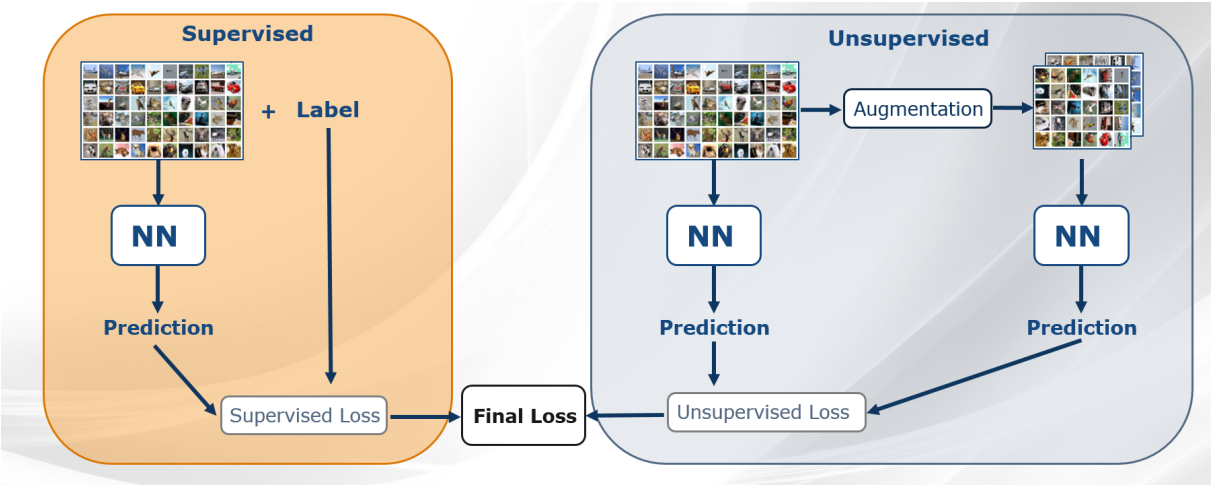
Furthermore, four networks are created, each trained from scratch on one of the four datasets. For the ImageNet network, an already trained open-source network is used. In a second step, these networks are used as pre-trained networks and each is fine-tuned on the two medical datasets.

Experiment



For smaller, more efficient networks, which could, for example, be implemented on an FPGA, no pre-trained networks are available. These networks must therefore be trained from scratch. For this, a learning method is needed that requires significantly fewer labeled data than "supervised learning." A learning method that fulfills this property is "self-supervised learning."
Self-supervised learning is another learning method for neural networks. The labels are derived from the input data and do not need to be explicitly specified.

As part of the internship, a promising approach was selected and implemented. The approach is called "Unsupervised Data Augmentation" (UDA), deals with the so-called "consistency loss" and self-supervised learning, and is illustrated using the following slide.



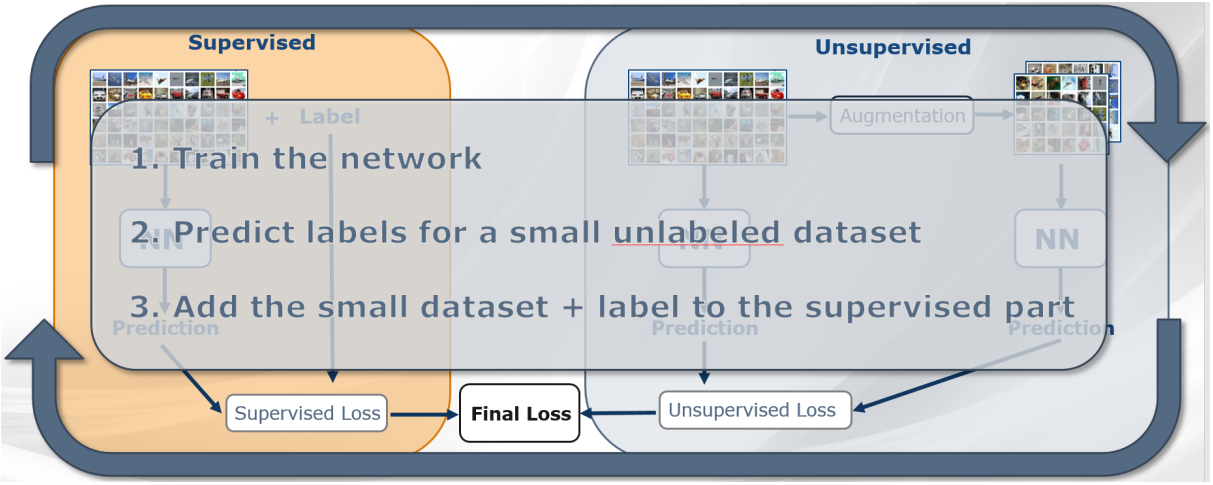
The left part of the figure describes the supervised part. This corresponds to classical supervised learning and serves to learn the distinction of objects based on a small number of labeled training data. The right part describes the unsupervised part. In the first path, the image is evaluated by the network, resulting in a probability distribution over the classes. In a second path, this image is first augmented (e.g., rotated) and then evaluated by the network. This also results in a probability distribution over the classes. A loss function is used to minimize the distance between these two distributions. This serves to learn the variations and manifestations of the objects.

With this approach, the following results can be achieved:

Method	Labelled Data	Unlabelled Data	Accuracy	Difference
SSL	4000 (8%)	46000	86.57%	3.55%
SSL	2000 (4%)	48000	79.66%	10.46%
Supervised	50000	-	90.12%	-
Supervised	4000	-	82.02%	9.97%

The best result for test accuracy is achieved by the classical supervised learning method with 50,000 labeled images. The goal would be to achieve the same test accuracy with the SSL method using significantly fewer labeled images. The results show that with only 4,000 labeled data and 46,000 unlabeled data, only a test accuracy of 86% is achieved. If the number of labeled data is further reduced, for example to 2,000, the test accuracy is only 79%.

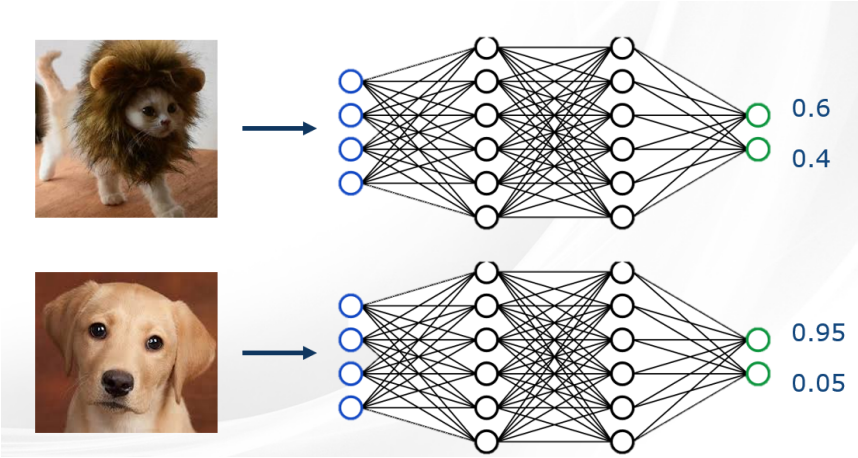
A first approach to improving this method is described using the following graphic:



For this, the network is first trained using the "UDA approach" until it reaches a previously defined test accuracy, e.g., 70%. Then, a small image dataset (e.g., 100 images) is extracted from the unlabeled dataset and evaluated by the network, whereby these images receive a label. This dataset and the labels are then added to the labeled dataset in the supervised part. This increases the proportion of labeled image data. The network is then further trained. These steps are then repeated.

Results of the improvement method:
With this approach, about 30% new, incorrect labels are initially added to the labeled dataset. This means that the network cannot improve sufficiently to achieve higher test accuracy.

Further improvement possibilities:
It is also being investigated whether test accuracy can be improved if only the image data and labels that have a high prediction probability (e.g., greater than 90%) for a class are added to the "supervised dataset."



Analysis:
How many images generate a higher prediction probability than 90%?
For this, the network is first trained on 2,000 labeled images and 48,000 unlabeled images until the network reaches a test accuracy of 70%.
Then, 30 images are first extracted from the unlabeled dataset and evaluated by the network. This step is repeated several times. It turns out that on average, 15 images have a higher prediction probability for a class than 90%.
This experiment was repeated for a number of 100 images. Here, on average, 26 images have a higher prediction probability of 90% for a class.

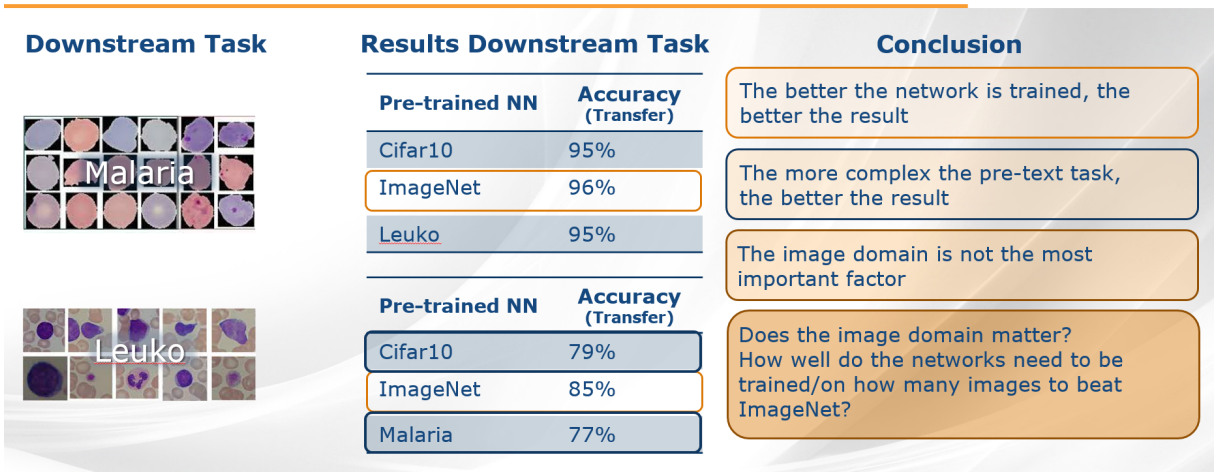
Are these labels always correct?
The results showed that a network must achieve at least a test accuracy of 70% for these labels to always be correct.
If the network achieves lower test accuracies, then even for high prediction probabilities (>90%), incorrect labels are present.
The analysis thus shows that it may make sense to only add images with a higher prediction probability than 90% to the labeled dataset in order to reduce the number of incorrect labels within this dataset.

What is the result?

Transfer Learning:
The results from the transfer learning experiment are shown on the following slide. It shows that the network pre-trained on the ImageNet dataset always produces the best results. This suggests that what matters most is how well a network has been trained/how good the already learned features are. It is not necessarily advantageous for the "pretext task" to have the same image domain as the downstream task. MVTec sells specially trained networks on industrial datasets. It would be interesting to see whether these networks achieve better results for an industrial downstream task than the ImageNet networks or whether even fewer training data are required. In this case, however, it would be relatively laborious to create these special pre-trained networks.

Transfer Learning

Conclusion



Self-supervised Learning:
The results from further adaptation have shown that test accuracy can be improved by 4-7%. Thus, for 2,000 labeled image data, a test accuracy of 85% can be achieved, which is an improvement of about 5% compared to the original UDA method. However, with this method, the 90% test accuracy achieved by the classical supervised method with 50,000 labeled image data cannot be reached.
The reason for this could be the unequal distribution of classes that are added to the labeled dataset through this adaptation. Thus, a different number of images are generated for each class within the training dataset. Further analysis has shown that, in the course of training, labels with a high prediction probability (>90%) are no longer correct. This in turn means that further improvement of the network is only possible to a limited extent.

Conclusion

Transfer Learning:
The experiments conducted here have shown that it is not necessarily advantageous to create pre-trained networks for specific applications. It would be interesting to test whether networks that have also been trained on billions of, for example, industrial data, are able to achieve better results for these industrial applications than the ImageNet networks or whether this leads to a further reduction in training data. In this case, however, it would be relatively laborious to create these special pre-trained networks.

Self-supervised Learning:
The approach of the UDA method (including the tested improvement possibilities) is not suitable if the amount of labeled data is to be further reduced, as the gap between the 90% test accuracy of the supervised method and the test accuracy that the network can still achieve with fewer labeled data becomes ever larger.
A further possible improvement in test accuracy could be achieved if an attempt is made to always add the same number of images for each class to the labeled dataset.
