Universität Leipzig
Fakultät für Mathematik und Informatik
Institut für Informatik

Problem-seminar Deep-Learning

Summary about

Computer Vision and its implications

based on the paper

ImageNet Classification with Deep Convolutional Neural Networks

by

Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton

Leipzig, WS 2017/18

vorgelegt von
Christian Stur

Betreuer: Ziad Sehili

Contents

1.0. Introduction

On the following pages, I will explain the paper "ImageNet Classification with Deep Convolutional Neural Networks" (Alex Krizhevsky and Sutskever 2012) and its broader context. I will explain certain principles of Neural Networks, then go in depths concerning Convolutional Neural Networks and the advantages and significance of the network presented in the above-mentioned paper.

1.1. Present significance

Today, Neural Networks and specifically Deep Neural Networks have become a center of attention in the TECH-Community. Yet, the theory of neural networks has begun in the early 1960s but didn't reach an applicable stage until recently. How did this change take place?

Over the past few decades, there were multiple parallel improvements in areas which have positively impacted the practicality of neural networks in general, but two main factors predominately made neural networks a practical solution to many problems that have been unsolved until recently.

Firstly, the massive advancements regarding the performance and reduced pricing of general processing units (= GPUs) has made it financially feasible, and time-wise doable to train neural networks in a reasonable amount of time. GPUs showed significant advantages to CPUs by having the capability to write and read directly from each other without accessing the hosts machine memory and therefore are way quicker in parallel processing tasks such as training neural networks.

Secondly, the amount, accessibility and variety of data-sets necessary for training neural network has widened and spread over community-based websites such as www.kaggle.com or LabelMe (N. Pinto 2008). Training a machine learning algorithm often requires thousands of individual data-points depending on the complexity and accuracy required to make a prediction. Neural networks require an even bigger amount of available data. Computer Vision, which is addressing a significantly more complex problem, requires an even broader number and variety of imagery to train a network with.

1.2. Motivation of Computer Vision

So why research a computers ability to see what is happening on an image? One general aim of technology is to deliver a certain service to their users in an automatic fashion without or minimal work from another person. Yet, a lot of tasks in today's world require for humans simple, but until recently, for machines impossible tasks. Such include identifying another person, categorizing objects or detecting anomaly in sight.

Simple forms of computer vision are already deployed in quality control of for example agricultural products, in which categorization takes place on conveyor belts. Yet these solutions mainly worked statistically, where manual human quality control often had to be done after and the system worked for only a very narrow complexity of problems. These areas have seen significant improvement.

Person identification has always been a big interest in computer vision as its use-cases include security, advertisement and retail. One of the first successful mass-deployment software working on the bases of computer vision became Instagram with their face-detection and their overlay functionality. Taking over an extent of social media and demonstrating that the degree of sophistication of computer vision reached a level on which reliable software can be built, made a broader audience aware of this technology and its future potential. Google (Alexander Toshev 2009) (Ivan Bogun 2015), Adobe (Xu 2017) (Yi 2017) and Facebook (Tsung-Yi Lin 2017) have invested heavily into research concerning computer vision.

Computer vision also is being incorporated into systems to progress faster in scientific discoveries such as fluid-dynamics (Lubor Ladicky 2016) or even early cancer detection (Álvarez Menéndez 2010) by using check-up X-Ray- or MRT-imagery.

Additionally, computer vision will be a main part of technologies which are on the brink of mass-deployment such as Tesla's self-driving cars (Lex Fridman 2017), which will require a very significant ability and accuracy at detecting and classifying objects, and Amazon's Go Stores (Kayla Backes 2017) are mainly based on object detection and facial recognition.

The potential for scientific discovery, business and the understanding of vision based "intelligence" seem to be deeply connected to the progress of computer vision.

2.0. Computer Vision

In this section, we will explore the bases of neural networks, the competition, the specific neural network presented in this paper (Alex Krizhevsky and Sutskever 2012), the current state and the potential future of computer vision.

2.1. Neural Networks

Neural networks have only recently become state of the art for computer vision in classification tasks (see 2.2.2). The base structure of a neural network is a neuron. It consists of incoming connections, outgoing connections (except for the last layer) and an activation function. The activation function describes what to output based on the input.

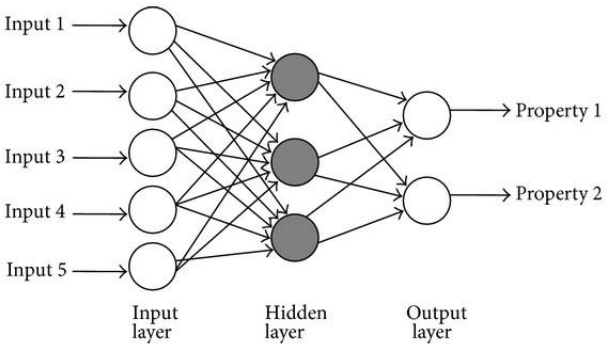


Figure 1) A basic neural network consists of an input layer, a hidden layer and an output layer. Input is being passed onto the hidden layer in which an output is computed based on the input value, the connection weight and the activation function.

A basic neural network consists of an input layer, a hidden layer and an output layer (which can also just be a single neuron. On each connection there is a weight placed, which represents the importance of the certain connection for the decision-making process. A decision made by the network can be of Boolean nature or also a classification task (see below). The last layer of such a network is directly connected to the possible “solutions”. In classification tasks the last layer is a special layer called softmax.

Supervised learning takes place in the network by receiving data in form of numerical features (for example pricing of a house, number of people, HUE values of an image) and the according label what is presented. As in the softmax the according label is given, the network begins to alter the weights slightly (in a step-size as big as the learning rate) in favor of receiving a high value in the softmax neuron with the given label. This redistributes the weights all the way back to the first layer. This process is called back-propagation in supervised learning.

The data is then presented to the network and when all individual data points have been given, an epoch has passed. While this process takes place, a small batch of data points not yet presented are sampled to test the accuracy of the prediction the network as a measurement to see how the network progresses. This is important as training can take hundreds of hours or even weeks even with modern GPUs to be fully trained and early experimentation altering hyperparameters (= global variables such as the learning rate), the network depth (= how many layers), or network width (= how many neurons per layer) can take place to lead to better results early on.

Ultimately, when the accuracy of prediction is high enough, a final bigger test-set is used to determine the overall accuracy of the network in making predictions.

2.2. Large Scale Visual Recognition Challenge (ILSVRC)

The ImageNet Large Scale Visual Recognition Challenge (ILSVRC) is a yearly competition by ImageNet, in which each participant presents a predictor-machine that outputs an array of what is most likely seen in the presented pictures. The most likely predicted object and the first 5 most likely predicted objects are then compared to the real label and for each wrong classification, the Top-1- and Top-5-Error-Rate are increased. The predictor with the lowest error-rate wins.

2.2.1. Goal

The specific goal of the ImageNet Classification Challenge is to establish under standardized conditions a benchmark to measure the progression of the scientific community regarding computer vision and specifically object detection.

2.2.2. Challenge Progression

The ILSVRC started 2010 and the winning team was a team using a support vector machine. The Top-5-Error Rate was at 28%, which means that at that time the best approach could only give a top-5 estimate of the correct label in 72% of the images. With such results, most similar complex problems concerning computer vision still seemed out of grasp of real world application such as identifying hundreds of street signs for self-driving cars with a high accuracy. Over the years the Top-5-Error-Rate steadily decreased with a significant drop in

2012. End of 2015 the error-rate dropped under the benchmark error-rate of 5%, which has been established by a small team of researchers as the error-rate for a human familiar with the dataset.

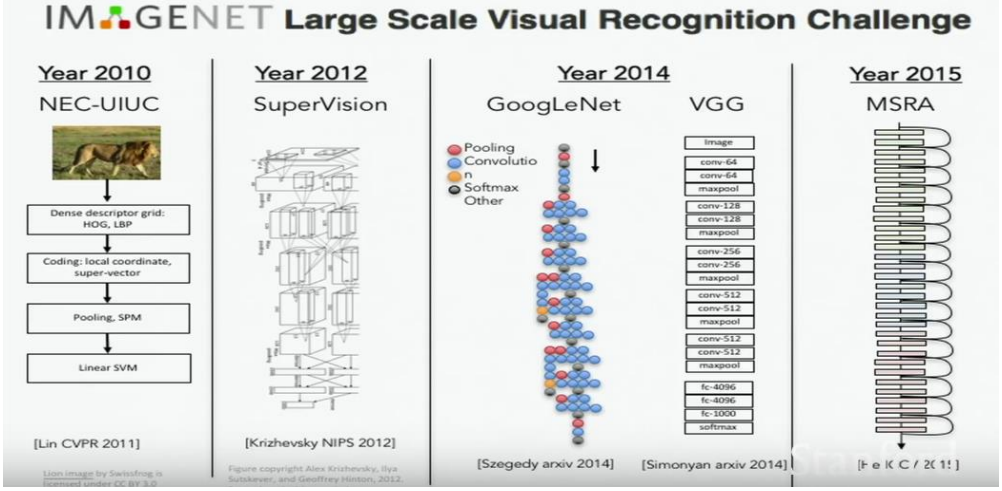


Figure 2**Error! Use the Home tab to apply 0 to the text that you want to appear here.**) ILSVRC winning teams over the years. 2010 a support vector machine won. From 2012 onwards only CNNs won.

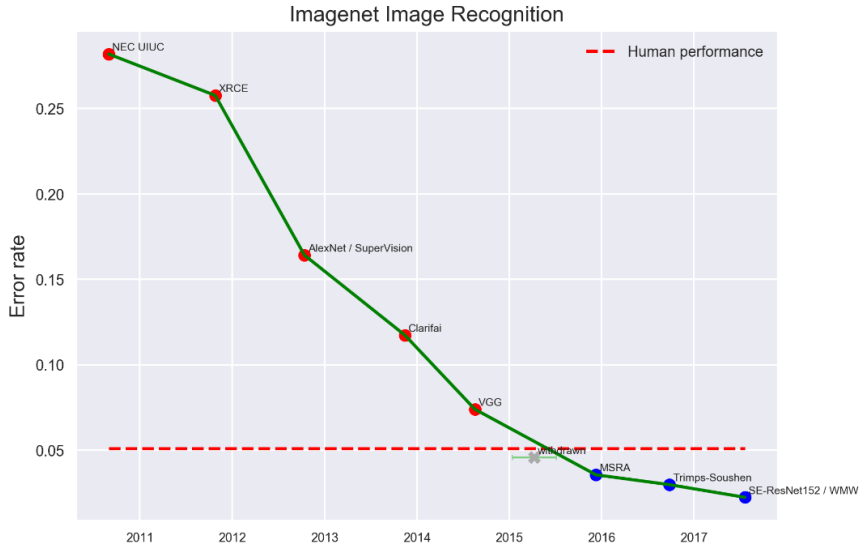


Figure 3) ILSVRC winners Top-5-Error-Rate over the years. The error rate dropped significantly over the years and even fell end 2015 below humans Top-5-Error-Rate.

2.2.2.1. Significance of 2012

2012, the CNN based network SuperVision reduced the Top-5-Error-Rate significantly from 26% to roughly 16%. Prior to this result, Support Vector Machines (=SVNs) have been the best at predicting labels for imagery and it was estimated that the improvement of SVNs would lead to the best predictors with slowly decreasing error-rates over the years as seen from 2010 to 2011. Yet 2012, a totally different approach which was yet not in the focus of main-stream science, won the competition and increased the interest about this emerging technology, which at that time still held a lot of undiscovered potential, while also reducing the base line of improvement to 16% as the Top-5-Error-Rate. Therefore 2012 became a stepping stone for computer vision.

2.2.3. The Data

The Data provided by ImageNet consist of 1000 categories with roughly 1000 images each. Additionally, to the 1.000.000 pictures provided, another 200.000 pictures were used as the validation set at the competition to calculate the error-rates.

The categories were selected by scientist to show typical flaws of networks, by for example including very similar categories as Australian terrier and Yorkshire terrier or using close-up pictures and zoomed-in pictures for the same label.

Figure 4) Left, 3 pictures of the flower toadflax are seen. The dataset includes these 3 pictures with the same level independently of their distance to the actual flower. The middle 2 pictures include a Yorkshire terrier, and the 2 pictures to the right include Australian terriers. The network must be able to differentiate between both dog breeds.

2.2.4. Issues

There is certain controversy about certain labels, which even when asking humans did lead to issues in labeling the imagery.

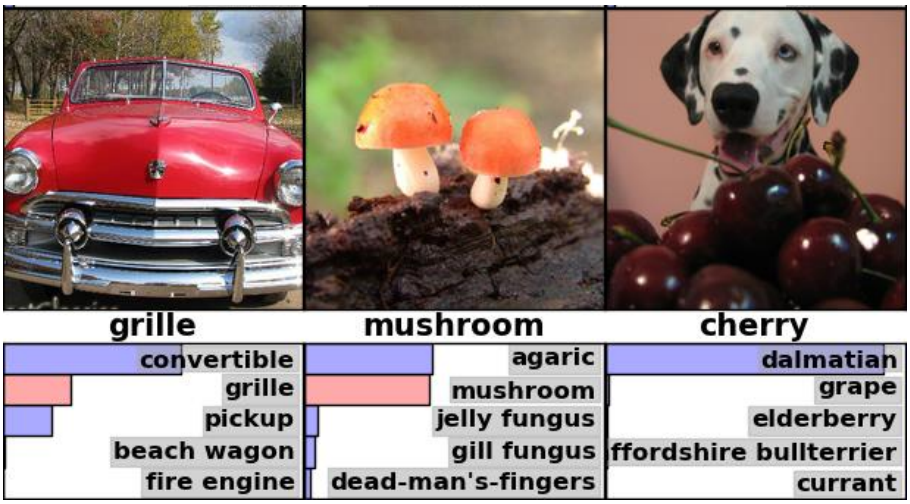


Figure 5) Here 3 pictures are given as input to the network and the probability of what is seen is hierarchically shown below. The label determined by humans are directly beneath the picture. On the left, a convertible is predicted but the human chosen label focused on the grille. The middle picture shows an agaric, a kind of mushroom, which is labeled by humans as mushroom. Here the network became overly specific in its prediction. The picture to the right has cherries and a Dalmatian in it. As most people focused on the cherries when labeling this picture, the predicted Dalmatian and the next 4 probabilities also ignoring the cherries, lead to an increase in the Top-1- and Top-5-Error.

Issues surrounding such grey areas are being solved by the community after each competition, but some reside unresolved as certain grey areas are intentionally and necessarily implemented to determine the outer bounds of the networks. The tasks become slightly more complex over the years by diversifying imagery. There is also a debate going on about including abstract words such as “labor” or “beauty” into the dataset.

2.3. Convolutional Neural Networks (CNNs)

CNNs are the focus of this paper and begin to increase in importance in computer vision. Here, we are going to look at what makes CNNs so special.

2.3.1. Structure of CNNs

The base structure of a CNN is pretty much the same to typical feed-forward networks but have differences normally in the first few layers right after the input neurons. Convolution describes the summarization of a highly complex web of connections fusing more and more together as the layers of the network progresses. The middle and end-part of these networks are often similarly or identically structured as other neural networks with a few fully connected layers towards the end and a softmax as the last layer.

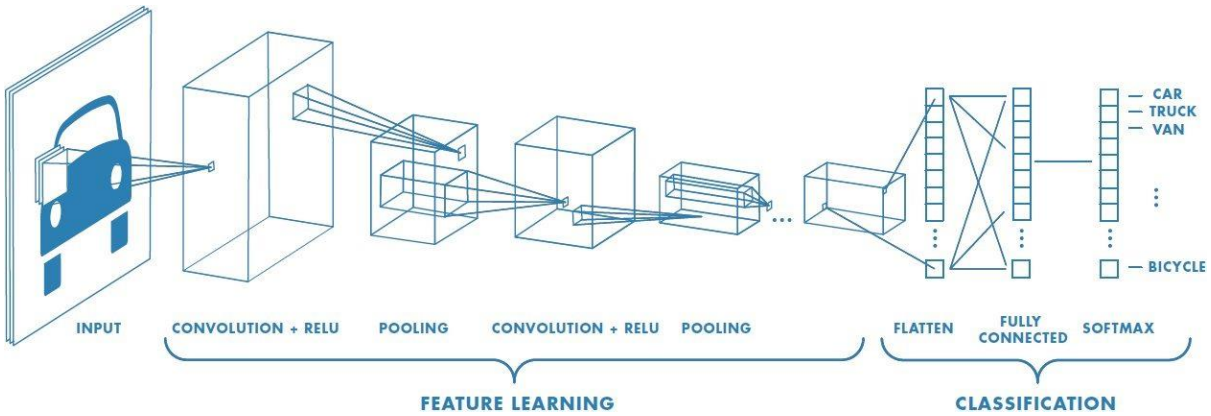


Figure 6) The architecture of a CNN. The neurons in the first layer get access to a small part of the input imagery. The layers then alternate between convolutional layers and pooling layers. The last few layers are fully connected, followed by softmax decision layer.

In the first convolutional layer, each neuron is only connected to a part of the field which holds the input parameters. This results into a single neuron in the first layer to only "see" part of the entire picture, which forces it specifically in classification tasks to abstract as much information as possible from this input field. This has proven to lead to better results specifically dealing with pictures where generalization becomes a priority. Each input consists of 256x256 pixels (in this network but also used generally) which results into 65.536 individual monochromatic values multiplied by the 3 colors red, green and blue per picture. Now multiplying this number times 1000 categories with each about 1000 pictures, gives a total of 65.536.000.000 in which it must find individual patterns of information that lead a network to stimulate the correct category inside of the softmax layer (= decision layer).

These networks need a lot of processing time in adjusting their weights with every new input picture provided. As smaller training time becomes significantly more important dealing with such large network, the activation function of these networks need to be easy to compute. Which is why, Rectified Linear Unit (= ReLU) has become the activation function of choice. It works very straight forward by outputting the highest positive value from the input or 0, whichever is a more positive number.

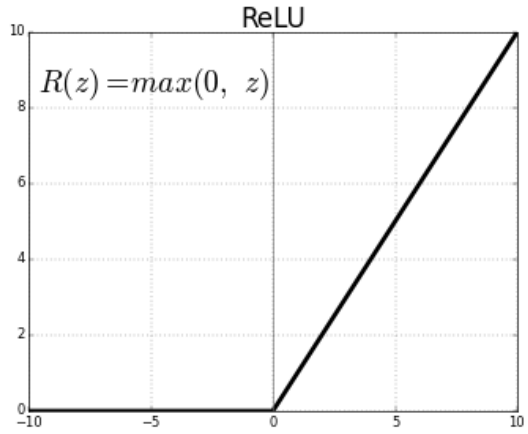


Figure 7) ReLU-activation function. On the x axis is the input value and on the y axis is the returned output value.

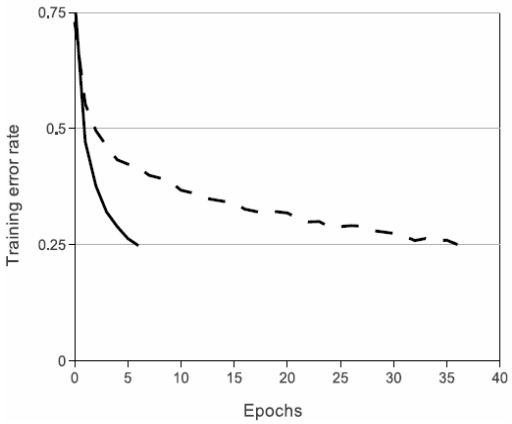


Figure 8) Both lines show the decline in training error rate over epochs. The dotted line shows the network with tanh activation function. The solid line shows the network with ReLU activation function.

Between each convolutional layer, there are additional pooling layers, which summarize and filters the responses. Pooling has as an input a small region of the previous layer and is therefore like the convolutional layers, yet it reduces the total amount of passed on neurons.

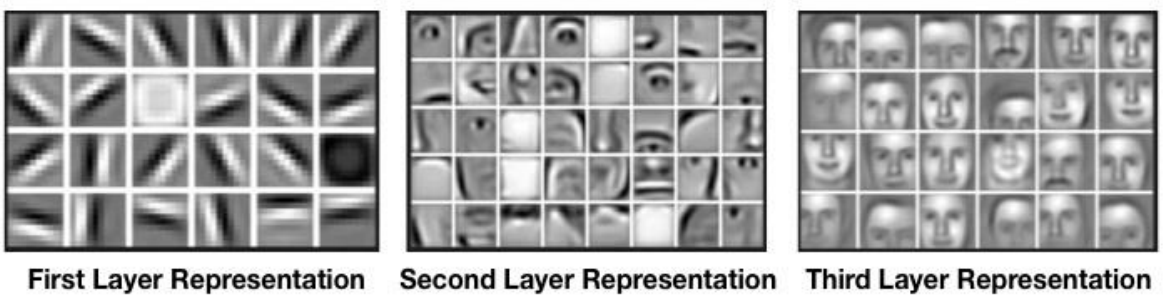


Figure 10) Left, the visualized output of the first layer of a convolutional neural network is shown. In the second layer (middle) the simple features are summed up and the network begins to generate more complex features such as ears, eyes and noses. In the third layer the features show the highest complexity which are shortly before the softmax layer, which is used to differentiate between faces.

2.3.2. Advantages of CNNs

Since CNNs have taken over as the predominantly winning model in the ILSRVC. Why is this the case? First, a pictures label is often unrelated to the exact positioning of the object of interest and in convolutional neural network the position becomes significantly less relevant as the first layers do not focus on position. Second, CNNs have the big advantage to generalize way better than fully connected networks early on which makes the first few layers of such networks often very similar to each other independently of what category it tries to classify. This has been figured out quite early on and taken advantage of in transfer learning. Transfer learning takes a fully trained network and only retrains the last few layers while leaving the weights in the frontal network unaltered. This becomes very handy for smaller projects in computer vision, with people not having the resources, time, experience with CNNs, or hardware to train classifiers to a high degree. Therefore, smaller groups can build CNNs based on already existing world-class networks such as Inception by Google (Google 2018).

2.3.3. Disadvantages of CNNs

So why aren't CNNs good for everything? CNNs tend to generalize better but also require even bigger amounts of data. Linear regression requires the least amount of data to lead to quite well results for low complexity problems. Machine Learning approaches such as SVN require already a significant increase in data to function properly. Neural Networks require often by a tenfold higher amount of data and this number increases with CNNs, especially if they become deeper. With too little data CNNs lead to bad results.

2.4. Alex Krizhevsky et al's Neural Network

The network supervision had a complex creation process based on many experiments with CNNs. For a CNN a lot of data needs to be given as an input and the amount of individual data points increases with each additional category. Training the network therefore required an even larger amount then the 1 million provided images. This was partly done by data augmentation which lead to a standardized input and 2048 times more input images. The main hurdle of such complex networks becomes the generalization aspect. Each added category and image can lead to overfitting (= the network adapting so well that it only leads to increased accuracy in prediction of already seen data). The team came up with different techniques to increase the generalization skill (= high accuracy on prediction of yet unseen pictures) of the network.

2.4.1. Data Augmentation

The presented pictures of ImageNet are in all shapes and sizes, with individual resolutions and formats. First, the image was cut down to a squared frame from the center of the image. This image furthermore then was reduced in resolution to 256x256-pixel image for standardization purpose. To increase the total number of individual pictures, the picture was sub-sampled and 224x224-pixel images cut from the 256x256 imagery, resulting in 1024 new pictures per original picture with a minor shapeshift which also helped to train the network more generally about the position of the object. These pictures were then reflected horizontally to double that amount leading to a total of 2048 new pictures generated from one original. Lastly, the mean RGB-value was calculated for each position and subtracted from each image to normalize the values (benefits networks learning).

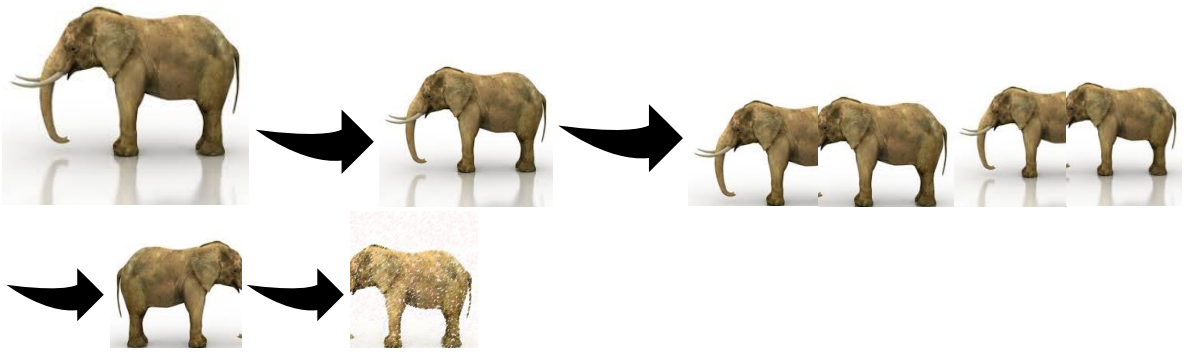


Fig 11) Left the original picture, reduced in resolution to a 256x256-pixel image and then sub-sampled into 1024 new 224x224-pixel images. Left the subsampled image is additionally flipped to double the amount of data-points. Lastly, the mean RGB-value of all images is subtracted from the individual sub-sampled image.

2.4.2. General architecture

The presented network from the paper (Alex Krizhevsky and Sutskever 2012) is a CNN with 650.000 individual neurons, 7 layers + softmax, 60.000.0000 individual parameters and over 630.000.000 connections. The first 5 layers are convolutional layers alternating with pooling layers with 2 final fully connected layers and a softmax layer for the final categorization task.

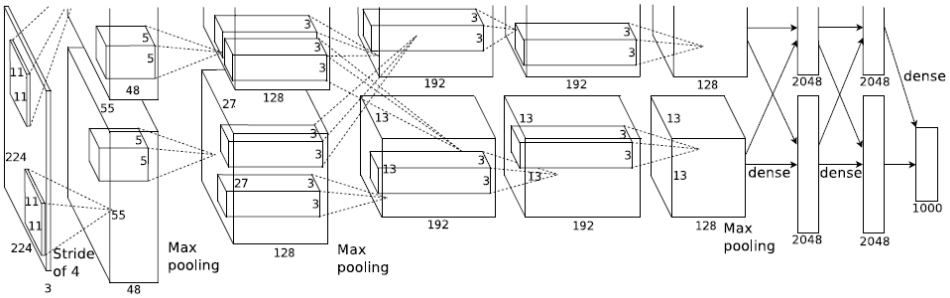


Figure 12) This is the base structure of the CNN SuperVision which won 2012s ILSVRC. The network is split onto 2 GPUs which are only connected between the second and third layer and the 5th, 6th and 7th layer. The input image is reduced in quality and subsampled to a 224x224-pixel tensor with a stride of 11x11-pixel. Layers 1, 3, 4 are convolutional layers. Layer 1, 2 and 5 are pooling layers. Layer 6 and 7 are fully connected layers followed by a softmax layer.

2.4.3. Specialties

The community working with CNNs ran into multiple issues regarding the too low volume of data available for real world applications and/or had issues in overfitting. With the 1 million provided pictures and the image augmentation the first issue became significantly smaller, but overfitting was still an issue which has been addressed with the following specialties.

2.4.3.1. Split Network

Through continuous experimentation with the size of the neural network (adding, removing layers; increasing width and depth) the team concluded that the networks performance would be better if it expanded outside of the memory range provided by a single GPU. Two GTX 580 3GB GPUs were therefore used to train the network. Half of each layer was placed on each of the GPUs. Originally, the intent was to fully connect each layer with the next, so that the split would be only on the hardware part and not noticeable in the interaction of the network. But they discovered that separating individual parts of the network (especially in the middle layer), improved performance. Layer 3, 4 and 5 were therefore separated and half of each layer did not become influenced by the other half of the layer.

Visualizing the fully trained network by inputting random scattered grey values and outputting the results after each individual layer shows the internal process of each layer. The network itself began to separate not only in hardware and structure but also began to separate in functionality into 2 separate parts. After training, one half of the layer 3 to 5 processed mostly color while the other half of layer 3 to 5 processed mainly saturation and angle. This functional split lead to an increase of accuracy and a reduction of overfitting. Later, the entire network was multiple times retrained and the split reoccurred over time.

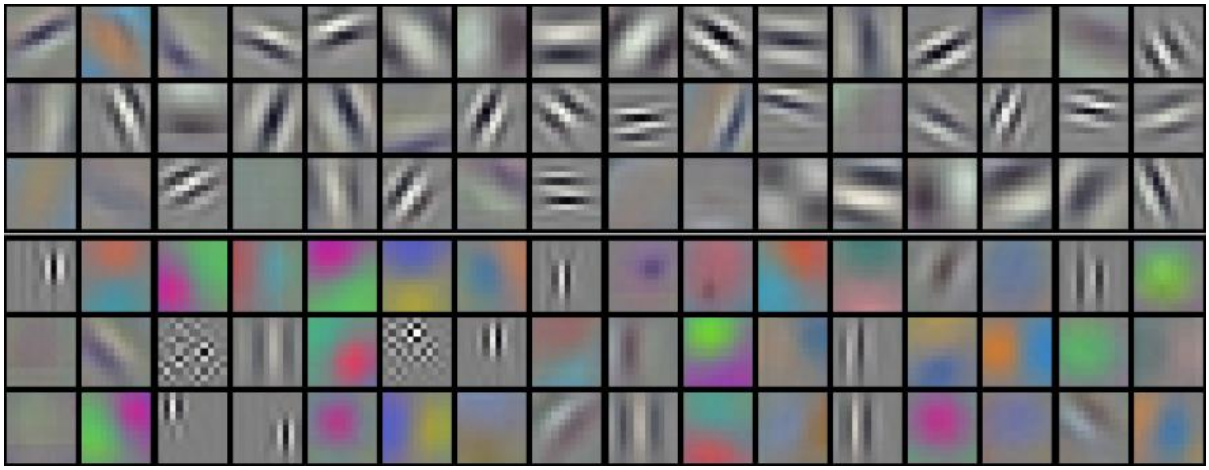


Figure 13) Each individual square represents the output of the fifth pooling layer of the network when 4 neurons project onto a white canvas. The top 3 rows represent the first half of the fifth layer and the lower 3 rows show the output of the second half of the fifth layer. The top rows are dominated by black and white and general line direction while the bottom layers clearly focus on color variation.

2.4.3.2. ReLU and pooling layers

The choice of using a ReLU- function for neural networks seemed at that time controversial as there was little empirical evidence supporting that this truly lead to better results. But as ReLU showed that it leads to significantly reduced training time and time being a factor, it became the activation function of choice. As ReLU only works with positive values and basically ignores negative values and passes on each positive value, it basically just passes information and turning each negative value into a 0.

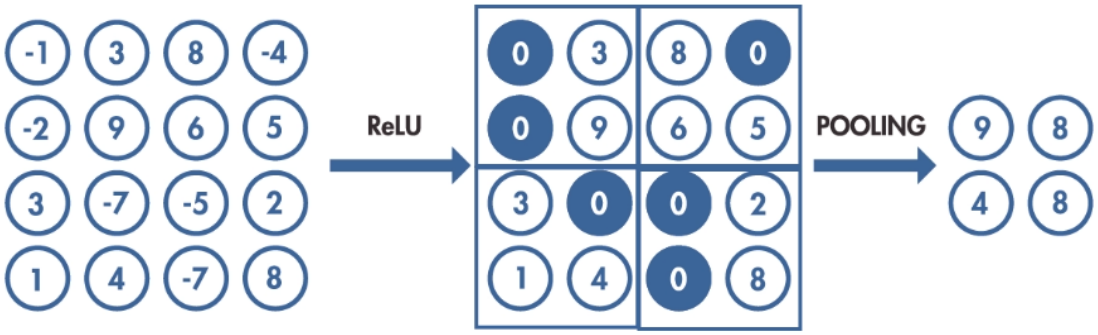


Figure 14) On the left, the normalized brightness values are shown. Through the ReLU function, each negative value is turned into a 0. Then in pooling layer, 4 next to each other neurons are seen as input and their maximum value is passed on as new value.

Yet through the addition of pooling layers, in which a stride (= squared convolution of the previous layer as input) was reduced to a smaller layer, had the effect of creating a neural network in which each neuron competed in each layer against its surrounding neurons to pass on the information into the next layer. This selection process passed on the strongest signals into the deeper layers of the network. Additionally, pooling reduced the overall information flow, reducing complexity and increasing the generalization capabilities of the network.

2.4.3.3. Local Response Normalization

The local competitive character of the neurons was then globalized over each layer by implementing Local Response Normalization. The average input values were calculated for each layer and imbedded into the local response by increasing the output for above average firing neurons and decreasing the output for below average firing neurons.

$$b^i_{x,y} = a^i_{x,y} / \left(k + \alpha \sum_{j=\max(0,i-n/2)}^{\min(N-1,i+n/2)} (a^j_{x,y})^2 \right)^{\beta}$$

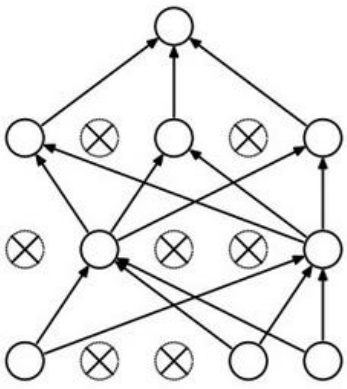
Fig 15) Formula for Local Response Normalization. k, n, α, β are hyperparameters set globally and set by experimentation with a validation set to k= 2, n= 5, α=0.0001 and β=0.75. b is the normalized response of the neuron and a is the unnormalized response value. x and y indicate the position of the neuron inside of the network and N is the total number of neurons inside of this layer.

Local Response Normalization increased overall accuracy and increased generalization.

2.4.3.4. Dropout

Dropout is a technique in which before each epoch, a part of the network is shut down. At the beginning of each epoch, a random number generated determines if in this epoch this individual neuron will be responsive or not. On the first 2 layers of this network dropout was applied with a 50% chance of becoming unresponsive. So why render half of 2 layers of neurons non-responsive?

While networks are trained, often singular features in the picture begin to dominate as a signal and as in the pooling layers, non-dominant information becomes filtered out, second and third degrees important signals can be filtered out as well. To avoid this loss of eventually valuable information, the dominating factors which summarize themselves in features represented by neurons, are ignored. By doing this, the network will try to back-propagate without being able to focus the weight-alteration on a few singular neurons, but instead begin to even out the network, to generate more complex answers relying on a bigger multitude of individual features generated by the neurons.



(b) After applying dropout.

Fig 16) Each neurons chance to dropout (= becoming unresponsive) is initially set. The network further on in the training must find new ways to generate the answer requested by the back-propagation process.

Obstructing the single model, into basically a new model every new epoch, mathematically means to average the weights of all epochs/ new models into a new meta-model.

As an effect, the network becomes better at generalizing, reduces co-adaption between neurons, initialization becomes less relevant to the network and it becomes more resilient to changes. Dropout is only applied in the training phase and removed during performance.

2.4.4. Results

The final training run took 7 days on 2 GPUs. 2012, the committed result achieved a Top-1 Error-Rate of 26,7% and a Top-5-Error-Rate of 15,3%. By these measures, SuperVision won 2012 the ILSVRC.

Furthermore, the team went into specifics regarding the quality of their predictions. The activation of the softmax category to predict the image is an absolute value. By looking for images that have the next higher absolute score in the corresponding category, the most closely related image can be found. The distance between these 2 values is called Euclidean distance.

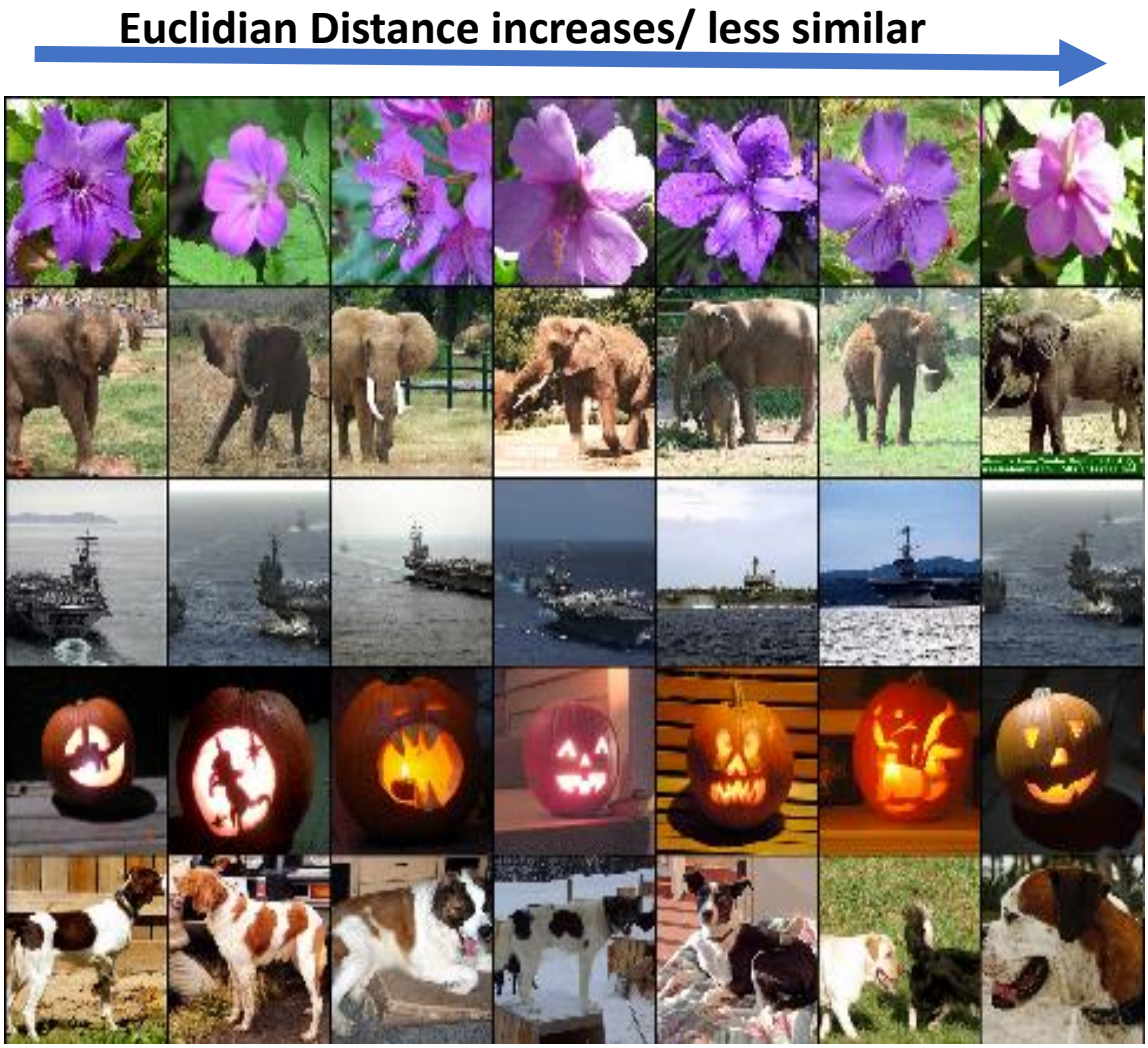


Fig 17) The left picture is the picture with the highest score for the specific category. The pictures right of it are the most similar pictures or the pictures with the lowest Euclidean distance to the first picture.

The similarity appears to be independent of shading, background, angle, entirety of the object or even specific motif (notice the pumpkins massive distance to each other). This indicates that the network slowly began to “understand” the important parts of the object, which give the object its name.

3.0. Summary

Summarizing, computer vision has already done a great leap 2012 into becoming a practical solution for a wide set of problems and progressed ever since to a degree of sophistication through new combination of techniques such as CNN with dropout, network splitting, intralayer competition and increased depth of the network. Computer vision and neural networks harbor immense potential in a multitude of services.

4.0 References

Alex Krizhevsky and Sutskever, Ilya and Hinton, Geoffrey E. 2012. "ImageNet Classification with Deep Convolutional Neural Networks." (Curran Associates, Inc.).

Alexander Toshev, Ameesh Makadia, Kostas Daniilidis. 2009. "Shape-based Object Recognition in Videos Using 3D Synthetic Object Models ." *Computer Vision and Pattern Recognition.*

Álvarez Menéndez, .de Cos Juez, Sánchez Lasheras, .Álvarez Riesgod. 2010. "Artificial neural networks applied to cancer detection in a breast screening programme." In *Mathematical and Computer Modelling*, Volume 52, Issues 7–8, October 2010, Pages 983-991.

Google. 2018. *https://github.com/google/inception.* 2 1. Accessed 2 1, 2018. https://github.com/google/inception.

Ivan Bogun, Anelia Angelova, Navdeep Jaitly. 2015. "Object Recognition from Short Videos for Robotic Perception ." *CoRR, vol. abs/1509.01602.*

n.d. *kaggle.* Accessed 3 1, 2018. http://kaggle.com.

Kayla Backes, Alex Polacco. 2017. "THE AMAZON GO CONCEPT: IMPLICATIONS, APPLICATIONS, AND ."

Lex Fridman, Benedikt Jenik, Bryan Reimer. 2017. "Arguing Machines: Perception-Control System Redundancy."

Lubor Ladicky, SoHyeon Jeong, Markus Gross, Marc Pollefeys, Barbara Solenthaler. 2016. "Data-driven Fluid Simulations using Regression Forests."

N. Pinto, D.D. Cox, and J.J. DiCarlo. 2008. "Why is real-world visual object recognition hard? ." *PLoS computational.*

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, Piotr Dollar. 2017. "Focal Loss for Dense Object Detection." *International Conference on Computer Vision* .

Xu, N., Price, B., Cohen, S., Yang, J., Huang, T. 2017. "Deep GrabCut for Object Selection." *British Machine Vision Conference* .

Yi, L., Guibas, L., Hertzmann, A., Kim, V., Su, H., Yumer, E. 2017. "Learning Hierarchical Shape Segmentation and Labeling from Online Repositories." *SIGGRAPH.*

Websites

- http://image-net.org
- http://mathworks.com
- http://stanford.edu

Books

Deep Learning – by Ian Goodfellow, Yoshua Bengio, Aaron Courville

Pictures

- *Olga Russakovsky et al 2015*
- *http://www.image-net.org*
- *www.mathworks.com/*
- *www.free3d.com*
- *www.eff.org*