Research Seminar in Data Science
Paper Reviews

Christoph Schmidl
s4226887
c.schmidl@student.ru.nl

June 9, 2018

Reviewed papers:

- **Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images** by Anh Nguyen et al. (Presented by Douwe van Erp)
- **Listen, Attend and Spell** by William Chan et al. (Presented by Lars Kuijpers)

1 Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images

1.1 Summary

In the paper "Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images" Anh Nguyen et al describe the fact that discriminative deep neural networks predict certain classes with high confidence based on images which are completely unrecognizable to humans. This effect is based on the assumption that these "fake" images contain certain patterns which have been learned in the first layers of the networks to produce these fake images. The authors present three different approaches to generate these fooling images, namely an evolutionary algorithm which is using direct encoding, an evolutionary alogorithm which is using indirect encoding (compositional pattern-producing network - CPPN) and gradient ascent. These three approaches are used in conjunction with two different datasets, namely MNIST and ImageNet and are used to generate fooling images which the corresponding DNN should classify with high confidence. The authors show that discriminative deep neural networks are prone to classify fooling images with high confidence and tried to overcome this problem by changing network architectures, combining architecturse and adding fooling images as a n + 1 class to the original dataset although the last approach only worked a little on the MNIST dataset. The paper showed that computer vision and human vision still differ alot from each other and that there's room for exploitation.

1.2 Evidence

The authors cite 32 different research papers and seem to base their work on relevant prior work like "Intriduing properties of neural networks" by C. Szegedy et al. Based on the reference list you can see that there are numerous papers which have been written by well-known authors like G.E. Hinton, I.J. Goodfellow and Y. LeCun which is giving this paper more credibility. Classical papers like "Imagenet classification withh deep convolutional neural networks" by A. Krizhevsky are also included.

1.3 Strengths

The paper is written in a clear way and the experimental setup is easy to follow. Using three different approaches for generating the fooling images on two different but well-know datasets seems to make alot of sense to me and makes it easy to reproduce the results. The supplementary material is also a big plus which seems to fill the gap in terms of details if the reader is more interested into that. Discovering the fact that deep neural networks can be tricked to "recognize" certain images with high confidence although they look nothing like the training samples when it comes to human perception, is the main reason to read this paper and was an interesting read for me.

1.4 Weaknesses

I do not know why the authors mention the fact that the generated, CPPN "fooling" images have been submitted to an art contest. It seems to be out of scope of this paper and has no further contribution to the main point of the paper.

1.5 Evaluation

I guess this paper would get accepted to a conference or journal because it reveals an interesting fact, namely that deep neural networks can be fooled by synthetic images which do not look like any sample from the training set or like a natural images. This could lead to lethal surprises if you think about safety with regards to self-driving cars and computer vision.

1.6 Comments on the quality of the writing

The writing style was formal but also written in a clear way. I liked the division of subsections and their clear subtitles. The division made it easy for me to get a first, quick overview of the paper and then go into the details.

1.7 Queries for discussion

- The authors concentrate on "fooling" discriminative models and state that generative models would be harder to fool. Have there been any advances in generative models to overcome the fact that these models do not scale well to high-dimensionality in order to test them as well?

2 Listen, Attend and Spell

2.1 Summary

Kaiming He et al address the degradation problem by introducing a deep residual learning framework. The degradation problem occurs with increased network depth. When the network starts converging the accuracy gets saturated and then degrades rapidly. In contrast to a plain network architectures like VGG16, a deep residual network is able to maintain a relatively high accuary even when the number of layers varies from 100 to 1000. Kaiming He et Hal hypothesize that it is easier to optimize the residual mapping than to optimize the original, unreferenced mapping. Residual learning is realized by so called "shortcut connections". They evaluate their results based on experiments with the imagenet and cifar-10 datasets which are about image classification and the PASCAL and MS COCO dataaset which are about object detection and object segmentation. Residual networks seem to yield better performance on all datasets when the number of layers start increasing into the hundreds.

2.2 Evidence

The authors cite 49 different research papers to back up their paper. Among the different citations are well knoen researchers like C.M. Bishop, I.J. Goodfellow, G.E. Hinton, Y. LeCun and J. Schmidhuber. Out of these 49 papers, only 3 are also published by K. He himself. The papers seem to be well picked and show the problems of training networks with increased depths and already known solutions on how to tackle these problems. Based on the fact that none of the cited papers includes the name "residual", I assume that this paper is the first one which came up with residual network structures.

2.3 Strengths

The authors explain in detail what the main problem is when it comes to training very deep neural networks. The degradation effect seems to be of main interest and the residual network structure seems to be a novel approach. In the section about "shortcut connections" the authors also mention that there seems to be a similar approach to their shortcut connection implementation which has been developed during the same time span by Schmidhuber. Schmidhuber's idea is called "highway networks". The authors state out that the main differences between their idea and the one proposed by Schmidhuber. I guess that was done because Schmidhuber is known for claiming credit for ideas which seem similar to his own but differ slightly and is known for confronting researcher publicly during conferences to explain themselves.

2.4 Weaknesses

I could not find any real discussion, conclusion or future work section. This made it kind of frustraing for me to read because I wanted to start to follow the main three-step approach of reading research papers.

2.5 Evaluation

I assume that this paper was accepted for a conference or journal because it offers a novel idea which seems promising. Mainting the same amount of hyperparamters as so called "plain networks" but being able make the network way deeper at the same time and keeping a high accuracy seems to be a very tempting idea to try out. Given the fact that the authors also won different competitons with their approach is another reason why this paper is interesting. Well known datasets like imagenet, cifar-10 and coco have been chosen to validate the approach which makes it easy to reproduce.

2.6 Comments on the quality of the writing

The technical detail seems to be sound and it is well written when you overlook the fact that the discussion, conclusion and future work section is missing.

2.7 Queries for discussion

- The authors state that they did not use maxout/dropout for their networks with over 1000 layers although they assume that this may improve their results. Has there been further work on this assumption?