# Research Seminar in Data Science
Paper Reviews

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

June 10, 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

William Chan et al propose a two-component neural network called "Listen, Attend and Spell (LAS)" that can lean to transcribe speech utterances to characters. The authors are using a two-component approach where a so called "listener" is a pyramidal recurrent neural network encoder and a so called "speller" which represents an attention based recurrent neural netword decoder. The novel idea behind this approach is the fact das LAS does not make independence assumptions in the label sequence and it does not rely on hidden markov models. The authors tested their network on three million Google voice search utterances. The state-of-the-art model on this type of dataset was a so called "CLDNN-HMM" system with a word error rate (WER) of 8.0%. LAS achieves 14.1% WER without a dictionary or a language model and later on 10.3% WER with LAS + Sampling + LM Rescoring.

### 2.2 Evidence

The authors cite 38 different research papers to back up their paper. Among the different citations are well known researchers like Ilya Sutskever, Alex Krizhevsky, Geoffrey Hinton, Juergen Schmidhuber and Andrew Ng. Most of the papers handle the topic of speech recognition using recurrent neural networks and sequence to sequence learning. Without looking any further into the content of the cited papers I assume that the proposed approach is based on a solid foundation.

### 2.3 Strengths

I guess the main strenghts of the paper is its novelty based on the fact that their approach does not use the concepts of phonemes and it does not rely on pronunciation dictionaries or hidden markov models.

### 2.4 Weaknesses

I could not find any weaknesses regarding the content or the writing style of this paper. It requires some prior natural language processing knowledge and knowledge about search algorithms like beam search but I guess that is not really a weakness.

### 2.5 Evaluation

I assume that this paper was accepted for a conference or journal because it offers a novel idea which seems to have potential to beat the state-of-the-art approaches with fine tuning in the future.

## 2.6 Comments on the quality of the writing

The technical detail is sufficient and the writing style is concise. Mathematical background was covered sufficiently.

## 2.7 Queries for discussion

- Towards the end of section 4 ”Experiments”, the authors state that ”convolutional filters could lead to improved results, as they have been reported to improve performance by 5% relative WER on clean speech and 7% relative on noisy speech compared to non-convolutional architectures”. Did the authors do further investigation on this statement?