InFoMM
Industrially Focused Mathematical Modelling

UNIVERSITY OF OXFORD

Sparse and Low-Rank Regularisation for Robust Deep Nets

Constantin Octavian Puiu

Deep Neural Networks (DNNs) are mathematical models of an ensemble of neurons. DNNs are obtained by stacking multiple layers of neurons, where the information typically flows in one direction (see Figure 1). The main property that DNNs try to achieve is having a "correct" output for a previously unseen input, after initially learning how to classify inputs using a set of training data. DNNs are used for classification tasks (eg. distinguishing between different objects given two pictures) or regression tasks (eg. predicting the stock price movement). We focus on classification tasks, in particular image classification.

DNNs achieve state-of-art performance by learning highly complex maps from the input space to the output space, using a large amount of training data. However, in order to learn such maps, DNN models tend to have many more parameters per unit of data than other Machine Learning Models. This causes the DNNs to be excessively sensitive to small, but well chosen perturbations in the input. Adversaries can find and exploit such perturbations to cause clandestine misclassification for any input sample, without notably altering it. We call such carefully perturbed samples "*adversarial samples*", and the process of creating these samples is referred to as "*adversarial attacks*". We refer to the DNN's ability to withstand such attacks as "*robustness*". National Physical Laboratory (NPL) are interested in understanding the robustness of DNNs. Since intuition suggests that DNN models are vulnerable because they are over-parameterised, our aim is to investigate the effect on robustness of two different methods which reduce over-parametrisation: Sparse and Low-Rank Regularisation. We seek to answer the question:

- Is any sparse or low-rank DNN model more robust than its non-sparse or full-rank counterpart?

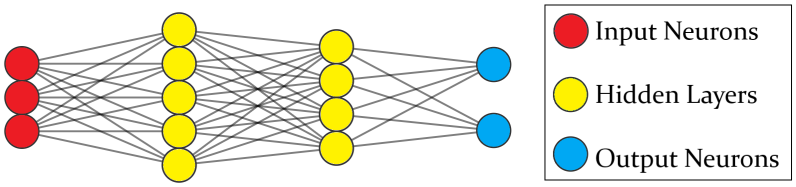

Figure 1 – Example of Deep Neural Net Architecture.

Robustness and Adversarial Attacks

What are adversarial attacks? We present an example in Figure 2. The adversary takes a legitimate input, perturbs it by a small well-chosen amount, and obtains an adversarial sample which is miss-classified as a Chainsaw by a particular DNN, even though the picture clearly shows a dog.

Legitimate Input Image, Classified as Dog

Carefully Selected Adversarial Perturbation (Intensity Magnified by a Factor of x 100)

Adversarial Image, Classified as Chainsaw

Figure 2 – Concept of Adversarial Attack.

UKRI Engineering and Physical Sciences Research Council

In Figure 3, we show the process of generating an adversarial sample in more detail. Adversary samples tend to be consistent across different architectures, training subsets, and hyper-parameters. Thus, we assume that the adversary knows the architecture of the DNN being attacked. To understand adversarial attacks, we have to think of each input as being a vector with a number of dimensions equal to the number of pixels. Thus, any perturbation of the input is also a vector, having a direction (the attack direction) and a magnitude. The goal of the adversary is then to find the perturbation vector with minimal magnitude that causes misclasification. In the Fast Gradient Sign Method (FGSM), the first step in generating an adversarial sample is to fix the attack direction. Using the benign target sample and the architecture of the DNN, the adversary estimates the attack direction that locally causes the largest change in the DNN output. Having selected this sensitive direction, a magnitude is then chosen, and a perturbation generated. This perturbation is added to the benign image in the hope to obtain a misclassification. If the perturbation was not enough to get the legitimate input misclassified, we repeat the process until we succeed. A local robustness measure to FGSM attacks can simply be given by the smallest distance along the FGSM direction which is needed to misclassify a given input. This can be made global by considering more samples.

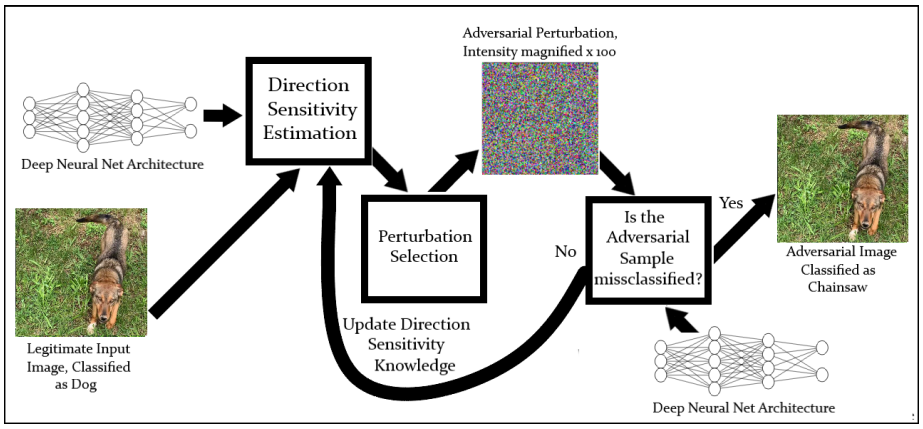

Figure 3 – Generating Successful Adversarial Attacks.

Sparse Networks and Low-Rank Networks

The idea of sparse DNNs is shown in Figure 4. A Sparse DNN is a DNN where a number of connections are removed. A Low Rank regularized DNN can be thought of as a network where the output of each layer is forced to live in a much smaller dimensional space than it otherwise would. Limited numerical experiments in the literature (with no theoretical proof) may suggest that relatively robust models are simultaneously sparse and low-rank. However, this is not to say that the most robust model is necessarily sparse or low-rank. In the other direction, we want to see if sparsity/low-rank also implies robustness.

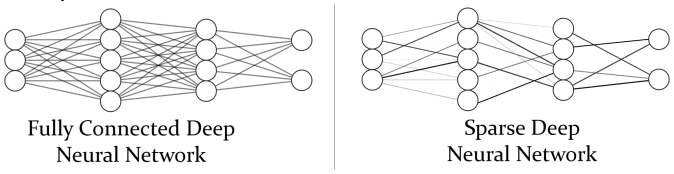

Figure 4 – The Concept of Sparse Deep Nets.

We show theoretically that for Linear Classifiers (which can be thought of as ancestors of DNNs), the most widely used technique which induces sparsity can lead to a degradation in robustness to FGSM attacks. In the case of DNNs, we have shown that we cannot conclude, based on the available theory, that sparse DNNs are always more robust. Our comprehensive numerical results show that sparsity can damage the robustness of DNNs. Thus, we conclude that while the most robust DNN model may be sparse (remains to be proved or disproved), there exist sparsification techniques which do not enhance robustness in DNNs, and some can cause more damage than good. In the case of low-rank, we show numerically that it is possible to have low-rank inducing methods which marginally reduce robustness, thus drawing analogous conclusions.

Dr Andrew Thompson, *Senior Research Scientist*, at NPL said:

> *"NPL are increasingly advising clients on how to make their ML trustworthy, and robustness is one important aspect of this. Constantin's project has provided valuable insight into which methods are successful in making deep networks more robust."*