Certifying Properties of Deep Neural Networks by Taking them into Shallow Waters

Konstantinos Barmpas – Master Thesis Project Poster
ETH Zürich (Swiss Federal Institute of Technology) | Data Analytics Lab | Imperial College London | Daedelean AI

ABSTRACT

Over the last decades, complex deep neural networks have revolutionized Artificial Intelligence (AI) research. These models can now achieve impressive performances on various complex tasks like recognition, detection and image semantic segmentation, achieving accuracy close to, or even better, than human perception. However, these neural networks require to be both deep and complex and this complexity constitutes a danger for the safety verification (certification) and interpretability of a neural network model.

This project explores the certification properties of complex neural networks by taking them into "shallow waters". First, a detailed investigation of efficient model distillation techniques is conducted. Then, using the shallow models trained with these distillation methods, several of their properties are further explored, among them adversarial robustness and their performance under parameter reduction procedures. Finally, by combining network's convex relaxation with model compression, the certification area of shallow student models (derived from either normally or robustly trained teacher networks) is researched. Through these experimental results, it is empirically demonstrated and proved that model distillation leads to shallow models with larger certification areas than their equivalent complex teacher networks. Therefore, based on this thesis evidence, shallow distillated networks constitute a possible solution to the safety and interpretability issues that modern complex Artificial Intelligence (AI) models face.

MODEL DISTILLATION

During the distillation experiments, two deep complex teacher neural networks were trained using the SVHN dataset (with 4 convolutional layers) and the CIFAR-10 dataset (with 6 convolutional layers). A series of distillation experiments were conducted using the techniques described in [2] & [3]. It was found that the shallow student models need to be both deep and convolutional and with many parameters to achieve high performance.

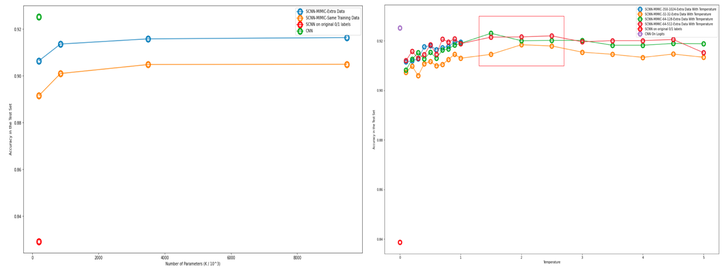

Figure 8.11: Graphical Illustration of SCNN Accuracy Results For Different Temperatures - SVHN Dataset.

Table 8.7: SCNN Experiments Two Convolutional Layers - SVHN

Name	Number of Parameters	Temperature	Temp. Accuracy	Logits Accuracy
SCN-32-32	220k	2.0	91.845%	90.646%
SCN-64-128	850k	1.5	92.306%	91.361%
SCN-64-512	3.5M	2.5	92.202%	91.583%

Table 8.8: SCNN Experiments Two Convolutional Layers - CIFAR10

Name	Number of Parameters	Temperature	Temp. Accuracy	Logits Accuracy
SCN-64-128	650k	4.5	76.570%	73.450%
SCN-64-512	2.5M	2.0	77.140%	75.120%
SCN-128-1024	6M	3.5	76.820%	75.460%

PARAMETER REDUCTION

The next step was to investigate whether the number of parameter can be decreased without affecting the accuracy of the best student distillated models. ”Pruning Filters” method [4] was implemented to the best accuracy student models for each dataset separately.

Table 8.10: Pruning Filters Experiments - CIFAR10 Dataset

Name	Number of Parameters	Accuracy Test Set
CNN	550k	82.560%
SCNN-MIMIC-CIFAR10-h64-512_2.0	2.5M	77.140%
SCNN-MIMIC-CIFAR10-h64-512_2.0_new	880k	77.140%

ADVERSARIAL ROBUSTNESS

In the next part of the experiment, the shallow distillated models (filter pruned) with the best accuracy in both SVHN and CIFAR10 datasets were tested for their out-of-the-box robustness to adversarial non-targeted attacks. Their performance was compared to the performance of their corresponding teacher models under the same series of attacks. It was found that for small ε hence small perturbations, the student model performs better than the teacher models.

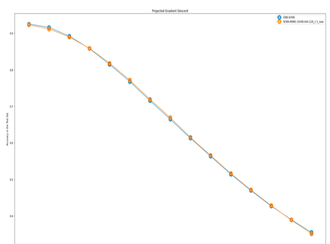

Figure 8.19: Projected Gradient Descent (PGD) Attacks in the SVHN dataset with ε at [0.0 - 0.014]. Orange is the student model after filter pruning and Blue is the teacher model.

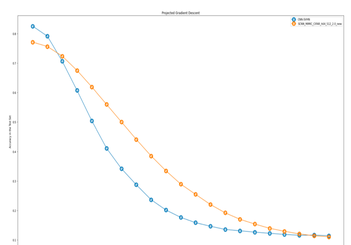

Figure 8.22: Projected Gradient Descent (PGD) Attacks in the CIFAR-10 dataset with ε at [0.0 - 0.02]. Orange is the student model after filter pruning and Blue is the teacher model.

CERTIFICATION AREA

Finally, using the conclusions from the previous experiments, two models with the same architecture were trained (one normally – baseline and one robustly using the ”Convex Outer Adversarial Polytope” technique [1]) and were used as the teacher models for distillation experiments in the SVHN dataset. This part’s research goal was to investigate the effect that distillation has on the certification area of a neural network.

The certification area for the Baseline Distillated Model was larger compared to the Baseline Model. This observation shows that another advantage of the distillation process of normally trained models is an increase in their certification area.

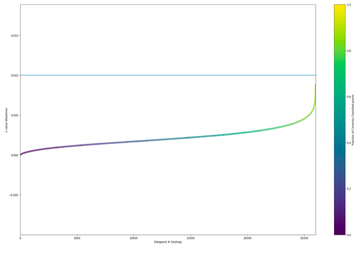

Figure 8.30: Maximum ε distances to the decision boundary of each data point of the test set in increasing ε order for the Baseline Model

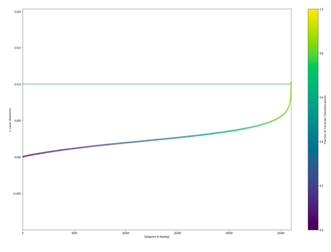

Figure 8.31: Maximum ε distances to the decision boundary of each data point of the test set in increasing ε order for the Baseline Distillated Model

From the experiments below, it is shown that the less the number of parameters of the student model, the better the certification area. This observation is independent on the way the teacher is trained (normally or robustly).

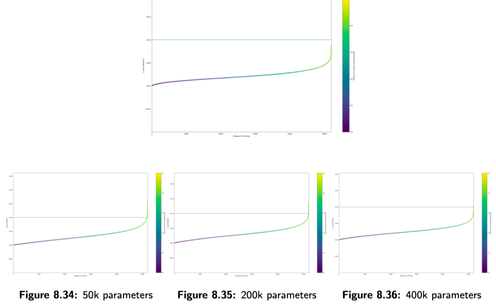

Figure 8.34: 50k parameters **Figure 8.35:** 200k parameters **Figure 8.36:** 400k parameters

Figure 8.37: Maximum ε distances to the decision boundary of each data point of the test set in increasing ε order for the Baseline Model (Up) and Baseline Distillated Models (Down)

On the contrary, the certification area for the Robust Model was much larger compared to the Robust Distillated Model. Different ways to improve the robustness and certification area of the Distillated Robust model were tested. The one that worked successfully was Robustly Re-Training for a few epochs using the technique from [1]. The question we tried to answer here was whether distillation is an important part after all. For this reason, the shallow network with the same architecture as the Robust Model (Retrained Distillated) model was trained robustly until we achieved a similar accuracy in the test set for a valid comparison and calculated its certification area.

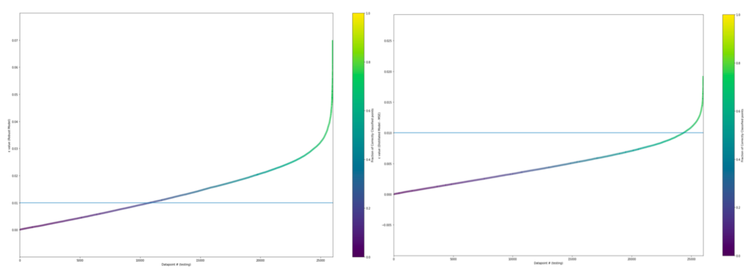

Figure 8.47: Maximum ε distances to the decision boundary of each data point of the test set in increasing ε order. (Left) the Robust Model - Accuracy: 74.220% and (Right) the Robust Distillated (Matching the Logits) Model - Accuracy: 74.589%

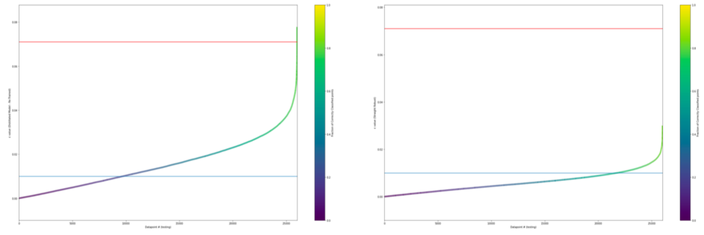

Figure 8.48: Maximum ε distances to the decision boundary of each data point of the test set in increasing ε order. (Left) the Robust Distillated (Matching the Logits - Re-Trained) Model - Accuracy: 78.423% and (Right) the Robustly Shallow Net - Accuracy: 80.015%

CONCLUSIONS

- The empirical affirmation of distillation conclusions described by [2] and a clean comparison between two distillation techniques: [2] and [3].
- The benefits of parameter reduction of shallow neural networks in terms of their computational cost and certification area.
- The extension of the prior work by [1]: Proving that distillation can result in shallow neural networks with enhanced certification areas compared to their deep complex teacher models. It was shown that if distillation is part of a neural network’s training process, it will increase the network’s certification area.

MAIN REFERENCES

[1]: Eric Wong and J. Zico Kolter. Provable Defenses against Adversarial Examples via the Convex Outer Adversarial Polytope. 2018.

[2]: Jimmy Ba and Rich Caruana. *Do deep nets really need to be deep? In NIPS*. 2014

[3]: Geoffrey Hinton Geoffrey Hinton and Jeff Dean. *Distilling the Knowledge in a Neural Network*. 2015.

[4]: Igor Durdanovic Hanan Samet Hans Peter Graf Hao Li, Asim Kadav. *PRUNING FILTERS FOR EFFICIENT CONVNETS*. 2017.