Knowledge Distillation via Smoothed Models and Adversarial Robustness

Adina Katz Yael Reches

March 24, 2022

1 Introduction

Neural networks can achieve great performance in tasks like image classification but can be very vulnerable to adversarial perturbations on their input. Nemcovsky, Zheltonozhskii, Baskin, Chmiel, Bronstein, and Mendelson [2020] showed that combining smoothing along with randomization approaches and adversarial training can improve the robustness to adversarial attacks. In this project we study the effect of knowledge distillation from the smoothed models of Nemcovsky et al. [2020] with adversarial training and investigate if it is possible to create a student model that is equal or even better in adversarial robustness.

1.1 Knowledge Distillation

Knowledge Distillation (KD) is the idea of model compression by training a small model (student) with a trained network (teacher) such as the small model can learn the exact behavior of the bigger network. We used the KD method that was developed by Hinton, Vinyals, and Dean [2015], this method is known as soft targets. The class probabilities that is produced by the trained model is used as "soft targets" for training the small model by a softmax function as

$$p(z_i, T) = \frac{exp(z_i/T)}{\sum_j exp(z_j/T)}$$

where z_i is the logit for the i-th class and T is a temperature, higher value for T will produce a softer probability distribution over classes. The knowledge distillation loss function that was used in the training process:

$$L = (1 - w) \cdot L_{CE}(z_s, y) + (w \cdot T^2) \cdot L_{CE}(p_t, p_s)$$

where y is the ground truth, z_s is the student logit, T is a temperature as described above, w is the distil weight, L_{CE} is the Cross Entropy loss function and p_t and p_s are teacher and student softmax products.

The student model trained against both the ground truth and the smoothed models outputs, aka teacher model, and the weight given to each loss calculated was determined by the distill weight variable.

1.2 The Model

The teacher model we used was the best performing smoothed model that was found by Nemcovsky et al. [2020], it included randomization of the neural network via colored noise injection (CNI). CNI is an implicit form of regularization, in which a noise vector sampled from the learned distribution is added to the weights in order to improve network robustness. M samples were taken from the learned distribution which created a cumbersome model and in order to aggregate all those samples they used randomized smoothing, averaging the outputs of the classifier over some random distribution.

To conclude, the teacher model is a Smooth CNI model with architecture of Wide-ResNet28 and with $M = 512$.

The backbone architecture of the student model is Wide-ResNet28. We used Adam optimizer with learning rate = 0.0001, weight decay = 0.0001, momentum = 0.9 and with MultiStepLR scheduler.

Table 1: Results on CIFAR-10 with WideResNet-28, accuracy score on clean and perturbed validation sets. The Smooth model is the teacher model and the KD model is the student model. The parameters used in training: epochs = 100, learning rate = 0.0001, distil weight = 0.5, loss = Cross Entropy, optimizer = ADAM and MultiStepLR scheduler.

Method	**Accuracy**, %
Smooth prediction smoothing	88.68
Smooth soft prediction smoothing	88.53
KD prediction smoothing	**91.21**
KD soft prediction smoothing	**91.22**

Table 2: Results on CIFAR-10 with PGD attack and WideResNet-28, accuracy score on the validation set. The Smooth model is the teacher model and the KD model is the student model. The parameters used in training: epochs = 100, learning rate = 0.0001, distil weight = 0.75, perturb distil weight = 0.25, loss = Cross Entropy and optimizer = ADAM.

Method	**Accuracy, %**	
	Clean	**PGD-10**
Smooth prediction smoothing	88.68	63.67
Smooth soft prediction smoothing	88.53	63.48
KD prediction smoothing	79.9	**46**
KD soft prediction smoothing	82.13	**45.83**

1.3 The Objective

This project came to pass as a continuation of the Nemcovsky et al. [2020] paper. The smoothed model is a non-deterministic, cumbersome model which achieved significant performance boost for perturbations.

In our project we explore the possibility of using Knowledge distillation via the smoothed model, to create a student model, that is deterministic and can perform on perturbed data similar to the smoothed model or even better.

2 Experiments

We conducted experiments to test if with knowledge distillation a student model can generalize over perturbed data when learning from a smoothed model. We used two types of aggregation methods for the outputs generated by the smoothed model as described in Nemcovsky et al. [2020]. The first, prediction smoothing, a voting scheme only considering the classification of each of the M samples, discarding the predicted probabilities of each class, divided by M. The second, soft prediction smoothing, as opposed to prediction smoothing, this output fully considers the predicted class probabilities of each of the M samples.

2.1 Knowledge Distillation

We started by training the student model on clean data with no perturbations, against both types of aggregated teacher outputs as described in Section 2. The student learned from both the teacher and the ground truth with distil_weight = 0.5, giving the same weight to each. As can be seen in (Table 1) the student model out-preformed the teacher model on the clean data and so we can determine that the knowledge distillation frame work is successful.

2.2 Knowledge Distillation with Adversarial Training

Next we combined the KD training with adversarial training as was done by Nemcovsky et al. [2020]. The teacher model we used for training the student model was a Wide-ResNet28, using the CNI model which was trained adversarially on CIFAR-10, for 100 epochs under projected gradient descent (PGD) attack with $k = 10$. We also adversarially trained the student model for 100 epochs under PGD

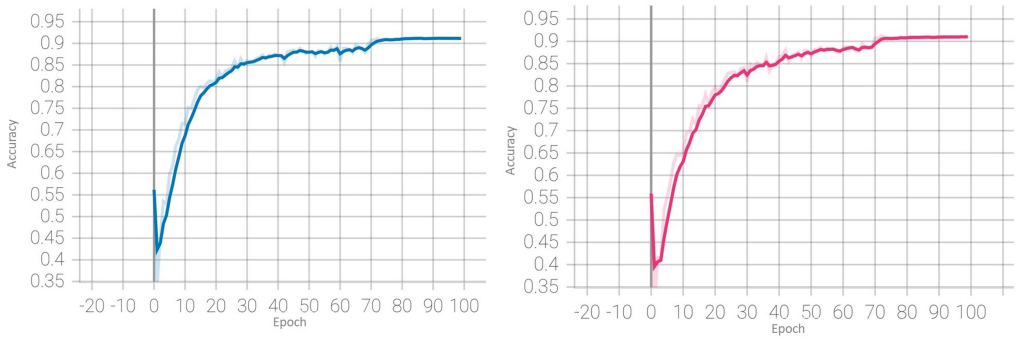


Figure 1: Accuracy of the distilled model trained both on the ground truth and the teacher's predictions outputs with respect to number of epochs. On the left the teacher outputs used was prediction smoothing and on the right the teacher outputs used was soft prediction smoothing outputs.

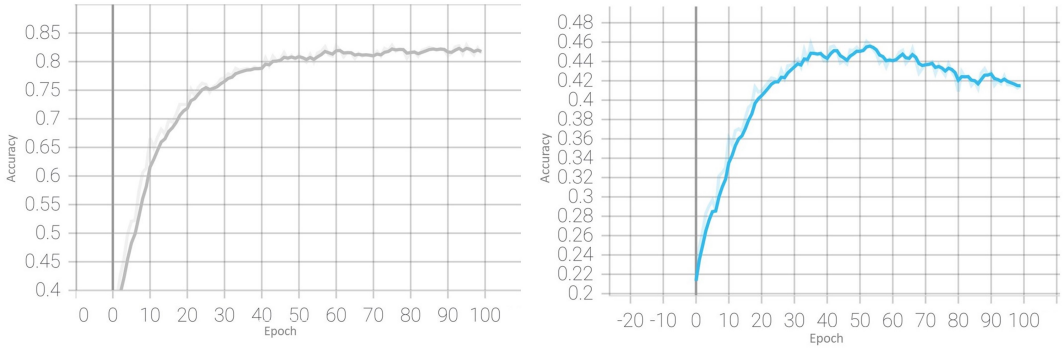


Figure 2: Accuracy of the distilled model trained with prediction smoothing outputs under PGD attack with respect to number of epochs. The PGD attacks parameters used: $k = 10$ and $\epsilon = 8/255$. On the left is the accuracy results on clean data and on the right is the accuracy results on perturbed data.

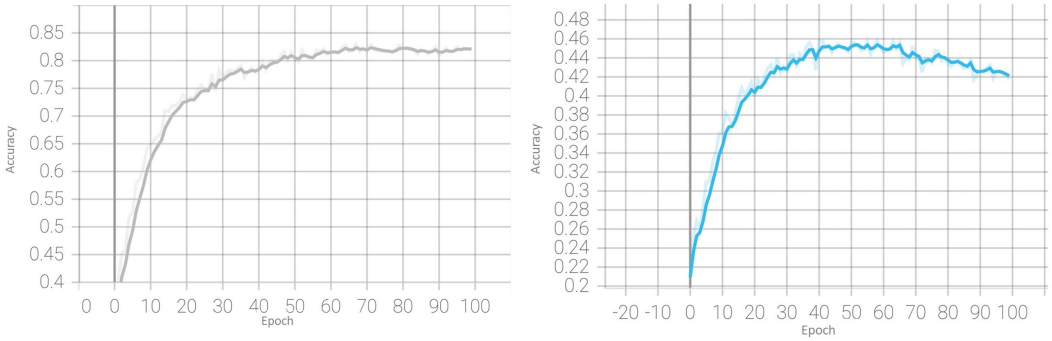


Figure 3: Accuracy of the distilled model trained with soft prediction smoothing outputs under PGD attack with respect to number of epochs. The PGD attacks parameters used: $k = 10$ and $\epsilon = 8/255$. On the left is the accuracy results on clean data and on the right is the accuracy results on perturbed data.

Table 3: Results of transfer attack, where the KD student model is the distilled model that was trained on CIFAR-10 with PGD attack. The target model is the model that was tested on the attack that was created by the attack model. The epsilon parameter used in the attacks was $\epsilon = 8/255$.

Attack model	**Target model**	**Accuracy, %**	
		Clean	**PGD-10**
CNI	CNI	88.67	63.88
KD student	CNI	88.76	**63.4**
Smoothed CNI	Smoothed CNI	42.81	36.58
KD student	Smoothed CNI	42.74	**32.95**
CNI	KD student	79.92	**69.3**
Smoothed CNI	KD student	79.92	**77.97**

attack with $k = 10$ on CIFAR-10, for each batch the student was trained on the clean data and the perturbed data. We controlled the weight for each, teacher output and the ground truth, with the distil weight and the perturb distil weight variables, best results were achieved with $distil\_weight = 0.75$ and $perturb\_distil\_weight = 0.25$. The results that can be seen in (Fig. 2) and (Fig. 3) show that the student model was not able to generalize over the perturbed data, the validation accuracy results can be viewed in (Table 2).

3 Transfer Attack

There are two types of adversarial attacks, black-box attacks and white-box attacks. In white-box attacks the attacker has access to the model's parameters, while in black-box attacks, the attacker has no access to these parameters, i.e., it uses a different model or no model at all to generate adversarial images with the hope that these will transfer to the target model. Transfer-based attacks, are black-box attacks in which adversarial examples generated for a different model are probable to remain adversarial for the target model due to the transferability Papernot, McDanieland, and Goodfellow [2016].

In this section we explore transfer-based attacks, we wanted to see if although the student model didn't generalize on the perturbed data, it had learned a weakness in Nemcovsky et al. [2020] model that could be exploited to create a black-box attack. To test this hypothesis we generated the PGD attack on the student model that best performed on perturbed data and then tested the attack on both the CNI and Smoothed CNI models.

As can be seen in (Table 3) the distilled model was able to create an effective black-box attacks over the CNI and Smoothed CNI models. The attack results are similar to the white-box attack results that both models displayed and can be seen in (Table 3).

Also, in (Fig. 4) it can be seen that the KD student model's accuracy reaches minimum accuracy while learning the attack, within $k < 10$ steps, unlike the CNI and Smoothed CNI models. We then continued to check if the reverse transfer-based attack had merit, so we created an attack on the CNI and Smoothed CNI models and transferred it to the distilled model. In (Table 3) it can be seen that although the distilled model was unable to generalize over the perturbed data, as we show in Section 2.2, it performed well against the transfer attack created by the CNI and Smoothed CNI models.

4 Conclusions

In this project we explored the effect of distilling a student model from a smoothed model that was adversarially trained, to see if the student could generalize over perturbed data as well as the smoothed model. We showed that using knowledge distillation was unsuccessful in regards to the smoothed model adversarial robustness. We suggest that the adversarial robustness stems from the combination of the model architecture weights as a whole and therefore can not be distilled via only the models outputs to a simpler model architecture. Another possibility is that the adversarial robustness is achieved via the sampling technique (randomization) which isn't implemented in the student model, hence the student

The following graphs display the results of the accuracy with respect to k, for $k = 10$, of the attack model while creating the perturbed data. Each figure represents a different attack model as described in the captions.

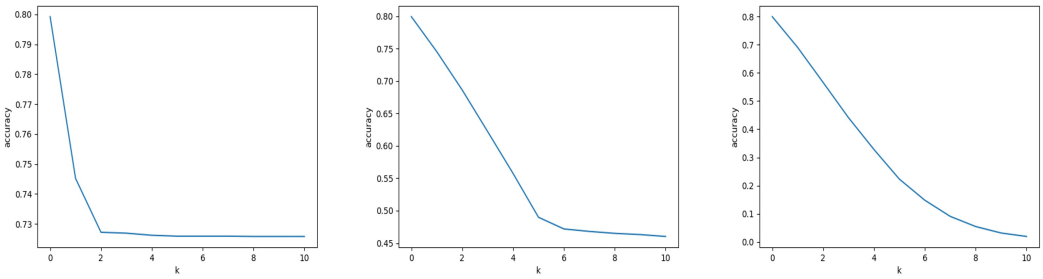
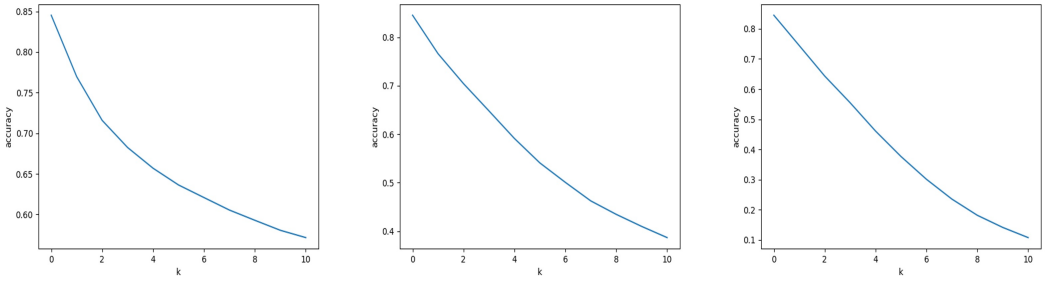


Figure 4: Accuracy results when the attack model is the **KD student** model. On the left the results when $\epsilon = 2/255$, in the middle $\epsilon = 8/255$ and on the right $\epsilon = 30/255$.



,

Figure 5: Accuracy results when the attack model is the **CNI** model. On the left the results when $\epsilon = 2/255$, in the middle $\epsilon = 8/255$ and on the right $\epsilon = 30/255$.

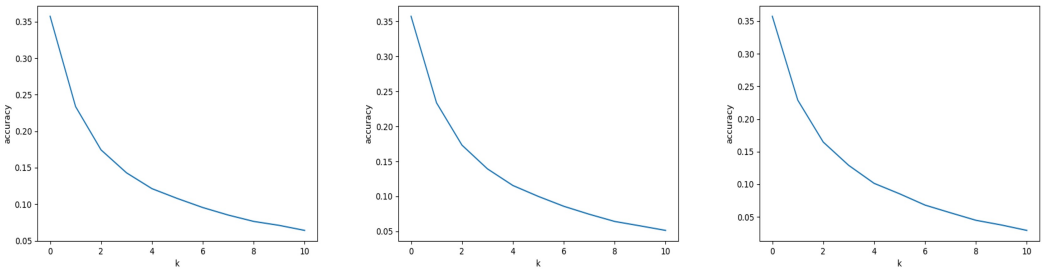


Figure 6: Accuracy results when the attack model is the **Smoothed CNI** model. On the left the results when $\epsilon = 2/255$, in the middle $\epsilon = 8/255$ and on the right $\epsilon = 30/255$.

model could not perform as well on perturbed data as the smoothed model.

Concluding the transfer attack section, seeing that the distilled model is a smaller one hence less computationally expensive, being able to produce an effective black-box attack is valuable. Also, the distilled model performed well as the target model when an attack was transferred from the CNI or Smoothed CNI models. To conclude, the results of the transfer-based attack experiment show promise and should be further researched.

References

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. 2015. URL https://arxiv.org/pdf/1503.02531.pdf.

Yaniv Nemcovsky, Evgenii Zheltonozhskii, Chaim Baskin, Brian Chmiel, Alex M. Bronstein, and Avi Mendelson. Adversarial Robustness via Noise Injection in Smoothed Models. 2020. URL http://arxiv.org/abs/1312.6199.

Nicolas Papernot, Patrick McDanieland, and Ian Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *NeurIPS*, 2016. URL https://proceedings.neurips.cc/paper/2019/file/32508f53f24c46f685870a075eaaa29c-Paper.pdf.