Knowledge Distillation: Current Understanding and Future Directions in an Adversarial Context

Luke Tianyou Zhuo

Introduction

Convolutional neural networks have been widely employed in image classification tasks. However, deeper and more complicated networks are often too large to run on devices with limited memory and computing resources like wearable devices and mobile phones. Consequently, researchers have sought to boost model accuracy while minimizing model size. Knowledge distillation (KD) achieves this goal by distilling the knowledge from a large "teacher" model (or an ensemble of models) into a single small "student" model [1]. However, the necessity for the role of a larger teacher has been debated with the development of alternative approaches, such as reversed knowledge distillation and teacher-free knowledge distillation [2][3]. I explore the reasons for the success of both "vanilla" and alternative approaches to knowledge distillation and build on the applications of knowledge distillation to adversarial training - training to improve model robustness to intentional adversarial noise - by presenting the approach I have devised for applying teacher-free distillation to an adversarial context.

Methods

I trained models for reversed and teacher-free distillation on CIFAR-10 as described in [2] using stochastic gradient descent (SGD) for 200 epochs with initial learning rate of 0.1, divided by 5 at epochs 60, 120, and 160, and with momentum of 0.9 and weight decay of 5e-4. The models for adversarial teacher-free distillation were trained on CIFAR-10 in accordance with [3], with SGD, initial learning rate of 0.1, and exponential learning rate decay at a rate of 0.9 for 50 epochs with early stopping. Model training for the vanilla approach to distillation was on MNIST with SGD with initial learning rate of 1e-2 and exponential learning rate decay at a rate of 0.95 for 50 epochs and momentum of 0.9 and weight decay of 1e-5.

Vanilla Knowledge Distillation

Figure 1: Accuracy and Number of Parameters for Large and Small (no KD and KD) Models

Model	Shape		
Large Model	(fc1): Linear(in_features=28*28, out_features=1200, bias=True) (fc2): Linear(in_features=1200, out_features=1200, bias=True) (fc3): Linear(in_features=1200, out_features=10, bias=True))		
	Number of Parameters	**Accuracy**	
	2,392,800	98.98	
Small Model	(fc1): Linear(in_features=28*28, out_features=400, bias=True) (fc2): Linear(in_features=400, out_features=10, bias=True)		
	Number of Parameters	**Accuracy (no KD)**	**Accuracy (KD)**
	317,600	98.06	98.73

Knowledge Distillation:

- Perform distillation by minimizing KD loss function: sum of cross entropy loss H between hard labels q and hard predictions p and the divergence in student model's soft predictions p_τ^t(softmax "probabilities") for each class from teacher's soft predictions p_τ
- KL–divergence typically used for divergence loss with softmax confidence parameter τ and hyperparameter α balancing cross entropy and divergence loss: $L_{KD} = (1-\alpha)H(q,p) + \alpha D_{KL}(p_\tau^t, p_\tau)$

Dark Knowledge:

- The small model attains higher accuracy following distillation; it has gained no parameters but has gained knowledge (Figure 1)
- This knowledge - privileged information on similarity among classes - is known as "dark knowledge" [2]

Reversed Knowledge Distillation

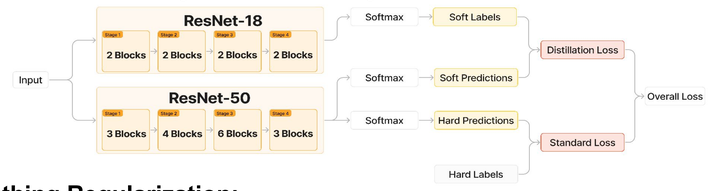
Figure 2: Reversed Knowledge Distillation (Re-KD) Accuracy (mean ± std over 3 runs)

Teacher: baseline	Student: baseline	Re-KD (S→T)
95.29±0.09	94.98±0.10	95.46±0.12

Reversed Knowledge Distillation (Re-KD):

- ResNet50 is the larger teacher model and ResNet18 is the smaller student model
- Distillation from the pretrained ResNet18 to the ResNet50 suggests improvement in spite of distillation being from the smaller (and less accurate) model (Figure 3)

Figure 3: Reversed Knowledge Distillation Visualized



Label Smoothing Regularization:

- Re-KD's greater accuracy than the baseline ResNet50 suggests dark knowledge from a larger model is not the sole driver of better accuracy (Figure 2)
- Distillation replaces hard labels with smoothed labels, regularizing model training
- Distillation acts as a learned label smoothing regularization with the smoothing distribution coming from the model being distilled [2]

Teacher-Free Knowledge Distillation

Figure 4: Teacher-Free Knowledge Distillation Accuracy (mean ± std over 3 runs)

ResNet18: Baseline	Tf KD-reg	Tf KD-self
94.98±0.10	95.39±0.15	95.42±0.13

Knowledge Distillation via Manually Designed Regularization (Tf KD-reg):

- For K classes, assign probability $a \in [0.9,1]$ to correct class, and $(1-a)/(K-1)$ to incorrect classes, before applying softmax to form "virtual teacher" on softmax outputs (Figure 5) [2]
- Knowledge distillation performed from manually-designed virtual "teacher" to ResNet18

Knowledge Distillation via Self-Training (Tf KD-self):

- Distillation performed from pretrained ResNet18 to another ResNet18 (self-distillation) [2]

Vanishing Gradient Problem:

- Tf KD-reg and Tf KD-self both have higher accuracy than ResNet18 baseline (Figure 4)
- With either a model itself as a teacher or no teacher entirely, teacher-free methods resemble Re-KD and benefit from the aforementioned label smoothing regularization
- Examining gradients for teacher-free approaches also indicates knowledge distillation mitigates the vanishing gradient problem known to plague deep neural networks [3]
- Significantly larger average gradients for layers in the Tf KD-reg and Tf KD-self models compared to those for a standard ResNet18 (Figure 6)

Figure 5: Manually Designed Virtual Teacher Probabilities. Probability of the correct class and incorrect classes (for a=0.9) at each temperature τ, the parameter controlling softmax confidence

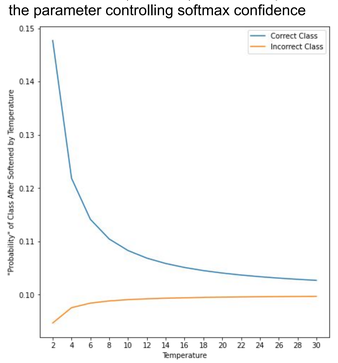
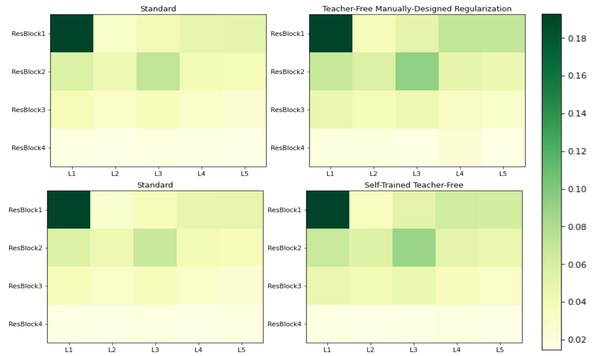


Figure 6: Heatmap Comparison of Average Gradients between Teacher-Free Knowledge Distillation Approaches and ResNet18 without Knowledge Distillation. Average gradients for each layer at each stage (ResBlock) of the ResNets



Novel Adversarial Teacher-Free Distillation

Figure 7: Novel Application of Teacher-Free Distillation Approaches to an Adversarial Context Compared to Adversarially Trained Distillation Approaches Highlighted in [4]

Method of Knowledge Distillation	Clean Accuracy	AutoAttack Accuracy
ResNet18 (AT) Student, No Teacher	84.22	46.99
ResNet18 Student, ResNet18 (AT) Teacher	82.70	47.66
ResNet18 (AT) Tf KD-reg	82.02	48.28
ResNet18 (AT) Tf KD-self	81.53	48.83

Adversarial Training:

- Using Projected Gradient Descent attack with 7 iterations (PGD-7) to add intentional noise to data with intent to lead the model to misclassify data
- Aim of attacks is noisy data x' where $x' = \arg\max_{||x'-x||<\epsilon} (H(f(x'), y))$

Adversarial Training Manually Designed Regularization (AT Tf KD-reg):

- Virtual teacher (soft outputs p_τ^d) distillation to adversarially trained model
- Loss function: $L_{KDreg} = (1-\alpha)H(q, p(x')) + \alpha D_{KL}(p_\tau^d, p_\tau(x'))$

Adversarial Self-Training (AT Tf KD-self):

- Distillation from adversarially pretrained ResNet18 (soft outputs p_τ^t) to another ResNet18 (self-distillation) being trained on adversarial data
- Loss function: $L_{KDself} = (1-\alpha)H(q, p(x')) + \alpha D_{KL}(p_\tau^t(x), p_\tau(x'))$

Label Smoothing:

- Model robustness measured using model accuracy on data perturbed by standard adversarial evaluation AutoAttack (Figure 7)
- Both adversarially trained teacher-free distillation approaches introduced are promising, with higher AutoAttack accuracy than the adversarially trained ResNet18 mentioned in [4] and the standard ResNet18 student with adversarially trained ResNet18 teacher highlighted in [4]. (Figure 7)

Conclusions and Further Investigation

Knowledge distillation enables improved accuracy in a student model. This stems from factors which include dark knowledge, addressing of the vanishing gradient problem, and label smoothing regularization, as evidenced by the results of vanilla, reversed, and teacher-free self-training and manually designed regularization approaches to knowledge distillation. The approaches devised to adversarially train teacher-free knowledge distillation models have showed initialling promising results in promoting robustness against adversarial noise. Delving into the use of teacher-free knowledge distillation in an adversarial setting is an area in which I hope to both further examine with greater computation resources and open up to others to explore [6].

References

[1] Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. "Distilling the knowledge in a neural network." (2015).

[2] Yuan, Li, et al. "Revisit knowledge distillation: a teacher-free framework." (2019).

[3] Zhang, Linfeng, et al. "Be your own teacher: Improve the performance of convolutional neural networks via self distillation." Proceedings of the IEEE/CVF International Conference on Computer Vision (2019).

[4] Maroto, Javier, Guillermo Ortiz-Jiménez and Frossard, Pascal "On the benefits of knowledge distillation for adversarial robustness" (2022).

[5] Goibert, Morgane and Elvis Dohmatob. "Adversarial Robustness via Label-Smoothing"

[6] https://github.com/lukezhuo/KnowledgeDistillationResNetArchitectures