Does Knowledge Distillation Really Work?

Samuel Stanton Pavel Izmailov Polina Kirichenko Alex Alemi Andrew Gordon Wilson

NYU Google AI

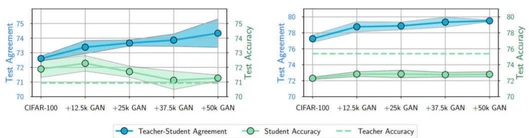
Paper Code

Overview

We show that while knowledge distillation can improve student generalization, it does not typically work as it is commonly understood:

- There is significant discrepancy between the predictive distributions of the teacher and the student.
- Difficulties in optimization are a key reason for why the student is unable to match the teacher.
- Details of the dataset are important for distillation fidelity but a better dataset can make optimization even more difficult.

Fidelity of knowledge distillation (ResNets on C-100)



(a) Self-distillation (b) Ensemble distillation

What is knowledge distillation?

Goal: train a student model to mimic predictions of a teacher model.

Distillation loss: $\mathcal{L}_s = \alpha \mathcal{L}_{\mathrm{NLL}} + (1-\alpha)\mathcal{L}_{\mathrm{KD}}$

$$\mathcal{L}_{\mathrm{NLL}}(\mathbf{z}_s, \mathbf{y}) := -\sum_{j=1}^{c} y_j \log \sigma_j(\mathbf{z}_s), \quad \mathcal{L}_{\mathrm{KD}}(\mathbf{z}_s, \mathbf{z}_t) := -\tau^2 \sum_{j=1}^{c} \sigma_j\left(\frac{\mathbf{z}_t}{\tau}\right) \log \sigma_j\left(\frac{\mathbf{z}_s}{\tau}\right).$$

What is fidelity?

Distillation fidelity — the ability of a student to match a teacher's predictions.

$$\text{Average Top-1 Agreement} := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{\arg\max_j \sigma_j(\mathbf{z}_{t,i}) = \arg\max_j \sigma_j(\mathbf{z}_{s,i})\},$$

$$\text{Average Predictive KL} := \frac{1}{n}\sum_{i=1}^{n} \mathrm{KL}\left(\hat{p}_t(\mathbf{y}|\mathbf{x}_i) \,||\, \hat{p}_s(\mathbf{y}|\mathbf{x}_i)\right),$$

Why does fidelity matter?

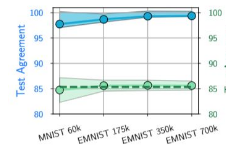
- For large ensembles good distillation fidelity implies better accuracy
- We may want to transfer properties beyond accuracy:
  - Uncertainty calibration
  - Fairness of predictions
- Scientific understanding of knowledge distillation

When is knowledge transfer successful?

Set α = 0, add unlabeled examples

- With enough data the student learns to match the teacher predictions
- self-distillation does not improve generalization

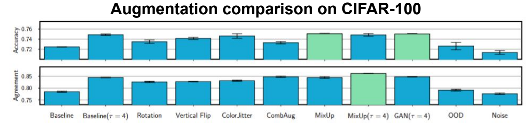
LeNet-5 on MNIST



Identifiability hypothesis: are we showing the student the wrong data?

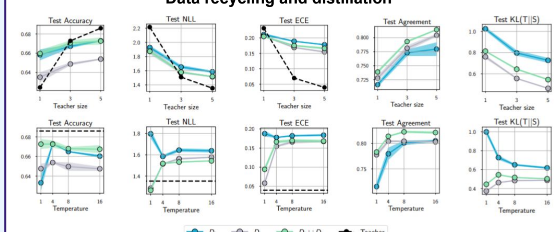
Data augmentation has a substantial effect on distillation fidelity and student accuracy.

Augmentation comparison on CIFAR-100



- Best policies for student accuracy (*MixUp, GAN*) are not best for distillation fidelity
- The best policy for fidelity (*MixUp T=4*) only achieves 85% agreement between the teacher and the student
- *Noise* and *OOD* data are not helpful for distillation

Data recycling and distillation

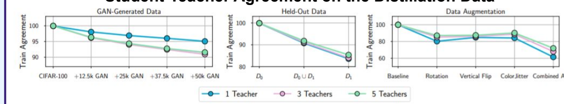


- Recycling data used to train the teachers is *better* than using new data for student accuracy
- Using new data leads to better distillation fidelity

Optimization hypothesis: are we solving the distillation problem well?

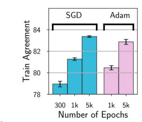
Increasing the support of the distillation dataset makes the distillation objective increasingly difficult to optimize.

Student-Teacher Agreement on the Distillation Data

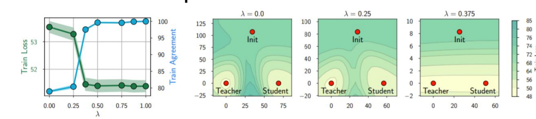


- We observe a major drop in the **train** agreement on the data used for distillation as we increase the size of the distillation dataset
- Same holds when we use strong data augmentation policies

We cannot resolve the issue by training longer or using a different optimizer.



Optimization and distillation



(b) Initialization effect (c) Loss Visualization

- If we initialize the student close to the teacher weights, it can recover a trivial optimal solution.
- Initializing the student randomly, we converge to suboptimal minima of the distillation loss surface.

Results on ImageNet and NLP

We verify that our observations hold on ImageNet and IMDB datasets.

Dataset	Teach. Size	Teach. Acc. (↑)	Stud. Acc. (↑)	Agree. (↑)	KL (↓)
IMDB	1	79.361 (0.132)	80.353 (0.198)	86.488 (0.521)	0.124 (0.012)
	3	81.807 (0.129)	81.129 (0.057)	89.832 (0.349)	0.064 (0.003)
	5	**82.216 (0.207)**	**81.167 (0.196)**	**90.793 (0.180)**	**0.052 (0.001)**
ImageNet	1	0.748 (0.001)	0.753 (0.001)	0.855 (0.001)	0.217 (0.002)
	3	0.764 (0.001)	0.755 (0.001)	0.878 (0.001)	0.157 (0.001)
	5	**0.767 (0.001)**	**0.756 (0.001)**	**0.884 (0.001)**	**0.142 (0.001)**