

# Does Knowledge Distillation Really Work?

Samuel Stanton, Pavel Izmailov, Polina Kirichenko, Alexander A. Alemi, Andrew Gordon Wilson, NeurIPS 2021



## Дистилляция: лосс-функция

$$\mathcal{L}_s := \alpha \mathcal{L}_{NLL} + (1 - \alpha) \mathcal{L}_{KD}, \alpha \in [0, 1)$$

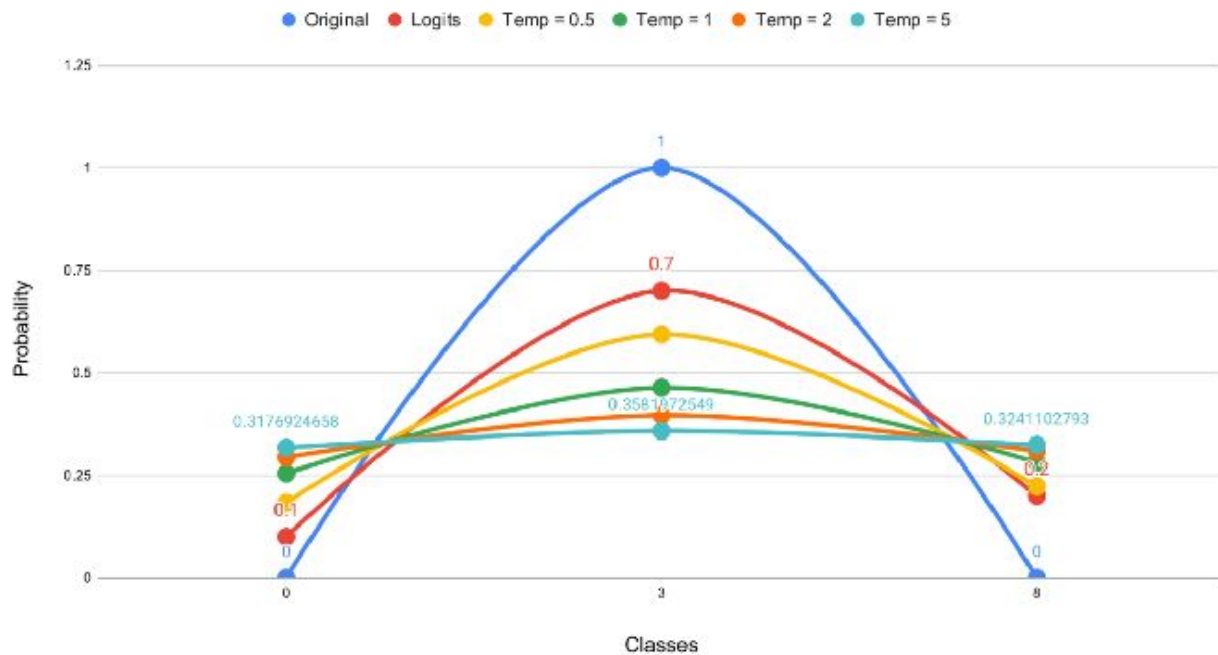
$$\mathcal{L}_{NLL}(\mathbf{z}_s, \mathbf{y}) := - \sum_{j=1}^c y_j \log \sigma_j(\mathbf{z}_s)$$

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

$$\sigma_i(\mathbf{z}) = \frac{\exp(z_i)}{\sum_j \exp(z_j)}, \mathbf{z} := f(\mathbf{x}, \theta) - \text{logits}$$

# Температура

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

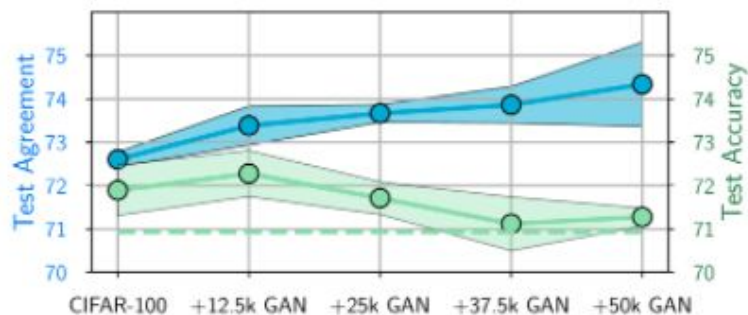




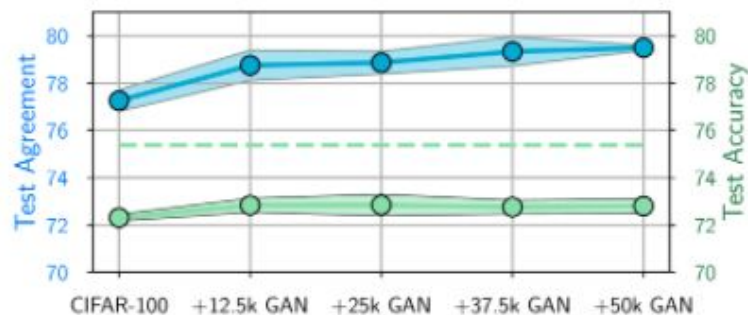
## Мотивация

- **Улучшение обобщающей способности ученика:** часто увеличить верность учителю == увеличить обобщающую способность.
- **Интерпретируемость:** возможность интерпретировать структуру данных, замеченную моделью-учителем.
- **Понимание:** разделение обобщающей способности ученика и его верности учителю поможет понять, как работает дистилляция.

## Дистилляция ResNet-56 в ResNet-56



(a) Self-distillation



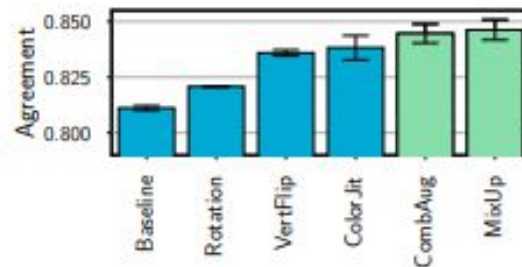
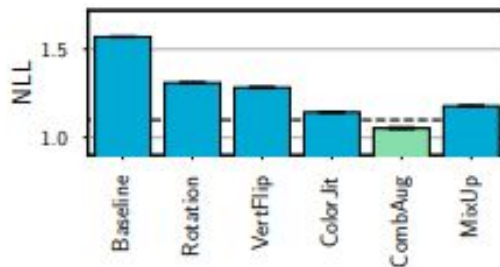
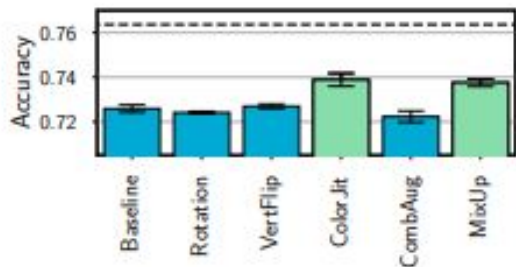
(b) Ensemble distillation


● Teacher-Student Agreement    ● Student Accuracy    - Teacher Accuracy

# Почему ученик так неверен учителю?

- Архитектура

Дистилляция 5-компонентного ансамбля VGG-16 в VGG-16 с разными аугментациями.





## Почему ученик так неверен учителю?

- Низкая способность ученика к обобщению
  - Низкая верность учителю наблюдается даже при дистилляции в такую же сеть.


# Почему ученик так неверен учителю?

- Слишком простой/маленький датасет.
- Специфика области, из которой взяты данные.

Дистилляция ансамбля BiLSTM в BiLSTM и ансамбля ResNet-56 в ResNet-56 на разных датасетах

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	<b>82.216 (0.207)</b>	<b>81.167 (0.196)</b>	<b>90.793 (0.180)</b>	<b>0.052 (0.001)</b>
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	<b>0.767 (0.001)</b>	<b>0.756 (0.001)</b>	<b>0.884 (0.001)</b>	<b>0.142 (0.001)</b>





## Почему ученик так неверен учителю?

- Используем неправильные данные для дистилляции: добиться высокой верности на тренировочных данных недостаточно, чтобы иметь высокую верность на тестовых данных.

## Правильные ли данные мы используем?

Дистилляция 5-компонентного ансамбля ResNet-56 в ResNet-56 с разными аугментациями.



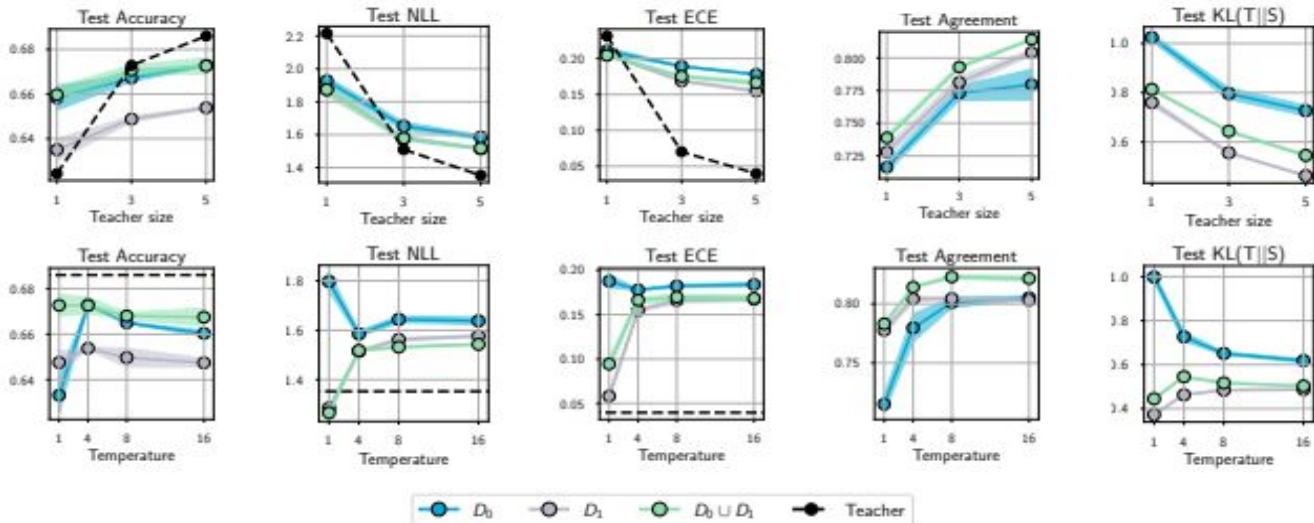



## Правильные ли данные мы используем?

- **Может быть, мы показываем ученику не те данные?**
  - Из-за аугментаций распределение обучающих данных для ученика и для учителя не совпадают.
  - Повторное использование данных, на которых обучался учитель, для обучения ученика, означает, что данные больше не являются независимыми случайно выбранными из одного распределения.

# Правильные ли данные мы используем?

Дистилляция ансамбля ResNet-56 в ResNet-56. В верхней строке: температура=4, меняется размер ансамбля, в нижней: размер ансамбля = 3, меняется температура.





# Почему ученик так неверен учителю?

- **Сложности оптимизации:**  
ученик не согласен с учителем даже на тренировочном датасете.

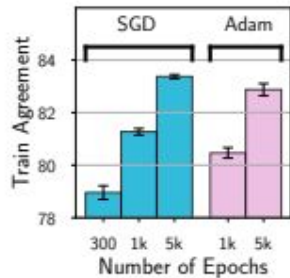
# Может ли ученик быть согласен с учителем на тренировочных данных?

Согласие с учителем на тренировочном датасете для предыдущих экспериментов.

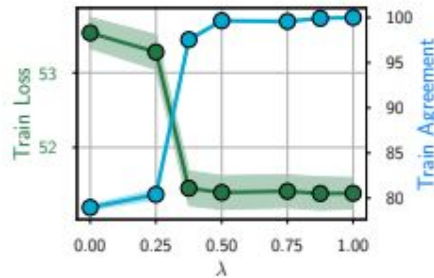


# Почему ученик не согласен с учителем на тренировочных данных?

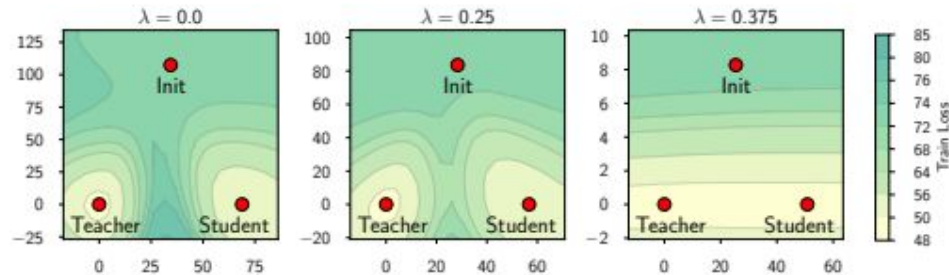
Дистилляция ResNet-20 в ResNet-20.



(a) Optimizer effect



(b) Initialization effect



(c) Loss Visualization



## Итого

- Высокая доля верных ответов ученика на тестовых данных не означает высокую верность учителю.
- Оптимизационная задача в дистилляции сложная.
- Увеличение датасета для дистилляции за пределы тренировочного датасета учителя увеличивает долю верных ответов ученика, но делает оптимизационную задачу сложнее.