Does Knowledge Distillation Really Work?

Морозов Никита
Шапкин Антон
Аюпов Шамиль

Knowledge Distillation: recap

Имеем задачу классификации и нейросеть s. Используем стандартную кросс-энтропийная функция потерь, учим сеть предсказывать верный класс:

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

z — логиты на выходе сети
y — one-hot метки класса
$\sigma_i(\mathbf{z}) := \exp(z_i) / \sum_j \exp(z_j)$ — софтмакс

Knowledge Distillation: recap

Теперь же пытаемся обучить сеть-ученика s повторять выходы учителя t

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

τ — температура, отвечает за “мягкость” меток.

При единичной температуре минимизация эквивалентна минимизации

$$\mathrm{KL}(\hat{p}_t(\mathbf{y} \mid \mathbf{x}) \| \hat{p}_s(\mathbf{y} \mid \mathbf{x}))$$

Fidelity и с чем его едят

Хотим измерить, насколько согласованы предсказания ученика и учителя:

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

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

Common sense: ученик работает хорошо, потому что обучается повторять поведение учителя. Однако бывает наоборот: ученик по итогу плохо согласован с учителем, однако работает лучше него.

Вопросы исследования

- Как связаны fidelity и качество ученика на тестовых данных?
- Почему возникают сложности с получением высокого fidelity?
- Does knowledge distillation really work?

банальный вопрос
овальный ответ

Fidelity vs Accuracy

self-distillation — в качестве ученика и учителя выступают сети абсолютно одинаковой архитектуры

Обучаем учителя на части MNIST, затем для дистилляции добавляем новые данные.

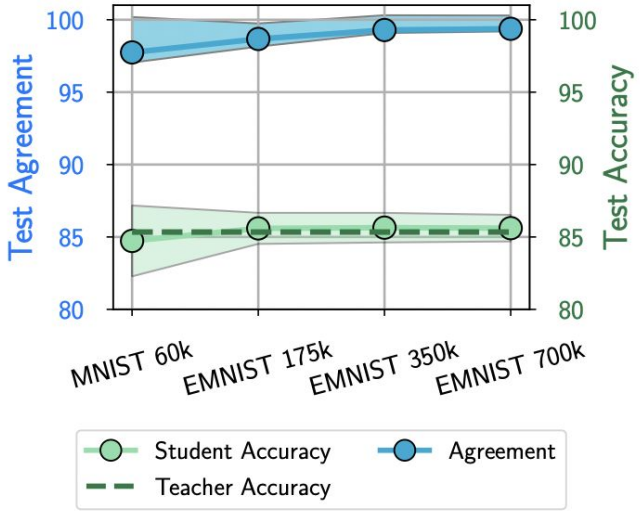
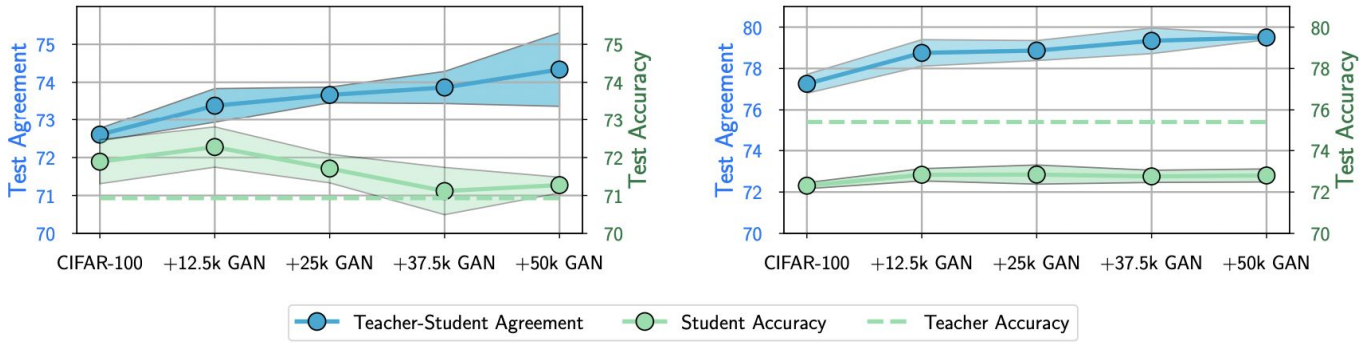


Figure 2: LeNet-5 self-distillation on MNIST with additional distillation data. The shaded region corresponds to $\mu \pm \sigma$, estimated over 3 trials.

Fidelity vs Accuracy



(a) Self-distillation

(b) Ensemble distillation

Figure 1: **Evaluating the fidelity of knowledge distillation.** The effect of enlarging the CIFAR-100 distillation dataset with GAN-generated samples. **(a)**: The student and teacher are both single ResNet-56 networks. Student fidelity increases as the dataset grows, but test accuracy decreases. **(b)**: The student is a single ResNet-56 network and the teacher is a 3-component ensemble. Student fidelity again increases as the dataset grows, but test accuracy now slightly increases. The shaded region corresponds to $\mu \pm \sigma$, estimated over 3 trials.

Чем вызван низкий fidelity?

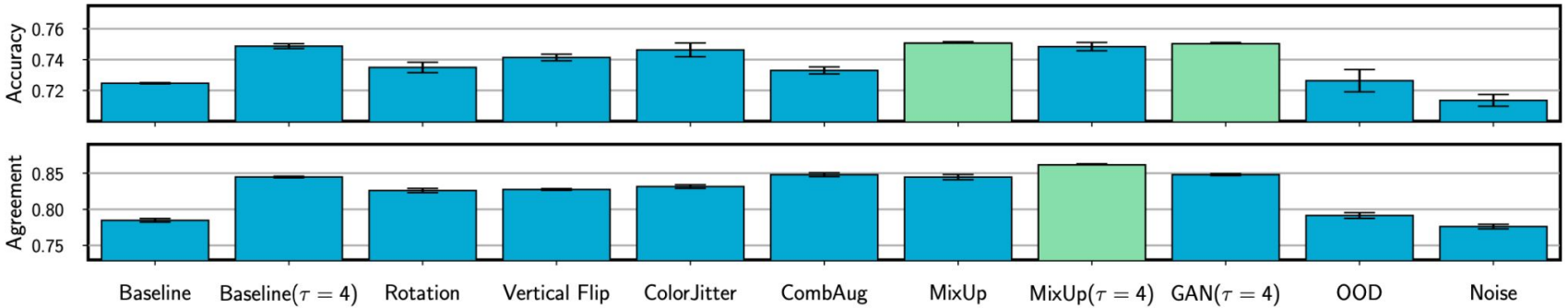
Авторы вкратце проходятся по нескольким гипотезам и отбрасывают их: student capacity, network architecture, dataset scale, data domain.

Обратим внимание на две другие:

- Identifiability — нам не хватает обучающих данных, чтобы сматчить учителя и ученика на отложенной выборке.

- Optimization — мы не справляемся с задачей оптимизации, то есть ученик не согласен с учителем даже на обучающей выборке.

Identifiability: аугментации



Baseline: horizontal flip + random crop. Дистиллируем ансамбль из 5 сетей.

Наблюдения: лучшие аугментации с точки зрения agreement и accuracy разные, далекие от распределения над данными аугментации работают не очень (OOD, Noise), максимальный agreement все еще довольно скромный.

Вывод: увеличение количества примеров слабо помогает студенту выучить поведение учителя.

Identifiability: recycling hypothesis

Может быть мы показываем студенту не те данные?

Разобьем обучающую выборку на $\mathcal{D}_0$ и $\mathcal{D}_1$. Учителя будем учить на $\mathcal{D}_0$, для дистилляции попробуем $\mathcal{D}_0, \mathcal{D}_1, \mathcal{D}_0 \cup \mathcal{D}_1$.

Увеличится ли fidelity, если дистиллировать ученика на тех данных, которые учитель не видел при обучении?

Почему может стать лучше? Когда мы проводим дистилляцию на тех же данных, на которых учитель обучался, не выполняется предположение, что данные для дистилляции являются i.i.d. семплами из совместного распределения на объектах и метках.

Identifiability: recycling hypothesis

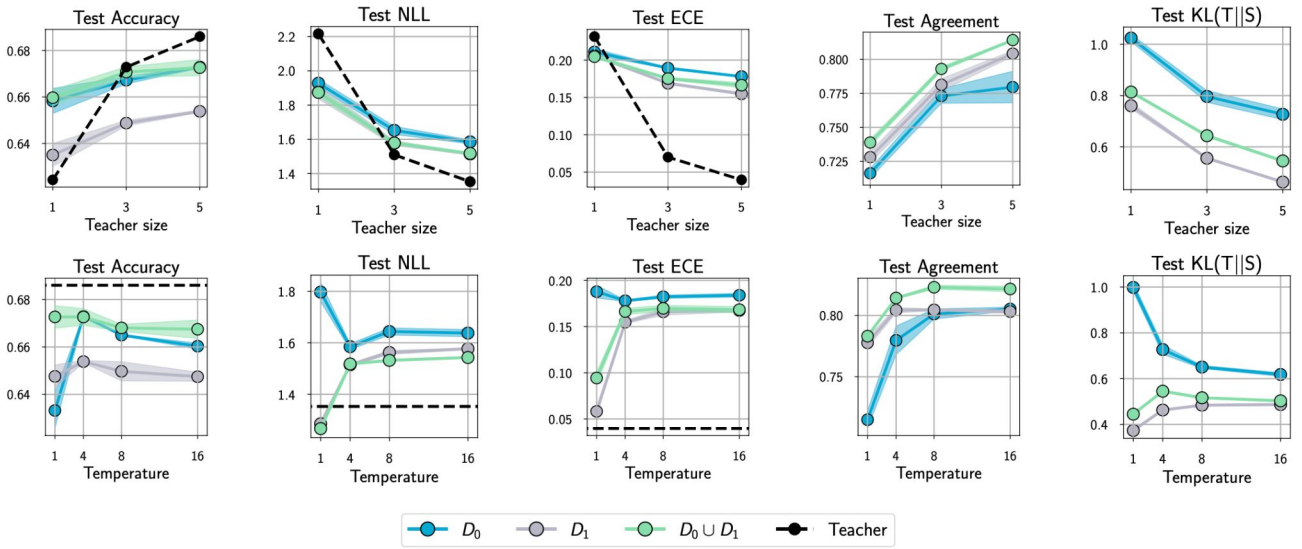
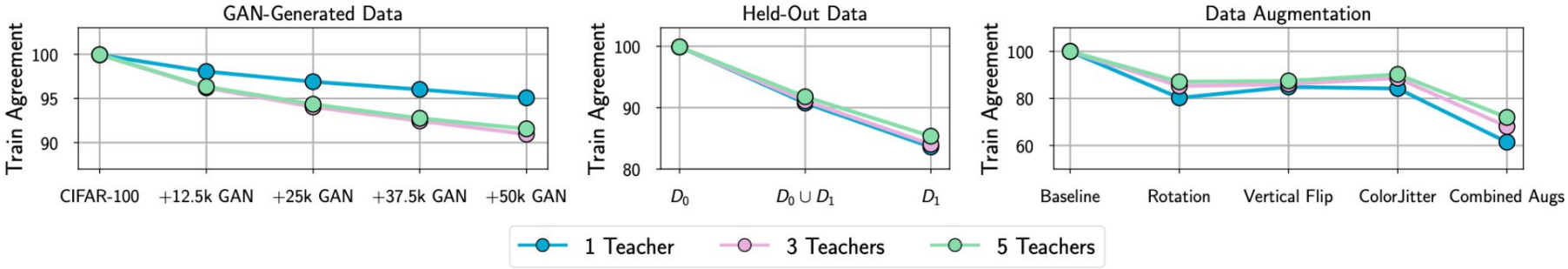


Figure 4: **Data recycling and distillation**: results on subsampled CIFAR-100. **Top:** We fix the temperature ($\tau = 4$) and vary the number of ensemble components (m), comparing students distilled on the same dataset as the teacher ($\mathcal{D}_0/\mathcal{D}_0$), a reserved dataset ($\mathcal{D}_0/\mathcal{D}_1$), or both ($\mathcal{D}_0/\mathcal{D}_0 \cup \mathcal{D}_1$). Distilling on both produces the best result, while distilling on $\mathcal{D}_0$ increases accuracy and decreases fidelity, relative to $\mathcal{D}_1$. **Bottom:** We repeat the experiment, but fix $m = 3$ and vary τ. The shaded region corresponds to $\mu \pm \sigma$, estimated over 3 trials.

Optimization: train agreement

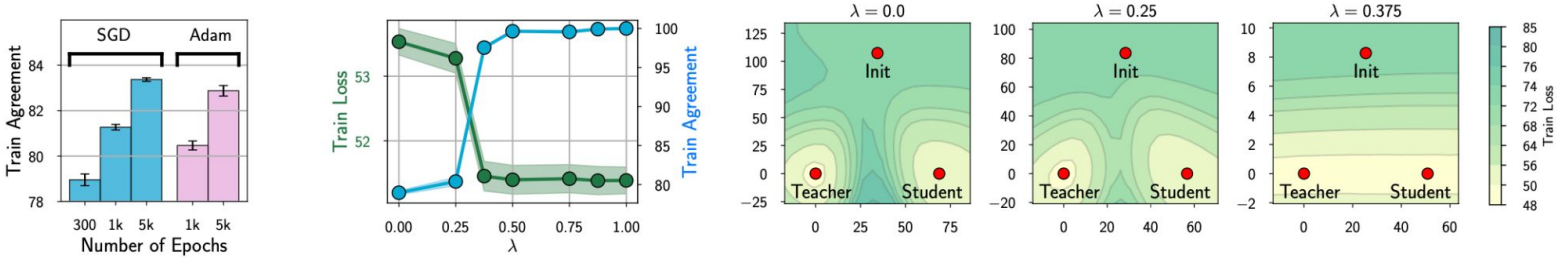
Если не виноваты данные, то что там с задачей оптимизации?

Давайте посмотрим на agreement на обучающей выборке для нескольких предыдущих экспериментов.



Чем больше и сложнее обучающая выборка, тем меньше train agreement (хотя с test agreement ситуация была обратная). Мы не справляемся найти ученика, который повторяет учителя даже на обучающей выборке! Вывод: наша общая интуиция о работе дистилляции имеет проблемы.

Optimization: больше экспериментов



(a) Optimizer effect (b) Initialization effect (c) Loss Visualization

Figure 6: **Optimization and distillation**: self-distillation with ResNet-20s with LayerNorm on CIFAR-100. **(a)**: Final train agreement for SGD and Adam optimizers. Training longer improves agreement, but it remains below 85% even after $5k$ epochs. **(b)**: Final train loss and agreement when the initialization is a convex combination of teacher and random weights, $\theta_s = \lambda\theta_t + (1-\lambda)\theta_r$. **(c)**: Projections of the distillation loss surface on the plane intersecting θ_t, the initial student weights, and the final student weights for different λ. When λ is small, the student converges to a suboptimal solution with low agreement. The uncertainty regions correspond to $\mu \pm \sigma$, estimated over 3 trials.

Не столь банальные ответы

- Лучшее accuracy ≠ Лучшее fidelity
- Задача оптимизации, решаемая при дистилляции, на самом деле сложная. Это является ключевой причиной, почему ученик имеет низкий fidelity.
- Трейдофф между сложностью оптимизации и качеством датасета: добавление данных для дистилляции за пределами обучающей выборки учителя позволяет найти более согласованного с учителем студента, но усложняет и так непростую задачу оптимизации.

Does knowledge distillation really work? In short: *Yes*, in the sense that it often improves student generalization. *No*, in that knowledge distillation often fails to live up to its name, transferring very limited knowledge from teacher to student.

Рецензия. Содержание и вклад

Статья исследует популярный метод сжатия / ускорения нейронных сетей – дистилляцию, опираясь на обобщающую способность и согласованность (насколько точно совпадают выход студента и учителя). Авторы провели большое количество экспериментов и выдвинули на основе этого несколько гипотез. Ключевым открытием является отсутствие корреляции между обобщающей способность (которую замеряли с помощью accuracy) и согласованностью (fidelity).

Рецензия. Сильные стороны

- Другие работы о дистилляции знаний предлагали свои подходы, исследовали, какой должна быть сеть-учитель и так далее. Однако какого-то общего исследования данного подхода (работает ли он как ожидается) до этой статьи не было.
- Статья актуальна, ведь задача сжатия нейронных сетей активно исследуется в наши дни. Это позволяет ускорить работу нейронной сети и снизить ресурсоемкость модели.
- Авторы провели обширный эмпирический анализ, эксперименты были проведены на разных данных (не только разных наборах, но и на разных типах данных: картинки, текст).

Рецензия. Слабые стороны

- Нет предложений как можно улучшить процедуру обучения модели студента, несмотря на интересные гипотезы в результате эксперимента.
- Исследования проведены только с классическим видом дистилляции знаний, однако, возможно, другие варианты вели бы себя по другому.

Рецензия

Статья написана на доступном языке, текст хорошо структурирован. Все сложные фрагменты сопровождаются комментариями и пояснениями.

Авторы предоставляют все необходимые детали для воспроизведения экспериментов. Кроме того, имеется хорошо организованный репозиторий с исходным кодом всех экспериментов на Github.

Оценка: 7
Уверенность в оценке: 4

Хакер. Описание

Задача – Key Word Spotting – здесь бинарная классификация (“есть ли слово в аудиозаписи?”)
Модель – Attention CRNN (paper)
Данные – Google Speech Commands
Метрика качества – AUC-FA-FR

Figure 1: .wav file in timeseries

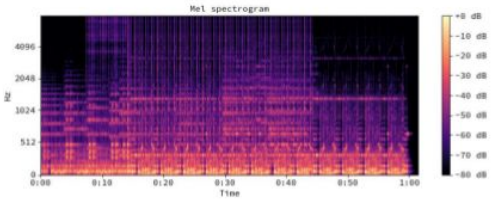


Figure 2: Mel Spectrogram Conversion

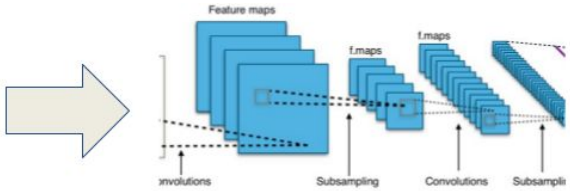
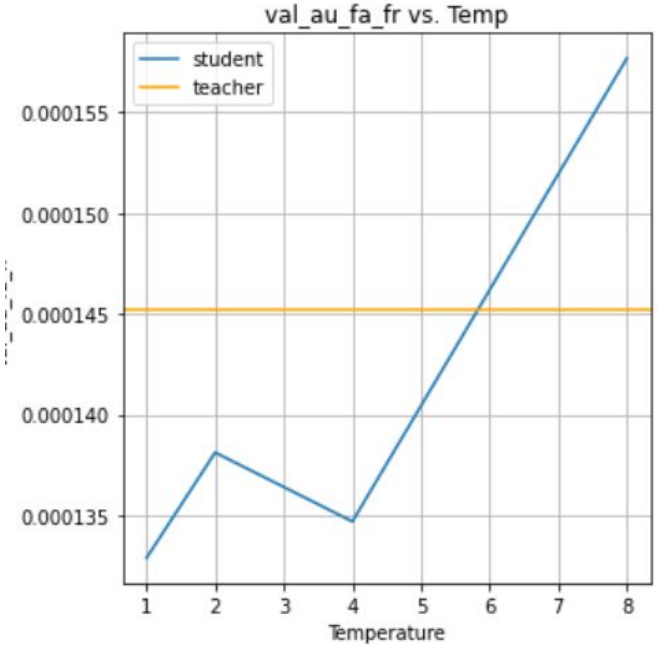
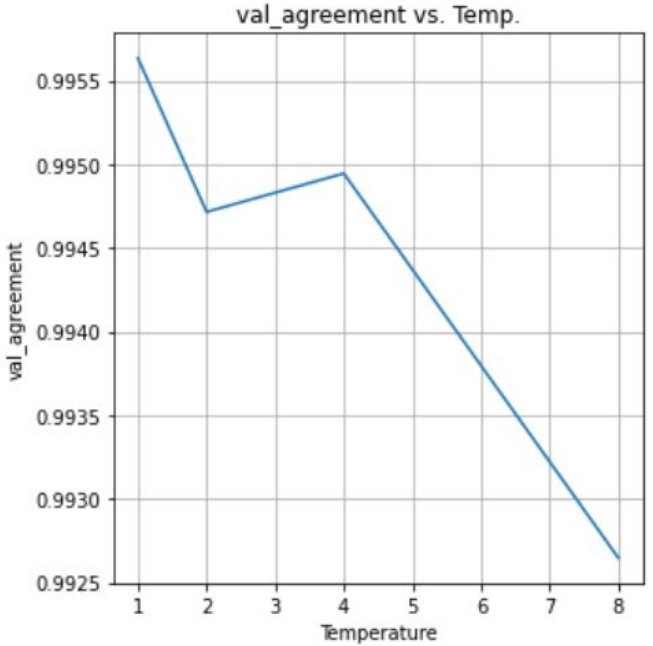


Figure 3: Classification

Эксперимент 1. Temperature

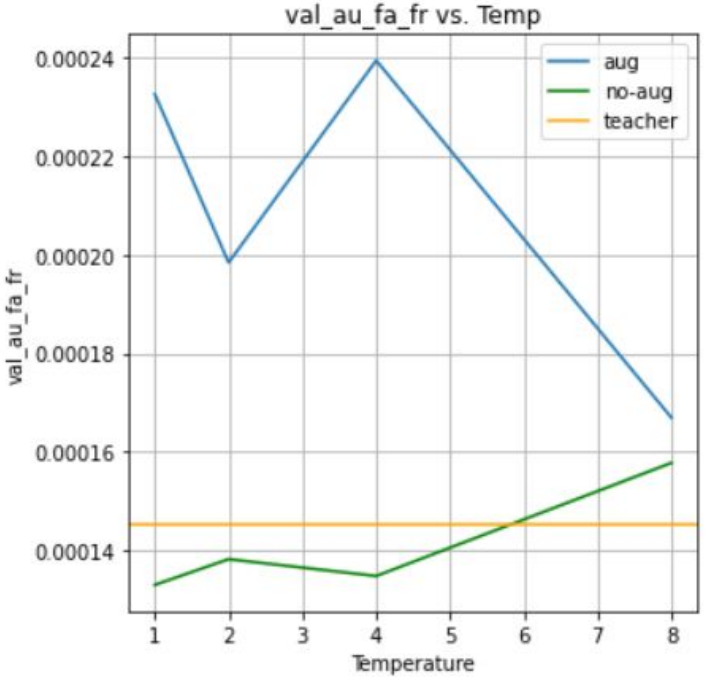
1. В self-distillation сеттинге, качество студента может быть выше учителя
2. Прямая зависимость качества от согласованности

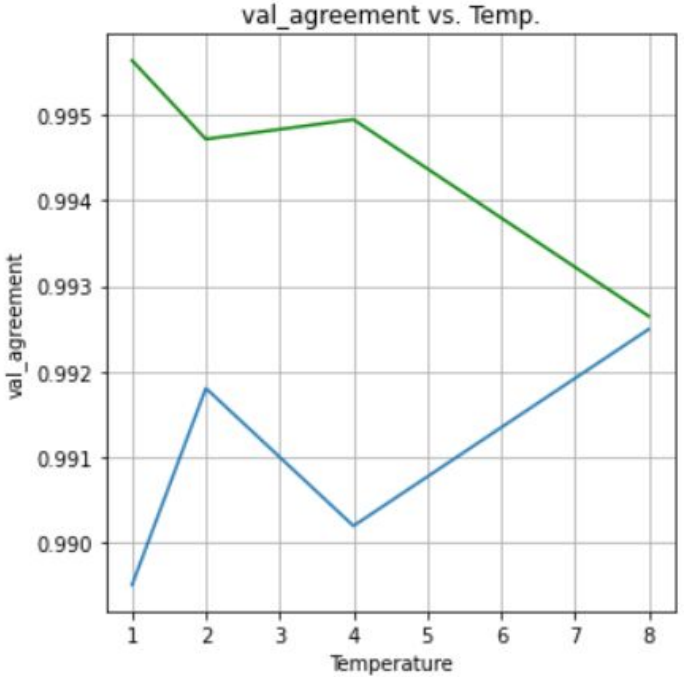




Эксперимент 2. Аугментации

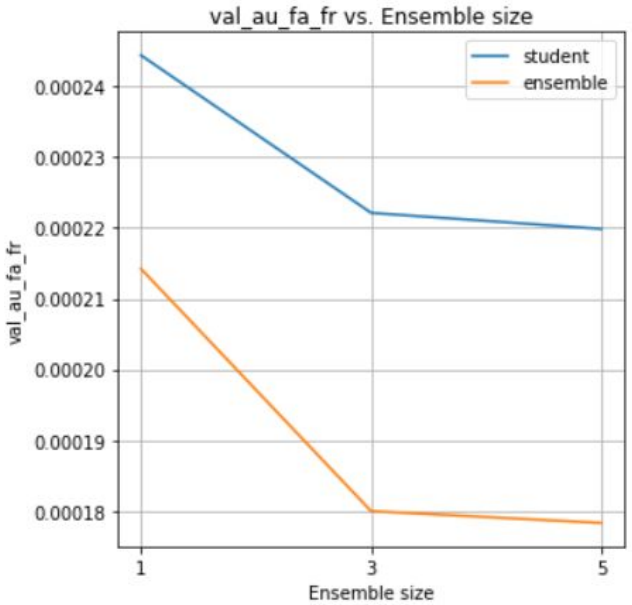
1. Тот же характер зависимости (качество - согласованность)
2. Качество с аугментациями пониже

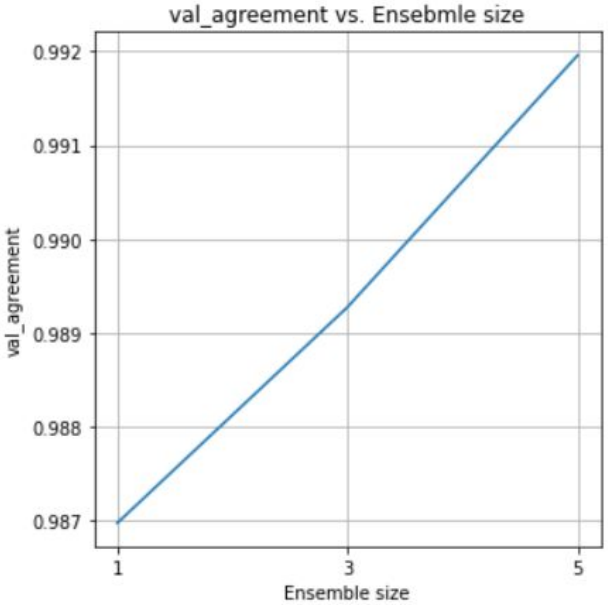




Эксперимент 3. Ансамбли

1. Снова прямая зависимость качества от согласованности.
2. Между студентом и ансамблем все равно gap.





Эксперимент 4. Другой способ дистилляции

Дистиллируем распределение attention’а

1. Качество и согласованность чуть похуже
2. Другой характер зависимости качества и согласованности!