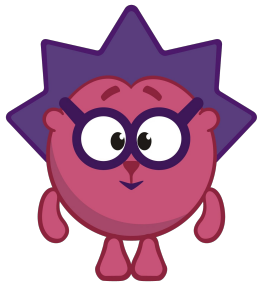
Does Knowledge Distillation Really Work?

Докладчик: Ольга Агапова
Рецензент: Дарья Барановская
Практик-исследователь: Артем Алекберов
Хакер: Артем Цыганов

Что вообще такое Knowledge Distillation?

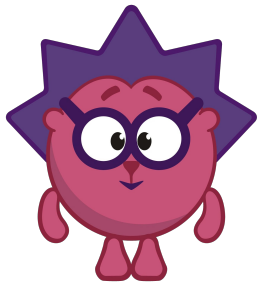
Проблема: большие модели не влезают в маленькие носители (например, в мобильное приложение),



Что вообще такое Knowledge Distillation?

Проблема: большие модели не влезают в маленькие носители (например, в мобильное приложение),

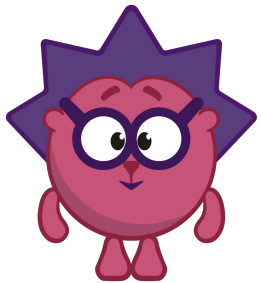
а еще дата-саентисты гораздо чаще занимаются улучшением маленьких моделей, и из-за этого их работа не обобщается на большие *(так сообщает автор изначальной статьи про K.D., а полезный видосик про это будет по ссылке в конце презентации)*



Что вообще такое Knowledge Distillation?

Идея: возьмем легковесную модель,

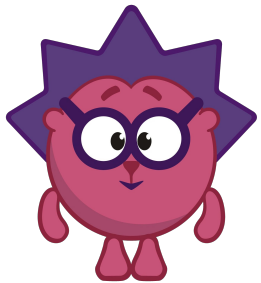
и научим ее имитировать поведение большой вычислительно сложной модели



Что вообще такое Knowledge Distillation?

Идея: возьмем легковесную модель, (ученик)

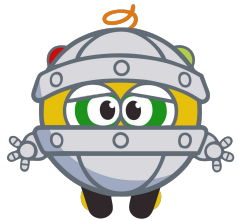
и научим ее имитировать поведение большой вычислительно сложной модели (учитель)



Как обучается ученик:

**подробные формулы на стр.3 статьи, ссылка в конце презентации*

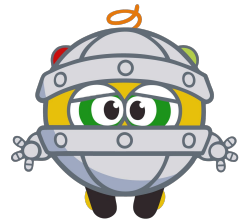
$$\mathcal{L}_s := \alpha \mathcal{L}_{\mathrm{NLL}} + (1-\alpha)\mathcal{L}_{\mathrm{KD}}$$



Как обучается ученик:

**подробные формулы на стр.3 статьи, ссылка в конце презентации*

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



лосс, который поощряет ученика копировать учителя

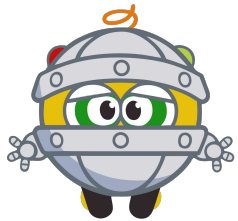
Как обучается ученик:

**подробные формулы на стр.3 статьи, ссылка в конце презентации*

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

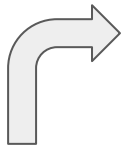
с точностью до константы – KL-дивергенция между эмпирическим распределением данных и распределением предсказаний ученика

лосс, который поощряет ученика копировать учителя



Какие метрики?

Для оценки имитирования учеником поведения учителя (*fidelity*):



среднее совпадение между ответами учителя и ученика по самому частому лейблу

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

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

Какие метрики?

Для оценки имитирования учеником поведения учителя (*fidelity*):

среднее совпадение между ответами учителя и ученика по самому частому лейблу

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

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

средняя KL-дивергенция между распределениями ответов учителя и ученика

Какие метрики?

Для оценки качества предсказаний ученика на незнакомых данных (*generalization*):

- top-1 accuracy *(кажется, тут речь идет тоже о top-1 лейбле)*
- expected calibration error (ECE)
- negative log-likelihood (NLL) *(та самая штука, которая похожа на KL-расстояние с точностью до константы)*

Что утверждают авторы нашей статьи?

KD работает, но:

Что утверждают авторы нашей статьи?

KD работает, но:

- разница между распределениями предсказаний учителя и ученика может быть больше, чем хотелось бы
- даже когда capacity позволяет ученику целиком повторять учителя, их результаты все равно различаются

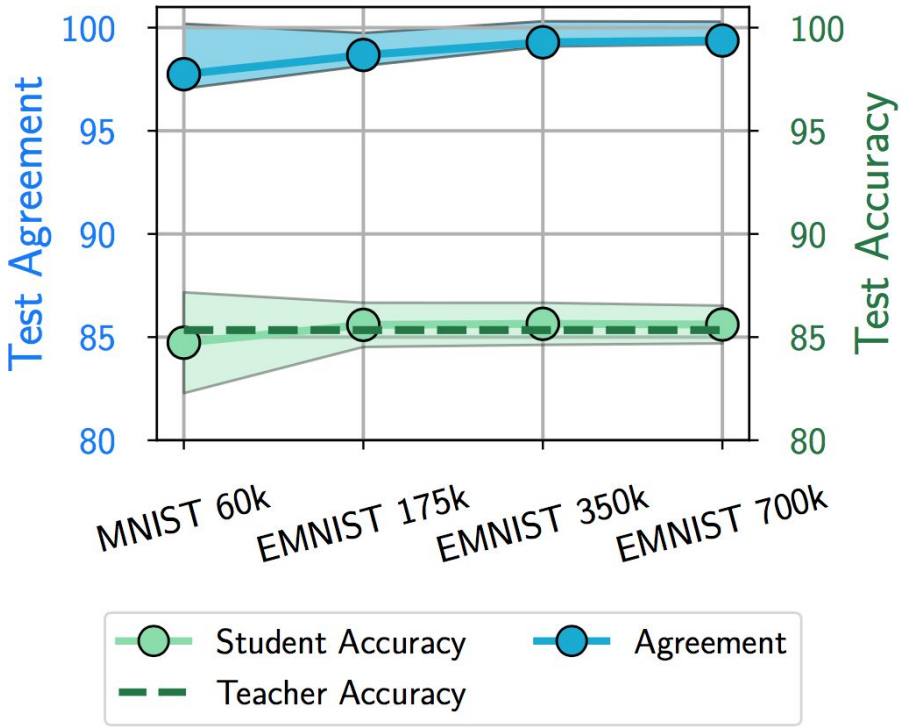
Что нам обещают в статье?

- показать, какие проблемы с оптимизацией не позволяют ученику в точности обучиться копировать учителя
- показать, какие нюансы в данных влияют на качество этого “мэтча”

- показать, почему точное “натаскивание” на учителя не обязательно дает хорошую генерализацию

1. *Knowledge* Distillation не очень хорошо сохраняет *Knowledge*

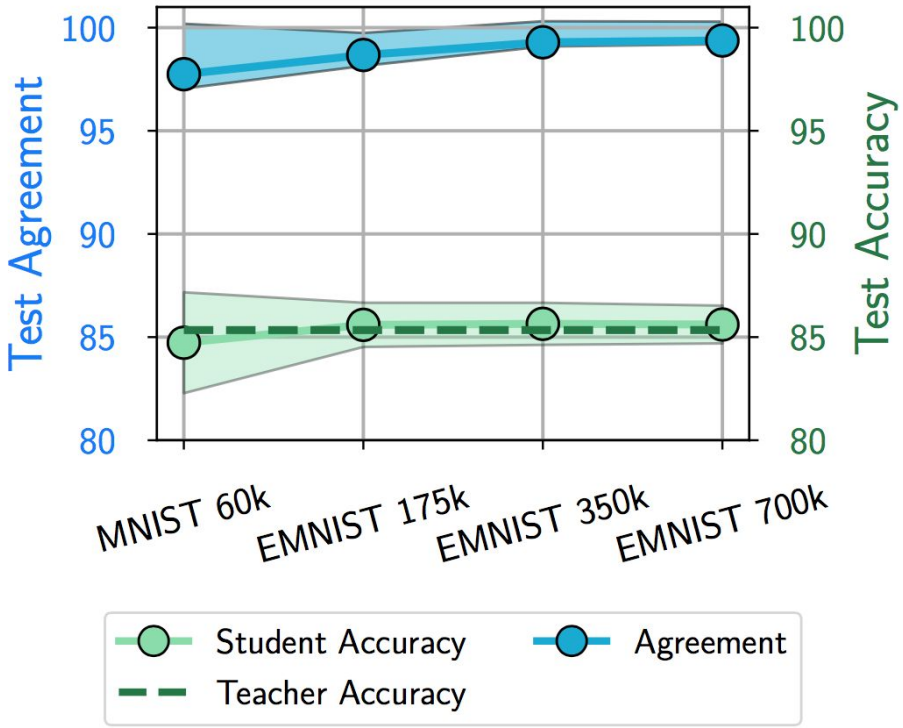
1. *Knowledge* Distillation не очень хорошо сохраняет *Knowledge*



Учитель – LeNet-5

Ученик – тоже LeNet-5

1. *Knowledge* Distillation не очень хорошо сохраняет *Knowledge*



Учитель – LeNet-5

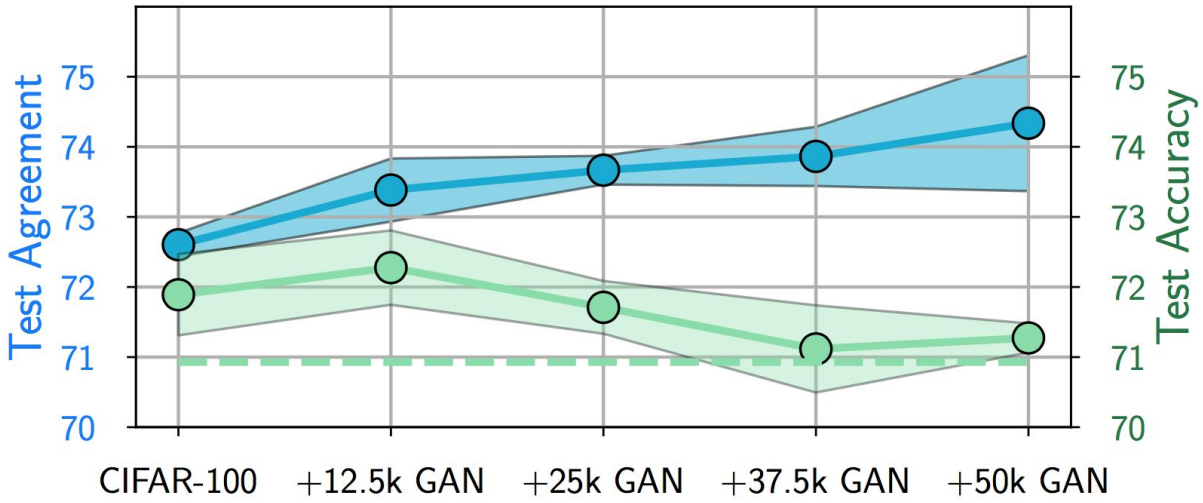
Ученик – тоже LeNet-5

Top-1 Agreement – 99%,

accuracy у учителя и ученика тоже почти одинаковая.

То есть и с fidelity, и с generalization все хорошо.

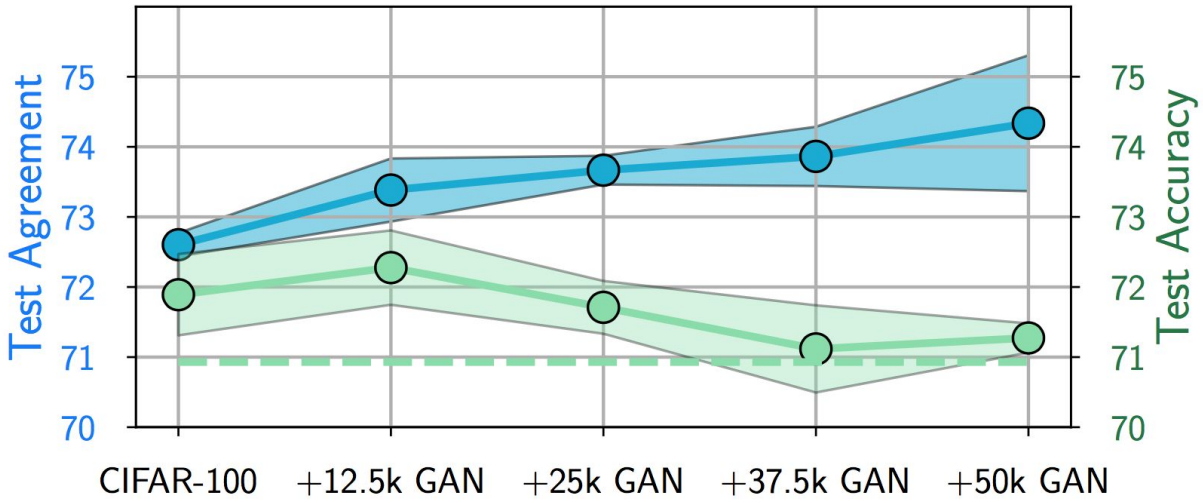
1. *Knowledge* Distillation не очень хорошо сохраняет *Knowledge*



Учитель – ResNet-56

Ученик – тоже

1. *Knowledge* Distillation не очень хорошо сохраняет *Knowledge*



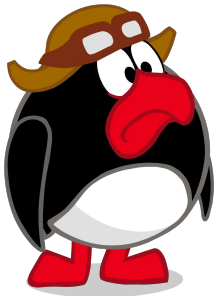
Учитель – ResNet-56

Ученик – тоже

Fidelity растет с увеличением кол-ва данных, но test accuracy падает

Что выяснили к этому моменту?

- при self-distillation (это когда ученик и учитель одинаковой архитектуры) ученик может превзойти учителя в плане accuracy, но только пожертвовав fidelity
- точно копирующий поведение учителя ученик его не превзойдет (логично)



Зачем вообще заботиться о fidelity, если accuracy и так нормальный?

- интерпретируемость

  утверждается, что если большую black-box модель утрамбовать в маленькую, то это поможет человеку понять закономерности, которые установила внутри black-box большая модель

Зачем вообще заботиться о fidelity, если accuracy и так нормальный?

- интерпретируемость

  утверждается, что если большую black-box модель утрамбовать в маленькую, то это поможет человеку понять закономерности, которые установила внутри black-box большая модель

- лучшая репрезентация “знания”

  авторы провели эксперимент, который показал, что у больших учителей (например ансамблей) и маленьких учеников часто получается большой gap в генерализации при оптимизации именно fidelity.

Почему бывает низкий fidelity?

- вместимость ученика
- архитектура сетей
- сложность и размерность данных
- data domain *(вроде происхождение данных)*
- идентифицируемость
- оптимизация

Почему бывает низкий fidelity?

- вместимость ученика
- архитектура сетей
- сложность и размерность данных
- data domain *(вроде происхождение данных)*

кратко изучено и описано в экспериментах в конце статьи

- идентифицируемость
- оптимизация

2. Идентифицируемость: правильные ли мы используем данные?

- Нужно ли использовать больше данных (пар типа “вход-ответ учителя”)?

  провели эксперимент: стали использовать больше разных аугментаций, выяснили, что аугментации, оптимальные для fidelity и для accuracy – разные;

2. Идентифицируемость: правильные ли мы используем данные?

- Нужно ли использовать больше данных (пар типа “вход-ответ учителя”)?

  провели эксперимент: стали использовать больше разных аугментаций, выяснили, что аугментации, оптимальные для fidelity и для accuracy – разные;

  и что разнообразие аугментаций к заметным улучшениям не ведет (agreement не выше 86%)

2. Идентифицируемость: правильные ли мы используем данные?

- Data Recycling Hypothesis

  гипотеза: использование для дистилляции тех же данных, на которых учили учителя, рискованно.

  эксперимент: разбили train data пополам, на **D-0** и **D-1**.

2. Идентифицируемость: правильные ли мы используем данные?

- Data Recycling Hypothesis

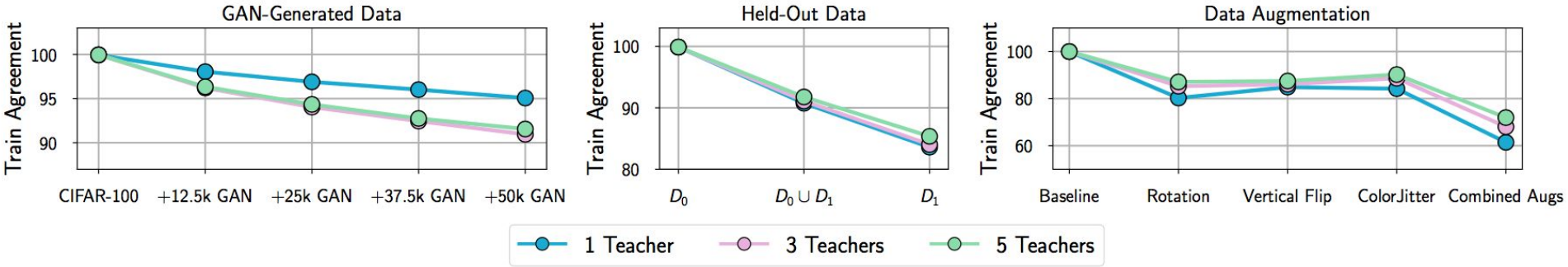
  гипотеза: использование для дистилляции тех же данных, на которых учили учителя, рискованно.

  эксперимент: разбили train data пополам, на **D-0** и **D-1**. На **D-0** обучили учителя, а потом сравнили трех учеников:

  - обученного на **D-0**
  - обученного на **D-1**
  - обученного на **D-0** и **D-1**

2. Идентифицируемость: правильные ли мы используем данные?

- Data Recycling Hypothesis
- результат эксперимента: гипотеза, что у ученика на D-1 будет выше fidelity, чем у ученика на D-0, подтверждается, но accuracy у него не выше
- ученик на обеих половинах сочетает в себе лучшие качества остальных, но это не дает большого прироста все равно (agreement 85%)



3. Оптимизация

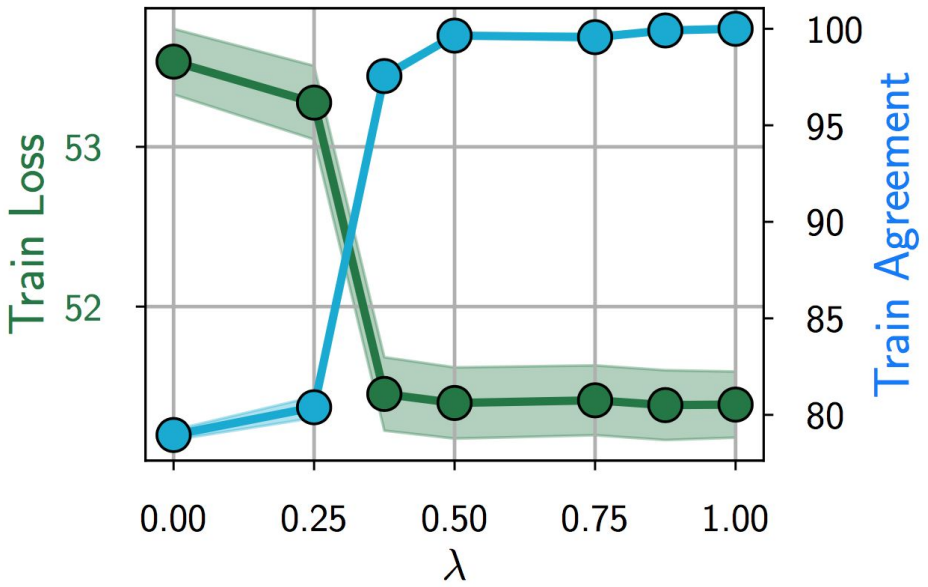
- Попробовали менять оптимизатор (SGD на Adam), fidelity только упал
- Больше эпох помогает, но не сильно *(как всегда)*

3. Оптимизация

- Попробовали менять оптимизатор (SGD на Adam), fidelity только упал
- Больше эпох помогает, но не сильно *(как всегда)*
- “У нас не получилось дистиллировать ResNet-20 на CIFAR-100, но есть ли другая постановка задачи, в которой получится высокий fidelity?”

3. Оптимизация

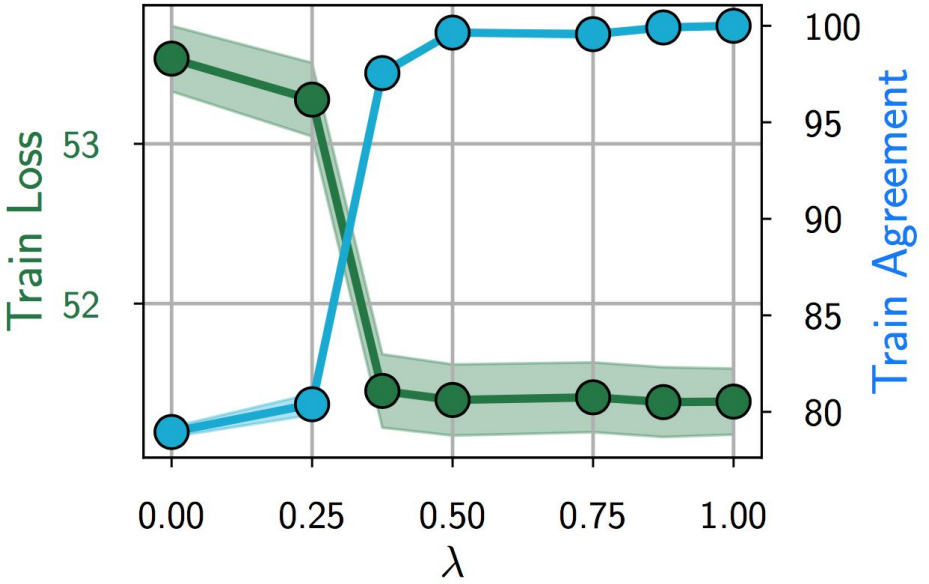
Идея: инициализировать веса ученика по-другому – либо весами учителя, либо рандомно, либо взвешенной суммой



3. Оптимизация

Идея: инициализировать веса ученика по-другому – либо весами учителя, либо рандомно, либо взвешенной суммой

Результат: если ученик инициализирован далеко от учителя (коэф. < 0.25), то он с ним ***не соглашается*** *(это довольно хороший ученик).*

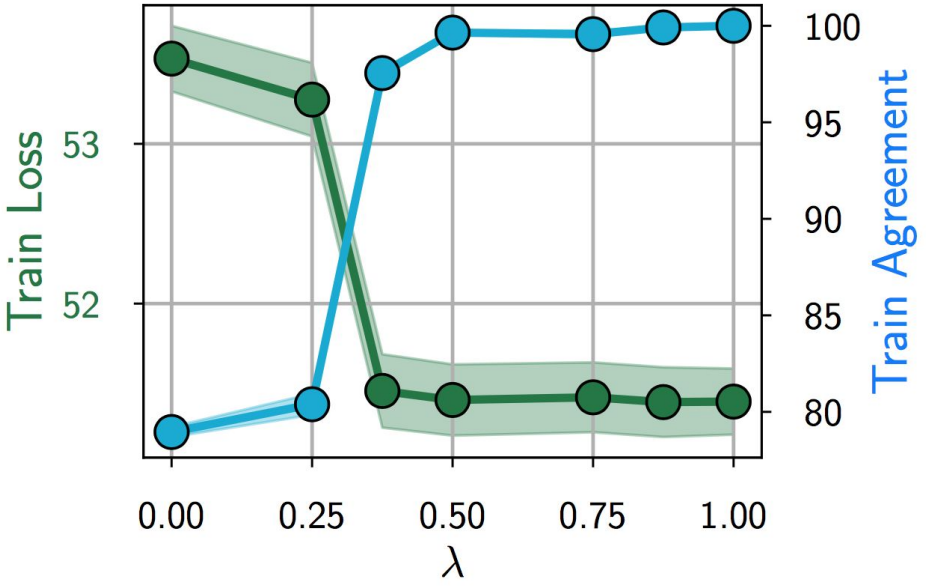


3. Оптимизация

Результат: если ученик инициализирован далеко от учителя (коэф. < 0.25), то он с ним ***не соглашается*** *(это довольно хороший ученик).*

Но при коэф. > 0.375, картина меняется и fidelity растет.

Оптимальная инициализация – с коэффициентом в районе 0.375



Выводы

Есть обмен между качеством дистилляции и сложностью оптимизации, и оптимизировать в этом процессе сложно

Хорошая fidelity не значит хорошая generalization

Рецензент

Положительные качества:

- Статья очень аккуратно написана, и ее легко читать. Вводящиеся термины подчеркнуты курсивом, есть обобщающие абзацы в конце разделов.
- Авторы проводят эксперименты по несколько раз для большей точности и предоставляют нам информацию об этом на графиках
- Экспериментов много и они интересные. (Например, статья также затрагивает вопрос инициализации весов ученика и аугментации данных)
- Структурированный код экспериментов на github

Отрицательные качества:

- Некоторые моменты в статье опущены (например, указаны формулы для измерения метрик fidelity, но формулы ECE придется гуглить)
- Эксперименты на Cifar и Mnist
- Несмотря на аккуратные выводы, авторами не дано каких-то конкретных рекомендаций по дистилляции нейросетей

Оценки рецензентов: 6, 7, 7, 7, 5

Уверенность рецензентов: 4, 4, 3, 4, 3

Моя оценка: 7

Уверенность: 4

Практик-исследователь

Хакер

Полезные ссылки

Сама статья: вот

Приятное видео про KD на 12 минут: вот