Машинное обучение (Machine Learning)

Distillation and Batch Normalization

Уткин Л.В.

Санкт-Петербургский политехнический университет Петра Великого

Model Distillation

Model Distillation

Jianping Gou et al. Knowledge Distillation: A Survey. arXiv:2006.05525

G. Hinton, O. Vinyals, J. Dean. Distilling the Knowledge in a Neural Network, arXiv:1503.02531

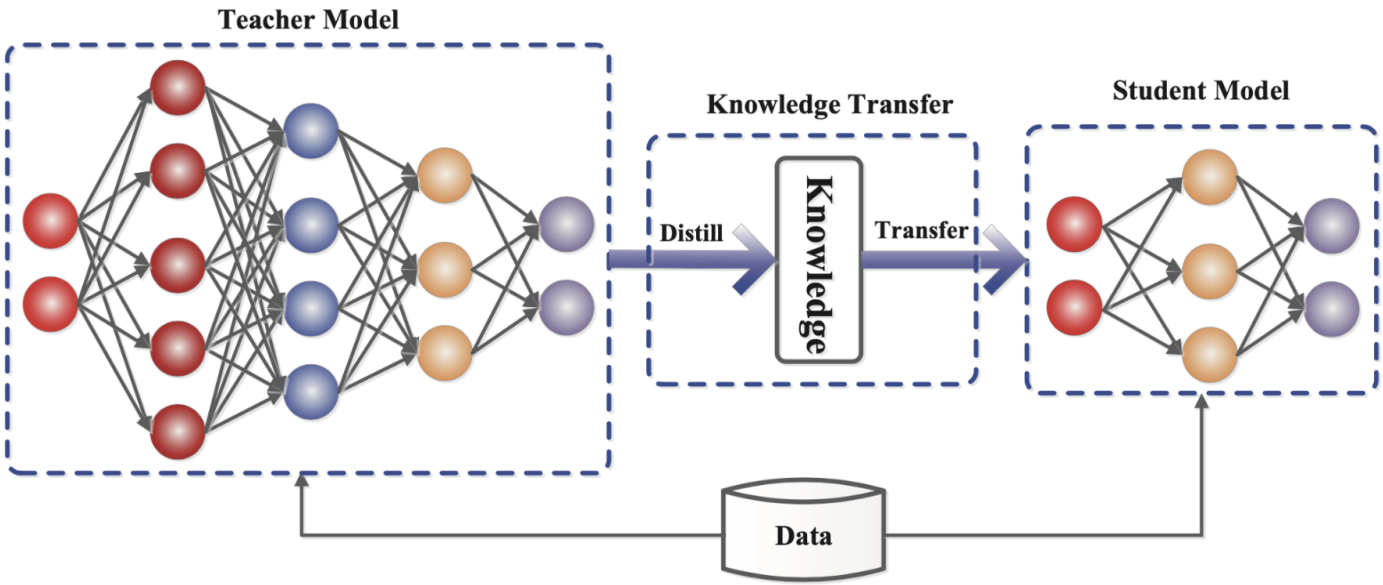
Distillation (idea)

- G. Hinton, O. Vinyals, J. Dean. Distilling the Knowledge in a Neural Network, arXiv:1503.02531.
- Мотивация: использовать результаты обучения “сложной” модели для обучения “простой”
- Отличие transfer learning и distillation: в последнем перенос обобщения (transfer of generalization)
- Понятия: сети учитель и студент, понятие температуры в softmax, “темные” знания (dark knowledge) и мягкие вероятности

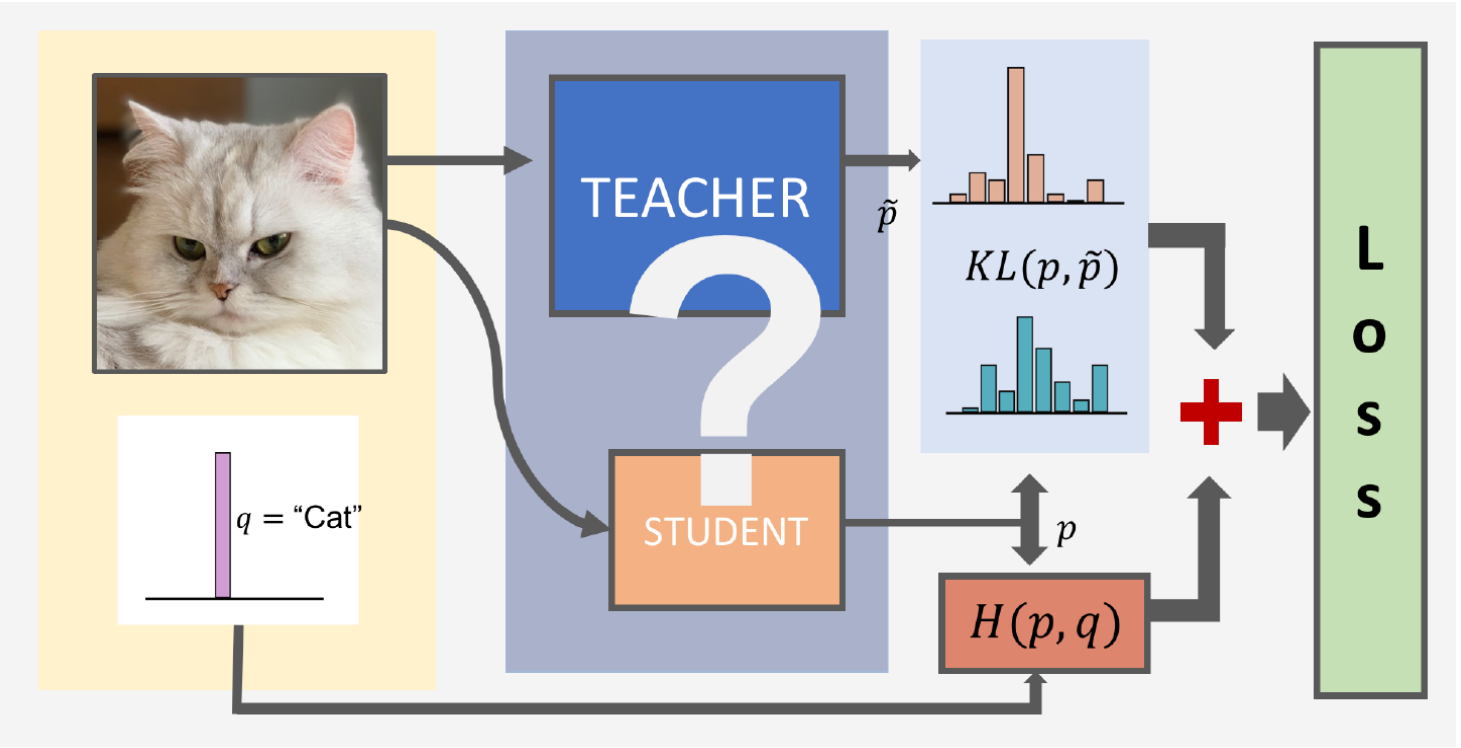
Distillation (модель учитель-студент)

- Учитель - это “сложная” глубокая нейронная сеть, которая была обучена на большом количестве данных (или любая другая модель - ансамбль) с хорошим обобщением
- Студент - это “простой” сеть, цель - выучить большинство обобщений учителя и оставаться “простой”

Distillation (модель учитель-студент)



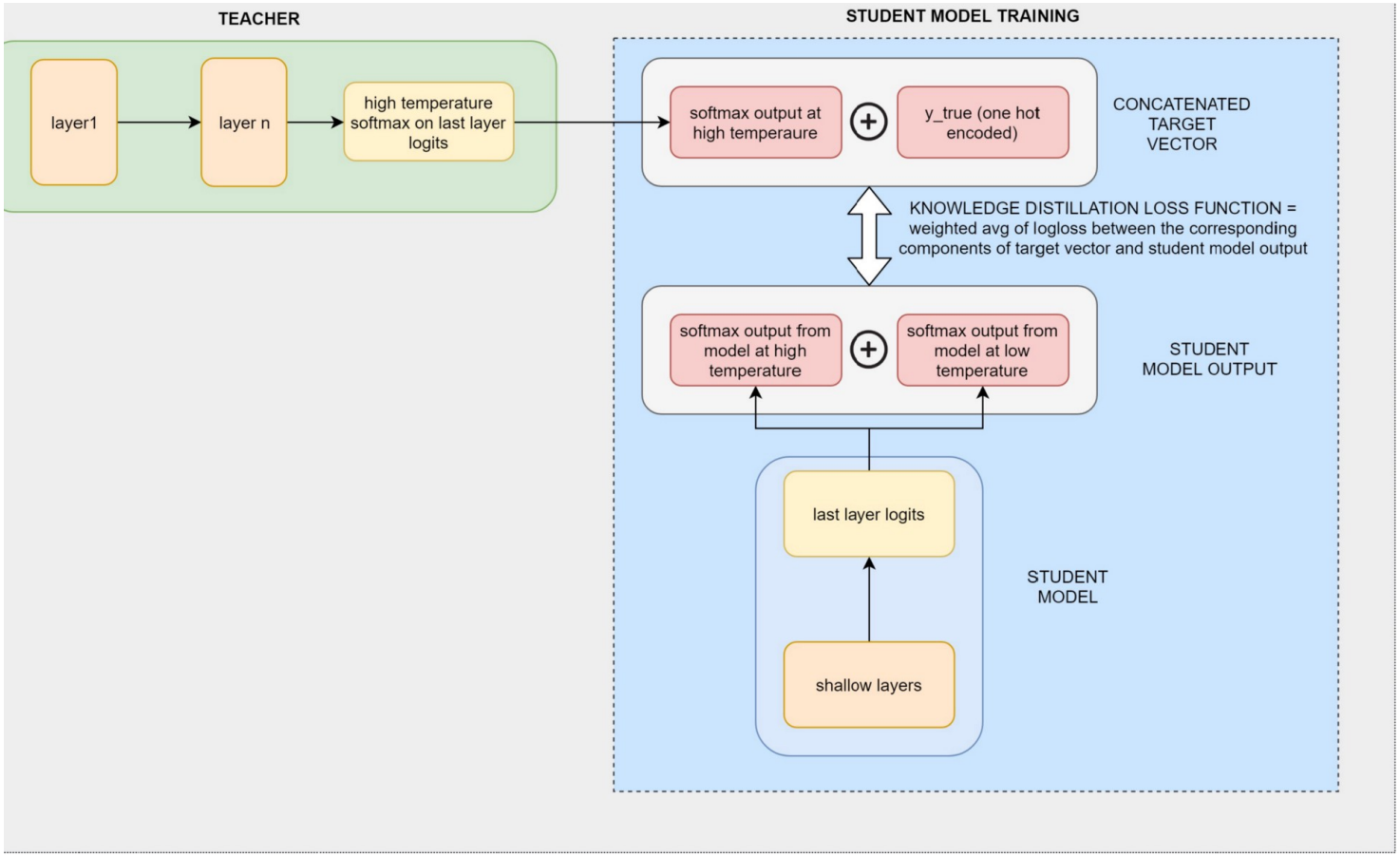
Distillation (модель учитель-студент)



J.H. Cho and B. Hariharan. On the Efficacy of Knowledge Distillation.

arXiv:1910.01348v1

Distillation (модель учитель-студент)



Distillation (температура)

- Softmax возвращает вероятности каждого класса от 0 до 1, и их сумма =1, целевой класс имеет высокую вероятность

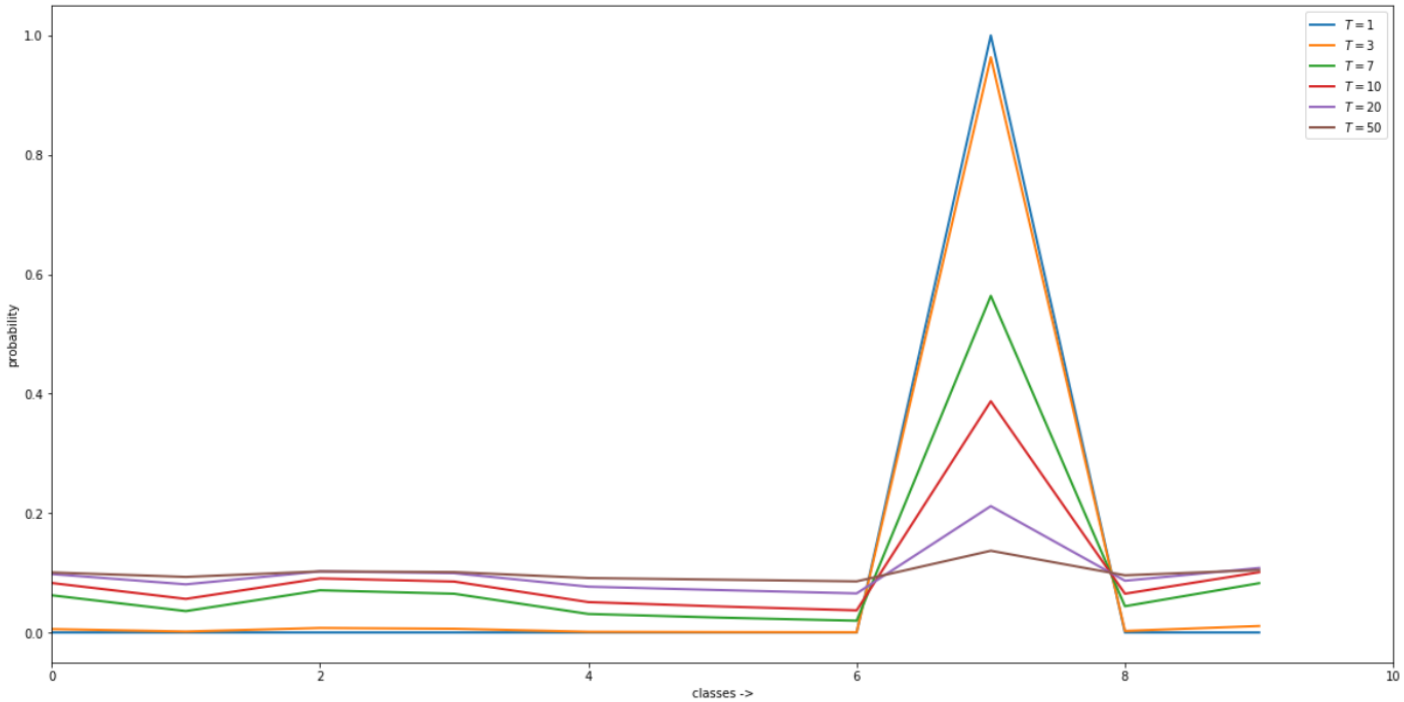
$$p_t(z_i) = \frac{\exp(z_i)}{\sum_j \exp(z_j)}$$

- Softmax с температурой

$$p_t^*(z_i, T) = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)}$$

- Больше температура - более размыты вероятности классов

Distillation (температура и MNIST)



Вероятности цифр MNIST, классифицируется цифра 7

Distillation (dark knowledge)

один семь или один?

Distillation (dark knowledge)

- Модель дает более высокую вероятность для 1 одновременно прогнозируя 7 при высокой T
- Человек не может количественно определить, насколько 7 выглядит ближе к 1, а “высокотемпературная” модель делает это
- Т.о. “высокотемпературная” модель обладает “темными” знаниями - в дополнение к предсказанию числа 7, она также хранит информацию о том, насколько это число 7 напоминает число 1
- “Низкотемпературная” модель (обычная модель) хороша для точных прогнозов, но теряем эти “темные” знания
- Основная идея distillation - передача “темных” знаний от обученного учителя к простой модели студента

Distillation (обучение студента)

- Модель студента обучается при той же высокой температуре, что и учитель
- Функция потерь для студента

$$L = \alpha L_{\text{cross entopy}} + (1 - \alpha) L_{\text{knowledge distil.}}$$

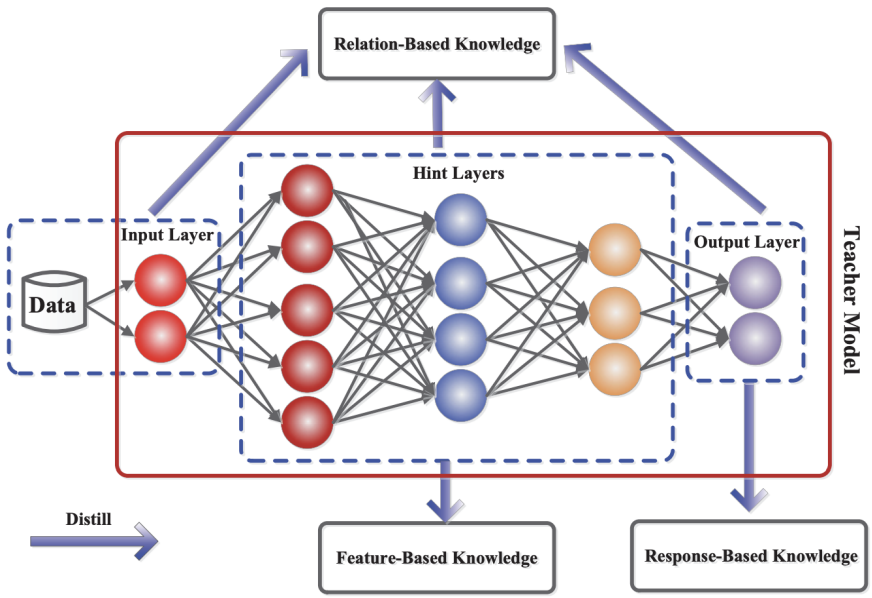
$$L_{\text{knowledge distil.}} = -\tau \sum_i p_t^*(z_i, T) \ln p_s^*(z_i, T)$$

- Модель студента тестируется с обычной активацией softmax (т.е. без температуры).

Типы знаний

- Response-based knowledge
- Feature-based knowledge
- Relation-based knowledge
- Source: https://arxiv.org/pdf/2006.05525.pdf

Типы знаний

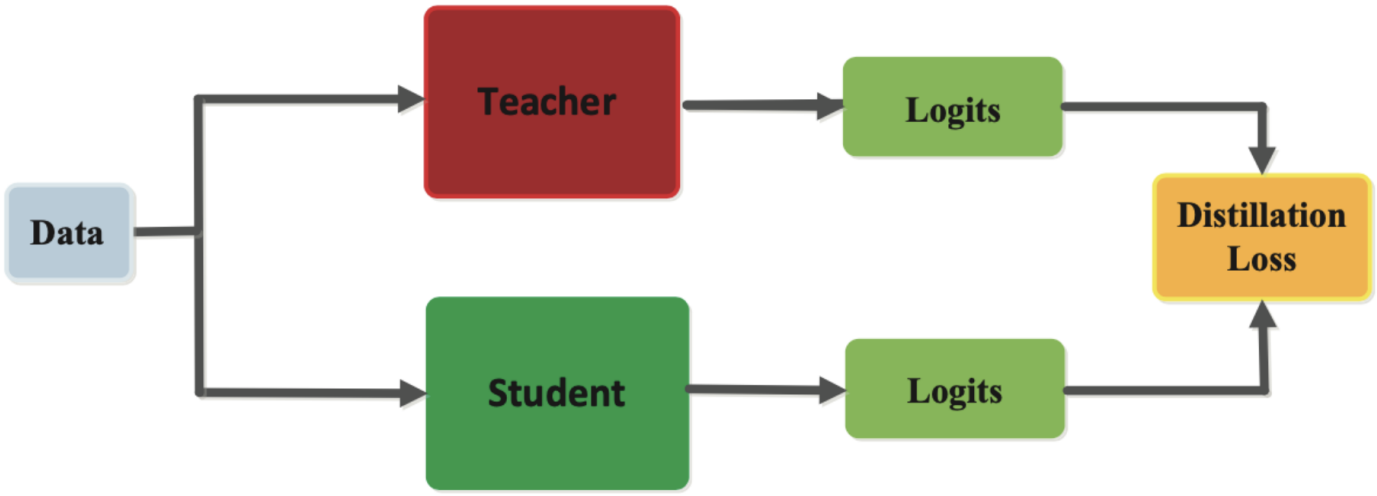


Source: https://arxiv.org/pdf/2006.05525.pdf

Response-based knowledge

- Знания, основанные на ответах, фокусируются на конечном выходном слое модели учителя.
- Гипотеза состоит в том, что модель ученика научится имитировать предсказания модели учителя, используя distillation loss, которая фиксирует разницу между логитами модели ученика и модели учителя соответственно.

Response-based knowledg (схема)

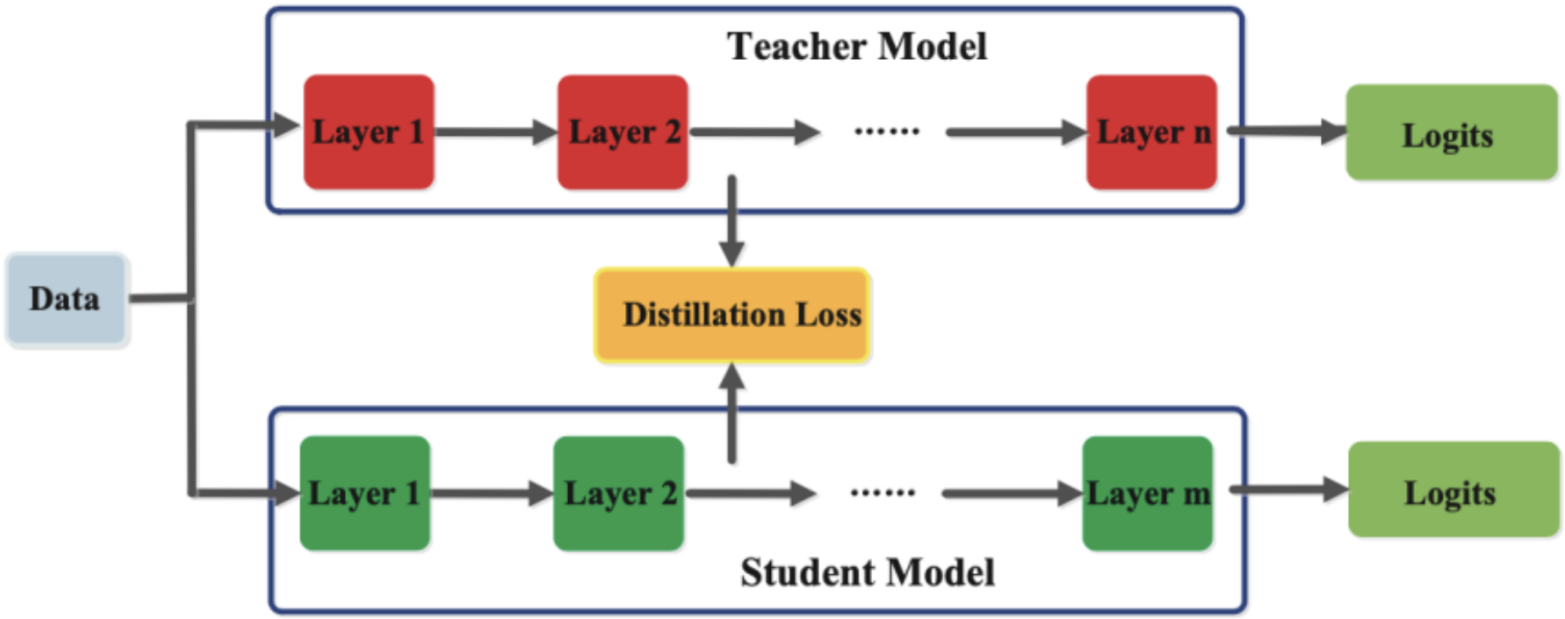


Source: https://arxiv.org/pdf/2006.05525.pdf

Feature-based knowledge

- Модель обученного учителя также фиксирует знания о данных в своих промежуточных слоях (для глубоких нейронных сетей).
- Промежуточные слои учатся различать определенные признаки, и эти знания можно использовать для обучения модели ученика.
- Цель - обучить модель ученика так, чтобы получить такие же карты признаков, что и модель учителя.
- Distillation loss минимизирует разницу между картами признаков моделей учителя и ученика.

Feature-based knowledge (схема)



Source: https://arxiv.org/pdf/2006.05525.pdf

Feature-based knowledge (обучение)

- Distillation loss:

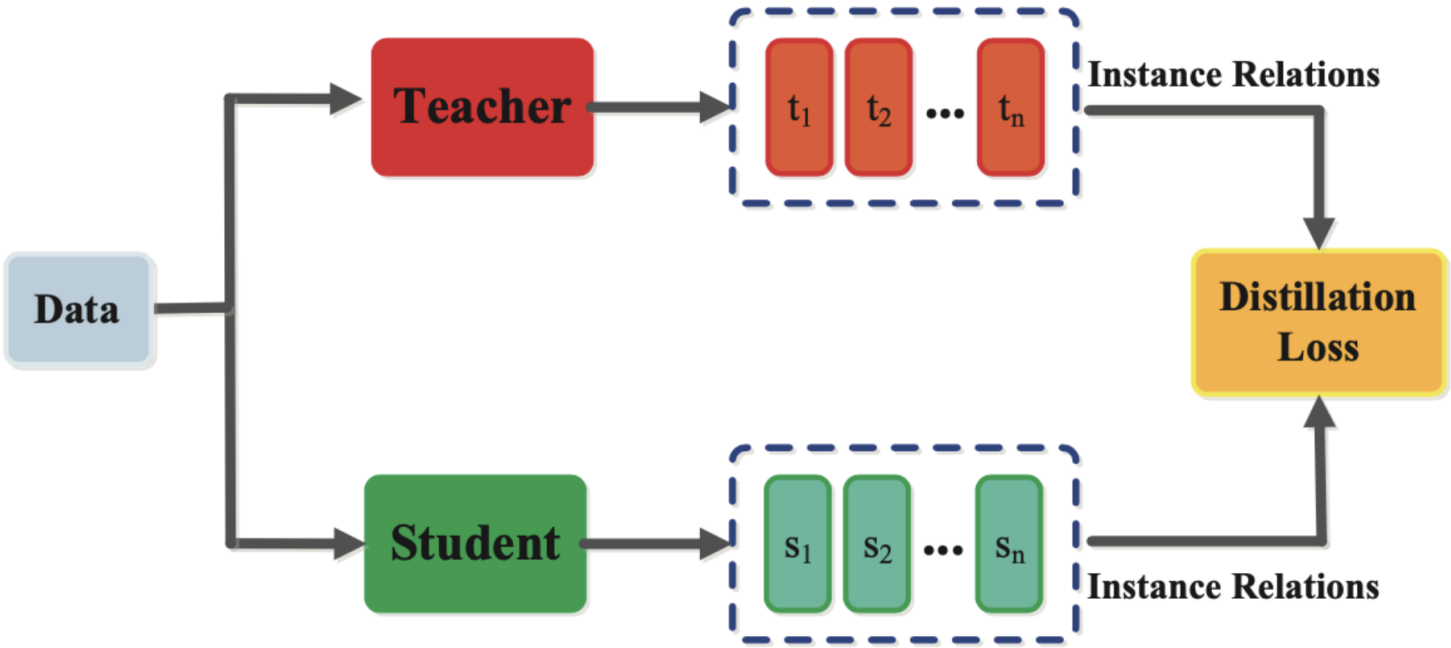
$$L(f_t(x), f_s(x)) = \mathcal{L}_F\left(\Phi_t(f_t(x)), \Phi_s(f_s(x))\right)$$

- $f_t(x)$ и $f_s(x)$ - карты признаков промежуточных слоев моделей учителя и студента
- $\Phi_t(f_t(x))$ и $\Phi_s(f_s(x))$ - функции трансформации, применяемые, когда карты признаков учителя и ученика имеют разный масштаб или размер
- $\mathcal{L}_F$ - функция схожести карт признаков учителя и студента (норма L_2 или L_1, кросс-энтропия и т.д.)

Relation-based knowledge

- В дополнение к знаниям, представленным в выходных слоях и промежуточных слоях нейронной сети, знания, которые отражают взаимосвязь между картами признаков, также могут использоваться для обучения модели ученика.
- Эти отношения можно смоделировать как корреляцию между картами признаков, матрицей подобия, эмбедингами или распределениями вероятностей на основе представлений признаков

Relation-based knowledge (схема)



Source: https://arxiv.org/pdf/2006.05525.pdf

Relation-based knowledge (обучение)

- Distillation loss на основе отношения карт **признаков**:

$$L(f_t(x), f_s(x)) = \mathcal{L}_{R1}\left(\Psi_t(\widehat{f_t}, \widetilde{f_t}), \Psi_s(\widehat{f_s}, \widetilde{f_s})\right)$$

- $f_t(x)$ и $f_s(x)$ - карты признаков промежуточных слоев моделей учителя и студента
- Пары карт признаков выбраны из модели учителя $\widehat{f_t}, \widetilde{f_t}$ и модели ученика $\widehat{f_s}, \widetilde{f_s}$
- $\Psi_t()$ и $\Psi_s()$ - функции близости пар карт признаков учителя и ученика
- $\mathcal{L}_{R1}$ - функция корреляции между картами признаков учителя и студента

Relation-based knowledge (обучение)

- Distillation loss на основе отношения **примеров**:

$$L(F_t, F_s) = \mathcal{L}_{R1}\left(\psi_t(t_i, t_j), \psi_s(s_i, s_j)\right)$$

- t_i, t_j и s_i, s_j - представления признаков для примеров учителя и студента
- $\psi_t()$ и $\psi_s()$ - функции близости пар представлений признаков
- $\mathcal{L}_{R1}$ - функция корреляции между представлениями признаков учителя и студента

Основные схемы distillation (offline distillation)

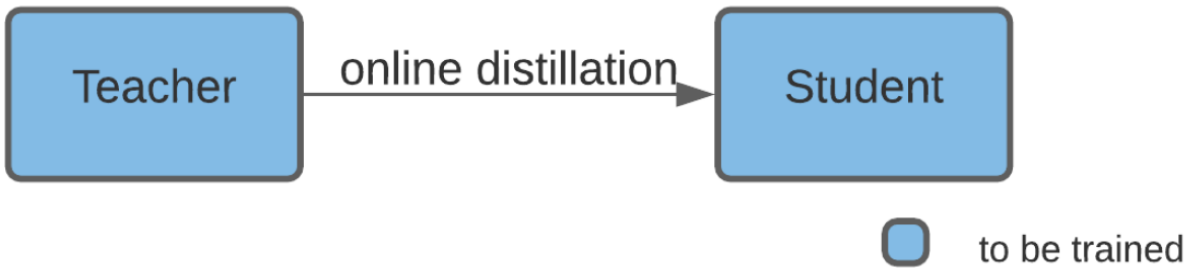
- Автономная дистилляция означает перенос знаний из обученной модели учителя в модель ученика.
- Процесс состоит из двух этапов:
  - обучение модели учителя перед дистилляцией;
  - знания в форме логитов или промежуточных признаков используются для обучения модели ученика.
- Концентрируется на улучшении различных аспектов передачи информации.
- Основное преимущество автономных методов - они просты и быстры в использовании.

Основные схемы distillation (offline distillation)

Основные схемы distillation (online distillation)

- Модели учителя и ученика обновляются одновременно в онлайн-дистилляции, и вся структура дистилляции знаний поддается обучению от начала до конца.
- Это новый способ заставить несколько нейронных сетей “сотрудничать” в глубоком взаимном обучении.
- Вариант онлайн-дистилляции, совместная дистилляция, используется для обучения крупномасштабной распределенной нейронной сети. Это процесс, в котором несколько моделей обучаются параллельно с одной и той же архитектурой.

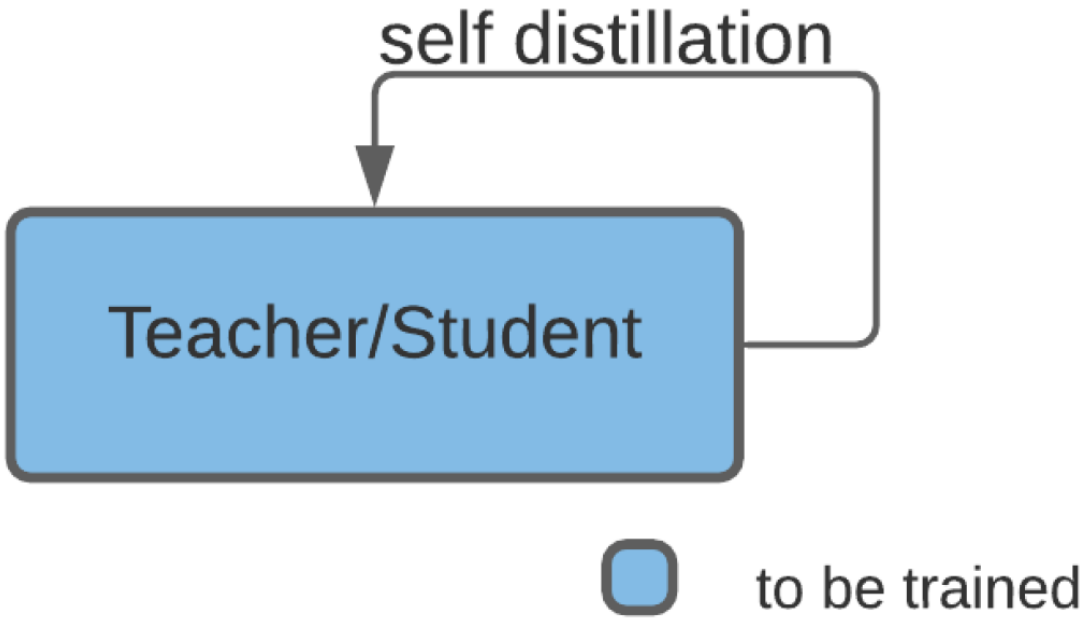
Основные схемы distillation (online distillation)



Основные схемы distillation (self-distillation)

- **Self-distillation**: одни и те же сети используются для моделей учителя и ученика при self-distillation.
- Это можно рассматривать как частный случай онлайн-дистилляции. В частности, знания “перегоняются” из более глубоких участков сети в ее shallow-участки.
- Особый вариант self-distillation, называемый “перегонкой” моментальных снимков, при котором знания из более ранних эпох сети (учителя) переносятся в ее более поздние эпохи (ученик) для поддержки контролируемого процесса обучения в той же сети.

Основные схемы distillation (self-distillation)



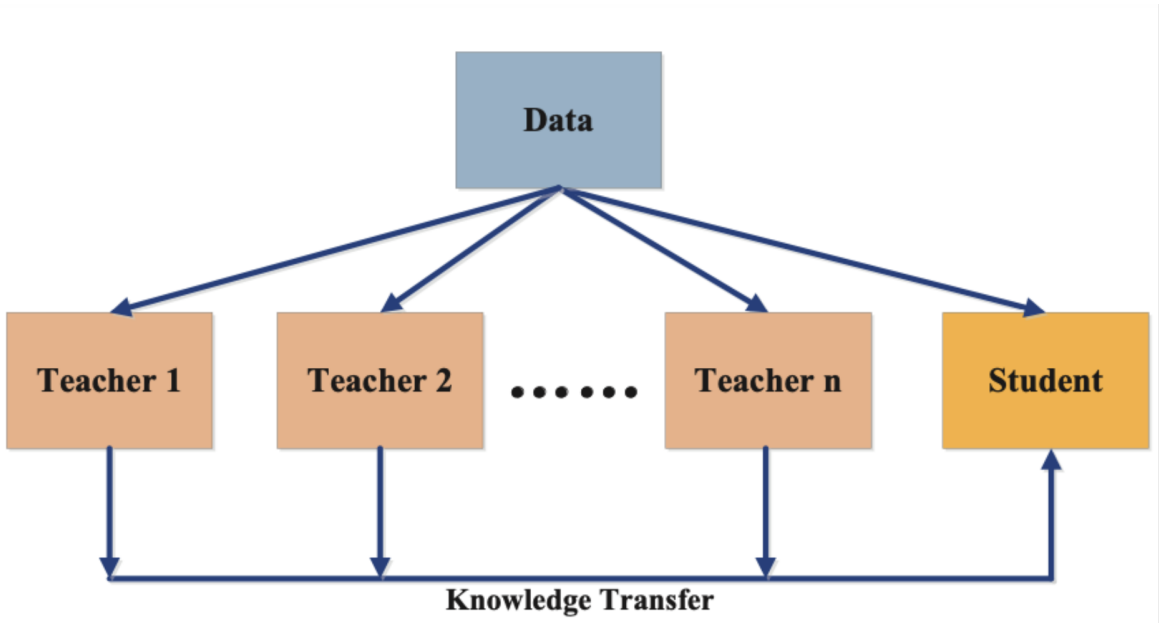
Основные схемы distillation

- Офлайн, онлайн и self-distillation могут быть интуитивно поняты с точки зрения обучения между учителем и учеником:
  - дистилляция в автономном режиме означает, что учитель учит ученика чему-то, чтобы ученик получил знания об этом;
  - онлайн-дистилляция относится к тому, когда и учитель, и ученик учатся вместе;
  - self-distillation относится к тому, когда учащийся усваивает знания самостоятельно.
- Эти три типа дистилляции можно комбинировать, чтобы дополнять друг друга в зависимости от индивидуальных преимуществ.

Multi-Teacher distillation

- Модель ученика получает знания от нескольких разных моделей учителей.
- Использование ансамбля моделей учителей может предоставить модели ученика различные виды знаний. Несколько учителей могут передавать разные виды знаний
- Знания от нескольких учителей можно объединить в качестве среднего ответа по всем моделям.

Multi-Teacher distillation (схема)

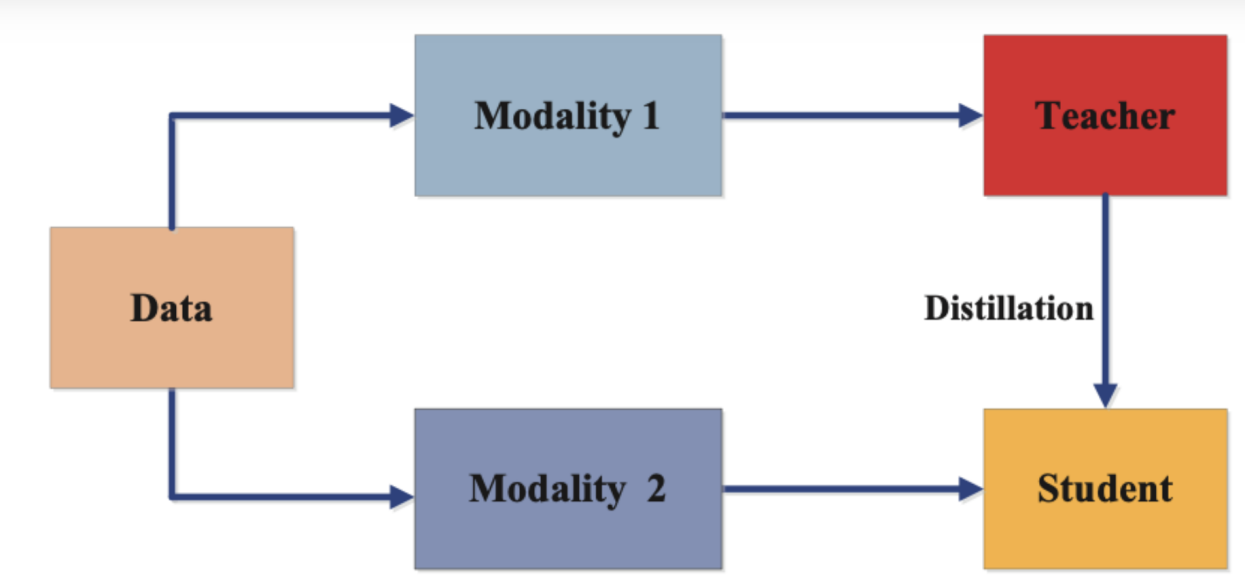


Source: https://arxiv.org/pdf/2006.05525.pdf

Cross-modal distillation

- Учитель обучается одной модальности, а его знания передаются ученику, которому требуются знания из другой модальности.
- Когда данные или метки недоступны для конкретных модальностей ни во время обучения, ни во время тестирования, что требует передачи знаний между модальностями.
- Например, знания учителя, обученного работе с размеченными данными изображения, можно использовать для дистилляции модели ученика с неразмеченной входной областью, такой как оптический поток, текст или аудио. В этом случае функции, извлеченные из изображений модели учителя, используются для supervised обучения модели ученика.

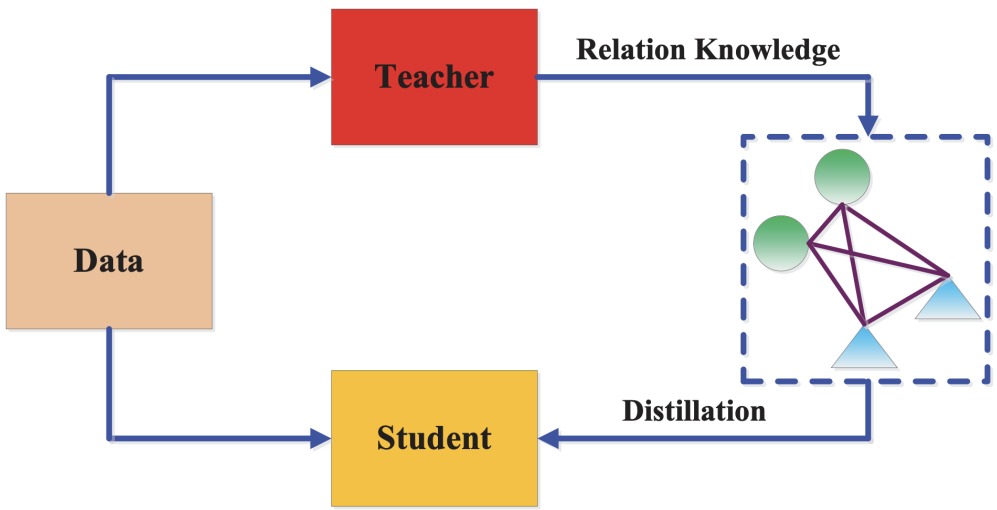
Cross-modal distillation (схема)



Source: https://arxiv.org/pdf/2006.05525.pdf

Graph-based distillation

- **Graph-based distillation**: фиксирует взаимосвязи внутри данных, используя графы, а не отдельные экземпляры знаний от учителя к ученику. Графы используются двумя способами:
  - как средство передачи знаний
  - для контроля передачи знаний учителя.



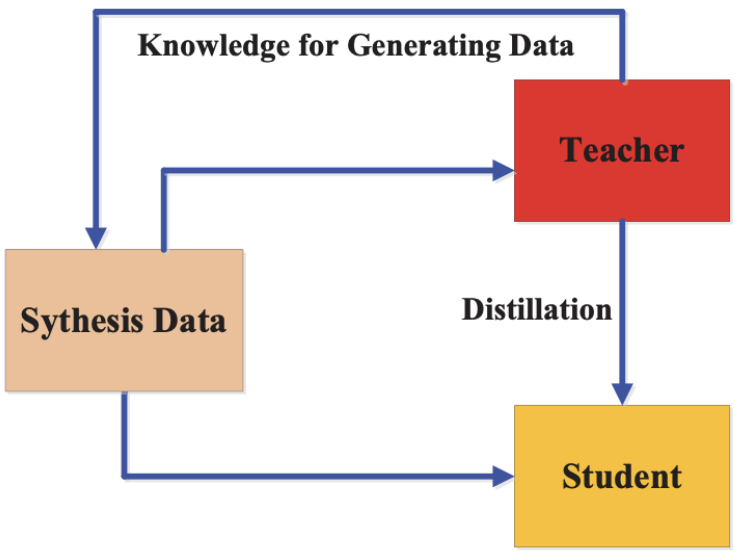
Source: https://arxiv.org/pdf/2006.05525.pdf

Attention-based distillation

- **Attention-based distillation** основана на передаче знаний из встроенных функций с использованием карт внимания.
- Главное в attention-based distillation - это определить карты внимания для согдания эмбедингов признаков в слоях нейронной сети. Тогда знания об эмбедингах передаются, используя функции карт внимания.

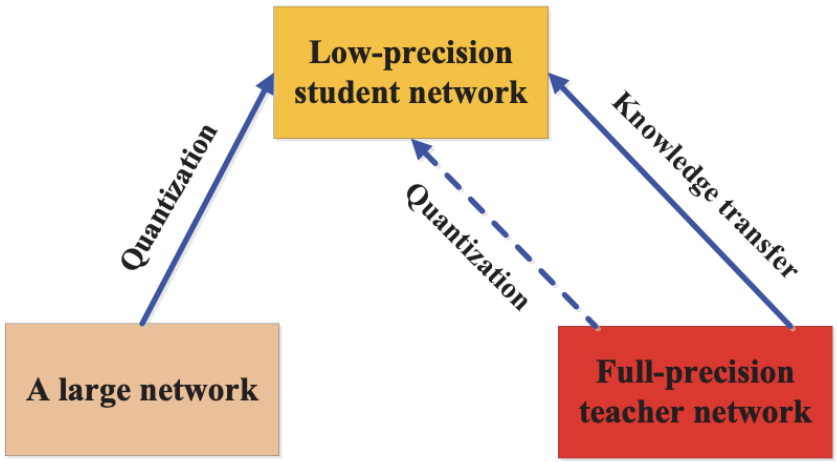
Data-free distillation

- **Data-free distillation**- на основе синтетических данных при отсутствии обучающего набора данных для студента. Синтетические данные обычно генерируются из представлений признаков предварительно обученной модели учителя. В других приложениях GAN также используются для создания синтетических обучающих данных.



Quantized distillation

- **Quantized distillation** - для передачи знаний из высокоточной модели учителя (например, 32-битной с плавающей запятой) в низкоточную студенческую сеть (например, 8-битную).



Source: https://arxiv.org/pdf/2006.05525.pdf

Lifelong и NAS distillation

- **Lifelong distillation** - основана на механизмах непрерывного обучения и метаобучения, при которых ранее полученные знания накапливаются и передаются для обучения в будущем. Обеспечивает эффективный способ сохранения и передачи полученных знаний без катастрофического забывания.
- **Neural architecture search-based distillation (NAS)** - AutoML подход, используется для определения подходящих архитектур моделей учеников, которые оптимизируют обучение на основе моделей учителей.

Dataset Distillation

Dataset Distillation

Ruonan Yu, Songhua Liu, Xinchao Wang. Dataset Distillation: A Comprehensive Review. arXiv:2301.07014

Tongzhou Wang et al. Dataset Distillation. arXiv:1811.10959

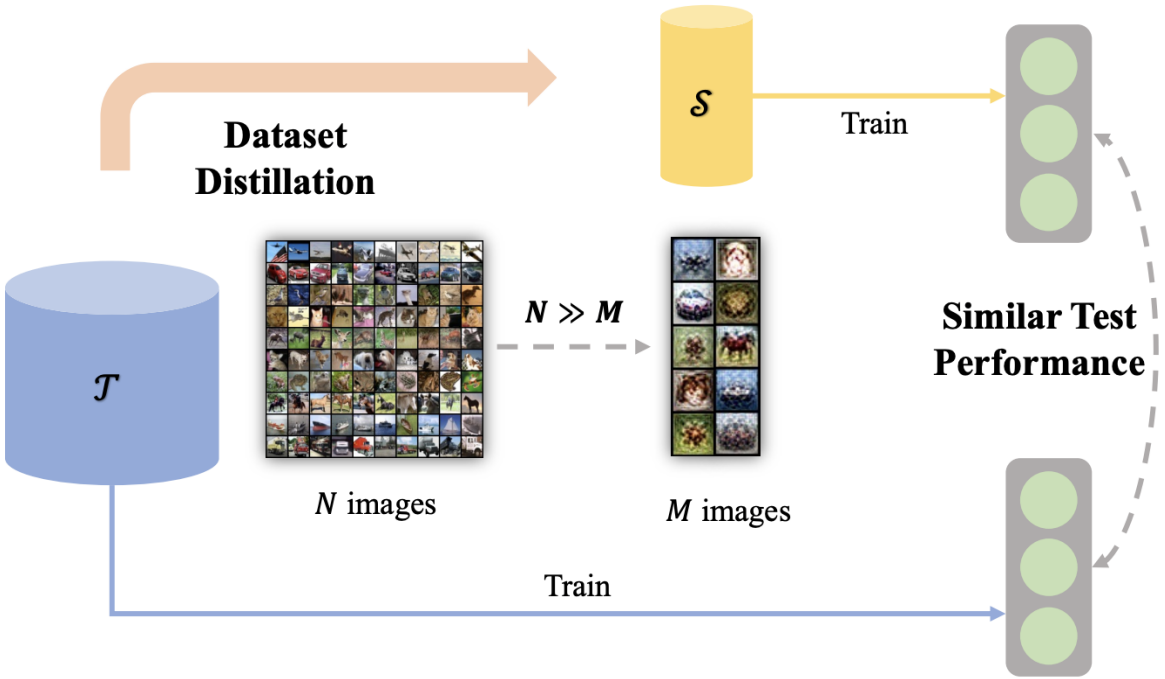
ttps://alexanderdyakonov.wordpress.com/2020/10/21/data-distillation/

Bo Zhao, Hakan Bilen. Dataset condensation with distribution matching. arXiv:2110.04181.

Dataset distillation - цель

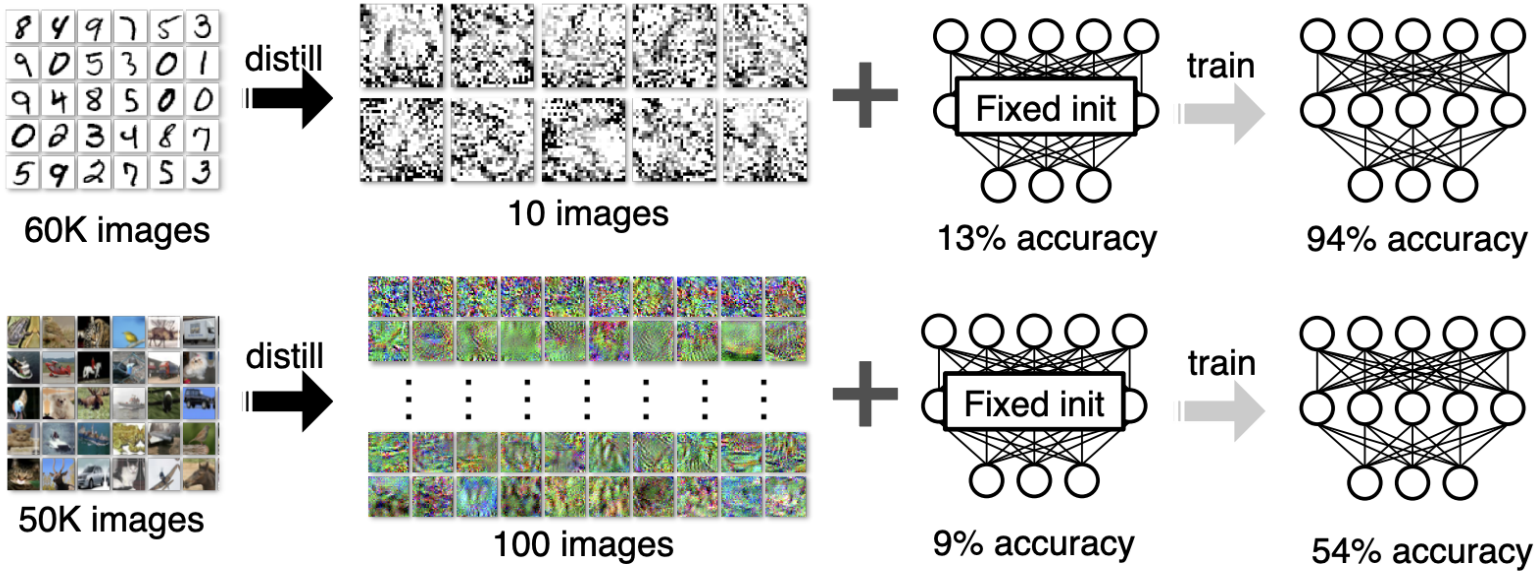
- **Дистилляция набора данных** направлена на создание небольшого информативного датасета из большого датасета, чтобы модели, обученные на этих датасетах, имели такую же тестовую эффективность, что и модели, обученные на исходном датасете.
- Аналогично **dataset condensation** (уплотнение данных), но DC чаще основан на генерации синтетических “малых” данных, а DD на подвыборке в новом представлении признаков

Dataset distillation - цель



https://arxiv.org/pdf/2301.07014.pdf

Dataset distillation - еще

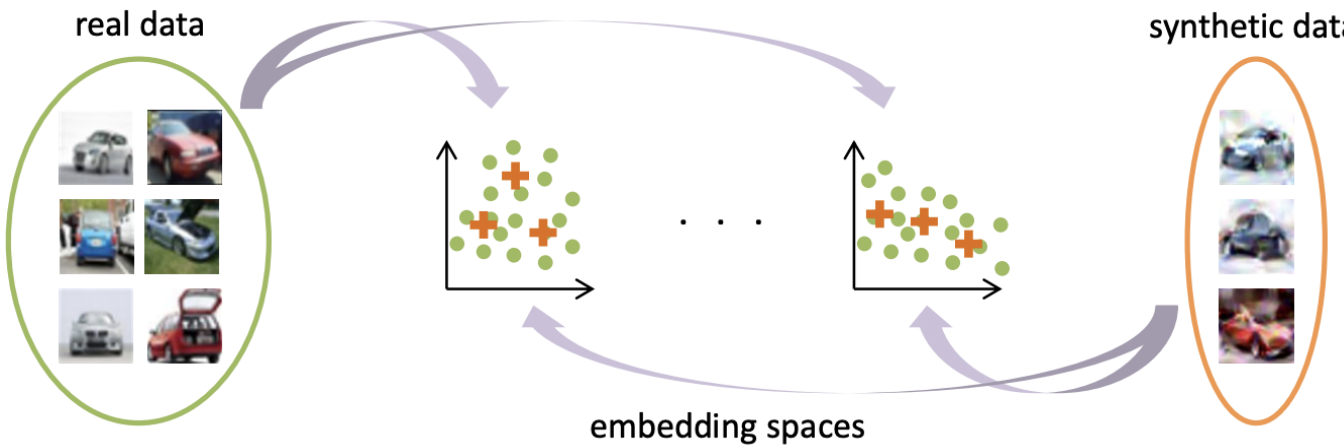


Dataset distillation on MNIST and CIFAR10

Tongzhou Wang et al. Dataset Distillation. arXiv:1811.10959

Интересный метод (1)

- Bo Zhao, Hakan Bilen. Dataset condensation with distribution matching. arXiv:2110.04181.
- Случайным образом отбираем реальные и синтетические данные, а затем встраиваем (embed) их при помощи случайно выбранных глубоких нейронных сетей.



Интересный метод (2)

- Учим синтетические данные, сводя к минимуму несоответствие распределений (the distribution discrepancy) между реальными и синтетическими данными в пространствах эмбединга.
- Пусть $\mathcal{T} = \{\mathbf{x}_i, y_i\}_{i=1}^{T}$ - большой датасет, $\mathcal{S} = \{\mathbf{s}_i, y_i\}_{i=1}^{S}$ - малый синтетич. датасет.
- Цель $\mathbb{E}_{\mathbf{x}\sim P_D}\left[l\left(\phi_{\theta^{\mathcal{T}}}(\mathbf{x}), y\right)\right] \simeq \mathbb{E}_{\mathbf{x}\sim P_D}\left[l\left(\phi_{\theta^{\mathcal{S}}}(\mathbf{x}), y\right)\right]$, где P_D - реальное распределение данных, l - функция потерь (кросс-энтропия), ϕ - глубокая нейронная сеть с параметрами θ, $\phi_{\theta^{\mathcal{T}}}$ и $\phi_{\theta^{\mathcal{S}}}$ - сетки обученные на $\mathcal{T}$ и $\mathcal{S}$

Maximum mean discrepancy

- Maximum mean discrepancy (MMD):

$$\mathbb{E}_{\vartheta \sim P_\vartheta} \| \frac{1}{|\mathcal{T}|} \sum_{i=1}^{|\mathcal{T}|} \psi_{\vartheta}(\boldsymbol{x}_i) - \frac{1}{|\mathcal{S}|} \sum_{j=1}^{|\mathcal{S}|} \psi_{\vartheta}(\boldsymbol{s}_j) \|^2$$

- где P_ϑ - распределение параметров сети, ψ_ϑ - функция с параметрами ϑ переводящая $\mathbf{x}$ в эмбединг меньшей размерности.
- Применяем дифференцируемую сиамскую аугментацию $A(\cdot, \omega)$ к реальным и синтетическим данным, где $\omega \sim \Omega$ - параметр аугментации, такой как угол поворота.

Задача оптимизации

- Окончательно, получаем задачу

$$\min_{\mathcal{S}} \mathbb{E}_{\vartheta \sim P_\vartheta, \omega \sim \Omega} \| \frac{1}{|\mathcal{T}|} \sum_{i=1}^{|\mathcal{T}|} \psi_\vartheta(\mathcal{A}(\boldsymbol{x}_i, \omega)) - \frac{1}{|\mathcal{S}|} \sum_{j=1}^{|\mathcal{S}|} \psi_\vartheta(\mathcal{A}(\boldsymbol{s}_j, \omega)) \|^2.$$

- Обучаем синтетические данные $\mathcal{S}$, минимизируя разность между двумя распределениями в различных пространствах вложения путем семплирования ϑ.

Batch normalization

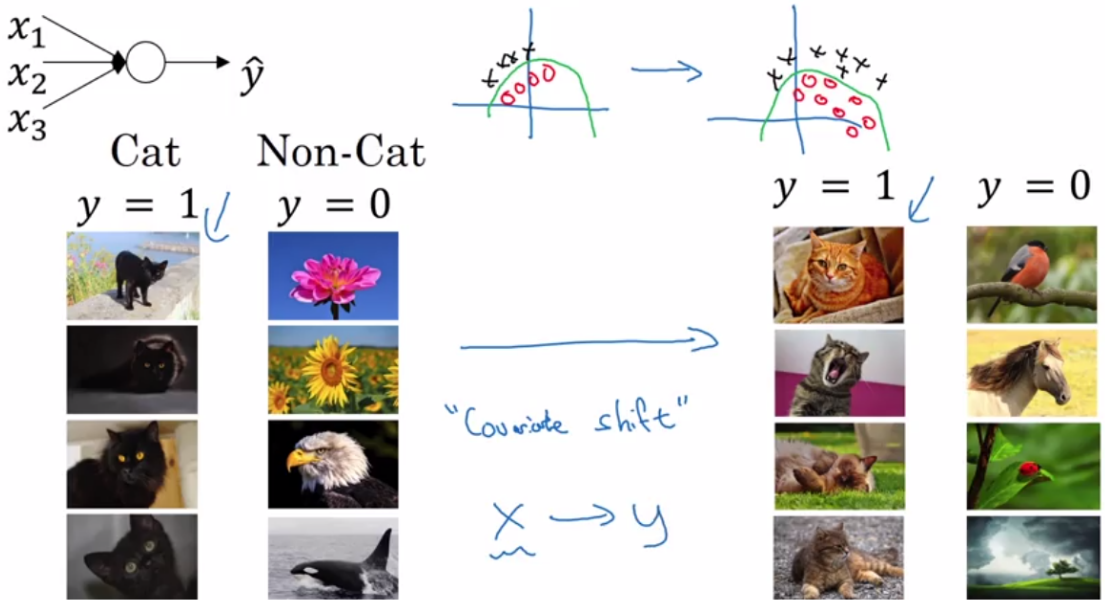
Batch normalization

Batch normalization (пакетная нормализация, BN) - зачем

- Есть функции активации от 0 до 1, а есть от 1 до 1000
- Если нормализуем входной слой, почему не сделать это для всех или части слоев
  - Это добавляет некторый шум к активациям аналогично dropout (регуляризация)
  - Уменьшает смещение
  - Делает слои сети более независимыми от других слоев
  - Высокая скорость обучения, т.к. нет очень больших или малых активации

BN - уменьшает смещение

Сеть по классификации кошек: обучаем только на черных кошках. Если применить сеть к цветными кошками, то будут ошибки. Обучающий и тестовый датасеты немного различаются. Batch normalization уменьшает смещение



Deeplearning.ai: Why Does Batch Norm Work? (C2W3L06)

BN - как

BN добавляет два обучаемых параметра к каждому слою: γ и β, что позволяет SGD выполнять денормализацию, изменяя только эти два веса для каждой активации, вместо потери стабильности сети путем изменения всех весов

Input: Values of x over a mini-batch: $\mathcal{B} = \{x_{1...m}\}$;
Parameters to be learned: γ, β
Output: $\{y_i = \mathrm{BN}_{\gamma,\beta}(x_i)\}$

$$\mu_{\mathcal{B}} \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i \qquad \text{// mini-batch mean}$$

$$\sigma^2_{\mathcal{B}} \leftarrow \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_{\mathcal{B}})^2 \qquad \text{// mini-batch variance}$$

$$\widehat{x}_i \leftarrow \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma^2_{\mathcal{B}} + \epsilon}} \qquad \text{// normalize}$$

$$y_i \leftarrow \gamma\widehat{x}_i + \beta \equiv \mathrm{BN}_{\gamma,\beta}(x_i) \qquad \text{// scale and shift}$$

S. Ioffe, C. Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift // arXiv:1502.03167v3

BN - зачем gamma и beta

- Если использовать BN в предобученной сети, то это изменит обученные веса (плохо)
- Поэтому нужно определить γ и β, чтобы отменить изменение выходных данных

Вопросы

?