Дистилляция знаний в глубоких сетях с применением методов выравнивания структур моделей

Выпускная квалификационная работа бакалавра

Михаил Сергеевич Олейник
Научный руководитель: к.ф.-м.н. О. Ю. Бахтеев

Кафедра интеллектуальных систем ФПМИ МФТИ
Специализация: Интеллектуальный анализ данных
Направление: 01.03.02 Прикладная математика и информатика

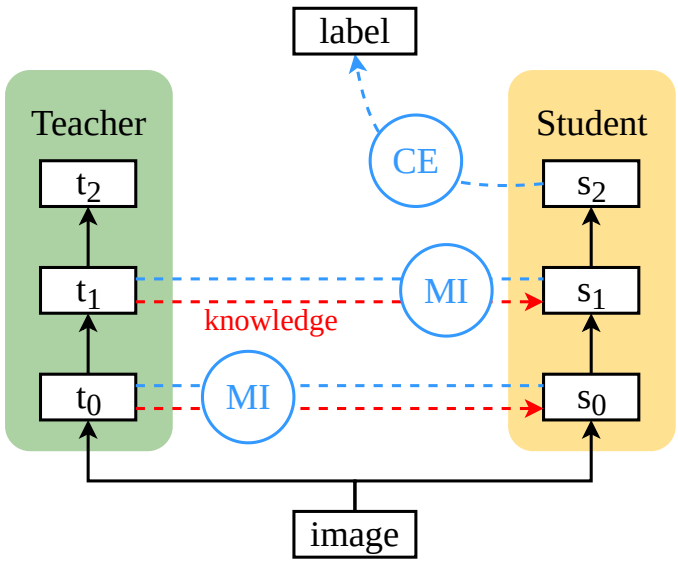
2024

Актуальность исследования

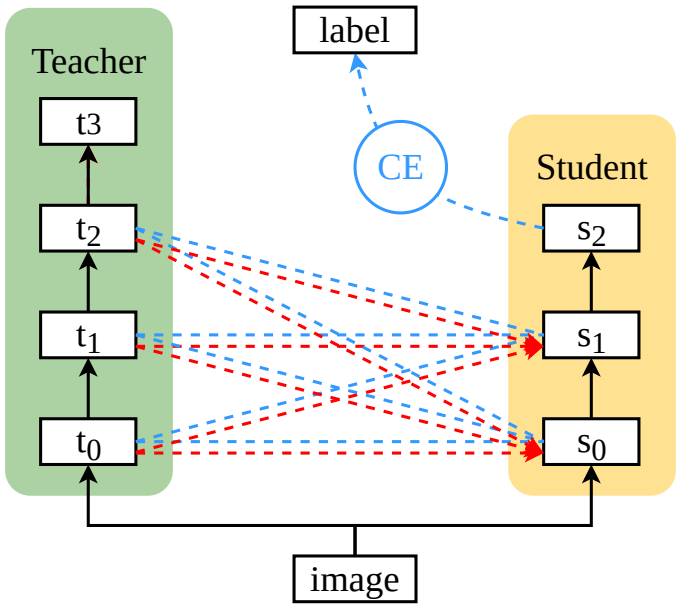
Проблема: если модели ученика и учителя имеют сильно отличающиеся архитектуры, то сложно провести дистилляцию знаний. Есть методы, с помощью которых это возможно сделать, но они дают малый прирост качества.

Цель: предложить метод дистилляции, который будет работать для разных архитектур, с разным количеством слоёв, и показывать лучшее качество.

Идея метода



Метод Sungsoo Ahn [1]



Наш метод

Слои объединяются не попарно, а каждый с каждым.

[1] Sungsoo Ahn et al. "Variational information distillation for knowledge transfer"

Постановка задачи классификации

Дана выборка для задачи классификации на K классов:

$$\mathfrak{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{m},\ \mathbf{x}_i \in \mathbb{R}^n,\ y_i \in \mathbb{Y} = \{1, \ldots, K\},$$

где $\mathbf{x}_i$ — данные, описывающие объект, взятые из распределения $p(\mathbf{x})$.
Обозначим:

- T — количество слоев в модели учителе,
- S — количество слоев в модели ученике,
- $\mathbf{t}_i$ — активации в i-м слое учителя,
- $\mathbf{s}_i$ — активации в i-м слое ученика.

Функция потерь ученика с дистилляцией

Функцию потерь ученика представим как:

$$\mathcal{L} = \beta \mathcal{L}_{\text{CE}} - (1 - \beta) \sum_{i=1}^{T} \sum_{j=1}^{S} \lambda_{i,j} I(\mathbf{t}_i, \mathbf{s}_j),$$

где:

- $\mathcal{L}_{\text{CE}}$ — функция потерь для решения задачи классификации (кросс-энтропия),
- $I(\mathbf{t}_i, \mathbf{s}_j)$ — взаимная информация,
- $\beta \in (0; 1)$ и $\lambda_{i,j} \in [0; 1]$ — необучаемые параметры (гиперпараметры).

Максимизация взаимной информации

Метод вариации нижней границы:

$$I(\mathbf{t};\mathbf{s}) = H(\mathbf{t}) - H(\mathbf{t}|\mathbf{s}) = H(\mathbf{t}) + \mathbb{E}_{\mathbf{t},\mathbf{s}}[\log p(\mathbf{t}|\mathbf{s})]$$
$$= H(\mathbf{t}) + \mathbb{E}_{\mathbf{t},\mathbf{s}}[\log q(\mathbf{t}|\mathbf{s})] + \mathbb{E}_{\mathbf{s}}[D_{\mathrm{KL}}(p(\mathbf{t}|\mathbf{s})||q(\mathbf{t}|\mathbf{s}))]$$
$$\geq H(\mathbf{t}) + \mathbb{E}_{\mathbf{t},\mathbf{s}}[\log q(\mathbf{t}|\mathbf{s})]. \quad (1)$$

Вид вариационного распределения, если слой t — свёрточный:

$$-\log q(t|s) = -\sum_{c=1}^{C}\sum_{h=1}^{H}\sum_{w=1}^{W}\log q(t_{c,h,w}|s) =$$
$$= \sum_{c=1}^{C}\sum_{h=1}^{H}\sum_{w=1}^{W}\log \sigma_c + \frac{(t_{c,h,w} - \mu_{c,h,w}(s))^2}{2{\sigma_c}^2} + constant. \quad (2)$$

Постановка двухуровневой оптимизации

$$\mathfrak{D} = \mathfrak{D}_{\text{train}} \bigsqcup \mathfrak{D}_{\text{val}}.$$

Определим вектор $\boldsymbol{\lambda}$ из всех гиперпараметров задачи:

$$\boldsymbol{\lambda} = [\lambda_{0,0}, \ldots, \lambda_{i,j}, \ldots, \beta].$$

Все обучаемые параметры — $\mathbf{w}$.
И это определяет задачу двухуровневой оптимизации:

$$\min_{\boldsymbol{\lambda}} \quad \mathcal{L}_{\text{val}}(\hat{\mathbf{w}}(\boldsymbol{\lambda}), \boldsymbol{\lambda}),$$

$$\text{s.t.} \quad \hat{\mathbf{w}}(\boldsymbol{\lambda}) = \arg\min_{\mathbf{w}} \mathcal{L}_{\text{train}}(\mathbf{w}, \boldsymbol{\lambda}).$$

Данные вычислительного эксперимента

Tiny	VeryTiny
Свёрточный 2D слой (3×3, 4 фильтра)	Свёрточный 2D слой (3×3, 8 фильтров)
Свёрточный 2D слой (3×3, 8 фильтров)	Свёрточный 2D слой (3×3, 16 фильтров)
Свёрточный 2D слой (3×3, 16 фильтров)	Свёрточный 2D слой (3×3, 32 фильтра)
Полносвязный слой (64 нейрона)	Полносвязный слой (64 нейрона)
Полносвязный слой (10 нейронов)	Полносвязный слой (10 нейронов)

Архитектура моделей Tiny и VeryTiny, слои заданы поочередно, сверху вниз

Датасеты:

- CIFAR10
- FashionMNIST

Модели:

- Учитель: Tiny
- Ученик: VeryTiny

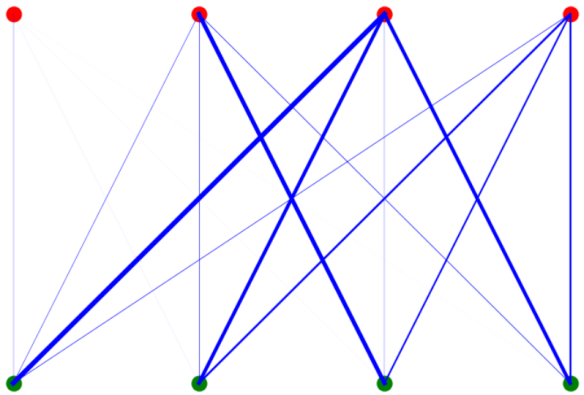
Результаты эксперимента

Метрика качества: точность.

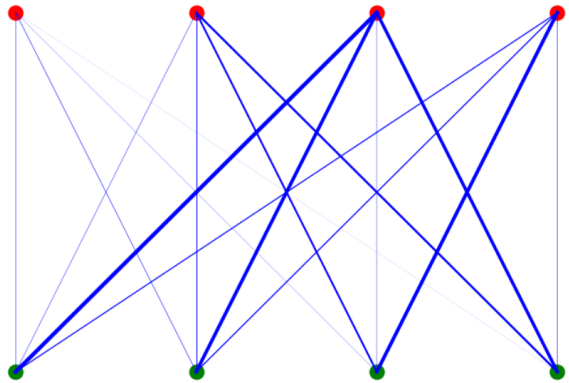
Метод	CIFAR-10	Fashion-MNIST
Без дистилляции	0.541	0.839
Дистилляция Хинтона	0.563	0.849
Дистилляцией Sungsoo Ahn	0.591	0.852
Наш метод, все связи	0.590	0.853
Наш метод, случайные гиперпараметры	0.595	**0.859**
Наш метод, вероятностная оптимизация	**0.608**	0.857

Значение качества экспериментов на датасетах CIFAR-10 и Fashion-MNIST

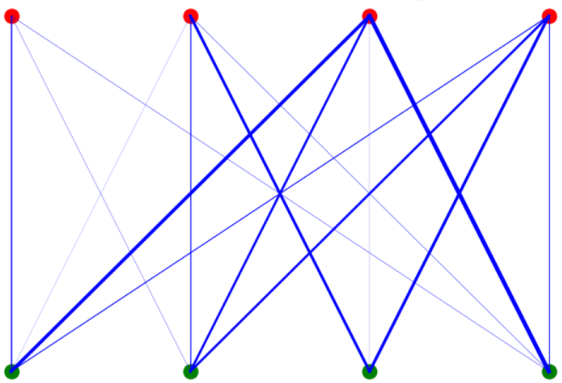
Дальнейшие исследования



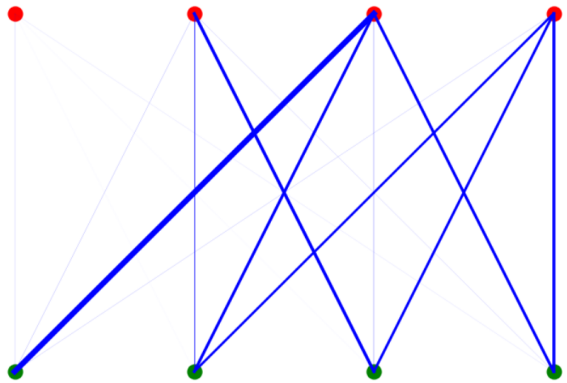
a)



b)



c)



d)

Рис.: Иллюстрация коэффициентов у четырех лучших моделей по качеству. Зелёные точки — слои ученика, красные — слои учителя. Чем толще линия, тем больше коэффициент у соответствующей связи.

Выносится на защиту

- Предложен метод дистилляции, максимизирующий взаимную информацию между всеми слоями ученика и учителя.
- Продемонстрировано, что предложенный метод достигает лучшего качества, чем базовые методы дистилляции знаний.
- Были проведены эксперименты, показавшие эффективность предложенного метода.