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

Цели 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 и с разным количеством слоёв, предложить для него алгоритм подбора гиперпараметров.

Постановка задачи

Дана выборка для задачи классификации на K классов:

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

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

Постановка задачи

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,
- β и $\lambda_{i,j}$ — гиперпараметры.

Схема метода

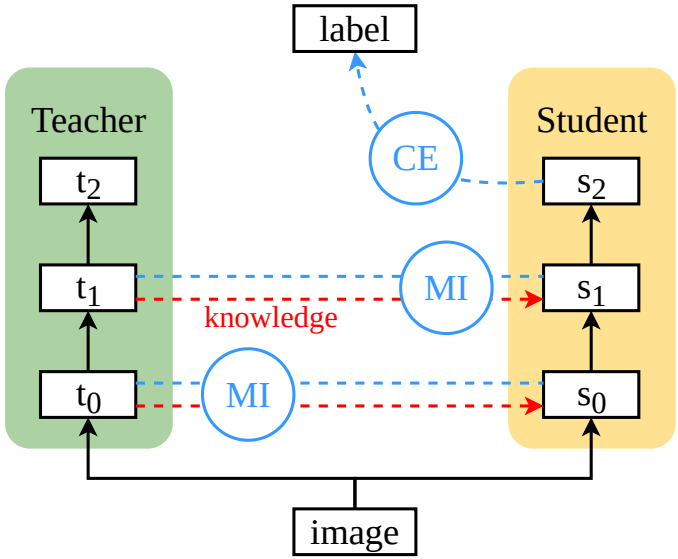


Рис.: Базовый метод [a]

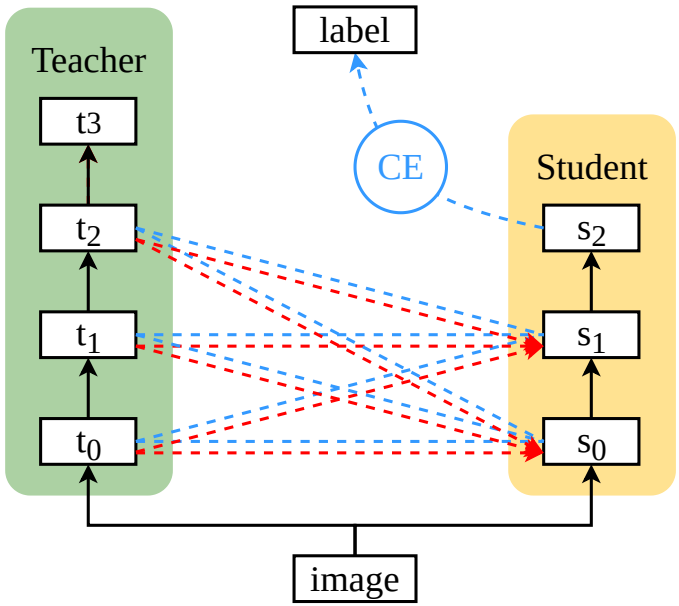


Рис.: Предлагаемый метод

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

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

Взаимная информация

Метод вариации нижней границы:

$$I(t,s) = H(t) - H(t|s) \geq H(t) + E_{t,s}[\log q(t|s)]. \qquad (2)$$

Вариационное распределение:

$$-\log q(t|s) = -\sum_{c=1}^{C}\sum_{h=1}^{H}\sum_{w=1}^{W} \log q(t_{c,h,w}|s) =$$

$$= \sum_{c=1}^{C}\sum_{h=1}^{H}\sum_{w=1}^{W} \log \sigma_c + \frac{(t_{c,h,w} - \mu_{c,h,w}(s))^2}{2\sigma_c^2} + constant. \qquad (3)$$

Обучаемые параметры:

$$\sigma_c^2 = \log\left(1 + e^{\alpha_c}\right) + \epsilon$$

$$\mu_{c,h,w}(s) = \mu(s)_{c,h,w}$$

Двухуровневая оптимизация

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

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:

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

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

Вычислительный эксперимент

Датасеты:

- CIFAR10
- FashionMNIST

Модели:

- ConvVeryTiny
- ConvTiny

Метрики:

- accuracy

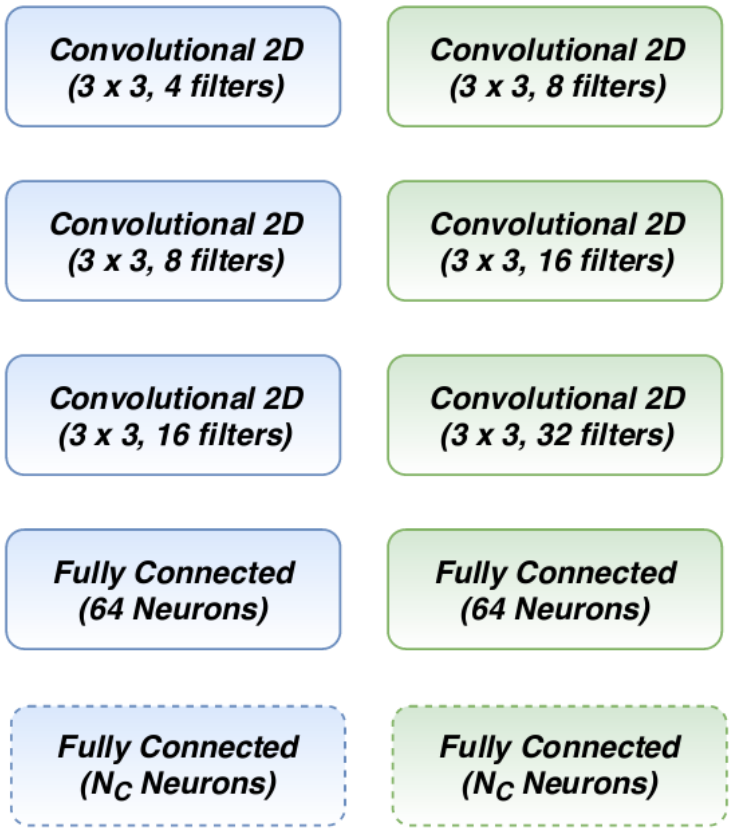


Рис.: Схема моделей ConvVeryTiny и ConvTiny

Эксперимент 1

Учитель: ConvTiny.
Ученик: ConvVeryTiny.

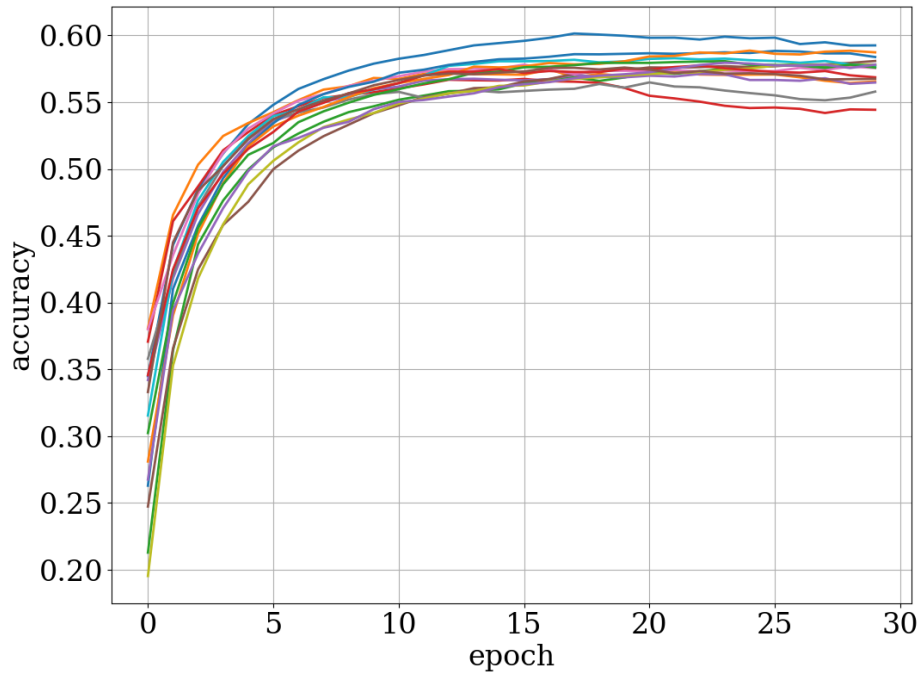
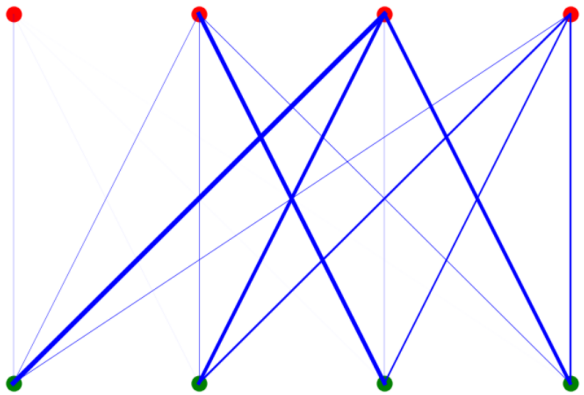


Рис.: Точность от эпохи при дистилляции каждый с каждым

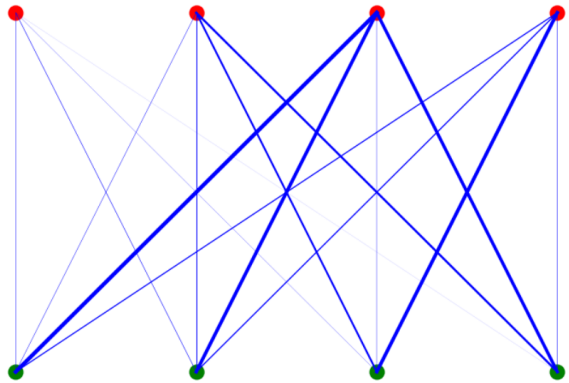
Дистилляция	—	Хинтона	попарная	каждый с каждым
Учитель	0.58	—	—	—
Ученик	0.54	0.56	0.58-0.59	0.58-0.59

Таблица: Сравнение качества моделей на тестовой выборке

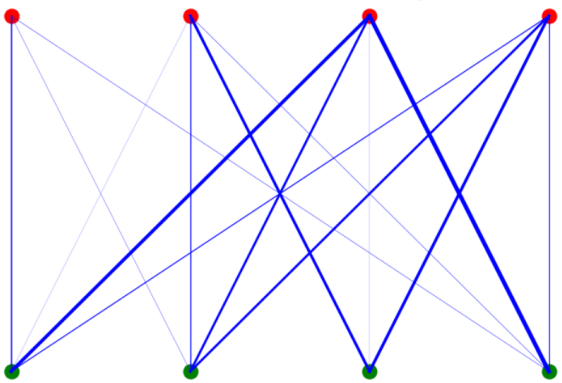
Эксперимент 1



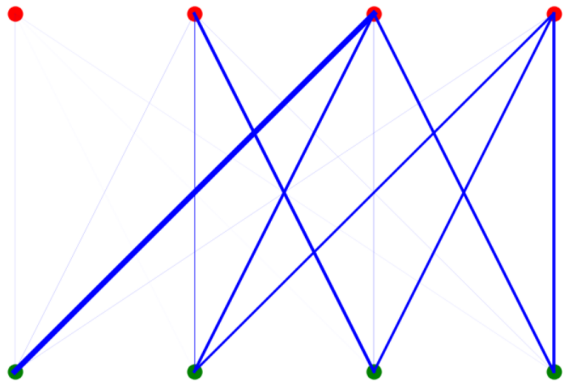
a)



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.

Заключение

Был предложен метод дистилляции знаний, который можно применить к моделям с разным количеством слоев и/или разными архитектурами, который выдает большее качество, чем дистилляция Хинтона. Однако, к недостаткам данного подхода можно отнести большие требования по памяти и времени, если модель учителя и/или ученика имеет большое количество слоёв. В дальнейших планах стоит более тщательное изучение влияния связей между слоями на итоговый результат, что потенциально и может свести недостатки подхода к минимуму.

Литература

- Sungsoo Ahn et al. "Variational information distillation for knowledge transfer"
- Nikolaos Passalis, Maria Tzelepi, and Anastasios Tefas. "Heterogeneous knowledge distillation using information flow modeling"
- M Gorpinich, O Yu Bakhteev, and VV Strijov. "Gradient methods for optimizing metaparameters in the knowledge distillation problem"
- Liu Hanxiao et al. "DARTS: Differentiable Architecture Search"