

Дистилляция знаний в глубоких сетях с применением методов выравнивания структур моделей

Михаил Сергеевич Олейник

Московский физико-технический институт

Научный руководитель: О. Ю. Бахтеев

2023

Проблема: сложно проводить дистилляцию, если ученик и учитель имеют сильно отличающиеся архитектуры. А если не сложно, то это дистилляция по Хинтону.

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

- ▶ 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"

Постановка задачи

Дана выборка для задачи классификации на K классов:

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

$$\mathcal{D} = \mathcal{D}_{\text{train}} \sqcup \mathcal{D}_{\text{test}}.$$

Обозначим:

- ▶ \mathbf{f} — модель учителя, обученная на $\mathcal{D}_{\text{train}}$
- ▶ \mathbf{g} — модель ученика, которую предстоит обучить
- ▶ T — количество слоев в учителе
- ▶ S — количество слоев в ученике
- ▶ t_i — активации в i -м слое учителя
- ▶ s_i — активации в i -м слое ученика

Постановка задачи

Функцию потерь ученика представим как:

$$\mathcal{L} = \beta \mathcal{L}_{\text{task}} - (1 - \beta) \sum_{i,j=1}^{T,S} \lambda_{i,j} I(t_i, s_j),$$

Где:

- ▶ $\mathcal{L}_{\text{task}}$ — функция потерь для решения задачи классификации (кросс-энтропия),
- ▶ $I(t_i, s_j)$ — взаимная информация,
- ▶ β и $\lambda_{i,j}$ — гиперпараметры.

Взаимная информация

Метод вариации нижней границы:

$$I(t, s) = H(t) - H(t|s) \geq H(t) + E_{t,s}[\log q(t|s)]. \quad (1)$$

Вариационное распределение:

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

Обучаемые параметры:

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

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

Схема метода

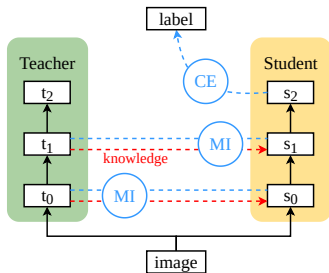


Рис.: Базовый метод ^a

^aSungsoo Ahn et al.
"Variational information
distillation for knowledge
transfer"

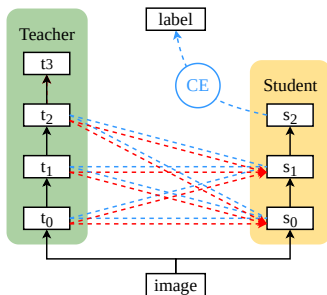


Рис.: Предлагаемый метод

$$\mathcal{L} = \beta \mathcal{L}_{\text{CE}} - (1 - \beta) \sum_{i,j=1}^{T,S} \lambda_{i,j} I(t_i, s_j) \quad (3)$$

Вычислительный эксперимент

Датасеты:

- ▶ CIFAR10

Модели:

- ▶ ConvVeryTiny
- ▶ ConvTiny
- ▶ ResNet10
- ▶ ResNet18

Метрики:

- ▶ accuracy

Convolutional 2D
(3 x 3, 4 filters)

Convolutional 2D
(3 x 3, 8 filters)

Convolutional 2D
(3 x 3, 8 filters)

Convolutional 2D
(3 x 3, 16 filters)

Convolutional 2D
(3 x 3, 16 filters)

Convolutional 2D
(3 x 3, 32 filters)

Fully Connected
(64 Neurons)

Fully Connected
(64 Neurons)

Fully Connected
(N_C Neurons)

Fully Connected
(N_C Neurons)

Рис.: Схема моделей
ConvVeryTiny и ConvTiny

Эксперимент 1

Учитель: ConvTiny.
Ученик: ConvVeryTiny.

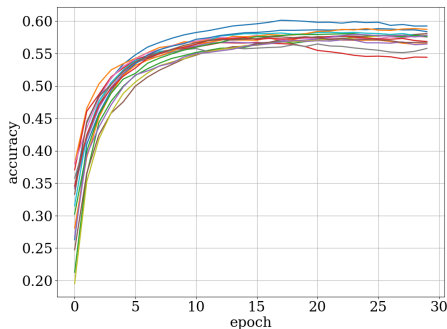
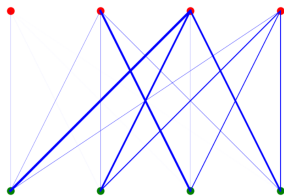


Рис.: Точность от эпохи при
дистилляции каждый с каждым

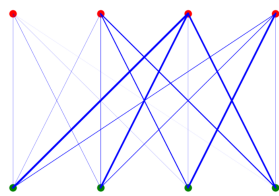
| Дистилляция | — | Хинтона | попарная | каждый с каждым |
|-------------|------|---------|-----------|-----------------|
| Учитель | 0.58 | — | — | — |
| Ученик | 0.54 | 0.56 | 0.58-0.59 | 0.58-0.59 |

Таблица: Сравнение качества моделей на тестовой выборке

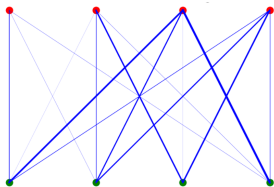
Эксперимент 1



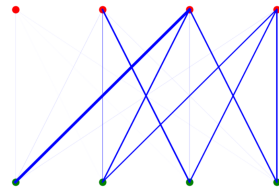
a)



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.

Эксперимент 2

Учитель: ResNet10.

Ученик: ConvTiny.

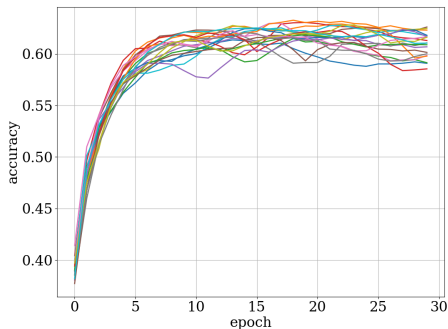
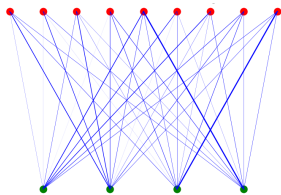


Рис.: Точность от эпохи при дистилляции каждый с каждым

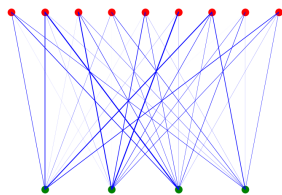
| | | | |
|-------------|------|---------|-----------------|
| Дистилляция | — | Хинтона | каждый с каждым |
| Учитель | 0.68 | — | — |
| Ученик | 0.58 | 0.60 | 0.63 |

Таблица: Качество моделей при дистилляции на тестовой выборке

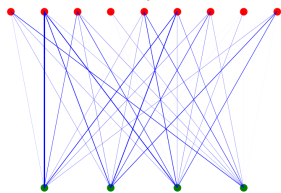
Эксперимент 2



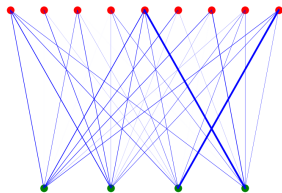
a)



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.

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.