Дистилляция знаний в глубоких сетях и выравнивание структур моделей

Михаил Сергеевич Олейник

Московский физико-технический институт

Курс: Моя первая научная статья

Эксперт: О. Ю. Бахтеев

Консультант: М. Горпинич

2023

Цели исследования

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.

Предлагаемый подход

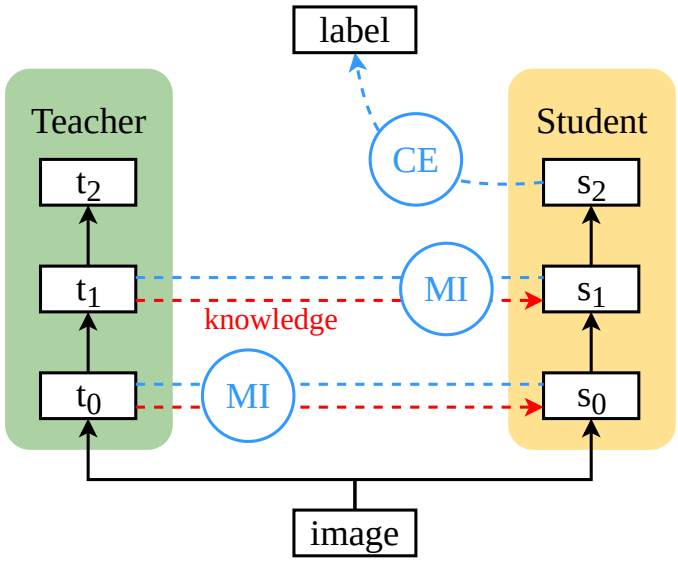


Рис.: Базовый метод

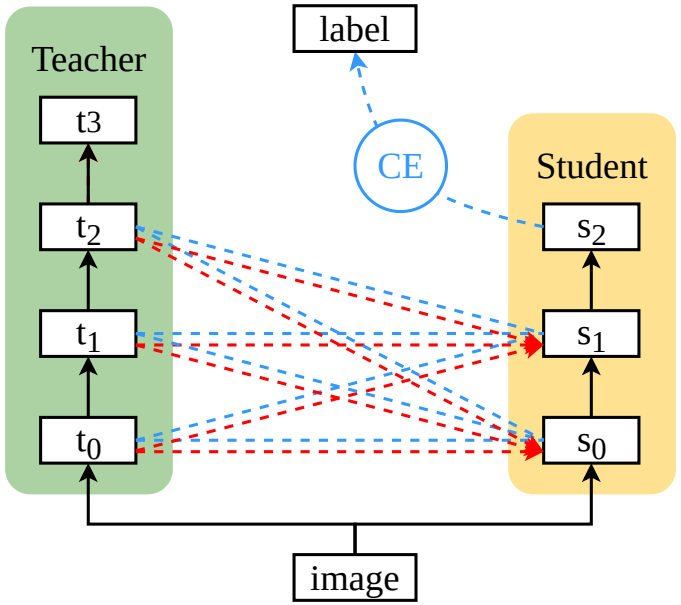


Рис.: Предлагаемый метод

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

$$\forall j \hookrightarrow \sum_{i=1}^{T} \lambda_{i,j} = 1 \quad (2)$$

T, S - количество слоёв учителя и ученика,
$I(t_i, s_j)$ - взаимная информация,
β и $\lambda_{i,j}$ — гиперпараметры.

Взаимная информация

Метод вариации нижней границы:

$$I(t,s) = H(t) - H(t|s) \geq H(t) + E_{t,s}[\log q(t|s)]. \quad (3)$$

Как будет выглядеть вариационное распределение?

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

$$= \sum_{c=1}^{C}\sum_{h=1}^{H}\sum_{w=1}^{W} \log \sigma_c + \frac{(t_{c,h,w} - \mu_{c,h,w}(s))^2}{2\sigma_c^2} + constant. \quad (4)$$

Обучаемые параметры скрываются здесь:

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

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

Эксперимент

Выборка: CIFAR10.
Модели: 3 conv, 2 linear.
Метрика: accuracy.

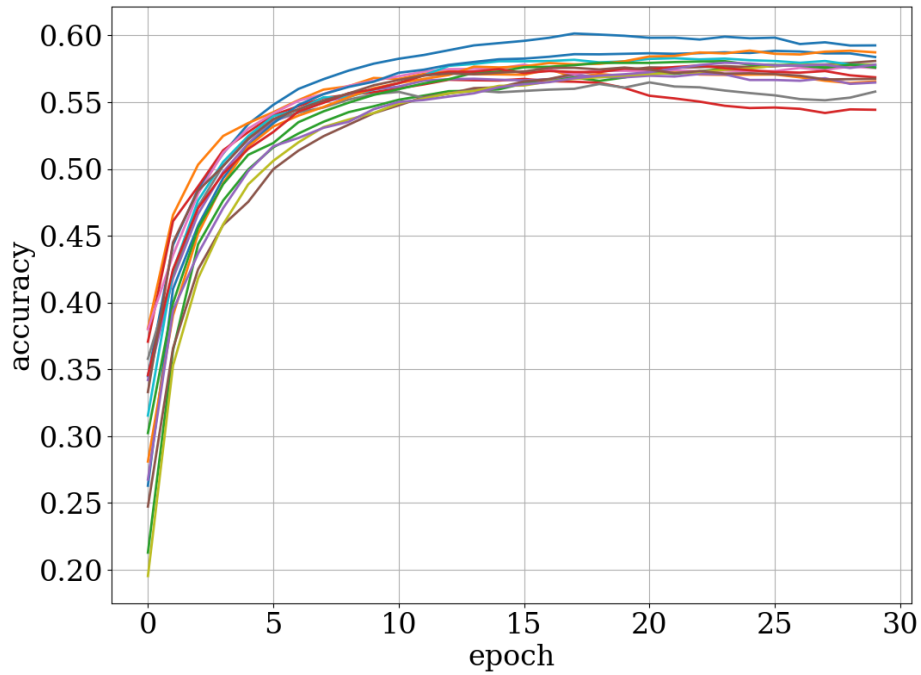
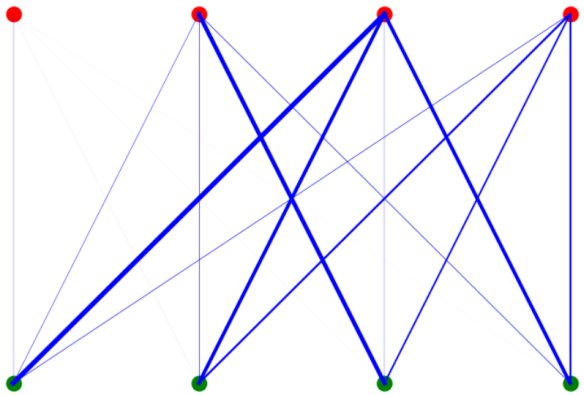


Рис.: Точность от эпохи при дистилляции каждый с каждым

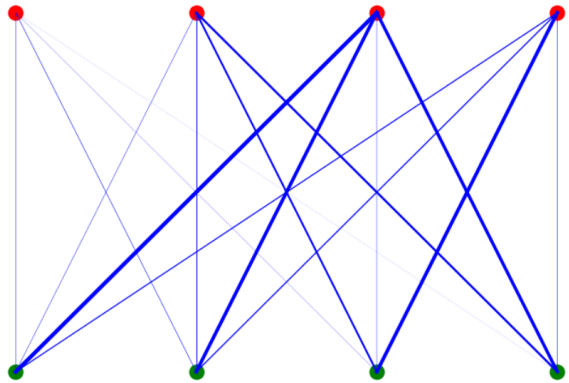
Дистилляция	—	Хинтона	попарная	каждый с каждым
Учитель	0.58	—	—	—
Ученик	0.54	0.56	0.58-0.59	0.58-0.59

Таблица: Сравнение качества моделей на тестовой выборке

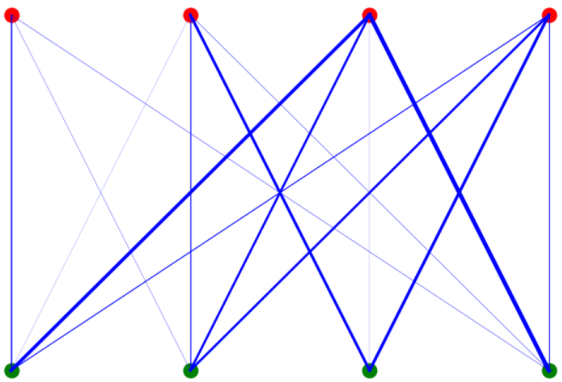
Ограничения на коэффициенты



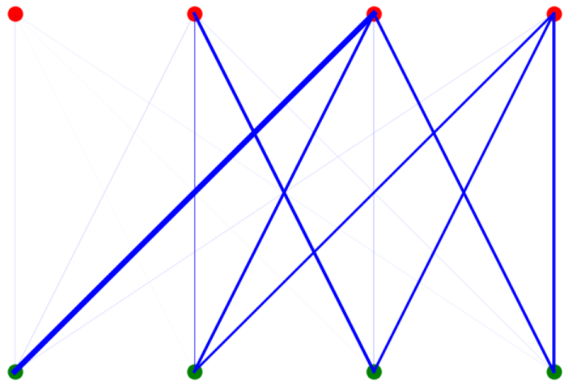
a)



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.

Промежуточные выводы

- Между учителем и учеником мало различий, не можем увидеть разницу в качестве у разных подходов. Пробуем с учителем побольше.
- Какие-то закономерности с коэффициентами присутствуют. Утверждать рано, продолжаем изучать.