Дистилляция знаний в глубоких сетях

Михаил Олейник
МФТИ
oleinik.ms@phystech.edu

М. Горпинич
МФТИ
gorpinich.m@phystech.edu

О. Ю. Бахтеев
МФТИ
bakhteev@phystech.edu

Аннотация

Дистилляция знаний позволяет повысить качество модели, называемой учеником, не увеличивая её число параметров, а используя модель большего размера, называемой учителем. Однако, в случае разных архитектур и несовпадения количества слоев у учителя и ученика, распространённые методы не применимы. Одним из подходов, который позволяет решать задачу для разных архитектур, является максимизация взаимной информации. Мы предлагаем улучшение этого подхода, которое позволит проводить дистилляцию и для моделей с разным количеством слоёв. Мы сравниваем наш метод с остальными с помощью вычислительного эксперимента. Также проводим анализ гиперпараметров и выводим ограничения на них, при которых достигается наибольшее качество.

Ключевые слова: дистилляция модели · байесовский вывод · глубокое обучение · вариационная оптимизация

1 Введение

Глубокие нейронные достигли больших успехов в задачах машинного зрения, обработки естественного языка и других. Однако, лучшие результаты достигают модели с большим количеством параметров, из-за этого их трудно встроить в системы с небольшими вычислительными мощностями, например, мобильные телефоны. Если подобрать размер модели под целевую платформу, уменьшив количество параметров, то сильно потеряем и в качестве.

Одним из подходов, которые позволяют не теряя сильно в качестве, получить модель с меньшим количеством параметров, является дистилляция знаний. Этот подход использует большую предобученную на необходимой задаче модель, называемую учителем, данные о слоях которой переносятся определенным образом в модель меньшего размера, называемую учеником. Перенос чаще всего выражается в дополнительном слагаемом в функции потерь ученика.

Так, в работе [1] предлагается переносить знания с последнего слоя модели. К недостаткам этого метода можно отнести то, что мы игнорируем информацию из остальных слоев учителя, а она может быть ценной. В работах ...

Однако, большинство подходов либо неэффективно работают, либо не могут быть применимы к случаям, когда модели имеют разное количество слоёв или разную архитектуру. Также возникают сложности в случае, когда количество параметров в слое учителя сильно больше, чем в соответствующем слое ученика, как показано в работе [2].

Больший интерес представляют подходы, которые можно применить, если учитель и ученик имеют разные архитектуры. В работе [3] моделируется информационный поток в учителе, который имитирует ученик. В работе [4] используется максимизация взаимной информации между парами соответствующих слоёв. В основе этого метода используется вариационный подход [5]. Наша работа предлагает улучшение данного метода, давая возможность проведения дистилляции при разном количестве слоев у учителя и ученика. Также мы проводим анализ гиперпараметров алгоритма, на примере работы [6].

2 Постановка задачи

Дан датасет для задачи многоклассовой классификации, с количеством классов K:

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

Датасет разбит на обучающую и тестовую выборку: $\mathfrak{D} = \mathfrak{D}_{\text{train}} \bigsqcup \mathfrak{D}_{\text{test}}$.

Дана модель учителя $\mathbf{f}$, обученная на $\mathfrak{D}_{\text{train}}$. Дана модель ученика $\mathbf{g}$, которую предстоит обучить.

Пусть T — количество слоев в модели учителя, S — количество слоев в модели ученика. Обозначим как t_i и s_j — активации в i-м слое учителя и в j-м слое ученика, соответственно.

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, $I(t_i, s_j)$ — взаимная информация, β и $\lambda_{i,j}$ — гиперпараметры.

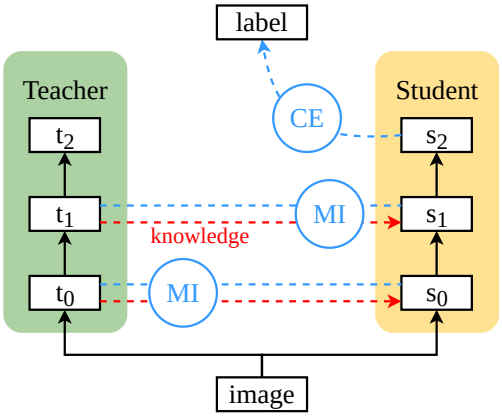


Рис. 1: Базовый метод.

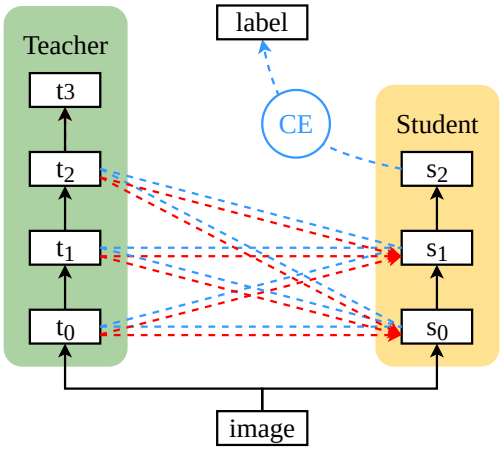


Рис. 2: Предлагаемый метод.

2.1 Взаимная информация

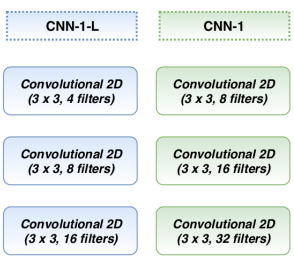
Взаимная информация

2.2 Алгоритм

Алгоритм.

3 Вычислительный эксперимент

Для проведения экспериментов использовался датасет CIFAR10, разбитый на обучающую и тестовую выборку. Использовались модели ResNet10, ResNet18, ConvTiny и ConvVeryTiny. Последние две - свёрточные нейронные сети, с тремя свёрточными и двумя полносвязными слоями, как описано на схеме 3.



3.1 Базовый эксперимент

В качестве модели учителя была выбрана модель ResNet18, в качестве учеников — ConvTiny и ConvVeryTiny. В качестве метрики была выбрана точность (accuracy).

Цель — сравнение качества учеников с и без дистилляции Хинтона, а также зависимость качества дистиллированных моделей от гиперпараметров β и T, что демонстрируется на графиках 4, 5 и 6 соответственно.

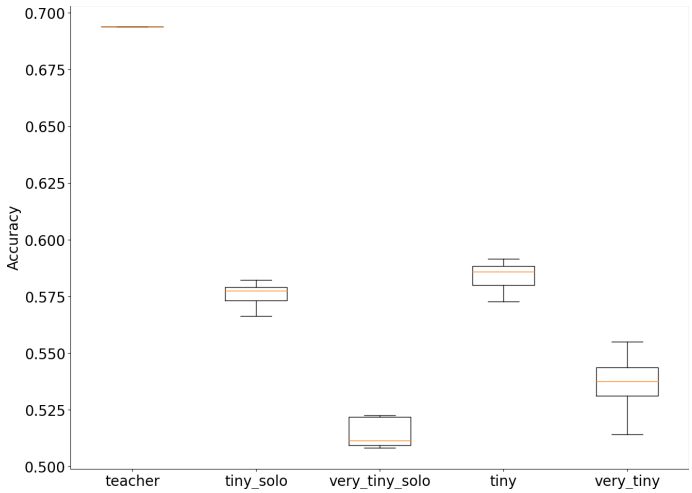


Рис. 4: Сравнение качества моделей.

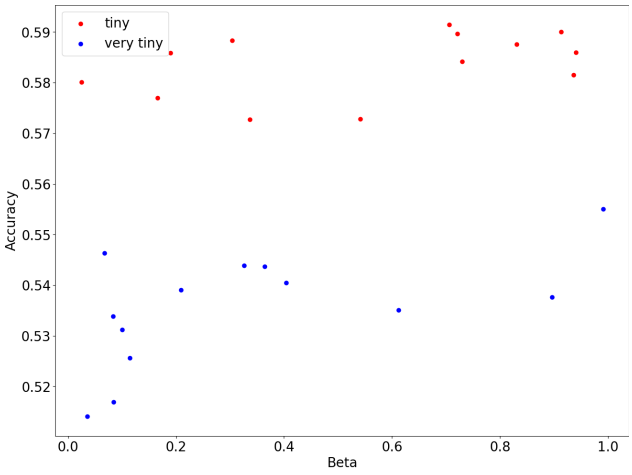


Рис. 5: Зависимость качества от параметра β.

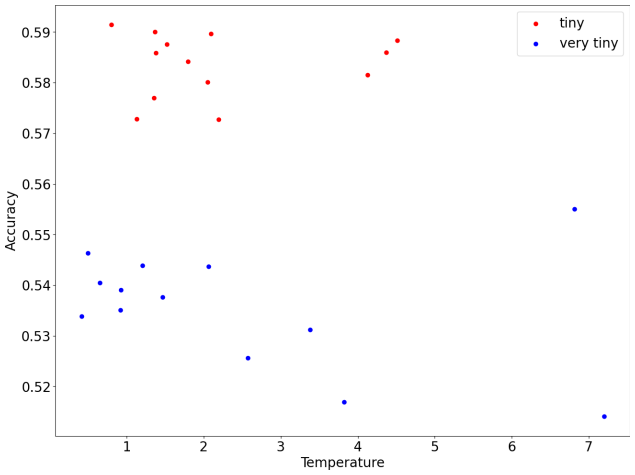


Рис. 6: Зависимость качества от параметра T.

3.2 Основной эксперимент

3.2.1 Дистилляция из ConvTiny

Был проведен эксперимент по дистилляции из ConvTiny в ConvVeryTiny. Были рассмотрены разные методы: Хинтона, попарная, и предлагаемый метод. В таблице 1 представлены значения точности для моделей без дистилляции и при дистилляции упоминаемыми методами.

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

Таблица 1: Качество моделей при дистилляции на тестовой выборке, учитель — ConvTiny, ученик — ConvVeryTiny.

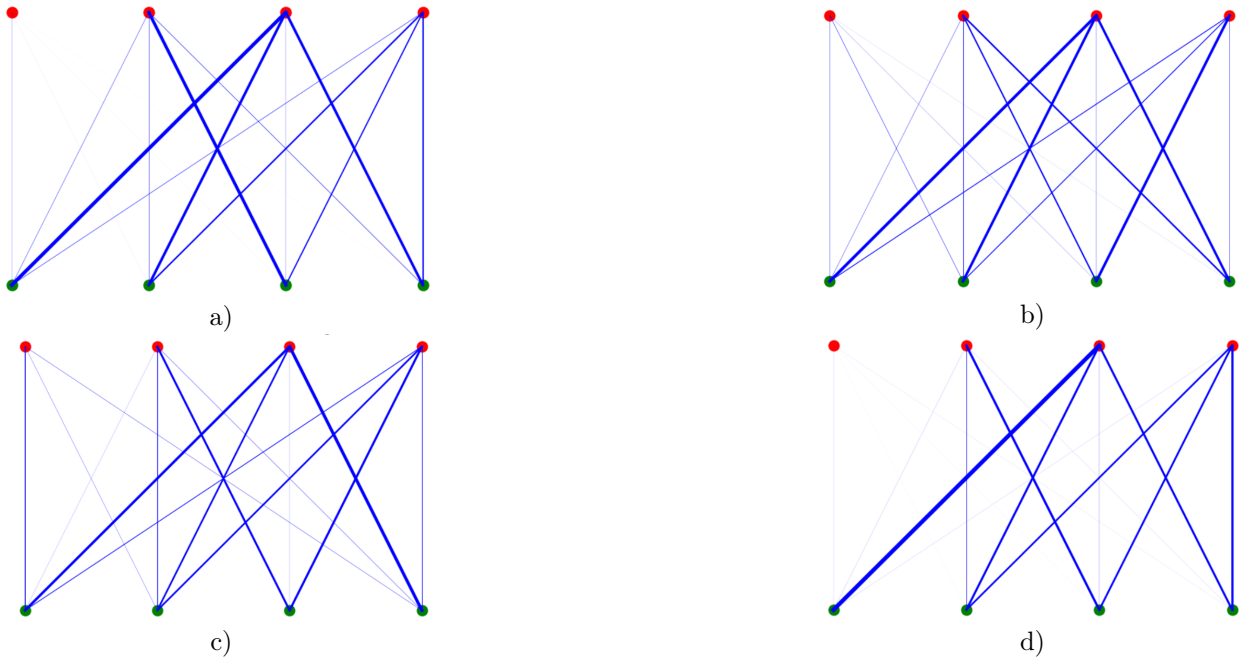


Рис. 7: Иллюстрация коэффициентов у четырех лучших моделей по качеству. Зелёные точки — слои ученика, красные — слои учителя. Чем толще линия, тем больше коэффициент у соответствующей связи.

При предлагаемом методе дистилляции "каждый с каждым"выбирались случаные гиперпараметры $\lambda_{i,j}$ и β, на графике 8 можно увидеть значение метрики на тестовой выборке для разных гиперпараметров, на рисунке 7 визуализированы значения $\lambda_{i,j}$ в лучших по качеству запусках. Наблюдается общий паттерн, что связи с третьим слоем учителя (самым большим по количеству параметров) больше, чем с остальными слоями учителя.

3.2.2 Дистилляция из ResNet10

При дистилляции из ResNet10 в ConvVeryTiny качество ухудшилось. Это связано с тем, что сеть учителя стала слишком сложной для ученика, данное явление рассматривается в работе [2].

В таблице 2 представлены значения метрик при дистилляции из ResNet10 в ConvTiny.

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

Таблица 2: Качество моделей при дистилляции на тестовой выборке, учитель — ResNet10, ученик — ConvTiny.

4 Заключение

Заключение.

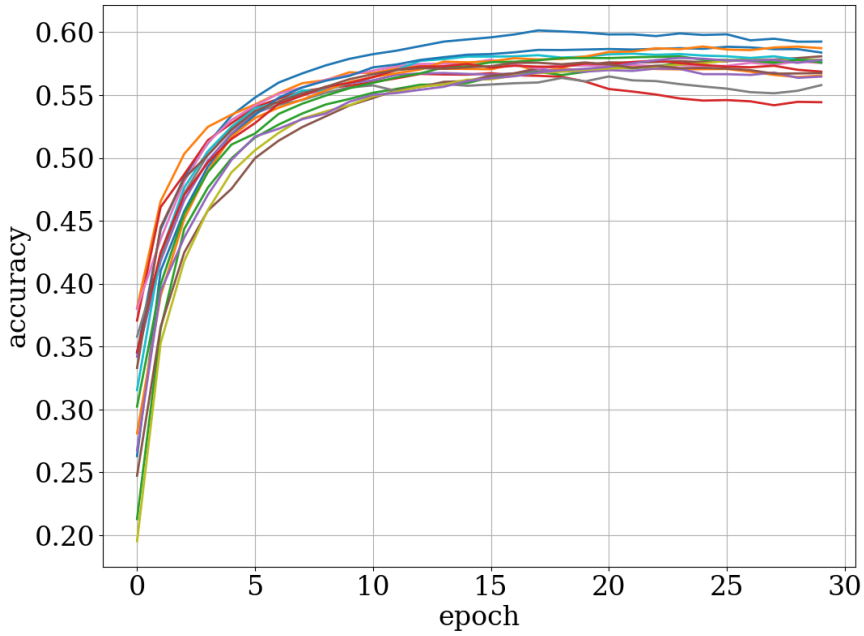


Рис. 8: Точность от эпохи при дистилляции каждый с каждым

Список литературы

[1] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531, 2015.

[2] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In Proceedings of the AAAI conference on artificial intelligence, volume 34, pages 5191–5198, 2020.

[3] Nikolaos Passalis, Maria Tzelepi, and Anastasios Tefas. Heterogeneous knowledge distillation using information flow modeling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2339–2348, 2020.

[4] Sungsoo Ahn, Shell Xu Hu, Andreas Damianou, Neil D. Lawrence, and Zhenwen Dai. Variational information distillation for knowledge transfer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2019.

[5] David Barber and Felix Agakov. The im algorithm: a variational approach to information maximization. Advances in neural information processing systems, 16(320):201, 2004.

[6] M Gorpinich, O Yu Bakhteev, and VV Strijov. Gradient methods for optimizing metaparameters in the knowledge distillation problem. Automation and Remote Control, 83(10):1544–1554, 2022.