Оптимизация процесса обучения при ограниченном объеме вычислительных ресурсов

Н. С. Мокрецов

Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина)

mokrecovnikita6374@gmail.com

Т. М. Татарникова

Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина)

tm-tatarn@yandex.ru

Аннотация. **Рассматривается проблема оптимизации моделей глубокого обучения для их использования в приложениях реального времени на устройствах с ограниченными ресурсами. Для решения этой проблемы предлагается использовать дистилляцию знаний, которая заключается в переносе знаний из точной, но громоздкой модели нейронной сети в более компактную. Описываются различные архитектуры такого подхода, их недостатки и преимущества. Приведён анализ эффективности дистилляции знаний с выявлением важных закономерностей для достижения желаемого результата при использовании такого подхода. Описывается реализация и результаты применения применительно к свёрточным нейронным сетям визуальной классификации.**

Ключевые слова: нейронные сети, глубокое обучение, оптимизация модели, дистилляция знаний

I. Введение

Применение алгоритмов глубокого обучения стало основой многих успехов в области искусственного интеллекта, включая различные приложения в области компьютерного зрения, обучения с подкреплением и обработки естественного языка. Благодаря множеству новейших методов, таких как остаточные соединения и пакетная нормализация, стало легко тренировать модели с миллиардами параметров на мощных кластерах графических процессоров (Graphics Processing Unit, GPU) или тензорных процессоров (Tensor Processing Unit, TPU).

Например, требуется менее десяти минут, чтобы обучить модель ResNet – глубокую сверточную нейронную сеть распознавания образов на популярном датасете, содержащем миллионы изображений. Требуется не более полутора часов, чтобы обучить мощную модель BERT для решения задач, основанных на обработке естественных языков.

Несмотря на то, что крупномасштабные модели глубокого обучения достигли ошеломляющих успехов, их огромная вычислительная сложность и требования к размерам хранилищ делают их использование в приложениях реального времени большой проблемой, особенно на устройствах с ограниченными ресурсами, таких как системы видеонаблюдения и автономные автомобили.

Для решения этой проблемы сообществом было предложено большое количество методов, алгоритмов и архитектур моделей, оптимизирующих её размеры. Чаще всего используются различные модификации алгоритмов удаления лишних нейронных связей из слоёв сети или повторное использование параметров для разных задач. Такие подходы в общем случае помогают достичь желаемых результатов, но обладают рядом недостатков, таких как уменьшение устойчивости модели к шуму, непредсказуемому уровню потери информации, отсутствием универсальности, либо невозможности применения таких подходов в конкретном случае.

II. Дистилляция знаний

В настоящее время, в качестве подхода для оптимизации модели нейронной сети, особое внимание уделяется идее дистилляции знаний. Суть данного подхода заключается в переносе знаний из точной, но громоздкой модели нейронной сети (учителя) в более компактную (ученика), с учётом ограничения вычислительных ресурсов или размеров чипа, на которых планируется запускать оптимизированную версию модели. На данный момент предложено большое количество различных подходов реализации дистилляции знаний. С точки зрения архитектуры модели, они делятся на два вида: основанные на выводе модели и основанные на внутренней структуре весов нейронной сети.

В случае архитектуры, основанной на выводе модели, сравниваются «логиты» учителя и ученика при одних и тех же входных данных. Схематично это представлено на рис. 1.

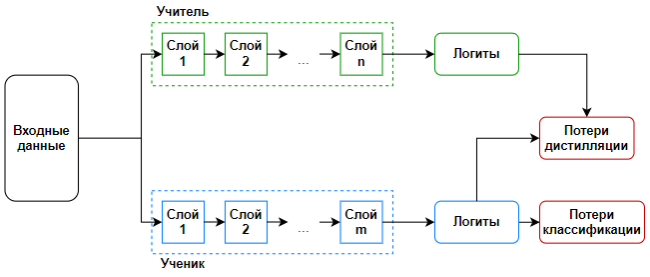

Рис. 1. Архитектура дистилляции знаний, основанная на выводе модели

Логит – это логарифмическая функция, используемая для преобразования вероятности в линейный интервал. В нейронах сети классификации, логиты представляют собой выходные значения последнего слоя сети, которые представляют собой неотличимые вероятности классов.

Эти значения используются для предсказания класса объекта и обычно преобразуются в вероятности с помощью функции активации softmax, которая гарантирует, что все вероятности суммируются до 1.

Одной из особенностей решений с данной архитектурой является то, что в большинстве случаев, для достижения желаемой точности модели ученика, обучение должно происходить в течении большого количества эпох, а логиты моделей должны быть основаны на одних и тех же изображениях. Однако, они отличаются универсальностью, так как не подразумевают изменений во внутренней структуре нейронной сети или анализ отдельных частей её скрытых слоёв, которые имеют уникальный характер для каждой модели или семейства моделей.

Архитектура, в которой дистилляция знаний реализуется на уровне внутренней структуры нейронной сети, использует передачу знаний в рамках весов, карт активаций и прочих свойств весов скрытых слоёв нейронной сети учителя и ученика (рис. 2).

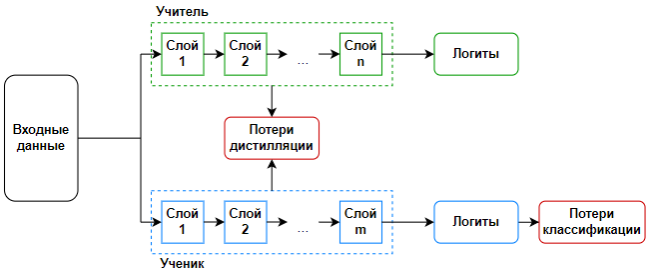

Рис. 2. Архитектура дистилляции знаний, основанная на внутренней структуре модели

Подходы дистилляции знаний, основанные на внутренней структуре модели, предоставляют ученикам дополнительную информацию путем оптимизации карт активации промежуточного слоя (intermediate activation maps) учеников, чтобы они были похожи на карты учителя. В данном случае, необходимо сделать конкретный выбор относительно того, какие слои модели учителя и ученика должны соответствовать друг другу, а, следовательно, эти подходы зависят от архитектуры используемых моделей, что усложняет достижение желаемого уровня универсальности данного подхода.

A. Достоверность

Подробный анализ идеи дистилляция знаний выявил важные закономерности для достижения необходимого результата при использовании описываемого подхода [5]. Зачастую, происходит так, что несмотря на то, что дистилляция знаний может повысить качество обобщения (способность модели делать правильные предсказания на новых данных после завершения обучения) модели ученика, не достигается основная цель, заложенная в идею дистилляции знаний – остается большое расхождение между предсказательными распределениями учителя и ученика. Подобная ситуация наблюдается даже в тех случаях, когда ученик способен идеально соответствовать учителю.

Для обозначения свойства соответствия предсказаний ученика предсказаниям учителя используется термин достоверности (fidelity). Анализ моделей дистилляции знаний, предшествующих описываемому исследованию, показывает, что уровень достоверности ученика в них достаточно мал.

Авторами были описаны причины, по которым уровень достоверности ученика играет важную роль для достижения желаемого уровня метрик его модели. Повышение достоверности ученика является наиболее очевидным способом уменьшения уровня расхождения обобщений в модели учителя и модели ученика. Между достоверностью и обобщением присутствует явная корреляция, как показано на рис. 3, в следствии чего можно сделать вывод, что повышение достоверности играет ключевую роль даже в том случае, когда разработчика модели интересует исключительно качество её обобщений.

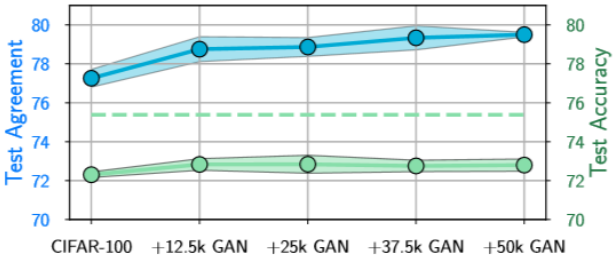

Рис. 3. Показатели достоверности и точности модели ученика [5]

Кроме того, повышение уровня достоверности улучшает модель с точки зрения интерпретируемости и надёжности. Способность выполнять перенос представлений данных из больших моделей, которые в меру своей громоздкости являются «черными ящиками», в более простые интерпретируемые модели могут способствовать выявлению закономерностей в данных, которые являются неочевидными.

B. Визуальные объяснения

В настоящее время используется разновидность подходов дистилляции знаний с архитектурой, использующей выводы нейронной сети, построенных на идее визуальных объяснений. Такое решение повышает уровень достоверности ученика за счёт ввода в модель дополнительной информации о том, почему учителем было принято конкретное решение.

Одним из методов получения визуального объяснения является Gradient-weighted Class Activation Mapping (Grad-CAM) [6]. Суть метода заключается в использовании градиентов какого-либо целевого концепта, например, для модели классификации объекта на изображении – «кот», протекающих в последний свёрточный слой нейронной сети, для последующего построения локализирующей карты, которая подсвечивает область изображения, за счет которой объекту в области изображения присваивается определённый класс. За счет локализирующих карт мы получаем информацию о дискриминативных областях входного изображения – важные части изображения, которые используются для классификации или распознавания объектов.

Этот метод позволяет выявить важные нейроны, которые после присвоения им имён [7] предоставляют человекочитаемые текстовые объяснения решений модели, которые позволяют разработчику проанализировать её внутреннюю структуру.

Отличительной чертой Grad-CAM в сравнении со своими предшественниками является универсальность, с точки зрения применимости к любой свёрточной нейронной сети без внесения изменений в их архитектуру или повторного переобучения.

Для получения класс-дискриминационной локализирующей карты методом Grad-CAM для выбранного класса $L^c_{\text{Grad-CAM}}$, применяется следующая последовательность вычислений. Сначала вычисляется отношение градиента оценки для класса – вектор, который показывает, насколько изменение входных данных влияет на оценку для определенного класса, к *k*-той свёрточной карте признаков. Оценка для класса – это число, которое показывает, насколько вероятно, что входное изображение принадлежит определенному классу. По итогу, получается частичная линеаризация α_k^c – вес, показывающий важность выбранной карты признаков A^k для класса *c*

$$\alpha_k^c = \frac{1}{HW}\sum_i^H\sum_j^W \frac{\partial y^c}{\partial A_{ij}^k},$$

где *c* – индекс класса; *k* – индекс карты признаков; *H* – высота карты признаков в пикселях; *W* – ширина карты признаков в пикселях; y^c – оценка вероятности класса *c* (вывод модели до применения softmax); *A* – карта признаков.

После этого, по отношению ко всем картам признаков в нейронной сети, получается взвешенная комбинация с последующим применением к ней ReLU

$$L^c_{\text{Grad-CAM}} = \text{ReLu}\left(\sum_k \alpha_k^c A^k\right). \quad (1)$$

В результате получаем тепловую карту того же размера, что и карты активации. Применение ReLU важно для выделения только тех пикселей, значимость которых должна быть повышена для присваивания изображению правильного класса. Без применения ReLU локализирующие карты начинают выделять лишние части изображения, которые не относятся области, на котором находится объект нужного класса.

C. Улучшенная дистилляция знаний с объяснениями

Для улучшения результатов дистилляции знаний, с учетом описанных выше проблем, связанных с достоверностью модели ученика, было предложено решение улучшенной дистилляции знаний, использующие визуальные объяснения для настройки более точного соответствия модели учителя – explanation-enhanced Knowledge Distillation – $\left(e^2KD\right)$ [8].

Для достижения этого, оптимизируется не только классическая потеря дистилляции знаний, но и схожесть в объяснениях моделей учителя и ученика. Применение такого подхода позволяет увеличить точность ученика и соответствие между его моделью и моделью учителя, убедиться, что ученик даёт схожие объяснения, то есть выдаёт правильные ответы по тем же причинам, что и учитель. Более того, этот метод независим от деталей конкретной архитектуры моделей, количества данных для обучения и позволяет использовать вычисленные заранее объяснения учителя, что ускоряет процесс обучения нейронной сети.

Для увеличения точности также используется модель объяснений GradCAM.

Для обучения модели используется следующая функция потерь:

$$\Im = \Im_{KD} + \lambda\Im_{\text{exp}},$$

где $\Im_{KD}$ – классическая функция потерь дистилляции знаний; $\Im_{\text{exp}}$ – потери схожести объяснений; λ – весовой коэффициент потери схожести объяснений.

Функция потерь классической дистилляции знаний минимизирует KL-дивергенцию между учителем (*T*) и учеником (*S*):

$$\Im_{\text{KD}} = -\tau^2\sum_{j=1}^n \sigma_j\left(\frac{z_T}{\tau}\right)\log\sigma_j\left(\frac{z_S}{\tau}\right),$$

где τ – температура; σ – функция softmax; *z* – результирующие логиты модели.

Температура – параметр, который используется для регулировки степени «мягкости» или «жёсткости» функции softmax, влияя на распределение вероятностей классов. Чем выше значение температуры, тем более равномерное распределение вероятностей классов, а при низких значениях температуры вероятности классов будут более сконцентрированы на одном классе.

Функция потери схожести объяснений имеет следующий вид:

$$\Im_{\text{exp}} = 1 - sim\left(Exp\left(T, x, \hat{y}_T\right), Exp\left(S, x, \hat{y}_T\right)\right),$$

где *sim* – функция схожести; $Exp\left(H, x, \hat{y}_T\right)$ – объяснения модели *H*; $\hat{y}_T$ – класс, определенный моделью учителя.

При дистилляции знаний применительно к свёрточным нейронным сетям, в качестве объяснений модели используется класс-дискриминационная карта, полученная методом Grad-CAM по формуле (1).

III. Результаты эксперимента на разных архитектурах дистиляции знаний

В ходе исследования проведены эксперименты с использованием различных подходов дистилляции знаний. Тестировались как архитектура дистилляции знаний, основанная на выводе модели (рис. 1), так и архитектура, основанная на внутренней структуре модели (рис. 2) при различном проценте дистилляции

знаний (4%, 15% и 100%) из модели учителя в модель ученика. В экспериментах на архитектуре рис. 2 проверялась целесообразность применения модели e^2KD в связке с объяснениями, полученными методикой Grad-CAM.

Дистилляции знаний выполнялась из модели ResNet-34 в ResNet-18.

В качестве функции схожести использовалось косинусное сходство – мера сходства между двумя векторами, которая вычисляется как косинус угла между ними. Мера показывает, насколько похожи два вектора по направлению, но не учитывает их длину. Косинусное сходство вычисляется как скалярного произведения двух векторов и деления его на произведение длин двух векторов. Результат вычисления косинусного сходства находится в диапазоне от -1 до 1, где 1 означает, что вектора идентичны, а -1 означает, что вектора противоположны.

Визуально результаты применения такого подхода показаны на рис. 4. На нем в первом столбце представлено исходное изображение, во втором объяснение учителя, в третьем объяснения ученика при использовании классического подхода дистилляции знаний и на последнем объяснение ученика при добавлении $\left(e^2KD\right)$. Если при классической дистилляции знаний объяснения ученика и учителя заметно отличались, то за счет такого подхода они максимально схожи.

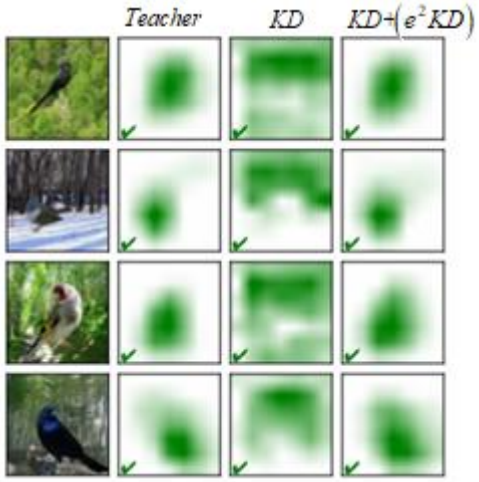

Рис. 4. Визуальные объяснения моделей [8]

Результаты, отражающие изменения в точности (Accuracy, Acc) и согласованности (Agreement, Agr) получаемой модели ученика, показаны в табл. I.

ТАБЛИЦА I. ИЗМЕНЕНИЕ В ТОЧНОСТИ И СОГЛАСОВАННОСТИ МОДЕЛИ УЧЕНИКА

Процент дистилляции	**4%**		**15%**		**100%**	
Характеристика	Acc	Agr	Acc	Agr	Acc	Agr
ResNet-18	23,3	24,8	47,0	50,2	68,9	78,8
KD	49,8	55,5	63,1	71,9	71,8	81,2
KD+e²KD	**54,9**	**61,7**	**64,1**	**73,2**	**71,8**	**81,6**

Результаты эксперимента показали, что архитектуры дистилляции знаний, основанные на внутренней структуре модели, являются более точными и согласованными в сравнении с архитектурой, основанной на выводе модели.

IV. ЗАКЛЮЧЕНИЕ

Показано, что дистилляция знаний является эффективным решением проблемы оптимизации моделей нейронных сетей для использования их на устройствах с ограниченными ресурсами.

Эффективность дистилляции доказана в ходе эксперимента, проведенного на глубоких свёрточных нейронных сетях ResNet-34 в ResNet-18.

Результаты эксперимента показали важность достоверности модели ученика по отношению к модели учителя, а также создание моделей объяснений, использование которых приводит к большей точности и согласованности модели ученика.

В дальнейшем планируется изучить влияние дистилляции данных и для других моделей нейронных сетей, где используются другие архитектуры и обрабатываются другие типы данных (тексты на естественном языке, векторы данных, аудиозаписи).

СПИСОК ЛИТЕРАТУРЫ

[1] He F., Liu T., Tao D. Why ResNet works? Residuals generalize // IEEE Transactions on Neural Networks and Learning Systems 31(12), 2020. С. 5349-5362.

[2] Wu B., Dai X., Zhang P., Wang Y., Sun F., Wu Y., Keutzer K. FBNet: Hardware-Aware Efficient ConvNet Design via Differentiable Neural Architecture Search // IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, США, 2019. С. 10726-10734. doi: 10.1109/CVPR.2019.01099.

[3] Devlin J., Chang M. W., Lee K., Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding, Minnesota, США // Proceedings of NAACL-HLT, 2019. С. 4171-4186.

[4] Hinton G.E., Vinyals O., Dean, J. Distilling the Knowledge in a Neural Network. 2015. 9 с. ArXiv, abs/1503.02531.

[5] Stanton S., Izmailov P., Kirichenko P., Alemi A.A., Wilson A.G. Does Knowledge Distillation Really Work? 2021. 21 с. ArXiv, abs/2106.05945.

[6] Selvaraju R.R., Cogswell M., Das A., Vedantam R., Parikh D., Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization // IEEE International Conference on Computer Vision (ICCV), Venice, Италия, 2017. С. 618-626. doi: 10.1109/ICCV.2017.74.

[7] Bau D., Zhou B., Khosla A., Oliva A., Torralba A. Network Dissection: Quantifying Interpretability of Deep Visual Representations // IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, США, 2017. С. 3319-3327. doi: 10.1109/CVPR.2017.354.

[8] Parchami-Araghi A., Bohle M., Rao S., Schiele B. Good Teachers Explain: Explanation-Enhanced Knowledge Distillation. 2024. 21 p.

[9] Bohle M, Fritz M., Schiele B. B-cos Networks: Alignment Is All We Need for Interpretability // IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. С. 10319-10328. doi:10.1109/CVPR52688.2022.01008.