Дистилляция знаний с сохранением геометрии признакового пространства

Т. М. Татарникова

Институт информационных технологий и программирования
Санкт-Петербургский государственный университет аэрокосмического приборостроения

tm-tatarn@yandex.ru

А. С. Раскопина

Институт информационных технологий и программирования
Санкт-Петербургский государственный университет аэрокосмического приборостроения

raskopina.anastasia@yandex.ru

***Аннотация*. В задачах сжатия и ускорения нейронных сетей широко применяется дистилляция знаний, позволяющая переносить информацию из более сложной модели-учителя в компактную модель-ученика. Большинство существующих методов дистилляции ориентированы на согласование логитов или отдельных признаков, однако структура признакового пространства при этом обычно не контролируется. В то же время геометрия представлений, отражающая взаимное расположение классов в пространстве признаков, играет важную роль в формировании разделимости данных и обобщающей способности модели. В данной работе рассматривается подход к дистилляции знаний с учетом геометрии признакового пространства. Предлагается согласовывать между моделью-учителем и моделью-учеником структуру межклассовых отношений, описываемую матрицей сходства центров классов. На основе этой идеи вводится дополнительный геометрический критерий в функцию потерь дистилляции. Проведены предварительные экспериментальные исследования на задачах классификации изображений, демонстрирующие, что учет геометрии представлений позволяет сохранять структуру признакового пространства при обучении компактной модели. Полученные результаты подтверждают перспективность геометрически согласованной дистилляции знаний и открывают возможности для дальнейшего развития методов сжатия нейронных сетей.**

Ключевые слова: дистилляция знаний, сжатие нейронных сетей, глубокое обучение, представления нейронных сетей, оптимизация

I. Введение

Современные глубокие нейронные сети демонстрируют высокую точность в широком спектре задач, включая компьютерное зрение, обработку речи, анализ текстов и другие области машинного обучения. Архитектуры глубоких моделей, такие как сверточные и трансформерные сети, позволяют эффективно извлекать сложные закономерности из данных, что приводит к значительным успехам в задачах классификации, детекции и сегментации.

Однако повышение точности моделей, как правило, сопровождается увеличением их вычислительной сложности, числа параметров и требований к памяти. Современные высокоточные модели могут содержать десятки и даже сотни миллионов параметров, что делает их использование затруднительным в условиях ограниченных вычислительных ресурсов. Это особенно актуально для мобильных устройств, встраиваемых систем и приложений реального времени, где критически важны скорость работы и энергоэффективность. Таким образом, возникает необходимость разработки методов, позволяющих уменьшить размер моделей без существенной потери качества.

Одним из наиболее распространённых и эффективных подходов к решению данной проблемы является дистилляция знаний (knowledge distillation). В рамках этого подхода знания, полученные крупной и высокоточной моделью-учителем, передаются более компактной модели-ученику. Классическая постановка задачи предполагает обучение модели-ученика не только на истинных метках, но и на «мягких» предсказаниях учителя, содержащих информацию о распределении вероятностей между классами. Такой подход позволяет передать скрытые зависимости между классами и улучшить обобщающую способность модели-ученика [1].

В дальнейшем были предложены многочисленные расширения данного подхода. Помимо согласования выходных логитов, исследователи рассматривали методы, основанные на выравнивании внутренних представлений, активаций промежуточных слоёв, а также статистических характеристик признаков. Эти методы позволяют более глубоко передавать знания, содержащиеся в модели-учителе, и зачастую приводят к дополнительному улучшению качества.

Несмотря на достигнутый прогресс, большинство существующих методов дистилляции сосредоточены на локальных аспектах представлений, таких как отдельные признаки или предсказания для конкретных объектов. При этом глобальная структура признакового пространства, формируемого нейронной сетью, как правило, явно не учитывается. Между тем именно геометрия этого пространства играет ключевую роль в формировании разделимости классов, устойчивости модели к шуму и способности к обобщению [2, 3].

В признаковом пространстве хорошо обученной модели объекты одного класса обычно группируются в компактные кластеры, тогда как объекты разных классов располагаются на значительном расстоянии друг от друга. Кроме того, между различными классами формируются определённые отношения сходства, отражающие их семантическую близость. Эти межклассовые отношения содержат важную информацию, которая не полностью передаётся через выходные вероятности модели.

В последние годы внимание исследователей привлекают методы, направленные на сохранение структурных свойств признакового пространства, таких как расстояния между объектами, угловые отношения и межклассовые связи. К таким подходам относятся методы реляционной дистилляции, сохраняющие отношения между объектами [4], а также методы, ориентированные на согласование сходства и корреляций признаков [5, 6]. Согласование геометрии представлений может позволить более полно перенести знания из модели-учителя в модель-ученика.

Тем не менее, существующие методы, ориентированные на геометрию, часто требуют сложных вычислений, работают с парами или тройками объектов и могут быть чувствительны к выбору выборок. Это ограничивает их практическое применение, особенно в задачах с большим числом классов.

В данной работе предлагается метод дистилляции знаний, основанный на согласовании геометрической структуры признакового пространства на уровне классов. В отличие от существующих подходов, метод использует центры классов как компактное представление структуры пространства и обеспечивает согласование межклассовых отношений между моделью-учителем и моделью-учеником.

Основной вклад работы заключается в следующем:

- предложен метод геометрически согласованной дистилляции, учитывающий взаимное расположение классов;
- разработан критерий согласования межклассовых отношений на основе центров признаков;
- проведено экспериментальное исследование, подтверждающее эффективность предложенного подхода по сравнению с классическими методами дистилляции.

Предложенный метод позволяет передавать не только информацию о правильных ответах, но и структуру данных, сформированную в признаковом пространстве модели-учителя, что делает его перспективным направлением для дальнейших исследований.

II. Предлагаемый метод

Предлагаемый метод направлен на сохранение геометрической структуры признакового пространства при переносе знаний от модели-учителя к модели-ученику. В отличие от классических подходов дистилляции, которые в основном ориентированы на согласование выходных предсказаний или отдельных признаков, данный метод делает акцент на сохранении взаимного расположения классов в пространстве представлений.

В современных нейронных сетях признаки, извлекаемые на промежуточных слоях, формируют некоторое многомерное пространство, в котором объекты одного класса, как правило, группируются в компактные области, а объекты разных классов располагаются на различном расстоянии друг от друга. Таким образом, структура этого пространства несёт важную информацию о взаимосвязях между классами, их сходстве и различиях. Эта информация не полностью отражается в выходных вероятностях модели, что ограничивает эффективность традиционных методов дистилляции.

Основная идея предлагаемого метода заключается в явном моделировании этой структуры. Для этого каждый класс представляется в виде центра признаков – усреднённого вектора, полученного по всем объектам данного класса. Такие центры можно интерпретировать как «репрезентативные точки» классов в признаковом пространстве. Они позволяют перейти от анализа отдельных объектов к анализу структуры пространства в целом.

На первом этапе вычисляются центры классов для модели-учителя. Полученный набор центров отражает геометрию признакового пространства, сформированного учителем, включая взаимное расположение классов и их относительную близость. Это пространство можно рассматривать как эталонное, поскольку модель-учитель обладает более высокой выразительной способностью и, как правило, формирует более качественные представления.

Аналогичная процедура выполняется для модели-ученика. Несмотря на меньшую ёмкость, модель-ученик также формирует собственное признаковое пространство, однако его структура может существенно отличаться от структуры учителя. В частности, классы могут располагаться менее упорядоченно, иметь меньшую разделимость или не отражать истинные взаимосвязи между данными.

Для количественного описания геометрии пространства вводится матрица межклассового сходства. Эта матрица отражает степень близости между центрами различных классов и тем самым описывает их взаимное расположение. Элементы матрицы характеризуют, насколько два класса похожи с точки зрения их признаковых представлений. Таким образом, матрица сходства выступает компактным и информативным представлением глобальной структуры пространства.

Ключевая идея метода заключается в согласовании этих матриц для модели-учителя и модели-ученика. В процессе обучения модель-ученик оптимизируется таким образом, чтобы её матрица межклассового сходства была максимально близка к аналогичной матрице учителя. Это означает, что ученик не просто воспроизводит правильные ответы, но и перенимает структуру взаимосвязей между классами.

Для реализации данного подхода в функцию потерь вводится дополнительный геометрический критерий, который измеряет различие между матрицами сходства. Этот критерий дополняет стандартную функцию потерь классификации и выступает в роли регуляризатора, направленного на структурное согласование представлений. Баланс между классификационной и геометрической составляющими регулируется с помощью соответствующего коэффициента.

Важно подчеркнуть, что предлагаемый метод ориентирован на сохранение именно глобальной структуры признакового пространства. В отличие от методов, выравнивающих отдельные признаки или пары объектов, данный подход учитывает отношения между всеми классами одновременно. Это позволяет более полно передавать знания, содержащиеся в модели-учителе.

Ещё одним важным преимуществом является архитектурная независимость метода. Поскольку используются агрегированные представления (центры классов), отсутствует необходимость в послойном согласовании признаков между моделями. Это упрощает применение метода и делает его универсальным для различных архитектур и сценариев обучения.

Кроме того, использование центров классов позволяет снизить влияние шума и выбросов в данных, так как усреднение по множеству объектов делает представления более устойчивыми. Это особенно важно в задачах с большим числом классов или сложной структурой данных.

Таким образом, предложенный метод обеспечивает перенос не только информации о правильных ответах, но и более глубоких знаний о структуре данных, закодированной в признаковом пространстве модели-учителя. Это приводит к формированию более организованных, интерпретируемых и потенциально более устойчивых представлений у модели-ученика.

III. Экспериментальные исследования

В данной работе проведено экспериментальное исследование эффективности различных методов дистилляции знаний на задаче классификации изображений. В качестве датасета использовался CIFAR-100, содержащий 100 классов объектов с небольшим числом обучающих примеров на класс, что делает задачу достаточно сложной и чувствительной к качеству представлений.

В качестве модели-учителя была выбрана архитектура ResNet-50, обладающая большей емкостью и способностью извлекать более сложные признаки. В роли модели-ученика использовалась более компактная архитектура ResNet-18. Обе модели обучались в одинаковых условиях для обеспечения корректного сравнения: число эпох обучения составляло 30, размер батча – 64, оптимизационный алгоритм и параметры обучения были одинаковыми для всех рассматриваемых методов.

Сначала была обучена модель-учитель, которая затем использовалась для передачи знаний в процессе обучения модели-ученика. Для оценки эффективности дистилляции были рассмотрены следующие варианты обучения модели-ученика:

- обучение без дистилляции (baseline);
- классическая дистилляция знаний на основе согласования логитов;
- дистилляция с согласованием признаковых представлений;
- предложенный метод дистилляции с учетом геометрии признакового пространства.

Во всех случаях обучение проводилось с использованием одних и тех же обучающих и тестовых выборок. Качество моделей оценивалось по метрике Top-1 accuracy на тестовой части датасета. Для каждого метода фиксировалось наилучшее достигнутое значение точности в процессе обучения.

Результаты экспериментов приведены в табл. 1.

ТАБЛИЦА I.

Метод обучения	**Top-1 точность (%)**	**Сохранение геометрии признаков**	**Комментарий**
Модель-учитель	72.81	Высокое	Учитель на базе ResNet-50
Обучение без дистилляции	74.64	Низкое	Базовая модель-ученик
Классическая дистилляция знаний	76.13	Среднее	Перенос логитов учителя
Дистилляция по признаковым представлениям	76.00	Среднее	Согласование признаковых представлений
Предлагаемый геометрический метод	75.81	Высокое	Согласование межклассовых отношений

Полученные результаты показывают, что все рассмотренные методы дистилляции превосходят базовое обучение модели-ученика без передачи знаний от учителя. Базовая модель ResNet-18, обученная без дистилляции, достигла точности 74.64%, тогда как использование классической дистилляции повысило точность до 76.13%. Метод согласования признаковых представлений также продемонстрировал улучшение качества, обеспечив точность 76.00%. Предлагаемый геометрически согласованный подход достиг 75.81%, что также превышает результат базовой модели.

Таким образом, учет дополнительной информации от модели-учителя положительно влияет на обучение компактной сети. Наилучший результат в проведенных экспериментах показала классическая дистилляция, обеспечив прирост 1.49 процентного пункта относительно baseline. Метод согласования признаков дал сопоставимый прирост в 1.36 процентного пункта, а предложенный геометрический метод улучшил результат базовой модели на 1.17 процентного пункта.

Следует отметить, что в рассматриваемой экспериментальной постановке точность модели-учителя составила 72.81%, что оказалось ниже точности базовой модели-ученика. Несмотря на это, все методы дистилляции оказались полезными. Это позволяет предположить, что даже в случае, когда учитель не превосходит ученика по итоговой точности, он все равно может передавать полезную информацию о структуре распределения классов и внутренней организации признакового пространства.

На рис. 1 представлено сравнение динамики функции потерь для различных методов обучения. Видно, что все методы демонстрируют устойчивое снижение функции потерь, что свидетельствует о корректной сходимости процесса обучения.

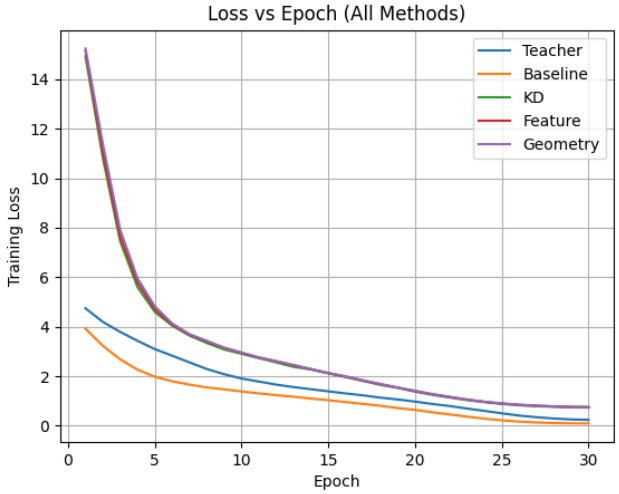

Рис. 1. Сравнение динамики функции потерь для различных методов обучения: базового обучения, классической дистилляции, дистилляции признаков и предложенного геометрического подхода

Базовая модель (без дистилляции) характеризуется наиболее быстрым снижением функции потерь, что связано с оптимизацией только по задаче классификации. В то же время методы дистилляции (KD, feature и geometry) имеют более высокие значения функции потерь на начальных этапах обучения, что объясняется наличием дополнительных регуляризирующих слагаемых.

По мере обучения различия между методами уменьшаются, и все кривые сходятся к близким значениям. При этом методы дистилляции обеспечивают более высокую итоговую точность по сравнению с базовым обучением, несмотря на более сложную структуру функции потерь.

Предлагаемый геометрический метод не показал наилучшего результата по метрике точности, однако продемонстрировал устойчивое улучшение по сравнению с обучением без дистилляции. Это подтверждает, что учет межклассовых отношений в признаковом пространстве является перспективным направлением для переноса знаний. Вместе с тем полученные результаты указывают на необходимость дальнейшей настройки веса геометрического критерия и более детального исследования способов согласования структуры признакового пространства между учителем и учеником.

IV. Заключение

Полученные результаты демонстрируют, что использование различных стратегий дистилляции знаний оказывает существенное влияние как на точность модели-ученика, так и на структуру формируемых признаковых представлений. Базовая модель, обученная без дистилляции, показывает достаточно высокую точность (74.64%), что подтверждает способность архитектуры эффективно обучаться напрямую по меткам. Однако при этом наблюдается низкий уровень сохранения геометрии признакового пространства, что указывает на отсутствие явного контроля над внутренней структурой представлений.

Классическая дистилляция знаний, основанная на передаче логитов учителя, обеспечивает наилучший результат по точности (76.13%), что свидетельствует о высокой эффективности данного подхода с точки зрения улучшения обобщающей способности модели. Аналогично, дистилляция по признаковым представлениям демонстрирует сопоставимый уровень точности (76.00%), подтверждая, что прямое согласование внутренних активаций также является эффективной стратегией переноса знаний.

В то же время предложенный геометрически согласованный метод показывает несколько более низкую точность (75.81%) по сравнению с классическими подходами дистилляции, однако обеспечивает значительно более высокий уровень сохранения геометрической структуры признакового пространства. Это указывает на то, что модель-ученик не только приближается к выходам учителя, но и воспроизводит взаимное расположение объектов в пространстве признаков, включая межклассовые отношения.

Таким образом, полученные результаты подтверждают наличие компромисса между максимизацией точности и сохранением геометрической структуры представлений. В отличие от традиционных методов дистилляции, ориентированных преимущественно на выходные распределения или отдельные признаки, предложенный подход позволяет явно контролировать структуру пространства признаков, что может быть особенно важно для задач, чувствительных к геометрии представлений, таких как кластеризация, поиск похожих объектов и перенос обучения.

В целом, разработанный метод демонстрирует конкурентоспособные результаты по точности при одновременном обеспечении более структурированных и интерпретируемых представлений, что делает его перспективным направлением для дальнейших исследований в области структурной дистилляции и обучения представлений.

Список литературы

[1] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," arXiv preprint arXiv:1503.02531, 2015.

[2] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 8, pp. 1798–1828, 2013.

[3] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," Journal of Machine Learning Research, vol. 10, pp. 207–244, 2009.

[4] W. Park, D. Kim, Y. Lu, and M. Cho, "Relational knowledge distillation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3967–3976.

[5] F. Tung and G. Mori, "Similarity-preserving knowledge distillation," in Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2019, pp. 1365–1374.

[6] B. Peng, X. Jin, J. Liu, D. Li, Y. Wu, Y. Liu, S. Zhou, and Z. Zhang, "Correlation congruence for knowledge distillation," in Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2019, pp. 5007–5016.