На правах рукописи

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Фигурнов Михаил Викторович

ВЕРОЯТНОСТНЫЙ МЕТОД ДЛЯ АДАПТИВНОГО ВРЕМЕНИ ВЫЧИСЛЕНИЙ В НЕЙРОННЫХ СЕТЯХ

РЕЗЮМЕ
диссертации на соискание учёной степени
кандидата компьютерных наук НИУ ВШЭ

Москва — 2019

Диссертационная работа выполнена в Национальном исследовательском университете «Высшая школа экономики».

Научный руководитель: Ветров Дмитрий Петрович, к.ф.-м.н., профессор-исследователь, Национальный исследовательский университет «Высшая школа экономики».

Тема диссертации

В диссертационной работе разработан вероятностный метод для пространственной адаптации вычислительного времени популярной модели компьютерного зрения — свёрточной нейронной сети. Применение этого метода повышает вычислительную эффективность и интерпретируемость.

Актуальность темы. В последние годы в мире наблюдается взрывной рост объёмов собираемых данных. В связи с этим возрастает актуальность методов машинного обучения, позволяющих автоматически извлекать закономерности из данных. В задачах машинного обучения предполагается, что объекты реального мира описаны с помощью *признаков*, а также что имеется *обучающая выборка*, полученная из генеральной совокупности объектов. В задаче обучения с учителем для объектов обучающей выборки также известны истинные *метки* и требуется восстановить зависимость меток от признаков. *Качество* полученного решения обычно оценивается точностью — долей правильно определённых меток на тестовой выборке. На сегодняшний день наиболее успешны именно методы обучения с учителем, хотя разметка обучающей выборки может оказаться крайне трудозатратной. Альтернативой этому подходу является обучение без учителя, в котором обучающая выборка состоит лишь из признаков объектов. Цель обучения без учителя — получение более компактного и информативного описания объектов, которое затем может использоваться, например, для обучения с учителем по меньшей размеченной выборке [1].

Популярным способом решения упомянутых задач машинного обучения является вероятностное моделирование. В случае обучения с учителем вероятностная модель задаёт распределение над метками при условии наблюдаемых данных. Для обучения без учителя в модель, как правило, вводятся латентные (ненаблюдаемые) переменные, определяющие факторы вариации данных. Параметры вероятностной модели настраиваются при помощи метода максимального правдоподобия, используя обучающую выборку и градиентные методы оптимизации. Во многих случаях правдоподобие модели с латентными переменными не может быть подсчитано аналитически. Тогда применяются вариационные методы, такие как вариационная нижняя оценка на правдоподобие.

Успех методов машинного обучения принципиально зависит от информативности признакового описания объектов. Одними из наиболее

сложных с точки зрения построения признакового описания объектами являются высокоразмерные неструктурированные данные: изображения, звуки, тексты, графы и т.д. При этом объём именно таких данных растёт с огромной скоростью в связи с распространением интернета и социальных сетей. К началу 2010-х годов были разработаны методы извлечения признаков из этих данных, основанные на экспертных знаниях о предметных областях. Например, в задачах обработки изображений широко использовались признаки SIFT [2] и HOG [3], а при обработке звука — признаки MFCC [4]. К сожалению, информативность таких признаков оставалась неудовлетворительной для решения практически важных задач, а отсутствие очевидных способов их улучшения привело к стагнации качества методов [5; 6].

В последние пять лет глубинное обучение (deep learning) стало наиболее эффективным способом работы с высокоразмерными неструктурированными данными [7]. Глубинное обучение предлагает использовать многослойные (глубинные) признаковые описания объектов, задаваемые нейросетями с десятками и сотнями слоёв. При этом архитектура нейросети выбирается исходя из особенностей данных. Так, для обработки изображений популярны свёрточные нейронные сети (СНС) [8], а для работы со звуками и текстами — рекуррентные нейронные сети (РНС) [9]. Как правило, последний слой нейронной сети соответствует ответу на поставленную задачу, например, вероятностному распределению над метками. Все параметры модели, число которых может достигать миллиардов [10], настраиваются при помощи стохастических градиентных методов оптимизации, максимизирующих правдоподобие вероятностной модели. Таким образом, глубинное обучение рассматривает параметрические модели, выбираемые исходя из особенностей данных, и сравнительно простые методы обучения.

Ключевыми факторами успеха глубинного обучения стало создание сверхбольших *размеченных* обучающих выборок, таких как ImageNet [6], и развитие вычислительных технологий, в частности, видеоускорителей. В 2012 году команда из Торонто успешно обучила свёрточную нейронную сеть (СНС) для задачи классификации изображений [11]. Команде удалось существенно улучшить качество работы по сравнению со всеми предыдущими подходами, не использующими нейросети. Вскоре после этого СНС стали важнейшим элементом систем компьютерного зрения. Использование СНС позволило значительно продвинуться в решении задач понимания сцены (распознавания образов), таких как классификация

изображений, идентификация объектов, детекция объектов и семантическая сегментация. При этом оказалось, что улучшение качества работы может быть достигнуто путём наращивания объёма вычислений, в первую очередь за счёт увеличения глубины (числа слоёв) СНС. Так, упомянутая СНС 2012 года состояла из 8 слоёв, а остаточная сеть, предложенная в 2015 году, — из 152 слоёв [12].

Несмотря на прорыв в качестве решения задач, у модели СНС имеется ряд недостатков:

1. СНС имеют огромную вычислительную стоимость, в основном определяемую свёрточными слоями (более 80% времени вычислений). Современные СНС используют десятки миллиардов операций с плавающей запятой для обработки одного изображения. Подобные вычислительные требования существенно усложняют использование СНС во многих случаях: обработка видеопотока в режиме реального времени, применение в устройствах без мощных видеоускорителей, а также в устройствах, где энергопотребление играет решающую роль.
2. СНС плохо интерпретируемы. Сложная структура моделей, большое число параметров и вычислений приводят к тому, что классические методы анализа моделей неприменимы к СНС. Из-за этого применение СНС затруднено в областях, где высока цена ошибки и требуется возможность валидации решения системы человеком. На сегодняшний день разработан ряд методов для интерпретации уже обученных СНС [13; 14]. Однако актуальной задачей является разработка более интерпретируемых СНС.

Для решения этих проблем в диссертационной работе используется предположение, что СНС *пространственно избыточны*, то есть применение части слоёв сети в некоторых пространственных позициях не является необходимым для получения высокого качества работы. Таким образом, методы, позволяющие пропустить часть свёрточных слоёв в некоторых пространственных позициях, могут улучшить соотношение между скоростью и качеством работы СНС. Кроме того, если пропускаемые пространственные позиции выбираются под конкретный объект, получаемые карты объёма вычислений повышают интерпретируемость СНС: области, которым выделяется больше вычислений, являются более важными для решаемой задачи. Такой механизм аналогичен биологическим системам зрения, которые тратят больше времени на анализ важных частей представленного изображения [15].

Механизм пространственного варьирования объёма вычислений может быть рассмотрен как модель внимания. Существующие в настоящее время модели внимания, применимые к СНС, обладают значительными недостатками. Так, «glimpse-based» модели внимания [16—19] не применимы ко многим классам задач (детекция объектов, сегментация изображений, генерация изображений); мягкие модели пространственного внимания (soft spatial attention models) [20; 21] не позволяют снизить объём вычислений; модели жёсткого внимания (hard attention models) [20; 22] настраиваются при помощи метода REINFORCE [23], который существенно затрудняет обучение сети.

Целью данной работы является разработка метода улучшения соотношения между скоростью обработки и качеством СНС.

Для достижения данной цели решены следующие **задачи**:

1. Разработан перфорированный свёрточный слой, позволяющий пространственно варьировать и снижать объём вычислений.
2. Метод адаптивного времени вычислений [24], предложенный ранее для РНС, применён для пространственной адаптации глубины (числа слоёв) СНС под конкретный объект.
3. Построена вероятностная модель адаптации пространственной глубины СНС и предложен способ её обучения.

Основные результаты и выводы

Научная новизна работы заключается в том, что впервые установлены следующие положения:

1. Сокращение пространственной избыточности промежуточных представлений сети позволяет повысить скорость работы СНС.
2. Пространственная адаптация глубины (числа слоёв) СНС в зависимости от объекта улучшает соотношение между скоростью и качеством работы СНС, а также повышает интерпретируемость модели.
3. Варьирование глубины СНС может осуществляться вероятностной моделью с латентными переменными.

Практическая значимость. Полученные результаты расширяют область практической применимости СНС за счёт улучшения соотношения между скоростью и качеством работы и повышения интерпретируемости.

Методология и методы исследования. Использована методология глубинного обучения, аппарат вероятностного моделирования, языки программирования Python, CUDA, MATLAB, библиотеки NumPy, MatConvNet, TensorFlow.

Достоверность результатов обеспечивается детальным изложением используемых методов, алгоритмов, доказательствами теорем, а также описанием экспериментов и публикацией исходного кода, что обеспечивает воспроизводимость.

Основные положения, выносимые на защиту:

1. Метод перфорирования свёрточных сетей, позволяющий пространственно варьировать объём вычислений в СНС.
2. Метод пространственно-адаптивного времени вычислений для настройки глубины (числа слоёв) СНС в зависимости от объекта и пространственной позиции.
3. Вероятностная модель с латентными переменными для адаптации глубины СНС, а также метод стохастической вариационной оптимизации для настройки модели.
4. Экспериментальная валидация предложенных методов, включающая сравнение с аналогами.

Личный вклад в положения, выносимые на защиту. Результаты получены диссертантом лично. В работах по теме диссертации диссертантом предложены ключевые научные идеи, реализованы и проведены эксперименты, написан текст статей. Результаты из подраздела 4.4 работы «PerforatedCNNs: Acceleration through Elimination of Redundant Convolutions» (NIPS 2016) получены Айжан Ибрагимовой и не включены в текст диссертации. Вклад остальных соавторов заключается в рецензировании программного кода экспериментов, технической помощи в постановке экспериментов, обсуждениях полученных результатов, правках текста статей, постановке решаемой задачи и общем руководстве исследованиями.

Публикации и апробация работы

Во всех публикациях по теме диссертации соискатель является главным автором.

Публикации повышенного уровня.

1. *Figurnov M., Ibraimova A., Vetrov D. P., Kohli P.* PerforatedCNNs: Acceleration through Elimination of Redundant Convolutions //

Advances in Neural Information Processing Systems 29. 2016. P. 947–955. Конференция ранга A*, индексируется SCOPUS.

2. *Figurnov M., Collins M. D., Zhu Y., Zhang L., Huang J., Vetrov D., Salakhutdinov R.* Spatially Adaptive Computation Time for Residual Networks // The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017. P. 1039–1048. Конференция ранга A*, индексируется SCOPUS.
3. *Figurnov M., Sobolev A., Vetrov D.* Probabilistic adaptive computation time // Bulletin of the Polish Academy of Sciences: Technical Sciences. 2018. Vol. 66, no. 6. P. 811–820. Журнал индексируется Web of Science (Q2) и SCOPUS (Q3).

Прочие публикации.

1. *Figurnov M., Vetrov D. P., Kohli P.* PerforatedCNNs: Acceleration through Elimination of Redundant Convolutions // International Conference on Learning Representations (ICLR) Workshop. 2016.

Доклады на конференциях и семинарах.

1. Семинар научной группы байесовских методов, г. Москва, 20 февраля 2015 г. Тема: «Ускорение свёрточных нейронных сетей».
2. Рождественский коллоквиум по компьютерному зрению, Сколтех, г. Москва, 28 декабря 2015 г. Тема: «PerforatedCNNs: Acceleration through Elimination of Redundant Convolutions».
3. Семинар ИППИ РАН «Структурные модели и глубинное обучение», г. Москва, 21 марта 2016 г. Тема: «Acceleration of Convolutional Neural Networks through Elimination of Redundant Convolutions».
4. Международная конференция по обучению представлений «International Conference on Learning Representations 2016», дополнительная секция (воркшоп), г. Сан-Хуан, Пуэрто-Рико, США, 3 мая 2016 г. Тема: «PerforatedCNNs: Acceleration through Elimination of Redundant Convolutions».
5. Международная конференция по нейронным системам обработки информации «Conference on Neural Information Processing Systems 2016», основная секция, г. Барселона, Испания, 7 декабря 2016 г. Тема: «PerforatedCNNs: Acceleration through Elimination of Redundant Convolutions».
6. Семинар компании OpenAI, г. Сан-Франциско, Калифорния, США, 1 марта 2017 г. Тема: «Spatially Adaptive Computation Time for Residual Networks».

7. Семинар научной группы байесовских методов, г. Москва, 10 марта 2017 г. Тема: «Spatially Adaptive Computation Time for Residual Networks».
8. Международный саммит «Машины могут видеть», г. Москва, 9 июня 2017 г. Тема: «Spatially Adaptive Computation Time for Residual Networks».
9. Международная конференция по компьютерному зрению и распознаванию образов «IEEE Conference on Computer Vision and Pattern Recognition 2017», основная секция, г. Гонолулу, Гавайи, США, 22 июля 2017 г. Тема: «Spatially Adaptive Computation Time for Residual Networks».
10. Рождественский коллоквиум по компьютерному зрению, Сколтех, г. Москва, 26 декабря 2017 г. Тема: «Spatially Adaptive Computation Time for Residual Networks».

Содержание работы

Объем и структура работы. Диссертация состоит из введения, четырёх глав и заключения. Полный объём диссертации составляет 116 страниц, включая 30 рисунков и 7 таблиц. Список литературы содержит 167 наименований.

Во **введении** обосновывается актуальность проводимых в данной диссертационной работе исследований, формулируется цель и задачи работы, излагается научная новизна работы и положения, выносимые на защиту.

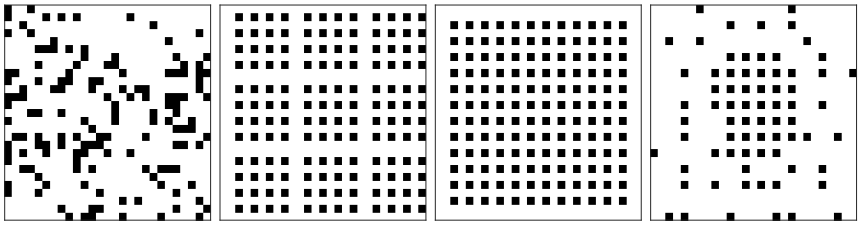
Первая глава носит обзорный характер и состоит из двух частей. В первой части приводится обзор методов глубинного обучения, в частности свёрточных нейронных сетей (СНС). Описываются решаемые СНС задачи: классификация и сегментация изображений, детекция объектов и др. Формулируется задача обучения с учителем, в которой на сегодняшний день наиболее эффективны методы глубинного обучения. Излагаются методы стохастической оптимизации и алгоритм обратного распространения ошибки для настройки параметров (обучения) нейронных сетей. Также описаны методы инициализации параметров нейронных сетей, существенно влияющие на итоговое качество работы из-за невыпуклости оптимизируемого функционала. Затем рассматриваются наиболее распространённые слои нейронных сетей: полносвязный слой, различные функции активации, слой дропаута, софтмакс. Далее описывают-

ся слои, специфичные для СНС: свёрточный слой, слой пулинга, батч-нормализация и др. Приводится историческая справка о конкурсе по классификации изображений ImageNet и примеры свёрточных нейронных архитектур, наилучшим образом зарекомендовавших себя на практике: AlexNet, VGG-16, остаточная сеть (ResNet). Во второй части главы рассматриваются методы обучения параметров случайных переменных, требуемые для настройки нейросетей со стохастическими переменными. Описывается метод REINFORCE, применимый к широкому классу вероятностных распределений, но обладающий высокой дисперсией градиентов; трюк репараметризации, применимый лишь для узкого класса непрерывных переменных, но имеющий низкую дисперсию градиентов; релаксация Гумбель-Софтмакс, позволяющая обучать параметры дискретных переменных при помощи репараметризации.

Во **второй главе** предлагается метод перфорации СНС, позволяющий ускорить работу СНС за счёт уменьшения пространственной избыточности. Метод назван по аналогии с методом перфорации циклов, ускоряющим программы за счёт пропуска некоторых итераций в циклах.

Сначала описывается *перфорированный свёрточный слой*, являющийся модификацией обычного свёрточного слоя и имеющий такие же размерности входа, выхода и тензора весов. Ключевой гиперпараметр перфорированного свёрточного слоя — *маска перфорации*, подмножество пространственных позиций выхода свёрточного слоя. *Степенью перфорации* называется доля пространственных позиций, не лежащих в маске перфорации. Значения выходов слоя в пространственных позициях из маски перфорации вычисляются точно, то есть равны значениям обычного свёрточного слоя. Значения остальных пространственных позиций интерполируются значением ближайшей позиции, вычисленной точно. Возможны и другие методы интерполяции, например, замена пропущенных значений на нули. Из определения ясно, что перфорированный свёрточный слой является обобщением стандартного свёрточного слоя. Эквивалентность с ним достигается в том случае, если маска перфорации содержит все пространственные позиции.

Предлагается несколько способов генерации масок перфорации, то есть выбора подмножества пространственных позиций, рис. 1. *Равномерная* маска получается равновероятным выбором позиций без возвращения. Её недостатком является то, что полученные позиции часто образуют компактные группы, что увеличивает среднее расстояние до точек из маски перфорации. *Решётчатая* маска строится как декартово произве-



а) Равномерная б) Решётчатая в) Структурная г) Важностная

Рис. 1 — Примеры масок перфорации.

дение подмножеств позиций по каждой из координат, выбираемых псевдослучайной схемой генерации последовательностей целых чисел [25]. Если число позиций делит нацело размерность выхода, то маска представляет собой регулярную решётку. В противном случае в решётке присутствуют нерегулярности. *Структурная* маска перфорации содержит позиции, используемые в следующем слое пулинга максимальное число раз. Эта маска основана на наблюдении, что при некоторых параметрах слоя пулинга, например, при ядре размера 3×3 и шаге 2 (такие параметры применяются в сетях Network in Network и AlexNet), различные выходы свёрточного слоя используются разное число раз. Наконец, *важностная* маска перфорации учитывает относительный вклад пространственных позиций в функцию потерь. Назовём *важностью позиции* на выходе свёрточного слоя для конкретного изображения приближение первого порядка по Тейлору абсолютной величины изменения функции потерь при замене истинного значения в этом выходе на ноль. Важность всех выходных позиций может быть эффективно подсчитана при помощи алгоритма обратного распространения ошибки. *Важность пространственной позиции* определим как сумму важностей позиций по всем каналам, усреднённую по объектам обучающей выборки. Важностная маска перфорации содержит пространственные позиции с наибольшей важностью. На выборке ImageNet в важностной маске преобладают центральные позиции, поскольку классифицируемый объект как правило центрирован. Также автоматически появляется решётка, характерная для структурной маски перфорации.

Преимуществом перфорированного свёрточного слоя является возможность эффективной реализации, то есть получение сокращения времени вычислений, близкое к снижению числа операций. Для этого предлагается произвести сведение вычисления слоя к матричному умножению: из входного тензора вырезаются подтензоры, соответствующие выход-

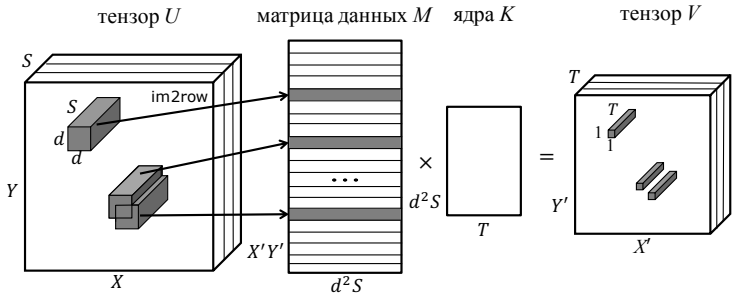


Рис. 2 — Сведение вычисления свёрточного слоя к матричному умножению.

ным значениям из маски перфорации, и помещаются в строки матрицы данных (см. рис. 2). Умножение матрицы данных на матрицу ядра позволяет получить точные значения свёрточного слоя в пространственных позициях из маски перфорации. Интерполяция пропущенных значений выполняется *неявно* путём индексации операций чтения в следующем слое. За счёт этого метод также достигает сокращения потребляемой памяти.

Экспериментальная валидация метода производится на задаче классификации изображений. Сначала предложенные виды масок перфорации сравниваются на стандартной задаче ускорения одного свёрточного слоя сети AlexNet (выборка ImageNet). Показывается, что наилучшими являются решётчатая и важностная маски перфорации. Затем на сети Network in Network (выборка CIFAR-10) демонстрируется превосходство соотношения скорости и качества работы перфорации над более простыми базовыми методами, такими как увеличение шага свёрточного слоя и уменьшение разрешения входного изображения. Наконец, перфорация используется для ускорения целых сетей AlexNet и VGG-16 на выборке ImageNet, что свидетельствует о применимости перфорации для ускорению СНС большого размера. Рассматриваемые сети можно ускорить в два раза на CPU и на GPU с увеличением ошибки не более, чем на 2,6%. Теоретические ускорения как правило близки к эмпирическим, что доказывает эффективность предлагаемой реализации перфорированного свёрточного слоя.

В **третьей главе** рассматривается метод пространственно-адаптивного времени вычислений (ПАВВ) для остаточных сетей. Данный метод позволяет фокусировать вычислений на важных областях изображения, рис. 3.

Сначала метод адаптивного времени вычислений (АВВ) [24], предложенный ранее для РНС, применяется к остаточным сетям, распростра-

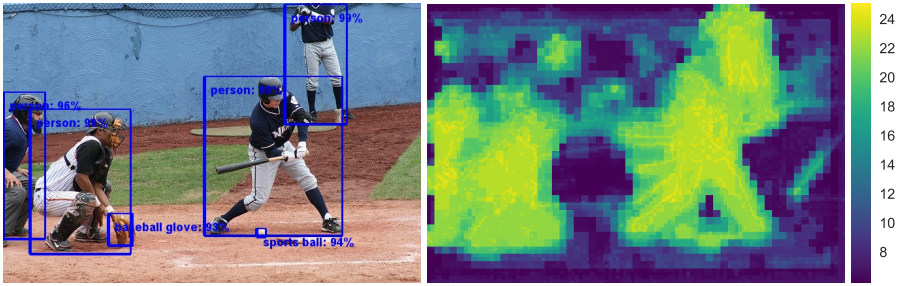


Рис. 3 — Детектированные объекты (слева) и карта стоимости вычислений предлагаемого метода ПАВВ (справа) для валидационного изображения выборки данных COCO. Метод ПАВВ использует больше вычислений в областях изображения, которые похожи на объект.

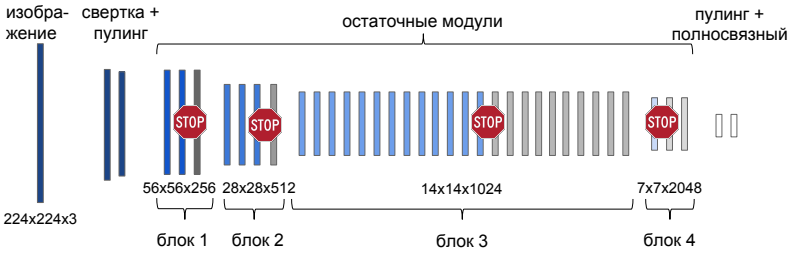


Рис. 4 — Метод адаптивного времени вычислений для остаточной сети ResNet-101.

нённой архитектуре CHC. Остаточная сеть состоит из *остаточных модулей*, функций вида $U^l = U^{l-1} + f(U^{l-1})$, где $f(U^{l-1})$ — свёрточная нейронная подсеть, называемая *остаточной функцией*. Последовательность остаточных модулей с одинаковыми размерностями выхода будем называть *остаточным блоком*. В методе АВВ каждый остаточный модуль дополнительно возвращает *вероятность остановки*, число из отрезка $[0; 1]$. Остаточные модули и их вероятности остановки вычисляются последовательно. Как только кумулятивная сумма вероятностей остановки достигает единицы, все последующие остаточные модули в текущем блоке пропускаются. Определим распределение остановки как вычисленные вероятности остановки, где последнее значение заменено на *остаток*. Величина остатка выбирается исходя из условия нормировки вероятностного распределения. Выход блока определяется как взвешенная сумма выходов остаточных модулей с весами, равными соответствующим значениям вероятности. Наконец, вводится *стоимость вычислений*, равная сумме числа выполненных остаточных модулей и величины остатка. Минимиз-

зация стоимости вычислений увеличивает вероятность остановки у всех модулей, кроме последнего, что приводит к более ранней остановке. Стоимость вычислений используется как регуляризатор для исходной функции потерь. Описанный метод применяется к каждому остаточному блоку сети независимо, а стоимости вычислений блоков суммируются. Таким образом, в каждом блоке выполняются лишь первые несколько модулей, рис. 4. Доказывается, что метод ABB обобщает модель остаточной сети.

Предлагается метод ПАВВ, идея которого заключается в применении ABB к каждой пространственной позиции блока. У каждой позиции имеется своя вероятность остановки, а выход блока в некоторой пространственной позиции определяется как взвешенная комбинация значений остаточных модулей блока в этой же позиции. Пространственная позиция остаточного модуля называется *активной*, если её кумулятивная сумма вероятностей остановки не превышает единицу. Ясно, что на выходы блока влияют лишь значения в активных позициях, поэтому разумно вычислять только эти значения. Значения остаточных функций в неактивных позициях доопределяются нулями, что эквивалентно копированию предыдущих значений. Эффективная реализация остаточного модуля, вычисляемого лишь в активных позициях, может быть осуществлена при помощи перфорированного свёрточного слоя, предложенного во второй главе, с заменой пропускаемых значений на ноль. Доказывается, что метод ПАВВ является обобщением метода ABB, а значит, и остаточной сети.

Для экспериментальной валидации методы ABB и ПАВВ применяются к модели остаточной сети ResNet-101. Описывается проблема *мёртвых остаточных модулей*, присущая ABB и ПАВВ: при «неправильной» инициализации модели последние остаточные модули блоков никогда не используются. Для её решения предлагается несколько эвристик инициализации.

Сначала рассматривается задача классификации изображений (выборка ImageNet). В качестве базового метода для сравнения ABB и ПАВВ используется неадаптивная остаточная сеть с близким числом операций с плавающей запятой. При увеличении разрешения на этапе теста, что является стандартной практикой, ПАВВ превосходит ABB и базовые методы по соотношению точности и числа операций. Показано, что превосходство ПАВВ сохраняется при увеличении разрешения на этапе обучения.

Далее решается задача детекции объектов на выборке COCO. В ней используются изображения с высоким разрешением, например, $1000 \times$

600, что значительно превосходит стандартные 224×224 для классификации ImageNet. Механизм ПАВВ позволяет уменьшить время вычислений для малоинформативного фона. Используется метод детекции объектов Faster R-CNN [26], в котором остаточная сеть, извлекающая признаки, заменяется на ПАВВ. Показано, что такой подход достигает лучшего соотношения между скоростью и величиной усреднённой средней точности (mean average precision, mAP) по сравнению с базовым подходом, состоящим в использовании неадаптивной модели остаточной сети ResNet для извлечения признаков. Пример детекций и карты стоимости вычислений приведён на рис. 3.

В конце главы приводятся эксперименты, показывающие, что карты стоимости вычислений ПАВВ хорошо коррелируют с визуальной значимостью. Для этого используется большая выборка cat2000, полученная путём демонстрации изображений людям и измерения позиций фиксаций их глаз. Целевая карта получается как сглаженная гистограмма этих позиций. Используются модели ПАВВ, заранее обученные на ImageNet и COCO; дообучение на задаче предсказания визуальной значимости не проводится. Однако выполняется параметрическая обработка карт стоимости вычислений, направленная на сглаживание и учёт центрального смещения в целевых картах из выборки cat2000. Показывается преимущество обработанных карт стоимости вычислений над базовым методом, центрированной гауссианой. Таким образом ПАВВ автоматически обучается фокусироваться на тех регионах изображения, которые кажутся важными людям, что повышает интерпретируемость СНС.

В **четвертой главе** предлагается метод вероятностного адаптивного времени вычислений. В его основе лежит вероятностная модель, в которой дискретные латентные переменные задают число выполняемых итераций. Метод АВВ является эвристической релаксацией предлагаемой вероятностной модели, существенным недостатком которой является разрывность функции потерь. Из-за этого метод АВВ не может использоваться, например, совместно с трюком репараметризации, требующим гладкой функции потерь.

В начале главы разрабатывается математический аппарат для стохастического MAP-вывода в дискриминативных вероятностных моделях. Сначала излагается метод вариационной оптимизации [27; 28] для максимизации функции $f(z)$ дискретной или непрерывной переменной. Он

основан на вариационной оценке

$$L(\phi) = \mathop{\mathbb{E}}_{q(z|\phi)} f(z) \leq \mathop{\mathbb{E}}_{q(z|\phi)} \max_z f(z) = \max_z f(z), \tag{1}$$

верной для любого вспомогательного распределения $q(z|\phi)$. Неравенство переходит в равенство, когда вспомогательное распределение есть дельта-функция в аргмаксимуме $f(z)$. Пусть величина $L(\phi)$ может быть подсчитана с приемлемой вычислительной стоимостью. Тогда предлагается максимизировать $L(\phi)$ при помощи градиентных методов оптимизации.

Для невычислимой аналитически или слишком вычислительно затратной функции $L(\phi)$ предлагается новый метод *стохастической вариационной оптимизации*. В случае репараметризуемого распределения $q(z|\phi)$ предлагается провести репараметризацию. Для дискретного распределения $q(z|\phi)$ предлагаются два варианта: использование метода REINFORCE, либо применение релаксации Гумбель-Софтмакс и обучение при помощи трюка репараметризации. В любом из этих случаев становится возможным подсчёт стохастического градиента целевой функции.

Рассмотрим дискриминативную вероятностную модель $p(y, z|x) = p(y|x, z)p(z)$, где x — объект, y — целевая метка, а z — латентная переменная. Здесь $p(y|x, z)$ — правдоподобие ответа при условии объекта и латентной переменной, которое может задаваться, к примеру, нейронной сетью. *Задача MAP-вывода* состоит в нахождении значения латентной переменной z^*, максимизирующего апостериорное распределение $p(z|x, y) = \frac{p(y,z|x)}{p(y|x)}$. Для решения этой задачи можно использовать вариационную оптимизацию со вспомогательным распределением $q(z|x, \phi)$, не зависящим от истинной метки, что позволяет использовать его на этапе тестирования.

Далее предлагается вероятностный метод для адаптивного времени вычислений. *Блок адаптивных вычислений* — это вычислительный модуль, который выбирает число итераций в зависимости от входа. Отдельными итерациями могут быть, например, слои нейронной сети. Предполагается, что выходы итераций блока имеют одинаковую размерность. В зависимости от конкретного вида латентных переменных блок может быть *дискретным*, *пороговым* или *релаксированным*. Виды блоков совместимы, то есть параметры модели, обученной с одним блоком могут быть протестированы с другим. После каждой итерации подсчитывается *вероятность остановки*, число из отрезка $[0; 1]$, которое является параметром латентной переменной. В дискретном блоке после каждой итерации

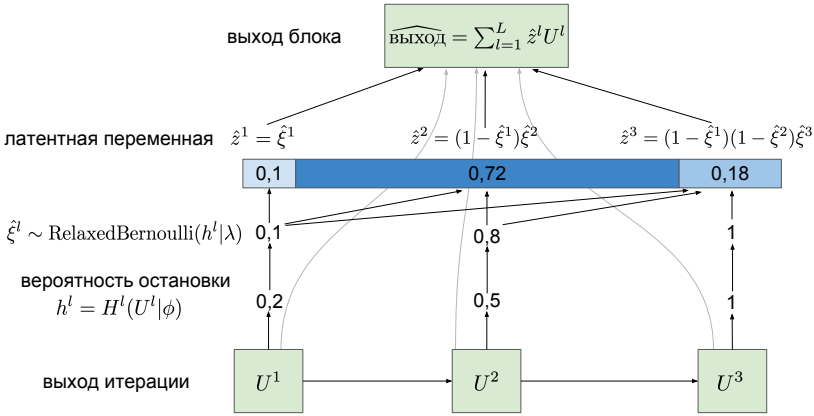


Рис. 5 — Релаксированный блок адаптивных вычислений.

генерируется переменная Бернулли, индикатор остановки, и в случае единичного исхода вычисления прерываются. В пороговом блоке остановка происходит, когда вероятность остановки превосходит $0{,}5$. Релаксированный блок получается из дискретного заменой распределения Бернулли на релаксированное распределение Гумбель-Софтмакс, рис. 5. В этом случае индикатор остановки принимает значения из отрезка $[0; 1]$. Выходом релаксированного блока является взвешенная комбинация выходов итераций, где веса получаются процессом ломания палки значений индикаторов остановки. Модель с релаксированным блоком вычислений допускает обучение стохастическим градиентным спуском при помощи трюка репараметризации.

Предположим, что в нейронной сети содержатся несколько блоков адаптивных вычислений, каждому из которых соответствует своя латентная переменная, задающая число итераций. Для каждой латентной переменной выбирается априорное усечённое геометрическое распределение (усечение проводится по максимальному числу итераций). Затем выполняется стохастический MAP-вывод для числа итераций, в котором вспомогательное распределение задаётся индикаторами остановки. Итоговый функционал качества имеет два слагаемых: среднее лог-правдоподобие правильного ответа по вспомогательному распределению и линейный штраф за ожидаемое число итераций. Этот функционал аналогичен получаемому в модели ABB, однако вместо эвристической стоимости вычислений используется непосредственно мат. ожидание числа итераций.

В конце раздела приводятся примеры применения предложенного метода к нейросетевым архитектурам. Для остаточных сетей предлага-

ется пространственно-адаптивная версия метода. Каждой пространственной позиции блока остаточной сети сопоставляется блок адаптивных вычислений. Итерации адаптивных вычислений соответствуют остаточным модулям. Получаемый метод является вероятностным аналогом ПАВВ. В случае рекуррентных сетей построение проводится аналогично методу АВВ [24]: блок адаптивных вычислений используется на каждом шаге по времени и выбирает число обновлений сети.

Экспериментальная валидация проводится на остаточных сетях ResNet-32 и ResNet-110 для задачи классификации CIFAR-10. Сначала показывается, что параметры релаксированной модели (использующей релаксированные адаптивные блоки вычислений) совместимы с дискретной и пороговой моделями. Для этого в процессе обучения релаксированной модели её параметры тестируются в дискретной и пороговой моделях. При этом целевая функция, точность и число операций отличаются незначительно. Затем обучение релаксированной модели сравнивается с обучением дискретной модели методом REINFORCE. Варьируется число латентных переменных, для чего пространственные позиции объединяются в группы, каждой из которых сопоставляется единственная переменная. Показано, что оба способа обучения показывают сравнимые результаты при числе латентных переменных менее ста, однако с ростом числа переменных метод REINFORCE не позволяет успешно обучить модель, что связано со слишком большой дисперсией градиентов. Обучение при помощи релаксации допускает использование вплоть до 1344 переменных. Релаксированная модель и метод ПАВВ обладают схожим соотношением числа операций и точности. Преимуществом вероятностного метода является возможность выполнения тестирования в пороговом режиме, имеющем крайне простую реализацию, без потери качества.

В **заключении** приведены основные результаты работы:

1. Разработан новый метод ускорения свёрточных нейронных сетей, основанный на перфорированном свёрточном слое, который позволяет пространственно варьировать объём вычислений. Показано, что перфорированный свёрточный слой может быть эффективно реализован как на CPU, так и на GPU. Предложено несколько видов масок перфорации, не зависящих от входного объекта и проведено их экспериментальное сравнение. При помощи разработанного метода достигнуто ускорение свёрточных нейронных сетей AlexNet и VGG-16 в несколько раз. Сокращение про-

странственной избыточности представлений свёрточной нейронной сети позволяет улучшить соотношение между скоростью и качеством работы.

2. Метод адаптивного времени вычислений, использованный ранее для рекуррентных нейронных сетей, применён к остаточным сетям. Полученный метод позволяет варьировать число слоёв в остаточных сетях в зависимости от входного объекта. Разработан метод пространственно-адаптивного времени вычислений, позволяющий выбирать различное число слоёв для пространственных позиций. Доказано, что этот метод является обобщением предыдущего. Для эффективной реализации метода используется перфорированный свёрточный слой, в котором маска перфорации зависит от объекта. Экспериментально показано преимущество пространственно-адаптивной версии метода для улучшения соотношения между скоростью и качеством работы остаточных сетей. Наилучшие результаты получены при обработке изображений высокого разрешения. Также показано, что карта стоимости вычислений может использоваться как модель человеческого визуального внимания.
3. Предложена вероятностная модель адаптивного времени вычислений, позволяющая адаптировать число слоёв в моделях глубинного обучения, таких как свёрточные нейронные сети. Разработан метод обучения этой модели, основанный на стохастической вариационной оптимизации и релаксации дискретных переменных Гумбель-Софтмакс. Исходный метод адаптивного времени вычислений является эвристической релаксацией предложенной модели. Показано, что предлагаемый метод позволяет получить результаты, аналогичные методу адаптивного времени вычислений, однако имеет более простую реализацию. Тем самым доказана возможность использования вероятностных моделей для адаптации глубины свёрточных нейронных сетей.

Список литературы

1. *Bengio Y.*, *Courville A.*, *Vincent P.* Representation learning: A review and new perspectives // IEEE transactions on pattern analysis and machine intelligence. — 2013. — Vol. 35, no. 8. — P. 1798–1828.
2. *Lowe D. G.* Object recognition from local scale-invariant features // Conference on Computer Vision and Pattern Recognition. — 1999. — Vol. 2. — P. 1150–1157.
3. *Dalal N.*, *Triggs B.* Histograms of oriented gradients for human detection // Conference on Computer Vision and Pattern Recognition. — 2005. — Vol. 1. — P. 886–893.
4. *Murty K. S. R.*, *Yegnanarayana B.* Combining evidence from residual phase and MFCC features for speaker recognition // IEEE signal processing letters. — 2006. — Vol. 13, no. 1. — P. 52–55.
5. *Furui S.* 50 years of progress in speech and speaker recognition research // ECTI Transactions on Computer and Information Technology (ECTI-CIT). — 2005. — Vol. 1, no. 2. — P. 64–74.
6. ImageNet Large Scale Visual Recognition Challenge 2016 (ILSVRC2016) Results / http://image-net.org/challenges/LSVRC/2016/results. — 2016.
7. *LeCun Y.*, *Bengio Y.*, *Hinton G.* Deep learning // Nature. — 2015. — Vol. 521, no. 7553. — P. 436–444.
8. *LeCun Y.*, *Boser B.*, *Denker J. S.*, *Henderson D.*, *Howard R. E.*, *Hubbard W.*, *Jackel L. D.* Backpropagation applied to handwritten zip code recognition // Neural computation. — 1989. — Vol. 1, no. 4. — P. 541–551.
9. *Hochreiter S.*, *Schmidhuber J.* Long short-term memory // Neural computation. — 1997. — Vol. 9, no. 8. — P. 1735–1780.
10. *Shazeer N.*, *Mirhoseini A.*, *Maziarz K.*, *Davis A.*, *Le Q.*, *Hinton G.*, *Dean J.* Outrageously large neural networks: The sparsely-gated mixture-of-experts layer // International Conference on Learning Representations. — 2017.
11. *Krizhevsky A.*, *Sutskever I.*, *Hinton G. E.* Imagenet classification with deep convolutional neural networks // Advances in Neural Information Processing Systems. — 2012.
12. *He K.*, *Zhang X.*, *Ren S.*, *Sun J.* Deep Residual Learning for Image Recognition // Conference on Computer Vision and Pattern Recognition. — 2016.
13. *Yosinski J.*, *Clune J.*, *Nguyen A.*, *Fuchs T.*, *Lipson H.* Understanding neural networks through deep visualization // ICML Deep Learning Workshop. — 2015.
14. *Nguyen A.*, *Dosovitskiy A.*, *Yosinski J.*, *Brox T.*, *Clune J.* Synthesizing the preferred inputs for neurons in neural networks via deep generator networks // Advances in Neural Information Processing Systems. — 2016. — P. 3387–3395.

15. *Rensink R. A.* The dynamic representation of scenes // Visual cognition. — 2000. — Vol. 7, no. 1–3.

16. *Larochelle H.*, *Hinton G. E.* Learning to combine foveal glimpses with a third-order Boltzmann machine // Advances in Neural Information Processing Systems. — 2010.

17. *Mnih V.*, *Heess N.*, *Graves A.*, [et al.]. Recurrent models of visual attention // Advances in Neural Information Processing Systems. — 2014.

18. *Ba J.*, *Mnih V.*, *Kavukcuoglu K.* Multiple object recognition with visual attention // International Conference on Learning Representations. — 2015.

19. *Jaderberg M.*, *Simonyan K.*, *Zisserman A.*, *Kavukcuoglu K.* Spatial transformer networks // Advances in Neural Information Processing Systems. — 2015.

20. *Xu K.*, *Ba J.*, *Kiros R.*, *Cho K.*, *Courville A.*, *Salakhutdinov R.*, *Zemel R. S.*, *Bengio Y.* Show, attend and tell: Neural image caption generation with visual attention // International Conference on Machine Learning. — 2015.

21. *Sharma S.*, *Kiros R.*, *Salakhutdinov R.* Action Recognition using Visual Attention // International Conference on Learning Representations Workshop. — 2016.

22. *Bengio E.*, *Bacon P.-L.*, *Pineau J.*, *Precup D.* Conditional Computation in Neural Networks for faster models // International Conference on Learning Representations Workshop. — 2016.

23. *Williams R. J.* Simple statistical gradient-following algorithms for connectionist reinforcement learning // Machine learning. — 1992.

24. *Graves A.* Adaptive Computation Time for Recurrent Neural Networks // arXiv. — 2016.

25. *Graham B.* Fractional Max-Pooling // arXiv. — 2014.

26. *Ren S.*, *He K.*, *Girshick R.*, *Sun J.* Faster R-CNN: Towards real-time object detection with region proposal networks // Advances in Neural Information Processing Systems. — 2015.

27. *Staines J.*, *Barber D.* Variational Optimization // arXiv. — 2012.

28. *Staines J.*, *Barber D.* Optimization by Variational Bounding // ESANN. — 2013.