



ОПТИМИЗАЦИЯ НЕЙРОННЫХ СЕТЕЙ ДЛЯ РАСПОЗНАВАНИЯ ОБРАЗОВ: МЕТОДЫ И ПОДХОДЫ

Смирнов Алексей Владимирович

старший преподаватель, Белорусский государственный университет
г. Минск, Беларусь.

Аннотация

Распознавание образов является одной из ключевых задач в области искусственного интеллекта и машинного обучения. Глубокие нейронные сети достигли выдающихся результатов в этой сфере, однако их высокая вычислительная сложность и большой объем памяти, необходимые для обучения и работы, остаются серьезным препятствием. Данная статья посвящена анализу современных методов и подходов к оптимизации нейронных сетей для распознавания образов, включая **квантование весов, прунинг и дистилляцию знаний**. В работе рассматриваются теоретические основы каждого метода, их практическое применение и влияние на точность и производительность модели. Особое внимание уделено тому, как эти методы позволяют создавать более легкие и быстрые модели, пригодные для развертывания на мобильных устройствах и встраиваемых системах с ограниченными ресурсами.

Ключевые слова: нейронные сети, распознавание образов, оптимизация, квантование, прунинг, дистилляция знаний, глубокое обучение, компьютерное зрение.

1. Введение: Проблема вычислительной эффективности в глубоком обучении

В последние годы глубокие нейронные сети (DNN) стали доминирующим инструментом в задачах распознавания образов, достигнув сверхчеловеческой точности в таких областях, как классификация изображений, обнаружение объектов и сегментация. Однако эти достижения сопровождаются существенным ростом вычислительных требований. Современные модели, такие как VGG, ResNet и Inception, содержат миллионы и миллиарды параметров, что делает их непригодными для использования на устройствах с ограниченными вычислительными мощностями, таких как смартфоны, дроны или IoT-устройства. Это создаёт острую необходимость в разработке методов оптимизации, которые могут сократить размер модели и ускорить её работу, не жертвуя при этом точностью.

2. Методы оптимизации нейронных сетей

Существует несколько основных подходов к оптимизации нейронных сетей, которые можно использовать как по отдельности, так и в комбинации.

2.1. Квантование весов

Квантование — это процесс уменьшения точности чисел, используемых для представления весов и активаций в нейронной сети. Вместо традиционного использования 32-битных чисел с плавающей запятой (FP32), веса и активации могут быть представлены 16-битными (FP16) или даже 8-битными (INT8) целыми числами.

Принципы работы: Квантование уменьшает объем памяти, необходимый для хранения модели, и позволяет использовать более быстрые целочисленные операции. Это приводит к значительному сокращению времени инференса (вычисления предсказаний).

Виды квантования:

Квантование после обучения (Post-Training Quantization): Наиболее простой метод, который не требует переобучения модели. Веса и активации квантуются после завершения обучения.

Квантование с учетом обучения (Quantization-Aware Training): Более сложный метод, при котором модель обучается с учетом квантования, что помогает минимизировать потерю точности.

2.2. Прунинг (обрезка) модели

Прунинг — это метод уменьшения размера нейронной сети путем удаления наименее важных связей, нейронов или целых фильтров. Цель прунинга — найти и удалить избыточные части модели.

Виды прунинга:

Неструктурированный прунинг: Удаляет отдельные веса в модели, которые имеют значения, близкие к нулю. Это приводит к очень разреженной матрице весов, но требует специализированного программного обеспечения для ускорения.

Структурированный прунинг: Удаляет целые нейроны или фильтры. Это более эффективно с точки зрения аппаратной оптимизации, так как не меняет структуру тензоров, что позволяет использовать стандартные библиотеки.

2.3. Дистилляция знаний (Knowledge Distillation)

Дистилляция знаний — это метод, при котором знания из большой, сложной, "учительской" модели передаются в меньшую, "ученическую" модель. Ученическая модель обучается не только на "жестких" метках (правильных ответах), но и на "мягких" ответах (распределении вероятностей классов), которые генерирует учитель.

Принципы работы: Большая модель-учитель способна улавливать более тонкие зависимости в данных. Передавая эти знания, мы можем обучить меньшую модель достигать почти такой же точности, как и большая, при значительно меньшем размере.

3. Практическое применение и влияние на производительность

Комбинирование этих методов позволяет достичь наилучших результатов. Например, можно сначала обучить большую модель, затем провести дистилляцию знаний в меньшую модель, а после этого выполнить квантование и прунинг.

Сравнение методов:

Квантование: Даёт наибольший выигрыш в скорости и размере, но может привести к небольшой потере точности.

Прунинг: Значительно уменьшает размер модели, но эффект на скорость может быть менее заметным без специальной аппаратной поддержки.

Дистилляция: Помогает сохранить высокую точность при переходе к меньшей модели, делая последующие шаги оптимизации более эффективными.

Заключение

Оптимизация нейронных сетей является критически важной задачей для широкого внедрения глубокого обучения в реальные приложения. Методы **квантования, прунинга и дистилляции знаний** предоставляют мощный инструмент для создания компактных и быстрых моделей, способных работать на устройствах с ограниченными ресурсами. Это открывает новые возможности для развития технологий компьютерного зрения в сферах мобильных приложений, робототехники и интернета вещей.

Литература

1. Han, S., Mao, H., & Dally, W. J. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*, 2015.
2. Hinton, G., Vinyals, O., & Dean, J. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

3. Jacob, B. et al. Quantization and training of neural networks for efficient integer-arithmetic-only inference. *Proceedings of the IEEE conference on computer vision and pattern recognition*, **2018**.
4. Gholami, A. et al. A survey of quantization methods for efficient neural network inference. *arXiv preprint arXiv:2102.04616*, **2021**.
5. Wang, Z., & Chen, G. Recent advances in network pruning. *arXiv preprint arXiv:2006.14798*, **2020**.
6. Polino, A. et al. Model compression and acceleration for deep neural networks: A survey. *Neurocomputing*, **2020**, 409, 1-17.
7. Deng, J. et al. ImageNet: A large-scale hierarchical image database. *CVPR*, **2009**.
8. Sze, V. et al. Efficient processing of deep neural networks: A tutorial and survey. *Proceedings of the IEEE*, **2017**, 105(12), 2295-2329.
9. He, K. et al. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, **2016**.
10. Simonyan, K., & Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, **2014**.