Федеральное государственное автономное образовательное учреждение высшего образования «Московский Физико-Технический Институт (Национальный исследовательский университет)»

Факультет управления и прикладной математики
Кафедра машинного обучения и цифровой гуманитаристики

ДИПЛОМНАЯ РАБОТА

Исследовательский проект на тему

”Непрерывные представления для акселерации глубоких нейронных сетей”

Выполнил студент группы M05-216б, 2 курс,
Рябыкин Алексей Сергеевич

Научный руководитель:
кандидат технических наук Ирина Сергеевна Асеева

Москва 2024

РЕФЕРАТ

Дипломная работа содержит 11 страниц, N рисунок, N таблицу, N использованных источников.

СТРУКТУРНЫЙ ПРУНИНГ, НЕПРЕРЫВНОЕ ПРЕДСТАВЛЕНИЕ НЕЙРОННЫХ СЕТЕЙ, ВИЗУАЛЬНЫЕ ТРАНСФОРМЕРЫ, КЛАССИФИКАЦИЯ, ТЕОРЕМА НАЙКВИСТА-ШЕННОНА-КОТЕЛЬНИКОВА

Дипломная работа посвящена исследованию современных методов ускорения нейронных сетей, в частности, структурного прунинга. Исследована проблема оценки емкости отдельных модулей глубоких нейронных сетей с целью более эффективного и автоматизированного прунинга. Исследованы различные техники прунинга визуальных трансформеров. (КРУТО ЕСЛИ ЭТО ПОЛУЧИТСЯ ВСЁ)

Теоретическая часть работы содержит исследовательский обзор статей для задачи структурного прунинга, непрерывных представлений нейронных сетей и прунинга трансформеров в разрезе анализа размерности полносвязных слоев и голов.

В практической части содержатся результаты поставленных экспериментов, их анализ и постановка последующих исследований.

Contents

ВВЕДЕНИЕ

TBD

ОСНОВНАЯ ЧАСТЬ

1 ТЕОРЕТИЧЕСКАЯ ЧАСТЬ

1.1 Обзор литературы

1.1.1 Метод оценки важности атомарных единиц нейронных сетей

Рассмотреть статьи [4] [7] [1] [3] [2]. Основные проблемы таких подходов в равномерности прунинга, если он локальный, и отсутствии понимания емкости моделей, если прунинг глобальный. На практике чаще всего в первую очередь происходит профайлинг, чтобы наиболее эффективно ускорить нейронную сеть на конкретном девайсе (мобильные устройства: телефоны, часы и прочее, видеокарта). Понимая емкость каждого слоя совместно с временем его исполнения становится возможно построить наиболее эффективную (с точки зрения ускорения и сохранения качества) финальную модель. Так же проблемой является то, что все эти методы, даже второго порядка и с аналатическими оценками сохраненных параметров, требуют дообучения, чтобы вернуться к качеству оригинальной модели. Потерянная емкость способна приводить к деградациям, требующим более сложного процесса обучения, чем для модели с изначальной емкостью (для задач, связанных с генерацией изображений: denoising, super-resolution, тд, возможны смещения по цвету, регулярности, например, из-за потери емкости ConvTranspose рис. и другие).

1.1.2 Динамический ресемплинг в непрерывных представлениях нейронных сетей

Рассмотреть статьи [6] [5]. Непрерывное представление дает возможность сразу менять число атомарных единиц нейронных сетей (нейронов, фильтров, голов). Кроме этого, семплирование может быть рассмотрено как дискретизация сигнала, что связывает этот процесс с теоремой Найквиста-Шеннона-Котельникова. Эта связь может оказаться полезной для определения оптимальной частоты дискретизации как оценки емкости модели нейронной сети. Приведенные гипотезы опробованы на двух основных задачах: классификация изображений (модели: ResNet18, ViT, OmniVec, датасеты: ImageNet, CIFAR-10), суперрезолюция (модели: EDSR, SRCNN, HAT-L, датасеты: Set5, B100)

1.2 Особенности поставленных экспериментов

1.2.1 BatchNorm

1.2.2 SpectralNorm

1.2.3 Multi-head Attention

2 ПРАКТИЧЕСКАЯ ЧАСТЬ

2.1 Используемые инструменты

- Python 3
- PyTorch
- TorchVision
- Matplotlib
- hugging-face
- Plotly
- transformers
- TorchIntegral
- TorchPruning
- diffusers
- wandb

2.2 Экспериментальные результаты

2.3 Выводы

2.4 Дальнейшая работа

”

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Chen Y., Zhou Z., Yan J. Going Beyond Neural Network Feature Similarity: The Network Feature Complexity and Its Interpretation Using Category Theory. — 2023. — arXiv: 2310.06756 [cs.LG].
2. Dong X., Chen S., Pan S. J. Learning to Prune Deep Neural Networks via Layer-wise Optimal Brain Surgeon. — 2017. — arXiv: 1705.07565 [cs.NE].
3. Frantar E., Singh S. P., Alistarh D. Optimal Brain Compression: A Framework for Accurate Post-Training Quantization and Pruning. — 2023. — arXiv: 2208.11580 [cs.LG].
4. Hessian-Aware Pruning and Optimal Neural Implant / S. Yu [et al.]. — 2021. — arXiv: 2101.08940 [cs.CV].
5. Integral Neural Networks / K. Solodskikh [et al.] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). — 06/2023. — P. 16113–16122.
6. Neural Network Compression via Low Frequency Preference / C. Zhang [et al.] // Remote Sensing. — 2023. — Vol. 15, no. 12. — ISSN 2072-4292. — DOI: 10.3390/rs15123144. — URL: https://www.mdpi.com/2072-4292/15/12/3144.
7. Zhu M., Gupta S. To prune, or not to prune: exploring the efficacy of pruning for model compression. — 2017. — arXiv: 1710.01878 [stat.ML].

ПРИЛОЖЕНИЯ

ПРИЛОЖЕНИЕ 1

ПРИЛОЖЕНИЕ 2