Университет ИТМО

Лабораторная работа №2
«Оптимизация разработанной нейронной сети»

по дисциплине: Технологии нейросетевых вычислений

вариант: Классификатор снимков с пневмонией

Выполнил: Неграш Андрей, P34301

Преподаватель: Старобыховская Анастасия Александровна

Санкт-Петербург
2023

1. Цель

Оптимизировать (ускорить) реализованную модель в рамках ЛР1, применяя один из классических подходов с лекции.

2. Выполнение

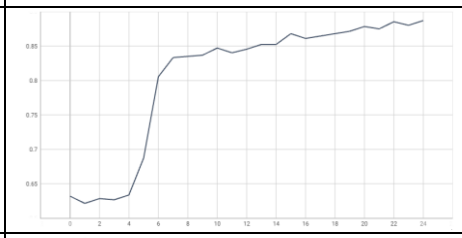
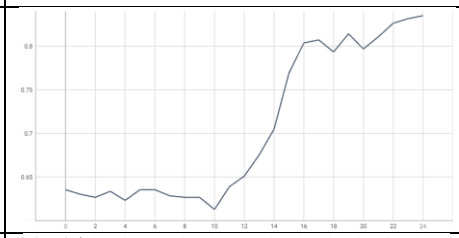
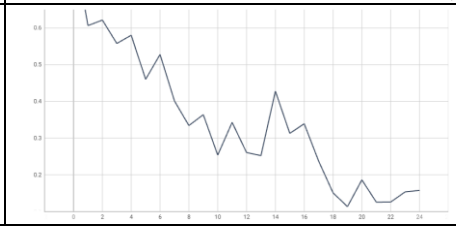
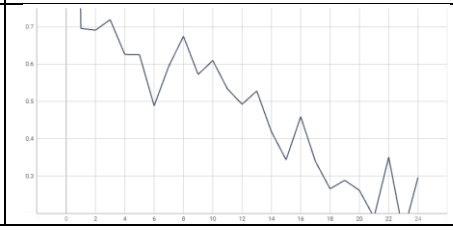
В качестве метода оптимизации был выбран итеративный прунинг (по l2-норме). Для его реализации в блокноте был написан следующий код:

```
for name, module in cnn.named_modules():
    if isinstance(module, torch.nn.Conv2d):
        prune.ln_structured(module, name='weight', n=2, amount=0.2, dim=1)
    elif isinstance(module, torch.nn.Linear):
        prune.ln_structured(module, name='weight', n=2, amount=0.4, dim=1)
```

Согласно нему из модели AlexNet, которую я использовал для обучения, в слоях *Conv2d* количество параметров будет сокращено на 20%, а в *Linear* слоях будет сокращаться 40% параметров.

3. Результаты работы

Для оценки итоговых результатов работы и сравнения характеристик моделей с прунингом и без него, заполним следующую таблицу:

Характеристика	**Модель**	
	Исходная	**Оптимизированная**
Точность модели на инференсе	0.887	0.837
Время работы модели на инференсе	7.463	6.246
Размер модели	233.081 МБ	466.123 МБ
Вес файла модели	233.1 МБ	466.1 МБ
График точности (accuracy)		
График функции потерь (loss)		

4. Вывод

Итак, в процессе данной лабораторной работы я провёл оптимизацию итеративным прунингом для разработанной ранее нейронной сети, классифицирующей рентгеновские снимки по признаку наличия на них пневмонии. Согласно полученным данным и проведённому сравнительному анализу относительно исходной модели можно сделать следующие выводы:

- Точность модели уменьшилась, но всё ещё на удовлетворительном уровне (более 80%)
- Время работы оптимизированной модели сократилось, что и требовалось
- Размер модели и, соответственно, файла её сохранения, увеличился в 2 раза, что, вероятно, вызвано усложнённой структурой модели после прунинга
- Графики точности и функции потерь ухудшились (стали гораздо менее плавными) видимо из-за достаточно большого количества не учитываемых параметров сети и, вероятно, недообучения за 25 эпох

Для сокращения общего веса модели, как мне кажется, стоило бы использовать комбинацию из прунинга и квантизации, однако это могло бы ещё больше повлиять на итоговую точность. В таком случае количество сокращаемых во время прунинга параметров стоило бы уменьшить.