Отчет по второму домашнему заданию по курсу “Глубинное обучение в обработке звука”

Выполнил: Михаил Малафеев

Что реализовал:

- dark knowledge distillation
- dynamic quantization
- streaming

Какие эксперименты провел (всё указано в сравнении с baseline model):

- dark knowledge distillation (MACs в 10 раз уменьшился, FLOPs в 12 раз)
- qint8 (сжатие в чуть более 3 раз по размеру модели)
- fp16 (по какой-то причине модель стала только больше)
- dark knowledge distillation + qint8 (использование памяти уменьшилось в 20 раз от бейзлайна, MACs аналогично dark knowledge distillation)
- dark knowledge distillation + fp16 (fp16 не дал профита, то есть всё так же, как и для dark knowledge distillation)

Все эксперименты, логи и соответствующие id весов на гугл диске для каждого эксперимента можно получить из ноутбука.

При distillation получилась модель по качеству не хуже исходной, но за счет большего количества эпох (110 против 20) при обучении. Тем не менее, качество получилось не потерять. Квантизация до qint8 давала +-5% лосса от исходной модели. Интересно, что для fp16 модель не уменьшила, а даже увеличила свой размер) Возможно, это связано с тем, что пришлось слишком много дополнительной информации в процессе квантизирования, но это все равно странно и неочевидно. Совместно дистилляция и квантизация дали компрессию в 20 раз и в 10 раз ускорили модель. Кроме того, был реализован Streaming, отображение работы которого имеется в ноутбуке с экспериментами. На нем вероятность ключевого слова близка к 1 как раз к части аудиозаписи с ключевым словом.

Боже, благослови KWS! (в сравнении с первой дз)