KWS

Были проведены следующие эксперименты:

1. **Dark Knowledge Distillation**

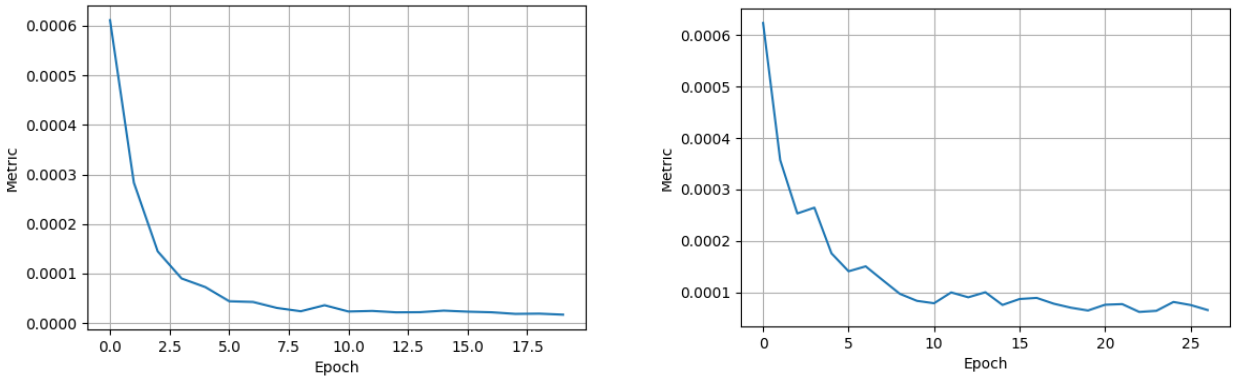
   Модель студента получается как сжатая модель учителя так, чтобы размер и время работы достаточно сократились. В данном случае число каналов свертки уменьшилось с 8 до 2, число слоев gru с 2 до 1 и размер скрытого слоя с 64 до 22.

   Модель обучается на 2 лосса - обычный и дистилляционный. Тут подбирается температура для софтмакса и соотношение, в котором лоссы влияют на итог. Оказалось хорошо брать большую температуру (10), соотношение в котором учитываются лоссы по-разному влияло на обучение, в последнем эксперименте лосс студента брался с весом 0.6.

   Также на обучение сильно влиял lr, его пришлось сильно увеличить.

   Результаты: Память сжалась в 11.7 раз, скорость - в 10.3 раза.

   Сначала модель била лосс 5e-5 * 1.1, но потом что-то поломалось (поменялся сид) и до нужного качества не доходила.



а) Учитель б) Студент

Графики обучения

2. **Dark Knowledge Distillation + small Attention**

   Здесь мы оставляем в attention только один линейный слой, размер

скрытого слоя - 24, остальные параметры такие же, как в первом эксперименте. У меня не получилось придумать, как адекватно сравнивать похожесть выходов attention разного размера, поэтому лосс оставила таким же. Лосс студента брался с коэффициентом 0.3, обычный - с 0.7.

Результаты: Память сжалась в 12.9 раз, скорость в 10.2.

Порог 5e-5 * 1.1 достигнут за 25 эпох.

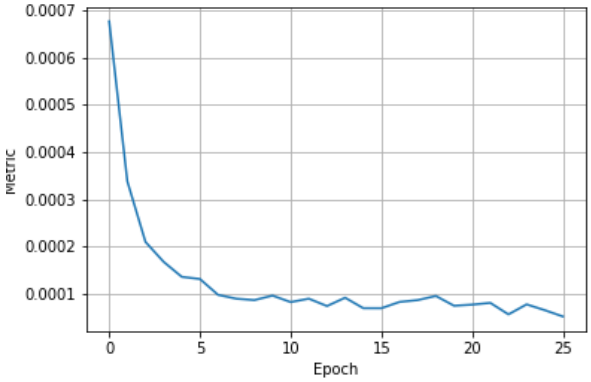


График обучения студента

3. **Dark Knowledge Distillation + fp16**

   Здесь берем такие же параметры, как в первом эксперименте, так как это практически единственный вариант, в котором macs уменьшается в нужное число раз. Обучаем модель с mixed precision.

   Результаты: Память сжалась в 19.2 раза, скорость в 10.3

   Порог 5e-5 * 1.1 достигнут за 19 эпох.

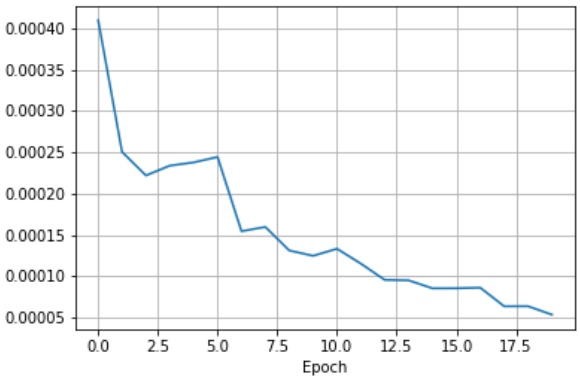
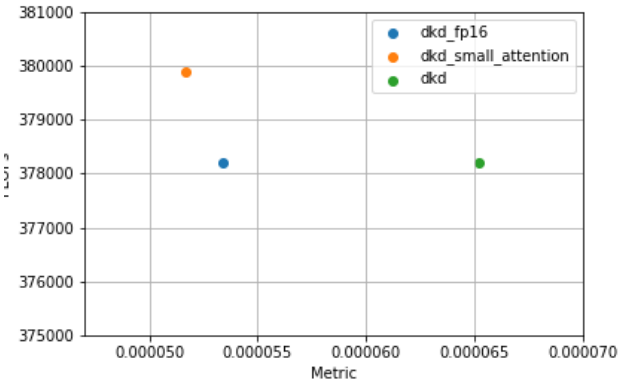


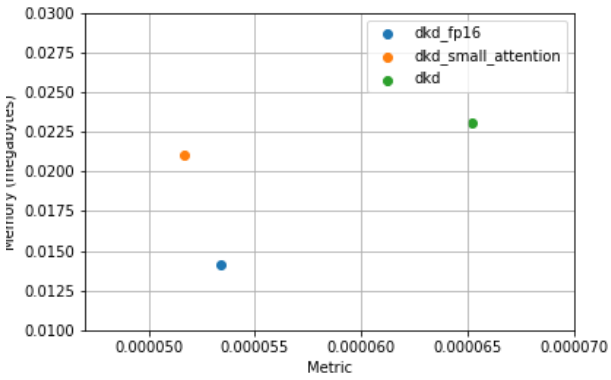
График обучения студента

Вывод: dark knowledge distillation может решать задачу сама по себе, далее уже можно изменять архитектуру и добавлять что-то для большего ускорения или уменьшения памяти. С fp16 обучение происходит сильно стабильнее и возможно достигнуть лучшего качества.

Еще графики:



а) metric-flops



б) metric-memory

4. Streaming

Мне не удалось установить ffmpeg на ubuntu так, чтобы запустился хотя бы исходный файл stream.py, поэтому на вход подается не звук с микрофона, а готовая запись.