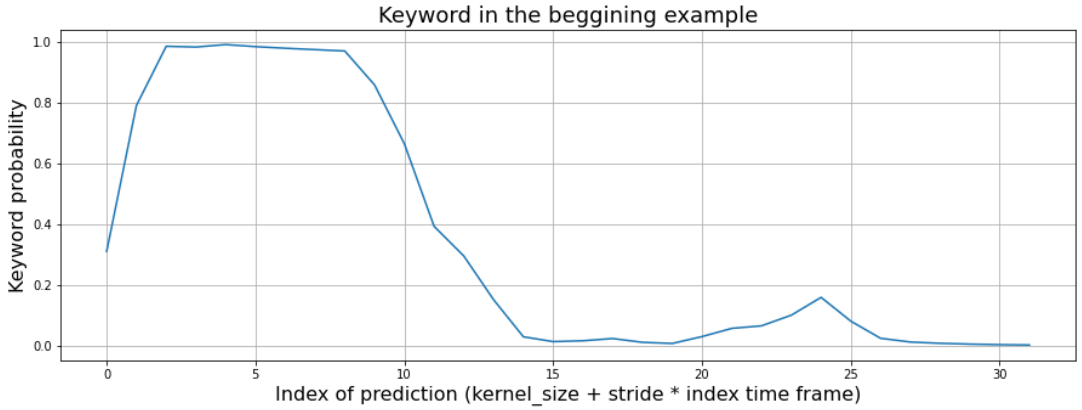
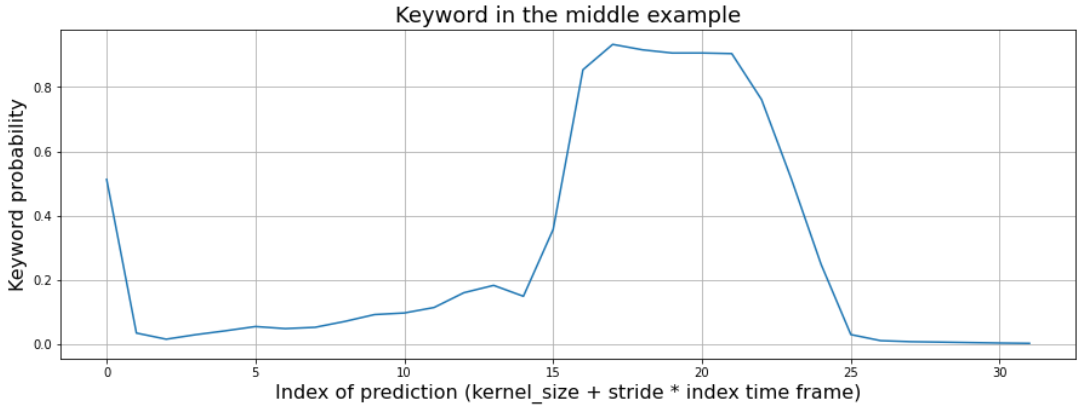
Отчет по KWS

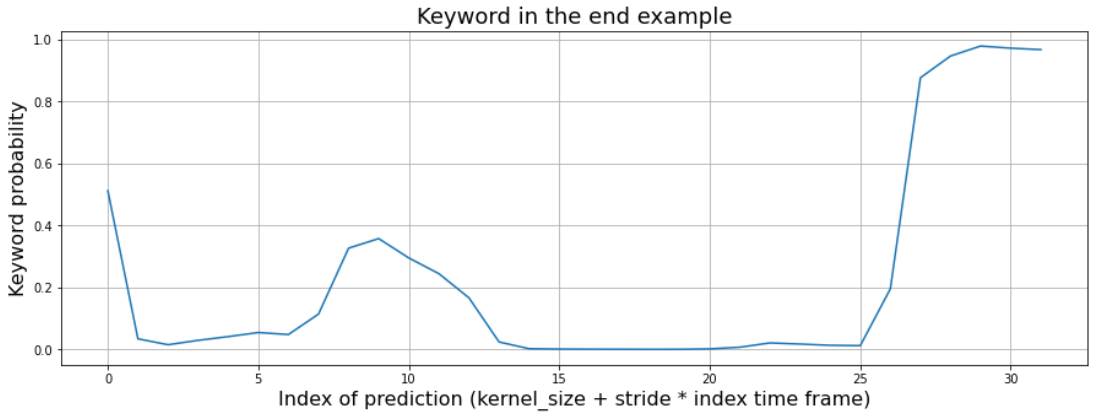
Болотин Арсений

Streaming.

Я реализовал подход, когда в сеть можно подавать аудио произвольными кусками. Проверил свою имплементацию следующим образом: взял 2 сэмпла без ключевого слова и 1 сэмпл с ключевым словом, затем я склеил их в разном порядке и подал в сеть. Построил графики вероятности ключевого слова в текущем окне в зависимости от времени и получил ожидаемые результаты.







Ускорение и сжатие.

Для начала я попробовал просто квантизацию и получил ускорение в 5.8 раз и сжатие в 3.1 раза. Далее я использовал дистилляцию. Подобрав параметры при ограничении с нужным качеством - получил уже почти необходимое - 10.48 и 8.98 соответственно. Далее я соединил два подхода и получил сжатие в 12.94 раза. Ускорение корректно непонятно как посчитать, но оно точно не хуже, чем было.

Также я попробовал сеттинг дистилляция + fp16 квантизация, но получил результаты хуже.