

# Отчет по KWS

## Барановская Дарья

(Ни на что не намекаю, но у нас в продакшене два человека месяц сжимали модель, уменьшили ее в три раза, и за это уже всех похвалили)

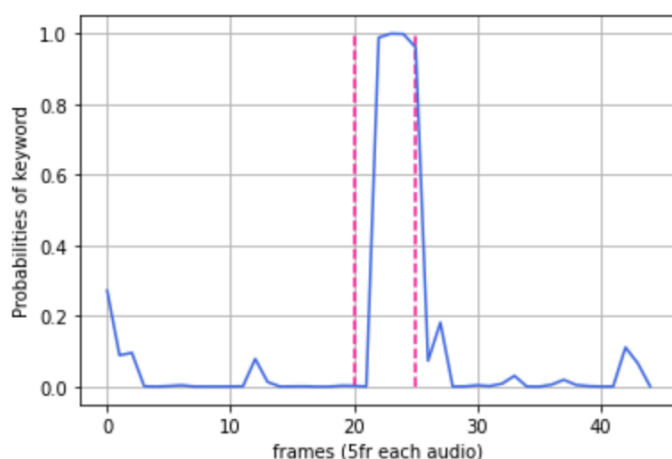
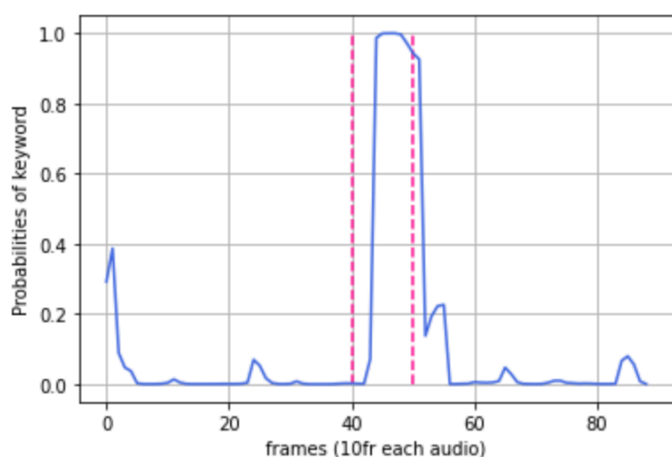
[https://wandb.ai/darya\\_baranovskaya/kws\\_hw?workspace=user-darya\\_baranovskaya](https://wandb.ai/darya_baranovskaya/kws_hw?workspace=user-darya_baranovskaya)

[https://drive.google.com/drive/folders/1gOD46jcxXMYI\\_hA8pabXdNN\\_PHuyIWfh?usp=sharing](https://drive.google.com/drive/folders/1gOD46jcxXMYI_hA8pabXdNN_PHuyIWfh?usp=sharing)

## Стриминг

Для стриминга написана отдельная модель, наследующаяся от CRNN, так что все натренированные модели могут быть использованы со стримингом. Для тестирования стриминга надо включить флаг `model.streaming = True` в модели. Для проверки работы я собрала 9 случайных из валидационного сета, в 8 из которых нет ключевого слова, а в одном, которое стоит посередине - есть, следовательно при правильной работе график будет выглядеть как пик в середине. Для проверки работы можно использовать функцию `check_streaming`, где `step_size` - длина фрейма, входящего на каждом шаге в модель.

Ниже продублирую графики из ноутбука с максимальной длиной окна 30 и `step_size` 10 (меньше размера ядра) и 20 соответственно



## Сжатие модели

Для сжатия я использовала: дистилляцию, квантизацию и прунинг.

Параметры исходной модели:







### Дистилляция:

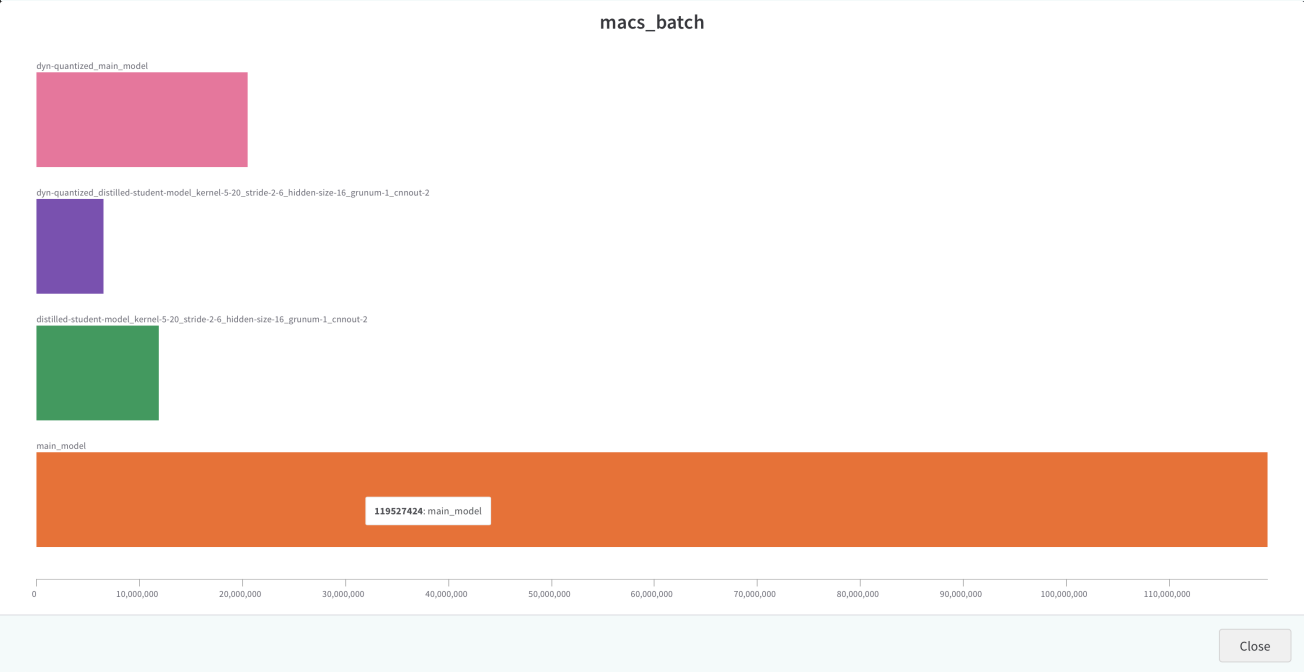
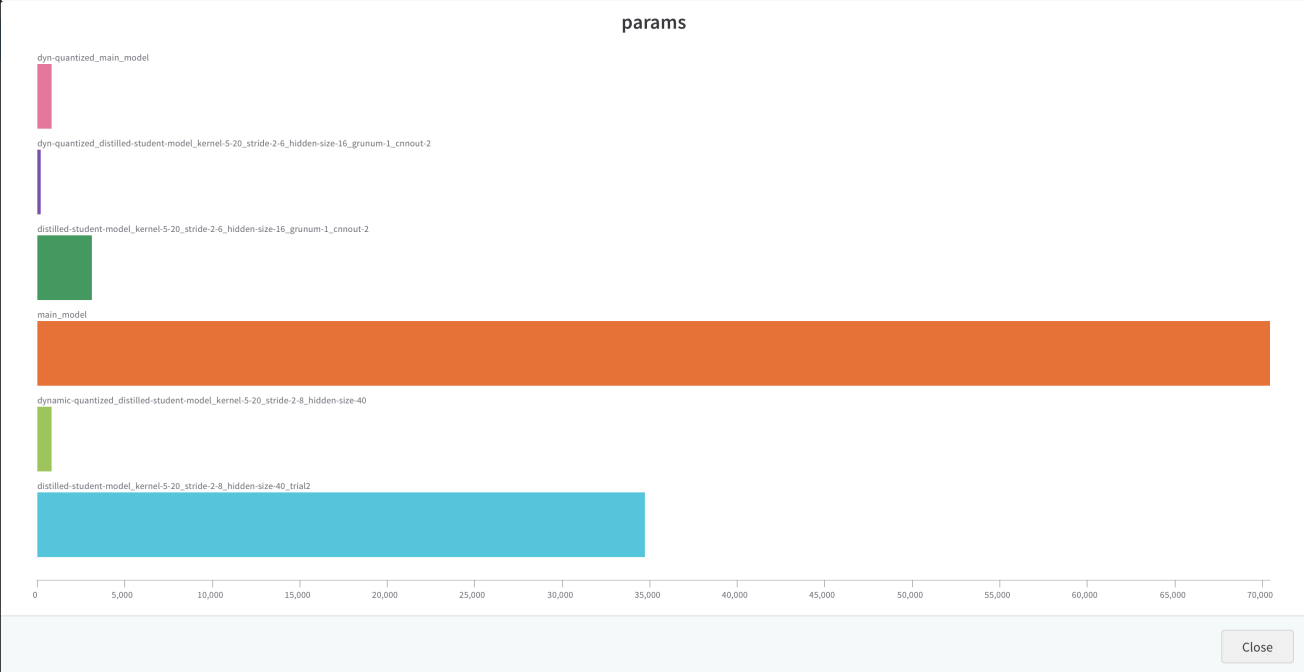
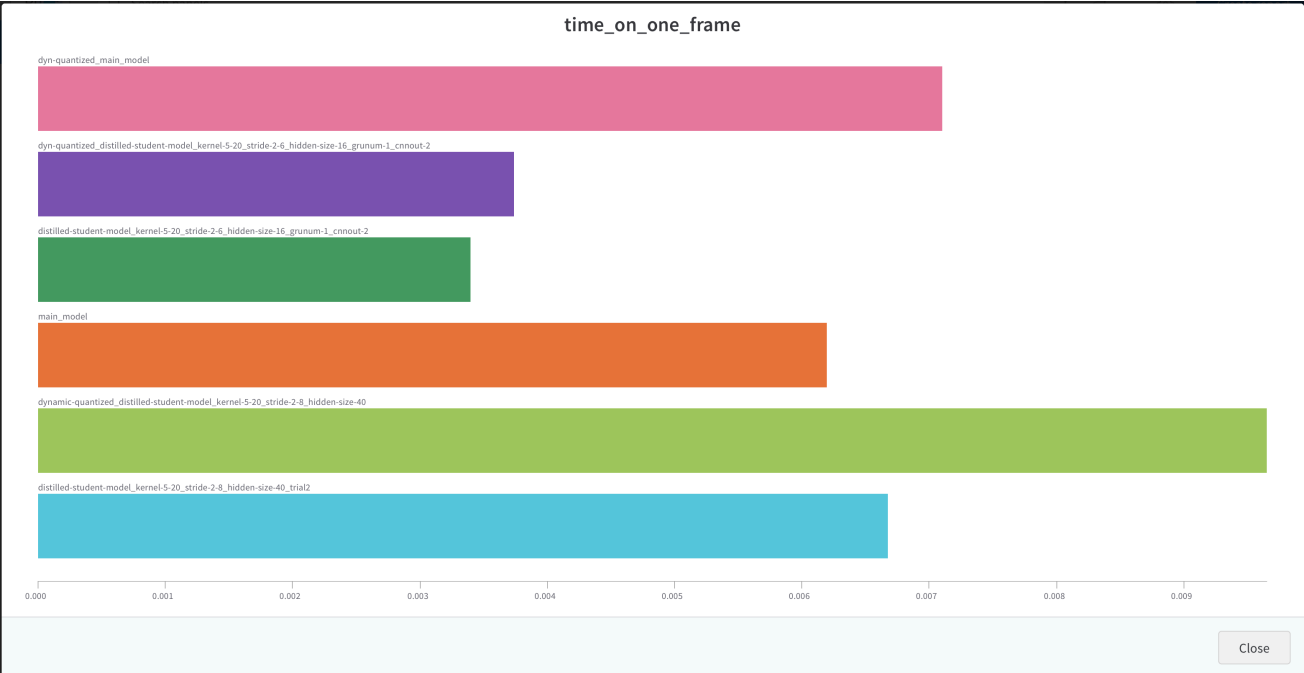
Попробовала множество различных конфигураций, вначале пробовала не слишком большое сжатие. Одной из лучших при небольшой дистилляция стала модель `distilled-student-model kernel-5-20 stride-2-8 hidden-size-40 trial2` ([https://wandb.ai/darya\\_baranovskaya/kws\\_hw/runs/3e82vogx](https://wandb.ai/darya_baranovskaya/kws_hw/runs/3e82vogx)). Название pth файла `distilled_model.pth`. Ее значение на валидационной метрике было даже лучше чем у основной модели, так что для следующих попыток дистилляции я использовала ее как учителя. Обучала 40 эпох

Максимальное сжатие в 10.08 раз по времени (`mac_batches`) (`119.527.424 / 11.852.800`) я смогла достичь при запуске модели `distilled-student-model kernel-5-20 stride-2-6 hidden-size-16 grunum-1 cnnout-2` ([https://wandb.ai/darya\\_baranovskaya/kws\\_hw/runs/10kkzyil](https://wandb.ai/darya_baranovskaya/kws_hw/runs/10kkzyil)). Название pth файла `distilled_model_smalest1.pth`. Обучала я ее 200 эпох с различными модификациями температуры, изменением scheduler, lr и тд

### Квантизация:

Логичнее всего применять квантизацию после всех других видов сжатия, ведь ни дистилляция, ни прунинг с интами не работают. Я квантовала практически все модели, которые дистиллировала а также основную модель. Стоит заметить, что квантизация дает сжатие примерно в 3.4 раза по размеру файла. Так квантизация основной модели сжала файл с 285КБ, до 83КВ. А самая маленькая модель, прошедшая дистилляцию и квантизацию имеет вес 47КВ.

Имя	Дата изменения	Размер
 <code>distilled_model_smalest1.pth</code>	Сегодня, 22:37	47 КБ
 <code>distilled_model.pth</code>	Сегодня, 22:39	142 КБ
 <code>dyn-quantized_distilled_model_smalest.pth</code>	Сегодня, 22:37	14 КБ
 <code>dyn-quantized_distilled_model.pth</code>	Сегодня, 22:39	56 КБ
 <code>dyn-quantized_main_model.pth</code>	Сегодня, 22:39	83 КБ
 <code>main_model.pth</code>	Сегодня, 22:38	285 КБ



## Прунинг:

Прунинг показал наихудшие результаты и в финальном сжатии я его даже не использовала. Я делала структурный прунинг, как для основной так и для дистиллированной модели, а потом дообучала модель, но при прунинге 0.3 параметров качество основной модели очень сильно падает (например при прунинге достаточно большой модели distilled model с 34731 параметрами, у которой при этом лучшее качество, качество ухудшается в 100 раз). Следовательно, по моему мнению, в данной задаче разумнее делать дистилляцию чем мучаться с прунингом.

Вывод: наилучшая модель с ускорением в 10.08 раз по времени (mac\_batches) (119.527.424 / 11.852.800 ) и 20.35 раз по памяти (285KB / 14KB) была получена двухэтапной дистилляцией + квантизацией в qint8.

Также были попробованы конфигурации: дистилляция, квантизация, прунинг, дистилляция + прунинг, дистилляция + квантизация, дистилляция + прунинг + квантизация

Далее будет небольшая табличка по самым основным моделям:

	MAC	Params	Фактический вес файла	Мое название модели в wandb	Название pth файла	Ссылка на wandb
Основная модель	11952742	70443	285KB	main_model	main_model	
Несильная дистилляция	50087936	34731	142KB	distilled-student-model_kernel-5-20_stride-2-8_hidden-size-40	distilled_model.pth	<a href="https://wandb.ai/darya_baran_ovskaya/kws_hw/runs/35rln5xp">https://wandb.ai/darya_baran_ovskaya/kws_hw/runs/35rln5xp</a>
Самая маленькая модель (только дистилляция)	11852800	3117	47KB	distilled-student-model_kernel-5-20_stride-2-6_hidden-size-16_gru_num-1_cnn_out-2	distilled_model_smallest1	<a href="https://wandb.ai/darya_baran_ovskaya/kws_hw/runs/10kkzyil">https://wandb.ai/darya_baran_ovskaya/kws_hw/runs/10kkzyil</a>

	MAC	Params	Фактический вес файла	Мое название модели в wandb	Название pth файла	Ссылка на wandb
Самая маленькая модель (дистилляция + квантизация)	11852800	3117	14KB	dyn-quantized_distilled-student-model_kernel-5-20_stride-2-6_hidden-size-16_groupnum-1_cnn_out-2	dyn-quantized_distilled_model_smallest.pth	<a href="https://wandb.ai/darya_baranovskaya/kws_hw/runs/19kcbxz8">https://wandb.ai/darya_baranovskaya/kws_hw/runs/19kcbxz8</a>

Отдельная глава, посвященная моей любви к



Был холодный осенний день, я как обычно проводила его за домашкой. Ближе к вечеру, датасфера начала неистово лагать и я оставила ее в покое на пару часов и пошла прогуляться. Однако, вернувшись домой, я с ужасом обнаружила, что моя сетка, которая должна была отучиться 80 эпох упала на 8-й из-за какого-то прикола датасферы с wandb. «Не беда» решила я, налила себе чайку и села перезапускать модельку. Ох, если бы я знала тогда, чем обернется история. К сожалению, каждый воскресный вечер с домашками провожу не только я, но и мои коллеги из Шада, которые забирают все гри себе. В итоге, простолюдинам как я не

остается ничего кроме как терпеть и ждать, пока какой-нибудь шадовец случайно упустит гпу, и она

Execute error: Servant gl.1 not allocated: Internal error

достанется кому-то другому. Так я ждала 40 минут. Дождалась, запустила, но сетка работать так и не начала из-за того, что датасфера часто виснет при num\_workers > 0. Пришлось останавливать зависнувшую тренировку и перезапускать data\_loaders. Потом снова возникла проблема с wandb, но уже другая, перезапуском клетки она не решилась, но я точно помнила, что она решается обновлением страницы и нажала

кнопку обновить в уголке браузера. Fatal Error. Следующие 30 минут я пытаюсь реанимировать файл, к которому датасфера обращается как к несуществующему, параллельно ноя в чате, где меня весьма смешно но чуть-чуть жестоко троллили. Скачать файл, восстановить чекпоинт датасферы, обновить страницу и тд - ничего не работает. Тимофей подсказал мне обратиться в чат поддержки, предупредив, что ответят мне только если утром. Однако, хотя бы здесь мне повезло - ответили за пару минут. Методы людей из поддержки вначале тоже были весьма суровыми и неэффективными. Но я молила о помощи и мне помогли. Я была вне себя от счастья. Ура! Теперь снова можно запускать модельку!

так, я спровоцировал закрытие проекта,  
попробуйте открыть его заново и проверим  
доступен ли файл

**Darya Baranovskaya**

✓ 00:30

Не доступен

Вы могли бы, пожалуйста, может быть  
достать мой файл с предыдущего  
чекпоинта или его сохранение полчаса  
назад, иначе я потеряю много часов  
работы?

✓ 00:31

попробуйте правой кнопкой по файлу и  
нажать Download

**Darya Baranovskaya**

Он скачивается, но при загрузке обратно в  
датасферу все та же проблема

файл битый и не открывается нигде

файла вроде частично восстановился, но  
возможно там пропали ячейки после того  
места, где он оборвался

посмотрите пожалуйста

00:45

возможно теперь сработает откат на  
чекпоинт

00:45

**Darya Baranovskaya**

✓ 00:46

Спасибо большое!!!!!!

Пару изменений исчезли, но это нестрашно  
Спасибо!

