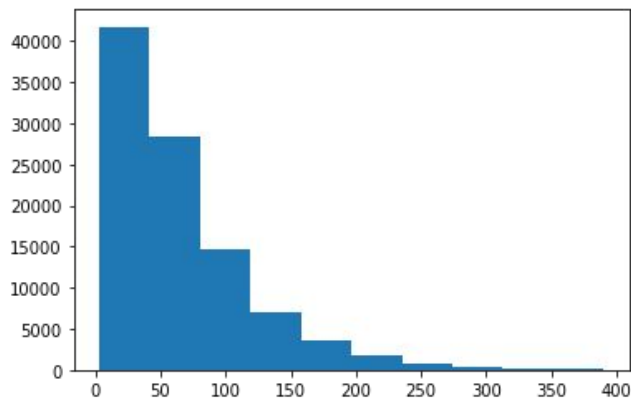


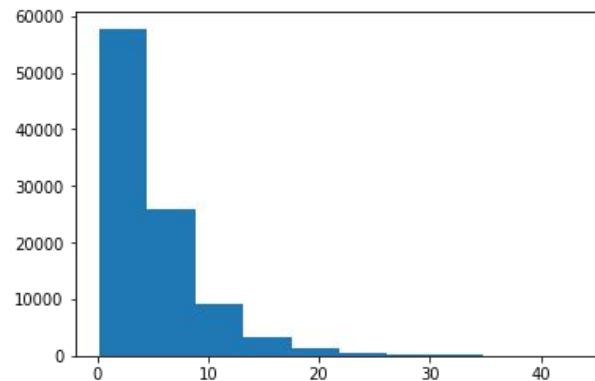
Решение задачи kaggle ASR

Анализ данных

- 1) Проблемы с разметкой - символы не из русского алфавита (1238 примера)
- 2) Домен данных - аудиокниги
- 3) На обучающей выборке и выборке для предсказаний данные по длине текста и речи распределены одинаково







Длина текста в target



Длина речевого сигнала в source

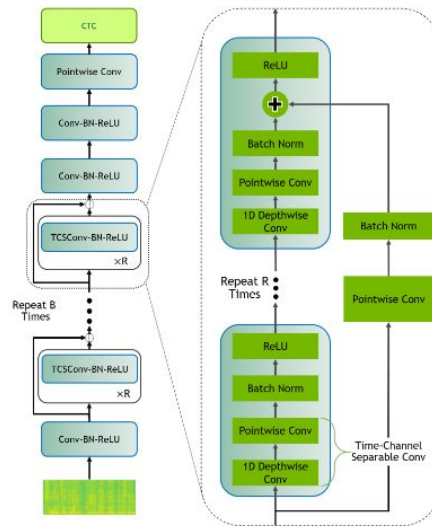
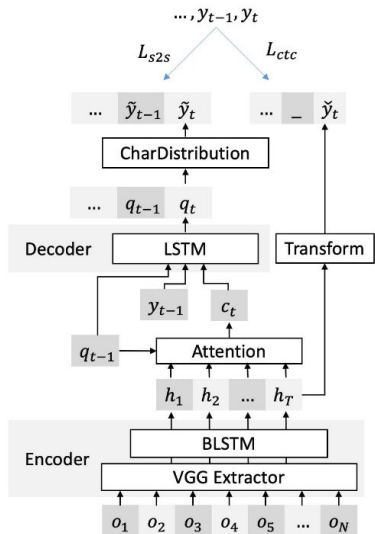
Выбор baseline

- 1) Мел спектрограмма, mfcc коэффициенты
- 2) Предсказание на уровне символов
- 3) Выбор легкой модели которую можно относительно быстро обучить с нуля
- 4) Анализ <https://paperswithcode.com/sota/speech-recognition-on-librispeech-test-clean>

Speech Recognition on Librispeech Test Clean						
29	QuartzNet15x5	2.69	×	QuartzNet: Deep Automatic Speech Recognition with 1D Time-Channel Separable Convolutions	 	2019
30	LAS (no LM)	2.7	×	SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition	 	2019

LAS и quartznet15x5

- 1) Анализ процесса обучения из оригинальных статей
- 2) Выбор базовых параметров оптимизатора и планировщика
- 3) Переобучение на одном батче
- 4) quartznet: Удалось выбить public cer 19+
- 5) Анализ оригинальной статьи: quartznet 5x5, novograd, CosineAnnealingLR



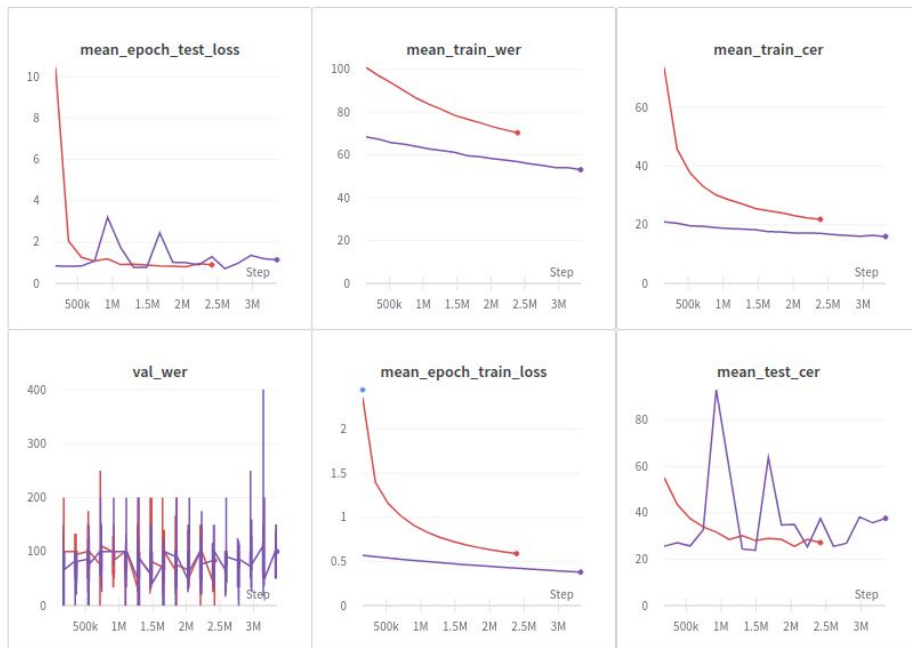
Логирование

[Mikhail Ivankin](#)



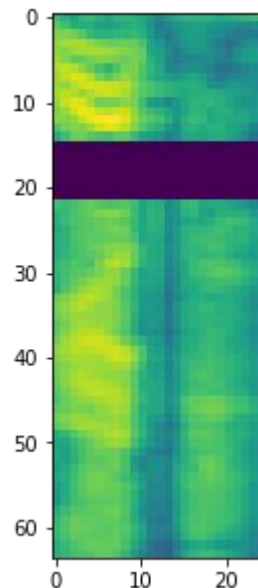
1) wandb

▼ Section 1



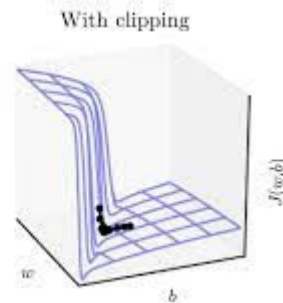
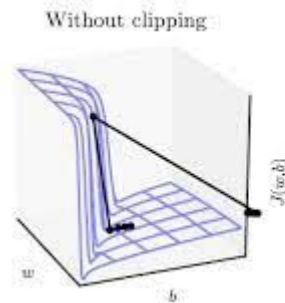
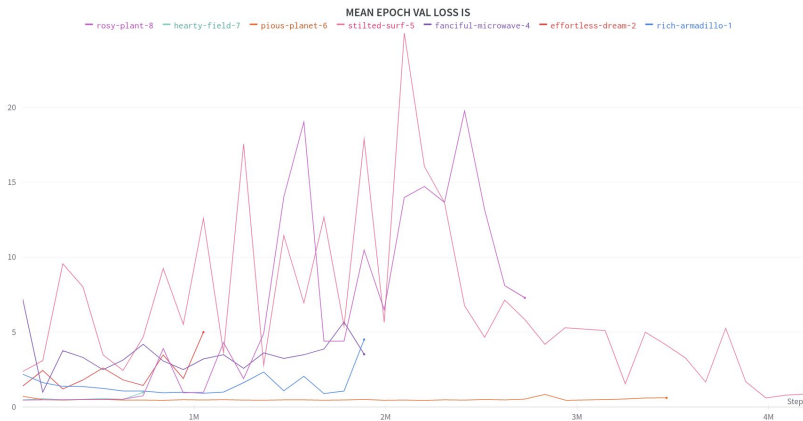
Аугментация данных

- 1) Первые послышки без аугментации данных
- 2) cutout
- 3) Подбор параметров: freq_mask=15, time_mask=30
- 4) Проблемы с time_mask
- 5) Аугментация во временной области: speed, noise, pitch, volume, time stretch



Клипирование

- 1) Большой перепад значений функции потерь при высокой скорости обучения
<https://forums.developer.nvidia.com/t/need-suggestions-on-the-gradient-explosion-of-nemos-quartznet/155894>
- 2) Для локальных минимумов перезапуск обучения с малым значением скорости обучения
- 3) ...
- 4) PROFIT!!! Но значения ser, wer в окрестности train
- 5) Клипирование градиента по норме
- 6) Подбор параметра: Дальнейшее обучение с клипированием градиента по норме по уровню 2



Finetuning модели

- 1) Языковая модель kenlm 3,4,6 n-gram на тексте датасета
- 2) Beam Search + lm, final rescoring, shallow fusion
- 3) Проблема с разметкой, анализ с помощью spell checker: редкие имена, отсутствующие слова в тексте или речи, слова разделенные на части пробелом
- 4) Фоновая музыка
- 5) Аугментация Background (librosa, openslr MUSAN)
- 6) Финальный пайплайн: x, speed, noise, background, volume + cutout(15, 30)

Валидация

- 1) Попытка понять переобучается модель на паблице или нет
- 2) Валидация на части sberbank golos dataset
- 3) Проблема с доменной областью

Что не успел

- 1) Voice activity detector
- 2) conformer, transformer
- 3) BPE
- 4) Нейросетевые языковые модели

Спасибо за внимание

[Mikhail Ivankin](#)



▼ Section 1

