Sirius Speech-to-Text

Реализация DNN модели для распознавания русской речи

Сириус, 25 марта 2021

Наша команда

Екатерина Чуйкова

Ментор

Максим Находнов

Полина Таранцова

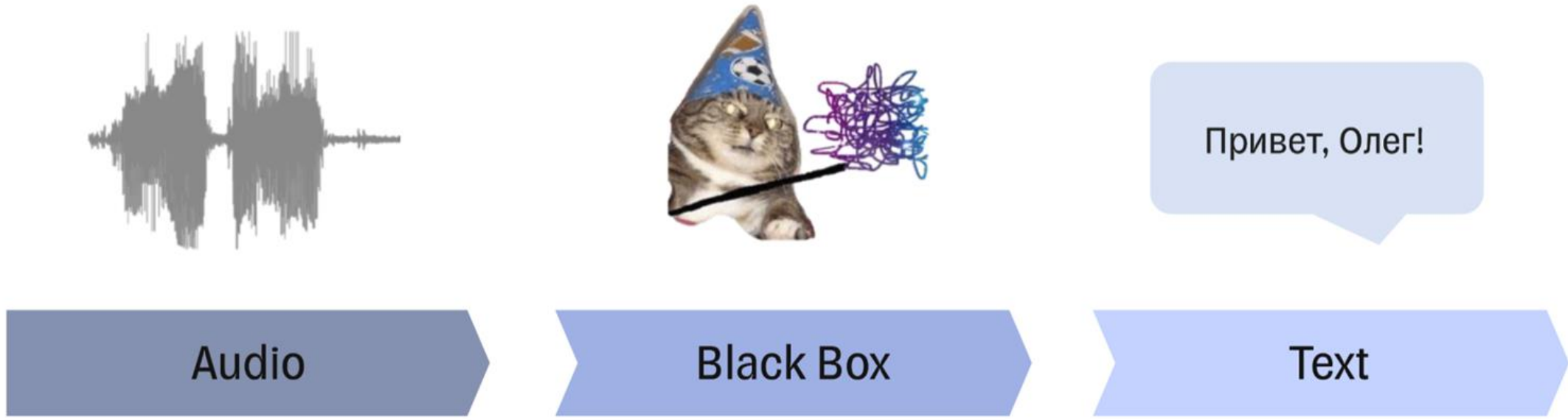
Оля Коломытцева

Саша Николаев

Задача

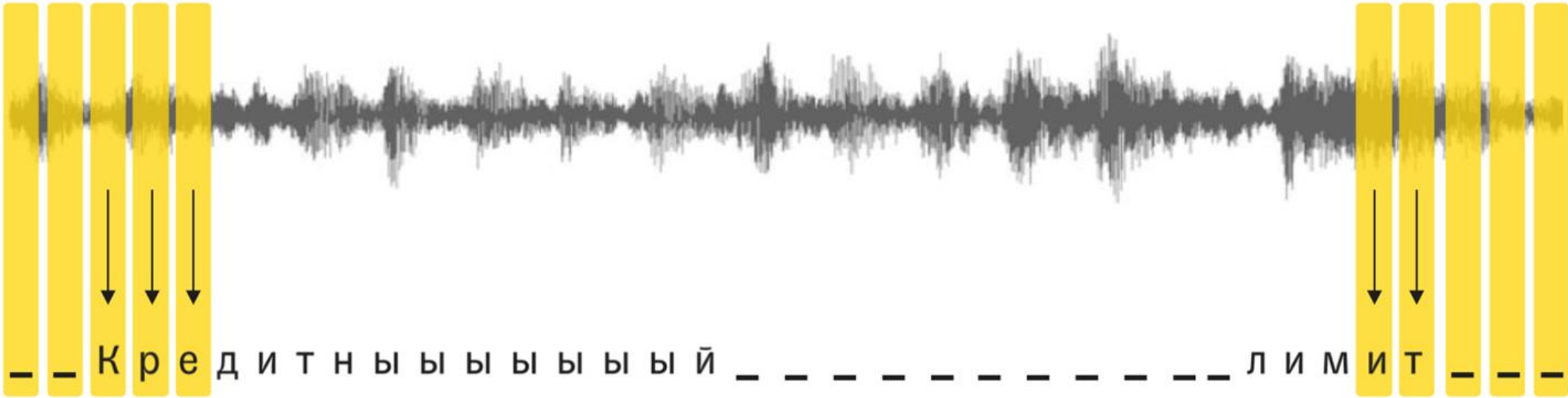
#stt #russian #speech #text #nlp #deeplearning #ai #state-of-the-art #commonvoice #openstt

Speech-to-Text распознавание русской речи



Тип задачи

#supervised_learning #end-to-end #sequence-to-sequence



Кредитный лимит

Метрика качества

#wer #word #error #rate #swap #delete #insert #nwords

Основная метрика speech2text - WER

$$WER = \frac{S + D + I}{N}$$

S –количество замен

D –количество удалений

I –количество вставок

N – количество слов

Привет олег закажи мне новую карту

_____ олег закажи не мне новую парту

$$WER = \frac{1 + 1 + 1}{6} = 0.5$$

Датасет

#data #datascience #speech&texts_makethebest

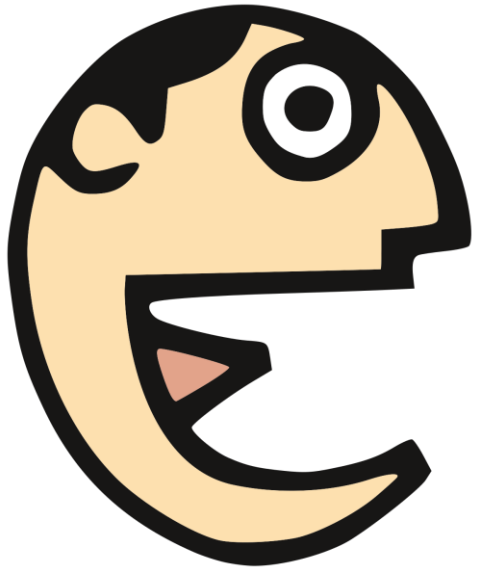
- Наборы данных:
  - Common Voice (4 GB)
  - Russian LibriSpeech (9 GB)
  - OpenSTT (40 GB radio + 40 GB audiobooks)

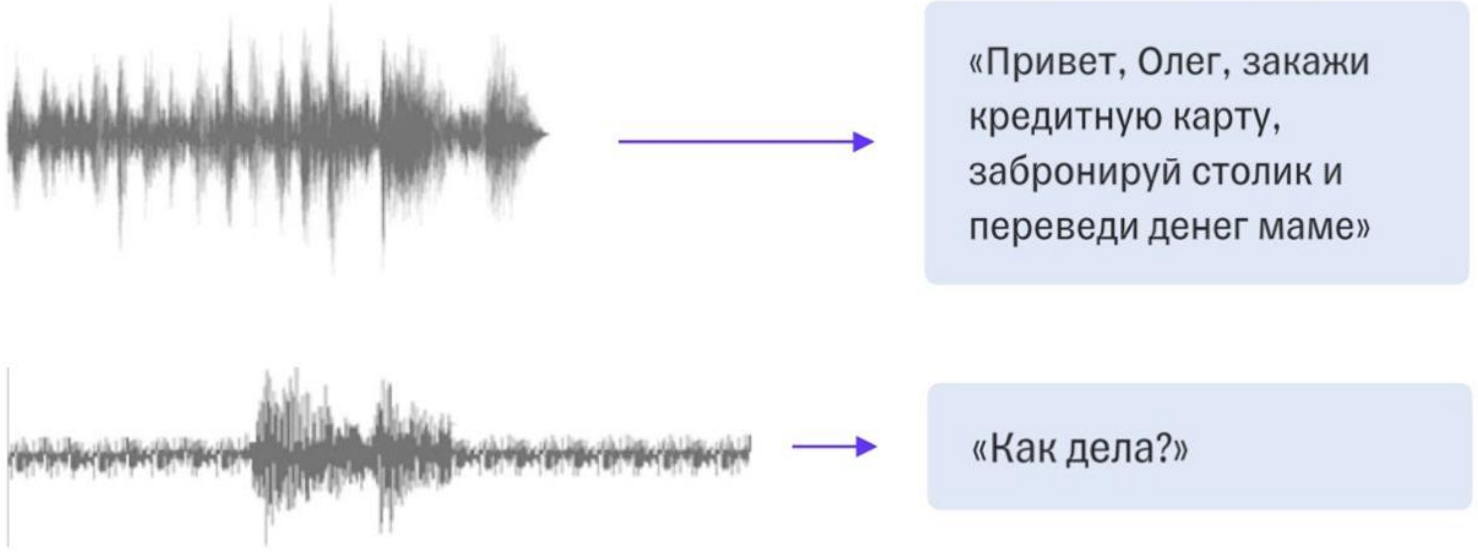
- Итоговая выборка:
  - Train: 1.3kk аудиозаписей длиной 1730 часов
  - Test: 23k аудиозаписей длиной 33 часа

Анализ и обработка исходных данных

#log #mel #spectro #sample_rate

- Единая частота звука - 8000 Hz
- Максимальная длина аудио - 10 s
- Модель обучается на мел-спектрограммах





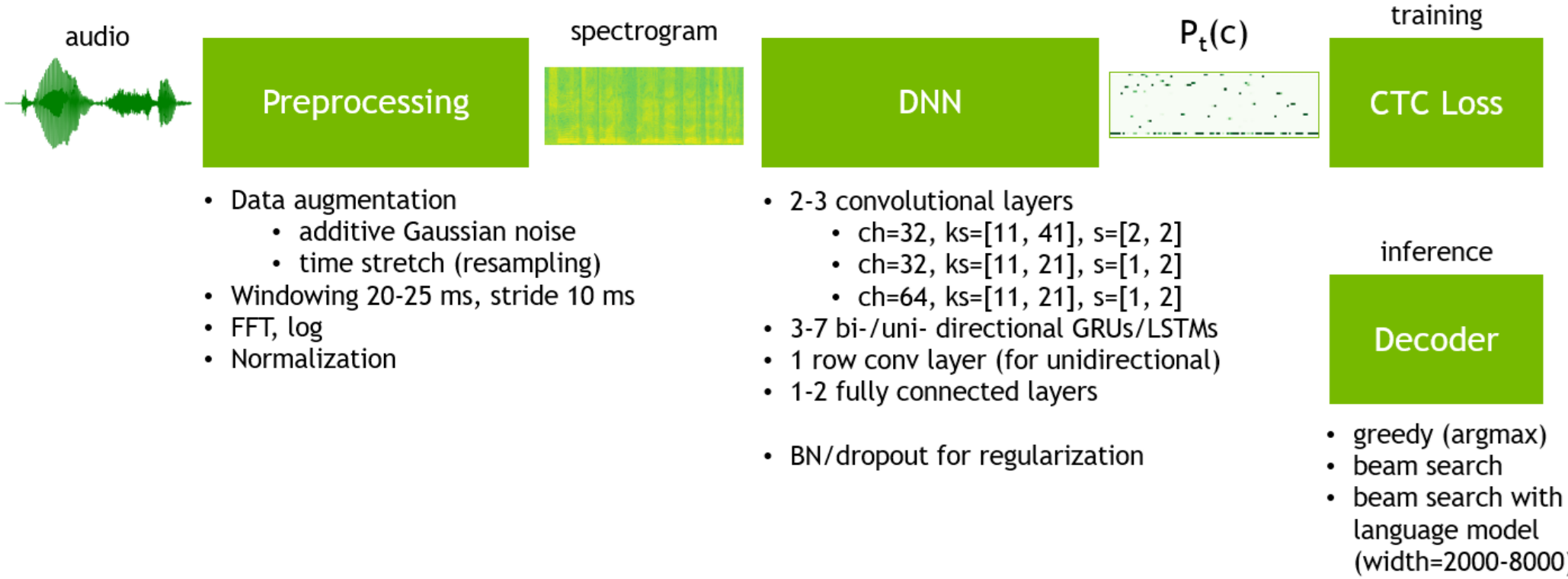
Модель - DeepSpeech2

#model #deepspeech2 #loss #train

~~Катя подготовила рыбу с такой архитектурой~~

✅ Оптимальна по времени обучения и качеству распознавания

✅ Достаточно проста в понимании и реализации

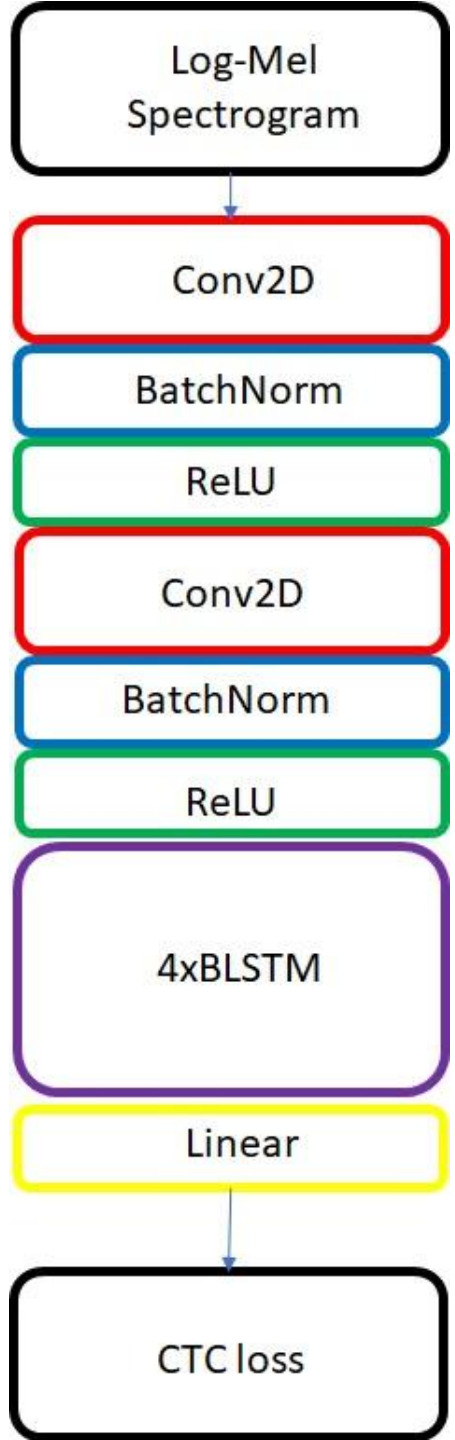


Модель - DeepSpeech2

#model #deepspeech2 #loss #train

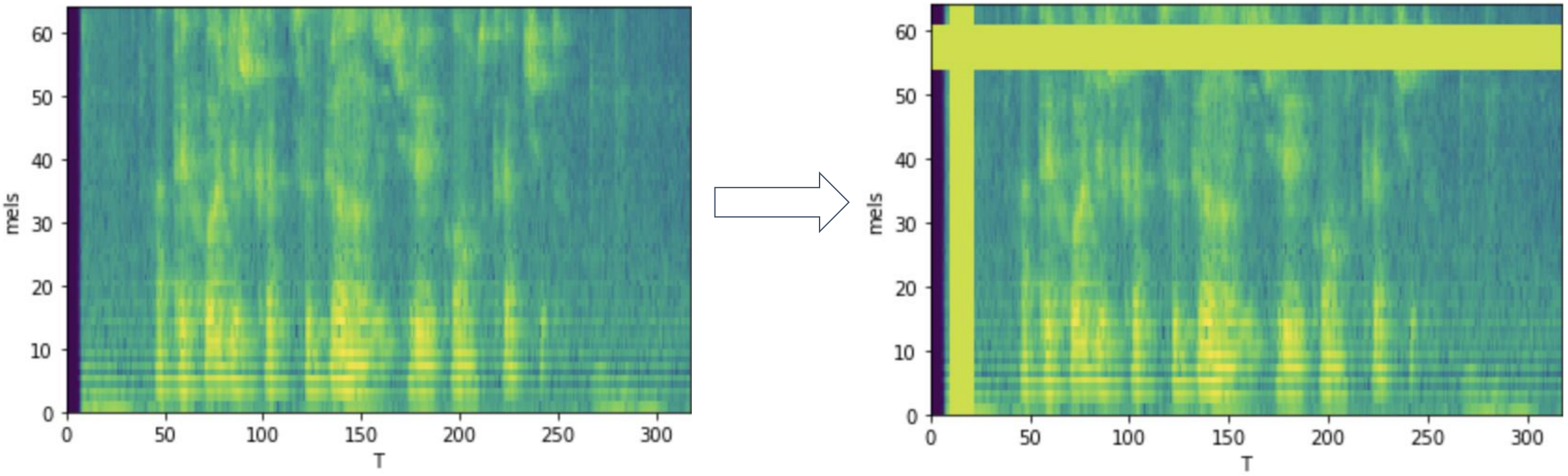
Параметры модели:

- 64 мел-фильтра
- 2 слоя сверток с BatchNorm
- Bidirectional LSTM с 4 слоями
- Размерность входа LSTM - 512



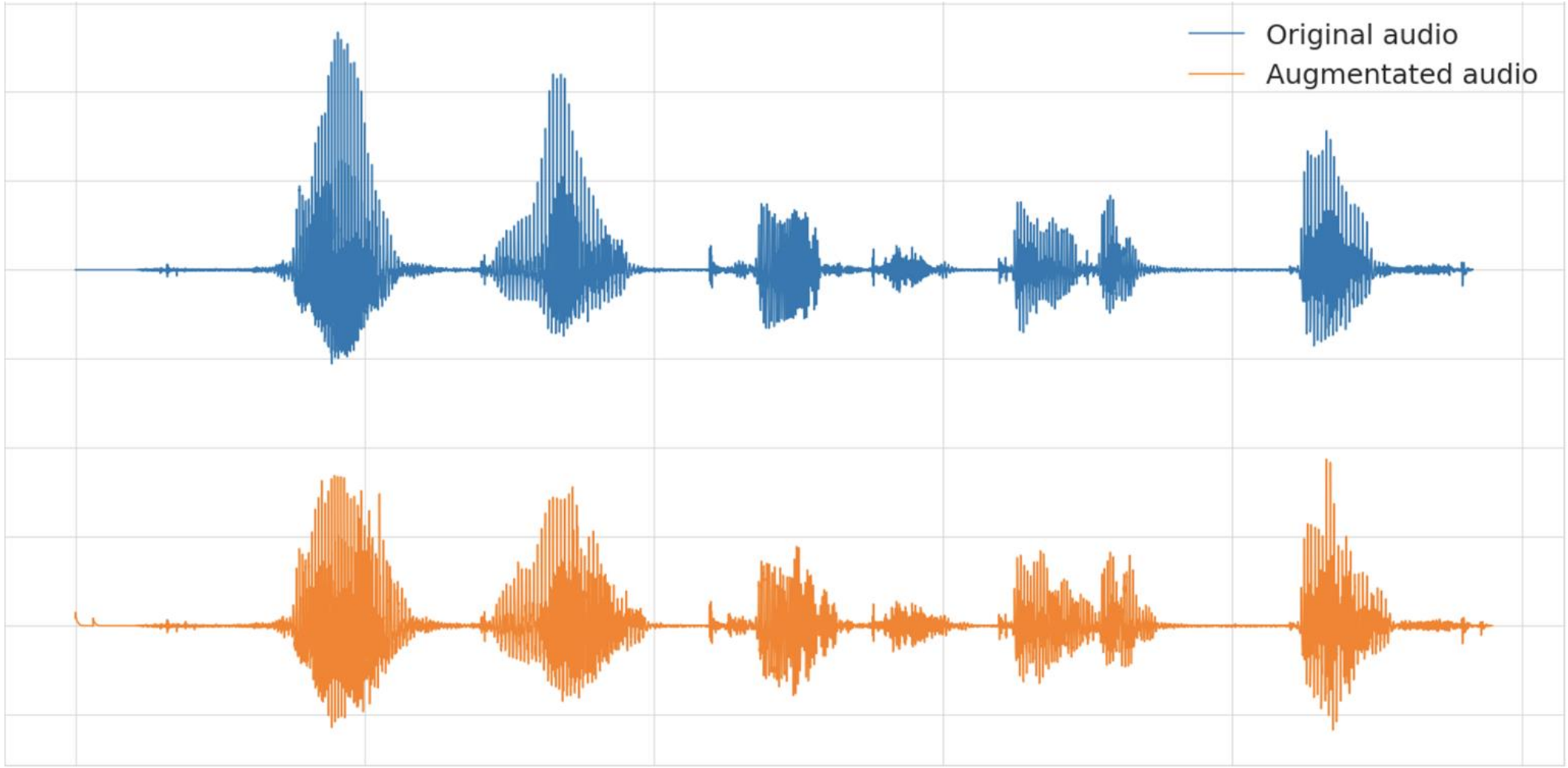
Spectrogram Augmentations

#model #augmentations #spectrogram



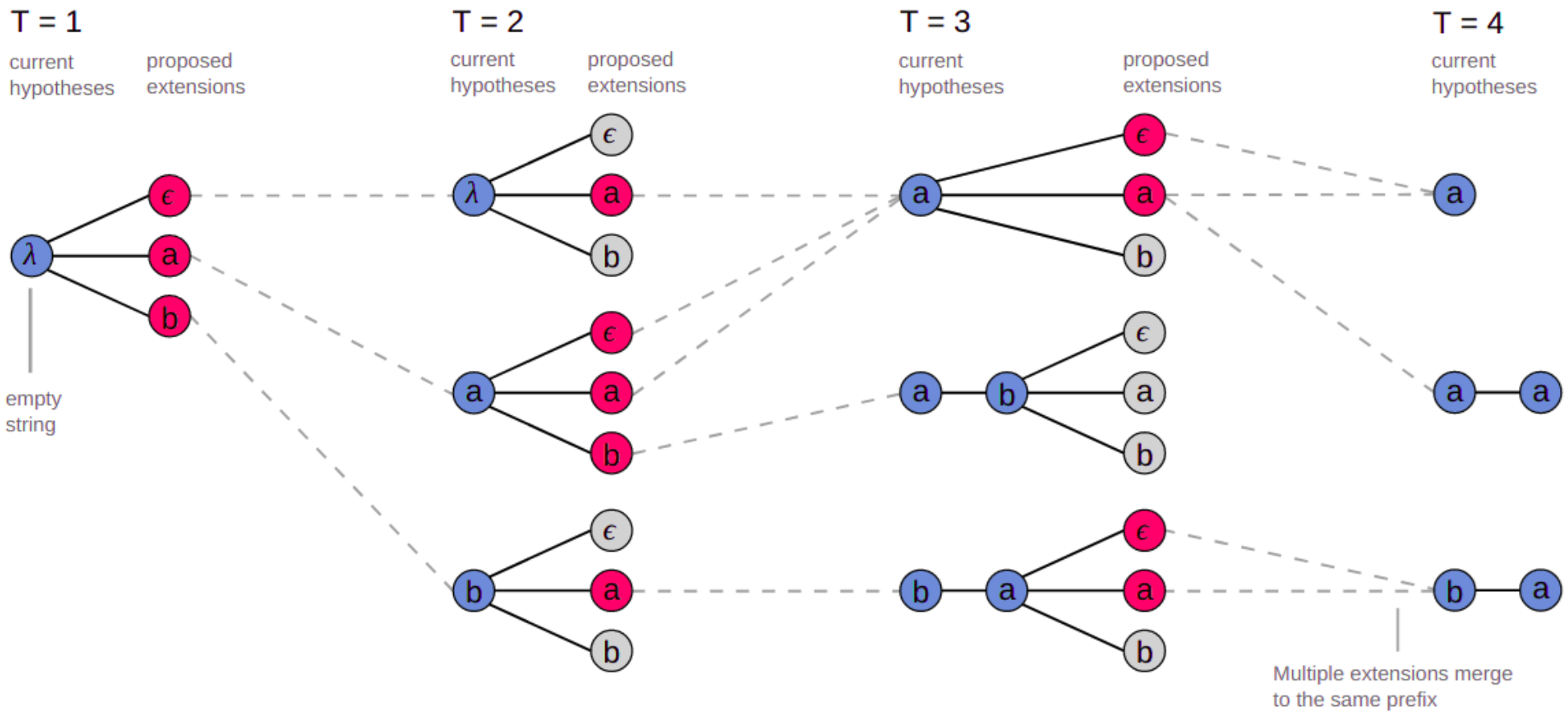
Audio Augmentations

#model #augmentations #audio



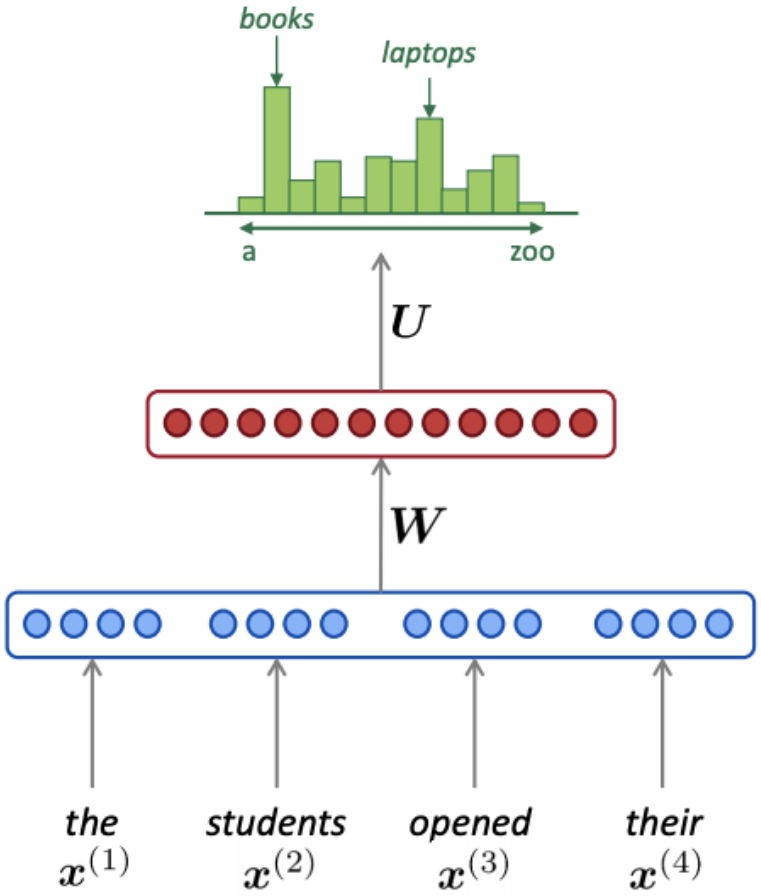
Модель: Beam Search

#model #beam_search



Модель: Beam Search + LM

#model #beam_search #LM



Example	Probability
The cat sat on the mat	0.95
The cat sad on the mat	0.20
High wind tonight	0.97
Large wind tonight	0.31

Стратегия обучения

- Learning rate = 2e-4
- Optimizer = Adam
- Scheduler = Exponential (decay=0.9)
- Аугментации спектрограм в частотной и временной областях
- Аугментации аудио:
    - нелинейный шум, временная задержка, dcshift

Inference

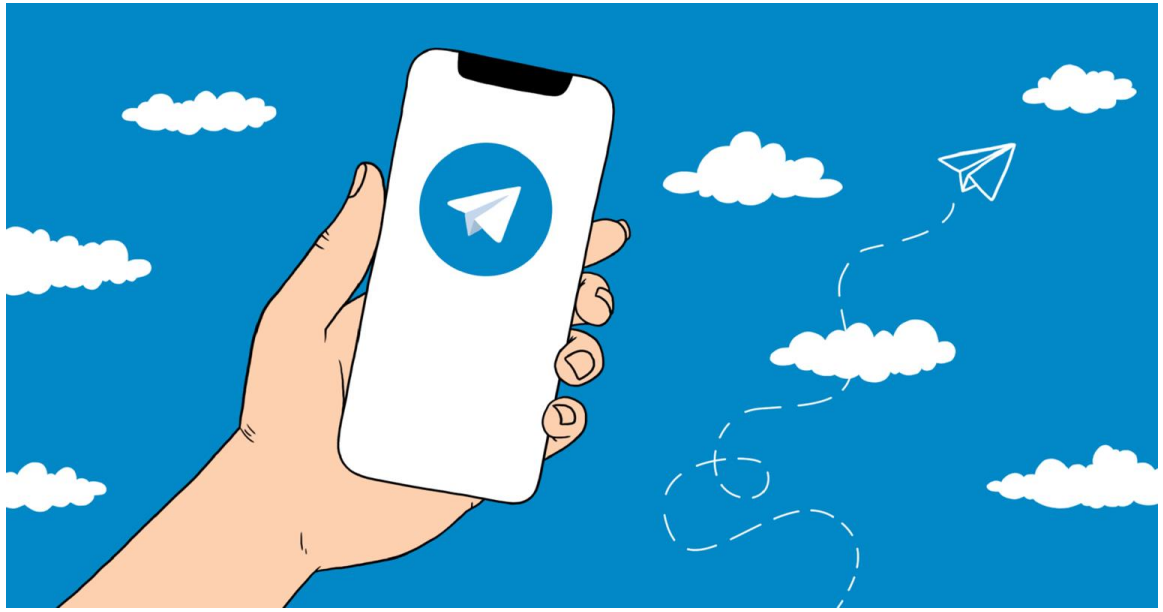
- BeamSearch
  - 200 наиболее вероятных гипотез
- KenLM
  - Статистическая n-gram модель
  - Обучена на Common Crawl (20 GB)
- Shallow fusion:
  - Топ-20 гипотез из BeamSearch ранжируются с помощью предобученной трансформер модели (Facebook-FAIR's WMT'19)

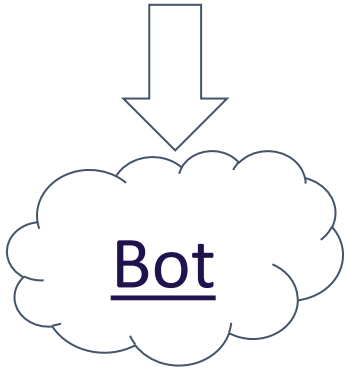
Результаты

	Common Voice	OpenSTT	LibriSpeech
Baseline	34%	—	—
Baseline + LM	**29%**	82%	80%
Our model	**29%**	**39%**	**69%**

MVP

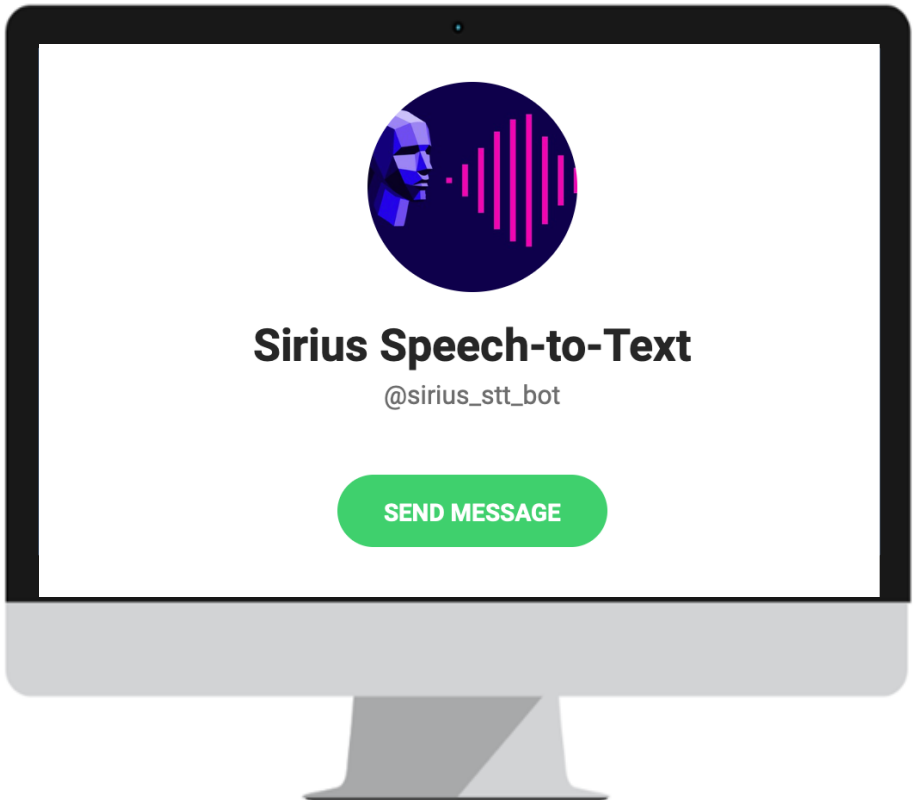
#telegram #bot #stt #can_test_on_my_phone

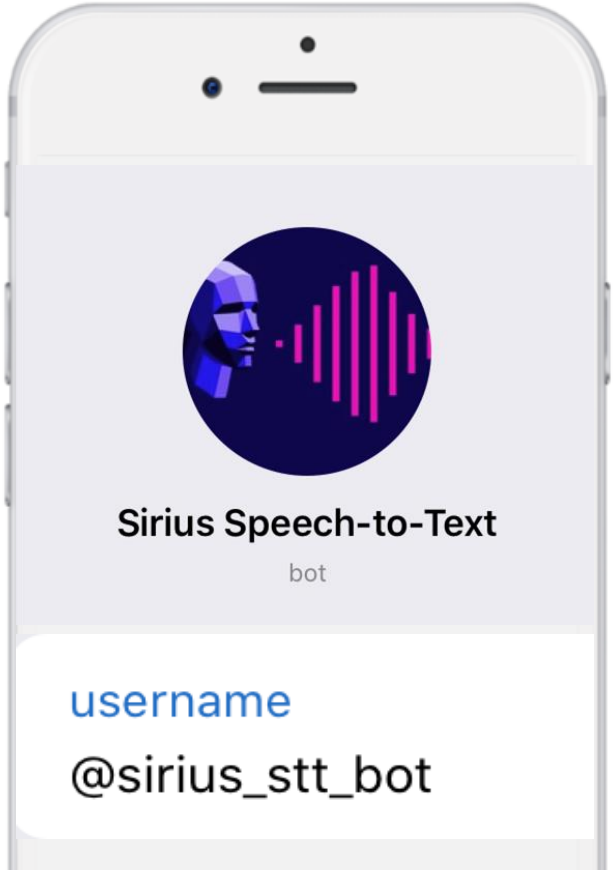




Демонстрация

#telegram #bot #stt #can_test_on_my_phone





Код

#code #ipynb #dataset #tensorboard #decode

Реализованы модули —> Model:

- **datasets** — реализация датасета аудиофайлов
- **audio_utils** — аугментация аудио и спектрограмм
- **decoding** — beam search и greedy decoding
- **deepspeech** — реализация модели
- **logging** — логгер процесса обучения
- **optimization** — train loop и метрики
- **inference** — api для инференса модели

Что дальше

#more_data #augmentations #model

Для улучшения качества модели можно:

- Добавить больше обучающих данных
- Лучше подобрать стратегию обучения
- Использовать больше аугментаций
- Усложнить архитектуру модели

Полезные продукты:

- Замена голосового сообщения на текстового
- Делать субтитры к видео
- Делать караоке
- Добавить модель перевода с русского языка на английский, чтобы общаться с иностранцами, не зная английский

Выводы

#conclusion #amazon #speech2text

- Кататься на амазоновских тачках - круто!
- Каждый член команды реализовал архитектуру Deepspeech
- Обучили классную модель на большом количестве данных и научились решать задачу speech2text, изучили основные подходы
- Научились работать с аудио: предобрабатывать, считать спектрограммы, аугментировать аудио
- Освоили работу с докером
- Провели эксперименты с аугментацией, языковой моделью, beamsearch, разными датасетами

Обзор работ в этой области

#scientific_staff #feel_smart #i_read_books #scientist

- Deep Speech 2: End-to-End Speech Recognition in English and Mandarin [1]
- First-Pass Large Vocabulary Continuous Speech Recognition using Bi-Directional Recurrent DNNs [2]
- SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition [3]
- On Using Monolingual Corpora in Neural Machine Translation [4]
- Facebook FAIR's WMT19 News Translation Task Submission [5]
- LONG SHORT-TERM MEMORY [6]

Спасибо за внимание!

Вопросы?