Идентификация диктора по голосу

Ефимов Владислав
Имеев Мерген
Руденко Дмитрий

Постановка задачи

Построить модель, которая по звуковому сигналу позволит определить личность диктора из ограниченного пула.

Методы решения задачи

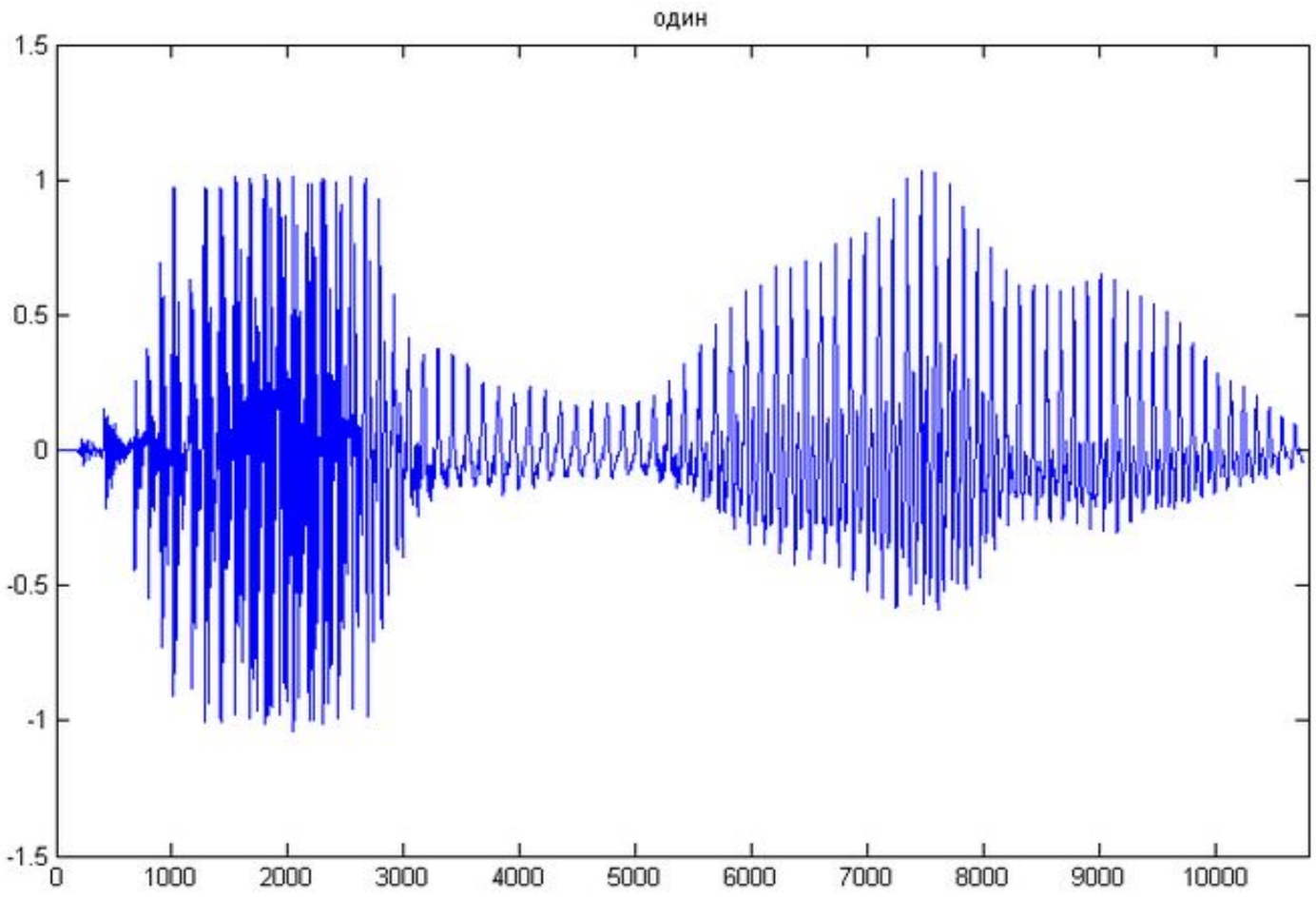
- Выделение признаков из звукового сигнала с помощью мел-кепстральных коэффициентов (MFCC).
- Обучение различных архитектур классификатора на полученных признаках.

Данные

- Датасет состоит из аудиокниг 5 различных дикторов.
- Каждая книга разбита на равное количество сэмпломв (для баланса классов)
- Семплы состоят из MFCC.
- Итоговый размер всего сета обучения: (744к*5, 39)

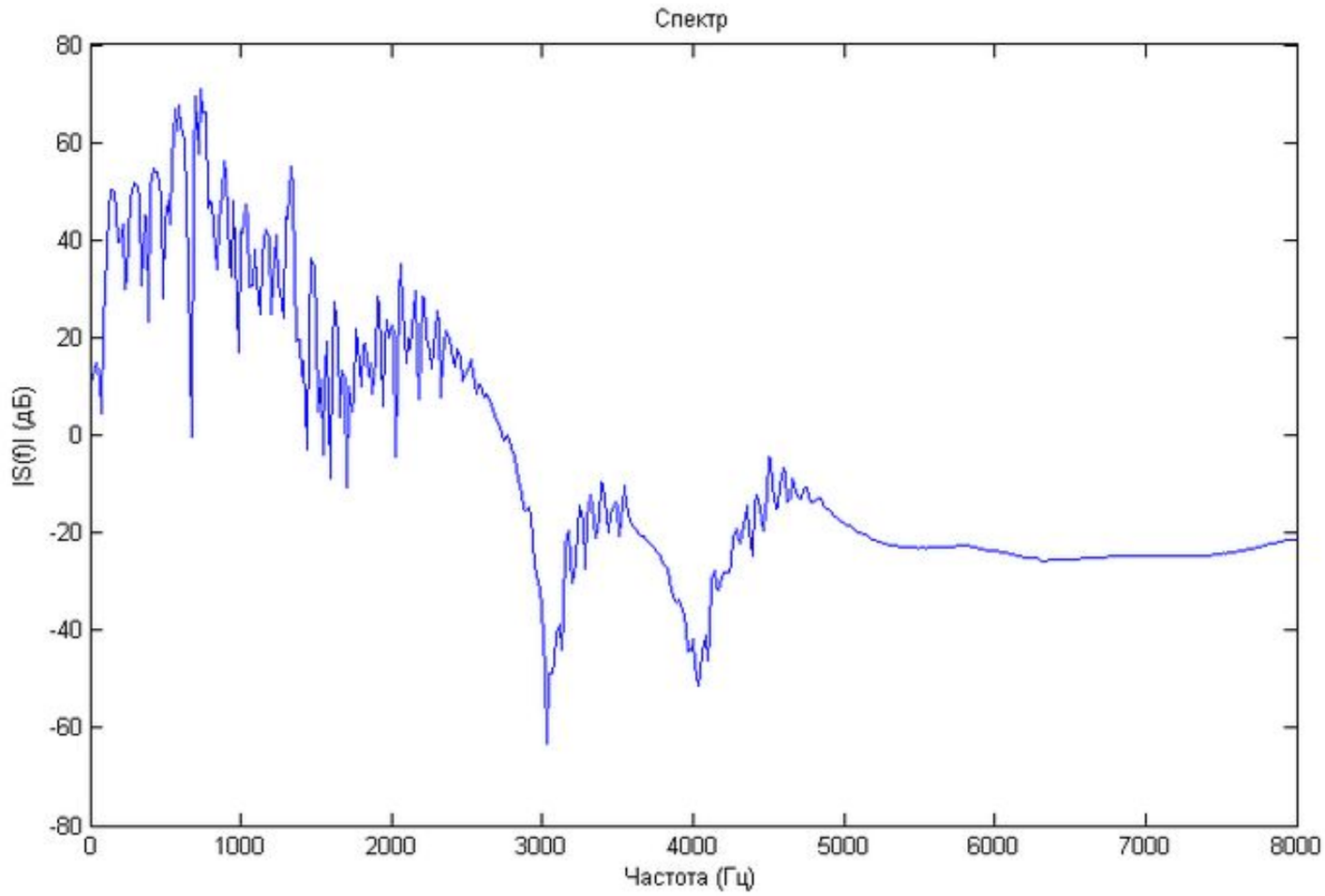
Изначальный сигнал

Произношение слова “один”:



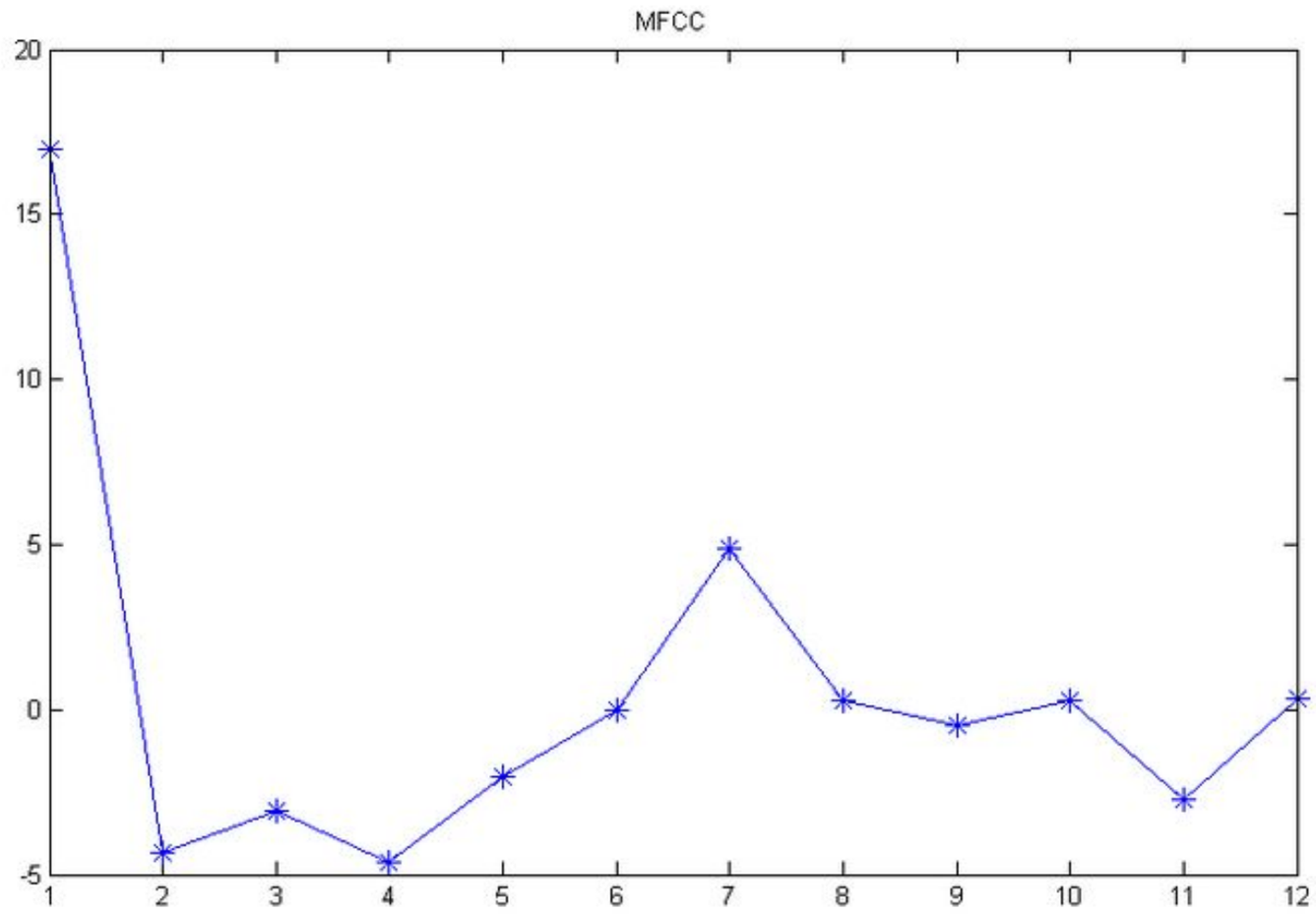
Изначальный сигнал

Преобразование Фурье:



Изначальный сигнал

Разбиваем по мел-шкале:



Модели

LightGBM (Baseline)

Bidirectionnel LSTM

Bidirectionnel GRU

Модели. LightGBM

В качестве baseline был взять LightGBM классикикатор.

Обучение на семплах из датасета (один семпл как отдельный элемент для обучения).

Качество на тесте: f1_score = 0.763

Сначала были дикие проблемы с переобучением на фоновой музыке. Решение- изменение частоты дискретизации (обучались на качестве записи), изменение датасета для получения единообразия.

Данные. RNN

Для рекуррентной сети пришлось подготовить данные тщательнее.

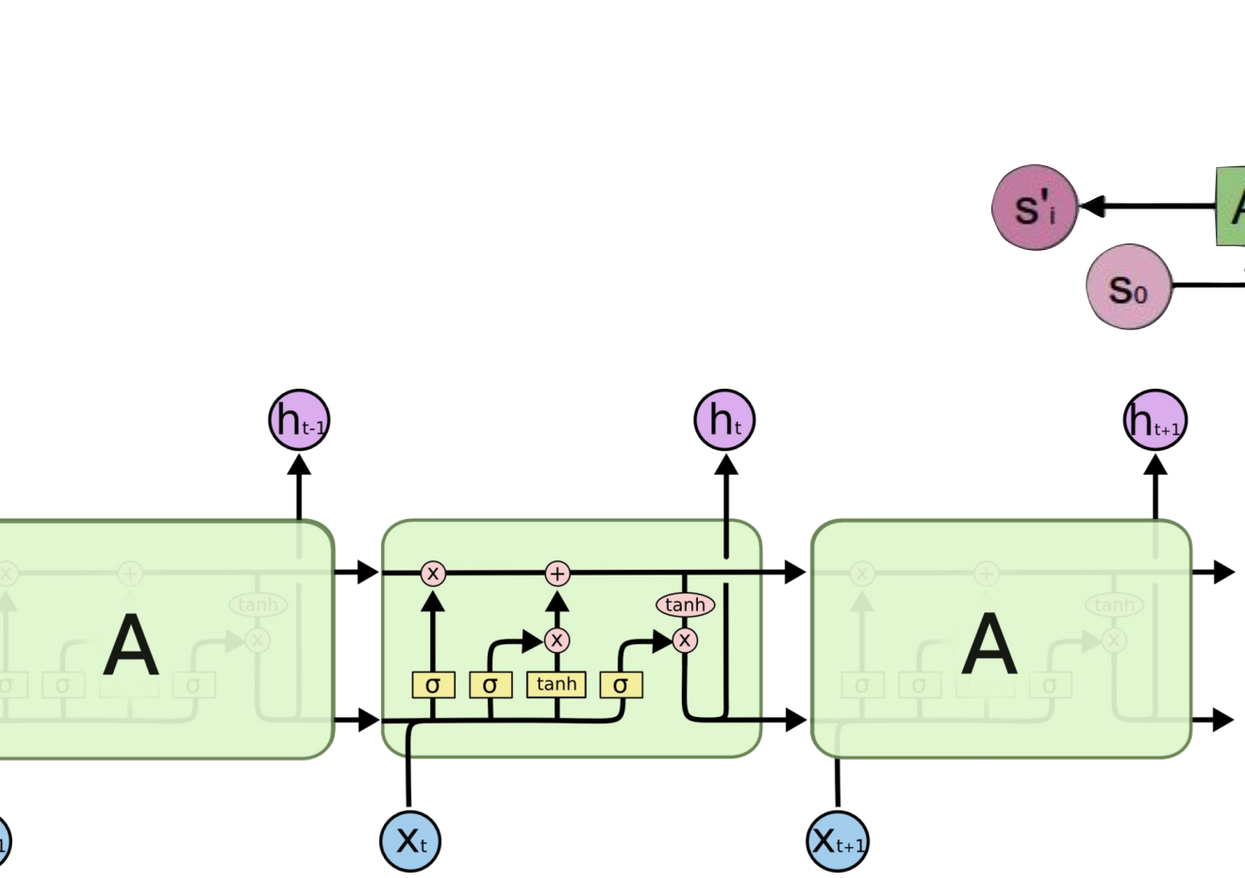
Последовательности длиной 10к (~4 мин)

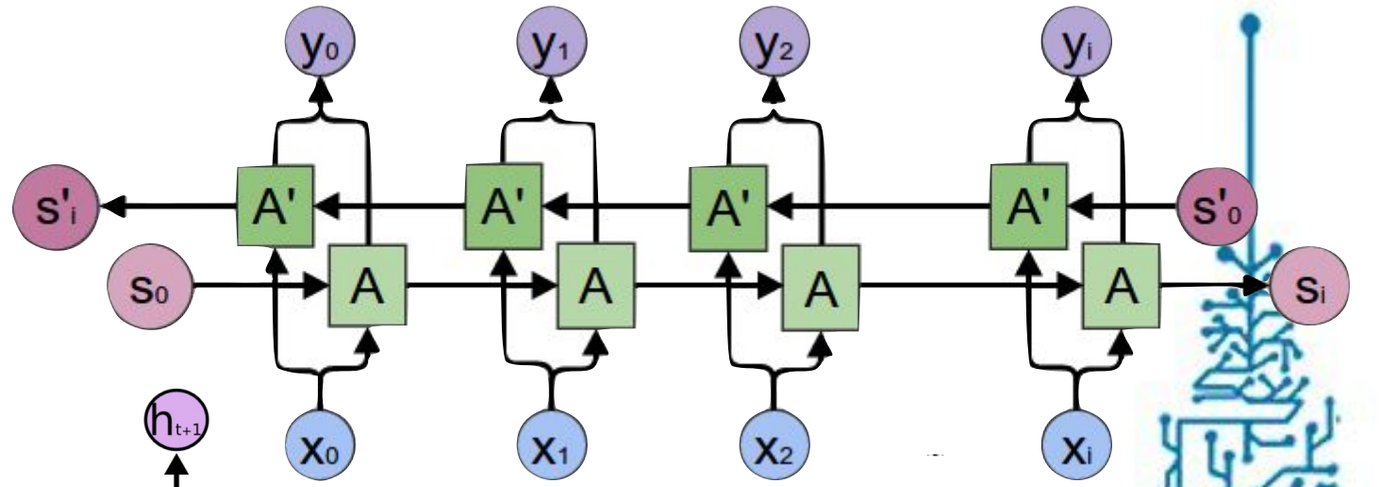
По 500 последовательностей на каждого диктора в обучении и валидации.

Модели. LSTM

Архитектура

Bidirectional, 3 скрытых слоя, размерность 25 +
Классификатор (полносвязный слой).

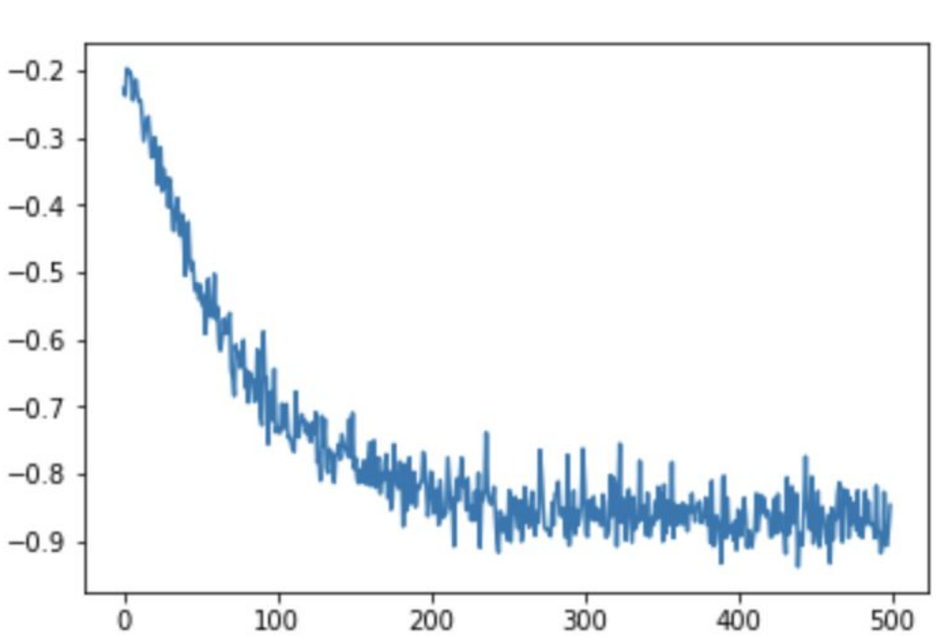


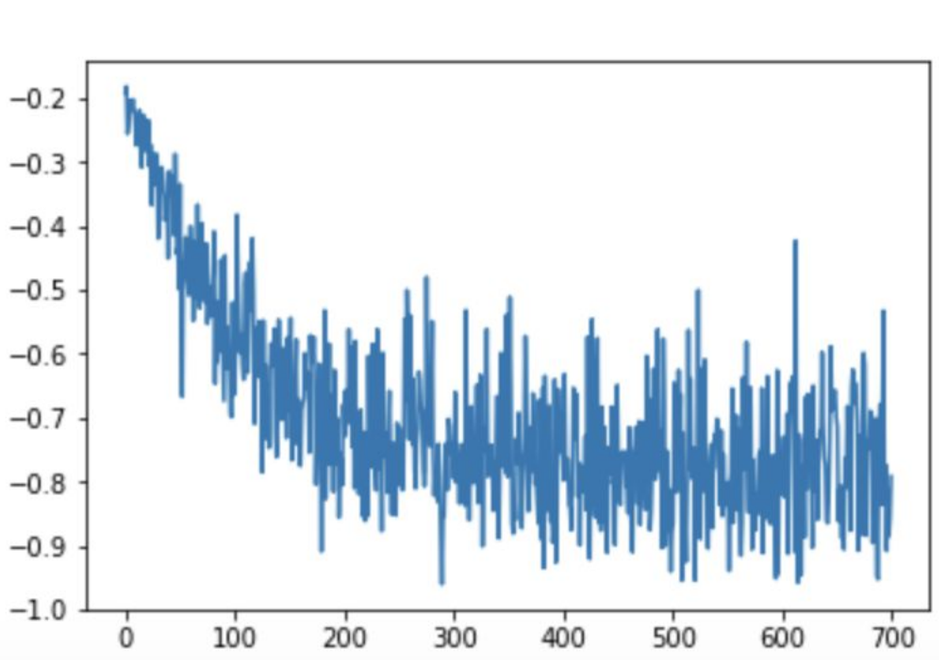


Модели. LSTM

Результаты

В связке с transfer learning удалось получить качество f1_score = 0.844

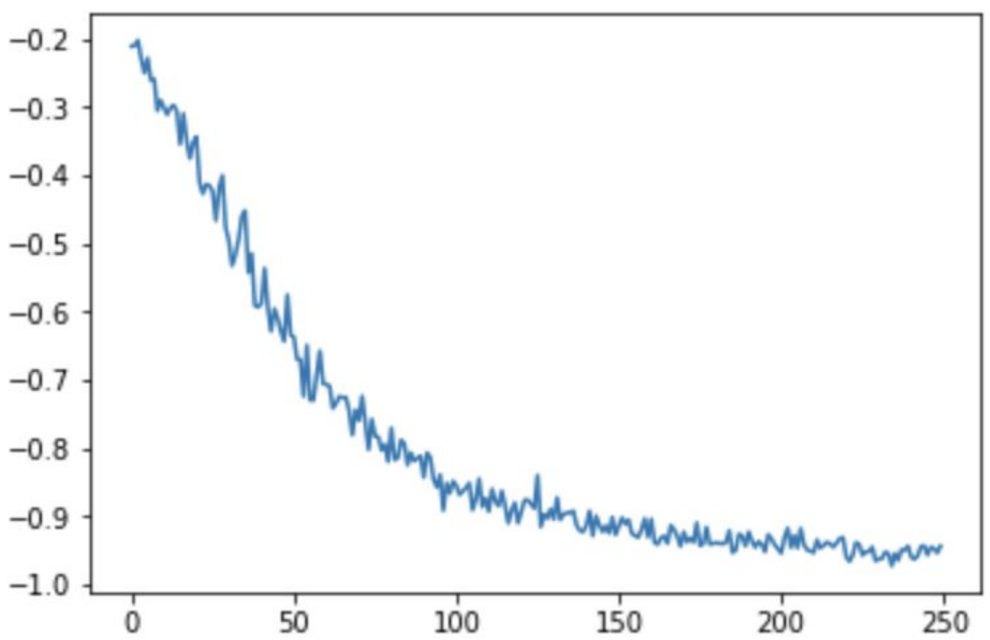


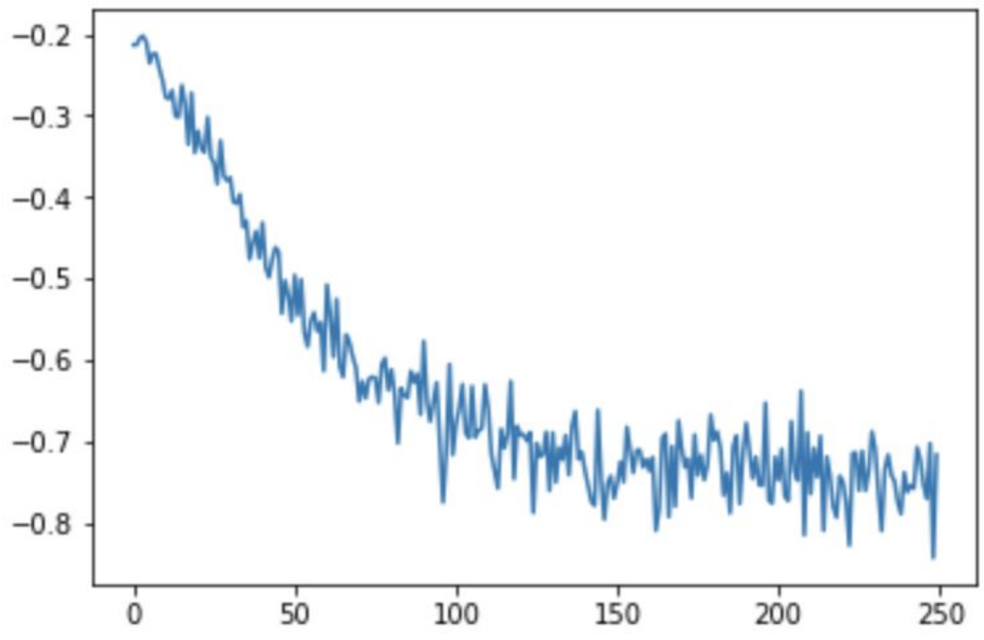


Модели. GRU

Дополнительно была рассмотрена архитектура GRU с аналогичными LSTM параметрами + Transfer Learning

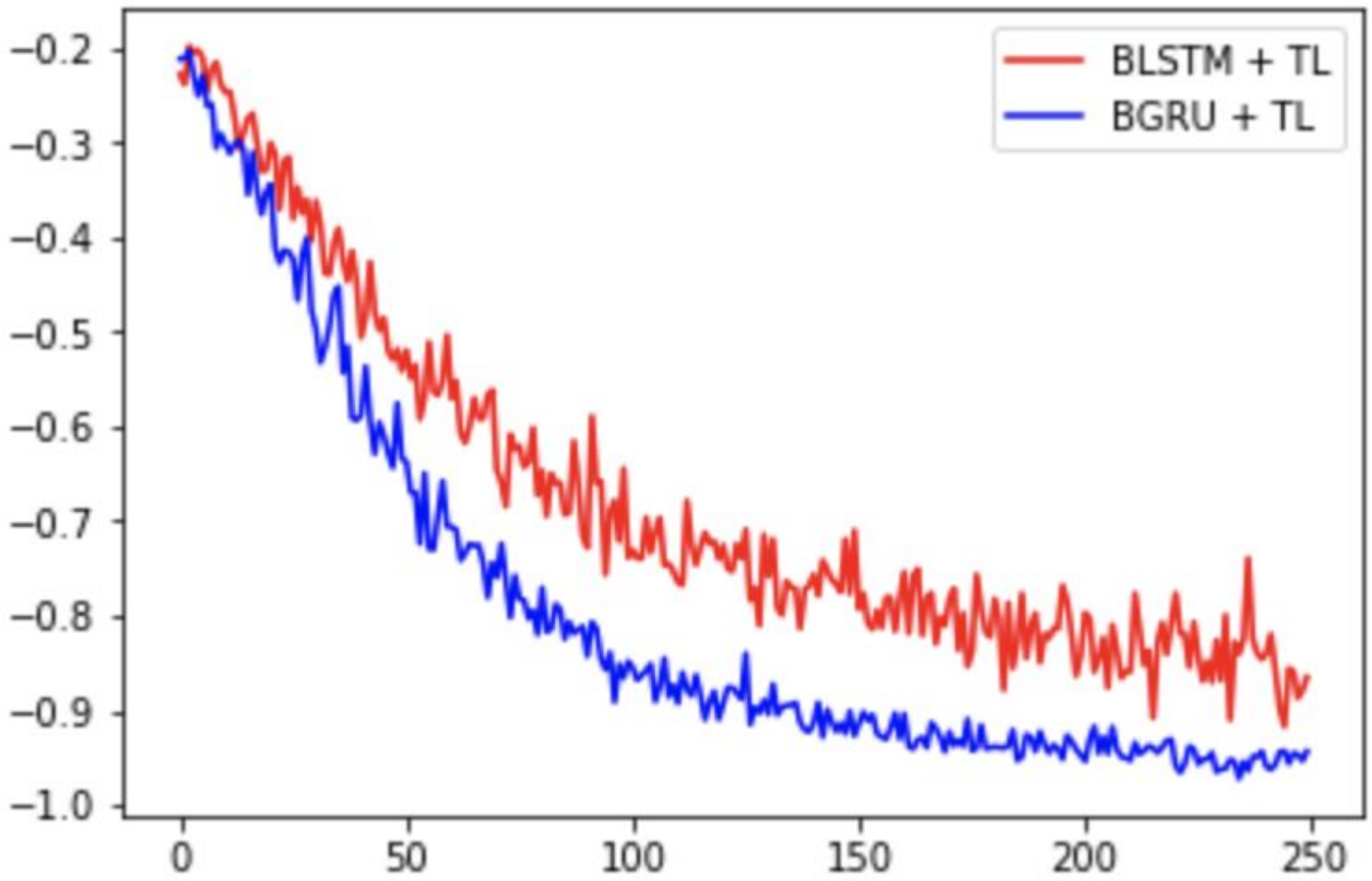
Качество на тесте: f1_score = 0.75





Модели. GRU

Сравнение темпов обучения BLSTM + TL & BGRU + TL



Выводы

Удалось получить результат значительно улучшивший baseline.

Остается острой проблема с исходными данными: очень чувствителен к качеству записи.

Спасибо за внимание!