Отчет по семестровому проекту

«Идентификация диктора по голосу»

Ефимов Владислав
Имеев Мерген
Руденко Дмитрий

Москва 2017

Постановка задачи.

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

Методы решения задачи.

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

Данные.

В качестве обучающей и валидационной выборок были взяти записи аудиокниг, озвучка которых была произведена различными дикторами. Всего было взято 5 дикторов. Изначально данные были представлены в формате mp3 и с различной частотой дискретизацией. В качестве предобработки данных было произведено декодирование в формат wav с единой частотой дискретизации.

Выделение признаков.

В качестве признаков брались мел-кепстральные коэффициенты, которые считались по окнам размера 25 мс с шагом 10 мс. В итоговый вектор признаков для конкретного окна брались первые 13 коэффициентов, а также их «траектории» (delta MFCC) и «ускорения» (delta-delta MFCC). Таким образом для каждого окна получался вектор из 39 коэффициентов.

Модели.

Получив векторное представление данных приходим к обычной задаче классификации. В качестве baseline модели рассматривался классификатор из библиотеки LightGBM. Помимо него рассматривались сети LSTM с различными параметрами, а также модификация последних — GRU.

Эксперименты.

В первых экспериментах наблюдались высокие показатели качества на обучения и очень низкие на валидации как в случае с прямым подходом, так и при обучении нейросети. Все это свидетельствовало о переобучении. Возможными причинами этого могло быть:

1) то, что изначально декодирование в wav формат проходило при увеличении частоты дискретизации относительно исходного формата, что могло вызывать переобучение на коэффициентах интерполяции алгоритма, который использовался для увеличения частоты дискретизации;
2) были записи очень плохого качества;
3) были дикторы, записи которых на обувающей и валидационной выборке на слух нельзя было идентифицировать как записи голоса одного и того же человека.

Далее эти недостатки были устранены редактированием изначальных данных (поиск новых качественных, замена старых записей), а также выбором небольшой общей частотой дискретизации в 16000 Гц.

В качестве меры качества при обучении использовался NLL Loss, также на обучении и тесте — f1 score взвешенный.

LightGBM

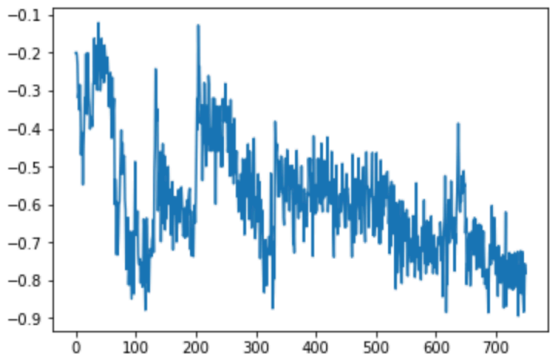
Как и упоминалось выше, в первой версии набора данных (некачественной) происходило переобучение: на обучающей выборке показатель f1score был порядка 0.98, на валидационной — менее 0.1. После исправления недочетов при формировании данных аналогичная метрика принимала значения 0.85 и 0.76 на обучающей и тестовой выборке соотвественно.

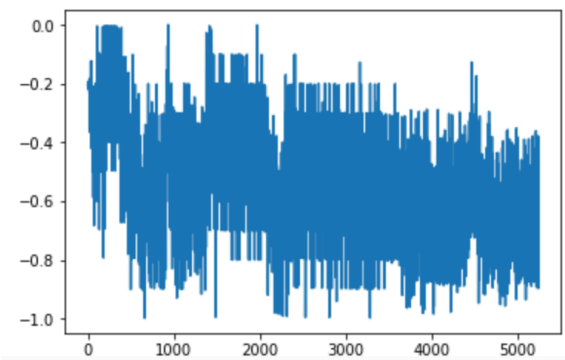
BLSTM

Была рассмотрена в качестве общей архитектуры следующая: размер скрытого слоя BLSTM брался равным 25, также после следовал полносвязный слой 25->5, в качестве нелинейности бралась ReLU, после полносвязного слоя брался Softmax. В частности были протестированы следующие конфигурации в дополнении к общей:

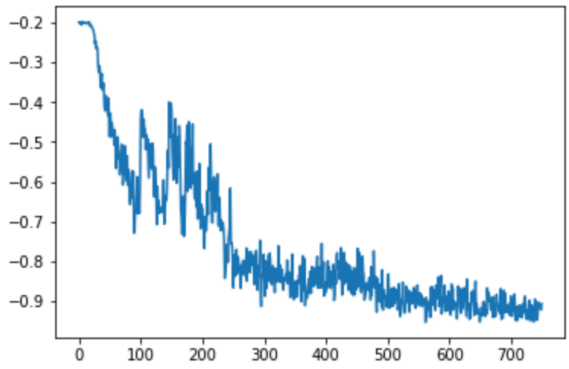
1) Количество скрытых слоев = 4.
2) Количество скрытых слоев = 4 + Dropout.
3) Количество скрытых слоев = 3 + Dropout.

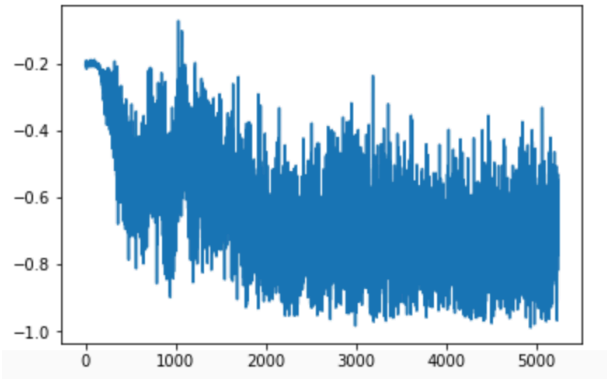
В случае 1) удалось получить качество на тестовой выборке в 0.73. Графики ошибки на обучении и на тесте во время обучения:



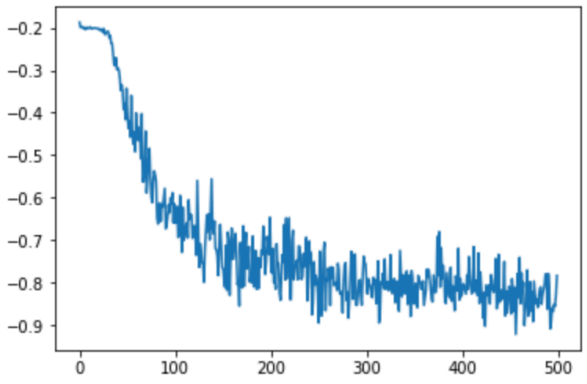


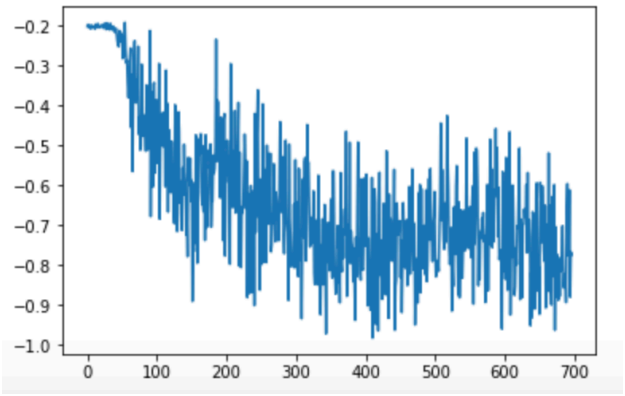
В случае с 2) качество не особо изменилось, однако график обучения стал лучше.



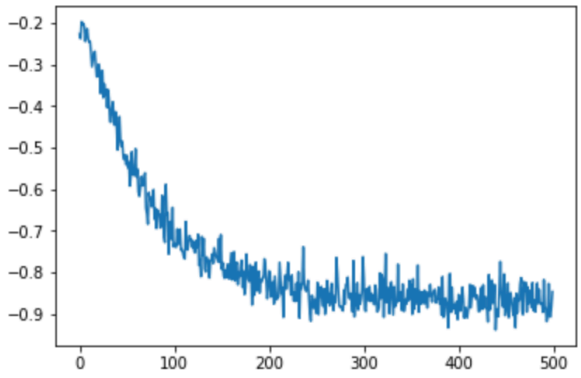


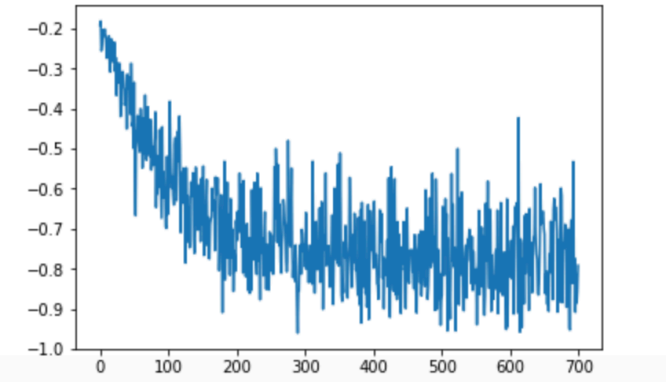
В случае 3) Удалось долиться значительного улучшения качества на тесте: 0.83. Графики:





Также для улучшения результата была использована идея transfer learning. Для предобученной из 3) LSTM был заново обучен классификатор (с теми же параметрами из 3)). Качество повысилось еще на 1%: 0.84. Графики обучения:

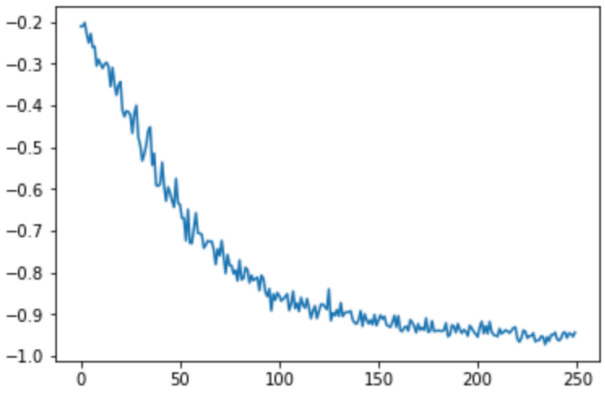


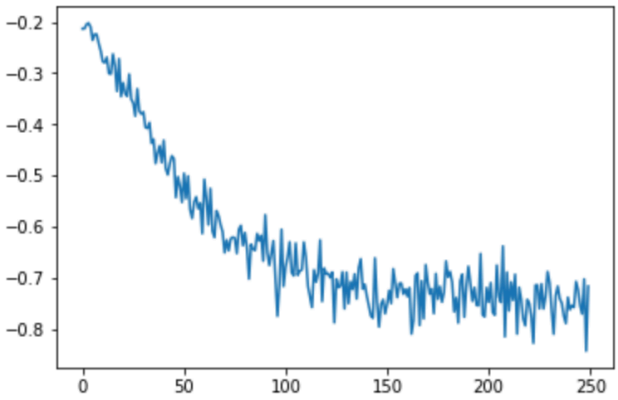


Также была попытка обучить LSTM с Dropout между скрытыми слоями, но сеть с такой архитектурой не обучалась.

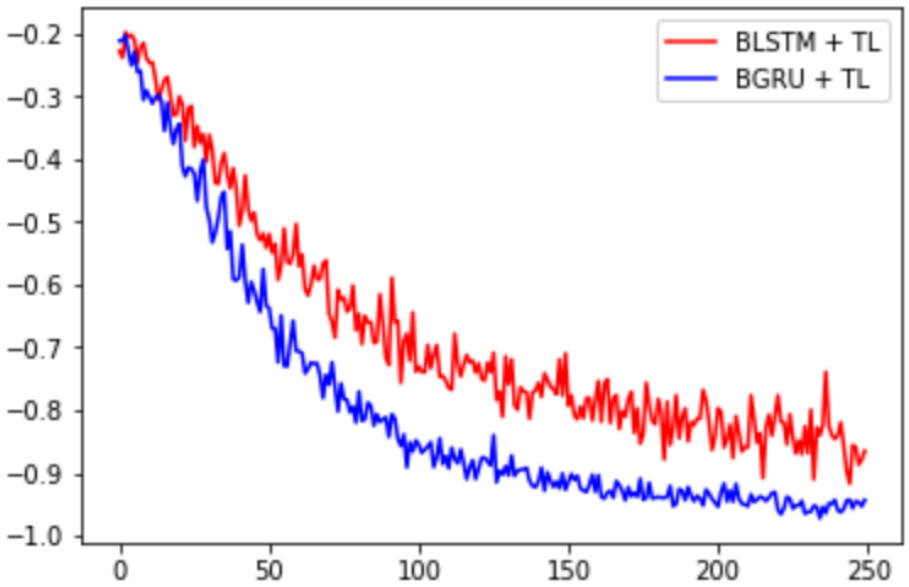
BGRU

Также была протестирована архитектура Bidirectional GRU с 3мя скрытыми слоями размеренности 25. А также была повторена идея 3) с Transfer Learning. Полученные результаты не превзошли baseline на тесте: f1 score = 0.75. Графики:





Также сравнили графики обучения BLSTM и BGRU:



Выводы.

Себя хорошо показала архитектура LSTM с 3мя скрытыми слоями. В дальнейшем хотелось бы попробовать обучить и протестировать сеть на большем числе дикторов. Однако, пока остается острой проблема с качеством исходных данных: плохое качество записи, oversampling приводят к плохим показателям обучения (переобучению).

Литература.

1. X. Huang, A. Acero, and H. Hon. Spoken Language Processing: A guide to theory, algorithm, and system development. Prentice Hall, 2001.
2. Larsson, Joel. "Optimizing text-independent speaker recognition using an LSTM neural network." (2014).
3. Olah, Christopher. "Understanding LSTM Networks. 2015." URL http://colah. github. io/posts/2015-08-Understanding-LSTMs/img/LSTM3-chain. png (2015).