Классификатор мужских и женских голосов на базе LibriTTS

Анализ данных

Датасет LibriTTS.

Для экспериментов использовались только аудиозаписи речей и id спикеров, по которым размечался датасет.

Частота дискретизации для всех записей 24 кГц.

Записи отличаются по длительности, но длительность мной никак не менялась.

Объемы выборок:

Train: 12479 мужских и 14109 женских семплов

Dev: 3128 мужских и 3520 женских семплов

Test: 1907 мужских и 2930 женских семплов

В целом, датасет можно считать сбалансированным по классам.

Метки классов: 0 – FEMALE, 1 – MALE.

Конструирование признаков

- Файл – feature_extractor_learn.ipynb
- По сырым данным сначала строились мел-спектрограммы, так как они представляют собой сжатое, но более информативное представление звукового сигнала. Параметры преобразования установленные вручную:

  n_fft = 1024 (установлено эмпирически)

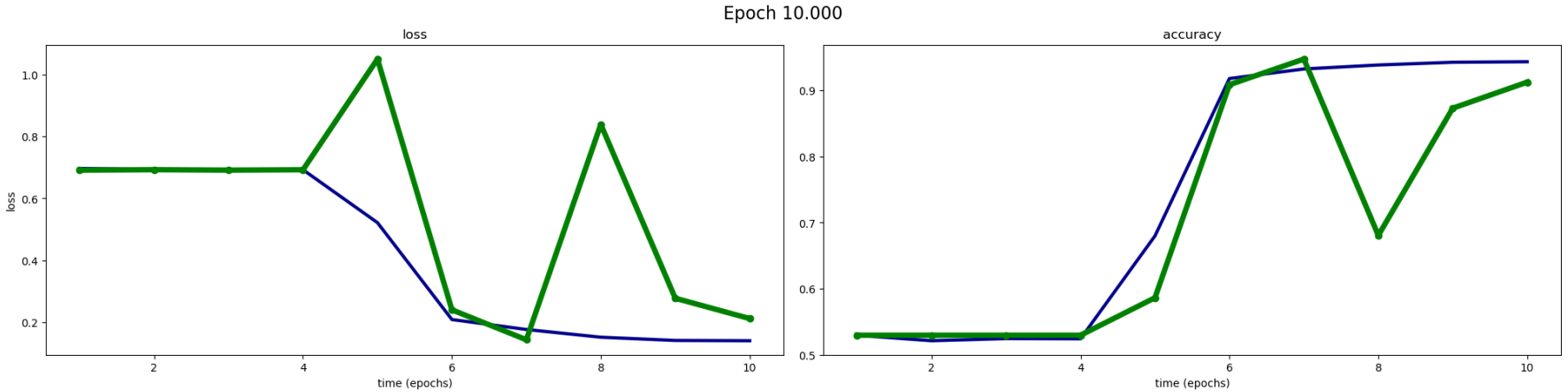
  n_mels = 128 (установлено эмпирически)

  Normalized = True (позволило не применять переход к децибелам)
- К тренировочному набору применялись аугментации: частотное и временное маскирование с целью повышения обобщающей способности конструктора признаков.
- В конце все спектрограммы обрезались до единого размера [128 x 128] и дальше с ними работали уже как с изображениями

Конструирование признаков

- В качестве конструктора признаков, который будет строить эмбеддинги звукозаписей на основании спектрограмм я решил взять ResNet18, так как сети ResNet являются популярными backbone для выделения признаков. Результатом данного этапа является обученная сеть ResNet18 сохраненная в файле ./parameters/model.pkl, которая входящий тензор [128, 128] кодирует в вектор длиной 8192.

На графике ниже представлен процесс обучения ResNet18.



Конструирование признаков

- Файл – feature_extractor_get_dataset.ipynb

Генерируем новый датасет, используя обученый backbone, преобразуя его в векторы длины 8192.

- Файл – voice_classifier.ipynb

Векторы такой размерности являются избыточными, так как многие их компоненты являются нулями. Для понижения размерности я воспользовался методом главных компонент (PCA). Количество компонент равно 128. Это оптимальное количество, установленное эмпирически.

Обученный PCA преобразователь сохранен в файле ./parameters/pca.pkl

Классификация

- Файл – voice_classifier.ipynb

В качестве алгоритма классификация был выбран градиентный бустинг CatBoostClassifier, так как ансамблевые методы являются на данный момент лучшими в задачах классификации.

Обученный классификатор сохранен по адресу ./parameters/clf_model.pkl

В конечном итоге имеем следующие результаты на test выборке:

	Precision	Recall	F1-score	Count
Female	0,97	0,99	0,98	2930
Male	0,98	0,96	0,97	1907
Accuracy		0,98		4837

Выводы

- В целом можно сказать результат получился неплохим, пайплайн по конструированию признаков на основе мел-спектрограммы и сети ResNet оказался удачным, так как позволил обучить классификатор с показателями точности выше 0,96.
- Для того чтобы улучшить результат, возможно, стоит попробовать более специализированные конструкторы признаков, такие как VGGish, OpenL3 или YamNet.