Speaker recognition

Журавская, Камалбеков, Козловцев
517 группа

Московский Государственный Университет имени М.В.Ломоносова
Факультет вычислительной математики и кибернетики
Кафедра математических методов прогнозирования

24 июня 2019 г.

Постановка задачи

Дано: образцы голосов N спикеров;
Обучающая и тестовая выборки;
У каждого спикера уникальный id.

Найти: для каждого образца в тестовой выборке либо подобрать id спикера, либо определить, что таких id в обучающей выборке нет.

Наивный алгоритм

- Признаки: mfcc из второго задания!
- Разделим пространство признаков на кластеры: K-Means!
- Найдем порог принадлежности кластеру: если расстояние от объекта до всех центров кластеризации больше порога, то считаем, что спикер новый, т.е. его не было в обучающей выборке.

Детали: Количество признаков 24. Количество кластеров $4N$*.
Проблемы: плохая разделимость, низкая скорость работы

*Speaker identification using mel frequency cepstral coefficients

VGGist

- Данные: VoxCeleb аудио + видео!
- Предобработка: ....
- Сеть: VGGish на GitHub

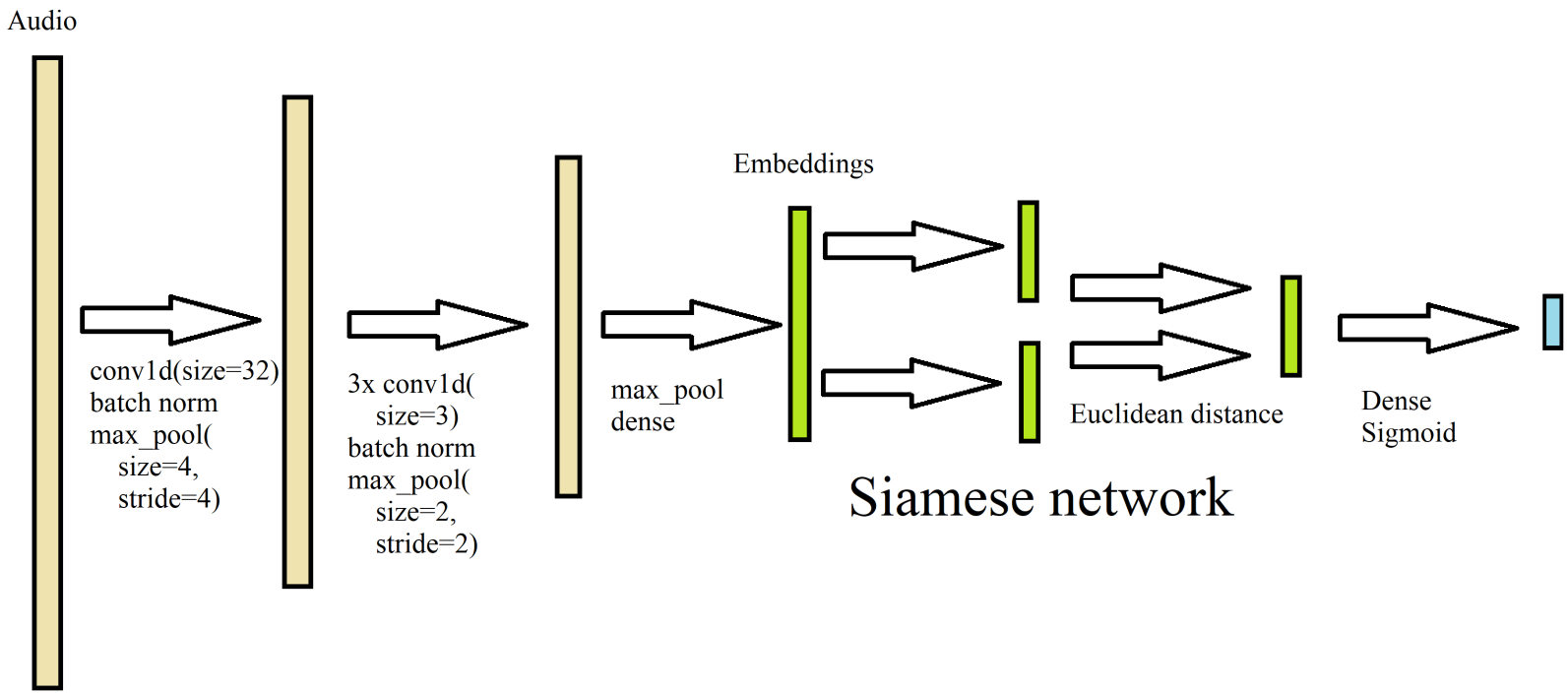
Проблемы: tf -> torch, свою сеть обучать долго: используем предобученную!

Итог: точность $\approx 20\%$ на 100 спикерах с наибольшим количеством образцов голоса.

Мы не сдаёмся

Building a Speaker Identification System from Scratch with Deep Learning

- Данные: LibriSpeech
- Архитектура:



Эксперименты

Датасет: LibriSpeech

- train-clean-100 – 100 часов «чистой» английской речи, 250 различных людей;
- train-clean-360 – 360 часов «чистой» английской речи;
- test-clean – 40 различных пользователей.

Идея: будем использовать кодировщик из предобученной «сиамской» сети для преобразования признакового пространства.

Эксперименты. Эмбеддинги

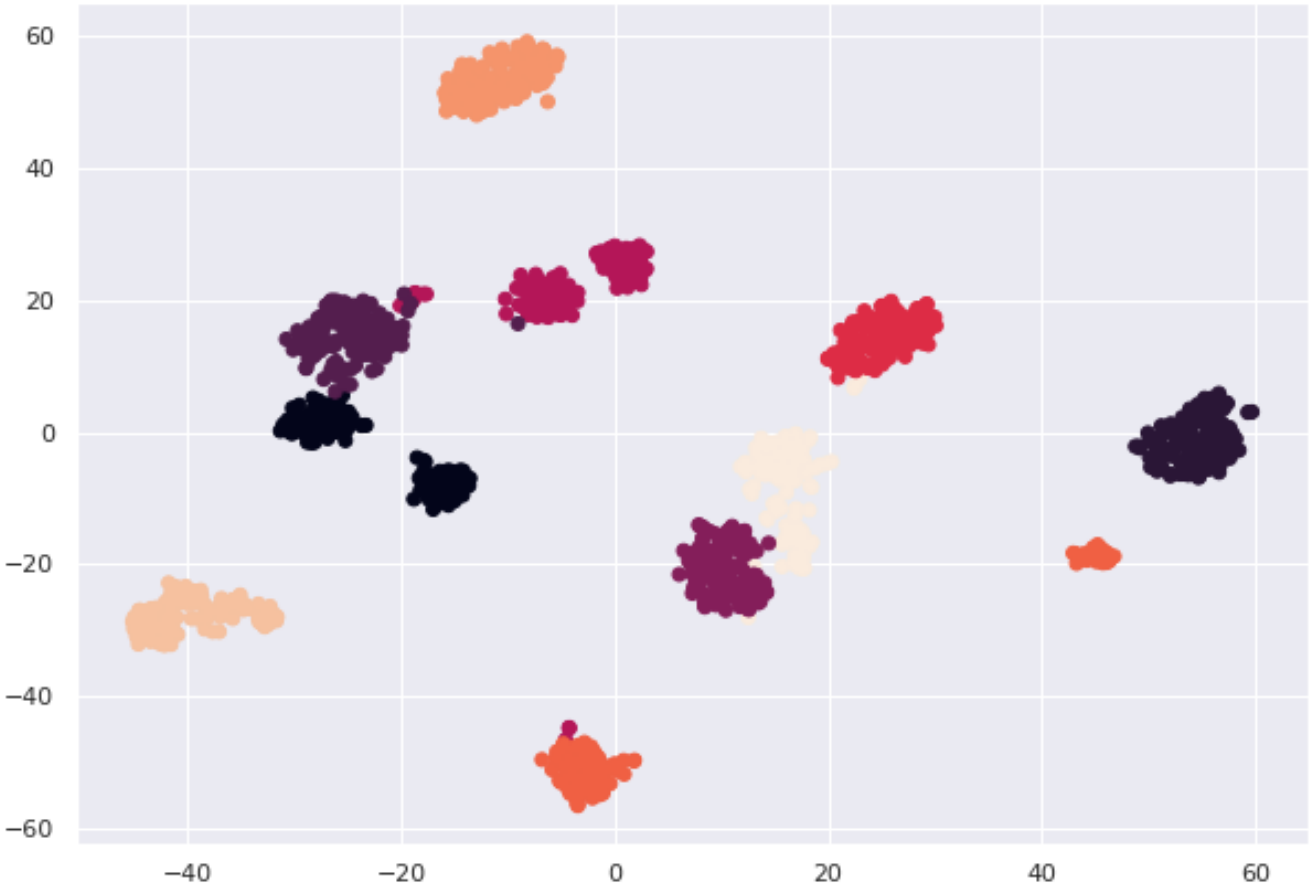


Рис.: Эмбеддинги тестовых записей

Эксперименты. K-shot learning

Как точно мы можем идентифицировать человека, если до этого наблюдали не более k элементов каждого класса? Классификация – KNN.

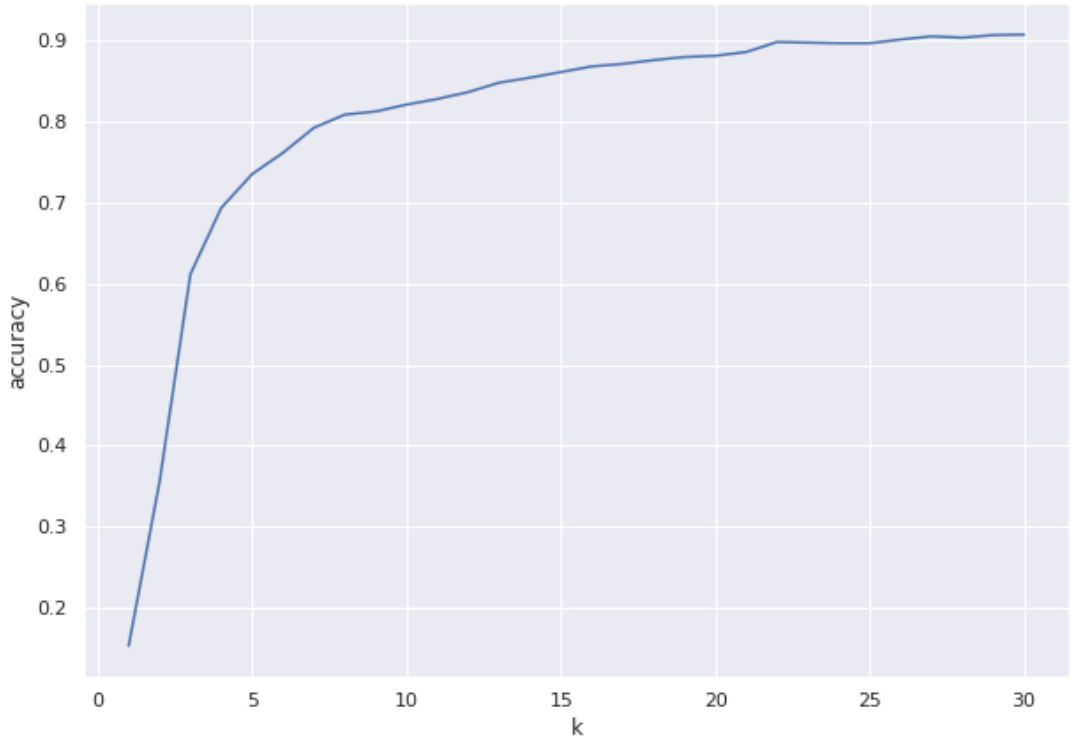


Рис.: Влияние k на качество модели

Эксперименты. Transfer Learning

Алгоритм	Accuracy
KNN на эмбеддингах	94.4%
KNN на усреднённых эмбеддингах	83.3%
Случайный лес	93.1%
Градиентный бустинг	91.2%

Таблица: Сравнение моделей

Заключение

Мы изучили существующие методы решения задачи распознавания спикера. Мы попробовали несколько методов распознавания на нескольких наборах данных. Получили опыт использования предобученных нейронных сетей.

Наилучший результат получен с использованием siamese networks, точность классификации на наборе данных LibriSpeech 94.4%!

Вклад участников:

Саша	идея, наивный алгоритм, презентация
Костя	сеть: выбор архитектуры, перенос весов
Тимур	данные: предобработка и эксперименты