Правительство Российской Федерации

Федеральное государственное автономное образовательное

учреждение высшего образования

Национальный исследовательский университет

«Высшая школа экономики»

Факультет гуманитарных наук

Образовательная программа

«Фундаментальная и компьютерная лингвистика»

Сонина Полина Алексеевна

**Сравнение работы методов машинного обучения: спектрограмма и звук**

Выпускная квалификационная работа

студентки 4 курса бакалавриата группы БКЛ141

|  |  |
| --- | --- |
| Академический руководитель образовательной программы | Научный руководитель |
| канд. филологических наук, доц. | старший преподаватель |
| Ю.А. Ландер | Г.А. Мороз |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |
| « » \_\_\_\_\_\_\_\_\_\_ 2018 г. | « » \_\_\_\_\_\_\_\_\_\_ 2018 г. |

Москва 2018

Оглавление

[1. Введение 1](#_Toc515192934)

[2. Обзор литературы 2](#_Toc515192935)

[3. Данные 2](#_Toc515192936)

[4. Этапы работы 3](#_Toc515192937)

[4.1. Подготовка данных 3](#_Toc515192938)

[4.2 Извлечение признаков (feature extraction) 3](#_Toc515192939)

[4.3 Базовые классификаторы 3](#_Toc515192940)

[4.4 Нейронные сети 3](#_Toc515192941)

[5. Анализ результатов 3](#_Toc515192942)

[6. Заключение 5](#_Toc515192943)

[Литература 6](#_Toc515192944)

[Приложение 7](#_Toc515192945)

1. Введение

На данный момент разработка систем для автоматического анализа устной речи является одним из основных направлений в области машинного обучения. Системы распознавания речи широко применяются для извлечения текстовой информации из аудио как в научных целях, так и в коммерческих.

Чтобы аудио данные можно было использовать при обучении модели, нужно провести извлечение признаков (feature extraction) из первоначального сигнала. Для извлечения признаков из аудио существует набор стандартных метрик, которые используются во многих существующих системах. Высокое качество работы современных систем обычно достигается за счёт обучения на большом количестве данных, а также за счёт доступности больших вычислительных ресурсов.

Существуют исследования, предлагающие альтернативный подход к анализу аудио данных. Подход заключается в представлении аудио в виде спектрограммы - изображения, показывающего зависимость спектральной плотности мощности сигнала от времени. Использование спектрограмм для классификации речевых единиц представляется возможным благодаря тому, что форманты, образующие речевые единицы, различимы на спектрограмме для человеческого глаза. Подобные исследования в большинстве своём проведены на англоязычных данных, для русскоязычных же данных конкретный подход недостаточно исследован.

Цель данной работы – исследовать подход к извлечению признаков, основанный на методах классификации изображений, и сравнить его работу с общепринятыми для аудио подходами. Так же, целью являлось максимизировать точность работы алгоритма классификации на малом объёме данных при небольших вычислительных затратах.

Задачи исследования состояли в:

* сборе небольшого количества аудио данных на русском языке;
* создании алгоритма для классификации данных, использующих стандартные способы извлечения признаков;
* создании алгоритма, где аудио данные представляются в виде спектрограмм, к которым применяются методы извлечения признаков из изображений;
* сравнении работы полученных алгоритмов;
* анализ ошибок, характерных для каждой из систем.

2. Обзор литературы

Обобщающие состояние области

Подходы использующие спектрограммы

Актуальные для задачи методы анализа изображений

Contemporary research papers on sound recognition are in abundance. However, papers focused on using spectrogram image processing are not that common.

(Washani, Sharma 2015) gives an overview of speech recognition research field, it’s problems, progress and current state. The paper stresses the importance of preprocessing stage for combating the most prominent problem - noise.

(Madan, Gupta 2014) reviews speech processing techniques focusing on feature extraction and classification. Methods compared in the paper are stated to be the most used in the field and are going to be useful for our research. MFCC combined with other techniques is suggested for feature extraction and discrete hidden markov model – for word classification.

(Dennis 2014) provides a relevant approach to sound events recognition. This thesis develops a method based on image processing-inspired feature extraction from the spectrogram combined with more traditional sound processing techniques. The results a quite successful but the research was focused on sound events and not speech recognition. However, development of similar models for speech recognition is suggested by the author as possible future work.

(Nguyen, Bui 2016) proposes a model based on feature extraction from spectrogram images for speech classification. The specific approach also allows adding training samples without retraining the model which is useful for big data tasks. The model showed satisfactory results on different speech datasets.

3. Данные

Для обучения экспериментальных моделей были собраны аудио данные в формате wav. На записях носители русского языка произносили числа от одного до пятидесяти (1-50). В результате было использовано 27 наборов данных, состоящих из 50 речевых единиц. Говорящие на записях – взрослые (21-55 лет) жители Москвы и Московской области, мужчины и женщины. Для записи были использованы встроенные микрофоны нескольких смартфонов, поэтому на качество аудио повлияли различия в чувствительности устройств. Также, запись проводилась в разных условиях, что привело к различным уровням зашумлённости данных. Качество собранных данных близко бытовым условиям использования технологии распознавания речи, например, в качестве пользовательских интерфейсов.

4. Этапы работы

Программы для работы с данными были написаны на языке Python 3. Были использованы дополнительные библиотеки:

* numpy, os, time, matplotlib – для общих задач;
* librosa – для работы с аудио данными;
* scipy, scikit-image – для работы с изображениями;
* scikit-learn – для создания и оценки обучаемых моделей;
* tensorflow – для создания искусственных нейронных сетей.

4.1. Подготовка данных

4.2 Извлечение признаков (feature extraction)

стандартные для аудио, для изображений из мел спектрограмм

4.3 Базовые классификаторы

4.4 Нейронные сети

5. Анализ результатов

Общие оценки

Сравнение ошибок

Сравнение затратности методов

6. Заключение

Литература

1. Schutte, K.T. (2009). *Parts-based Models and Local Features for Automatic Speech Recognition.*
2. Madan, A. & Gupta, D. (2014). *Speech Feature Extraction and Classification: A Comparative Review*. International Journal of Computer Applications (0975 – 8887) Volume 90 – No 9, March 2014
3. Dennis, J.W. (2014). *Sound Event Recognition in Unstructured Environments using Spectrogram Image Processing.*
4. Washani, N. & Sharma, S. (2015). *Speech Recognition System: A Review. International Journal of Computer Applications* (0975 – 8887) Volume 115 – No. 18, April 2015
5. Nguyen, Q.T. & Bui, T.D. (2016). *Speech classification using SIFT features on spectrogram images*. Vietnam Journal of Computer Science. November 2016, Volume 3, Issue 4, pp 247–257.

Приложение

Материалы в репозитории на Github:

<https://github.com/SoDipole/ML-sound-vs-picture>