Решение задачи распознавания звуковой информации с применением методов машинного обучения

Бармина Ольга Константиновна, НФИбд-01-19

Российский университет дружбы народов имени Патриса Лумумбы,

Факультет физико-математических наук,

Кафедра информационных технологий

Научный руководитель: к.ф.-м.н., доцент Хачумов М. В

Российский университет дружбы народов

Москва, 2023

Введение

Целью работы является изучение и применение современных методов машинного обучения для решения задачи распознавания речевых команд.

Задачи:

1. Аналитическое сравнение методов машинного обучения, включая нейросетевые подходы, применительно к задаче распознавания звуковых команд;
2. Анализ способов предварительной обработки аудио сигнала;
3. Исследование подхода к решению поставленной задачи, основанного на применении мел-спектрограмм;
4. Проведение экспериментальных исследований по распознаванию речевых команд на основе различных методов машинного обучения;
5. Сравнение и оценка полученных результатов.

Актуальность, новизна и области применения

Актуальность работы обуславливается появлением множества новых подходов к обработке и исследованию звуковой информации и расширением сфер применения.

Новизна работы заключается в исследовании современного подхода, основанного на применении мел-спектрограмм на расширенном наборе нейронных сетей (MobileNet, Xception, ShuffleNet, RegNet, SqueezeNet).

Области применения:

- Голосовые ассистенты,
- Сбор информации,
- Система “умный дом”.

Содержание

Глава 1. Постановка задачи, обзор сфер применения.

Глава 2. Исследование методов решения задачи распознавания звука.

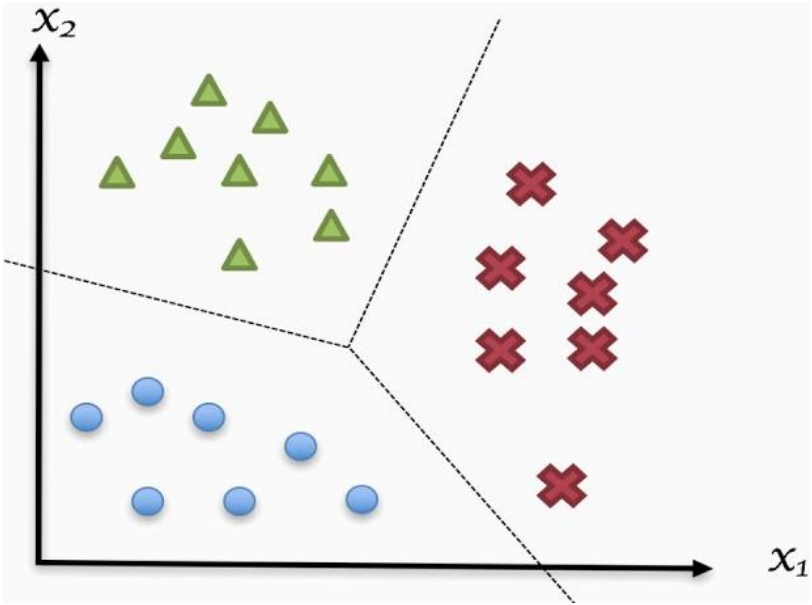
Глава 3. Исследование методов предобработки аудио.

Глава 4. Практическое исследование решения задачи.

Постановка задачи классификации

Дано: образ x, набор классов $y = \{y_1, y_2 \dots y_n\}$, обучающая выборка.

Задача классификации: нахождение функции $f: x \to y_i, i \in [1, n]$



Методы машинного обучения

В работе были использованы следующие методы машинного обучения:

- k-ближайших соседей,
- Наивный Байесовский классификатор,
- Метод опорных векторов,
- Дерево решений.

Нейросетевые:

- CNN со случайными весами,
- CNN с весами ImageNet,
- LSTM.

Рассмотрены основные принципы, преимущества и недостатки данных методов.

Выбор архитектур CNN

Проведен аналитический обзор, в результате которого выявлены следующие архитектуры CNN:

- DenseNet
- MobileNet
- MobileNetV2
- MobileNetV3
- Xception
- RegNet
- ShuffleNet
- ShuffleNetV2
- SqueezeNet

Отобранные архитектуры охватывают все современные принципы построения быстрых и точных сетей.

Способы предобработки аудио данных

Волновая форма:

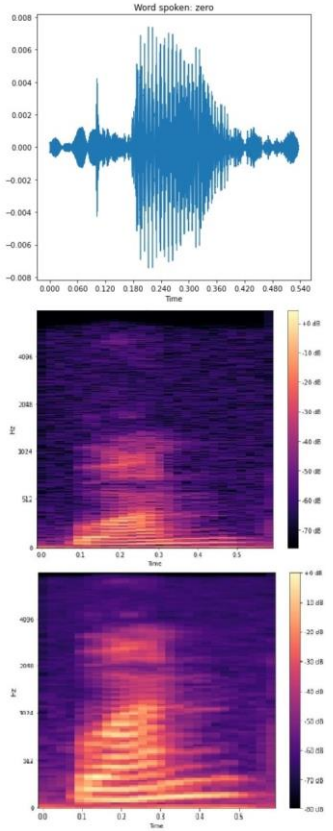
- Демонстрирует отношение амплитуды ко времени.

Спектрограмма:

- Разложение волны на частотные полосы,
- Демонстрирует отношение частоты ко времени,
- Цвет каждой точки зависит от амплитуды.

Мел спектрограмма:

- Направлена на сохранение информации, хорошо воспринимаемой человеческим слухом.
- Вычисляется по обычной спектрограмме.
- Перевод частоты из Гц в мел: $m = 2595 \ln(1 + \frac{f}{700})$.



Формирование данных

Датасет **AudioMNIST**: записи произношения чисел от 0 до 9 на английском языке.

Структура:

- 30000 записей, 60 дикторов.
- Каждое число произнесено 50 раз каждым диктором.

Преобразование данных:

- Дублирование на 3 канала.
- Нормализация в интервал [0; 1].
- Логарифмирование $LogMel = \ln(m + \varepsilon),\ \varepsilon > 0.$
- Разделение 7:2:1.

Построение и параметры моделей

CNN со случайными весами:

- Метрики precision, recall, и F-мера,
- Обучение на 15 эпохах.

CNN с весами ImageNet:

- Дополнительный полносвязанный слой,
- Обучение на 10 эпохах с замороженными весами,
- Разморозка последних 30% слоев, обучение на 15 эпохах.

Обучение со случайно заданными весами

Основная архитектура	Модель	Precision	Recall	F-мера
DenseNet	DenseNet121	0,9821	0,9822	0,9821
	DenseNet169	**0,9856**	**0,9856**	**0,9856**
	DenseNet201	0,9837	0,9836	0,9836
MobileNet	-	0,9648	0,9647	0,9647
MobileNetV2	-	0,9844	0,9826	0,9835
MobileNetV3	MobileNetV3Large	0,9817	0,9794	0,9806
	MobileNetV3Small	0,9705	0,9678	0,9691
Xception	-	0,9716	0,9715	0,9715
RegNet	RegNetX002	0,9461	0,9441	0,9450
	RegNetX004	0,9389	0,9377	0,9382
	RegNetX006	0,9506	0,9503	0,9504
	RegNetY002	0,9444	0,9441	0,9442
	RegNetY004	0,9333	0,9310	0,9321
	RegNetY006	0,9587	0,9587	0,9587
ShuffleNet	-	0,9259	0,9114	0,9186
ShuffleNetV2	-	0,9519	0,9439	0,9478
SqueezeNet	-	0,9786	0,9778	0,9772

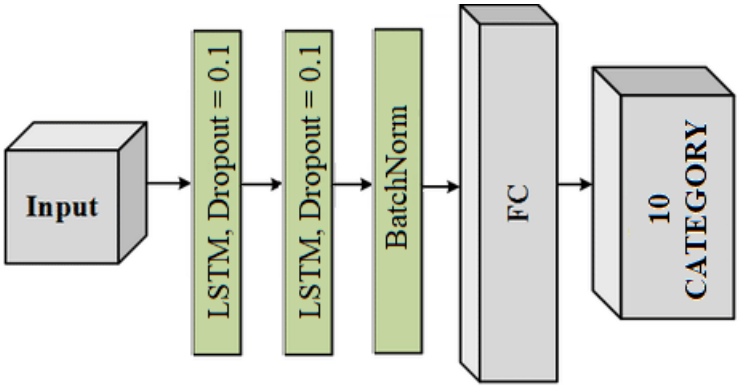
Обучение с весами ImageNet

Основная архитектура	Модель	Precision с fine-tune	Recall с fine-tune	F-мера с fine-tune
DenseNet	DenseNet121	0,9801	0,9801	0,9801
	DenseNet169	0,9733	0,9733	0,9733
	DenseNet201	**0,9837**	**0,9838**	**0,9837**
MobileNet	-	0,9774	0,9812	0,9793
MobileNetV2	-	0,9742	0,9738	0,9740
MobileNetV3	MobileNetV3Large	0,9668	0,9668	0,9668
	MobileNetV3Small	0,9469	0,9459	0,9463
Xception	-	0,9734	0,9732	0,9733
RegNet	RegNetX002	0,9617	0,9614	0,9615
	RegNetX004	0,9641	0,9642	0,9641
	RegNetX006	0,9703	0,9703	0,9703
	RegNetY002	0,8995	0,8954	0,8975
	RegNetY004	0,9025	0,9017	0,9021
	RegNetY006	0,9754	0,9755	0,9754

Сравнение с другими методами машинного обучения

LSTM:

- Подвид RNN сети,
- Сохранение информации о предыдущих состояниях,
- Обучение на 25 эпохах.



Классификатор	Precision	Recall	F-мера
GaussianNB	0.5603	0.5276	0.5231
Kneighbors	0.9619	0.9617	0.9617
SVC	**0.9828**	**0.9829**	**0.9828**
NuSVC	0.9782	0.9782	0.9782
LinearSVC	0.9593	0.9593	0.9593
RandomForest	0.9339	0.934	0.9338
ExtraTrees	0.7141	0.693	0.6879
DecisionTree	0.4513	0.4175	0.4266
LSTM	**0.936**	**0.934**	**0.934**

Выводы по результатам экспериментов

- Качество обучения всех CNN моделей достаточно высоко, среднее значение f-score – 0.9619,
- Лучшей архитектурой для распознавания звуковых команд оказалась DenseNet,
- 7 архитектур CNN продемонстрировали лучшие показатели при обучении на ImageNet, и 7 – со случайными весами,
- LSTM уступает большинству CNN моделей в качестве распознавания, но превосходит в скорости обучения,
- Нейросетевые классификаторы продемонстрировали лучшие показатели.

Выводы

- **Рассмотрены подходы** к решению задачи распознавания звуковых команд.
- **Исследован метод предобработки** аудио на основе мел-спектрограммы.
- **Выполнено аналитическое сравнение** современных методов машинного обучения, включая нейросетевые подходы,
- **Проведены эксперименты**, показавшие перспективность применения нейросетей, обученных на изображениях, для решения задачи распознавания звука.
- Планируется проведение **дальнейших расширенных исследований** с использованием больших выборок данных, содержащих зашумленную звуковую информацию.

Основные источники

1. **Бармина О. К. Решение задачи распознавания звуковой информации с применением методов машинного обучения // Информационно-телекоммуникационные технологии и математическое моделирование и высокотехнологичных систем: материалы Всероссийской конференции с международным участием. Москва, РУДН, 17-21 апреля 2023 г. – М.: РУДН, 2023**

2. Becker Sören, Ackermann Marcel, Lapuschkin Sebastian, Müller Klaus-Robert, and Samek Wojciech. Interpreting and explaining deep neural networks for classification of audio signals // arXiv preprint arXiv:1807.03418. — 2018.

3. Palanisamy Kamalesh, Singhania Dipika, and Yao Angela. Rethinking CNN models for audio classification // arXiv preprint arXiv:2007.11154. — 2020.

4. Huang Gao, Liu Zhuang, Van Der Maaten Laurens, and Weinberger Kilian Q. Densely connected convolutional networks // Proceedings of the IEEE conference on computer vision and pattern recognition. — 2017. — P. 4700–4708.

5. Kaiser Lukasz, Gomez Aidan N, and Chollet Francois. Depthwise separable convolutions for neural machine translation // arXiv preprint arXiv:1706.03059. — 2017.

6. Abadi Martín, Agarwal Ashish, Barham Paul, Brevdo Eugene, Chen Zhifeng, Citro Craig, Corrado Greg S, Davis Andy, Dean Jeffrey, Devin Matthieu, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems // arXiv preprint arXiv:1603.04467. — 2016.

Introduction

The aim of this work is to study modern machine learning methods to be applied to the problem of speech command recognition.

The relevance of the work is due to the emergence of a variety of new approaches for processing and examining audio information and the expansion of its application areas.

Objectives:

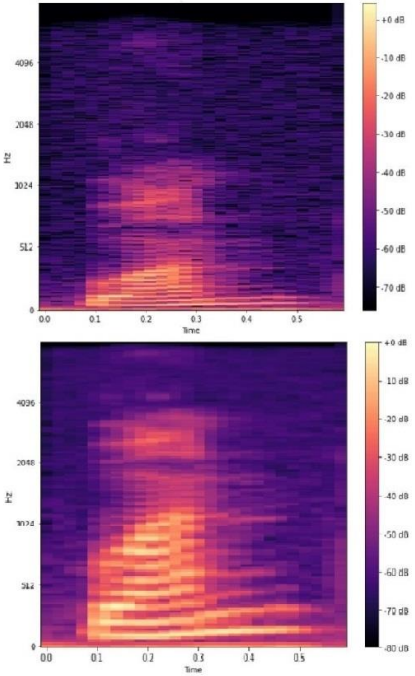
1. Analytical comparison of machine learning methods as applied to the task of audio command recognition;
2. Analysis of audio signal preprocessing methods;
3. Research on a mel-spectrogram-based approach to the problem;
4. Conducting experimental research;
5. Comparison and evaluation of the results obtained.

Experimental research

AudioMNIST dataset: recordings of pronunciation of numbers 0 to 9 in English.

Mel spectrogram:

- Designed to preserve information well-perceived by the human ear.
- Frequency conversion from Hz to mel: $m = 2595 \ln(1 + \frac{f}{700})$.
- Repeat over 3 channels.



Architectures:

- DenseNet
- MobileNet
- MobileNetV2
- MobileNetV3
- Xception
- RegNet
- ShuffleNet
- ShuffleNetV2
- SqueezeNet

Conclusion

- The performance of all CNN models is rather high,
- The best architecture for sound command recognition was DenseNet,
- 7 CNN architectures showed better performance when training on ImageNet, and 7 with random weights,
- LSTM is inferior to most CNN models in recognition quality, but superior in learning speed,
- Neural Network classifiers showed better performance,
- Further studies using large samples of data containing noisy audio information are planned.