

МИНОБРНАУКИ РОССИИ

федеральное государственное бюджетное образовательное учреждение высшего образования

«Санкт-Петербургский государственный технологический институт (технический университет)»

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| УГНС | | 09.00.00 | Информатика и вычислительная техника | | |
| Направление подготовки | | 09.03.01 | Информатика и вычислительная техника | | |
| Направленность (профиль) | |  | Автоматизированные системы обработки информации и управления | | |
| Форма обучения | |  | очная | | |
|  | |  |  | | |
| Факультет | |  | Информационных технологий и управления | | |
| Кафедра | |  | Систем автоматизированного проектирования и управления | | |
| Учебная дисциплина | |  | ИИвАС | | |
| Курс | IV | | | Группа | 413 |

Отчёт по лабораторной работе № 2

Вариант № 11

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Исполнитель: |  |  | |  |  | |
| обучающийся группы 413 |  |  | |  | Мезенцева Анастасия Олеговна | |
|  |  |  | |  | Беналлал Абдель-Малек | |
|  |  |  |  | | |
| Проверил(и): |  |  | |  | Тетерин Михаил Александрович | |
|  |  | (дата, подпись) | |  |  | |
|  |  |  | |  |  | |

Содержание.

[Цель работы 3](#_Toc188706618)

[Ход работы 3](#_Toc188706619)

[1. Теоретическая база 3](#_Toc188706620)

[2. Описание разработанной системы 4](#_Toc188706621)

[3. Результаты работы и тестирования системы 5](#_Toc188706622)

[Выводы 5](#_Toc188706623)

# Цель работы

Научиться работать с предобученными моделями и на основе предобученных эмбеддингов строить новые модели.

Задание:

Скачать предложенный датасет, получить готовые эмбеддинги и построить на их основе нейросеть и оценить качество модели.

|  |  |  |  |
| --- | --- | --- | --- |
| № варианта | Датасет | Эмбеддинги | Источник |
| 11 | 1 | VGGish (по выбору) | Мел-спектрограмма |

# Ход работы

## Теоретическая база

**VGGish** - предварительно обученная нейронная сеть, разработанная Google для извлечения аудио-эмбеддингов. Она основана на архитектуре VGG и обучена на большом наборе аудиоданных. VGGish преобразует мел-спектрограммы в компактные векторные представления (эмбеддинги), которые сохраняют важную информацию о звуке и могут быть использованы для классификации, поиска и других задач анализа аудио. Использование предварительно обученной модели позволяет сократить время обучения и повысить точность классификации.

**Эмбеддинги** (или векторные представления) – способ представления дискретных переменных (например, слов, изображений, аудио) в виде непрерывных векторов. Они позволяют "перевести" категориальные данные в числовой формат, понятный для машинного обучения. Главная идея заключается в том, чтобы объекты с похожим значением или смыслом имели близкие векторные представления в векторном пространстве.

**Мел-спектрограммы** представляют собой визуальное представление аудиосигнала, которое отражает распределение энергии звука по частотам во времени. Шкала мел используется для имитации восприятия звука человеческим ухом, что делает мел-спектрограммы более эффективными для анализа аудио, чем обычные спектрограммы. Они позволяют выделить характерные особенности звука, которые могут быть связаны с определенной тематикой видео (например, наличие музыки, речи, звуков природы и т.д.).

Для определения тематики видео по аудио-эмбеддингам используется **нейронная сеть**. Нейронная сеть обучается на наборе данных, состоящем из аудио-эмбеддингов и соответствующих им тематик видео. В процессе обучения сеть настраивает свои параметры, чтобы максимально точно предсказывать тематику видео по входным эмбеддингам.

## Описание разработанной системы

Алгоритм работы системы:

1. Отбор видео из дата-сета для загрузки. Модуль: generate\_download\_list.py. На вход принимается дата-сет, происходит его обработка: проверка id и наличия видео на устройстве, по результатам проверки происходит разделение дата-сета на три части: скачанные видео, видео с некорректным id, не скачанные видео с корректным id. Для последней части URL видео сохраняются в текстовый файл.
2. Загрузка первой минуты аудиодорожки для каждого видео из текстового файла при помощи yt-dlp. Скачивание только первой минуты позволяет уменьшить занимаемое на диске пространство, однако зачастую именно в первую минуту автор устно определяет тему видео.
3. Извлечение эмбеддингов. Модуль: embeddings.ipynb. Модули vggish\_slim, \_params, \_input, mel\_features принадлежат готовой модели VGGish. Каждый сегмент (0.96с) аудиодорожки описывается 128-мерным вектором, то есть каждая минута аудио представляется тензором 62х128. Каждый сегмент дополнительно разделяется на части по 25мс со смещением 10мс, для каждой части вычисляется спектрограмма, которую и обрабатывает VGGish для получения эмбеддингов.
4. Обучение и тестирование нейросети. Модуль: classification.ipynb. Обучение нейросети происходит на полученных на предыдущем шаге эмбеддингах. Выборка предварительно разбивается случайным образом на тестовую, валидационную и тренировочную. После обучения и тестирования выводится результат: точность и матрица несоответствий (confusion matrix).

Архитектура:

1. Входные данные: первая минута аудиодорожек видео, категории видео
2. Выходные данные: вероятности каждой из четырех тематик видео
3. Модель нейронной сети имеет 6 слоев. Первый слой определяет форму данных, подающихся в модель: эмбеддинги vggish для одной минуты аудио. Второй слой – слой свертки (Conv1D). Размер тензора после слоя 62х32. Третий слой, GlobalAveragePooling1D(), выполняет глобальное усреднение по временной оси, что позволяет уменьшить размерность данных, сохраняя при этом важные характеристики. Размер тензора: 32. Четвертый слой, DropOut, используется для регуляризации модели, случайным образом обнуляя 30% входных данных во время обучения, чтобы предотвратить переобучение. Размер тензора: 32. Следующий слой Dense является полносвязным. Размер тензора: 64. Затем еще один слой DropOut и слой Dense, который также является полносвязным, размер тензора: 4. Функция активации softmax используется для получения вероятностей для каждого класса.
4. Функция потерь: Сategorical Сross-Entropy
5. Оптимизатор: Adam
6. Инфраструктура: yt\_dlp, ffmpeg. Для загрузки wav используется soundfile. Метод подсчета мел-спектрограммы скопирован из примера VGGish.

## Результаты работы и тестирования системы

Вывод в процессе обучения:

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

Результат тестирования представлен точностью (accuracy) и матрицей несоответствий (confusion matrix).

Изображение выглядит как текст, чек, Шрифт, белый

Автоматически созданное описание

Точность модели составила 65%

# Выводы

В ходе работы была обучена и протестирована нейронная сеть для определения категории видео по мел-спектрограмме первой минуты аудиодорожки.

Точность нейронной сети составила 65%. В общем случае точность выше 80% считается хорошей, а ниже 70% требует улучшения модели. Однако даже полученная точность значительно выше 36% (вероятность угадывания при выборе самой распространенной категории видео).

Для повышения точности нейросети можно сделать следующее:

1. Увеличить дата-сет
2. Увеличить длину анализируемой аудиодорожки