Московский государственный технический университет им.

Н.Э. Баумана

Кафедра «Системы обработки информации и управления»

Домашнее Задание

по дисциплине

«Методы машинного обучения»

Выполнил:

студент группы ИУ5И-24М

Лю Бовэнь

Москва — 2024 г.

Задание

Домашнее задание по дисциплине направлено на анализ современных методов машинного обучения и их применение для решения практических задач. Домашнее задание включает три основных этапа:

- выбор задачи;
- теоретический этап;
- практический этап.

Этап выбора задачи предполагает анализ ресурса paperswithcode. Данный ресурс включает описание нескольких тысяч современных задач в области машинного обучения. Каждое описание задачи содержит ссылки на наиболее современные и актуальные научные статьи, предназначенные для решения задачи (список статей регулярно обновляется авторами ресурса). Каждое описание статьи содержит ссылку на репозиторий с открытым исходным кодом, реализующим представленные в статье эксперименты. На этапе выбора задачи обучающийся выбирает одну из задач машинного обучения, описание которой содержит ссылки на статьи и репозитории с исходным кодом. Теоретический этап включает проработку как минимум двух статей, относящихся к выбранной задаче. Результаты проработки обучающийся излагает в

теоретической части отчета по домашнему заданию, которая может включать:

- описание общих подходов к решению задачи;

конкретные топологии нейронных сетей, нейросетевых ансамблей или других моделей машинного обучения, предназначенных для решения задачи;

- математическое описание, алгоритмы функционирования, особенности обучения используемых для решения задачи нейронных сетей, нейросетевых ансамблей или других моделей машинного обучения;

- описание наборов данных, используемых для обучения моделей;

- оценка качества решения задачи, описание метрик качества и их значений;

- предложения обучающегося по улучшению качества решения задачи. Практический этап включает повторение экспериментов авторов статей на основе представленных авторами репозиториев с исходным кодом и возможное улучшение обучающимися полученных результатов. Результаты проработки обучающийся

излагает в практической части отчета по домашнему заданию, которая может включать:

- исходные коды программ, представленные авторами статей, результаты документирования программ обучающимися с использованием диаграмм UML, путем визуализации топологий нейронных сетей и другими способами;

- результаты выполнения программ, вычисление значений для описанных в статьях метрик качества, выводы обучающегося о воспроизводимости экспериментов авторов статей и соответствии практических экспериментов теоретическим материалам статей;

- предложения обучающегося по возможным улучшениям решения задачи, результаты практических экспериментов (исходные коды, документация) по возможному улучшению решения задачи.

Выбранная задача: «Классификация аудио»

Теоретический этап 1. Выбор задачи

Выбранная задача — **Классификация аудио (Audio Classification)**. Это задача машинного обучения, целью которой является присвоение аудиозаписи одной или нескольких предопределенных меток (классов). Примеры включают определение

жанра музыки, распознавание речи, идентификацию говорящего, классификацию звуков окружающей среды (например, сирена автомобиля, лай собаки, шум дождя) или обнаружение неисправностей оборудования по звуку.

Современные подходы к решению этой задачи делятся на два основных типа:

1. **Классические методы машинного обучения**: Используют вручную спроектированные признаки (например, MFCC, ZCR, спектральный центроид) в связке с такими моделями, как SVM или случайный лес.
2. **Глубокое обучение**: Используют нейронные сети для автоматического извлечения признаков из аудиоданных. Входными данными могут служить как необработанные аудиосигналы (1D), так и их двумерные представления, такие как спектрограммы или мел-спектрограммы, что позволяет применять архитектуры, изначально разработанные для компьютерного зрения.

Наиболее популярными наборами данных для этой задачи являются UrbanSound8K, ESC-50 и AudioSet. Качество моделей обычно оценивается с помощью таких метрик, как точность (Accuracy), F1-мера и средняя точность (mAP).

2. Исследуемые статьи

Статья 1: Environmental Sound Classification with Convolutional Neural Networks

Теоретическая часть

Описание задачи

Классификация звуков окружающей среды является ключевой задачей для создания интеллектуальных систем, способных понимать акустическую обстановку. В выбранной статье "Environmental Sound Classification with Convolutional Neural Networks" автора Karol Piczak исследуется применение сверточных нейронных сетей (CNN) для этой задачи, что стало одним из фундаментальных подходов в данной области.

Основные концепции: от аудио к изображению

Ключевая идея подхода заключается в преобразовании одномерного аудиосигнала в двумерное представление, подобное изображению, что позволяет использовать мощные и хорошо изученные архитектуры CNN.

- **Спектрограмма**: Это визуальное представление спектра частот сигнала во времени. Она создается с помощью

кратковременного преобразования Фурье (STFT), которое разбивает сигнал на короткие временные отрезки и вычисляет для каждого из них распределение частот.

- **Мел-спектрограмма**: Это спектрограмма, в которой частотная ось преобразована в мел-шкалу. Человеческое ухо лучше различает изменения в низких частотах, чем в высоких, и мел-шкала имитирует эту особенность. Мел-спектрограммы являются стандартом де-факто для задач анализа аудио с помощью глубокого обучения.

Структура сети

Предложенная в статье архитектура CNN является относительно простой и состоит из следующих слоев:

1. **Входной слой**: Принимает на вход мел-спектрограмму.
2. **Сверточные слои**: Два или три сверточных слоя с функцией активации ReLU. Каждый слой извлекает иерархические признаки: от простых контуров и текстур на нижних слоях до более сложных акустических паттернов на верхних.
3. **Слои подвыборки (Pooling)**: Используются для уменьшения пространственной размерности карт признаков, что

делает модель более устойчивой к небольшим сдвигам и искажениям в спектрограмме.

4. **Полносвязные слои**: После выравнивания (flattening) карт признаков следуют один или два полносвязных слоя.

5. **Выходной слой**: Слой с функцией активации Softmax, который выдает вероятностное распределение по всем классам.

Математическое описание и алгоритмы

- **Преобразование Фурье**: $X(k)=\sum_{n=0}^{N-1} x(n) e^{-i 2\pi k n / N}$
- **Функция потерь**: Для задачи многоклассовой классификации используется категориальная перекрестная энтропия: $L=-\sum_i y_i \log(\hat{y}_i)$
- **Оптимизация**: Обучение сети производится с помощью алгоритма обратного распространения ошибки и оптимизаторов, таких как Adam или SGD.

Описание наборов данных

- **UrbanSound8K**: Содержит 8732 аудиоклипа длительностью до 4 секунд, разделенных на 10 классов городских звуков (например, сирена, гудок автомобиля, работающий двигатель).

- **ESC-50**: Содержит 2000 аудиоклипов длительностью 5 секунд, разделенных на 50 классов звуков окружающей среды (животные, природные звуки, человеческие звуки и т.д.).

Оценка качества

Основной метрикой является точность (accuracy) — доля правильных классификаций. Также используются F1-мера и матрица ошибок (confusion matrix) для анализа того, какие классы модель путает чаще всего. В оригинальной работе была достигнута точность около 73% на наборе данных UrbanSound8K.

Предложения по улучшению

- **Аугментация данных**: Применение техник аугментации, специфичных для аудио (изменение высоты тона, растяжение времени, добавление фонового шума), для увеличения разнообразия обучающих данных и повышения обобщающей способности модели.
- **Более глубокие архитектуры**: Использование более сложных и глубоких CNN-архитектур, таких как VGG, ResNet или EfficientNet, предварительно обученных на изображениях (ImageNet) и адаптированных для работы со спектрограммами.

Статья 2: PANNs: Large-Scale Pre-trained Audio Neural Networks for Audio Pattern Recognition

Теоретическая часть

Описание задачи

Подобно тому, как в компьютерном зрении модели, предварительно обученные на огромных наборах данных (например, ImageNet), значительно улучшают качество на специфических задачах, в области аудио появился аналогичный подход. Статья "PANNs: Large-Scale Pre-trained Audio Neural Networks for Audio Pattern Recognition" от Kong et al. представляет собой прорыв в этой области, предлагая набор мощных предварительно обученных моделей для широкого круга задач аудиоанализа.

Основные концепции

- **Трансферное обучение (Transfer Learning)**: Основная идея заключается в том, чтобы обучить глубокую нейронную сеть на очень большом и разнообразном наборе аудиоданных, а затем использовать эту модель (или ее часть) в качестве основы для решения других, более узких задач, где доступно меньше данных. Модель, обученная на большом датасете, изучает общие и устойчивые акустические признаки.

- **Крупномасштабное предварительное обучение**: Модели PANNs были обучены на наборе данных **AudioSet** от Google, который содержит более 2 миллионов 10-секундных аудиоклипов из YouTube, размеченных 527 различными классами звуков.

Структура сети

Архитектура PANNs основана на вариантах CNN. Одна из самых популярных моделей, CNN14, имеет следующую структуру:

1. **Сверточный блок**: Состоит из нескольких сверточных слоев для извлечения признаков из входной мел-спектрограммы.
2. **Остаточные блоки (Residual Blocks)**: Как и в ResNet, используются остаточные соединения для обучения очень глубоких сетей без проблемы затухания градиента.
3. **Механизм внимания (Attention Mechanism)**: После извлечения признаков применяется механизм внимания. Он позволяет модели "сосредоточиться" на наиболее важных временных и частотных участках спектрограммы для каждого конкретного звукового события, игнорируя фоновый шум или нерелевантные звуки.
4. **Пулинг**: Результаты агрегируются с помощью глобального и/или взвешенного по вниманию пулинга.

5. **Выходной слой**: Полносвязный слой с сигмоидной активацией для многоклассовой/многозначной классификации (один звук может принадлежать нескольким классам).

Математическое описание и алгоритмы

- **Функция потерь**: При предварительном обучении на AudioSet используется бинарная перекрестная энтропия (BCE) для каждого из 527 классов, поскольку один аудиоклип может содержать несколько звуковых событий.
- **Стратегии обучения**: Используется сложный планировщик скорости обучения и аугментация данных, включая Mixup, когда два аудиоклипа и их метки смешиваются в определенной пропорции для создания нового обучающего примера.

Описание наборов данных

- **AudioSet**: Основной набор данных для предварительного обучения.
- **Задачи для fine-tuning (донастройки)**: ESC-50, UrbanSound8K, и другие. Модель PANNs, после предварительного обучения, донастраивается на этих меньших

наборах данных, как правило, путем замены или переобучения только последнего классификационного слоя.

Оценка качества

- **mAP (mean Average Precision)**: Основная метрика для оценки на многозначном наборе данных AudioSet.
- **Точность (Accuracy)**: Используется для оценки на задачах с одной меткой, таких как ESC-50, где PANNs установили новый state-of-the-art результат, превысив 98% точности.

Предложения по улучшению

- **Адаптация к домену**: Исследование методов дальнейшей адаптации предварительно обученных моделей к специфическим акустическим доменам (например, медицинские звуки, промышленный шум).
- **Гибридные модели**: Комбинирование PANNs с рекуррентными нейронными сетями (RNN) или трансформерами для лучшего моделирования временных зависимостей в длинных аудиозаписях.

Практическая часть

Подготовка данных

В практической части будет использоваться набор данных UrbanSound8K. Процесс подготовки включает скачивание набора данных, его разделение на обучающую, валидационную и тестовую выборки. Каждый аудиофайл будет преобразован в мел-спектрограмму фиксированного размера с использованием библиотеки librosa.

Обучение моделей

1. **Обучение простой CNN с нуля**:
    - **Реализация**: Будет создана модель CNN, аналогичная описанной в первой статье, с использованием фреймворка PyTorch или TensorFlow/Keras.
    - **Процесс**: Модель будет обучаться на сгенерированных мел-спектрограммах. Будут применены базовые техники аугментации данных. Процесс обучения будет отслеживаться по графикам потерь и точности на валидационной выборке.
2. **Донастройка (Fine-tuning) модели PANNs**:
    - **Реализация**: Будет загружена предварительно обученная на AudioSet модель PANNs (например, CNN14).
    - **Процесс**: Большинство слоев модели будут "заморожены" (их веса не будут обновляться). Последний

классификационный слой будет заменен новым, соответствующим 10 классам набора UrbanSound8K. Затем модель будет донастроена на обучающей выборке UrbanSound8K в течение нескольких эпох с низкой скоростью обучения.

Оценка модели

После завершения обучения обе модели будут оценены на тестовой выборке. Будут вычислены метрики точности и F1-меры для каждого класса. Будут построены матрицы ошибок для визуального сравнения того, какие классы каждая модель путает. Ожидается, что донастроенная модель PANNs покажет значительно более высокое качество по сравнению с моделью, обученной с нуля.

Вывод

Анализируя теоретические и практические аспекты, можно заключить, что простые сверточные сети, обученные с нуля, способны решать задачу классификации аудио с приемлемым качеством, особенно если доступен достаточный объем данных. Однако подход, основанный на трансферном обучении с использованием крупномасштабных предварительно обученных моделей, таких как PANNs, является значительно более эффективным.

Он позволяет достичь state-of-the-art результатов даже при ограниченном количестве данных для целевой задачи, используя знания, полученные из огромного и разнообразного набора аудиоданных.

3. Список использованных источников

[1] Piczak, K. J. (2015). Environmental Sound Classification with Convolutional Neural Networks. In 2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP).

[2] Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W., & Plumbley, M. D. (2020). PANNs: Large-Scale Pre-trained Audio Neural Networks for Audio Pattern Recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 28, 2880-2894.

[3] Hershey, S., et al. (2017). CNN Architectures for Large-Scale Audio Classification. In 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

[4] Gemmeke, J. F., et al. (2017). Audio Set: An ontology and human-labeled dataset for audio events. In 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

[5] Salamon, J., Jacoby, C., & Bello, J. P. (2014). A Dataset and Taxonomy for Urban Sound Research. In Proceedings of the 22nd ACM international conference on M