СВЁРТОЧНЫЕ НЕЙРОННЫЕ СЕТИ ДЛЯ ОБРАБОТКИ РЕЧИ

Прокопеня А.С.

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Азаров И.С. – доцент, д. т. Н..

В данной работе рассматривается применение свёрточных нейронных сетей для задачи распознавания речи. Предложен метод распознавания изображения спектрограммы звукового сигнала с помощью свёрточных нейронных сетей. Реализованы алгоритмы для предварительной обработки входных данных. Оценено качество распознавания, проведено сравнение двух предложенных моделей распознавания речи посредством применения свёртоных нейронных сетей.

В работе рассмотрены свёрточные нейронные сети (СНС) для обработки речи, т. к. сети такого вида в настоящее время показывают один из лучших результатов в области распознавания изображений. Такая особенность СНС позволяет рассматривать не временную реализацию звукового сигнала (речи), а ее спектрограмму, которая распознается как изображение. Важной особенностью свёрточных нейронных сетей является устойчивость к изменениям масштаба и смещениям изображения, что, очевидно, имеет место в случае спектрограмм речевых сигналов [2].

Первую модель:

Чтобы реализуемая свёрточная нейронная сеть показывала необходимые результаты, необходимо сформировать несколько директорий, в которых будут храниться данные об аудио сигналах:

1.Директория тренировочных данных.

2.Директория проверочных данных.

3.Директория тестовых данных.

С целью минимизации ошибки, обучение требуется проводить на массиве данных от 1000 до 10000 [2]. В данной модели использовалась выборка данных 100 изображений спектрограмм для тренировки, 50 изображений для проверки. В качестве функций активации в разных слоях использовались «сигмоида» и ReLu. Разберем структуру сети:

- Сверточный слой, выделяющий 32 признака;
- Ядро 3×3;
- Функция активации ReLU;
- Слой подвыборки (MaxPooling) с ядром 2×2;
- Слой преобразования массива в вектор;
- Полносвязный слой из 32 нейронов, функция активации ReLU;
- Слой Dropout против переобучения;
- Полносвязный слой из 10 нейронов; активационная функция «сигмоида».

Результаты исследования модели

Для получения статистических данных были проведены эксперименты в которых создавалось по 5 нейронных сетей для 30,40,50 эпох с количество свертоных слоев от 1 до 4. Представленная нейронная сеть показала точность около 80 – 98% в тестовой выборке.

Вторая модель:

Вторая модель была спроектирована для задачи распознавания эмоций по голосу. Применение таких нейросетей находит себя во множестве областей профессиональной деятельности человека: разработке вспомогательных роботов, автономных транспортных средств, оборудовании для нейро-обратной связи и т. д. [1].

Для начала необходимо провести предобработку звука. Для этого необходимо преобразовать звуковую волну в цифровой вид. Для этого выполняется процедура дискретизации звуковой волны. После выполнения операции дискретизации на выходе будет получен массив чисел, каждое из которых представляет амплитуду звуковой волны через интервалы 1/8000 секунды.

В качестве особенностей исходного сигнала используются MFCC (Mel-Frequency Cepstral Coefficients, мел-частотные кепстральные коэффициенты) [1]. Для извлечения MFCC необходимо разложить звуковую волну на отдельные составляющие. Далее необходимо вычислить логарифм мощности сигнала в каждом мелокне и произвести дискретное косинусное преобразование.

Построение нейронной сети и ее обучение

Свёрточная нейронная сеть имеет следующую архитектуру:

1. Первый уровень обработки представляет собой два свёрхточных слоя и max pooling с 128 фильтрами размера 5, имея на выходе 216×128×2 нейронов.

2. Второй уровень обработки состоит из трех слоев свёртки также с 128 фильтрами. На выходе второго слоя модель имеет 216×128×3 нейронов.

3. Последний уровень обработки представляет собой полносвязный слой, который производит классификацию по 10 эмоциям. Структура нейронной сети представлена на рисунке 1.

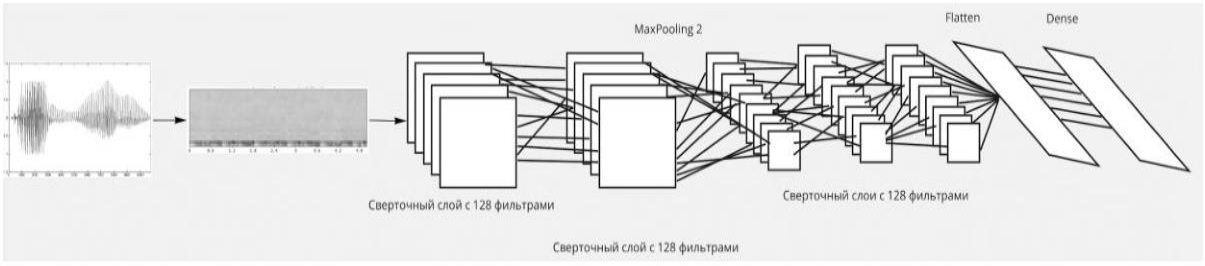

Рисунок 1 – структура свёрточной нейронной сети

Результаты:

В качестве входных данных использовался набор данных от RAVDESS [1], состоящий из аудиозаписей 24 актеров, которые содержат различные эмоции: счастье, злость, грусть, отвращение, спокойствие, удивление. Данный набор данных был разделен на тренировочные и тестовые данные в пропорции 80/20. В результате моделирования точность данной системы составила 73%.

В результате проведенного анализа, а также сравнения выбранных моделей свёрточных нейронных сетей для задачи распознавания речи получены следующие результаты: точность первой модели варьируется от 80%-98%, вторая же модель показывает результат с точностью в 73%. Из этого следует, что, подход к построению свёрточной нейронной сети предложенный в первой модели является более точным.

Список использованных источников:

1. Научная электронная библиотека [Электронный ресурс]. – Режим доступа: https://www.elibrary.ru/item.asp?id=40872772. – Дата доступа: 17.04.2021.

2. Белоруцкий Р.Ю., Житник С.В. Распознавания речи на основе свёрточных нейронных сетей // Вопросы радиоэлектроники. 2019.№4. С. 47-52.