ПРИМЕНЕНИЕ ПОЛНОСВЯЗНЫХ НЕЙРОННЫХ СЕТЕЙ В ЗАДАЧАХ РАСПОЗНАВАНИЯ ЭМОЦИЙ В РЕЧИ

Краснопрошин Д.В.

Белорусский государственный университет информатики и радиоэлектроники

Минск, Республика Беларусь

Вашкевич М.И. – доктор техн. наук

Аннотация. Экспериментально исследуется возможность применения полносвязных нейронных сетей для классификации эмоций в человеческой речи. Представлен вариант реализации классификатора на основе трехслойной полносвязной нейронной. Показано, что полученная модель позволяет определять эмоции с точностью до 48.5%.

Ключевые слова. нейронные сети, глубокое обучение, распознавание, цифровая обработка сигналов, машинное обучение.

Введение

Одной из существенных задач, связанных с разработкой эффективного человеко-машинного взаимодействия, является создание интерфейса, который бы максимально приближался к естественным условиям. Решение этой задачи требует, чтобы компьютер мог воспринимать текущую ситуацию и реагировать соответственно. Важным аспектом такого восприятия является умение компьютера понимать эмоциональное состояние пользователя.

Среди основных способов выражения человеческих эмоций важная роль отводится его речи. За последние годы было проведено большое количество исследований по распознаванию (классификации) эмоций на основе речи [1-2].

Существуют различные варианты решения данной проблемы. В частности, можно отметить подходы, основанные на использовании нейронных сетей, линейного дискриминантного классификатора, метода опорных векторов и т. д. [1].

В данной работе предлагается подход для классификации человеческих эмоций с использованием полносвязной нейронной сети.

Набор данных

Для исследования использовался набор данных Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [3].

Данный набор включает записи от 24 актеров (12 мужчин и 12 женщин), представленных по 104 высказывания на каждого актера (60 речевых и 44 песенных). В рамках нашей работы мы ограничились использованием речевых высказываний, что дало нам доступ к 1440 аудиофайлам в формате WAV (16 бит, 48 кГц). RAVDESS содержит в себе различные эмоциональные состояния, такие как нейтральность, спокойствие, счастье, грусть, гнев, страх, удивление и отвращение. Следует отметить, что эмоциональные состояния были представлены на двух уровнях громкости, что способствует более эффективному обучению моделей в условиях повседневной разнообразной эмоциональной динамики в реальных сценариях общения.

Извлечение признаков

В данной работе анализ речевых характеристик базировался на использовании мел-частотных кепстральных коэффициентов (МЧКК) [2]. Процесс вычисления МЧКК относится к методам кратковременного анализа речевого сигнала, которые включают разбиение сигнала на короткие фреймы или сегменты. В финальный набор исходных признаков включались следующие характеристики: среднее значение МЧКК (34 признака), среднеквадратичное отклонение MFCC (34 признака), среднее значение первой и второй производных от МЧКК (68 признаков), их среднеквадратичное отклонение (68 признаков), а также коэффициент асимметрии, эксцесс и межквантильный размах (по 34 признака для каждой характеристики соответственно). Таким образом, для каждого аудиофайла мы получаем 306-компонентный вектор признаков МЧКК. Процесс извлечения признаков проиллюстрирован на рисунке 1.

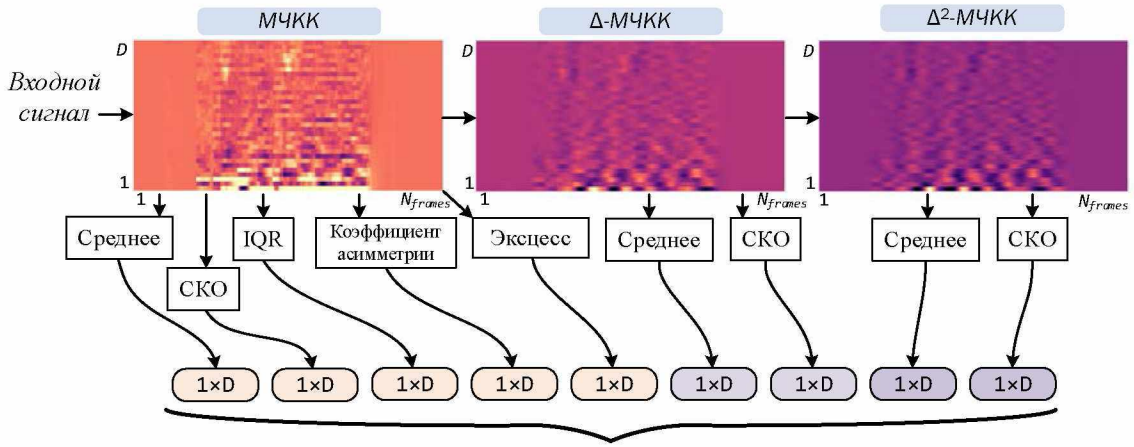

Рисунок 1 – Схема формирования вектора признаков

Разработка классификатора на основе нейронной сети

В данном исследовании был разработан и оценен классификатор эмоций в речи на основе полносвязной нейронной сети, реализованной с использованием библиотеки PyTorch.

Была построена 3-х слойная полносвязная нейронная сеть:

- **первый слой** (исходный вектор признаков, 256 выходных признаков); Функция активации: ReLU

- **второй слой** (256 входных признаков, 128 выходных признаков); Функция активации: ReLU

- **третий слой** (128 входных признаков, 8 выходных признаков (предсказания вероятности классов)). Функция активации: SoftMax. В качестве функции оптимизации использовался градиентный метод Adam.

Стоит отметить, что функция активации ReLU (Rectified Linear Unit) – одна из наиболее популярных функций активации в нейронных сетях, определяется следующим выражением:

$$ReLU(x) = max(0, x), \tag{1}$$

где x – входной сигнал. Функция ReLU просто возвращает входное значение, если оно положительное, иначе возвращает ноль. Это позволяет модели эффективно обучаться и избежать проблемы затухающего градиента.

Функция активации $softmax$ широко используется в многоклассовой классификации, где модель должна предсказать «вероятности» принадлежности к различным классам. Её математическая формула выглядит следующим образом:

$$softmax(x_i) = \frac{exp(x_i)}{\sum_j exp(x_j)} \tag{2}$$

где x_i – входной сигнал для класса i, а $softmax(x_i)$ – «вероятность» принадлежности к классу i. $Softmax$-функция преобразует вектор входных значений $\mathbf{x} = [x_1, x_2, \ldots, x_C]$ в «вероятности», где сумма всех вероятностей для всех классов равна единице. Это позволяет модели легко интерпретировать результаты как вероятности классов.

Для оптимизации параметров сети был выбран алгоритм Adam. Метод оптимизации Adam (*Adaptive Moment Estimation*) – это алгоритм оптимизации, который сочетает в себе идеи из алгоритмов градиентного спуска и адаптивного шага. Согласно методу Adam необходимо вычислить сглаженную (усредненную) версию градиента

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t, \tag{3}$$

где g_t – градиент функции потерь, вычисленный на шаге t, m_t – сглаженный градиент на шаге t, β_1 – параметр, отвечающий за «скорость» усреднения. Аналогичным образом вычисляется сглаженная версия среднеквадратичного градиента функции потерь:

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2, \tag{4}$$

где v_t – среднеквадратичное значение градиента функции потерь на шаге t, β_2 – параметр, отвечающий за «скорость» усреднения значений g_t^2. Обновление параметров нейронной сети θ_t выполняется по следующему правилу [4]:

$$\theta_t = \theta_{t-1} - \frac{\eta}{\sqrt{v_t + \epsilon}} m_t. \quad (5)$$

η – скорость обучения, ϵ – малое число, вводимое для численной стабильности.

В оригинальной статье, где был представлен метод Adam рекомендуется брать следующие значения параметров: $\beta_1 = 0{,}9$, $\beta_2 = 0{,}999$, $\epsilon = 10^{-8}$.

В качестве функции потерь для обучения классификатора была выбрана перекрестная энтропия, так как она хорошо подходит для многоклассовых задач, включая классификацию эмоций. Её математическая формула выглядит следующим образом:

$$CrossEntropyLoss(y, \hat{y}) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} y_{ij}\, log(\hat{y}_{ij}). \quad (6)$$

где y – истинные метки классов в виде унитарного кода, $\hat{y}$ – предсказанные вероятности классов моделью, N – количество примеров в выборке, C – количество классов. Функция (3) вычисляет ошибку между истинными метками и предсказанными вероятностями для каждого класса. Она стремится минимизировать расхождение между распределением вероятностей истинных меток и предсказанных вероятностей модели.

Оценка классификатора

Для итоговой оценки качества модели вычисляли среднее арифметическое (невзвешенное) полноты (*unweighted average recall, UAR*). UAR – это показатель, используемый для измерения общей производительности модели многоклассовой классификации, вычисляет средний уровень запоминания по всем классам, придавая каждому классу одинаковую важность без учета классового дисбаланса:

$$UAR = \frac{1}{N_c}\sum_{i=1}^{C}\frac{A_{ii}}{\sum_{j=1}^{C} A_{ij}} \quad (7)$$

где A – матрица спутанности (confusion matrix).

Значение UAR находится в диапазоне от 0 до 1.

Эксперимент проводили в три этапа:

1) подготовка обучающей выборки;

2) обучение и тестирование классификатора;

3) оценка модели с использованием метрики UAR.

Для оценки производительности классификатора использовался метод перекрестной проверки по k-блокам (k-fold cross-validation). В данной работе данных были разбиты на блоки следующим образом (в скобках указаны номера актеров):

- блок 0: (2, 5, 14, 15, 16);
- блок 1: (3, 6, 7, 13, 18);
- блок 2: (10, 11, 12, 19, 20);
- блок 3: (8, 17, 21, 23, 24);
- блок 4: (1, 4, 9, 22).

Такой порядок разбиения был предложен в [3]. Выбранная стратегия заключается в том, что каждый блок должен содержать одинаковое количество случайно выбранных образцов для каждого класса. При этом должно выполняться условие, что каждый актер представлен либо в обучающей, либо в валидационной выборке, но не в обоих.

Результаты

В результате построения и обучения модели был получен классификатор, точность предсказаний которого при использовании тестового набора данных и вышеуказанной метрики качества достигала 48.5%.

На рисунке 2 показана мультиклассовая матрица спутывания представляющая собой таблицу или диаграмму, показывающая точность прогнозирования классификатора в

отношении двух и более классов. Ячейки таблицы заполняются количеством прогнозов классификатора. Правильные прогнозы идут по главной диагонали от верхнего левого угла в нижний правый.

Рисунок 2 – Мультиклассовая матрица спутывания

Вывод

Анализ полученных результатов показал, что небольшая полносвязная нейронная сеть способна справляться с распознавание эмоций в речи. Тем не менее, для более комплексных входных данных (большее количество актеров разного пола и возрастов) этого метода может оказаться недостаточно. В связи с этим для решения обозначенной задачи, возможно, следует попробовать более сложные модели. Таковыми, например, являются скрытые Марковские модели, сверточные нейронные сети и долговременная память (особая разновидность архитектуры рекуррентных нейронных сетей, способная к обучению долговременным зависимостям), поскольку они лучше отражают временную динамику, включенную в речь человека.

Список использованных источников:

1. L. Chen, X. Mao, Y. Xue, and L. Cheng, “Speech emotion recognition: Features and classification models,” Digital Signal Processing. Vol. 22, No. 6, pp. 1154-1160, 2012.
2. D. Ververidis and C. Kotropoulos, “Emotional speech recognition: Resources, features, and methods,” Speech Communication. Vol. 48, No. 9, pp. 1162-1181, 2006.
3. Multimodal Emotion Recognition on RAVDESS Dataset Using Transfer Learning/ C. Luna-Jiménez, D. Griol, Z. Callejas, R. Kleinlein, J.M. Montero, F. Fernández-Martínez // Sensors. – 2021. – vol. 21. – pp. 1 – 29.
4. Николенко, С., Кадурин, А., Архангельская, Е. Глубокое обучение – СПб.: Питер, 2019. – 480 с.
5. Kingma D. P., Ba J. Adam: A method for stochastic optimization //arXiv preprint arXiv:1412.6980. – 2014.