УДК 004.934.2+534.784

ПРИМЕНЕНИЕ РЕКУРЕНТНЫХ НЕЙРОННЫХ СЕТЕЙ С ДОЛГОЙ КРАТКОСРОЧНОЙ ПАМЯТЬЮ ДЛЯ ЗАДАЧИ РАСПОЗНАВАНИЯ ЭМОЦИЙ ПО РЕЧИ

Краснопрошин Д.В., аспирант кафедры ЭВС

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

Вашкевич М.И. – докт. техн. наук, доцент

Аннотация. Экспериментально исследуется возможность применения рекуррентных нейронных сетей с долгой краткосрочной памятью нейронных сетей для классификации эмоций в речи. Представлены варианты реализации классификатора на основе рекуррентных сетей с одним, двумя и тремя скрытыми слоями. В качестве исходных речевых признаков использовались мелчастотные кепстральные коэффициенты. Наилучший результат классификации показала трехслойная LSTM-сеть, продемонстрировав значением метрики точности UAR=32%. Также предложены варианты улучшения модели, которые, потенциально могут улучшить качество классификации.

Ключевые слова. Нейронные сети, глубокое обучение, распознавание эмоций в речи, рекуррентные нейронные сети, обработка речевого сигнала.

Введение

Распознавание эмоций по речи является актуальной задачей в области обработки естественного языка и вычислительной психологии. В последние годы были предложены различные подходы, основанные, как на статистических методах (классическое машинное обучение), так и на нейросетевых архитектурах, для решения этой задачи. Одним из таких методов является использование рекуррентных нейронных сетей (РНС) с долгосрочной краткосрочной памятью (LSTM), которые эффективны при работе с временными рядами данных, такими как аудиофайлы. В этой работе мы экспериментально исследуем возможность применения LSTM для задачи распознавания эмоций в речи с использованием набора Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [1, 2].

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. В финальный набор исходных признаков включались нормализованные коэффициенты МЧКК (34 признака).

Разработка модели классификации на основе LSTM

В данном исследовании были разработаны и оценены классификаторы эмоций в речи на основе LSTM, реализованных с использованием библиотеки PyTorch.

Так как речевой сигнал представляет собой последовательность, не менее широкое распространение в задачах распознавания эмоций получили различные модификации РНС. Формально взаимосвязь речевого сигнала и РНС можно описать следующим образом. Речевой сигнал $s(t)$ представляет собой непрерывную функцию времени. В процессе обработки он дискретизируется с частотой f_s, что даёт последовательность дискретных отсчётов:

$$s[n] = s(nT_s), T_s = \frac{1}{f_s}. \quad (1)$$

В контексте задачи распознавания эмоций на вход РНС обычно подаются временные последовательности акустических признаков, характеризующих речевой сигнал. К таким признакам, как было упомянуто выше, относятся и МЧКК.

Так, с помощью оконного преобразования Фурье и вычисления акустических признаков (в данном случае МЧКК), речевой сигнал преобразуется в последовательность векторов:

$$X = [x_0, x_1, \dots, x_T], x_t \in R^d, \quad (2)$$

где сигнал T – длина последовательности, а d – размерность признакового пространства.

РНС моделируют временные зависимости в последовательности X с помощью скрытых состояний h_t, которые передаются между слоями нейронной сети:

$$h_t = \sigma(W_h h_{t-1} + W_x x_t + b_h), \quad (3)$$

где $h_t \in R^m$ – скрытое состояние на шаге t, $W_h \in R^{m \times m}$ и $W_x \in R^{m \times d}$ – матрицы весов, $b_h \in R^m$ – вектор смещения, $\sigma(\cdot)$ – функция активации [3].

Сеть обучается так, чтобы скрытые состояния h_t содержали информацию о предыдущих входных значениях, позволяя моделировать контекст в речевом сигнале.

На основе скрытых состояний РНС формируются выходные предсказания, например, вероятность принадлежности речевого фрагмента к определённому эмоциональному классу:

$$y_t = g\left(V_h h_t + b_y\right), \tag{4}$$

где V_h – матрица весов, $g(\cdot)$ – функция активации, b_y – вектор смещения.

Схематическое представление РНС показано на рисунке 1.

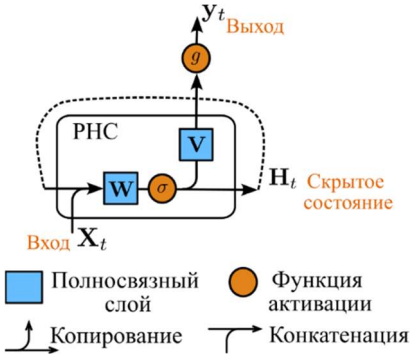

Рисунок 1 – Архитектура простой РНС

LSTM является улучшенной разновидностью РНС, которая решает проблему исчезающего градиента, позволяя эффективно моделировать долгосрочные зависимости в последовательных данных [3]. Математически, LSTM-ячейку можно описать набором уравнений для вычисления состояния и выхода на каждом временном шаге t:

$$\begin{gathered} f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right)(forgetgate), \\ i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)(inputgate), \\ \hat{C}_t = tanh(W_C \cdot [h_{t-1}, x_t] + b_C)(candidatememorycell), \\ C_t = f_t \odot C_{t-1} + i_t \odot \hat{C}_t(memorycell), \\ o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)(outputgate), \\ h_t = o_t \odot tanh(C_t)(output), \end{gathered} \tag{4}$$

где f_t, i_t, o_t – забывающее, входное и выходное затворы, соответственно, C_t – состояние памяти ячейки на текущем шаге, h_t – выход сети на текущем шаге, $\sigma itanh$ – сигмоидная и гиперболическая тангенс-функции активации, $\odot$ – операция поточечного умножения, W, b – обучаемые веса и смещения.

В работе предлагается модель, которая представляет собой однонаправленную многослойную LSTM-сеть. На вход подаётся последовательность признаков, представленных нормированными МЧКК (34 признака), которая обрабатывается рекуррентным блоком LSTM с заданным числом слоёв и размерностью скрытого состояния.

Далее предлагается два способа формирования итогового вектора признаков. В первом случае, после прохождения через LSTM извлекается последнее скрытое состояние последнего слоя, которое используется в качестве вектора признаков. Во втором случае, предлагается конкатенация последних скрытых состояний, полученных на каждом слое LSTM.

Итоговый вектор признаков передается через слой dropout (метод регуляризации, при котором случайные нейроны временно "отключаются" во время обучения, чтобы предотвратить переобучение и улучшить обобщающую способность модели) на вход полносвязному слою, а затем через softmax (функция активации, которая преобразует выходные значения модели в вероятностное распределение по классам, обеспечивая, чтобы их сумма была равна единице) для получения распределения вероятностей по классам.

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 распознавания по всем классам, придавая каждому классу одинаковый вес. Значение UAR находится

в диапазоне от 0 до 1. Для оценки производительности классификатора использовался метод перекрестной проверки по k-блокам (k-fold cross-validation). Подробно с метрикой оценки качества, а также схемой разбиения данных на блоки можно ознакомиться в [2].

Описание эксперимента

Одной из ключевых гипотез данного исследования является влияние скрытых состояний нижних уровней LSTM на итоговую точность классификации. Для проверки гипотезы использовались различные архитектуры сети с разным количеством слоев LSTM и числом скрытых нейронов. После чего анализировалось, как изменения в структуре модели влияет на результаты классификации.

Для проведения сравнительного анализа использовалось два вида вектора признаков, сформированных LSTM: 1) скрытое состояние последнего слоя и 2) конкатенация последних скрытых состояний, полученных на каждом слое. Это необходимо для проверки вышеупомянутой гипотезы.

Отдельное внимание уделено инициализации весов и смещений. Веса входов нейронной сети и скрытых состояний РНС инициализируются с помощью методов Ксавье (начальная установка весов нейронной сети, который сохраняет дисперсию сигналов на входе и выходе слоя, способствуя стабильному и быстрому обучению) и ортогональной инициализации (способ задания начальных весов РНС, при котором матрица весов формируется как ортогональная, что помогает сохранить стабильность градиентов при обучении глубоких рекуррентных моделей) соответственно.

Также в ходе экспериментов установлено, что модели с инициализацией смещений равных единице для забывающих гейтов и нулями для других слоев показывают заметное улучшение производительности по сравнению с моделями, в которых все смещения инициализируются одинаково (например, нулями или случайными значениями). Предполагается, что связано это с улучшением способности сети управлять забыванием важных временных зависимостей в данных.

Результаты экспериментов

В результате обучения моделей получены классификаторы, точность предсказаний которых при использовании тестового набора данных и вышеуказанной метрики качества достигала 32.10%.

В таблице 1 представлены результаты классификации для различных конфигураций модели (число и размер скрытых слоев, тип итогового вектора признаков.

Таблица 1 – Результаты классификации для различных вариантов РНС

Название	Число слоев	Число скрытых нейронов	Число параметров (данные с последнего слоя)	UAR, последний скрытый слой, %	Число параметров (конкатенация скрытых слоев)	UAR, конкатенация скрытых слоев, %
RNN-l1-h32	1	32	8 968	22,40		
RNN-l1-h64		64	26 120	26,04		
RNN-l1-h128		128	85 000	**31,38**		
RNN-l2-h32	2	32	17 416	23,11	17 672	25,20
RNN-l2-h64		64	59 400	27,08	59 912	28,84
RNN-l2-h128		128	217 096	31,58	218 120	**31,71**
RNN-l3-h32	3	32	25 864	23,96	26 376	23,96
RNN-l3-h64		64	92 680	29,30	93 704	26,24
RNN-l3-h128		128	349 192	**32,10**	351 240	32,03

И таблицы видно, что сконкатенированный вектор признаков, пусть и незначительно, но все же негативно влияет на результаты классификации, и более, того, усложняет архитектуру модели за счет увеличения числа параметров модели. Также из результатов, описанных в таблице 1 можно заключить, что по мере усложнения архитектуры LSTM (добавления числа скрытых слоев и увеличения числа скрытых нейронов) точность классификации увеличивается.

Вывод

В работе показано, что однонаправленные рекуррентные нейронные сети с долгосрочной краткосрочной памятью (LSTM) в их базовых конфигурациях могут использоваться для решения задач распознавания эмоций в речи. Также экспериментально доказана важность правильной инициализации весов и смещений, а также конфигурации гиперпараметров (число скрытых слоев РНС и число скрытых нейронов на каждом слое). Выполнена проверка гипотезы о влиянии скрытых состояний нижних слоев РНС на производительность модели в случае их конкатенации в итоговый вектор признаков. В дальнейшем предлагается попробовать улучшить предложенную архитектуру сети с помощью механизма внимания. Добавление механизма внимания потенциально может позволить модели сфокусироваться на наиболее информативных фрагментах временной последовательности, усиливая вклад ключевых моментов речи в распознавание эмоций и тем самым повышая точность классификации.

Список использованных источников:

1. 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.

2. Краснопрошин Д. В., Вашкевич М. И. Метод распознавания эмоций в речевом сигнале с использованием машины опорных векторов и надсегментных акустических признаков // Доклады БГУИР. – 2024. – Т. 22. – №. 3. – С. 93-100.

3. Николенко, С., Кадурин, А., Архангельская, Е. Глубокое обучение – СПб.: Питер, 2019. – 480 с.

UDC 004.934.2+534.784

APPLICATION OF RECURRENT NEURAL NETWORKS WITH LONG SHORT-TERM MEMORY FOR SPEECH EMOTION RECOGNITION TASKS

Krasnoproshin D.V. PhD Student at the Department of Electronic Computing Facilities

Belarusian State University of Informatics and Radioelectronics, Minsk, Republic of Belarus

Vashkevich M.I. – Dr.Sc.

Annotation. The possibility of applying recurrent neural networks with long-term memory for classifying emotions in human speech is experimentally studied. Variants of classifier implementation based on recurrent networks with one, two and three hidden layers are presented. Mel-frequency cepstral coefficients (MFCCs) were used as initial speech features. A three-layer LSTM network showed the best classification results, with an accuracy metric (UAR) of 32%. Possible improvements to the model have also been proposed, which could potentially enhance the quality of the classification.

Keywords. Neural networks, deep learning, speech emotion recognition, recurrent neural networks, speech signal processing.