УДК 004.934.2+534.784

ПРИМЕНЕНИЕ LSTM МОДЕЛЕЙ С МЕХАНИЗМОМ ВНИМАНИЯ И ОСТАТОЧНЫМИ СВЯЗЯМИ ДЛЯ ЗАДАЧИ КЛАССИФИКАЦИИ ЭМОЦИЙ В РЕЧИ

Краснопрошин Д.В., аспирант

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

Вашкевич М.И. – д-р. техн. наук, доцент

Аннотация. Экспериментально исследуется возможность применения LSTM с механизмом внимания и остаточными связями для классификации эмоций в речи. В качестве исходных речевых признаков использовались мелчастотные кепстральные коэффициенты и хромограммы. Наилучший результат классификации показала ResLSTM-SA-сеть со скрытой размерность в 128 нейронов, продемонстрировав значением метрики точности UAR=56.4%. Также предложены варианты улучшения модели, которые, гипотетически могут улучшить качество распознавания.

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

Введение. Классификация эмоций по речевому сигналу является не до конца нерешенной актуальной задачей в области обработки естественного языка. Существуют различные подходы и методы решения данной задачи, основанные, как на классических алгоритмах машинного обучения (статистические методы), так и на нейросетевых моделях. Одним из таких методов является использование рекуррентных нейронных сетей (РНС) с долгосрочной краткосрочной памятью (LSTM), которые эффективны при работе с временными рядами данных, коими и являются аудио записи. В этой работе мы экспериментально исследуем возможность применения LSTM с механизимом внимания и остаточными связями для задачи распознавания эмоций в речи с использованием датасета IEMOCAP (Interactive emotional dyadic motion capture database (Интерактивная база данных для захвата эмоциональных диадических движений)) [1].

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. Для улучшения разделительных свойст были также использованы хромограммы. Извлечение хромограмм - это метод представления звука в обработке аудио, который переносит весь частотный спектр в 12 контейнеров (bin), соответствующих 12 полутонам музыкальной октавы. Этот метод отбрасывает информацию об октаве, фокусируясь исключительно на гармоническом содержании (высоте тона), что делает его устойчивым к изменениям тембра и инструментовки [3]. В финальный набор речевых признаков включались нормализованные коэффициенты МЧКК (34 признака) и хромограммы (12 признаков).

Разработка модели классификации на основе LSTM. В данном исследовании были разработаны и оценены классификаторы эмоций в речи на основе LSTM с механизмом внимания, представленной в [4].

Стоит отметить, что одним из основных недостатков классической LSTM модели является то, что ее внутренняя структура не предполагает механизма приоритизации отдельных элементов временного контекста.

Для того, чтобы решить данную проблему авторами была предложена модификация путем внедрения механизма мягкого внимания. Его добавление позволяет модели вычислять веса «важности» для различных временных шагов, что способствует получению более информативного выходного вектора признаков, на основании которого выполняется классификация эмоции [4].

В данной работе мы предлагаем модификацию архитектуры LSTM с механизмом мягкого внимания путем включения дополнительного слоя LSTM с остаточными связями. Этот слой обогащает временное представление входных признаков до их обработки основным LSTM на основе механизма внимания, тем самым повышая способность модели улавливать долгосрочные зависимости. Более того, для повышения информативности признакового пространства были также использованы хромограммы, которые отсутствуют в оригинальной работе [4].

Схема предлагаемой архитектуры показана на рисунке. 1.

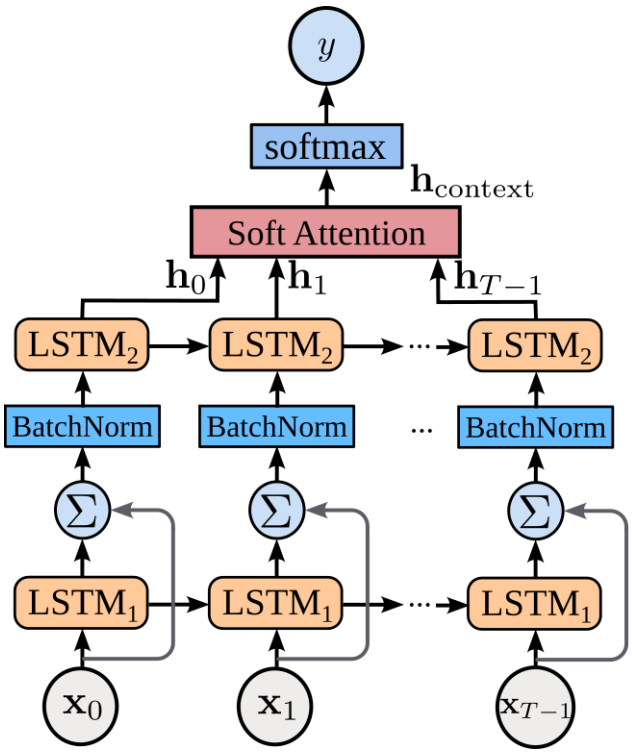

Рисунок 1 – Схематическое представление LSTM с механизмом мягкого внимания и остаточными связями

Отличительной особенностью предлагаемой архитектуры является то, что размерность скрытого состояния первого слоя LSTM соответствует размерности входных признаков (d = 46). Такая конструкция обеспечивает остаточные связи, которые позволяют модели обогащать входную последовательность контекстной информацией, сохраняя при этом исходную структуру признаков.

Остаточные связи, благодаря согласованности своей размерности с размерностью входных данных в первом слое LSTM, позволяют аддитивно объединять исходные признаки с их контекстно обогащенными аналогами. Это сохраняет исходную информацию, позволяя при этом основанной на механизме внимания LSTM сосредоточиться исключительно на извлечении эмоциональных паттернов более высокого порядка.

Оценка классификатора. Для оценки качества классификатора использовалось 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 более объективной оценки также использовался метод перекрестной проверки по k-блокам (k-fold cross-validation). Это было реализовано с помощью протокола Leave-One-Session-Out (LOSO). LOSO- это широко используемый, строгий протокол перекрестной проверки для оценки моделей распознавания эмоций в речи на наборе данных IEMOCAP, гарантирующий тестирование модели на дикторах, с которыми она никогда не сталкивалась во время обучения.

Описание эксперимента. Одной из ключевых гипотез данного исследования является влияние размерности скрытого состояния LSTM на итоговую точность классификации. Для проверки гипотезы использовалась предложенная архитектура сети с разным числом скрытых нейронов. После чего анализировалось, как изменения в структуре модели влияет на результаты классификации.

Для инициализации весов и смещений нейронной сети был применен метод Ксавье (начальная установка весов нейронной сети, который сохраняет дисперсию сигналов на входе и выходе слоя, способствуя стабильному и быстрому обучению). В дополнение к этому была применена ортогональная инициализация. Это способ задания начальных весов LSTM, при котором матрица весов формируется как ортогональная, что помогает сохранить стабильность градиентов.

Результаты экспериментов. В результате обучения моделей получены классификаторы, точность предсказаний которых при использовании тестового набора данных и вышеуказанной метрики качества достигала 56.40%.

В таблице 1 представлены результаты классификации для различных конфигураций модели (число и размер скрытых слоев).

Таблица 1 – Результаты классификации для различных вариантов LSTM с механизмом мягкого внимания

Название	Число скрытых нейронов	Число параметров (данные с последнего слоя)	UAR, последний скрытый слой, %
ResLSTM-SA-h32	32	28 000	54,6
ResLSTM-SA-h32	64	46 800	55,3
ResLSTM-A-h128	128	108 900	**56,40**

Из результатов, описанных в таблице 1 можно заключить, что по мере усложнения архитектуры LSTM (добавления числа скрытых слоев) точность классификации увеличивается.

Вывод. В работе показано, что LSTM с механизмом мягкого внимания и остаточными связями могут использоваться для решения задач распознавания эмоций в речи. Также экспериментально доказана важность правильной инициализации весов и смещений, а также конфигурации гиперпараметров (размерность скрытого состояния). Выполнена проверка гипотезы о влиянии размерности скрытого состояния на производительность модели. В дальнейшем предлагается попробовать усилить предложенную архитектуру сети с помощью помощью использования большего числа векторов внимания, а также поиска других оптимальных гиперпараметров (скорость обучения, коэффициент затухания весов, размер батча и др). Увеличение количества векторов внимания и нахождение оптимальных гиперпараметров модели предположительно может позволить модели сфокусироваться на наиболее информативных фрагментах временной последовательности и тем самым повысить точность распознавания. Более того, планируется предобучить предложенную модель на более простых датасетах, таких как TESS и RAVDESS, что по задумке авторов, может существенно улучшить качества финального классификатора.

Список использованных источников:

1. Busso, C.; Bulut, M.; Lee, C.-C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J. N.; Lee, S. & Narayanan, S. S. (2008), 'IEMOCAP: interactive emotional dyadic motion capture database.', Language Resources and Evaluation 42 (4), 335-359.

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. D. P. Ellis, "Chroma feature analysis and synthesis," Online Resource, 2007. [Online]. Available: https://www.ee.columbia.edu/ dpwe/resources/matlab/chroma- ansyn/

4. Mirsamadi S., Barsoum E., Zhang C. Automatic speech emotion recognition using recurrent neural networks with local attention. Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 05–09 March 2017. New Orleans, 2017, pp. 2227–2231.

UDC 004.934.2+534.784

APPLICATION OF LSTM MODELS WITH ATTENTION MECHANISM AND RESIDUAL CONNECTIONS FOR SPEECH EMOTION RECOGNITION TASKS

Krasnoproshin D.V. graduate student

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

Vashkevich M.I. – Dr.Sc.

Annotation. The feasibility of using an LSTM model with an attention mechanism and residual connections for classifying emotions in speech is experimentally investigated. Mel-frequency cepstral coefficients (MFCCs) and chroma features were used as initial speech features. The ResLSTM-SA network with a hidden size of 128 neurons demonstrated the best result, yielding UAR metric = 56.4%. Model enhancements that could potentially improve performance are also proposed.

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