РЕКУРРЕНТНАЯ НЕЙРОННАЯ СЕТЬ НА ОСНОВЕ LSTM-БЛОКОВ С МЕХАНИЗМОМ ВНИМАНИЯ ДЛЯ РАСПОЗНАВАНИЯ ЭМОЦИЙ В РЕЧИ

Краснопрошин Д. В., Вашкевич М. И.
Кафедра ЭВС, Белорусский государственный университет информатики и радиоэлектроники
Минск, Республика Беларусь
E-mail: daniil.krasnoproshin@gmail.com, vashkevich@bsuir.by

В работе предлагается механизм мультивекторного мягкого внимания для рекуррентных нейронных сетей на основе LSTM для задачи распознавания эмоций в речи. Эксперименты проводились на наборе данных RAVDESS. Для автоматизированного подбора оптимальных гиперпараметров сети использовался метод байесовской оптимизации. Результаты экспериментов показывают, что увеличение количества векторов внимания с одного до 64 приводит к улучшению среднего значения метрики UAR на 0.9%, что является статистически значимым результатом и подтверждает целесообразность использования предложенного механизма внимания в архитектурах на основе LSTM для задачи классификации эмоций в речи.

Введение

Распознавание эмоций в речи является одной из ключевых задач в области обработки естественного языка. В настоящее время продолжается активный поиск архитектур нейронных сетей, которые имеют умеренную сложность и высокую точность работы [1–3].

Целью данного исследования является разработка мультивекторного механизма мягкого внимания и количественная оценка данного механизма на эффективность LSTM-моделей при распознавании эмоций в речи. Предполагается, что увеличение числа векторов внимания позволит модели учитывать более разнообразные аспекты эмоционального контекста.

I. Система распознавания эмоций на основе РНН

Для параметризации речевого сигнала в работе используются мел-частотные кепстральные коэффициенты (МЧКК). Сигнал разделяется на короткие фреймы (кадры) после чего из каждого фрейма извлекается n-мерный вектор МЧКК [2]. В результате речевой сигнал преобразуется в последовательность векторов:

$$X = [\mathbf{x}_0, \mathbf{x}_1 \dots \mathbf{x}_{T-1}], \mathbf{x}_t \in \mathbb{R}^n, \tag{1}$$

где T – длина последовательности.

За основу в работе берется архитектура РНС с мягким механизмом внимания, предложенная в [3]. Входная последовательность подается в РНС типа LSTM [4], которая формирует последовательность скрытых состояний $\mathbf{h}_0, \mathbf{h}_1, \dots, \mathbf{h}_T$. Далее рассчитывается взвешенная сумма скрытых состояния, которая называется вектором контекста:

$$\mathbf{h}_{wp} = \sum_{t=0}^{T-1} \alpha_t \mathbf{h}_t, \tag{2}$$

где α_t – весовые коэффициенты отражающие значимость вектора состояния $\mathbf{h}_t$ в формировании вектора контекста.

На основании вектора контекста $\mathbf{h}_{wp}$ выполняется классификация эмоции. Т. е. вектор $\mathbf{h}_{wp}$ подается на полносвязный слой с активационной функцией softmax. Общая схема описанного подхода представлена на рис. 1.

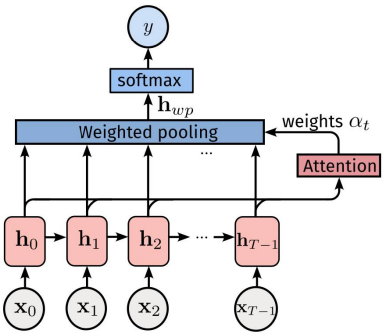

Рис. 1 – Классификация эмоций при помощи РНС с механизмом внимания

В оригинальной работе [3] для формирования весов использовался механизм мягкого внимания (англ. *soft attention*):

$$a_t = \text{softmax}(e_t) = \frac{\exp(e_t)}{\sum_{t=0}^{T-1} \exp(e_t)}, \tag{3}$$

где $e_t = \mathbf{u}^T \mathbf{h}_t$ – оценка внимания (англ. *attention score*), где $\mathbf{u}$ – вектор внимания (обучаемый параметр).

II. Мультивекторный механизм мягкого внимания

В настоящей работе предлагается мультивекторный механизм мягкого внимания, согласно которому оценка внимания рассчитывается по формуле

$$e_t = \text{maxpool}_{1D}\left(\begin{bmatrix} \mathbf{u}_1^T \\ \mathbf{u}_2^T \\ \dots \\ \mathbf{u}_{N_{att}}^T \end{bmatrix} \mathbf{h}_t\right), \tag{4}$$

где N_{att} – число векторов внимания, а $\mathbf{u}_i$ – набор векторов внимания, каждый из которых отвечает

за независимое взвешивание векторов скрытых состояний.

Следует отметить, что обычный механизм мягкого внимания является частным случаем мультивекторного механизма внимания, когда $N_{att} = 1$.

III. Описание эксперимента

Для экспериментальной проверки мультивекторного механизма внимания использовалась однослойная РНС с LSTM-блоками. Входные данные представляют собой нормализованные МЧКК размерностью 1×34, размерность скрытого состояния РНС выбиралась равной 64. Количество векторов внимания изменялось в диапазоне от одного до 64: $N_{att} \in \{1, 2, 4, 8, 16, 32, 64\}$.

Для автоматизированного поиска оптимальных гиперпараметров (скорость обучения, размера батча, вероятность дропаута, скорость «затухания» весов и количество циклов в планировщике косинусного отжига) использовался метод байесовской оптимизации, реализованный в библиотеке Optuna языка Python.

Оптимизация выполнялась для базовой модели с одним вектором внимания. Полученные оптимальные параметры использовались далее для всех экспериментов с различным числом векторов внимания.

Эксперименты проводились на наборе данных RAVDESS [1], включающем аудиозаписи с восьмью эмоциональными категориями. Для каждой конфигурации модели обучение повторялось 10 раз с различными начальными значениями весов. Для оценки качества использовался показатель UAR (*Unweighted Average Recall*). Результаты экспериментов приведены в табл. 1.

Табл. 1. Результаты экспериментов

Model	UAR	UAR (max)
LSTM-v-1	0,5510 ± 0,0071	0,5625
LSTM-v-2	0,5566 ± 0,0104	0,5742
LSTM-v-4	0,5548 ± 0,0065	0,5618
LSTM-v-8	0,5593 ± 0,0079	0,5710
LSTM-v-16	0,5586 ± 0,0112	**0,5775**
LSTM-v-32	0,5581 ± 0,0109	0,5729
LSTM-v-64	**0,5598 ± 0,0083**	0,5742

На рис. 2 представлены сглаженные распределения значений UAR для моделей с одним и 64-мя векторами внимания. Модель с большим числом векторов внимания демонстрирует более высокие значения UAR.

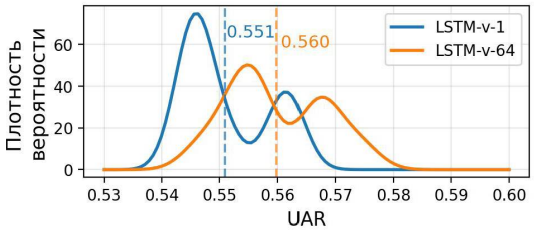

Рис. 2 – Сглаженные распределения точностей моделей: LSTM-v-1 и LSTM-v-64

Статистический анализ различий между моделями проводился с использованием t-теста для независимых выборок. Полученные результаты показали наличие статистически значимой разницы ($p < 0{,}05$) между моделью с одним и 64 векторами внимания. Среднее улучшение UAR составило 0,9%.

В целом, рост точности можно объяснить тем, что каждый вектор внимания учится фокусироваться на различных аспектах эмоциональных признаков – интонации, тембре, спектральной энергии и т. д. Таким образом, использование нескольких таких векторов создаёт более богатое представление входного сигнала, повышая способность модели выделять эмоционально релевантные сегменты речи.

Несмотря на то что абсолютное улучшение метрики невелико (порядка 1%), оно является стабильным и статистически подтверждённым. Это указывает на потенциал дальнейших исследований в данном направлении.

IV. Заключение

В работе проведено исследование влияния количества векторов внимания на качество распознавания эмоций в речи с использованием LSTM-блоков. Эксперименты на датасете RAVDESS показали, что увеличение числа векторов внимания до 64 приводит к статистически значимому росту метрики UAR. Данный результат подтверждает, что мультивекторный механизм мягкого внимания способствует более эффективному обучению эмоциональных представлений в речевом сигнале.

V. Список литературы

1. C. Luna-Jiménez, et al., “Multimodal emotion recognition on RAVDESS dataset using transfer learning,” Sensors, vol. 21, 2021, pp. 1-29.
2. D. V. Krasnoproshin and M. I. Vashkevich “Speech emotion recognition method based on support vector machine and suprasegmental acoustic features,” Doklady BGUIR, vol. 22, no. 3, 2024, pp. 93-100. (In Russain)
3. S. Mirsamadi, E. Barsoum and C. Zhang “Automatic speech emotion recognition using recurrent neural networks with local attention, Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017, pp. 2227-2231.
4. S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural Computation, vol. 9, no. 8, 1997, pp. 1735–1780.