УДК 004.89

МАСКИРОВАННОЕ МУЛЬТИМОДАЛЬНОЕ ВНИМАНИЕ ДЛЯ РАСПОЗНАВАНИЯ ЭМОЦИЙ И ОЦЕНКИ ВАЛЕНТНОСТИ

Волошина Т.А. (Университет ИТМО)
Научный руководитель – к.т.н. Махныткина О.В (Университет ИТМО)

Исследования выполнены за счет финансирования университета ИТМО в рамках НИР №622281 «Разработка методов и алгоритмов для мультимодального распознавания валентности высказываний и доминантности дикторов в полилогах».

Введение. Эмоции играют важную роль в жизни и взаимодействии людей. Они характеризуют наше отношение к тому или иному объекту или событию. Если сгруппировать эмоции по категориям, то можно определить общую эмоциональную направленность выражения, или другими словами, валентность. В условиях огромного количества данных компании заинтересованы в автоматизации сбора отзывов из социальных сетей [1] о предоставляемых товарах и услугах, а также для создания рекомендательных систем. Автоматический анализ валентности может также применятся в сферах медицины и психологии, для контроля эмоционального состояния. Например, в статье [2] показано создание бота-терапевта на основе распознавания эмоций. В статьях [3, 4] приводятся примеры различного применения чат-ботов для отслеживания психологического здоровья, способных определять, например, депрессию.

Основная часть. Цель данной работы — построение модели для автоматической оценки эмоций и валентности на основе аудио, видео и текстовой информации и использовании маскированного внимания модели BERT [5] с целью улучшения качества распознавания. Существует множество подходов для автоматического распознавания эмоций и валентности в тексте и речи. Предварительная обработка каждой из модальностей, разновидности представления информации, используемая модель машинного обучения - все факторы значительно влияют на конечный результат распознавания эмоций.
Для извлечения признаков из изображений выделяются лица с помощью библиотек Dlib и OpenCV. Каждый 5-ый кадр добавляется в группы длиной 10 с перекрытием (окном) между группами в 5 кадров. Аудио признаки представлены в виде объединенного набора MFCC [6] и GeMAPs [7] признаков. Текст преобразован в эмбеддинги с помощь BERT модели.
Признаки по каждой модальности преобразовывались к единой размерности с помощью сверточных слоев (CNN для текста и аудио, LSTM для видео). Полученные наборы объединены в единый вектор признаков, на который накладывается маска внимания от BERT модели, полученной при обработке текста. Таким образом, аудио и видео модальности расширяют текстовую модальность.
Для обучения и тестирования используются следующие наборы данных: MOSI [8], CMU-MOSEI [9], MELD [10], IEMOCAP [11], которые содержат видео записи монологов и диалогов из сериала и открытых источников. Разметка данных сопровождается текстовой аннотацией и содержит оценку по валентности или эмоциональности.

Выводы. Результатами данной работы являются реализованные модели для распознавания эмоций и оценки валентности, обученные на каждом датасете. Также были улучшены state-of-the-art результаты по датасетам CMU-MOSEI, MOSI и MELD в задачах распознавания эмоций и оценки валентности при гораздо более простой архитектуре модели по сравнению с SOTA реализациями. Для сравнения использовались метрики взвешенная точность (weighted accuracy) и взвешенный F1.

Список использованных источников

1. Jansen, J. Twitter Power: Tweets as Electronic Word of Mouth. / & Zhang, M., Sobel, K., Chowdury, A. - doi:2169-2188. 10.1002/asi.21149. // JASIST. - 2009. - № 60.
2. Zygadło, A. Text-based emotion recognition in english and polish for therapeutic chatbot. / Kozłowski, M., & Janicki, A. - doi:10.3390/app112110146. // Applied Sciences (Switzerland), 11(21). - 2021.
3. Abd-alrazaq, A. A. An overview of the features of chatbots in mental health: A scoping review. / Alajlani, M., Alalwan, A. A., Bewick, B. M., Gardner, P., & Househ, M. - doi:10.1016/j.ijmedinf.2019.103978 // International Journal of Medical Informatics. - 2019. - № 132.
4. Valstar, M. AVEC 2016: Depression, Mood, and Emotion Recognition Workshop and Challenge. / & Pantic, M. & Gratch, J. & Schuller, B. & Ringeval, F. & Lalanne, D. & Torres Torres, M. & Scherer, S. & Stratou, G. & Cowie, R. - doi:3-10. 10.1145/2988257.2988258. - 2016.
5. Kaicheng Yang. Cross-Modal BERT for Text-Audio Sentiment Analysis. / Hua Xu, Kai Gao. - doi:10.1145/3394171.3413690. - 2020.
6. S. Davis. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. / P. Mermelstein. - doi: 10.1109/TASSP.1980.1163420. // IEEE Transactions on Acoustics, Speech, and Signal Processing. vol. 28. no. 4. pp. 357-366. - 1980.
7. Eyben, Florian. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. / Scherer, Klaus & Schuller, Björn & Sundberg, Johan & Andre, Elisabeth & Busso, Carlos & Devillers, Laurence & Epps, Julien & Laukka, Petri & Narayanan, Shrikanth & Truong, Khiet. - 10.1109/TAFFC.2015.2457417. // IEEE Transactions on Affective Computing. 7. 1-1. - 2015.
8. Zadeh, Amir. MOSI: Multimodal Corpus of Sentiment Intensity and Subjectivity Analysis in Online Opinion Videos. / Zellers, Rowan & Pincus, Eli & Morency, Louis-Philippe. - 2016.
9. Zadeh, AmirAli. Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph. 2236-2246. / Liang, Paul & Poria, Soujanya & Cambria, Erik & Morency, Louis-Philippe. - doi:10.18653/v1/P18-1208. - 2018.
10. Poria, Soujanya. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. / Hazarika, Devamanyu & Majumder, Navonil & Naik, Gautam & Cambria, Erik & Mihalcea, Rada. - 2018.
11. Busso, C. IEMOCAP: Interactive emotional dyadic motion capture database. Language Resources and Evaluation. 42. 335-359. / Bulut, Murtaza & Lee, Chi-Chun & Kazemzadeh, Abe & Mower Provost, Emily & Kim, Samuel & Chang, Jeannette & Lee, Sungbok & Narayanan, Shrikanth. - doi:10.1007/s10579-008-9076-6. - 2008.