Секция «17.3 Искусственный интеллект и анализ данных в космических исследованиях»

Применение методов глубокого обучения для автоматической классификации эмоциональных состояний

Научный руководитель – Шишкин Алексей Геннадиевич

Петров Михаил Александрович
Студент (специалист)
Московский государственный университет имени М.В.Ломоносова, Факультет космических исследований, Москва, Россия
E-mail: pmishha@mail.ru

Введение

Распознавание эмоциональных состояний по аудио- и видеосигналам является актуальной задачей в области интеллектуального анализа данных и систем человеко-машинного взаимодействия. Учет эмоциональной составляющей речи позволяет анализировать не только семантическое содержание высказывания, но и его просодические и интонационные характеристики, что повышает качество оценки психофизиологического состояния человека.

Подобные технологии могут применяться в системах поддержки принятия решений, при оценке когнитивной нагрузки и уровня стресса персонала, а также при анализе рисков ошибок оператора, вызванных эмоциональным состоянием. В частности, при длительных космических миссиях мониторинг эмоционального состояния космонавтов может способствовать раннему выявлению неблагоприятных психологических состояний, потенциально влияющих на качество принимаемых решений и безопасность экипажа.

Цель работы

Целью работы является исследование и сравнительный анализ нейросетевых подходов к распознаванию эмоций по аудио- и видеоданным, а также оценка влияния механизмов внимания и мультимодального объединения признаков на качество классификации эмоционального состояния человека.

Материалы и методы

В качестве экспериментальной базы используется мультимодальный набор данных IEMOCAP (Interactive Emotional Dyadic Motion Capture) [ссылка], содержащий синхронизированные аудио- и видеозаписи с экспертной разметкой эмоциональных состояний.Отметим, что в данных наблюдается дисбаланс распределения классов, что учитывается при построении и оценке эффективности моделей распознавания эмоций.

Предобработка аудиоданных включает приведение сигналов к единому формату, нормализацию и извлечение информативных спектральных признаков на основе мел-частотного анализа. Для видеоданных выполняется извлечение кадров, их масштабирование и нормализация перед подачей в нейронную сеть.

В рамках работы исследуется применимость нейросетевых архитектур, сочетающих сверточные и рекуррентные компоненты для моделирования пространственно-временных

зависимостей в аудиосигнале. Для повышения качества распознавая эмоций используются механизмы локального и глобального внимания, позволяющие модели адаптивно выделять наиболее информативные фрагменты последовательности.

Мультимодальность моделей реализуется посредством нескольких стратегий объединения модальностей, включая прямую конкатенацию признаков и более сложные механизмы их интеграции. В том числе для обеспечения согласованного обучения моделей применяется механизм адаптивной балансировки вкладов модальностей на основе анализа функций потерь и норм градиентов.

Оценка качества классификации проводится с использованием стандартных метрик классификации, включая F1-меру, точность и показатели, основанные на ROC-анализе.

При численных экспериментах использовались следующие модели:

- CRNN на сыром аудиосигнале;
- CRNN с использованием MFCC;
- CRNN с локальным механизмом внимания;
- CRNN с локальным и глобальным вниманием;
- мультимодальная модель (аудио + видео);
- мультимодальная модель с адаптивной балансировкой градиентов.

Результаты

Проведённые численные эксперименты показывают, что использование мел-частотных кепстральных коэффициентов существенно повышает эффективность распознавания по сравнению с обучением на исходном аудиосигнале. Использование механизмов локального и глобального внимания способствует улучшению способности модели выделять релевантные временные участки и учитывать глобальный контекст высказывания.

Совместное использование аудио- видео- модальностей приводит к повышению качества классификации, особенно для эмоциональных состояний со схожими акустическими характеристиками. Применение адаптивной балансировки вкладов модальностей обеспечивает более устойчивое обучение модели и способствует формированию более достоверной информационной характеристики эмоционального состояния.

Полученные результаты свидетельствуют о перспективности применения подобных моделей в задачах мониторинга эмоционального состояния человека, что может быть использовано в том числе в аэрокосмических задачах.

Источники и литература

1) А. Г. Шишкин, «Распознавание эмоционального состояния человека с помощью методов глубокого обучения,» Информатика и системы управления, pp. 45-62, 2022.
2) R. R. S. K. [S. Latif, "Survey of Deep Representation Learning for Speech Emotion Recognition," IEEE Transactions on Affective Computing, Vols. 14, No. 2, pp. 1634-1654, 2023.
3) R. F. B. R. M. E. N. M. A. S. B. Z. S. Trigeorgis G., "ADIEU FEATURES? END-TO-END SPEECH EMOTION RECOGNITION USING A DEEP CONVOLUTIONAL RECURRENT NETWORK," in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016.

4) H. T. G. W. Zhang S., "Speech emotion recognition using deep convolutional neural network and bidirectional long short-term memory," in Proceedings of the 2017 ACM International Conference on Multimedia (ACM MM), 2017.

5) W. C. L. Yan W.-J., "How Fast Are the Leaked Facial Expressions: The Duration of Micro-Expressions," Journal of Nonverbal Behavior, Vols. 37, No. 1, p. 1-19, 2013.

6) He K., «Deep Residual Learning for Image Recognition,» в Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016), Las Vegas, 2016.

7) Poria S., «A review of affective computing: From unimodal analysis to multimodal fusion,» Information Fusion, т. 37, p. 98-125, 2017.

8) Baltrusaitis T., «Multimodal Machine Learning: A Survey and Taxonomy,» IEEE Trans. Pattern Anal. Mach. Intell., T. 1, No 2, p. 423-443, 2019.

9) А. Г. Шишкин, Методы цифровой обработки и распознавания речи, Издательский Дом "Инфра-М", 2023

10) M. B. C. L. A. K. E. M. S. K. J. C. S. L. a. S. N. C. Busso, "IEMOCAP: Interactive emotional dyadic motion capture database," vol. 42, no. 4, pp. 335-359, December 2008.