УДК 004.93

ОПТИМИЗАЦИЯ ПОТРЕБЛЕНИЯ ПАМЯТИ В ЗАДАЧЕ РАСПОЗНАВАНИЯ АУДИОСОБЫТИЙ С ИСПОЛЬЗОВАНИЕМ АНСАМБЛЯ РЕКУРРЕНТНЫХ НЕЙРОННЫХ СЕТЕЙ

Сурков М.К. (ИТМО)

Научный руководитель – кандидат технических наук, Романенко А.Н. (ИТМО)

Введение. На сегодняшний день задача распознавания аудиособытий является актуальной и вызывает большой интерес как со стороны коммерческих компаний, разрабатывающих умные портативные устройства, так и со стороны научного сообщества, которое начало свое активное изучение данной задачи на ежегодной конференции Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop. Она заключается в том, чтобы во входной аудиозаписи длиной 10 с и по заранее зафиксированному списку аудиособытий определить, в каких частях аудио произошло каждое из событий. Отметим, что возникают ситуации, когда несколько событий происходят одновременно. Также существуют случаи, когда одно и то же событие повторяется несколько раз в разных частях записи, например лай собаки. Кроме того, во входной аудиозаписи каких-то событий может и не быть.

Основная часть. На данный момент большинство подходов опирается на архитектуру Convolutional Recurrent Neural Network (CRNN). Данная архитектура состоит из кодировщика [1,2,3], который преобразует входной сигнал в последовательность эмбеддингов. Затем полученные вектора используются в рекуррентной нейронной сети. Именно она совершает предсказания для каждого момента времени каждого аудиособытия. В случае использования компактных кодировщиков рекуррентный декодировщик может содержать подавляющее большинство обучаемых параметров сети. Подобные модели трудно применимы в активно развивающейся индустрии умных устройств (телефоны, часы, колонки), так как интеллектуальные устройства ограничены в вычислительных ресурсах, объеме оперативной и дисковой памяти. Как следствие, оптимизация потребляемых ресурсов нейронными сетями при решении поставленной задачи является актуальным направлением научных исследований. Идея подхода, предлагаемого в данной работе основана на замене одной тяжеловесной рекуррентной сети на ансамбль из N более компактных моделей, содержащих меньшее число обучаемых параметров. Эмбеддинги, полученные с помощью кодировщика обрабатываются каждой моделью из ансамбля. После чего, вектора, полученные на выходе рекуррентных нейронных сетей, конкатенируются и передаются заключительному линейному слою. Эксперименты показали, что предложенный подход позволяет существенно сократить размер модели, сохраняя при этом точность предсказаний.

Выводы. Предложенная модель была обучена и протестирована на данных из задачи DCASE 2024 Task 4 «Sound Event Detection with Weak Labels and Synthetic Soundscapes». Данный подход существенно сокращает размер модели, сохраняя точность предсказаний. Было произведено сравнение точности предсказаний предложенной модели с точностью базовой модели, состоящей из одной рекуррентной нейронной сети в качестве декодировщика.

Список использованных источников:

1. K. Koutini, J. Schl¨uter, H. Eghbal-zadeh, and G. Widmer, “Efficient training of audio transformers with patchout,” 2022.
2. S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, W. Che, X. Yu, and F. Wei, “Beats: audio pre-training with acoustic tokenizers,” in Proceedings of the 40th International Conference on Machine Learning, 2023, pp. 5178–5193.
3. X. Li and X. Li, “Atst: Audio representation learning with teacher-student transformer,” 2022.