УДК 004.89

ОПРЕДЕЛЕНИЕ КАЧЕСТВА РАСПОЗНАВАНИЯ АУДИООБЪЕКТА В ЗАВИСИМОСТИ ОТ МЕТА-ФИЧЕЙ ДАТАСЕТА

Устинова В.Е. (ИТМО)
Научный руководитель – кандидат технических наук, ассистент Ефимова В.А. (ИТМО)

Введение

Обнаружение звуковых событий представляет собой нетривиальную, но крайне востребованную задачу, находящую применение в самых разных областях: от мониторинга животных и диагностики технических неисправностей до анализа медицинских сигналов и распознавания аварийных или охранных звуков. В последние годы были предложены многочисленные модели, способные по аудиозаписи определять, какие звуковые события в ней присутствуют. Тем не менее, такие модели, как правило, ограничены набором классов, на которых они были обучены (как правило, это большой датасет аудиозаписей AudioSet [1]), и не способны напрямую распознавать ранее неизвестные или редкие звуки. В связи с этим задача дообучения существующих моделей на новых звуковых классах остаётся актуальной. Однако запись новых звуков требует ресурсов, и полезно заранее понимать, на какое качество распознавания можно было бы рассчитывать при том или ином количестве аудиозаписей в обучающей выборке, в связи с чем актуальной становится реализация модели, которая может по мета-фичам звука и по количеству аудиозаписей в обучающей выборке определить, какое качество распознавания данного звука будет после дообучения на этом датасете.

Основная часть

Предлагается модель, которая по мета-фичам аудиозаписи и по количеству аудиозаписей в обучающей выборке определяет, какое качество распознавания звука будет после дообучения на этом датасете. В качестве модели распознавания звука используется PANNS — Large-scale Pretrained Audio Neural Networks [2] как одна из наиболее популярных моделей в области аудио-распознавания (на основе сети CNN14, поскольку именно эта архитектура является бейзлайном в большинстве статей и показывает приемлемый баланс между качеством и ресурсоемкостью).

Используется часть PANNS, отвечающая за аудио-теггирование. Метрикой качества распознавания является mean Average Precision (mAP) [3].

В качестве мета-фичей аудиообъекта используется агрегированный по обучающему датасету эмбеддинг PANNS как наиболее информативное сжатое представление звука. Посредством дообучения на различных подвыборках и датасетах и определения качества mAP формируется мета-датасет, на котором обучается финальная модель регрессии.

Таким образом, в рамках эксперимента предлагается следующая последовательность действий: для каждого датасета звуков сначала определяются его мета-признаки, после чего для различных объемов обучающей выборки выполняется дообучение предобученной модели PANNS с целью распознавания целевого звука и последующим вычислением метрики mAP на тестовых данных. На основе полученных результатов формируется мета-датасет, где входными параметрами выступают мета-признаки датасета и размер выборки, а выходом — соответствующее значение mAP, после чего на этих данных производится обучение итоговой регрессионной модели для предсказания ожидаемой эффективности.

Выводы

Разработана модель, прогнозирующая итоговое качество распознавания звука после дообучения на основе мета-признаков аудиоданных и объема обучающей выборки. Предложенный подход позволяет априорно оценивать ожидаемую эффективность распознавания в зависимости от размера набора данных, что существенно упрощает планирование экспериментов.

Разработанный метод открывает перспективы для решения прикладных задач обнаружения новых аудиообъектов, включая биоакустические сигналы, сигналы аудитехногенного происхождения (сирены, тревоги) и прочие.

Литература

1. Gemmeke J. F. et al. Audio set: An ontology and human-labeled dataset for audio events //2017 IEEE international conference on acoustics, speech and signal processing (ICASSP). – IEEE, 2017. – С. 776-780.
2. Kong Q. et al. Panns: Large-scale pretrained audio neural networks for audio pattern recognition //IEEE/ACM Transactions on Audio, Speech, and Language Processing. – 2020. – Т. 28. – С. 2880-2894.
3. Fonseca E. et al. Audio tagging with noisy labels and minimal supervision //arXiv preprint arXiv:1906.02975. – 2019.