Автоматическая аннотация аудио в условиях ограниченного объема данных

Кузьмин Никита Валерьевич

Студент

Факультет ВМК МГУ имени М. В. Ломоносова, Москва, Россия

E-mail: s02170136@gse.cs.msu.ru

Научный руководитель *— Дьяконов Александр Геннадьевич*

Задача автоматической аннотации аудио может быть поставлена как генерация текстового описания для аудиофайла. Эта проблема очень важная, так как решение обычной задачи классификации не способно описывать физические свойства объектов и окружения, высокоуровневые знания (например, «в дверь постучали три раза»).

Алгоритм автоматической аннотации аудио состоит из двух частей: обработки аудио; генерации текста. На первый взгляд описание аудио может показаться очень похожей на аннотацию изображений, но на практике есть различия. Люди могут с легкостью описывать изображения, так как каждый объект имеет свою форму, цвет, размер. Большинство людей лучше знакомы с визуальным представлением информации, чем с аудио представлением. Рассмотрим отличия между стандартным представлением аудиоданных — Мел-спектрограммами и изображениями: в отличие от картинок, оси спектрограмм не несут один и тот же смысл, так как ось X определяет время, а ось Y определяет частоту; Мел-спектрограммы не обладают свойством локальности: это означает, что соседние пиксели не принадлежат одному объекту.

Результаты алгоритма автоматической аннотации аудио могут быть использованы в различных сферах: в медицине для людей с проблемами слуха, на производстве для более аккуратного описания экстренных ситуаций, для более детального анализа видео.

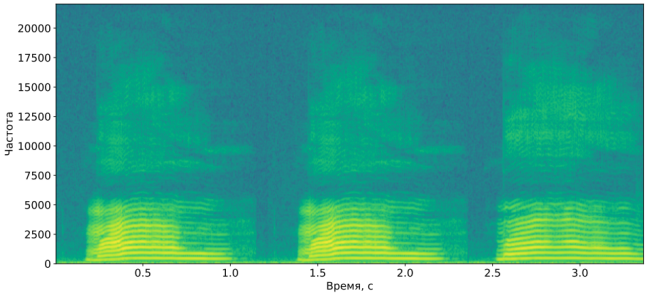
На момент написания этой работы было произведено не так много исследований по данной теме [1, 2, 3]. Объемы данных для обучения моделей, как следствие, сильно ограничены, поэтому основная цель работы заключалась в изучении и реализации различных методов аугментации аудиоданных.

В качестве базового алгоритма для решения задачи автоматической аннотации аудио был предложен нейросетевой подход, основанный на рекуррентных кодировщике и декодировщике [4]. В ходе экспериментов были исследовано множество подходов к аугментации аудиоданных и лучшие результаты получились при использовании

следующих алгоритмов аугментирования данных: Input, Manifold MixUp [5, 6]; overdrive; реверберация; изменение тональности; изменение скорости. С помощью расширения обучающей выборки удалось достигнуть улучшения результатов на 22% относительно базового алгоритма.

Дополнительно был реализован механизм внимания для улучшения базовой модели. Комбинация механизма внимания и аугменаций позволила получить результаты, превосходящие на 34% базовый алгоритм.

Иллюстрации



Пример спектрограммы аудиозаписи

Литература

1. Drossos K., Adavanne S., and Virtanen T., «Automated audio captioning with recurrent neural networks», CoRR, vol. abs/1706.10006, 2017.
2. Ikawa S. and Kashino K., «Neural audio captioning based on conditional sequence-to-sequence model» // Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019), 2019.
3. Wu M., Dinkel H., and Yu K., «Audio caption: Listen and tell» // ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019.
4. Cho K., Van Merrienboer B., Gulcehre C., Bougares F., Schwenk H., and Bengio Y., «Learning phrase representations using RNN encoder-decoder for statistical machine translation» // CoRR, vol. abs/1406.1078, 2014.

5. Zhang H., Cisse M., Dauphin Y. N., and Lopez-Paz D., «mixup: Beyond empirical risk minimization» // CoRR, vol. abs/1710.09412, 2017.
6. Verma V., Lamb A., Beckham C., Najafi A., Mitliagkas I., Courville A., Lopez-Paz D., and Bengio Y., «Manifold Mixup: Better Representations by Interpolating Hidden States» // arXiv e-prints, p. arXiv:1806.05236, 2018.