Рецензия статьи On Generative Spoken Language Modeling from Raw Audio

Авторы статьи: Kushal Lakhotia, Eugene Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Abdelrahman Mohamed, Emmanuel Dupoux

Авторами статьи была представлена новая задача Generative Spoken Language Modeling, нацеленная на извлечение лингвистических и акустических характеристик из сырого аудио (без текстовых данных). Задача делится на два уровня — лингвистическую и акустическую — и два этапа — получения представлений и генерации. Для решения задачи представлен базовый пайплайн из unit2Speech - unitLM - Speech2unit моделей. В статье введены новые метрики для оценки качества генерации аудио, использующие предобученные ASR и LM модели, а также zero shot probe метрики, оценивающие полученные представления.

Сильные стороны:

- Актуальность задачи: предложена новая задача обучения только с сырыми аудио без текста, при достижении полного unsupervised обучения генеративные модели можно будет использовать в языках с маленьким текстовым корпусом;
- Введены новые метрики, высоко коррелированные с человеческой оценкой;
- Метрики для этапов генерации (ASR-based) и получения представлений (zero shot) хорошо коррелируются, значит, ASR-based можно аппроксимировать с помощью легче получаемых zero shot probe метрик.

Слабые стороны:

- Для оценки моделей используются ASR и LM модели, для обучения которых необходим большой текстовый корпус (в статье берутся предобученные модели), на данной стадии пока не получается применить задачу к языкам с маленьким текстовым корпусом;
- При оценке ресинтеза аудио (u2S-S2u) на метрике Phone Error Rate наблюдается domain effect: на датасете LJ значения получаются везде ниже, чем на LibriSpeech.

S2u модели обучались на LJ, u2S — на LS, возможно, можно было бы дополнительно изучить вариант, когда и S2u, и u2S обучались бы на наборах из одного домена;

- При оценке представлений на лингвистическом уровне получаются средние результаты: достигается ошибка в 31.3% (spot-the-word-error rate) на наилучшей модели, а ошибки моделей (в качестве S2u моделей рассмотрены HuBERT, wave2vec и CPC) не сильно отделены от ошибки бейзлайна LogMel;

- Выбор моделей энкодеров: дополнительно можно было бы рассмотреть другие энкодеры, а также другие варианты квантизации эмбеддингов (кроме k-means).

Текст статьи: Для чтения статьи необходимо знать основные понятия из Глубинного обучения в обработке звука; также статья много опирается на другие статьи и даёт описания используемых моделей и метрик на поверхностном уровне.

Воспроизводимость: Авторы предоставляют хороший репозиторий с достаточно подробным описанием, приведены чекпоинты для всех рассмотренных в статье моделей. В приложении статьи описаны параметры обучения u2S модели. Также на `https://speechbot.github.io/gslm/` есть много примеров сгенерированных аудио.

Оценка 7; Уверенность 4