УДК 004.93'11

КОНТРАСТИВНОЕ ПРОГНОЗИРУЮЩЕЕ КОДИРОВАНИЕ ДЛЯ ЗАДАЧИ ИЗВЛЕЧЕНИЯ АКУСТИЧЕСКИХ ПРИЗНАКОВ

Лаптев А.А.
(Университет ИТМО)
Научный руководитель – к.т.н. Меденников И.П.
(Университет ИТМО")

В данном докладе рассматривается методика контрастивного прогнозирующего кодирования для обучения модели извлечения акустических признаков для задачи автоматического распознавания речи. Приводится обзор последних исследований применения методики и анализ качества получаемых признаков для целевой задачи.

Задачу автоматического распознавания речи можно декомпозировать на 3 части: извлечение акустических признаков из исходного цифрового аудиосигнала, преобразование последовательности признаков в промежуточное языковое представление с помощью акустической модели и применение к языковому представлению языковой информации с помощью языковой модели для получения финальной текстовой гипотезы. Наиболее часто используемые акустические признаки включают в себя наборы фильтров в лог-мел шкале и мел-кепстральные коэффициенты. Также часто используются дополнительные признаки, такие как частота основного тона, вектора признаков диктора (ivectors) и другие. Альтернатива вышеперечисленным признакам: обучать нейросетевую модель извлекать значимые признаки из исходного сигнала. За последние 2 года наблюдается прорыв в этом направлении: методика контрастивного прогнозирующего кодирования для обучения извлекающей модели. Данная работа рассматривает особенности этой методики и влияние полученных из обученной модели признаков на качество автоматического распознавания речи.

Контрастивное прогнозирующее кодирование (КПК) – это методика обучения скрытых представлений на манер авторегрессии (модель предсказывает будущий сигнал на основе скрытых представлений прошлых наблюдений). Одно из новшеств КПК в том, что одновременно, с использованием специальной функции потерь, обучаются 2 нейросети: кодировщик, преобразующий входной сигнал в скрытое представление сети, и контекстная модель, предсказывающая будущее представление на основе взаимной информации между сигналом, преобразованным кодировщиком и скрытыми представлениями авторегрессионной модели. Это позволяет обучать модель значительно быстрее, т.к. исходный сигнал содержит слишком много информации. Для предсказания следующего представления используется подвид негативного семплирования. Совсем недавно было предложено улучшение архитектуры модели: дискретизация софтмаксом Гамбела представлений кодировщика, чтобы иметь возможность использовать еще один преобразователь представлений, обученный на манер языковой модели. Такой подход позволяет учитывать глобальный контекст при получении финальных признаков.

Описанная выше методика была протестирована на практике (там, где это представлялось возможным). Было изучено влияние различных архитектур для компонент КПК, необходимое количество данных для обучения и применимость для специфичных задач автоматического распознавания речи.