АВТОМАТИЧЕСКОЕ РАСПОЗНАВАНИЕ СЛОВ ИЗ ОГРАНИЧЕННОГО СЛОВАРЯ НА ОСНОВЕ ВИЗУАЛЬНЫХ ПРИЗНАКОВ

Ткаченко Григорий Станиславович, магистрант кафедры компьютерных технологий Университета ИТМО, grtkachenko@gmail.com

Фильченков Андрей Александрович, к.ф.-м.н., доцент кафедры компьютерных технологий Университета ИТМО, aaafil@mail.ru

Аннотация

В работе была рассмотрена актуальная задача распознавания речи на основе визуальной информации. А именно был спроектирован и реализован алгоритм, который на ограниченном словаре находит произнесенную последовательность слов. Важно отметить, что полученная модель является speaker-independent, то есть она может быть применена к тем спикерам, данные которых не принимали участие в обучении модели.

Введение

В данный момент большое внимание уделяется задаче распознавания речи. Уже трудно представить многие продукты, такие как мобильные ассистенты, навигаторы или развлекательные сервисы без голосового ввода. Однако, в некоторых случаях (например, в метро или на музыкальном концерте) одной звуковой информации бывает недостаточно. И на помощь приходит визуальная информация (например, с фронтальной камеры телефона). В моей работе я обратился именно к этой задаче – распознавание речи на основе визуальной информации.

Поскольку задача довольно сложная, то было принято решение начать с более простой задачи – распознавание речи из ограниченного словаря. При этом полученная модель является speaker-independent, что в некоторых случаях позволит избежать подстройку под конкретного пользователя и упрощает ее последующее использование.

Источники данных

Исследование было начато с поиска данных. Для этой задачи было необходимо подобрать довольно специфические данные – видео с одним человеком, который продолжительно говорит в анфас. Для этого могли бы подойти видео с интервью или выпуск новостей, однако в таких видео

используемый словарь довольно большой, что не подходит под нашу задачу. Поэтому было решено воспользоваться уже существующими датасетами. Выбор был сделан в пользу датасета GRID [1], потому что он содержит большое число спикеров и каждый из них говорит большое количество слов.

Модель распознавания меток слов

Для обучения модели распознавания на полученных данных было необходимо построить некоторое соответствие между изображением и так называемыми виземами – единицами распознавания речи на основе визуальной информации. Для этого необходимо уметь получать вектор признаков по изображению, а также разметку по виземам по каждому фрейму. Рассмотрим каждый из этих этапов по отдельности.

Каждый фрейм видео представляет из себя лицо человека, на котором необходимо найти ключевые точки – координаты губ (пример результаты работы такого алгоритма представлен на Рисунке 1). При этом сейчас существует множество алгоритмов для выделения таких точек. Такие алгоритмы можно разделить на два типа – основанные на выделении контуров и на анализе значений пикселей. Среди алгоритмов первого типа можно выделить AAM [2], ASM [3], а также основанные на подгоне статистической модели формы лица к заданному изображению с помощью градиентного бустинга на регрессионных деревьях [4]. Среди алгоритмов второго типа наиболее яркими являются модели, использующие нейронные сети, в которые передаются значения пикселей из ROI. Выбор был сделан в пользу регрессионных деревьев, поскольку они дают высокую точность и очень высокую скорость работы. В работе использовалась реализация из открытой библиотеки dlib. Важно отметить, что после получения точек с каждого фрейма было выполнена их нормализация и последующее сглаживание, что дало прирост в точности алгоритма. Кроме того, был выполнен так называемый upsampling данных с 25fps до 100fps. Похожий прием широко используются в задачах распознавания речи [5], что также, как будет показано позже, дало прирост в точности алгоритма.

Следующим шагом является разметка обучающей выборки по виземам. Для этого необходимо получить для каждого фрейма соответствующую фонему и после выбрать соответствующую визему. Для первого шага было решено воспользоваться готовой моделью для получения распределения по фонемам, обученной на датасете TIMIT. Для второго шага была выбрана таблица соответствия фонем и визем из [6] соответсвующей работы.

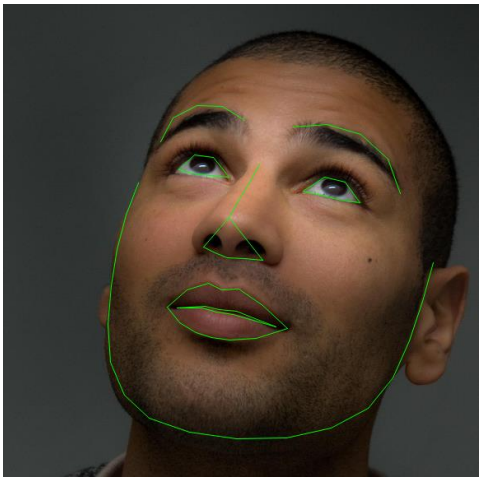

Рисунок 1: Результаты алгоритма распознавания меток слов

Таким образом, полученные на предыдущих шагах данные используются для обучения модели, которая будет соотносить визуальную информацию и виземы. В ходе построения этой модели возникает несколько проблем – если брать данные с одного фрейма, то их будет недостаточно для предсказания нужной виземы. Для этого было решено брать данные с некоторого окна фреймов. Но и в этом случае возникает проблема – данные на соседних фреймах сильно коррелируют друг с другом. Для решения этой проблемы было решено уменьшить размерность. Было взято два известных подхода – с помощью метода главных компонент и с помощью нейронной сети специального вида (autoencoder). По результатам экспериментов оба метода показали примерно одинаковую точность и были сильно лучше наивного подхода. Для эксперимента был взят autoencoder с 1 скрытым слоем размером 72 и для PCA был выбран размер также равный 72. В качестве ориентира для этого параметра использовалось правило, заключающееся в том, что остаются только компоненты с соответствующим собственным числом большим среднего среди всех собственных чисел (т.н. Kaiser rule [7]). Лучший же результат показала модель, которая на вход принимала сглаженные данные (экспоненциальное сглаживание второго порядка). Стоит отметить, что результаты могут быть лучше, если использовать более точную модель распознавания фонем (в нашем случае ее точность

была около 71%).

Невысокие результаты работы алгоритма, возможно, связаны с тем, что на выходе размечаются и первые фреймы слов, для которых нейронной сети не всегда хватает информации, чтобы правильно их разметить. Частично эта проблема может быть решена, если позволить нейронной сети прочитать весь вход, а уже после сгенерировать последовательность слов (sequence-to-sequence learning). Для этого опять же не подходят классические сети из-за фиксированного размера входа, но также и не подходят рекуррентные сети в чистом виде. Возможное решение для этой задачи – применение encoder-decoder LSTM. Это решение, по сути, сначала кодирует всю последовательность в некоторый вектор фиксированной длины (который, фактически, описывает входную последовательность), а после декодирует его в метки слов.

Результаты работы алгоритма приведены в таблице 1. В качестве тренировочной выборки было использовано множество из 30 спикеров, в то время как в тестовой находилось 5 спикеров. При этом записи одного и того же спикера не могли находиться одновременно в обучающей и тестовой выборках. Слово считалось правильно распознанным, если было верно определено общее количество слов во фразе и совпала выходная метка с реальной меткой слова. При том на 97% записей правильно определилось количество слов (на остальных же видео все распознанные вхождения слов считались ошибочными). Приведенная в таблице величина – точность (процент) угаданных меток по правилу выше.

Описание модели	**Точность распознавания**
EncoderHiddenSize = 128 DecoderHiddenSize = 50	73%
EncoderHiddenSize = 100 DecoderHiddenSize = 40	72%
EncoderHiddenSize = 110, dropout = 0.2 DecoderHiddenSize = 60, dropout = 0.1	75%
EncoderHiddenSize = 128, dropout = 0.2 DecoderHiddenSize = 40	76%

Таблица 1: Результаты алгоритма распознавания меток слов

Заключение

В большинстве работ, посвященных распознаванию речи по видеозаписи, авторы приводят качество своих моделей, которые обучены на аудио и визуальных признаках вместе и сравнивают их с моделями, анализирующими только аудио. Поэтому сравнивать работы между собой сложнее, чем, например, при работе со звуковой информацией. Все же существуют исследования, которые работали с тем же датасетом GRID corpus и анализировали возможность чтения по губам. Например, в работе [8] авторам удалось добиться точности 80%, используя только видео-признаки. Однако там все эксперименты были speaker-dependent, то есть тестовая и тренировочная выборки всегда брались с одного и того же спикера, в отличие от данной работы.

Литература

1. Cookea Martin, Jon Barker Stuart Cunningham Xu Shao. An audio-visual corpus for speech perception and automatic speech recognition. — 2006. — Access mode: http://laslab.org/upload/an_audio-visual_corpus_for_speech_perception_and_automatic_speech_recognition.pdf.
2. Timothy F. Cootes Gareth J. Edwards Christopher J. Taylor. Active Appearance Models. — 2008. — Access mode: http://www.comp.hkbu.edu.hk/~ymc/papers/conference/cis08_publication_version.pdf.
3. Le Thai Hoang, Vo Truong Nhat. Face Alignment Using Active Shape Model And Support Vector Machine // CoRR. — 2012. — Vol. abs/1209.6151. — Access mode: http://arxiv.org/abs/1209.6151.
4. Kazemi Vahid, Sullivan Josephine. One Millisecond Face Alignment with an Ensemble of Regression Trees // CVPR. — 2014.
5. Yuxuan Lan Richard Harvey, Theobald Barry-John. Insights into machine lip reading. — 2012. — Access mode: https://pdfs.semanticscholar.org/c573/c71213b46a2b966546c7b7848b5bbe0536ec.pdf.
6. Benedikt Lanthao. Facial Motion: a novel biometric? — 2010.
7. Jackson Donald A. Stopping rules in principal component analysis: a comparison of heuristical and statistical approaches. — 1993.
8. Wand Michael, Koutník Jan, Schmidhuber Jürgen. Lipreading with Long Short-Term Memory // CoRR. — 2016. — Vol. abs/1601.08188. — Access mode: http://arxiv.org/abs/1601.08188.