Mozilla CommonVoice and DeepSpeech

Written by AlephZero for CCExtractor, Google Code-In 2019

CommonVoice

CommonVoice is a project maintained by Mozilla. Its purpose is to create an open-source and multi-language dataset of voices, that can be used in order to train and develop speech recognition software.

How does CommonVoice gather its data? There is a website located at voice.mozilla.org where anyone can contribute, either by speaking a given sentence or by verifying other people's recordings. Once a recording has been verified as correct by at least two separate users, it is added to the dataset.

How large is the dataset? According to Mozilla, CommonVoice is the largest publicly available voice dataset of its kind. Specifically, at the time of writing it contains 29 languages with more in development, and almost 2000 validated hours of audio and corresponding transcriptions. In total the English dataset is 30 gigabytes in size, and the entire dataset is over 60 gigabytes.

Are there any alternative datasets? There are many proprietary and closed datasets companies use for their own speech recognition. However, there are a few more open-source datasets: LibriSpeech (1000h of English speech gathered from audiobooks), TED-LIUM (450h of audio gathered from TED talks), and some other smaller datasets gathered and published by organizations, universities or individuals. Additionally there are some publicly-available datasets that contains only a limited number of words, for example the Google Speech Commands dataset which contains only 10 simple words, and TIDIGITS which contains only digits.

DeepSpeech

DeepSpeech is an open-source speech recognition engine developed and maintained by Mozilla. It is based on a deep-learning model originally introduced in the research paper "Deep Speech: Scaling up end-to-end speech recognition" by Baidu, and uses data from the CommonVoice project for model training.

How does the model work? Mozilla used Tensorflow to develop their model, and based it on Baidu's DeepSpeech research paper, published in 2014. Baidu's system is based on recurrent neural networks, a type of neural network that can feed its outputs back as inputs to previous layers, thus creating a sort of "memory". As a result of this "memory", recurrent neural networks are often better fitted for problems that involve sequential data - such as text, or speech. Baidu's system also involves various optimization techniques, to allow it to run on multiple GPUs effectively thus decreasing training time.

How can you use DeepSpeech? There are two main uses for DeepSpeech:

Using pre-trained models Pre-trained models are speech recognition models already trained by Mozilla, which can be used without training / complicated setup, and are sufficient for most use-cases. DeepSpeech offers bindings for Python and Node.js, and is installed using "pip3 / npm install deepspeech", and optionally "pip3 / npm install deepspeech-gpu" for GPU optimization. Additionally there is a command line client, and third-party bindings for languages like Go and Rust. The pre-trained models can be downloaded from the DeepSpeech GitHub page, here: https://github.com/mozilla/DeepSpeech/releases.

Training custom models If pre-trained models are not sufficient (for example there is a need to use a custom dataset, or an unknown language) it is possible to create and train a custom model using DeepSpeech algorithms. This is done by cloning the GitHub repo at https://github.com/mozilla/DeepSpeech, installing tensorflow and running DeepSpeech.py with the parameters --train_files (training files), --dev_files (validation files), --test_files (testing files).

Are there any alternatives? Yes, there are many speech recognition libraries for Python, here are some of them:

- SpeechRecognition - Very popular open-source Python library for speech recognition, with support for multiple APIs such as Google Cloud Speech, Microsoft Azure Speech, IBM Speech to Text and CMU Sphinx.
- Google Cloud Speech API, Amazon Transcribe, Azure Speech - Very powerful and popular speech recognition APIs, however they are all paid and are part of GCP / AWS / Azure.
- CMU Sphinx - Open-source speech recognition toolkit, which supports continuous speech recognition from a microphone, as well as from files. It is only a toolkit, and it is needed to provide models and dictionaries in order to fully recognize speech.