PROJECT NOTES

DATASET EXPLORATION

Speech Commands Dataset. This contains around 65000 one-second sound files with commands like *Go*, *Yes* or *Stop*.

Reading
http://dkopczyk.quantee.co.uk/speech-nn/
The further information about the dataset can be found here.

Convert data to **spectrograms** or **MFCC** (Mel-Frequency Cepstral Coefficients)

In order to convert raw data to spectrograms we apply Short-time Fourier Transform (STFT)

Keras comes with a very convenient method fit_generator to preprocess data *on the fly*.

USE CNN

If you want to learn how to increase the accuracy of your speech recognition model even more, you can read about mixing Convolution Neural Networks with Recurrent Neural Networks (RNN)

About CNNs https://ujjwalkarn.me/2016/08/11/intuitive-explanation-convnets/

Kaggle speech recognition
https://www.kaggle.com/c/tensorflow-speech-recognition-challenge

https://www.kaggle.com/c/tensorflow-speech-recognition-challenge/kernels

https://www.kaggle.com/kcs93023/keras-sequential-conv1d-model-classification

---

https://stackoverflow.com/questions/52012242/event-loop-is-running-error-on-starting-jupyter-notebook -

One fundamental principle of deep learning is to do away with hand-crafted feature engineering and to use raw features. This principle was first explored successfully in the architecture of deep autoencoder on the "raw" spectrogram or linear filter-bank features,[72] showing its superiority over the Mel-Cepstral features which contain a few stages of fixed transformation from spectrograms. The true "raw" features of speech, waveforms, have more recently been shown to produce excellent larger-scale speech recognition results.[73]

Reading
https://blog.manash.me/building-a-dead-simple-word-recognition-engine-using-convnet-in-keras-25e72c19c12b
The same dataset, the same approach (maybe worse) more about the simple processes

Since computing MFCC is time consuming, we will do it only once. And for later times we will just load it from the saved files, which will buy us some time to do other things ;)

Manash Kumar Mandal
Sep 16, 2018

That might be the cause, try using Bidirectional RNN as well, I used upto 13 classes and the validation accuracy was 97% with 30k samples. Thanks again for reading!

Reading about speech recognition
https://medium.com/@ageitgey/machine-learning-is-fun-part-6-how-to-do-speech-recognition-with-deep-learning-28293c162f7a
We are taking a reading thousands of times a second and recording a number representing the height of the sound wave at that point in time.
“CD Quality” audio is sampled at 44.1khz (44,100 readings per second). But for speech recognition, a sampling rate of 16khz (16,000 samples per second) is enough to cover the frequency range of human speech.

But thanks to the Nyquist theorem, we know that we can use math to perfectly reconstruct the original sound wave from the spaced-out samples — as long as we sample at least twice as fast as the highest frequency we want to record.

That sort of true in some cases, but not for speech. Recognizing speech is a hard problem. You have to overcome almost limitless challenges: bad quality microphones, background noise, reverb and echo, accent variations, and on and on. All of these issues need to be present in your training data to make sure

the neural network can deal with them.

Дуже пізнавальне відео про speech recognition, але дууууже довге. Як то можна вислухати так довго?
https://www.youtube.com/watch?v=9dXiAecyJrY&feature=youtu.be&t=13874

Reading:
https://www.quora.com/What-kind-of-neural-network-should-be-used-for-speech-recognition-or-for-any-other-time-varying-data

Звідси:
https://github.com/kk7nc/RMDL/blob/master/docs/index.rst

Random Multimodel Deep Learning (RMDL) for Classification solves the problem of finding the best deep learning structure and architecture while simultaneously improving robustness and accuracy through ensembles of deep learning architectures.

Attempt at tracking states of the arts and recent results (bibliography) on speech recognition.
https://github.com/syhw/wer_are_we

---

https://missinglink.ai/guides/deep-learning-frameworks/tensorflow-speech-recognition-two-quick-tutorials/ - TensorFlow Speech Recognition: Two Quick Tutorials
Звідси:
https://www.tensorflow.org/tutorials/sequences/audio_recognition
https://github.com/pannous/tensorflow-speech-recognition

---

https://www.intechopen.com/books/from-natural-to-artificial-intelligence-algorithms-and-applications/convolutional-neural-networks-for-raw-speech-recognition
Стаття (чи що це) про використання снн з сирим розпізнаванням розмови і там дуже-дуже багато всякої інфи

The three major types of end-to-end architectures for ASR are attention-based method, connectionist temporal classification (CTC), and CNN-based direct raw speech model.

---

https://www.youtube.com/watch?v=9dXiAecyJrY&feature=youtu.be&t=13874 - ДУЖЕ КОРИСНЕ ВІДЕО! ПРО МОЮ ТЕМУ! 3 3 год 51 хв - 5:22

---

https://venturebeat.com/2019/02/21/google-ai-technique-reduces-speech-recognition-errors-by-29/
Звідси: (просто якась статейка)
As the paper's authors explain, most automatic speech recognition (ASR) systems jointly train three components: an acoustic model that learns the relationship between audio signals and the linguistic units that make up speech, a language model that assigns probabilities to sequences of words, and a mechanism that performs alignment the acoustic frames and recognized symbols. All three use a single neural network (layered mathematical functions modeled after biological neurons) and transcribed audio-text pairs, and as a result, the language model typically suffers degraded performance when it encounters words that infrequently occur in the corpus.

LibriSpeech dataset, after filtering out 500,000 sequences that contained only single-letter words and those that were shorter than 90 words. They found that, by correcting entries from the LAS, the speech correction model could generate an expanded output with "significantly" lower word error rate.

SpecAugment introduced by GoogleBrain - COOL FEATURE
https://ai.googleblog.com/2019/04/specaugment-new-data-augmentation.html

https://towardsdatascience.com/state-of-the-art-audio-data-augmentation-with-google-brains-specaugment-and-pytorch-d3d1a3ce291e
Implementation:
https://github.com/zcaceres/spec_augment
Another one:
https://github.com/shelling203/SpecAugment

---

https://paperswithcode.com/sota/speech-recognition-on-librispeech-test-clean - КЛАСНА РІЧ!
Пейпери з кодами, перераховані і посортовані різні методи, які використовуються для даної задачі, датасет LibriSpeech

https://realpython.com/python-speech-recognition/ - бібліотеки для під'єднання до мікрофона