NVIDIA Deep Speech

Piotr Ambroszczyk Łukasz Kondraciuk
Wojciech Przybyszewski Jan Tabaszewski

Abstract

This thesis focuses on implementing scripts for training DeepSpeech2 model for Automatic Speech Recognition. We try to reproduce results obtained by authors in their paper using `PyTorch` framework. We also experiment with obtaining dataset for Polish language and trying DeepSpeech2 model for it. Finally, we provide fully trained models for English and Polish together with statistics about how changing hyperparametrs and architecture impacts model's performance and accuracy.

Contents

1 Introduction

Our thesis is realized with the support of NVIDIA company, and its goal is to implement Automatic Speech Recognition (ASR) model DeepSpeech2 described in [2]. This model uses end-to-end deep learning method, as opposed to the complex algorithms (e.g. Hidden Marokv Models or acoustic models) which were used by many previous ASR systems. Since `PyTorch` framework is supported with CUDA, and is considered to be comfortable to work with, we decided to use this one. Moreover, the authors of DeepSpeech2 prepared it not only for recognizing English but also Mandarin. One of goals of our thesis is to experiment with applying it to Polish language as well.

DeepSpeech2 model consists of three groups of layers i.e. convolutional, recurrent and fully-connected. Any specific choice of a concrete number and type for each group affects both the model performance and accuracy. To determine the best hyperparamters we needed to run many experiments, collect their results, and finally analyze them. Of course, we had to do it both for English and Polish as these languages differ (e.g. unlike English, there is a strict correspondence between letters and phones in Polish).

Accuracy of the model depends not only on its implementation, but also on the size and diversity of used dataset. Therefore we needed to find appropriate one (paying attention to licenses and copyrights) and prepare it adequately. When it comes to English it was relatively easy - one can find lots of free data. However, for spoken corpus of Polish it was harder, but we managed to find hundreds of hours of Polish speech collected for university programs and from audiobooks. In [1] the authors present techniques for dataset augmentation by e.g. injecting noise. That's because working with specially prepared, clear vocal sounds does not indicate success with processing words and statements of lower quality, and best model's learning parameters can vary in different languages. We used these techniques to improve our model robustness.

The great size of dataset creates another problem - we need our model to be able to train on that data in reasonable time and then work in the real time. To achieve this we train our model on GPU, bearing in mind that `PyTorch` framework supports usage of CUDA devices. Moreover we used open-source libraries prepared by NVIDIA which made it possible to train one neural network over multiple GPUs. Another optimization which speeds computations up is using half precision floating point numbers (also known as FP16) instead of single precision.

To sum up, we present `PyTorch` scripts for training DeepSpeech2 model for ASR. We also present already trained models for English and Polish as

well as the results of our experiments justifying using specific hyperparameters and architecture solutions.

2 Basic model description

3 Additional extensions

4 Experiments on architecture and hyperparameters

5 Recognizing Polish language

5.1 Preparing Polish dataset

5.2 Model's architecture description

5.3 Comparison of model's performance on English and Polish

6 Conclusions

References

[1] Hannun et al. (2014) *Deep Speech: Scaling up end-to-end speech recognition*

[2] Hannun et al. (2015) *Deep Speech 2: End-to-End Speech Recognition in English and Mandarin*