University of Warsaw
Faculty of Mathematics, Informatics and Mechanics

Piotr Ambroszczyk
Student no. 385090

Łukasz Kondraciuk
Student no. 385775

Wojciech Przybyszewski
Student no. 386044

Jan Tabaszewski
Student no. 386319

NVIDIA Deep Speech

Bachelor's thesis
in COMPUTER SCIENCE

Supervisor:
dr Janina Mincer-Daszkiewicz
Instytut Informatyki

May 2019

Supervisor's statement

Hereby I confirm that the presented thesis was prepared under my supervision and that it fulfils the requirements for the degree of Bachelor of Computer Science.

Date Supervisor's signature

Authors' statements

Hereby I declare that the presented thesis was prepared by me and none of its contents was obtained by means that are against the law.

The thesis has never before been a subject of any procedure of obtaining an academic degree.

Moreover, I declare that the present version of the thesis is identical to the attached electronic version.

Date Authors' signatures

Abstract

The authors of this thesis focus on implementing scripts for training DeepSpeech2 model for Automatic Speech Recognition. We try to reproduce results obtained by Baidu Research in End-to-End Speech Recognition paper DODAC DOKUMENTACJE 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.

Keywords

Deep Speech, ASR, Neural Networks, Machine Learning, Python, PyTorch, NVIDIA, RNN, multi-GPU, FP16

Thesis domain (Socrates-Erasmus subject area codes)

11.3 Informatyka

Subject classification

D. Software

Tytuł pracy w języku polskim

NVIDIA Deep Speech

Contents

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]. Our model is going to use 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 plan 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 will have to run many experiments, collect their results, and finally analyze them. Of course, we are going 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 need to find appropriate one (paying attention to licenses and copyrights) and prepare it adequately. When it comes to English it looks easy – one can find lots of free data. However, for spoken corpus of Polish it is going to be harder as we should 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 are going to use 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 will have to train our model on GPU, bearing in mind that `PyTorch` framework supports usage of CUDA devices. Moreover we plan to use open-source libraries prepared by NVIDIA which make 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.

Chapter 1

Basic model description

Chapter 2

Additional extensions

Chapter 3

Experiments on architecture and hyperparameters

Chapter 4

Recognizing Polish language

4.1. Preparing Polish dataset

4.2. Model's architecture description

4.3. Comparison of model's performance on English and Polish

Chapter 5

Conclusions

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.

Bibliography

[1] Hannun et al. *Deep Speech: Scaling up end-to-end speech recognition*, Silicon Valley AI Lab 2014, https://arxiv.org/abs/1412.5567

[2] Baidu Research *Deep Speech 2: End-to-End Speech Recognition in English and Mandarin*, Silicon Valley AI Lab 2015, https://arxiv.org/abs/1512.02595

All the files were downloaded on January 16, 2019