Effects of Layer Freezing when Transfering DeepSpeech to German

Onno Eberhard[1] and **Torsten Zesch**

Language Technology Lab
University of Duisburg-Essen
[1]onno.eberhard@stud.uni-due.de

Abstract

In this paper, we train Mozilla's DeepSpeech architecture in various ways on Mozilla's Common Voice German language speech dataset and compare the results. We build on previous efforts by Agarwal and Zesch (2019) and reproduce their results by training the model from scratch. We improve upon these results by using an English pretrained version of DeepSpeech for weight initialization and experiment with the effects of freezing different layers during training.

1 Introduction

The field of automatic speech recognition is dominated by research specific to English. There exist plenty available text-to-speech models pretrained on (and optimized for) the English language. When it comes to German, the range of available pretrained models becomes much sparser. In this paper, we train Mozilla's implementation[1] of Baidu's DeepSpeech architecture (Hannun et al. 2014) on German speech data. We use transfer learning to leverage the availability of a pretrained English version of DeepSpeech. Our approach is to initialize all parameters of the network to the parameters of the English pretrained model. While training the model, we experiment with freezing the weights of different layers. The rationale for using transfer learning is not only that English and German are closely related languages. In fact, one could argue that they are very different in this context, because DeepSpeech is trained to directly infer written characters from audio data and English and German pronunciations of characters differ greatly. However, the first few layers of the DeepSpeech network are likely not infering the final output character, but rather lower lever features of the spoken input, such as phonemes, which are shared across different languages. Thus, the model should give better results when trained on a small dataset than a model trained from scratch, because it does not have to learn these lower level features again.

[1]https://github.com/mozilla/DeepSpeech

In addition to using transfer learning, we also train the whole model without initializing the weights to those of the English model, thereby reproducing a result from Agarwal and Zesch (2019). We see that the model trained using transfer learning performs better, which is in accordance with our hypothesis stated above.

2 Training

Radeck-Arneth et al. 2015 Heafield 2011

3 Results

4 Further Research

Citing Agarwal and Zesch 2019 here.

References

Agarwal, Aashish and Torsten Zesch (2019). "German End-to-end Speech Recognition based on DeepSpeech". In: *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019): Long Papers*. Erlangen, Germany: German Society for Computational Linguistics & Language Technology, pp. 111–119.

Hannun, Awni et al. (2014). *Deep Speech: Scaling up end-to-end speech recognition*. arXiv: `1412.5567 [cs.CL]`.

Heafield, Kenneth (July 2011). "KenLM: Faster and Smaller Language Model Queries". In: *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Edinburgh, Scotland: Association for Computational Linguistics, pp. 187–197. URL: `https://www.aclweb.org/anthology/W11-2123`.

Radeck-Arneth, Stephan et al. (2015). "Open Source German Distant Speech Recognition: Corpus and Acoustic Model". In: *Proceedings Text, Speech and Dialogue (TSD)*. Pilsen, Czech Republic, pp. 480–488.