Multilingual HW3 Report

Phoebe Li, Yu Zhong

Abstract

End-to-End (E2E) model is the promising method on speech recognition. However, E2E model performance poorly on the low-resources data. In this homework, we build up the E2E-ASR model and try to improve the performance on Chinese speech recognition. Specifically, we experimented with data set size, and maximum duration.

1 Introduction

Over the last decades, speech recognition has been through great development and the Automation Speech Recognition (ASR) technique is wildly adapted into smart home devices, such as Google Home, Amazon Alex, and Siri. However, the ASR still performs poorly on the low-resources data due to the lack of enough dataset and has trouble dealing with multilingual speech audio and the language with no word boundary. To improve the ASR performance on the multilingual data, the researchers recently introduced the E2E ASR model. The E2E model offers a much simpler and more compact solution compared to the conventional model, which has separate acoustic, pronunciation, and language models.(Sainath, 2019) In this report, we mainly observed how the E2E-ASR model performs on the Chinese dataset, and discuss the future improvement on the E2E model.

2 Training Stages

The training in total involves 13 stages. The stage 1 to stage 5 are data preparation. The stage 6 to stage 9 are language modeling phase while stage 10 to 11 are end-to-end ASR model. Stage 12 is used for decoding while stage 13 is used for scoring.

3 Dataset Description

The dataset utilizes the existing egs2 recipe named 'primewords_chinese'. The dataset is Chinese Mandrin corpus that contains 100 hours of speech data. It was recorded by smart mobile phones from 296 native Chinese speakers. The transcription accuracy is larger than 98%. The dataset size is around 9GB.

4 Metrics

To keep track of our model training process, we closely investigated the following sets of metrics.

1. Attention Plot

It tells us which parts of the input sentence has the model's attention while translating. Ideally, the attention plot should show with diagonal shape.

2. Word Error Rate (WER)

The WER is a metric abased on Levenshtein distance, The metrics is calculated based on the number of words that differ between a piece of machine-translated text and reference translation.

4. Character Error Rate(CER) CER is similar to WER but it operates on character instead of word.

4. Train Log

This file helps track the model architecture, loss and accuracy after each iteration.

5 Performance Analysis

5.1 Experiment 1: subset training

Since the dataset is relatively large and it took a long time to train, we firstly ran a small model using only a subset of data. For this part, we extracted 30% of overall dataset with language modeling. However, the model didn't improve WER value much.

5.2 Experiment 2: maximum duration

The second experiment we tried was to compare the model performance under different maximum duration in seconds. The change of maximum duration is closely related to the stage 4 processing, which

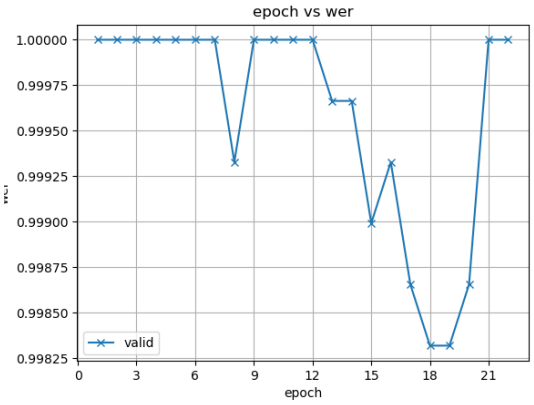

Figure 1: WER of subset training

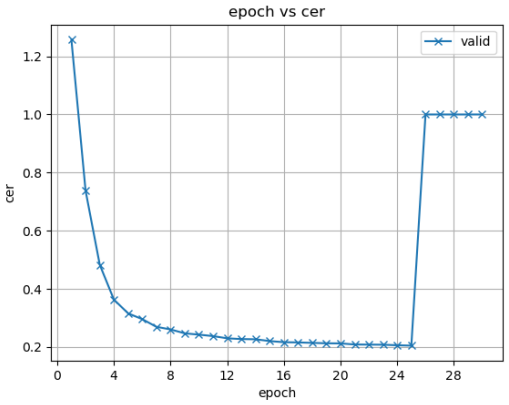

Figure 3: CER of full set training, max duration 20

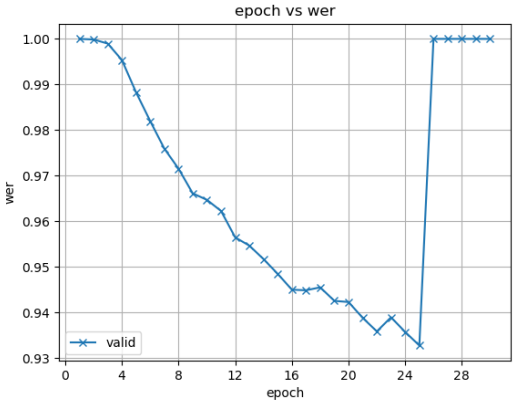

Figure 4: WER of full set training, max duration 20

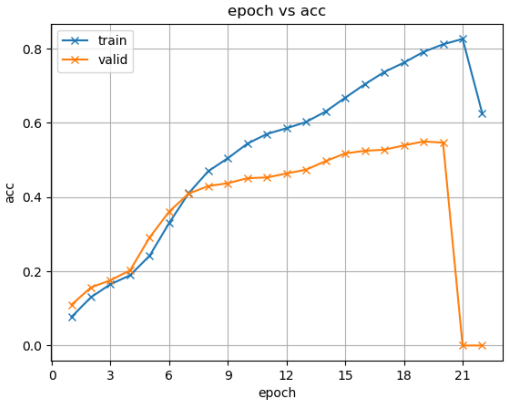

Figure 2: Accuracy of subset training

remove some extremely long sequences. The reason we tried this was because we noticed that with default maximum duration, the grad norm went nan around 25 epoch(figure 3 and figure 4). The output length could be larger than input length for CTC. To improve the model performance, we later changed the maximum duration to 15 seconds. Although the model improved slightly, the grad norm went nan after 34 epoch(figure 5 and figure 6).

6 Discussion

Based on our research, we found that ASR training with Chinese has overall very high WER value. We think this is due to that Chinese characters do not have spaces between. If one character was wrong then the entire sentence would be considered wrong. Therefore, we think that CER could be a better illustration of our model performance.

Secondly, for future work, we hope to improve further on the decoding stage. While we were reviewing our decoded transcripts, we found that one

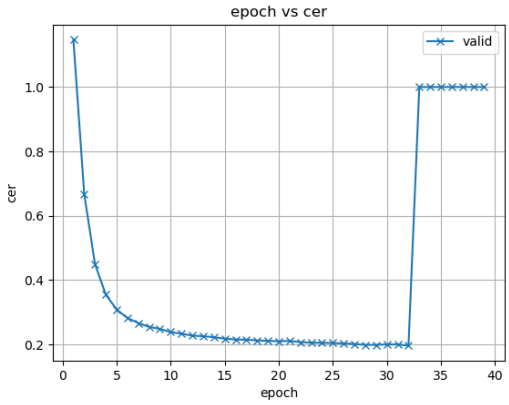

Figure 5: CER of full set training

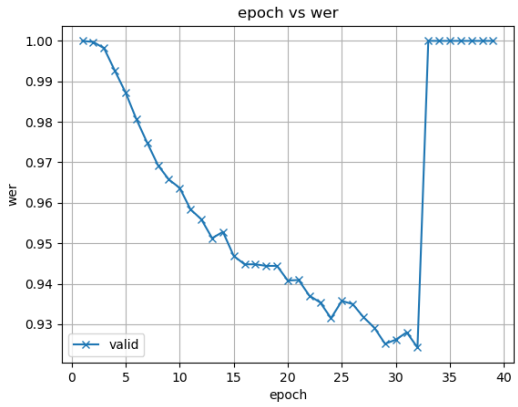

Figure 6: WER of full set training

problem with Chinese speech recognition is that one pronunciation corresponds to many different characters. Consequently, the decoded sentence sounds right when reading but it is wrong in writing. Therefore, we would propose to decode all voices into Pinyin and then use another model to convert the Pinyin to readable texts.

Lastly, to further improve the performance, we could implement some pretraining techniques. One example is to pretrain the decoder first for 15 epochs. During this process, we could also use teacher forcing algorithm to speed up the learning of decoders.

7 Summary

In summary, we conducted two experiments on our dataset and ASR models. We noticed that the E2E models required a relatively large dataset to improve the performance, which also indicates why E2E perform poorly on the low-resource data. In the second experiment we showed that by reducing the maximum duration, the model performance improved slightly. In the future experiment, we are planning to use Pinyin as the decoder to avoid the wrong prediction of the Chinese characters.

References

B. Li, T. N. Sainath, R. Pang and Z. Wu, "Semi-supervised Training for End-to-end Models via Weak Distillation," ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 2837-2841, doi: 10.1109/ICASSP.2019.8682172.