Performance of pretraining wav2vec model

Jiaqi Liu, Colin Tan

December 2020

Abstract

This report elaborates on optimization of wav2vec pre-training balance between training set size, accuracy and training time. The approach is to experiment with progressively larger source data sets to pre-train the wav2vec 2.0 representation that goes through an identical fine-tuning using a fixed dataset of the annotated data, and evaluate with accuracy and training time of the pre-trained model. Our experiment finds a unexplored relationship between pre-training set size, accuracy and training time of the wav2vec 2.0 model.

1 Introduction

Wav2vec is a CNN-based automatic speech recognition model. Facebook AI Research's 2019 paper *wav2vec: Unsupervised Pre-training for Speech Recognition* [3] concludes that pre-training substantially improves the performance of the wav2vec model. That is, the increase in the training data size leads to a decrease in word error rate (WER), which is a common metric of the performance of speech recognition. Therefore our assumption is that the increase of wav2vec pre-training dataset size would result in an increase in the testing accuracy of the fine-tuned model but also an increase in the training time.

Hence, we conducted experiments to find a balance between wav2vec unlabeled training data size, accuracy and training time of pre-training the model, by progressively adjusting the unlabeled training data size on the model. This was unexplored in Facebook's original wav2vec 2.0 paper *wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations* published a couple months ago [2].

2 Related Works

Our work is based on Facebook AI Research's 2019 paper [3] on wav2vec and 2020's paper on wav2vec 2.0 [2]. wav2vec and its successor are unsupervised pre-training methods for speech recognition that use masking pre-training with contrastive loss to learn representations of speech units' time and contents or the raw audio. We conduct the experiment using wav2vec 2.0 source code from the official github repository [4].

The 2019 paper demonstrates that increasing training set size substantially improves WER in simulated low-resource setups on the audio data of the full 960 hours LibriSpeech dataset, the 81 hours WSJ dataset, and the wav2letter++ with log-mel filterbank baseline. It also shows that wav2vec model on the full 960 hours LibriSpeech dataset performs approximately 20% better than wav2vec model on the WSJ dataset and nearly 40% better than the baseline [3]. This paper inspires us to investigate the balance between pre-training dataset size, WER and pre-training time of the wav2vec 2.0 model. Also, the outstanding performance of wav2vec model in LibriSpeech is the reason why we choose to use LibriSpeech daraset in the experiment.

Facebook AI Research's 2020 official blog *wav2vec 2.0: Learning the structure of speech from raw audio* [1] shows the relationship between the resulting WER and the amount of the annotated data from Libri-light dataset for fine-tuning the model pre-trained by the full 960 hours LibriSpeech benchmark. It demonstrates that wav2vec 2.0 can enable speech recognition models for settings where there is very little labeled training data for fine-tuning. It provides the technical support for us to use relatively small dataset for fine-tuning in the experiment.

3 Approach

We conduct the experiment with progressively larger unlabeled pre-training data sets to train the wav2vec 2.0 representation. The quality of the sound recording should be kept at the same level. The pre-trained representation will go through an identical Librispeech ASR fine-tuning using labeled dataset, then verified with WER metric and we record the training time as well.

3.1 Model Selection

The SOTA model wav2vec large (LV-60) was trained on 24 GPUs to get trained in viable timeframe. Considering the large model requires huge amount of resources to train, we determine to use the wav2vec 2.0 base model due to our computational constraint.

3.2 Dataset Selection

We use the LibriSpeech ASR corpus, which is a large-scale (approximately 1000 hours) corpus of 16kHz read English speech. The data is derived from read audiobooks from the LibriVox project, and has been carefully segmented and aligned. The acoustic models trained on LibriSpeech give lower error rate on the Wall Street Journal (WSJ) test sets than models trained on WSJ itself.

The pre-training dataset is the 100 hours clean subset from Librispeech ASR corpus. The pre-trained models are fine-tuned with 3 hour clean speech from Librispeech ASR corpus. It was previously demonstrated that wav2vec 2.0 enables speech recognition with very small size of fine-tuning dataset.

4 Experiments

4.1 Environment

The CNN architecture of the wav2vec 2.0 model means GPU provides a lot of boost in training performance, so we have done the training with a NVIDIA Tesla V100 GPU, which is the same GPU type used in Facebook's original wav2vec 2.0 paper [2]. We use the open-sourced implementation released by Facebook that uses PyTorch framework to train. The VMs are Google Cloud VMs with 2 vCPUs and 13 GB of RAM. They run PyTorch 1.6 with CUDA 11.

4.2 Evaluation

We record and compare the training time and fine-tuning time of the wav2vec 2.0 representation of different training dataset size, and compare the accuracy with WER metric as well. Final evaluation was done using LibriSpeech `test-clean` set.

4.3 Parameter Adjustment

We use the mentioned training dataset from LibriSpeech with total audio length varying from 10 hours, 20 hours, 30 hours, 40 hours, to 50 hours to train the wav2vec model separately from scratch. The split is done with selecting a subset of the records that sums up to the target playback length. Due to our computational constraint, we decide not to move forward with pre-training dataset size larger than 50 hours. Pre-training is set to run for 50 epochs.

Since the contrastive loss seems to explode with 50 hours of audio for pre-training, we have adjusted the base learning rate from 0.0005 to 0.0002 for the 50 hours experiment.

The fine-tuning was done using 3 hours of labeled LibriSpeech audio recordings. Finetuning is set to run for 150 epochs.

4.4 Time

From Figure 1 we can see that the training time scales linearly with training audio length used for pre-training. The fitted linear model $552.05x + 27.38$ has an $R^2 = 0.9958$.

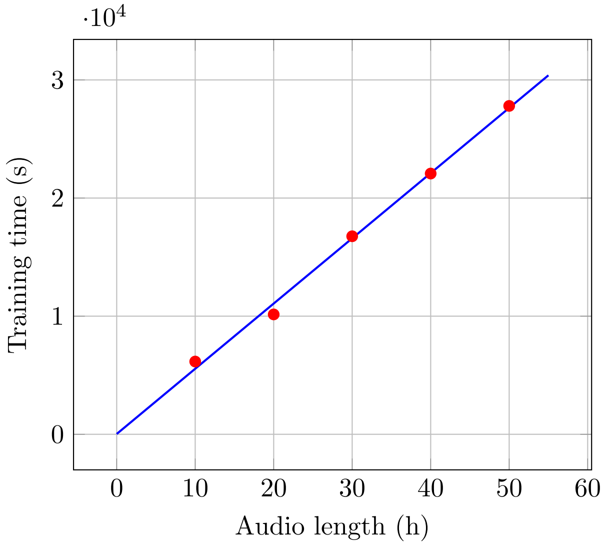

Figure 1: Pre-training time vs pre-training audio length

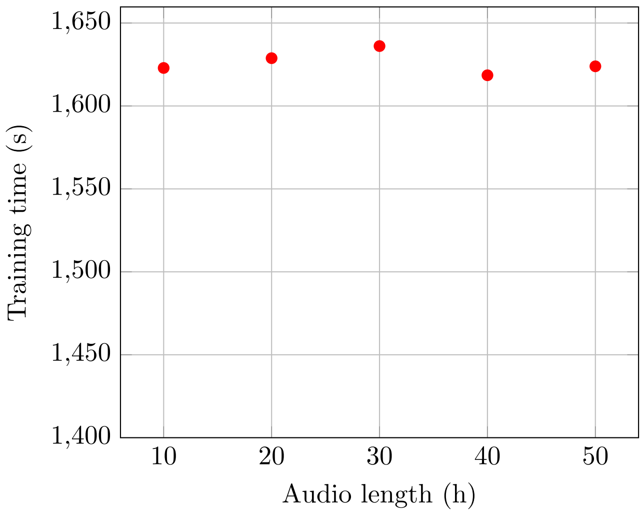

Figure 2: Fine-tuning time vs pre-training audio length

Figure 2, on the contrary, shows that pre-training audio length has little effect on the fine-tuning time cost of the model.

4.5 WER

In this analysis we have to discard the 50 hours datapoint, as using a smaller base learning rate seem to have made that incompatible to be compared with other data points. Using WER of 10 hours of pre-training as baseline, we can compare the fine-tuned model accuracy in Figure 3.

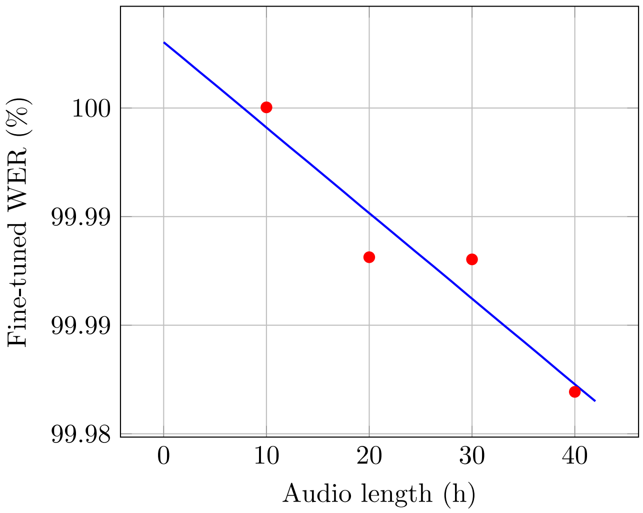

Figure 3: Pre-training time vs pre-training audio length

The linear model $100.003 - 0.000394x$ has an $R^2 = 0.9028$. The WER metrics are very far from SOTA performance, but that is expected since we are training the models with about 1/10 to 1/50 of pre-training data, and about 1/10 of the fine-tuning data, and also about 1/20 of epoch limits. We focus on the relative performance between different pre-training settings.

5 Conclusion

Our experiments report on the previously unexplored relationship between pre-training dataset size of wav2vec, the pre-training cost, and the fine-tuned model performance. We found linear relationships between pre-training dataset size and pre-training cost and fine-tuned model performance.

In before the marginal improvement start to diminish, we recommend users of wav2vec to incorporate as much pre-training data as their budget allows.

References

[1] Alexei Baevski, Alexis Conneau, and Michael Auli. Wav2vec 2.0: Learning the structure of speech from raw audio, 2020.

[2] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations, 2020.

[3] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition, 2019.

[4] Facebook AI Research Sequence to-Sequence Toolkit written in Python. wav2vec 2.0, 2020.