MaSo: Masking and Sorting for Self-supervised Representation Learning for Raw Audio

Nitish Joshi
160070017

Yash Shah
160050002

Abstract

In this project we work on representation learning from raw audio that can capture important information useful in various downstream tasks. We hypothesize that **context** and **order** are two important aspect of audio and introduce two self-supervised tasks of Masking and Sorting to capture them. We evaluate our representations on two different tasks: Speaker Identification and Automatic Speech Recognition. We also provide an analysis and discussion on the usefulness of these self-supervised tasks. We publicly release the code for this project.[1]

1 Introduction

The problem of representation learning lies at the heart of many modern deep learning techniques. Strong representations of the input can compactly capture the essential information which can be useful in various downstream tasks. Recently there has been a surge of interest in unsupervised learning, specifically self-supervised learning for representation learning. Self-supervised learning has been very effective in Computer Vision [Jing and Tian, 2019] and in Natural Language Processing [Devlin et al., 2019, Peters et al., 2018], but there has sparse work on this problem for speech.

For speech signals we hypothesize that **context** and **order** are two important aspects which can be useful in learning strong representation from raw audio. To facilitate learning of these two aspects, we introduce two self-supervised tasks: Masking and Sorting. These are inspired from previous success of similar methods in Computer Vision (Sorting: Lee et al. [2017]) and in Natural Language Processing (Masked Prediction: Devlin et al. [2019]). Masking refers to the task of reconstructing the original input from a masked representation of the input. Sorting refers to the task of predicting the original temporal order of a sequence from one of its random permutations.

We evaluate our representations on two different downstream tasks: Speaker Identification and Automatic Speech Recognition. We also compare against PASE [Pascual et al., 2019] and perform an analysis of combining our self-supervised task with the tasks introduced in PASE.

2 Related Work

In this section, we describe PASE [Pascual et al., 2019] which is closest in motivation to our project. PASE introduces seven different self-supervised tasks for learning good representations from speech signals by keeping a common encoder across all the tasks. The first task uses an auto-encoder objective which learns to reconstruct the original waveform from the encoded representation. The next three tasks are regression tasks to predict Log Power Spectrum (LPS), Mel-frequency Cepstral Coefficients (MFCC) and Prosody features. The next two tasks are based on Mutual Information Objective as in Ravanelli and Bengio [2018a]. Specifically, these tasks are binary discriminative tasks with the positive sample being drawn from the sentence and the negative from a random sentence.

[1] https://github.com/ys1998/maso

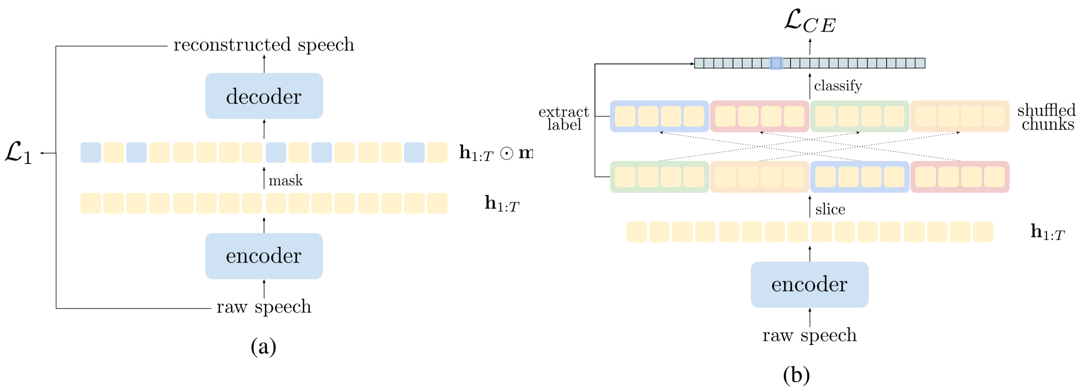

Figure 1: The figure on left depicts the masking task with $\mathcal{L}_1$ reconstruction loss. The right subfigure is the sorting task when input is divided into $m = 4$ equal parts.

The last task, sequence predictive coding is also a binary discriminative task with the positive and negative samples being drawn from future and past frames respectively.

3 Our Method

In this section we describe in detail the self-supervised tasks of Masking and Sorting.

Masking: In this self-supervised task, we use a decoder to reconstruct the original waveform from randomly masked features of input speech. Specifically, if $\boldsymbol{h}_{1:T}$ is the encoded representation of speech with length T, and $\boldsymbol{m}_{1:T}$ is a random vector of same size with each element being 0 with probability p (else it is set to 1) then the masked features are $\boldsymbol{h}_{1:T} \odot \boldsymbol{m}_{1:T}$. Here, $\odot$ represents the element-wise multiplication operation. We use $\mathcal{L}_1$ loss to train the decoder weights. The choice for $\mathcal{L}_1$ loss is similar to PASE and is motivated by robustness since speech distribution is very peaky. We provide an illustration of the masking task in Fig. 1a. In all our experiments, we set $p = 0.2$.

Sorting: In this self-supervised task, we predict the original order in a sequence from one of its random permutations. This encourages the encoder to learn features which are aware of the temporal information. We divide the input sequence of length T into m equal parts. Let S_m denote the array of all permutations of $(1..m)$ in increasing order. Then given $\boldsymbol{h}_{1:m}$ the encoded representation of speech divided into m parts, we randomly generate a permutation of $(h_1..h_m)$, say $r = (h'_1..h'_m)$. We train the model to predict $S_m.index(r)$ given r as the input where $index(.)$ function returns the index of the argument in the array. Note that this is essentially treating the task as a classification problem with $m!$ labels. We use $m = 4$ in our experiments to keep the labels space sufficiently small and provide an illustration of the sorting task in Fig. 1b.

4 Experiments

In this section we describe the implementation details of our model as well the datasets we use for each of the tasks.

Implementation Details: We built our code on that released for PASE.[2]. Our encoder is similar to that used by PASE - It is a fully convolutional network based on the recently proposed SincNet [Ravanelli and Bengio, 2018b]. The worker network for sorting is a feed-forward network with one hidden layer and the worker node for masking is a deconvolutional network. We use Adam [Kingma and Ba, 2014] optimizer in all our pre-training experiments with an initial learning rate of 0.0005 and train the model for 100 epochs.

Datasets: For the pre-training experiments we use Librispeech dataset [Panayotov et al., 2015] and extract 30 sec from each of the 251 speakers. For the speaker identification task we use VCTK [Veaux et al., 2017] and preprocess to extract only 15 sec of speech samples for each speaker - This makes

[2] https://github.com/santi-pdp/pase

Model	Speaker Identification Accuracy
No Pre-training	30.2%
Masking & Sorting (MaSo)	59.06%
PASE	99.33%
All Workers	98.66%

Table 1: Speaker Identification results on the VCTK dataset.

Model	Phoneme Accuracy
Masking & Sorting (MaSo)	20.69%
PASE	65.65%
All Workers	65.8%

Table 2: Automatic Speech Recognition (Phoneme Accuracy in this case) for the TIMIT dataset.

the task harder as well reduces the computational burden for the experiments. For Automatic Speech Recognition, we use TIMIT dataset [Zue et al., 1990] and follow the train-test split as in PASE.

5 Results

In this section we describe our main results along with analysis to better understand the self-supervided tasks of masking and sorting.

5.1 Speaker Identification

In Table 1, we report our results for speaker identification on VCTK dataset with 109 speakers. The first row denotes a model in which all weights are trained from scratch. The second row is using Masking and Sorting (MaSo) and the third row denotes our reproduced result using PASE. The last row uses all the self-supervised tasks in PASE and our work. In all these experiments we fine-tune the weights of the encoder along with the rest of the weights. As we can see, MaSo by itself does not perform very well - But it's combination with all PASE workers allows to get a reasonable good speaker identification accuracy.

5.2 Automatic Speech Recognition

In Table 2, we report the results for Automatic Speech Recognition on the TIMIT dataset. Following PASE, we use phoneme classification accuracy as the evaluation metric (instead of the usual metric of edit distance). The models in each row are same as described in Section 5.1. All these experiments use frozen weights for the encoder - thus the encoder only acts a pre-trained feature extractor. As we can see from the last row, MaSo combined with the worker tasks in PASE can help give a small improvement over the results of PASE.

5.3 Analysis

To understand the importance of fine-tuning encoder weights along with training the other weights, we show a comparison of the two cases for speaker identification on VCTK. We obtain an accuracy of **90.66%** with frozen encoder weights compared to **98.66%** with fine-tuning of the encoder weights. Thus, fine-tuning the encoder does help to give large improvement in performance. This observation is consistent with what is usually observed with pre-trained language models in Natural Language Processing such as BERT [Devlin et al., 2019].

Next, we take up the following question: Can we identify the crucial tasks in PASE which when combined with MaSo can give equally good performance? Based on the ablations presented in PASE, we choose MFCC and LPS prediction as the two tasks from PASE (giving a total of 4 self-supervised tasks compared to 7 in PASE). We report the results of this model on both speaker identification (with fine-tuning of encoder) and automatic speech recognition (with frozen encoder weights) in Table 3. As evident from the results, the combination of these tasks can give large jumps over MaSo thus

Model	Accuracy (VCTK)
MaSo	59.06%
MaSo Enhanced	98.66%

Model	Accuracy (TIMIT)
MaSo	20.69%
MaSo Enhanced	65.77%

Table 3: Analysis of adding MFCC and LPS tasks to MaSo (MaSo enhanced in table) for both speaker identification and phoneme classification.

performing equally if not better than PASE. This demonstrates that the MaSo combination can be powerful enough to replace up to 5 self-supervised tasks in PASE.

6 Conclusion

In this work, we presented two self-supervised tasks for representation learning for raw audio. The two tasks of Masking and Sorting by themselves are not good enough to learn good representation for speech signals - this could potentially be because these tasks are much harder than say in NLP owing to the continuous nature of speech. But these tasks when combined with a few fundamental self-supervised tasks can make the model powerful enough to potentially replace up to 5 tasks in PASE.

7 Future Work

Currently for the masking task, we employ feature masking rather than input masking. The reason for this is that we wanted to make best use of the smaller pre-training dataset that we use in the experiments. When more data is available, input masking can be an effective option as seen in BERT [Devlin et al., 2019]. The second possible direction could be exploring the loss for sorting task further - Currently the model is penalized even if there is partial match between predicted sequence and original sequence since it is treated as classification task. It could be possible to develop better versions of this loss which accommodate for such partial matches.

References

Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *ArXiv*, abs/1902.06162, 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke S. Zettlemoyer. Deep contextualized word representations. *ArXiv*, abs/1802.05365, 2018.

Hsin-Ying Lee, Jia-Bin Huang, Maneesh Kumar Singh, and Ming-Hsuan Yang. Unsupervised representation learning by sorting sequences. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 667–676, 2017.

Santiago Pascual, Mirco Ravanelli, Joan Serrà, Antonio Bonafonte, and Yoshua Bengio. Learning problem-agnostic speech representations from multiple self-supervised tasks. *ArXiv*, abs/1904.03416, 2019.

Mirco Ravanelli and Yoshua Bengio. Learning speaker representations with mutual information. *ArXiv*, abs/1812.00271, 2018a.

Mirco Ravanelli and Yoshua Bengio. Speaker recognition from raw waveform with sincnet. *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 1021–1028, 2018b.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An asr corpus based on public domain audio books. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210, 2015.

Christophe Veaux, Junichi Yamagishi, and Kirsten Macdonald. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit. 2017.

Victor Zue, Stephanie Seneff, and James R. Glass. Speech database development at mit: Timit and beyond. *Speech Communication*, 9:351–356, 1990.