Title: Self-supervised audio representation learning for mobile devices

Author: Marco Tagliasacchi, Beat Gfeller, Félix de Chaumont Quitry, Dominik Roblek (Google Research)

Summary: This paper is focused on self-supervision learning of audio representations for mobile devices. But it also can be modified for general-purpose use. More specifically, the authors proposed two self-supervised tasks that exploit the temporal context in the spectrogram domain. Firstly, TemporalGap estimates the distance between two audio segments, then Audio2Vec is described to reconstructing a spectrogram slice from past and future slices. Experiments with fully supervised methods prove the effectiveness of the proposed method over.

Good Things about This Paper

1. The topic Self-supervised audio representation is useful in some other areas with small labeled dataset like speech emotion recognition.
2. With analogy to Word2Vec, the proposed method is simple and easy to understand. The article itself is well written with no explicit ignorance on the table and figures.
3. The authors did a lot of experiments among many different methods.

Major Comments

1. The proposed model shows that pre-training and fine-tuning helps many downstream tasks, but they did not explain why. It may be better to discuss what the model is actually learning in the self-supervised tasks, because it is an important evaluation if we want to apply the method on other topics.
2. However, the most significant problem is, even though the result in the experiments seems great, but we can see from table 4, that the performance of this self-supervised method did not surpass the supervised model. The reasons and possible solution need further discussion here. So, it becomes a question that if we really need to pre train on big dataset.
3. It may be better to describe how the authors decide some hyper parameters like the number of epochs, fine-tuning process, optimizer choice, learning rate etc.
4. Finally, tough mentioned above that the authors run several different former models to evaluate their new method, the baseline seems not strong enough. There are still some low cost experiments remain undone like simple concatenation of the embeddings. For the downstream supervised tasks presented in the paper, there is one very similar article [1], which is even better because it usees speech data but not general audio data. What is the performance compared to this one?

[1] Milde, B., & Biemann, C. (2018). Unspeech: Unsupervised Speech Context Embeddings. INTERSPEECH.

Minor Comments

1. Fig. 2: The figures maybe too small.