Conformer-Based Self-Supervised Learning for Non-Speech Audio Tasks

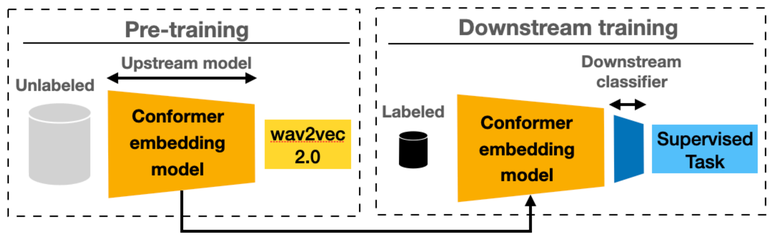
Sangeeta Srivastava[1], Yun Wang[2], Andros Tjandra[2], Anurag Kumar[3], Chunxi Liu[2], Kritika Singh[2], Yatharth Saraf[2]
[1]The Ohio State University, [2]Meta AI, [3]Reality Labs Research

Meta

Poster Number: 3268

1. Objective and Setup

- Learn audio representation using self-supervised wav2vec 2.0 task and conformer architecture
- Evaluate the generality of the representations on diverse non-speech audio tasks

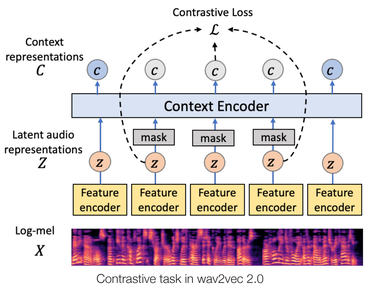


Our self-supervised learning pipeline

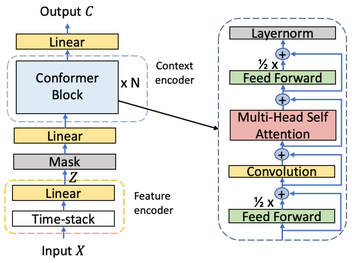
- 67k hours of de-identified non-speech sounds from Facebook
- Fine-tuning during downstream adaptation

2. Wav2vec 2.0 & Conformer

- **Task**: Identify true latent representation for a masked time step within a set of K + 1 candidates
- **Feature Encoder:** Stacked spectrogram into latent representation
- **Context Encoder:** Linear layer + N conformer blocks



Contrastive task in wav2vec 2.0

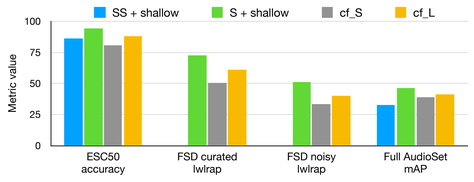


Conformer architecture and conformer block

- Self attention for global interactions
- Convolution for local correlations
- Conformer Small: cf_S (~18M)
  - 12 conformer layers, 256 encoder dim, feed-forward network dim 1024, heads 8
- Conformer Large: cf_L (~88M)
  - 12 conformer layers, 768 encoder dim, feed-forward network dim 1024. heads 12

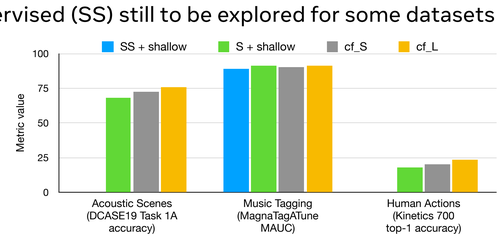
3. Sound Event Detection

- Baseline models
  - Self-supervised (SS) + shallow and supervised (S) + shallow
- cf_L outperforms SS baselines for ESC50 & full AudioSet
- cf_L is 7.6% (19.1%) worse than S baselines in ESC50 (FSD curated)
  - Pre-trained on AudioSet -> overlap with ESC50 & FSD in label-space

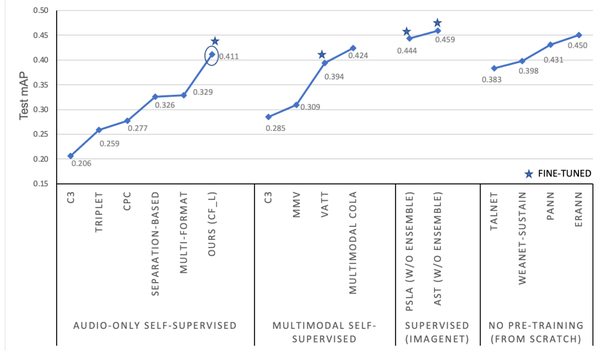


4. Other Non-Speech Audio Tasks

- Conformer models are competitive (if not better) compared to baselines
- Self-supervised (SS) still to be explored for some datasets



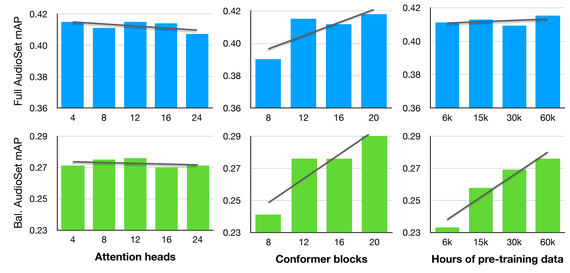
5. Prior Works on AudioSet



- cf_L outperforms SoTA in audio-only SS by 25% with mAP of 0.411
- Competitive even to the best multimodal SSL work on AudioSet
- Worse than ImageNet pre-trained models and some models trained from scratch

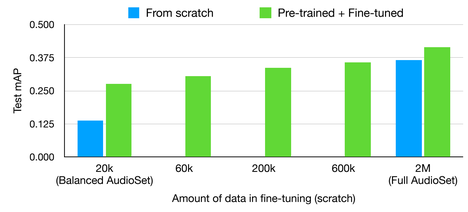
6. Hyperparameter Optimization

- No significant change in performance with the variation in heads
- Trendline shows a larger variation in Balanced AudioSet for variation in conformer blocks and pre-training data



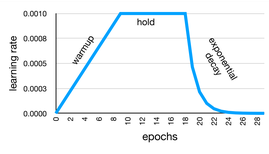
7. Effect of Pre-training

- Models trained from scratch worse than pre-trained counterparts
- Pre-training helps reduce the need for labeled data by two-thirds



8. Avoiding Overfitting

- 3-stage learning rate scheduler
  - warmup, hold, exponential decay



- Batch Size
  - high-resource datasets perform best with larger batches
- Dropout in output layer of pre-trained conformer

9. References

- Baevski, Alexei, et al. "wav2vec 2.0: A framework for self-supervised learning of speech representations." NeurIPS. 2020.
- Gulati, Anmol, et al. "Conformer: Convolution-augmented transformer for speech recognition." arXiv preprint arXiv:2005.08100. 2020.