Compositional Audio Representation Learning

Sripathi Sridhar, Mark Cartwright

New Jersey Institute of Technology

Correspondence: ss645@njit.edu

Highlights

- **We propose a source-centric representation learning framework for environmental sound**
- **We investigate the design space of these models**
- **Compositional audio representations outperform baselines on audio classification**
- **Label supervision helps learn better source-centric representations on synthetic data**
- **Proposed additional loss terms lead to better downstream performance**

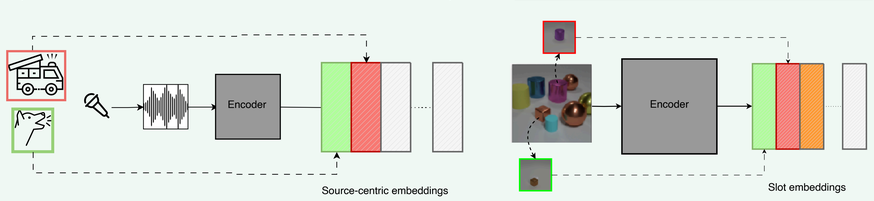
Motivation

- While human auditory perception is compositional, audio representations do not typically distinguish between constituent sources in the auditory scene
- Source-level reasoning may be useful for many tasks including source localization, individual ID, open-set classification, audio-language modeling, etc.
- Source separation models estimate sources from mixtures, but don't aim to learn semantic representations for downstream tasks
- Source-level inference in open-set scenarios needs universal source separation, which is a challenge
- Compositional audio representations could unlock applications in machine listening with greater flexibility and interpretability

What is Compositional Audio Representation Learning?

Goal: Map each source in the auditory scene to a distinct embedding, taking inspiration from object-centric learning in computer vision.

In computer vision, object-centric learning maps each object in the visual scene to a distinct embedding.



Synthetic data

Open-set Soundscapes (OSS)
- 260k 10s soundscapes synthesized from FSD50K sources
- 54 seen classes, 35 unseen classes

Open-set Tagging (OST)
- 500k 1s clips
- Window around center of each event in soundscape

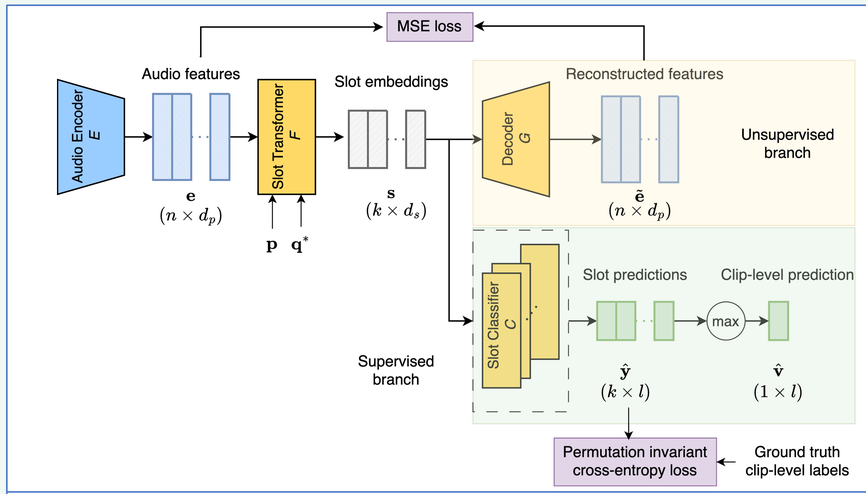
Research questions

1. Do the proposed unsupervised source-centric models improve downstream performance?
2. How does the reconstruction target affect downstream performance for unsupervised models?
3. How important are design choices such as the decoder and the loss function?
4. How does the model generalize to unseen classes?

Key findings

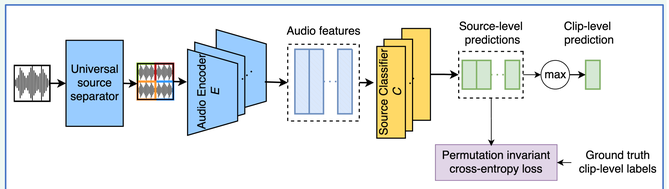
1. **Compositional audio representations outperform baselines and are useful for environmental sound classification**
2. **Target signal matters! Reconstructing features is significantly better than reconstructing spectrograms**
3. **The proposed disjointedness and sparsity loss terms lead to better downstream performance**
4. **Label supervision helps learn better source-centric representations on synthetic data**

Proposed source-centric representation learning model



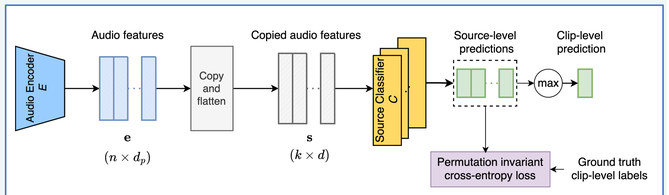
- Audio encoder generates clip-level audio features
- Slot transformer transforms them into source-centric latents, or slot embeddings
- Decoder reconstructs the target signal, which may be audio features or spectrograms
- The classifier makes a prediction per slot
- We evaluate supervised and unsupervised model variants
- *Disjointedness penalty (not shown):* encourage slots to be disjoint using a cosine similarity on slot embeddings
- *Sparsity penalty (not shown):* encourage each slot to encode a single source using L1 norm on slot predictions

Universal source separator baseline



- TDCN++ MixIT universal source separator (USS) trained from scratch on our dataset
- Freeze USS and audio encoder
- Permutation invariant training of source classifier, as source-level labels are assumed unavailable

AudioMAE baseline



- Simulates an ineffective slot transformer by copying and flattening the audio features as slot embeddings
- Freeze audio encoder
- Permutation invariant training of source classifier, as source-level labels are assumed unavailable

Models

- Audio encoder: AudioMAE pretrained on AudioSet
- Slot transformer: Inverted-attention transformer decoder
- Target types: AudioMAE features | Mel spectrograms
- Reconstruction decoder types: Transformer | MLP | CNN
- Alpha mask: Scalar per time-frequency region used to composite slot-level reconstructions

Training strategy

Supervised model: Train using cross-entropy, reconstruction, disjointedness, and sparsity.

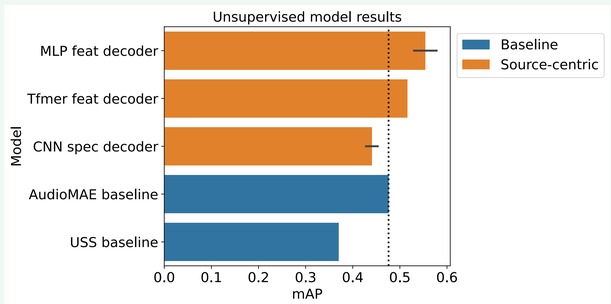
Unsupervised model:

- Train slot embeddings using reconstruction and disjointedness
- Train slot classifier on top of frozen embeddings using cross-entropy and sparsity

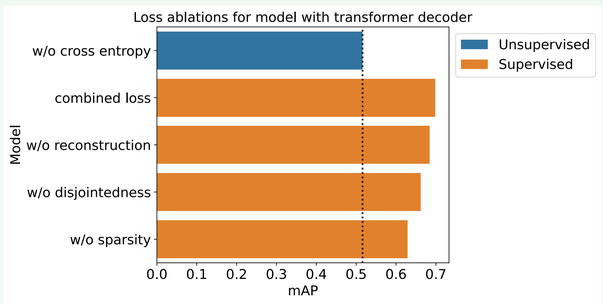
Results

RQ1. The proposed unsupervised source-centric models improve downstream performance

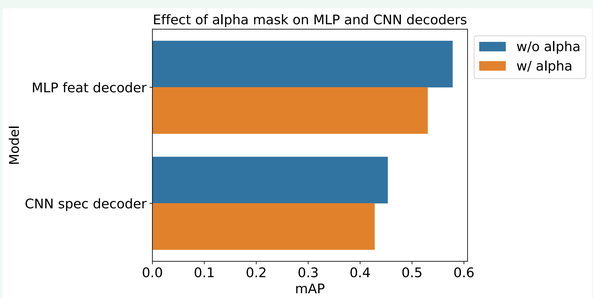
RQ2. Feature reconstruction is better than spectrogram reconstruction for downstream performance



RQ3a. Label supervision, disjointedness, and sparsity are useful for downstream performance



RQ3b. Unsupervised Models without alpha mask perform better on downstream classification



RQ4. Supervised models generalize to unseen classes