Towards Learning Semantic Audio Representations from Unlabeled Data

Aren Jansen, Manoj Plakal, Ratheet Pandya, Daniel P. W. Ellis, Shawn Hershey, Jiayang Liu, R. Channing Moore, Rif A. Saurous

Google AI

NIPS 2017 ML4Audio Workshop
AudioSet ([http://g.co/audioset](http://g.co/audioset))

- A large-scale collection of labeled sound examples
  - Like ImageNet for sound
- 2M+ ten-second excerpts from high-view count YT videos
- At least 120 human-verified examples for 500+ classes
- **Plus:** we released a state-of-the-art embedding model + code

[Gemmeke et al., Audio Set: An Ontology and Human-Labeled Dataset for Audio Events, ICASSP 2017]
● The Semantic Value of Unlabeled Audio

● Unsupervised Triplet Embeddings
  ○ 4 Unsupervised Triplet Sampling Methods

● Evaluation
  ○ Query-by-Example Sound Retrieval
  ○ Sound Event Classification
The Semantic Information in Unlabeled Audio

- **AudioSet gives**: “this recording is a dog bark”
- **This work**: What can we assert in the absence of that label?
  1. We can add Gaussian noise to the recording and it is still a dog bark.
  2. It is still a dog bark if it instead occurs 5 seconds from now, or has slightly higher pitch.
  3. It is still a dog bark if someone is simultaneously talking or a car is passing by.
  4. If the dog is barking now, it is probably also barking (or growling or panting) 5 seconds from now.
- Analogous to “self-supervised” approaches in computer vision community
● **Triplet Loss for Deep Metric Learning:**
  ○ Given: example triplets of form (anchor, positive, negative)
  ○ Estimate: map \( g \) to low-dimensional space where
    \[
    \text{Dist}(g(a), g(p)) + \text{margin} < \text{Dist}(g(a), g(n))
    \]

● **Typical use:** anchor and positive same class, negative different class

● **However:** can be use for any constraint of form “\( a \) is more like \( p \) than like \( n \)"
Sampling Method 1: Gaussian Noise

- Audio Perspective:
  - Semantic category is invariant to moderate noise

  ![Anchor and Positive Examples](image)

- Machine Learning Perspective:
  - Categories invariant to small perturbations in input space
  - Analogous to denoising autoencoder without the decoder
  - Opens up arbitrary encoder architecture

\[ \varepsilon_{tf} \sim N(0, \sigma^2) \]

\[ \text{positive}_{tf} = \text{anchor}_{tf} * (1 + |\varepsilon_{tf}|) \]
Semantic percept (of individual events) are invariant to arbitrary translations in time and (to some extent) shifts in frequency.

Sampling Method 2: Time/Frequency Translation

- Positive: random circular shift in time & random truncated shift in frequency
Sampling Method 3: Example Mixing

● **Audio Perspective:**
  ○ Mixtures preserve constituent sound categories

● **Machine Learning Perspective:**
  ○ Warp interpolation points towards individual examples
  ○ Like replacing Gaussian noise with real distractors, but interpolations safer than using random negatives

\[
\text{positive} = \text{anchor} + \alpha \cdot \text{negative}
\]
Sampling Method 4: Temporal Proximity

- Nearby sounds are likely to be same category or semantically related

![anchor](image1)
![positive: within $\Delta t$ seconds of anchor](image2)

(same clip for AudioSet)
Joint Training

- Combining all the above semantic constraints into a single model is trivial:
  - Randomly shuffle all training triplet sets together

- **Note**: one could also introduce per-source loss weighting or vary each sources sample sizes, but we only evaluate equal contribution
**Data:** AudioSet used for all training and evaluation (527 classes, 3M training segments, public eval set)

**Triplet Embedding Models:**
- Input: 96 frame X 64 mel band log mel spectrogram context windows (0.96 seconds)
- ResNet-50 CNN architecture
- 128-dimensional output embedding layer + L2 normalization (Euclidean → cosine)

**Evaluation Tasks:**
- Query-by-example sound event retrieval
- Sound event classification using shallow classifiers

**Topline:** fully-supervised triplet embedding

**Baseline:** input log mel spectrogram features

Query-by-Example Retrieval

- **For Each Class**: Rank target and nontarget example pairs by cosine distance
- **Metric**: Mean average precision (mAP) over the 527 AudioSet classes (Prior = 0.331)

![Bar chart showing mAP values for different methods]

- **Supervised Triplets**: 0.790
- **Log Mel Spectrogram**: 0.423
- **Gaussian Noise**: 0.478
- **T/F Translation**: 0.508
- **Example Mixing**: 0.489
- **Temporal Proximity**: 0.562
- **Joint Unsupervised**: 0.575

41% Recovery
Sound Event Classification

- Train shallow fully-connected (512 units) classifier using all AudioSet labeled data
- Metric: Mean average precision (mAP) over the 527 AudioSet classes (Prior = 0.003)

![Graph showing mAP for different methods and layers](image)

- 84% Recovery
- Individual Sampling Methods:
  - Supervised Triplets
  - Log Mel Spectrogram
  - Gaussian Noise
  - T/F Translation
  - Example Mixing
  - Temporal Proximity
  - Joint Unsupervised

Metric values:
- 1 layer FC: 0.288, 0.289
- 2 layers FC: 0.065, 0.102
Semi-Supervised Classification

- **Train Set**: Random 20 labeled examples/class = 0.5% of training data (3 trials)
  - Unsupervised triplet model trained on entire set without labels
- **Metric**: Mean average precision (mAP) over the 527 AudioSet classes

<table>
<thead>
<tr>
<th>Input Representation</th>
<th>Classifier Architecture</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Log Mel Spectrogram</td>
<td>Fully Connected (4x512)</td>
<td>0.032</td>
</tr>
<tr>
<td>Log Mel Spectrogram</td>
<td>ResNet-50</td>
<td>0.072</td>
</tr>
<tr>
<td>Joint Unsupervised Triplet</td>
<td>Fully Connected (1x512)</td>
<td>0.143</td>
</tr>
</tbody>
</table>

Log Mel Spectrogram + FC 1x512 trained with 100% labels gets 0.065
Layer 1 Convolutional Filters

- Nicely localized, qualitatively similar to supervised model
Conclusions

- We proposed a general strategy to eliciting semantic structure in learned audio representations.
- Allows pre-training arbitrarily complex neural networks on in-domain unlabeled data, reducing labeled data requirements.
- Compatible (and probably complementary) with other neural network architectures tailored to unsupervised audio modeling.