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, Inc.
Mountain View, CA, and New York, NY, USA
{arenjansen,plakal,ratheet,dpwe,shershey,jiayl,channingmoore,rif}@google.com

Abstract

Our goal is to learn semantically structured audio representations without relying on categorically labeled data. We consider several class-agnostic semantic constraints that are inherent to non-speech audio: (i) sound categories are invariant to additive noise and translations in time, (ii) mixtures of two sound events inherit the categories of the constituents, and (iii) the categories of events in close temporal proximity in a single recording are likely to be the same or related. We apply these constraints to sample training data for triplet-loss embedding models using a large unlabeled dataset of YouTube soundtracks. The resulting low-dimensional representations provide both greatly improved query-by-example retrieval performance and reduced labeled data and model complexity requirements for supervised sound classification.

1 Introduction

The last few years have seen great advances in nonspeech audio processing as popular deep learning architectures developed in the speech and image processing communities have been ported to this relatively understudied domain [17, 8, 4, 21]. However, these data-hungry neural architectures are not always matched to the available training resources in the audio domain; while unlabeled audio is easy to collect, manually labeling data for each new sound application remains notoriously costly and time consuming. We address this mismatch by exploiting semantic properties of sound using massive amounts of unannotated audio.

Recent efforts in the computer vision community have identified several class-independent constraints on natural images and videos that can be used to learn semantic representations [25]. For example, object categories are invariant to camera angle, and tracking unknown objects in videos can provide novel examples for the same unknown category [20]. For audio, we can identify several analogous constraints, which are not tied to any particular inventory of sound categories. First, we can apply category-preserving transformations to individual events of unknown type, such as adding Gaussian noise, translation in time within the analysis window, and small perturbations in frequency. Second, pairs of unknown sounds can be mixed to provide new, often natural sounding examples of both. Finally, sounds from within the same vicinity of a recording are likely to contain multiple examples of the same (or related) unknown categories.

Triplet loss-based deep metric learning [22, 19, 9] estimates nonlinear maps into a low-dimensional space where simple Euclidean distance can express any desired relationship between examples of the form X *is more like* Y *than like* Z. This is a natural framework to express the above semantic constraints because the inter-example relationship need not be anchored to a categorical judgment. To

This extended abstract is derived from a manuscript submitted for publication to ICASSP 2018.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

validate this approach, we train embedding models using state-of-the-art convolutional architectures on millions of triplets sampled from the *AudioSet* dataset [6], both with and without the use of the provided labels. We evaluate the learned embeddings as features for query-by-example sound retrieval and supervised sound event classification. Our results demonstrate that complex models can be trained from unlabeled triplets to produce representations that recover much of the performance gap between the raw log mel spectrogram inputs and fully supervised semantic embeddings.

2 Related Work

Lee et al. [13] applied convolutional deep belief networks to extract a representation for speech and music, but not general purpose nonspeech audio. More recently, a denoising autoencoder variant was used to extract features for environmental sound classification [23], but there was no explicit training mechanism to elicit semantic structure in their learned embeddings. Recent zero-resource efforts in the speech processing community have explicitly aimed to learn meaningful linguistic units from untranscribed speech [18]. Moreover, various forms of deep metric learning [16, 10, 24, 12] have been applied using these speech-specific constraints. Finally, so-called self-supervised approaches in the computer vision community are analogous to what we propose in this paper for audio [1, 14, 5, 20, 25]. Recent efforts have extended this principle of self-supervision to joint audio-visual models that learn speech and audio embeddings using semantic constraints imposed by the companion visual signal [2, 3, 7, 11].

3 Learning Algorithm

Our training procedure consists of two stages: (i) sampling training triplets from a collection of unlabeled audio recordings, and (ii) learning a map from input context windows extracted from spectrograms (matrices with F frequency channels and T frames) to a lower d-dimensional vector space using triplet loss optimization of convolutional neural networks.

3.1 Metric Learning with Triplet Loss

The goal of triplet loss-based metric learning is to estimate a map $g : \mathbb{R}^{F\times T} \to \mathbb{R}^d$ such that simple (e.g.) Euclidean distance in the target space correspond to highly complex geometric relationships in the input space. Training data is provided as a set $\mathcal{T} = \{t_i\}_{i=1}^N$ of example triplets of the form $t_i = (x_a^{(i)}, x_p^{(i)}, x_n^{(i)})$, where $x_a^{(i)}, x_p^{(i)}, x_n^{(i)} \in \mathbb{R}^{F\times T}$ are commonly referred to as the anchor, positive, and negative, respectively. The loss is given by

$$\mathcal{L}(\mathcal{T}) = \sum_{i=1}^{N} \left[\|g(x_a^{(i)}) - g(x_p^{(i)})\|_2^2 - \|g(x_a^{(i)}) - g(x_n^{(i)})\|_2^2 + \delta \right]_+ , \tag{1}$$

where $\|\cdot\|_2$ is L_2 norm, $[\cdot]_+$ is standard hinge loss, and δ is a nonnegative margin hyperparameter. Intuitively, the optimization is attempting to learn an embedding of the input data such that positive examples end up closer to their anchors than the corresponding negatives do by some margin. The map g can be defined by a d-unit output layer of any modern deep learning architecture. The optimization is performed with stochastic gradient descent, though training time is greatly decreased with the use of within-batch semi-hard negative mining [15]. Here, all examples in the batch are transformed under the current state of g, and the available negatives are reassigned to the anchors-positives to make more difficult triplets.

3.2 Triplet Sampling Methods

To set our performance topline, we consider the traditional fully-supervised triplet sampling strategy: for each class, we randomly sample anchor-positive pairs with the same class, and attach a negative with a different class. We can express our proposed unsupervised semantic constraints using the following triplet sampling strategies:

- **Gaussian Noise**: For each $x_i \in \mathbb{R}^{F\times T}$ contained in the provided set of unlabeled examples, we sample one (or more) anchor-positive pairs of the form (x_i, x_p), where we define element $x_{p,tf} = x_{i,tf}(1 + |\epsilon_{tf}|)$ for $\epsilon_{tf} \sim \mathcal{N}(0, \sigma^2)$, a Gaussian distribution with mean 0 and standard deviation σ (a model hyperparameter). Finally, we attach a third random example to each pair to be the triplet's negative.

- **Time and Frequency Translation**: For each $x_i \in \mathbb{R}^{F \times T}$ in the provided set of unlabeled examples, we sample one or more anchor-positive pairs of the form(x_i, x_p) where $x_p = \mathrm{Trunc}_S(\mathrm{Circ}_T(x_i))$. Here,$\mathrm{Circ}_T$ is a circular shift in time by an integer number of frames sampled uniformly from $[0, T-1]$. Trunc_S is a truncated shift in frequency by an integer number of bins sampled uniformly from $[-S, S]$(missing values after shift are set to zero). We again attach a random example as the negative.
- **Example Mixing**: Given a random anchor x_a and random negative x_n, both containing energies in each time-frequency cell, we construct positive $x_p = x_a + \alpha \cdot [E(x_a)/E(x_n)]x_n$, where $E(x)$ is the total energy of example x, and α is a hyperparameter.
- **Temporal Proximity**: We sample triplets of the form (x_a, x_p, x_n) where x_a and x_p are from the same recording, x_n is from a different recording. We can further impose the constraint that $|\mathrm{time}(x_a) - \mathrm{time}(x_p)| < \Delta t$, where $\mathrm{time}(x)$ is the start time of example x and Δt is a hyperparameter.

All triplet sets produced using the sampling methods outlined in this section can be simply mixed together for training a joint embedding that reflects them all to whatever degree possible.

4 Experiments

We evaluate the triplet sampling methods presented in Section 3.2 by using the resulting embeddings in two downstream tasks: (i) query-by-example retrieval of sound segments that contain the same event categories, and (ii) training shallow fully-connected sound event classifiers. We also perform a lightly supervised experiment using only a small fraction of the labeled data.

4.1 Dataset and Model Architecture

We use an internal version of Google's recently released *AudioSet* database of manually annotated sound events [6] for both training and evaluation. *AudioSet* consists of 10 second audio clips from YouTube videos, each labeled using a comprehensive ontology of 527 sound event categories. We use a total of 3 million clips for training, with at least 120 examples for every class. We compute 64-channel mel-scale, logarithmically-compressed spectrograms using an FFT window size of 25 ms with 10 ms step. We then process these per-recording spectrograms into nonoverlapping 0.96 second context windows, such that each training example is a $F = 64$ by $T = 96$ matrix. Following Hershey et al. [8], we use the ResNet-50 convolutional neural network architecture. Instead of the classification output layer, all of our triplet models use a 128-unit fully-connected linear output layer. This produces a vector of dimension d =128, which represents a factor of 48 reduction from the original input dimensionality of 64 × 96. We also employ the standard practice of length normalizing the network output before input to the loss function, making the squared Euclidean distance of the loss function equivalent to cosine distance.

4.2 Query-by-Example Retrieval

Our first evaluation task is segment-level query-by-example (QbE) retrieval. We begin by mapping each 0.96 second context window in the evaluation set to its corresponding 128-dimensional embedding vector and average these across each *AudioSet* segment to arrive at a segment-level embedding. For each sound event category, we sample 100 segments where it is present and 100 segments where it is not. We report the mean average precision (mAP) of separating within-class and across-class pairs using cosine distance in the embedding space (chance baseline is 0.331). Table 1 shows the retrieval performance for each of the evaluated representations (hyperparameters optimized on the development set). The fully-supervised topline uses explicitly labeled data to sample the triplets. The raw log mel spectrogram features are the baseline, and provide a measure of the inherent semantic structure exhibited by the input representation when no unsupervised mapping is learned. The recovery listed is the percentage of the baseline-to-topline performance gap recovered using each given unsupervised triplet embedding. We find that each unsupervised triplet method significantly

Table 1: Mean average precision for segment retrieval and shallow model classification using original log mel spectrogram and triplet embeddings as features.

Representation	QbE Retrieval		Classification (1 layer, 512 units)		Classification (2 layer, 512 units)	
	mAP	recovery	mAP	recovery	mAP	recovery
Supervised Triplets (topline)	0.790	*100%*	0.288	*100%*	0.289	*100%*
Log Mel Spectrogram (baseline)	0.423	*0%*	0.065	*0%*	0.102	*0%*
Gaussian Noise ($\sigma = 0.5$)	0.478	15%	0.096	14%	0.114	6%
T/F Translation ($S = 10$)	0.508	23%	0.108	19%	0.125	12%
Mixed Example ($\alpha = 0.25$)	0.489	18%	0.103	17%	0.122	11%
Temporal Proximity ($\Delta t = \infty$)	0.562	38%	0.226	72%	0.241	74%
Joint Unsupervised	0.575	41%	0.244	80%	0.259	84%

Table 2: Lightly supervised classifier performance averaged over three trials, each trained with a different random draw of 20 segments/class (totaling 0.5% of labeled data).

Representation	Model Architecture	mAP
Log Mel Spectrogram	Fully Connected (4x512)	0.032
Log Mel Spectrogram	ResNet-50	0.072
Joint Unsupervised Triplet	Fully Connected (1x512)	0.143

improves retrieval performance over the input features, with the joint unsupervised model improving mAP by 15% absolute over the input features.

4.3 Sound Classification

We use our various embeddings as features for training a shallow fully-connected neural networks (1 or 2 layers of 512 units each) using labeled AudioSet segments. The output layer consists of independent logistic regression models for each class. Following Gemmeke et al. [6], we compute segment-level scores (average of the frame-level predictions) for the evaluation set, and we again report the mean average precision for each feature type. Table 1 shows the classification performance for each feature, including the topline and baseline representations. We again find substantial improvement over the baseline in all cases, with temporal proximity the clear standout. Combining triplet sets provides additional gains, indicating the model's ability to encode multiple types of semantic constraints for downstream tasks. Notice that our fully-unsupervised training of a ResNet-50 triplet embedding model achieves 85% (0.244/0.288) the mAP of a fully-supervised ResNet-50 triplet embedding model, when both are coupled to a single hidden layer downstream classifier.

Finally, Table 2 shows performance of lightly supervised classifiers trained on just 20 examples per class (we report average performance over 3 training samples), which amounts to only 0.5% of the train set. We evaluate three models: (i) a ResNet-50 classifier model, (ii) a fully-connected (4 layers of 512 units each, which was optimal) model trained from log mel spectrograms, and (iii) a single layer (512) fully-connected model trained on the joint unsupervised triplet embedding (last line of Table 1). Since our unsupervised triplet embeddings are derived from the full AudioSet train set (as unlabeled data), a single layer classifier trained on top doubles the mAP of a full ResNet-50 classifier trained from raw inputs.

5 Conclusions

We have presented a new approach to unsupervised audio representation learning that explicitly designed to elicit semantic structure in the absence of labeled data. By sampling triplets based on a variety of audio-specific semantic constraints that do not require labeled data, we arrive at a representation that greatly outperforms the raw inputs on both near neighbor retrieval and classification of sound events. We found that the various semantic invariants are complementary, producing downstream classification improvements when combined to train a joint triplet loss embedding model. Finally, we demonstrated that our best unsupervised embedding provides a great advantage when training sound event classifiers in limited supervision scenarios.

References

[1] Pulkit Agrawal, Joao Carreira, and Jitendra Malik. Learning to see by moving. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 37–45, 2015.

[2] Relja Arandjelović and Andrew Zisserman. Look, listen and learn. *arXiv preprint arXiv:1705.08168*, 2017.

[3] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In *Advances in Neural Information Processing Systems*, pages 892–900, 2016.

[4] Emre Cakır, Giambattista Parascandolo, Toni Heittola, Heikki Huttunen, and Tuomas Virtanen. Convolutional recurrent neural networks for polyphonic sound event detection. *arXiv preprint arXiv:1702.06286*, 2017.

[5] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1422–1430, 2015.

[6] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: A strongly labeled dataset of audio events. In *Proceedings of ICASSP*, 2017.

[7] David Harwath, Antonio Torralba, and James Glass. Unsupervised learning of spoken language with visual context. In *Advances in Neural Information Processing Systems*, pages 1858–1866, 2016.

[8] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. CNN architectures for large-scale audio classification. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 131–135. IEEE, 2017.

[9] Elad Hoffer and Nir Ailon. Deep metric learning using triplet network. In *International Workshop on Similarity-Based Pattern Recognition*, pages 84–92. Springer, 2015.

[10] Herman Kamper, Micha Elsner, Aren Jansen, and Sharon Goldwater. Unsupervised neural network based feature extraction using weak top-down constraints. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 5818–5822. IEEE, 2015.

[11] Herman Kamper, Shane Settle, Gregory Shakhnarovich, and Karen Livescu. Visually grounded learning of keyword prediction from untranscribed speech. *arXiv preprint arXiv:1703.08136*, 2017.

[12] Herman Kamper, Weiran Wang, and Karen Livescu. Deep convolutional acoustic word embeddings using word-pair side information. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 4950–4954. IEEE, 2016.

[13] Honglak Lee, Peter Pham, Yan Largman, and Andrew Y Ng. Unsupervised feature learning for audio classification using convolutional deep belief networks. In *Advances in neural information processing systems*, pages 1096–1104, 2009.

[14] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2536–2544, 2016.

[15] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015.

[16] Gabriel Synnaeve, Thomas Schatz, and Emmanuel Dupoux. Phonetics embedding learning with side information. In *Spoken Language Technology Workshop (SLT), 2014 IEEE*, pages 106–111. IEEE, 2014.

[17] Naoya Takahashi, Michael Gygli, Beat Pfister, and Luc Van Gool. Deep convolutional neural networks and data augmentation for acoustic event detection. *arXiv preprint arXiv:1604.07160*, 2016.

[18] Maarten Versteegh, Roland Thiolliere, Thomas Schatz, Xuan-Nga Cao, Xavier Anguera, Aren Jansen, and Emmanuel Dupoux. The Zero Resource Speech Challenge 2015. In *Interspeech*, pages 3169–3173, 2015.

[19] Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. Learning fine-grained image similarity with deep ranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1386–1393, 2014.

[20] Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2794–2802, 2015.

[21] Yun Wang and Florian Metze. A first attempt at polyphonic sound event detection using connectionist temporal classification. In *Proc. of ICASSP*, 2017.

[22] Kilian Q Weinberger and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10(Feb):207–244, 2009.

[23] Yong Xu, Qiang Huang, Wenwu Wang, Peter Foster, Siddharth Sigtia, Philip JB Jackson, and Mark D Plumbley. Unsupervised feature learning based on deep models for environmental audio tagging. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(6):1230–1241, 2017.

[24] Neil Zeghidour, Gabriel Synnaeve, Nicolas Usunier, and Emmanuel Dupoux. Joint learning of speaker and phonetic similarities with Siamese networks. In *INTERSPEECH*, pages 1295–1299, 2016.

[25] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European Conference on Computer Vision*, pages 649–666. Springer, 2016.