Semi-supervised triplet loss based learning of ambient audio embeddings

Nicolas Turpault, Romain Serizel, Emmanuel Vincent

SEMI-SUPERVISED TRIPLET LOSS BASED LEARNING OF AMBIENT AUDIO EMBEDDINGS

Nicolas Turpault *Romain Serizel* *Emmanuel Vincent*

Université de Lorraine, CNRS, Inria, Loria, F-54000 Nancy, France

ABSTRACT

Deep neural networks are particularly useful to learn relevant representations from data. Recent studies have demonstrated the potential of unsupervised representation learning for ambient sound analysis using various flavors of the triplet loss. They have compared this approach to supervised learning. However, in real situations, it is common to have a small labeled dataset and a large unlabeled one. In this paper, we combine unsupervised and supervised triplet loss based learning into a semi-supervised representation learning approach. We propose two flavors of this approach, whereby the positive samples for those triplets whose anchors are unlabeled are obtained either by applying a transformation to the anchor, or by selecting the nearest sample in the training set. We compare our approach to supervised and unsupervised representation learning as well as the ratio between the amount of labeled and unlabeled data. We evaluate all the above approaches on an audio tagging task using the DCASE 2018 Task 4 dataset, and we show the impact of this ratio on the tagging performance.

Index Terms— Semi-supervised learning, triplet loss, audio tagging, audio embedding

1. INTRODUCTION

Sound carries a lot of information that humans can interpret even unconsciously. However, building a system that is able to automatically recognize sounds in real environments is far from trivial. In the recent years, there has been a growing interest for research in ambient sound analysis [1] motivated by applications in various domains including surveillance, smart cities or home assisted living. Audio tagging is a particular task of ambient sound analysis that consists of detecting the sound event classes that occur in an audio clip regardless of their start and end times. In the following, we aim to detect and classify domestic sounds in 10-second clips from the DCASE 2018 Task 4 dataset [2].

State-of-the-art methods for audio tagging rely on deep neural networks (DNNs), such as convolutional neural networks (CNNs) [3, 4], recurrent neural networks (RNNs) [5], or a combination of both usually referred to as CRNNs [6, 7]. These methods learn a discriminative representation of the data called an *embedding* which helps the classification. The embedding may be learned jointly with the classifier by minimizing a classification cost, or separately by methods such as siamese networks [8] or triplet networks [9] which rely on different costs. Triplet networks have been used for images [10], speech [11], and music [12] and recently for ambient sounds [13]. They are trained on triplets of data points consisting of an *anchor*, a *positive* sample, and a *negative* sample. All the above methods are supervised learning methods. Therefore they highly depend on the amount of labeled data. Annotating data is a tedious task. The availability of labeled datasets is usually the bottleneck that hinders the application of ambient sound analysis to real world problems.

Recently, there have been efforts to collect substentially larger datasets. Audioset [14] is an audio dataset including clips with overlapping events in real environments but labels are not carefully verified and often unreliable [14]. Freesound [15] is an open source platform that provides more reliable labels and can be used to build meaningful tagging datasets [16]. However, most clips in Freesound contain isolated events and little or no background noise. Therefore, the applicability of systems trained on Freesound to real-world applications is limited. Soundscape synthesis and data transformation [17] can be used to simulate more realistic environments or create large synthetic datasets from smaller labeled datasets.

In this paper, we use the DCASE 2018 Task 4 dataset which is a subset of Audioset where some labels have been manually verified and the others have been discarded. One solution to exploit a dataset without annotations is to use unsupervised learning approaches based on generative DNNs [18, 19], dimension reduction and clustering [20, 21], or triplet networks [13]. Jansen et al. proposed to train a triplet network with different unsupervised triplet sampling [1] strategies making use of data transformation and compared them to a fully supervised sampling strategy. They demonstrated that an unsupervised sampling strategy can achieve 84% of the mean average precision achieved by a supervised sampling strategy for ambient sound analysis task on Audioset. This approach does not take labels into account during the learning process.

Semi-supervised learning is a combination of supervised and unsupervised learning approaches that allows the exploitation of both labeled and unlabeled data [22]. It has usually been applied to ambient sound analysis by annotating new data recursively during consecutive passes through a supervised system [23, 24]. This pseudo-labeling approach provides a slight performance improvement. However clips that are given labels with low confidence in the early passes tend to remain with low confidence as training progresses through the passes. Recently, a task in the DCASE challenge has targeted semi-supervised approaches [2]. The winner used a mean-teacher system [25] which makes use of two networks. However, to the best of our knowledge, semi-supervised learning of triplet networks has not yet been studied.

In this paper we propose semi-supervised sampling strategies to create triplets and study their application to audio tagging. Sampling strategies have already been discussed for siamese networks [26].

This work was made with the support of the French National Research Agency, in the framework of the project LEAUDS "Learning to understand audio scenes" (ANR-18-CE23-0020) and the French region Grand-Est. Experiments presented in this paper were carried out using the Grid5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations (see `https://www.grid5000.fr`).

[1] In this paper, statistical sampling, not signal processing sampling

Here we start from the sampling strategies used by Jansen et al. [13]. We then propose a new way of selecting the positive sample in a triplet and combine it with the semi-supervised sampling strategy.

We first describe the problem in Section 2 then we explain the proposed solution in Section 3. The experimental setup is described in Section 4 before drawing some conclusions in Section 5.

2. PROBLEM DESCRIPTION

The problem of audio tagging consists of identifying which ambient sound classes are present in an audio clip regardless of their position in time. Let $\mathscr{C}$ be a set of C classes. We have a small set of labeled data $\mathscr{D}_{\mathscr{L}} = \{(\boldsymbol{x_l}, \boldsymbol{y_l})\}_{l=1}^{L}$ and a large set of unlabeled data $\mathscr{D}_{\mathscr{U}} = \{(\boldsymbol{x_u})\}_{u=1}^{U}$ where $\boldsymbol{x_i}, i \in \{u, l\}$, is a time-frequency representation of the input data and $\boldsymbol{y_l} = [e_1, ..., e_C]$ is a vector containing the labels with $e_c \in \{0, 1\}$ indicating whether event class c is present in the clip or not. Note that the vector of labels can contain multiple classes. Let $\mathscr{D} = \{\mathscr{D}_{\mathscr{L}}, \mathscr{D}_{\mathscr{U}}\}$ denote the set containing both labeled and unlabeled data. In the following, we aim to learn embeddings of both unlabeled and labeled data using the triplet loss.

2.1. Triplet loss

The aim of the triplet loss [9] is to find meaningful embeddings of the data that bring the anchor and the positive example closer than the negative example and the anchor. In the following, $(\boldsymbol{x^a}, \boldsymbol{x^p}, \boldsymbol{x^n})$ is a triplet defined by (anchor, positive, negative) samples and the cost function to be minimized is:

$$\sum_{x_i \in \mathscr{D}_{\mathscr{N}}} \left[||f(x_i^a) - f(x_i^p)||_2^2 - ||f(x_i^a) - f(x_i^n)||_2^2 + \delta\right]_+ \quad (1)$$

where $[]_+$ is the hinge loss, $||.||_2$ is the L_2 norm and δ is the margin. $f(x)$ is the embedding of x. This cost function differs from the pair-based costs used in siamese networks by ensuring a balanced number of negative and positive examples.

2.2. Supervised sampling strategy

Different triplet sampling strategies have been proposed in the literature, which exploit the data and the available labels. Different sampling strategies result in different embeddings. The simplest strategy tested by Jansen et al. [13] relies on labeled data only. In this strategy, the positive and negative samples are randomly chosen under the constraints that the positive sample has at least one label in common with the anchor and the negative sample has no label in common with the anchor.

2.3. Transformation-based unsupervised sampling strategy

Jansen et al. [13] also proposed an unsupervised sampling strategy, which does not make use of any label. In this strategy, a transformed version of the anchor is used as the positive sample and *semi-hard mining* is employed to obtain the negative sample. Semi-hard mining consists of choosing the negative sample that is closest to the anchor in the embedding space while being further away than the positive sample [27]. We describe below the different transformations used to obtain the positive sample.

Gaussian noise: The positive sample is created by adding Gaussian noise with a standard deviation of σ to the anchor.

Time and frequency translation: The positive sample is created by circularly shifting the anchor sample in time by an integer number of frames sampled uniformly from $[0, T-1]$, where T is the number of frames in the sample. This sample is then shifted in frequency by an integer number of bins sampled uniformly from $[-S, S]$ (missing values after shift are set to zero).

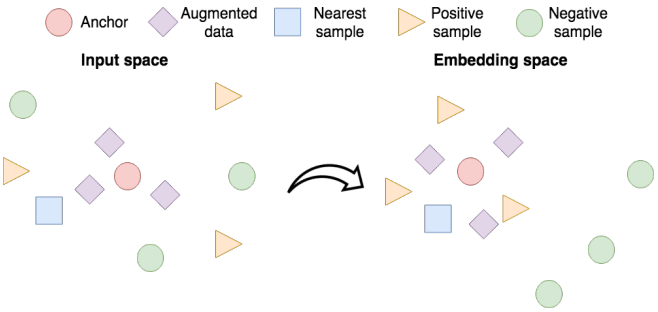

Fig. 1: Illustration of triplet sampling.

Temporal proximity: The positive sample is a sample from the same audio clip as the anchor sample with a time difference inferior to Δt between them.

Example mixing: The positive sample is a mix of the anchor and the negative samples using $x^p = x^a + \alpha[E(x^a)/E(x^n)]x^n$ where E means the energy. The negative sample is taken randomly in the dataset.

3. PROPOSED SEMI-SUPERVISED SAMPLING STRATEGIES

The supervised sampling strategy proposed by Jansen et al. [13] uses labeled data only, while their other approaches are fully unsupervised and discard the labels. By contrast, we propose two semi-supervised sampling strategies where labeled and unlabeled triplets are both used during training. Our strategies differ by the choice of the positive sample when dealing with unlabeled data.

3.1. Semi supervised strategy

In the semi-supervised setting, we use the supervised strategy when the label is available and the unsupervised strategy when the sample is unlabeled. The following unsupervised strategies to choose the positive sample are always combined with semi-hard mining to choose the negative sample in our experiments (See table 1).

3.2. Positive sample: transformed anchor

This strategy consists of generating the positive sample by transforming the anchor sample as above.

3.3. Positive sample: nearest sample

This strategy consists of taking the nearest sample in the input space.

$$x^a \in \mathscr{D}_{\mathscr{U}}$$
$$x^p \in \{x_i \in \mathscr{D} | x_i = argmin_{x_i}(||x_i - x^a||_2^2)\}$$

The idea is to use the sparsity and repartition of the data. Previous work has used it to regularize the embedding [28]. A transformed data sample can be seen as a close point in the input space. Therefore taking the nearest point in the input space is implicitly assuming these points should also be close in the representation space.

As can be seen in Fig. 1, our motivation is to search beyond the space spanned by the transformed data to create the embedding.

Sampling strategies		Unsupervised	Semi-supervised		Supervised	
		S1 [13]	S2	S3	S4	S5
Positive	Label		X	X	X	X
	Trans	X	X			
	Nearest			X		
Negative	*Label*		X	X		X
	S-hard	X	X	X	X	

Table 1: Sampling strategies for the different systems. Trans means transformation and S-hard means semi-hard mining

Indeed, when we have many samples in a small space, taking a transformed version of the anchor as the positive sample can make sense, but when the data points are sparse, it is interesting to take the nearest sample to enlarge the space spanned by positive samples in the input space. Figure 1 represents this latter case. However, it can happen that the nearest sample is closer than some transformed samples.

4. EXPERIMENTAL SETUP

4.1. Experimental description

In this work we are comparing 5 sampling strategies and a baseline system. Each sampling strategy differs from the others by the selection of the positive and the negative sample. The positive sample can be the nearest sample in the input space, a sample with a common label with the anchor or a transformed version of the anchor. The negative sample can be a sample wich does not have a common label with the anchor or chosen thanks to semi-hard mining [27]. The different strategies are summarized in Table 1: strategy 1 (S1) is the unsupervised sampling strategy suggested by Jansen et al. using data transformation, strategy 2 (S2) is a semi-supervised sampling strategy that uses transformations when no label is available, strategy 3 (S3) is a proposed sampling strategy which uses the nearest sample as positive when the anchor is unlabeled, strategy 4 (S4) uses labels to select the positive sample and semi-hard-mining to select the negative sample and strategy 5 (S5) relies on labels for both the positive and the negative sample selection.

4.2. Dataset

We use the dataset from DCASE 2018 task 4 [2]. It is composed of 10-second audio clips. Ten different classes of events are considered. The training set is composed of a labeled set and an unlabeled set. The labeled set is composed of 1,578 human-labeled audio clips. The unlabeled set contains 14,412 clips. These clips were annotated in Audioset, but since the labels have not been verified they have been discarded in this dataset. However, the clips in Audioset have been choosen so they should contain on of the ten classes considered and the distribution per class of sound event should be close to the distribution in the labeled set. Since the labels in Audioset can be very noisy, the distribution might not be exactly similar. The test set contains 288 labeled clips and the evaluation set contains 880 labeled clips. Both of these subsets have a similar distribution as the labeled training set and time coded labels. We are comparing audio tagging performance, therefore these labels are converted to clip-level labels (an event is present or not during the clip). We use the test set to validate our training and the evaluation set to perform the evaluation of the system.

4.3. Features

The audio signals are mono-channel and sampled at 44,100 Hz. From this we compute a fast Fourier transform on 25 ms windows and a step size of 10ms. We then compute log-mel spectrograms features with 64 filters. Every 10-seconds segment is then divided in 0.96 s subsegments during training as in Jansen et al. [13]. We assume that each 0.96 s subsegment has the same label as the full 10-seconds clip. However, this might be an issue for some classes which correspond to short events. The data transformation settings are similar to Jansen et al. [13] ($\sigma = 0.5$, $S = 10$, $\alpha = 0.25$ and $\Delta t = 10s$).

4.4. Model architecture

The baseline is a CRNN inspired by DCASE 2018 task 4. The CNN is composed of 3 convolutional layers with 64 3x3 filters, a pooling of 4 in frequency and no pooling in time. The RNN is a one layer bidirectional gated recurrent unit (BiGRU) with 64 cells followed by a fully connected layer with 10 cells to tag the audio clips, the final prediction is the average over the frames. The classification loss is the binary cross entropy.

In the case of embedding computation, a CNN similar to the baseline is trained using the triplet loss. The triplet loss margin is set to 1 for all models. We train a classifier using the embeddings as input. The classifier is a RNN using two layers of bidirectional gated recurrent unit (BiGRU) with 64 cells followed by a fully connected layer with 10 cells to tag the audio clips, the final prediction is the average over the frames. A single RNN layer was not enough parameters to be optimized to perform the classification on embeddings, two layers have been choosen which stays a small network. Note that this 2 layers configuration we tried with the baseline but it did not perform as well as a single RNN layer in that case.

All classifiers are trained on the 1,578 labeled files from DCASE 2018 task 4 training set.

4.5. Metric

The audio tagging performance is evaluated with F1-score at clip level [9] computed class-wise over the whole evaluation set:

$$F1_c = \frac{2TP_c}{2TP_c + FP_c + FN_c}$$

where TP_c, FP_c and FN_c are the number of true positives, false positives and false negative for a given class of event c over the whole evaluation set, respectively. The final score is the F1-score averaged over class regardless of the number of events per class (macro averaged):

$$F1_{macro} = \frac{\sum_{c \in \mathscr{C}} F1_c}{C}$$

where $\mathscr{C}$ is the classes ensemble and C the number of classes.

4.6. Results

The first experiment compare the performance of the different strategies depending on the amount of unlabeled data. In the first case we use 31,560 triplets. For S1, S2, S3 it corresponds using the same number of labeled and unlabeled triplets and taking each sample as an anchor once. For systems S4 and S5 we use each sample twice

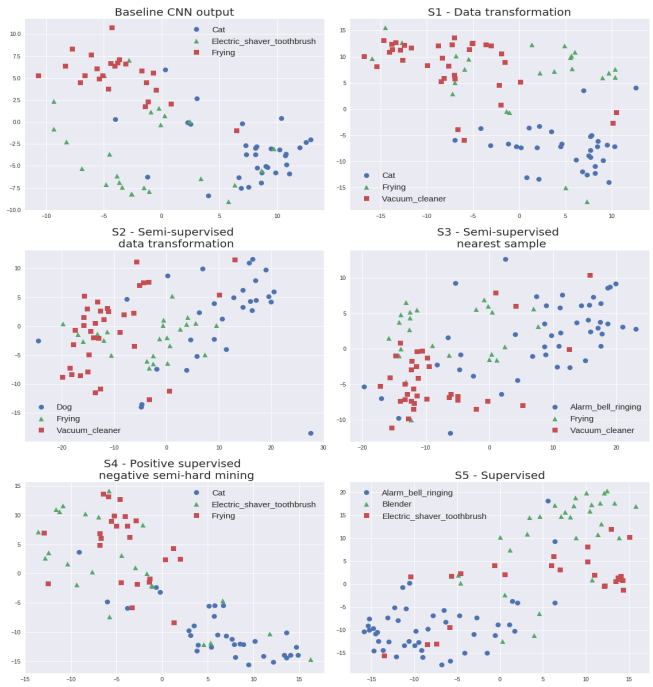

Fig. 2: Class separation for different embeddings (t-sne)

as an anchor but with different positive and negative samples. In the second and third case we use 85,780 and 159,900 triplets, respectively. For S1, S2 and S3, we use 15,780 labeled triplets, others are unlabeled triplets. For systems S4 and S5 we approximately use each sample as an anchor 4.5 times and 9 times for the second and third case, respectively, but with different positive and negative samples. Macro F1-score are presented in Table 2., the baseline has a macro F1-score of 49.61% and the classifier applied on log-mel spectrograms has a macro F1-score of 35.29%.

Sampling strategy / Nb triplets	S1	S2	S3	S4	S5
31,560	52.32	**54.35**	50.96	42.47	53.59
85,780	51.44	**54.57**	46.13	42.72	51.85
159,900	51.62	**52.27**	38.59	42.19	49.95

Table 2: Macro f-score on the evaluation set

In Figure 2 we show different embeddings learned with the different sampling strategies. For simplicity, we chose to show only the 3 more distinctive classes for each strategy on the graph. This figure presents embeddings obtained with systems trained using 31,560 triplets. We can make a correlation between the difference of results between systems obtain in Table 2 and embeddings in the graph.

All cases in Table 2 show the benefit of semi-supervised triplet sampling. Indeed, we can see that semi-supervised sampling using data transformation outperforms unsupervised sampling using data transformation. It means that labels bring meaningful information to obtain embeddings that are suitable for classification. In all cases it is better to take the transformed version of the data as positive sample than the nearest sample. This can be due to the lack of variability, to overcome this problem, we could take one of the nearest

	U	*7,890*	*15,780*	*19,725*	*23,670*
	L	*23,670*	*15,780*	*11,835*	*7,890*
S2		55.16±0.7	54.35±0.7	47.3±6.2	17.42±26.8
S3		51.58±2.2	50.96±2.0	30.02±19.8	0.0±0.0

Table 3: Macro f-score on the evaluation set with varying number of labeled (L) and unlabeled (U) triplets for strategies 2 and 3. Experiences have been launched 3 times to get 95% confidence score.

samples. S4 shows semi-hard mining does not seem suitable when taking a labeled positive sample. In the first case, we use only twice the same sample to create triplets, we are in a scenario where the supervised embedding makes sense. The second and third cases allow us to compare the different approaches and how to deal with more unlabeled data. Indeed, we see that sampling strategy S2 benefits from unlabeled data to a certain point. However, all other systems seem to be less adapted to the task when using more unlabeled data.

The unsupervised data transformation sampling strategy can be seen as a way of having a general embedding of the data, it is a good regularizer to be able to generalize in a task, and that can explain the performance of the semi-supervised data transformation sampling strategy. In the case of taking the nearest sample to create the positive, more data means closer samples. It can explain the reduction of performance when we gradually add unlabeled data. Indeed, in this case, the nearest sample can be closer to the anchor than a transformed sample. To overcome this problem, we can think about taking the nearest sample of the anchor with a constraint of being further away than transformed data samples. Results of S5 can be explained by an overfitting of the training and validation set because we see each anchor many times when we increase the number of triplets. We can also notice that most of the experiments outperform the baseline and the best experiment uses a semi-supervised setting with transformed data (S2) but without using all the unlabeled data.

In Table 3 we study the impact of the ratio between labeled and unlabeled data. We take 31,560 triplets with different proportions of labeled and unlabeled data. We can see the importance of labeled data in this experiment. S3 is much more sensitive to the number of labeled data than S2. We can see that when we do not have enough labeled data, performance completely drops. The amount of labeled data needed for a class seems to depend on the number, duration and overlap with other classes. When using more labeled data, performance is close to the supervised case with a small regularization from the unlabeled data, which can explain the best performance.

5. CONCLUSION

In this paper, we introduced semi-supervised sampling strategies to create triplets. We compared their performance and analyze the impact of different strategies to select the positive and negative samples. Semi-supervised sampling strategies were shown to outperform both supervised and unsupervised strategies. The ratio between labeled and unlabeled data can then have a great impact on the final performance. In the next step we could further investigate the interactions between the different methods to select positive and negative samples and explore their combination to improve the embeddings. This work could also be extended to a semi-supervised learning algorithm which can combine the triplet loss and the classification loss in order to benefit from both the embedding learning aspect and the classification targeting learning.

6. REFERENCES

[1] Tuomas Virtanen, Mark Plumbley, and Dan Ellis, *Computational Analysis of Sound Scenes and Events*, Springer, 2017.

[2] Romain Serizel, Nicolas Turpault, Hamid Eghbal-Zadeh, and Ankit Parag Shah, "Large-Scale Weakly Labeled Semi-Supervised Sound Event Detection in Domestic Environments," in *In Proc. DCASE Workshop*, 2018.

[3] Il-Young Jeong and Hyungui Lim, "Audio tagging system for DCASE 2018: focusing on label noise, data augmentation and its efficient learning," Tech. Rep., DCASE Challenge, 2018.

[4] Matthias Dorfer and Gerhard Widmer, "Training general purpose audio tagging networks with noisy labels and iterative self-verification," Tech. Rep., DCASE Challenge, 2018.

[5] Sharath Adavanne, Giambattista Parascandolo, Pasi Pertil, Toni Heittola, and Tuomas Virtanen, "Sound event detection in multichannel audio using spatial and harmonic features," in *In Proc. DCASE Workshop*, 2016.

[6] Yong Xu, Qiuqiang Kong, Wenwu Wang, and Mark D. Plumbley, "Large-scale weakly supervised audio classification using gated convolutional neural network," Tech. Rep., DCASE Challenge, Oct. 2017.

[7] Turab Iqbal, Qiuqiang Kong, Mark D. Plumbley, and Wenwu Wang, "Stacked convolutional neural networks for general-purpose audio tagging," Tech. Rep., DCASE Challenge, 2018.

[8] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Sckinger, and Roopak Shah, "Signature Verification using a "Siamese" Time Delay Neural Network," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 7, no. 4, pp. 8, 1993.

[9] 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 *CVPR*, Columbus, OH, USA, 2014, pp. 1386–1393, IEEE.

[10] Zuheng Ming, Joseph Chazalon, Muhammad Muzzamil Luqman, Muriel Visani, and Jean-Christophe Burie, "Simple Triplet Loss Based on Intra/Inter-Class Metric Learning for Face Verification," in *In Proc. ICCV*. 2017, pp. 1656–1664, IEEE.

[11] Herve Bredin, "TristouNet: Triplet loss for speaker turn embedding," in *In Proc. ICASSP*, New Orleans, LA, 2017, pp. 5430–5434, IEEE.

[12] R. Lu, K. Wu, Z. Duan, and C. Zhang, "Deep ranking: Triplet MatchNet for music metric learning," in *In Proc. ICASSP*, 2017, pp. 121–125.

[13] Aren Jansen, Manoj Plakal, Ratheet Pandya, Daniel P. W. Ellis, Shawn Hershey, Jiayang Liu, R. Channing Moore, and Rif A. Saurous, "Unsupervised Learning of Semantic Audio Representations," in *In Proc. ICASSP*, 2017, pp. 126–130, arXiv: 1711.02209.

[14] Jort F. Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter, "Audio Set: An ontology and human-labeled dataset for audio events," in *In Proc. ICASSP*, 2017.

[15] Eduardo Fonseca, Jordi Pons, Xavier Favory, Frederic Font, Dmitry Bogdanov, Andres Ferraro, Sergio Oramas, Alastair Porter, and Xavier Serra, "Freesound datasets: a platform for the creation of open audio datasets," in *In Proc. of the 18th International Society for Music Information Retrieval Conference (ISMIR 2017)*, 2017, pp. 486–493.

[16] Eduardo Fonseca, Manoj Plakal, Frederic Font, Daniel P. W. Ellis, Xavier Favory, Jordi Pons, and Xavier Serra, "General-purpose Tagging of Freesound Audio with AudioSet Labels: Task Description, Dataset, and Baseline," in *In Proc. DCASE Workshop*, 2018, arXiv: 1807.09902.

[17] Justin Salamon, Duncan MacConnell, Mark Cartwright, Peter Li, and Juan Pablo Bello, "Scaper: A library for soundscape synthesis and augmentation," in *In Proc. WASPAA*, New Paltz, NY, 2017, pp. 344–348, IEEE.

[18] 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 22*, Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, Eds., pp. 1096–1104. Curran Associates, Inc., 2009.

[19] Y. Xu, Q. Huang, W. Wang, P. Foster, S. Sigtia, P. J. B. Jackson, and M. D. Plumbley, "Unsupervised Feature Learning Based on Deep Models for Environmental Audio Tagging," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1230–1241, 2017.

[20] J. Salamon and J. P. Bello, "Unsupervised feature learning for urban sound classification," in *In Proc. ICASSP*, 2015, pp. 171–175.

[21] Tatsuya Komatsu, Takahiro Toizumi, Reishi Kondo, and Yuzo Senda, "Acoustic event detection method using semi-supervised non-negative matrix factorization with mixtures of local dictionaries," in *In Proc. DCASE Workshop*, 2016, pp. 45–49.

[22] Avital Oliver, Augustus Odena, Colin Raffel, Ekin D. Cubuk, and Ian J. Goodfellow, "Realistic Evaluation of Deep Semi-Supervised Learning Algorithms," in *In Proc. Workshop track - ICLR 2018*, 2018, arXiv: 1804.09170.

[23] Zixing Zhang and Bjorn Schuller, "Semi-supervised learning helps in sound event classification," in *In Proc. ICASSP*, Kyoto, Japan, 2012, pp. 333–336, IEEE.

[24] Pedro J Moreno and Shivani Agarwal, "An experimental study of semi-supervised EM algorithms in audio classification and speaker identification," in *In Proc. Workshop on the Continuum from Labeled to Unlabeled Data*, 2003, p. 10.

[25] Lu JiaKai, "Mean teacher convolution system for DCASE 2018 task 4," Tech. Rep., DCASE Challenge, 2018.

[26] Rachid Riad, Corentin Dancette, Julien Karadayi, Neil Zeghidour, Thomas Schatz, and Emmanuel Dupoux, "Sampling strategies in Siamese Networks for unsupervised speech representation learning," in *In Proc. Interspeech*, 2018, arXiv: 1804.11297.

[27] Florian Schroff, Dmitry Kalenichenko, and James Philbin, "FaceNet: A Unified Embedding for Face Recognition and Clustering," in *In Proc. CVPR*, 2015, pp. 815–823, arXiv: 1503.03832.

[28] Sunil Thulasidasan and Jeffrey Bilmes, "Acoustic classification using semi-supervised Deep Neural Networks and stochastic entropy-regularization over nearest-neighbor graphs," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, Mar. 2017, pp. 2731–2735, IEEE.