INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Do sound event representations generalize to other audio tasks? A case study in audio transfer learning

Anurag Kumar[†*], *Yun Wang*[‡*], *Vamsi Krishna Ithapu*[†], *Christian Fuegen*[‡]

[†]Facebook Reality Labs Research, Redmond, WA, USA
[‡] Facebook Applied AI Research, Menlo Park, CA, USA
{anuragkr,yunwang,ithapu,fuegen}@fb.com

Abstract

Transfer learning is critical for efficient information transfer across multiple related learning problems. A simple, yet effective transfer learning approach utilizes deep neural networks trained on a large-scale task for feature extraction. Such representations are then used to learn related downstream tasks. In this paper, we investigate transfer learning capacity of audio representations obtained from neural networks trained on a large-scale sound event detection dataset. We build and evaluate these representations across a wide range of other audio tasks, via a simple linear classifier transfer mechanism. We show that such simple linear transfer is already powerful enough to achieve high performance on the downstream tasks. We also provide insights into the attributes of sound event representations that enable such efficient information transfer.

Index Terms: transfer learning, representation learning, sound events, audio

1. Introduction

Building task-agnostic representations that can generalize across multiple learning problems has been critical in advancing and applying machine learning techniques in a variety of domains. However, often the design of neural networks are driven by low-level tasks in a given problem domain. For instance, a variety of robust deep networks for low-level tasks like visual object recognition, co-segmentation, etc., have been designed and thoroughly evaluated. While these carefully designed networks are very successful on such low-level tasks, the need for frameworks and algorithmic procedures that combine and *transfer* information across multiple low-level tasks to help tackle higher-level tasks (like multi-sensory scene parsing, activity understanding, etc.), remains a challenging problem. Moreover, this notion of combining or sharing knowledge is helpful for training systems under limited and noisy-labeled data as well [1, 2].

Transfer learning is possibly the best suited framework for building such shareable representations, and has been studied comprehensively in the domains of computer vision and natural language processing [3, 4, 5]. Taking vision as an example, transfer learning has been applied to a wide range of problems like scene understanding and action summarizing [6, 7, 8], few shot learning and noisy label learning [1, 2], to mention a few. Systems trained on low-level vision tasks such as object detection and classification serve as the source task from which the knowledge is transferred. This is mainly because of the availability of large-scale annotated datasets for such tasks.

In acoustics, specifically, in audio machine learning, transfer learning has been studied to a relatively lesser extent. One possible reason is the less obvious choice of low-level source task. Nevertheless, transfer learning has gained traction in recent times in the audio machine learning. It's been studied in isolated contexts such as sound event detection (SED) [9, 10, 11], music tagging [12] and emotion recognition [13]. Nevertheless, we do not yet understand the nuances of generalized representations that capture structural similarity between source and target audio tasks. While prior works have employed transfer learning for different audio tasks, knowledge transfer from a *single low-level audio task* to a variety of other audio tasks has not been studied comprehensively. This forms the key motivation of this paper. Clearly, the capability of shareable representations may depend entirely upon the choice of the tasks used for evaluation. We hypothesize that SED representations to have substantial capabilities to generalize to other related audio task. We choose SED as the source task for two reasons: *first*, sound event datasets are among the largest available audio datasets, thereby providing a large enough database for learning "robust" representations; *second*, learning sound events implicitly entails learning low-level acoustic phenomena, which, in principle, amounts to capturing a significant amount of information in an audio snippet. We refer to the SED as the *source* task and explore their generalization power to other audio*target* tasks. Besides, benchmarking capabilities of SED representations for audio transfer learning, we aim to provide interesting insights into the target tasks and the relationship between the target tasks and the source task.

Keeping the above motivations in mind, we standardized the transfer learning process in the following way. We train neural networks for a large scale SED task and transfer the representations obtained from these networks for any given audio to the target tasks. To reduce the bias in the design of the SED model itself, we train and analyze results through two separate networks. We constrain ourselves to training *linear classifiers* for each target task using the representations obtained from the SED networks. Linear classifiers allow a simple and grounded way to evaluate the efficacy of these audio representations for knowledge transfer. Even using a simple non-linear mapping for transfer limits us from disentangling the power of sound event representations vs. the power of non-linear transfer itself. Finally, Finally, we consider a variety of target tasks to help to better understand the effectiveness as well as the limitations of these audio representations obtained from SED models.

In Section 2 we introduce the networks used for SED, and Section 3 discusses the target tasks. We evaluate the transfer of event representations in Section 4, and provide insights with some visualizations in Section 5. Section 6 concludes the paper.

2. Source Task & Audio Representations

As said above, the source task is sound event detection, and representations are obtained from two state-of-the-art deep networks trained on the AudioSet [14] corpus, which contains 2 million training recordings of 527 types of sound events. The two

* Equal contribution.

http://dx.doi.org/10.21437/Interspeech.2021-347

models, TALNet [15] and WEANet-SUSTAIN [16], are briefly summarized below.

2.1. TALNet

TALNet [15] is a deep convolutional recurrent network for SED. The network takes logmel spectrograms as inputs; the logmel spectrograms have 40 frames per second and 64 frequency bins. The input features are passed through 10 convolutional layers, 5 pooling layers, and one bidirectional GRU layer. The output of the GRU has 10 frames per second, each being a $1,024$-dimensional vector. These vectors are further processed by a fully connected layer to calculate the probability of each type of sound event at each frame, and these probabilities are aggregated over an entire recording using a linear softmax pooling function to yield global event probabilities. We extract the $1,024$-dimensional output of GRU layer (averaged over time) as the learned transferable representation for any given input. Before using these representations to train linear classifiers, we first normalize them to have zero mean and unit variance across training data, then normalize each vector to have unit l_2-norm.

2.2. WEANet-SUSTAIN

The second network we use is WEANet-SUSTAIN [16]. This network also takes 64-dimensional logmel spectrograms as input, but the frame rate is 100 frames per second. The network is a fully convolutional neural network with a class-specific attention layer. The input is first processed by 4 blocks of layers (B1 to B4); each block consists of 2 convolutional layers followed by max pooling. These blocks are followed by 4 more blocks (B5 to B8) of only convolutional layers. At this stage we get *segment-level* outputs, which are then combined through a class-specific attention mechanism to produce a *recording-level* output. The network is trained using a sequential self-teaching approach leading to robust generalization. We use WEANet's $2,048$-dimensional hidden representation from the output of block B5 (average/max pooled over time and l_2-normalized) for transferring to target tasks.

3. Transfer Learning to Target Tasks

Our motivation here is to understand the knowledge transfer from SED to a variety of other audio downstream tasks, focusing on sounds, actions, music, etc., and with small as well as large-scale datasets. The representations from TALNet and WEANet are used to train **linear classifiers** for these tasks. This helps us in focusing on representative power of the learned representations for the target tasks rather than relying on strong classifiers to obtain good performance.

3.1. Sound Event Classification

Although SED on AudioSet is our source task, we also consider sound event classification on 3 other datasets as target tasks: ESC-50 [17], Urbansound [18] and FSDKaggle2019 [19]. The domain mismatch between these datasets and AudioSet makes transfer learning non-trivial. The FSDKaggle2019 dataset, in particular, is more challenging. Unlike the other two, FSDKaggle2019 is a multi-label dataset, where each recording can have more than one label. It also consists of a "curated" set and a "noisy label" set: the former contains audio recordings carefully labeled by humans, whereas the latter can have wrongly labeled audio examples. An estimated 60% of all the labels are wrong, making it a very challenging task.

Table 1: *Comparison with state-of-the-art methods on AudioSet*

Method	MAP	MAUC
Ford *et. al* [26]	0.380	0.970
TALNet [15]	0.386	0.971
WEANet-SUSTAIN [16]	0.398	0.972

3.2. Acoustic Scene Classification

Acoustic scenes are often composed of a mixture of sounds thereby exhibiting complex acoustic characteristics. While this implicit relation between sound events and acoustic scenes can provide some nuanced understanding of acoustic scenes, it remains to be seen if representations based on SED can capture enough information for good scene classification performance. Here we evaluate the transferability of SED to acoustic scene classification using Task 1a of the 2019 DCASE challenge [20].

3.3. Music Tagging

This target task aims at tagging audio recordings with different music genres, instruments, moods, etc. It adds on the variability of target labels considered in this work. We use the well-known MagnaTagATune dataset [21] for transfer evaluation. This is a multi-label dataset, where each recording can belong to a genre class as well as multiple instrument classes, at the same time. We use the top 50 tags of this dataset in our experiments.

3.4. Human Action Classification Using Audio

The goal of this task is to recognize human actions such as "ice skating" and "playing guitar" in video recordings. We use the most recent version of the Kinetics dataset (Kinetics700), a widely used benchmark for action classification [22]. It is a large-scale dataset with over 550k 10-second clips from 700 action classes. This problem has been primarily tackled from a visual perspective, although some multimodal approaches have also been proposed [23, 24].

In this paper, we explore *audio-only* recognition of human actions. This is interesting in several aspects. To the best of our knowledge, this is perhaps the *first work* that explicitly tries to link human actions and sound events. In principle, similar to ImageNet [25] based pre-trained models being used for visual-driven action classification, we hypothesize that pre-trained SED models can help advance the state of audio-driven action classification. Further, being a large-scale dataset with over 550k clips, transferring SED representations to this task via linear classifiers helps characterize the efficacy of direct classification of actions vs. action classification based on knowledge of sounds.

4. Experiments

4.1. Datasets and Setup

For fair comparison, we follow the standard training/validation/test split and use performance metrics defined for each dataset. When such information is unavailable, we follow the most prevailing setup from previous works. For ESC-50 and Urbansound, we perform 5-fold and 10-fold cross-validation following the predefined folds, and report the average accuracy across all folds. For FSDKaggle2019, the "public" test set is used for validation. For MagnaTagATune, we use the $12:1:3$ split for training, validation and testing, as was done in several prior works [29, 30, 31]. For Kinetics700, we take out $20,525$ examples from the training set to use as a validation set. All models are implemented in PyTorch, and hyperparameters are tuned using the validation sets.

Table 2: *Summary of the target tasks and the performance of TALNet and WEANet-SUSTAIN, compared with some previous works.*

Task	Dataset	# Classes	Metric		TALNet	WEANet	Prior Work (Uses TL?)	
Sound Events	ESC-50	50	Accuracy		91.0	94.1	94.7 [11]	✓
	Urbansound	10	Accuracy		85.2	85.2	85.1 [27]	✓
	FSDKaggle2019	80	lwlrap	Curated	72.0	72.8	54.2 [19]	✗
				Noisy	51.0	50.3	31.2 [19]	
Acoustic Scenes	DCASE2019	10	Accuracy		65.8	68.0	58.9 [11]	✓
Music Tagging	MagnaTagATune	50	MAUC		91.5	91.5	90.2 [12]	✓
Human Actions	Kinetics700	700	Accuracy	Top-1	15.9	18.0	21.9 [28]	✗
				Top-5	30.5	33.0	36.9 [28]	

4.2. AudioSet Models

The details of the TALNet and WEANet models trained on AudioSet can be found in [15] and [16] respectively. Table 1 shows the performance of the two models on AudioSet. In this work, we re-trained TALNet applying SpecAugment [32] to the inputs. We masked out one frequency band of at most 16 bins, and one time interval of at most 2 seconds. This improves the the mean average precision (MAP) from 0.359 in [15] to 0.386.

4.3. Results

Table 2 summarizes results for all target tasks. For brevity, we also show performance from prior works that most closely relate to the proposed approach. To our knowledge, no such transfer learning work exists for FSDKaggle2019 and Kinetics700 (shown by ✗ under the "Uses TL?" column); for the other tasks, the reported baselines are from prior works using transfer learning. However, these transfer learning processes are often much more complex compared to our simple linear classifiers trained on learned representations from TALNet and WEANet.

For the target tasks of sound event classification, the linear classifiers built upon TALNet and WEANet representations give similar performance as prior transfer learning works on the ESC-50 and Urbansound datasets. These numbers also come close to the state-of-the-art (SOTA) on these datasets. Note that, [27] applies transfer via networks pre-trained on images. On FSDKaggle2019, we compare with the baseline approach in [19], and our results are 34% and 63% superior.

For the target task of acoustic scene classification, audio representations from TALNet and WEANet give 6.9% and 9.1% better performance compared to [11]] (also trained on AudioSet).

On the music tagging task, TALNet- and WEANet-based representations lead to better performance compared to the transfer learning proposal from [12]. Interestingly, [12] uses a large-scale music tagging as source task (the Million Song Dataset [33]), which is very similar to the target task. While AudioSet also contains a fairly large number of music examples, this clearly shows that it is possible to construct good representations for music tagging via a more general-purpose source task like SED.

For Kinetics700, the performance of linear classifiers with audio representations from TALNet and WEANet is inferior to training an Xception model from scratch [28]. This is expected for a large-scale dataset such as Kinetics700. But it is noteworthy that these audio representations can give competitive results with just linear classifiers, illustrating the shared information between the two tasks – this has not been previously studied or observed. Some action classes, such as "rolling eyes" and "peeling banana", do not exhibit specific acoustic signatures and are hard to detect through any audio-only approach. The action "playing bagpipes" achieves the highest top-1 accuracy of 87.5% (using WEANet features). This is not surprising because the "bagpipes" event gets the highest performance on the source AudioSet task as well. In Sec. 5.2 we will provide some qualitative interpretation of the relationship between the source SED task and target Kinetics700.

5. Analysis and Visualizations

Overall, the results on target tasks shows that, in most cases, simple linear classifiers that transfer TALNet and WEANet representations can give competitive, or marginally better, results compared to previously published numbers on these datasets. To bring further insights, we provide some more analysis here.

We first show that the representations can capture semantics-driven proximity relationships among the target labels. We show this through the linear classification weights learned for each class in the target task. We also analyze the correlation between the target task labels and source task sound events, and illustrate that to a certain extent we can explain the sound events that contribute to specific target labels. Keeping page-limit in mind, we summarize the analysis for TALNet-based representations alone; similar results were obtained for WEANet representations.

5.1. Clustering of Target Labels

The weight matrix of the linear model learned for a target task is essentially a *condensed* representation of the target labels' semantics. We denote this as $W \in \mathbb{R}^{C \times D}$, where C is the number of classes in the target task and D is the dimensionality of audio representations.

Consider the music tagging task with TALNet representations as an example. The learned weight matrix W has a size of $50 \times 1,024$; each $1,024$-dimensional row vector essentially represents a music tag. If the TALNet representations do allow for learning the semantics of the music tags, then in this $1,024$-D space, semantically similar tags should be close to each other. Given this hypothesis, we perform a hierarchical clustering in this space. The resultant dendrogram is shown in Fig. 1. It clearly shows the hypothesized semantically meaningful grouping of classes. In particular, we see that synonymous tags such as "woman", "female", "female vocal", and "female voice" are clustered together. Similarly, instruments of classical music (*e.g.* "violin" and "harp") form a cluster, and so do words describing vibrant music (*e.g.* "drums", "beat", and "dance"). This shows that our setup and source task are robust in learning general task-agnostic (abstract) information about audio and sounds.

We performed a similar analysis for the human action recognition task, and the resulting dendrogram is shown in Fig. 2. Given the rather large number of action types in Kinetics700, we only show a few action names. Observe that we can recognize semantic clusters at both macro and micro levels. At the macro level, one can summarize each large cluster (marked by colors) with a few words. For example, most of the actions in teal are *housework*, and most actions in purple are *sports on land*. At a

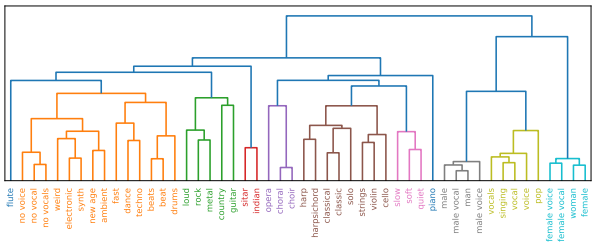

Figure 1: *Hierarchical clustering dendrogram of the MagnaTagATune music tags.*

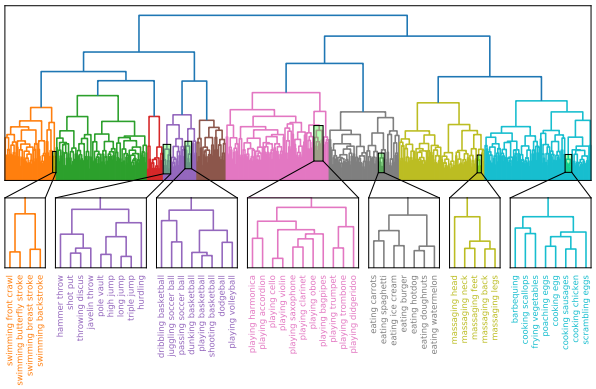

Figure 2: *Hierarchical clustering dendrogram of the Kinetics700 actions. To avoid clutter, action names are only shown for some small clusters.*

finer level of resolution, we can see small clusters representing *ball sports* and *track-and-field* sports form within the large purple cluster. Surprisingly, action classes such as massaging head, neck, back, legs, and feet, which do not correspond to sound signals, are also well clustered. A possible explanation is that these actions often come with audio tracks containing relaxing music, and the audio representations are able to exploit such acoustic cues to support the recognition of visual actions. Fig. 3 provides an alternate view by running t-SNE [34] on the learned class representations. The color coding follows the dendogram coloring scheme. Once again, we notice that closely related events cluster together. In summary, deep audio representations from large scale SED task may directly be used to learn semantic relationships among human-actions using a linear classification transfer methodology.

5.2. Correlation Between Target Labels and Sound Events

While we have shown that the audio representations from our source task contain adequate information for recognizing music tags and actions, it is hard to interpret and rationalize the evidence that the transfer learning models use to predict a target label. To understand this, we study the correlation between the target labels and sound events to see if predictions of target labels are often supported by the existence of certain sound events.

We compute the cosine similarity between the following two sets of vectors: $1,024$-D representations of music tags and actions, taken from the rows of the weight matrices of the two linear classifiers; and the $1,024$-D representations of the 527 AudioSet sound events, taken from the rows of the weight matrix of the final fully connected layer of TALNet. Before computing the cosine similarity, we perform mean-variance normalization.

A part of the resulting cosine similarity matrix is shown in Fig. 4. Rows represent sound events, and columns represent

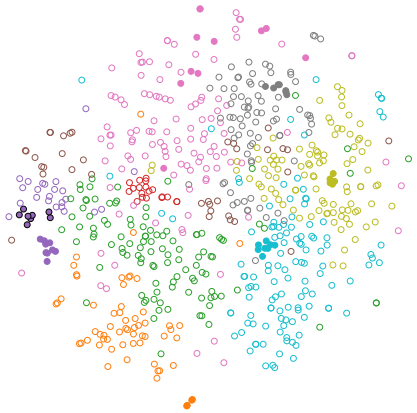

Figure 3: *t-SNE plot of the Kinetics action classes.*

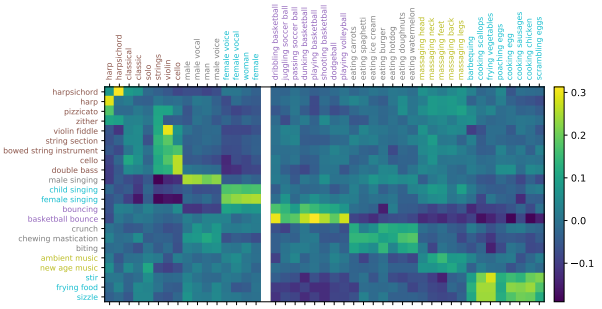

Figure 4: *Cosine similarity between some AudioSet sound events (rows) and MagnaTagATune music tags / Kinetics700 actions (columns).*

music tags and actions. The rows and columns are sorted according to the dendrograms produced by hierarchical clustering, so that similar music tags, actions, and sound events are next to each other. We can immediately recognize blocks of high similarity values (often ≥ 0.2), manifesting themselves in yellow and light green cells. Considering that two random vectors in a high-dimensional space are usually nearly orthogonal, cosine similarity values above 0.2 are remarkably large. This figure demonstrates that many music tags or actions can be explained by a single or a few sound events, and transfer learning is able to discover such correspondences. For example, various actions of cooking exhibit high similarity to events like "sizzle"; actions of eating are characterized by "chewing". Actions like massaging are often accompanied by relaxing music such as "new age music", although the similarity is not as high. Overall, the correlation analysis provides a quantifiable way to interpret the correspondences between music tags, actions, and sound events.

6. Conclusion

We demonstrated that it is possible to transfer knowledge from sound event detection (SED) task to a wide range of other audio tasks, including acoustic scene classification, music tagging and human action recognition. Using a simple linear classifier on audio representations obtained from SED models, we are able to achieve performance that is comparable or better than the state of the art on several datasets. The linear classification system also provides a lucid way to interpret the suitability of these representations for target tasks. By visualizing the classifier learned representations learned for target tasks, we found meaningful structures that reflect proximity relationships among music genres, instruments, moods, as well as various human actions. Lastly, it was possible to identify the unique sound events that contribute to the learning of downstream tasks.

7. References

[1] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, "Generalizing from a few examples: A survey on few-shot learning," *ACM Computing Surveys (CSUR)*, 2020.

[2] K. Lee, X. He, L. Zhang, and L. Yang, "CleanNet: Transfer learning for scalable image classifier training with label noise," in *IEEE CVPR*, 2018.

[3] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big Data*, 2016.

[4] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, "A survey on deep transfer learning," in *International Conference on Artificial Neural Networks*. Springer, 2018, pp. 270–279.

[5] X. Zhai, J. Puigcerver, A. Kolesnikov, P. Ruyssen, C. Riquelme, M. Lucic, J. Djolonga, A. Pinto, M. Neumann, A. Dosovitskiy *et al.*, "A large-scale study of representation learning with the visual task adaptation benchmark," *arXiv preprint arXiv:1910.04867*, 2019.

[6] G. Csurka, "Domain adaptation for visual applications: A comprehensive survey," *arXiv preprint arXiv:1702.05374*, 2017.

[7] A. R. Zamir, A. Sax, W. Shen, L. J. Guibas, J. Malik, and S. Savarese, "Taskonomy: Disentangling task transfer learning," in *IEEE CVPR*, 2018.

[8] A. B. Sargano, X. Wang, P. Angelov, and Z. Habib, "Human action recognition using transfer learning with deep representations," in *International Joint Conference on Neural Networks (IJCNN)*, 2017, pp. 463–469.

[9] A. Kumar, M. Khadkevich, and C. Fügen, "Knowledge transfer from weakly labeled audio using convolutional neural network for sound events and scenes," in *ICASSP*, 2018, pp. 326–330.

[10] P. Arora and R. Haeb-Umbach, "A study on transfer learning for acoustic event detection in a real life scenario," in *IEEE 19th International Workshop on Multimedia Signal Processing (MMSP)*, 2017, pp. 1–6.

[11] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *arXiv preprint arXiv:1912.10211*, 2019.

[12] J. Lee and J. Nam, "Multi-level and multi-scale feature aggregation using pretrained convolutional neural networks for music auto-tagging," *IEEE Signal Processing Letters*, 2017.

[13] S. Latif, R. Rana, S. Younis, J. Qadir, and J. Epps, "Transfer learning for improving speech emotion classification accuracy," *arXiv preprint arXiv:1801.06353*, 2018.

[14] J. Gemmeke, D. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio Set: An ontology and human-labeled dataset for audio events," in *ICASSP*, 2017, pp. 776–780.

[15] Y. Wang, J. Li, and F. Metze, "A comparison of five multiple instance learning pooling functions for sound event detection with weak labeling," in *ICASSP*, 2019, pp. 31–35.

[16] A. Kumar and V. Ithapu, "A sequential self teaching approach for improving generalization in sound event recognition," in *International Conference on Machine Learning (ICML)*. PMLR, 2020, pp. 5447–5457.

[17] K. Piczak, "ESC: Dataset for environmental sound classification," in *23rd ACM International Conference on Multimedia*, 2015, pp. 1015–1018.

[18] J. Salamon, C. Jacoby, and J. Bello, "A dataset and taxonomy for urban sound research," in *22nd ACM International Conference on Multimedia*, 2014, pp. 1041–1044.

[19] E. Fonseca, M. Plakal, F. Font, D. Ellis, and X. Serra, "Audio tagging with noisy labels and minimal supervision," *arXiv preprint arXiv:1906.02975*, 2019.

[20] A. Mesaros, T. Heittola, and T. Virtanen, "A multi-device dataset for urban acoustic scene classification," in *Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018)*, November 2018. [Online]. Available: https://arxiv.org/abs/1807.09840

[21] E. Law, K. West, M. Mandel, M. Bay, and J. Downie, "Evaluation of algorithms using games: The case of music tagging." in *ISMIR*, 2009, pp. 387–392.

[22] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev *et al.*, "The Kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[23] B. Ghanem, J. C. Niebles, C. Snoek, F. C. Heilbron, H. Alwassel, V. Escorcia, R. Krishna, S. Buch, and C. D. Dao, "The ActivityNet large-scale activity recognition challenge 2018 summary," *arXiv preprint arXiv:1808.03766*, 2018.

[24] F. Xiao, Y. J. Lee, K. Grauman, J. Malik, and C. Feichtenhofer, "Audiovisual SlowFast networks for video recognition," *arXiv preprint arXiv:2001.08740*, 2020.

[25] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *IEEE CVPR*, 2009, pp. 248–255.

[26] L. Ford, H. Tang, F. Grondin, and J. R. Glass, "A deep residual network for large-scale acoustic scene analysis." in *INTERSPEECH*, 2019, pp. 2568–2572.

[27] K. Palanisamy, D. Singhania, and A. Yao, "Rethinking CNN models for audio classification," *arXiv preprint arXiv:2007.11154*, 2020.

[28] Z. Qiu, D. Li, Y. Li, Q. Cai, Y. Pan, and T. Yao, "Trimmed action recognition, dense-captioning events in videos, and spatio-temporal action localization with focus on activitynet challenge 2019," *arXiv preprint arXiv:1906.07016*, 2019.

[29] S. Dieleman and B. Schrauwen, "End-to-end learning for music audio," in *ICASSP*, 2014.

[30] Z. Wang, S. Muknahallipatna, M. Fan, A. Okray, and C. Lan, "Music classification using an improved CRNN with multi-directional spatial dependencies in both time and frequency dimensions," in *International Joint Conference on Neural Networks (IJCNN)*, 2019, pp. 1–8.

[31] Q. Wang, F. Su, and Y. Wang, "Hierarchical attentive deep neural networks for semantic music annotation through multiple music representations," *International Journal of Multimedia Information Retrieval*, vol. 9, no. 1, pp. 3–16, 2020.

[32] D. Park, W. Chan, Y. Zhang, C. Chiu, B. Zoph, E. Cubuk, and Q. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[33] T. Bertin-Mahieux, D. P. Ellis, B. Whitman, and P. Lamere, "The million song dataset," 2011.

[34] L. Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.