LEARNING CONTEXTUAL TAG EMBEDDINGS FOR CROSS-MODAL ALIGNMENT OF AUDIO AND TAGS

Xavier Favory[†], *Konstantinos Drossos*[‡], *Tuomas Virtanen*[‡], *Xavier Serra*[†]

[†]Music Technology Group, Universitat Pompeu Fabra, Spain
[‡]Audio Research Group, Tampere University, Finland

ABSTRACT

Self-supervised audio representation learning offers an attractive alternative for obtaining generic audio embeddings, capable to be employed into various downstream tasks. Published approaches that consider both audio and words/tags associated with audio do not employ text processing models that are capable to generalize to tags unknown during training. In this work we propose a method for learning audio representations using an audio autoencoder (AAE), a general word embeddings model (WEM), and a multi-head self-attention (MHA) mechanism. MHA attends on the output of the WEM, providing a contextualized representation of the tags associated with the audio, and we align the output of MHA with the output of the encoder of AAE using a contrastive loss. We jointly optimize AAE and MHA and we evaluate the audio representations (i.e. the output of the encoder of AAE) by utilizing them in three different downstream tasks, namely sound, music genre, and music instrument classification. Our results show that employing multi-head self-attention with multiple heads in the tag-based network can induce better learned audio representations.

Index Terms— representation learning, multimodal contrastive learning, audio classification

1. INTRODUCTION

An effective way to learn audio representations that can be used for sound classification involves training deep neural networks (DNNs) on supervised tasks, using large annotated datasets [1, 2, 3]. However, these datasets require a considerable amount of effort to be built and are always limited in size, hindering the performance of learned representations. Recent research approaches explore and adopt unsupervised, self-supervised or semi-supervised learning methods for obtaining generic audio representations, that later can be used for different downstream tasks [4, 5, 6, 7]. The large amount of multimedia data available online is a great opportunity for these types of approaches to learn powerful audio representations.

In the natural language and image processing fields, both supervised and unsupervised approaches enabled the creation of powerful pre-trained models, that are often employed in many different tasks [8, 9, 10]. Contrastive learning recently got a lot of attention due to its success for the unsupervised pre-training of DNNs, enabling to learn flexible representation without the need of having labels associated with the content [10, 11]. Only the definition of positive and negative data pairs is required in order to learn a model that will produce a latent space reflecting semantic characteristics. The association of images and text can be exploited for learning embeddings [12], e.g. by using a self-attention mechanism to learn context sensitive text embeddings that are then aggregated into sentence embeddings. Similar approaches have been adopted for machine listening, for instance unsupervised pre-training of transformer models can improve speech recognition performance by employing contrastive predictive coding strategies [13, 14]. In [7], the authors employ a semi-supervised sampling strategy to create triplets for benefiting automatic tagging systems. However, these approaches consider just one modality, i.e. audio.

A cross-modal method for learning and aligning audio and tag latent representations (COALA) was presented in [15]. Latent representations were learnt using autoencoders, one aiming at encoding and reconstructing spectrogram representations of sounds, and the other focusing on encoding and reconstructing a set of tags represented as multi-hot vectors. The outcomes suggest that it is possible to leverage noisy, user-generated data of audio and accompanying tags, for learning semantically enriched audio representations that can be used for different classification tasks. However, in COALA the tag-based input representation was fixed, and therefore the tag-based encoder cannot generalize to now terms that have not been seen during training. This makes the approach loose the flexibility of contrastive representation learning.

In this work we propose a method for allowing the textual generalization of cross-modal approaches, using pre-trained word embedding models which project words into semantic spaces. We propose an attention mechanism for learning

The authors would like to thank all the Freesound users that have been sharing very valuable content for many years. X. Favory and K. Drossos are grateful for the GPUs donated by NVidia. K. Drossos and T. Virtanen wish to acknowledge CSC-IT Center for Science, Finland, for computational resources.

higher-level contextualized semantic representation similarly to [12]. However, our approach relies on accompanying tags instead of text, and therefore employs a simpler approach for computing semantic embeddings. The rest of the paper is as follows. In Section 2 we present our proposed method. Section 3 describes the utilized dataset, the tasks and metrics that we employed for the assessment of the performance, the baselines that we compare our method with. The results of the evaluation is presented and discussed in Section 4. Finally, Section 5 concludes the paper and proposes future research directions.

2. PROPOSED METHOD

Our method, illustrated in Figure 1, consists of an audio encoder, $e_{\text{a}}(\cdot)$, an audio decoder, $d_{\text{a}}(\cdot)$, a pre-trained word embedding model, $e_{\text{w}}(\cdot)$, and multi-head self-attention, $\text{Att}(\cdot)$. As input to our method, we employ a dataset of N_{B} paired exampled examples, $\mathbb{G} = \{(\mathbf{X}_{\text{a}}^{n_{\text{B}}}, \mathbf{x}_{\text{w}}^{n_{\text{B}}})\}_{n_{\text{B}}=1}^{N_{\text{B}}}$, where $\mathbf{X}_{\text{a}} \in \mathbb{R}^{T_{\text{a}} \times F_{\text{a}}}$ is a time-frequency audio representation of T_{a} audio feature vectors of F_{a} features, and $\mathbf{x}_{\text{w}} = \{x_{\text{w}}^i\}_{i=1}^{T_{\text{w}}}$ is a set of T_{w} tags, x_w^i, like "techno" and "dog bark", and associated with $\mathbf{X}_{\text{a}}$[1]. Att extracts from $\mathbf{x}_{\text{w}}$ an embedding containing the context of the tags, and by using a contrastive loss we align the output $\mathbf{z}_{\text{a}} = e_{\text{a}}(\mathbf{X}_{\text{a}})$ with the output of Att. $\mathbf{z}_{\text{a}}$ is also used as an input to $\mathbf{d}_{\text{a}}$, to reconstruct $\mathbf{X}_{\text{a}}$, similarly to [15], effectively infusing $\mathbf{z}_{\text{a}}$ with both semantic (from $\mathbf{x}_{\text{w}}$) and low-level acoustic information (from the reconstruction by $\mathbf{d}_{\text{a}}$). The code of our method is available online[2].

2.1. Audio encoding and decoding

The encoder e_{a} consists of $N_{\text{e-a}}$ cascaded 2D convolutional neural networks (CNNs), $\text{CNN}_{n_{\text{e-a}}}$, with $C_{n_{\text{e-a}}}^{\text{in}}$ and $C_{n_{\text{e-a}}}^{\text{out}}$ input and output channels, respectively, a square kernel of $K_{\text{e-a}}$ size, and $S_{\text{e-a}}$ stride. The CNNs process $\mathbf{X}_{\text{a}}$ is a serial fashion, and each $\text{CNN}_{n_{\text{e-a}}}$ is followed by a batch normalization process ($\text{BN}_{n_{\text{e-a}}}$), a rectified linear unit (ReLU), and a dropout (DO) with probability p_{a}, as

$$\mathbf{H}_{n_{\text{e-a}}} = \text{DO}(\text{ReLU}(\text{BN}_{n_{\text{e-a}}}(\text{CNN}_{n_{\text{e-a}}}(\mathbf{H}_{n_{\text{e-a}}-1})))), \quad (1)$$

where $\mathbf{H}_0 = \mathbf{X}_{\text{a}}$. $\mathbf{H}_{N_{\text{e-a}}} \in \mathbb{R}_{\geq 0}^{C_{N_{\text{e-a}}}^{\text{e-out}} \times T'_{N_{\text{e-a}}} \times F'_{N_{\text{e-a}}}}$ is flattened to a vector and given as an input to a layer normalization process (LN) [16] ($\text{FFN}_{\text{e-a}}$), as $\mathbf{z}_{\text{a}} = \text{LN}(\mathbf{h}_{N_{\text{e-a}}})$, where $\mathbf{h}_{N_{\text{e-a}}}$ is the flattened $\mathbf{H}_{N_{\text{e-a}}}$, and $\mathbf{z}_{\text{a}} \in \mathbb{R}^V$, with $V = C_{N_{\text{e-a}}}^{\text{e-out}} \cdot T'_{N_{\text{e-a}}} \cdot F'_{N_{\text{e-a}}}$, is the learned audio embedding by our method. $\mathbf{z}_{\text{a}}$ is used at the employed contrastive loss, in order to be aligned with the information contained at the associated tags $\mathbf{x}_{\text{w}}$, and as an input to d_{a} in order to encode low-level acoustic features in $\mathbf{z}_{\text{a}}$, similarly to [15].

The decoder d_{a} takes as an input $\mathbf{z}_{\text{a}}$ and processes it through a series of $N_{\text{e-a}}$ transposed 2D CNNs [17, 18], $\text{CNN}_{n_{\text{d-a}}}$, in a reverse fashion to e_{a}. Firstly $\mathbf{z}_{\text{a}}$ is turned to the matrix $\mathbf{Z}_{\text{a}} \in \mathbb{R}^{C_{N_{\text{e-a}}}^{\text{e-out}} \times T'_{N_{\text{e-a}}} \times F'_{N_{\text{e-a}}}}$ and then is processed by $\text{CNN}_{n_{\text{d-a}}}$ as

$$\mathbf{H}_{n_{\text{d-a}}} = \text{ReLU}(\text{BN}_{n_{\text{e-a}}}(\text{CNN}_{n_{\text{d-a}}}(\text{DO}(\mathbf{H}_{n_{\text{d-a}}-1})))), \quad (2)$$

where $\mathbf{H}_0 = \mathbf{Z}_{\text{a}}$ and $\text{BN}_{n_{\text{e-a}}}$ is the batch normalization process used after $\text{CNN}_{n_{\text{d-a}}}$. The reconstructed input, $\hat{\mathbf{X}}_{\text{a}}$, is obtained as $\hat{\mathbf{X}}_{\text{a}} = \sigma(\mathbf{H}_{N_{\text{d-a}}})$, where σ is the sigmoid function.

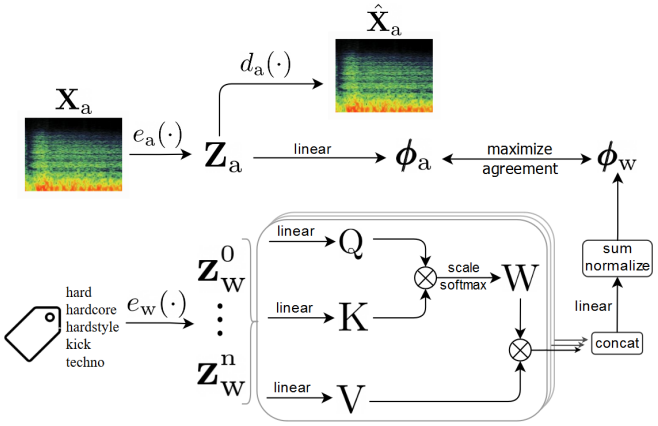

Fig. 1. Illustration of our method. $\boldsymbol{\phi}_{\text{a}}$ and $\boldsymbol{\phi}_{\text{w}}$ are aligned by maximizing their agreement through contrastive learning and, at the same time, $\mathbf{Z}_{\text{a}}$ is used for reconstructing back the original spectrogram input. Word embeddings are passed through a multi-head scaled dot-product self-attention layer in order to build higher-level semantic vectors that are finally aggregated into a single vector $\boldsymbol{\phi}_{\text{w}}$.

2.2. Multi-head, self-attention tags encoding

As our e_{w} we select a pre-optimized word embedding model, using an embedding dimensionality of F_{w}, which outputs $\mathbf{z}_{\text{w}}^{t_{\text{w}}} \in \mathbb{R}^{F_{\text{w}}}$. For processing the output of e_{w} we follow the recent proposal of multi-head, self-attention in the Transformer model, where a scaled dot-product attention mechanism is employed to extract the relevant information of each word in a sentence [19]. Since we use an unordered set of tags, we do not employ the positional embeddings for each tag. The multi-head self-attention attends on the set of encoded tags by the word embeddding model e_{w}, and extracts a contextual embedding of the set of tags. We use this contextual embedding as the latent representation of tags for our cross-modal alignment process with the contrastive loss.

Specifically, we employ three feed-forward neural networks, FNN_q, FNN_k, and FNN_v, FNN_o (without non-linearities) as our linear transformations for the query, key, value and output of the self-attention mechanism, respectively. We follow the original paper of the self-attention [19], we use H attention heads, and we apply the self-attention, Att, on the embeddings of tags, $\mathbf{Z}_{\text{w}}$. The output of Att is $\mathbf{O} \in \mathbb{R}^{H \times T_{\text{w}} \times F_{\text{w}}}$, according to [19]. Then, we concatenate

[1] For the clarity of notation, the superscript n_{B} is dropped here and for the rest of the document, unless it is explicitly needed.

[2] https://github.com/xavierfav/ae-w2v-attention

the result of each H, resulting to $\mathbf{O}' \in \mathbf{R}^{(H \cdot T_\text{w}) \times F_\text{w}}$. Finally, to process the output of each attention head H and obtain the contextual embedding for the input tags, $\boldsymbol{\phi}_\text{w}$, we employ FNN$_o$ and a layer normalization process, LN, and we calculate the contextual embedding, $\boldsymbol{\phi}_\text{w} \in \mathbb{R}^V$, as

$$\mathbf{O}' = \text{FNN}_o(\mathbf{O}) \text{ and} \tag{3}$$

$$\boldsymbol{\phi}_\text{w} = \text{LN}(\sum_{i=1}^{T_\text{w}} \mathbf{O}'_i) \ , \tag{4}$$

where $\mathbf{O}' \in \mathbb{R}^{T_\text{w} \times V}$ and FNN$_o$ has its weights shared along the $H \cdot T_\text{w}$ dimension.

2.3. Cross-modal alignment and optimization

To align the learned latent representations from audio and tags, we employ a constrastive loss and we maximize the agreement of $\boldsymbol{\phi}_\text{a}^{n_\text{B}}$ and $\boldsymbol{\phi}_\text{w}^{n_\text{B}}$ on minibatches $\mathbb{G}_\text{b}$ of size N_b, randomly sampled from our dataset $\mathbb{G}$. To reduce the mismatch between the two different modalities, we employ a feed-forward neural network (FNN$_\text{c-a}$) to process $\mathbf{z}_\text{a}^{n_\text{B}}$, as

$$\boldsymbol{\phi}_\text{a}^{n_\text{B}} = \text{FNN}_\text{c-a}(\mathbf{z}_\text{a}^{n_\text{B}}). \tag{5}$$

Then, we employ a contrastive loss between $\boldsymbol{\phi}_\text{a}^{n_\text{B}}$ and $\boldsymbol{\phi}_\text{w}^{n_\text{B}}$ and we jointly optimize e_a, d_a, FNN$_q$, FNN$_k$, FNN$_v$, and FNN$_o$ by minimizing the loss

$$\mathcal{L}(\mathbb{G}_\text{b}, e_\text{a}, d_\text{a}, \text{Att}) = \lambda_\text{a}\mathcal{L}_\text{a} + \lambda_\xi\mathcal{L}_\xi, \tag{6}$$

where $\mathcal{L}_\text{a}$ is the generalized Kullback-Leibler divergence between $\mathbf{X}_\text{a}$ and $\hat{\mathbf{X}}_\text{a}$, shown to work good for audio reconstruction [20, 21]. $\mathcal{L}_\xi$ is the contrastive loss between paired $\boldsymbol{\phi}_\text{a}^{n_\text{B}}$ and $\boldsymbol{\phi}_\text{w}^{n_\text{B}}$ and other examples in the minibatch, as defined in [10]. We use a temperature parameter τ for $\mathcal{L}_\xi$, and $\lambda_\star$ is a hyper-parameter used for numerical balancing of the two losses.

3. EVALUATION

We evaluate our method by assessing the performance of e_a as a pre-trained audio embedding extractor in different audio classification tasks. We utilize a different audio dataset for each task and we compare the performance of our e_a against COALA [15] and a set of hand-crafted MFCCs feature. We choose COALA because is the SOTA method that deals with cross-modal alignment of audio and text.

3.1. Pre-training dataset and data pre-processing

We use the same dataset as used in COALA, which consists of sounds and associated tags collected from Freesound platform [22]. We compute $F_\text{a} = 96$ log-scaled mel-band energies using sliding windows of 1024 samples (≈46 ms), with 50% overlap and the Hamming windowing function. Then, we select the spectrogram patch of size $T_\text{a} = 96$ that has maximum energy in each sample. This process leads to 189 896 spectrogram patches. 10% of these patches are kept for validation and all the patches are scaled to values between 0 and 1. We process the tags associated to the audio clips by firstly removing any stop-words and making any plural forms of nouns to singular. We remove tags that occur in more than 70% of the sounds as they can be considered less informative, and consider the C=1000 remaining most occurring tags. Then we train a continuous bag-of-words Word2Vec model [8] using these processed tags and we use this model as our e_w.

To have our method comparable with COALA and assess the impact of our proposed approach for learning contextual tags, we follow COALA and we use $N_\text{e-a} = 5$ CNN$_{n_\text{e-a}}$, with $K_\text{e-a} = 4$ and $S_\text{e-a} = 2$. We use $C_1^\text{in} = 1$ and $C_{n_\text{e-a}}^\text{out} = 128$. These hyper-parameters result to $V = 1152$ and e_a has approximately 2.4M parameters. The tag encoder takes a set of $T_\text{w} = 10$ maximum tags. It uses fully-connected linear transformations that retain the same dimension as the word embeddings, producing tag-based embedding vectors of the same dimension. We utilize two different F_w, one of 128 and another of 1152. The first of 128 is due to the fact that we are using a small scale vocabulary and we choose to have a small embedding size for e_w. The second, is for having $F_\text{w} = V$. Additionally, we employ two different set-ups for our Att, one with far and another with one attention head. We indicate the different combinations as "w2v-F_w-Hh", where H is the amount of attention heads. For example, "w2v-128-1h" means that we are using $F_\text{w} = 128$ and one attention head. We also employ the self-attention strategy with a simple mean aggregation of the tags' word embeddings, which we refer as w2v-128-mean and w2v-1152-mean, to assess the impact of Att when using a simpler aggregation of its output. Finally, we employ the 20 first mel-frequency cepstral coefficients (MFCCs) with their Δs and $\Delta\Delta$s as a low anchor, using means and standard deviations through time, and we term this case as MFCCs.

All our models are trained for 200 epochs, using a minibatch size N_B=128 and an SGD optimizer with a learning rate value of 0.005. We utilize the validation set to define the different λ's at Eq. (11) and the constrastive loss temperature parameter τ, to λ_a=5, λ_ξ=10, and $\tau = 0.1$. We use dropout probability of 0.25 for our e_a and d_a and 0.1 after the tag-based embedding model to avoid overfitting while training.

3.2. Audio-based classification

We assess the performance of the different embeddings extracted with e_a in three different audio classification tasks. For each task, we employ the pre-trained, with our method, e_a, and we extract audio embeddings $\mathbf{z}_\text{a}$ for the the audio data of the corresponding task. Then, adopt the corresponding training protocol of the task (e.g. cross-fold validation) and we optimize a multi-layer perceptron (MLP) with one hidden layer of 256 features, similar to what is used in [4, 15]. Finally, we

assess the performance of the classifier using the data and the testing protocol of each task. To obtain an unbiased evaluation of our method, we repeat 10 times the training procedure of the MLP in each task and report the mean accuracy. Additionally, in order to understand if the self-attention mechanism is actually beneficial for obtaining a better semantic embedding from tags, we perform a tag-based classification for comparing the different approach versions. For this purpose, we use the UrbanSound8K dataset and the tags of the associated samples from Freesound.

As the first task, we consider Sound Event Recognition (SER), where we use the UrbanSound8K dataset (US8K) [23]. US8K consists of around 8000 single-labeled sounds of maximum 4 seconds and 10 classes. We use the provided folds for cross-validation. We additonally consider the task of Music Genre Classification (MGC), where we use the fault-filtered version of the GTZAN dataset [24, 25] consisting of music excepts of 30 seconds, single-labeled split in pre-computed sets of 443 songs for training and 290 for testing. Finally, we also consider the Musical Instrument Classification (MIC) task. We use the NSynth dataset [26] which consists of more than 300k sound samples organised in 10 instrument families. However, because we are interested to see how our models performs with relatively low amount of training data, we sample from NSynth a balanced set of 20k samples from the training set which correspond to approximately 7% of the original set. The evaluation set is kept the same.

For the above tasks and datasets, we use non-overlapping audio frames that are calculated similarly to the pre-training dataset. These frames are given as input to the different models in order to obtain audio embeddings. In order to obtain fixed-length vectors, the audio embeddings are aggregated using a mean average statistic and finally used as an input to a classifier that is trained for each corresponding task. Additionally, resulting embedding and MFCCs vectors are standardized to zero-mean and unit-variance, using statistics calculated from the training split of each task.

4. RESULTS

Table 1 shows the performances of the different embeddings our MFCCs baseline and previous results from COALA [15]. The self-attention mechanism used in the tag-based network benefits the classification performance in SER and MGC. This indicates that our proposed method indeed results in learning a contextual embedding that can be effectively used for learning better general audio representations. For MIC however, we do not observe any benefit from it. A reason may be that the classification of a musical instrument does not rely much on the context of employed textual descriptions, at least for the musical instruments contained in the employed dataset for MIC. This means that that classifying instrument samples can be done by solely using representations learned by the audio autoencoder, without any semantic information. In [15], it was observed that the reconstruction objective was bringing important improvements in this case, and probably, the enrichment of semantics achieved with the alignment with the tag-based latent representation losses its benefits. Using more attention heads is able to bring better performance in SER and MGC. This suggests that the pre-trained word representation we use can benefit from more powerful attention mechanism. The impact of the embedding size of the word embeddings is not clear from our experiment. But, it shows that using different dimensions for the audio autoencoder and the tag-based encoder does not necessary hinders the contrastive alignment.

Table 1. Average mean accuracy for SER, MGC, and MIC. Additionally performances on US8K dataset using a tag-based classifier are reported in the last column.

	SER	**MGC**	**MIC**	**US8K-tag**
MFCCs	65.8	49.8	62.6	-
w2v-128-1h	71.5	61.3	68.9	79.2
w2v-1152-1h	72.1	61.5	68.6	**80.3**
w2v-128-4h	**73.5**	59.6	69.7	79.4
w2v-1152-4h	70.5	**63.4**	69.9	78.7
w2v-128-mean	71.3	60.4	70.0	79.7
w2v-1152-mean	71.1	60.7	68.4	78.5
COALA [15]	72.7	60.7	**73.1**	-

When using tags for performing the classification on US8K for SER, there is no benefit of using multiple attention heads, and the self-attention mechanism is only slightly improving the performance compared to the mean aggregation strategy. Moreover, the results show that audio-based classification is still not performing as well as the tag-based one, which could suggest that more powerful audio encoders could be better aligned with the semantics of the content, and produce better results.

5. CONCLUSION

In this work we present a method for cross-modal alignment of audio and tags. The proposed approach uses a pre-trained word embedding and learns contextual tag embeddings that are aligned with audio embeddings using contrastive learning. From audio samples and associated tags, our method is able to learn semantically enriched audio representation that can be used in different classification tasks. The embedding model produced is evaluated in three different downstream tasks, including sound event recognition, music genre and music instrument classifications. Over a previous similar method [15], the proposed approach relies on pre-trained word embeddings that grants the the advantage of being directly able to be used in a wider range of applications, such as cross-modal retrieval or zero-shot learning. Future work will focus on improving the model and evaluating in different scenarios, such as the one just mentioned.

6. REFERENCES

[1] J. F. Gemmeke et al., "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[2] J. Lee, J. Park, K. L. Kim, and J. Nam, "SampleCNN: End-to-end deep convolutional neural networks using very small filters for music classification," *Applied Sciences*, vol. 8, no. 1, 2018.

[3] J. Pons and X. Serra, "Musicnn: Pre-trained convolutional neural networks for music audio tagging," *arXiv preprint arXiv:1909.06654*, 2019.

[4] J. Cramer, H.-H. Wu, J. Salamon, and J. P. Bello, "Look, listen, and learn more: Design choices for deep audio embeddings," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 3852–3856.

[5] A. Jansen et al., "Unsupervised learning of semantic audio representations," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2018.

[6] H. Alwassel, D. Mahajan, L. Torresani, B. Ghanem, and D. Tran, "Self-supervised learning by cross-modal audio-video clustering," *arXiv preprint arXiv:1911.12667*, 2019.

[7] N. Turpault, R. Serizel, and E. Vincent, "Semi-supervised triplet loss based learning of ambient audio embeddings," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[8] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[10] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," *arXiv preprint arXiv:2002.05709*, 2020.

[11] P. H Le-Khac, G. Healy, and A. F Smeaton, "Contrastive representation learning: A framework and review," *arXiv preprint arXiv:2010.05113*, 2020.

[12] Y. Wu, S. Wang, G. Song, and Q. Huang, "Learning fragment self-attention embeddings for image-text matching," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019.

[13] D. Jiang et al., "Improving transformer-based speech recognition using unsupervised pre-training," *arXiv preprint arXiv:1910.09932*, 2019.

[14] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[15] X. Favory, K. Drossos, T. Virtanen, and X. Serra, "Coala: Co-aligned autoencoders for learning semantically enriched audio representations," in *International Conference on Machine Learning (ICML), Workshop on Self-supervised learning in Audio and Speech*, 2020.

[16] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016.

[17] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *International Conference on Learning Representations (ICLR)*, 2016.

[18] V. Dumoulin and F. Visin, "A guide to convolution arithmetic for deep learning," 2016.

[19] A. Vaswani et al., "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[20] K. Drossos et al., "Mad twinnet: Masker-denoiser architecture with twin networks for monaural sound source separation," in *2018 International Joint Conference on Neural Networks (IJCNN)*, 2018.

[21] S. I. Mimilakis et al., "Monaural singing voice separation with skip-filtering connections and recurrent inference of time-frequency mask," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[22] F. Font, G. Roma, and X. Serra, "Freesound technical demo," in *Proceedings of the 21st ACM international conference on Multimedia*, 2013.

[23] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 1041–1044.

[24] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on speech and audio processing*, vol. 10, no. 5, pp. 293–302, 2002.

[25] C. Kereliuk, B. L Sturm, and J. Larsen, "Deep learning and music adversaries," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 2059–2071, 2015.

[26] J. Engel et al., "Neural audio synthesis of musical notes with wavenet autoencoders," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 2017.