CROSS MODAL AUDIO SEARCH AND RETRIEVAL WITH JOINT EMBEDDINGS BASED ON TEXT AND AUDIO

Benjamin Elizalde[†⋆], *Shuayb Zarar*[†], *Bhiksha Raj*[⋆]

[†]Microsoft Research, [⋆]Carnegie Mellon University
Email: bmartin1@cmu.edu, shuayb@microsoft.com, bhiksha@cs.cmu.edu

ABSTRACT

Existing audio search engines use one of two approaches: matching text-text or audio-audio pairs. In the former, text queries are matched to semantically similar words in an index of audio metadata to retrieve corresponding audio clips or segments, while in the latter, audio signals are directly used to retrieve acoustically-similar recordings from an audio database. However, independent treatment of text and audio has precluded information exchange between the two modalities. This is a problem because similarity in language does not always imply similarity in acoustics, and vice versa. Moreover, independent modeling can be error prone especially for ad hoc, user-generated recordings, which are noisy in both audio and their associated textual labels. To overcome this limitation, we propose a framework that learns joint embeddings from a shared lexico-acoustic space, where vectors from either modality can be mapped together and compared directly. Thus, we improve semantic knowledge and enable the use of either text or audio queries to search and retrieve audio. Our results break new ground for a cross-modal audio search engine, and further exploration of lexico-acoustic spaces.

Index Terms— Joint Audio-Text Embedding, Cross Modal Retrieval, Audio Search Engine, Content-Based Audio Retrieval, Query by Example, Siamese Neural Network

1. INTRODUCTION

User-generated audio is shared on the web every day. Examples of these include recordings from digital personal assistants, security cameras, game streams, podcasts and social media. For emerging applications such as ambient sensing, video-content analysis and personalized multimedia services, it is increasingly important to search these recordings and retrieve audio segments or textual descriptions that describe acoustic events such as vehicles, nature, animal and human sounds. Typically, search engines that are used for this purpose utilize text queries to find semantically similar words in an index of audio metadata, and retrieve the corresponding audio [1, 2]. An alternative approach is content-based retrieval, where an audio clip is directly used as a query to match against acoustically similar recordings and retrieve audio [3, 4]. However, neither approach enables direct comparison between text-audio or audio-text pairs, nor learn to map together similarities of lexical semantics and acoustics.

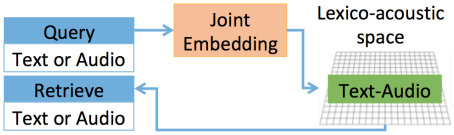

Fig. 1. Proposed framework enables cross-modal search and direct comparison of audio and text modalities. Shared latent space fuses lexical semantics with acoustic similarity.

Cross-modal search and retrieval is an approach that enables matching of text-audio or audio-text pairs, which can compensate for missing or ambiguous information in either modality [5]. This approach is useful because user-generated multimedia content is inherently noisy in both textual and auditory content. The labels, tags and descriptions are incomplete, subjective, lacking or wrong. And the inter- and intra- class acoustics are increasingly diverse. For instance, the sounds of *popcorn popping* and a *sink filling* are similar to the sounds of *fireworks* and a *bathtub filling*, respectively. Thus, utilizing information from one modality to improve the retrieval performance of the other is promising. Authors in [6] were one of the first to tackle this problem at scale. They pair audio with an associated text label and utilize it as a query to retrieve audio based on a combined scoring function. A limitation of this work is that the labels used in the query must match exactly those in the index. An earlier work had proposed a more extensible cross-modal approach that could name new sounds with existing labels or associate labels with existing sounds [7]. However, the hierarchical language model used in the paper limited semantic relations and, according to the authors, scalability. Neither of these methods find lexico-acoustic spaces in a data-driven way, which we aim to do in this paper. An overview of our approach is illustrated in Fig. 1.

From a technical perspective, learning to map text and audio together can help combine lexical relationships and acoustic similarities, which is interesting because similarly-annotated sounds are not always close in both lexical and acoustic semantics. For instance, textual labels *violin plays* and *violin crashes* both refer to violins, are close in lexical semantics, yet their acoustics are not. This is an intrinsic problem in audio and text, exacerbated due to the lack of proper lexicalization of acoustic phenomena [8]. Furthermore, au-

dio labels are highly subjective; a sound may be described differently depending on the listener. Authors in [7] create a connection between acoustic-similarity and lexico-semantic spaces in a many-to-many setup similar to our objective, but the relations are forced to be hierarchical. This restricts the ability to generalize or scale. Due to limited related work in mapping audio and text together, we looked into the computer vision domain. Similar problems are observed with images and text, and indeed there has been work to learn joint embeddings with these modalities [9, 10]. Drawing inspiration from these works, we propose to learn joint embeddings of audio and text; details of which are provided next.

2. METHODOLOGY

In this section, we present the modeling framework and describe the Siamese neural network (SNN) architecture.

2.1. Cross-Modal Search Framework

We propose a cross-modal search methodology, where a text or audio query can be used to retrieve audio or text. This is enabled by a shared latent space that combines lexical semantics with acoustic similarity, thus affecting the retrieved results. The shared latent space is learned in a data-driven way *via* a SNN [11]. Such networks have been used in the literature for audio-audio retrieval of videos [3], audio-audio human-fall detection [12], audio-image retrieval of videos [13].

2.2. SNN Embeddings and Loss Function

The SNN, shown in Fig. 2, consists of twin networks that have the same base architecture with shared weights. The weights are learned simultaneously at every parameter update. Each base network utilizes an input vector of any one modality – audio or text. For training, the SNN does not need any explicit class labels for each modality, but rather, labels are inferred per pair to be 1 or 0 depending on whether the inputs are from the same class or not. While the SNN is trained with pairs of audio and text examples, a similarity metric is computed for each pair. This metric is utilized by a loss function to enforce constraints that cause similar pairs to come together and different pairs to go apart. During the embedding stage, the trained base networks can utilize either audio or text vectors as input to produce a joint embedding vector.

The architecture of the base network is a feed-forward multi-layer perceptron network. It consists of 5 layers: the input layer of dimensionality 300, which takes either audio or text feature vectors, 3 dense layers of dimensionality 1024, 512 and 512, respectively, and the output layer of dimensionality 1024, which is the dimensionality of the joint embedding. The dense layers utilize a dropout rate of 0.5 and the ReLU activation function; $\max(0, x)$, where x is input to the function. We trained the SNN for 100 epochs using the Adam algorithm. We also tuned the hyper-parameters of the SNN to achieve good performance with the input features that are described in the next section.

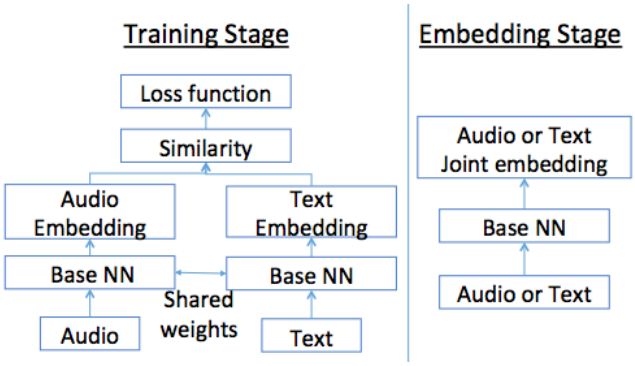

Fig. 2. Architecture of the SNN that is used to learn the shared space and compute joint embeddings of audio and text.

In the literature, contrastive loss ($\mathcal{L}_{CL}$) with Euclidean distance (ED) has been a popular loss function for training SNNs [11, 3]. However, this did not work in our audio-text setup. The idea behind this loss is that dissimilar points contribute to the training loss only if the similarity between them is within a margin. After inspecting the embedding from each twin network, we noticed that their values were very sparse and with similar non-zero values, which caused the computed distance to be close to zero even for negative pairs. We, unsuccessfully tried tuning the margin value, modifying the architecture and adding more epochs. Hence, we proposed the binary cross entropy loss $\mathcal{L}_{BCE}$ with an exponential negative Euclidean distance (d_w). The proposed distance, instead of having unbounded similarity values, forces the values to lie between 0 and 1; a higher value implies more similarity. Therefore, the $\mathcal{L}_{BCE}$ with similarity metric d_w is defined for an audio-text pair (a_i, t_i) as follows:

$$\mathcal{L}_{BCE} = -\frac{1}{N}\sum_{i}^{N} y_i log(d_w) - (1-y_i) log(1-d_w)$$

$$d_w = \exp\left(-\sqrt{\sum_{i}^{N}(a_i - t_i)^2}\right),$$

where y_i represents output samples produced by the SNN. We tried different combinations of $\mathcal{L}_{CL}$ and $\mathcal{L}_{BCE}$ with other similarity metrics such as distances $l1$ and $l2$, and their exponential variants, $\exp(-l1)$ and $\exp(-l2)$, respectively. However, as shown in Table 1, for our test data set described ahead in Section 3.1, $\exp(-l2)$ yielded the best retrieval performance; measured by the mean-average precision at 3 docu-

Table 1. The combination of $\mathcal{L}_{CL}$ and $l2$ did not work well for our setup. Therefore we employed $\mathcal{L}_{BCE}$ and d_w. MAP@3 scores for retrieving audio with joint embeddings.

	Similarity (%)			
Loss	$L1$	$L2$	$\exp(-L1)$	$\exp(-L2)$
$\mathcal{L}_{CL}$	32.0	15.4	12.9	14.6
$\mathcal{L}_{BCE}$	40.3	44.7	34.0	**61.2**

ments (MAP@3)]. Thus, we picked it for future experiments. We also tried different sizes of the input training data, from 100 to 5,000 audio-text pairs and found that 500 yielded the best trade-off between performance and processing speed to train the SNN in all our experiments.

3. EXPERIMENTS AND RESULTS

In this section, we evaluate the retrieval performance of the cross-modal search framework. We also present some qualitative results from searching the lexico-acoustic space.

3.1. Dataset

For evaluation of the proposed approach, we employed dataset from task-2 of the 2018 DCASE challenge [14] because it comprised both clean recordings and labels that could be used for retrieval. The dataset included audio clips from the Freesound library, which were annotated using a vocabulary based on the Google AudioSet Ontology. The training and test sets included $\sim 9.5k$ and $\sim 1.6k$ recordings, respectively, which were unequally distributed among 41 classes. The number of recordings per class ranged between 94-300. The duration of the recordings ranged between 0.3-30 seconds. The text labels comprised one or two words that described the class. 20% of the training data was used for validation. All audio recordings in this dataset were available as uncompressed PCM 16 bit, 44.1 kHz mono audio files.

3.2. Features

We used text features that exhibit linear substructure and similarities in a lexico-semantic space. Global vectors for word representation (GloVe) is one such embedding [15]. We employed GloVe features produced from a model that was separately trained on Wikipedia and the Gigaword corpus. Thus, we transformed text labels corresponding to each audio clip in our dataset into a word embedding of 300 dimensions. If the labels comprised more than one word, we computed the average GloVe vector. Feature vectors were eventually normalized to have unit magnitude for consistency with similarity metrics in the SNN.

We used two types of audio features: standard Mel-frequency cepstral coefficients (MFCCs) and state-of-the-art Walnet features [16, 17]. MFCCs were computed based on a sliding window of 25 ms and hop sizes of 10 ms. They included delta and double delta values, resulting in a feature-dimensionality of 300 per window. We also tried computing a set of acoustic features of up to 6,500 features [18], but they achieved lower performance. We averaged the MFCCs across all windows to produce one feature vector per audio clip. For the Walnet features, we computed a 128-dimensional logmel-spectrogram vector and transformed it *via* a convolutional neural network (CNN) that was trained separately on the AudioSet data. The network comprised 8 convolutional layers, resulting in an output feature vector of dimensionality 527. We used the intermediate outputs from the 8^{th} layer of the CNN, which had a dimensionality of 1024. Through principal component analysis (PCA), we reduced the dimensionality of the resulting feature vector to 300, which matched the text-based features [19]. Features were L2 normalized.

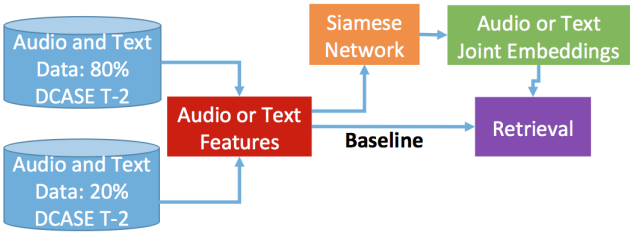

Fig. 3. Evaluation system used for cross-modal search.

3.3. Classifiers

Our evaluation system is summarized in Fig. 3. For retrieval experiments on the test data, audio and text features were processed in two ways: (1) direct retrieval, which formed the baseline, and (2) embeddings with the SNN followed by retrieval. To study the retrieval performance, we employed three classifiers: support-vector machine, multilayer perceptron and k-nearest neighbor classifier (kNN). The performance was similar across all three classifiers. Thus, we picked kNN (k=25) because it exhibited an intrinsic sense of neighborhood and had less parameters to tune.

3.4. Retrieval Performance

We first demonstrate that the baseline is insufficient for cross-modal retrieval. Recall that in this case, we train the classifiers with audio features (MFCCs, Walnet) and test them with text features (GloVe), and vice versa. Retrieval performance is measured in terms of MAP@3, standard in DCASE Task 2 [14], which is the mean of avg. precision scores for each query, utilizing the first 3 retrieved results [20]. As shown in Table 2, homologous training and test features yielded good performance: 56 and 72% for audio MFCCs and Walnet, respectively, and 100% for text GloVe features (100% is expected because text labels corresponding to different audio

Table 2. Cross modal search is possible with joint embeddings, which outperform the baseline features.

	Train	
Test Baseline	Audio (MFCC) Features	Text Features
Audio (MFCC) Features	56.0%	2.4%
Text Features	2.4%	100%
Test Baseline	Audio (Walnet) Features	Text Features
Audio (Walnet) Features	72.0%	2.4%
Text Features	2.4%	100%
Test JE	Audio (MFCC) JE	Text JE
Audio (MFCC) JE	61.2%	54.7%
Text JE	100%	100%
Test JE	Audio (Walnet) JE	Text JE
Audio (Walnet) JE	74.9%	71.3%
Text JE	100%	100%

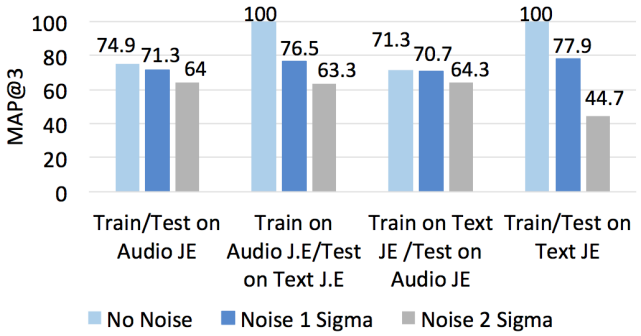

Fig. 4. Proposed SNN joint embeddings are robust to additive random Gaussian noise in the text features.

files were the same in training and test data). However, cross-modal search with opposite feature types during training and test yielded random performance (2.4% MAP@3).

Next, we show how SNN embeddings enable cross-modal retrieval. We processed MFCCs or Walnet, and GloVe features together to produce joint embeddings (JEs) and normalized them to have unit norm. We then employed the kNN classifier to study retrieval performance. Based on the results shown in Table 2, we make the following conclusions. First, training and testing on audio JEs yielded 61.2 or 74.9%, which outperformed the baseline. Second, training and testing on text JEs yielded 100%, which is consistent with the baseline. Third, cross-modal search results are not random; MAP@3 of 54.7 or 71.3% for training on text JEs and testing on audio JEs, and 100% for training on audio JEs and testing on text JEs. Fourth, better audio features (Walnet) improved performance. Results could be further boosted with a better tuning of the Walnet network, but this was out of the scope of this work. Overall performance was better than the CNN baseline provided by DCASE (70%).

Noise injection. To address the issue of text labels corresponding to different audio files being the same in training and test data, we altered the text labels. There are several ways to achieve this goal. We chose to add random Gaussian noise with one or two standard deviations to each of the *text features*. We re-trained the SNN on audio (Walnet) and noisy text features. The retrieval results with this new embedding are shown in Fig. 4. As expected, the performance degraded with more noise. Testing with text JEs resulted in a larger drop of 36-66% MAP@3 because noise in the input features affected both the training process of the SNN and test JEs. However, since audio JEs were affected only while training the SNN, the performance drop was lesser in the cases where these were used at test time. As part of future work, we hope to combine audio metadata together with the labels before extracting GloVe text features, which will represent a more realistic search set up and approach to adding noise.

3.5. Qualitative Results

We inspected the cross-modal retrieval results with some test examples. We also studied the retrieval performance when using out-of-vocabulary (OOV) labels and audio recordings. In this section, we present some qualitative results for the same.

As a first example, we considered the class label *gunshot*, computed its GloVe features, compared them against GloVe features of the other 40 classes using cosine similarity, and retrieved the following top four labels: *gunshot, tearing, applause, cough*. Next, we extracted JEs of the same label, compared the lexico-acoustic similarity against k nearest neighbors, and retrieved audio corresponding to the following four labels: *fireworks, gunshot, microwave oven, knock*. Another example was *meow*, which retrieved with text similarity: *meow, fart, cough* and with lexico-acoustic similarity: *meow, bark, trumpet*. From these results, the main conclusion is that although both approaches predicted the correct class, they retrieved different results, which suggests that the shared space includes knowledge that combines both modalities.

Our framework allows the use of OOV class labels for querying. For example, we considered the query *house*, which is not part of the training data and which has a more abstract meaning than any label in the training data. With text and lexico-acoustic similarity, the top five retrieved items were: *drawer, telephone, writing, gunshot, double bass* and *meow, cough, finger snapping, laughter, computer keyboard*, respectively. It is interesting to note that both results are arguably relevant, but are not the same.

In a reverse cross-modal search setup, our framework also permits the use of OOV audio clips for querying. For instance, we considered the audio query corresponding to *thunderstorm* downloaded from Freesounds [1]. When compared for acoustic similarity against audio features (Walnet) of the training set, the following four labels were retrieved: *fireworks, applause, tearing, fart*. However, with JEs and lexico-acoustic similarity, the following results were retrieved: *fireworks, cough, drawer open or close, gunshot*. Another example of audio query was *orchestra*, which retrieved *applause, cello, acoustic guitar, flute, fireworks, violin, clarinet* and *violin, trumpet, saxophone, flute, double bass, clarinet, cello* with acoustic and lexico-acoustic similarities, respectively. It is important to note that varying the value of k could sometimes affect the number of retrieved items and their order. Future work will involve ranking the results from the JEs .

4. CONCLUSIONS

We proposed a cross-modal search framework that enables us to retrieve audio recordings using either text or audio queries. It was built on a shared lexico-acoustic space that was learned in a completely data-driven way. The shared space thus enabled us to map together and directly compare state-of-the-art text and audio features for search and retrieval. Our results showed robustness against noise in the text labels. We also demonstrated good retrieval performance with out-of-vocabulary text or audio queries that were not found in the training data.

[1] https://freesound.org

5. REFERENCES

[1] Paul Over, Jon Fiscus, Greg Sanders, David Joy, Martial Michel, George Awad, Alan Smeaton, Wessel Kraaij, and Georges Quénot, "Trecvid 2014–an overview of the goals, tasks, data, evaluation mechanisms and metrics," in *Proceedings of TRECVID*, 2014, p. 52.

[2] Benjamin Elizalde, Rohan Badlani, Ankit Shah, Anurag Kumar, and Bhiksha Raj, "Nels-never-ending learner of sounds," *NIPS Workshop on Machine Learning for Audio*, 2017.

[3] Pranay Manocha, Rohan Badlani, Anurag Kumar, Ankit Shah, Benjamin Elizalde, and Bhiksha Raj, "Content-based representations of audio using siamese neural networks," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 3136–3140.

[4] Gordon Wichern, Jiachen Xue, Harvey Thornburg, Brandon Mechtley, and Andreas Spanias, "Segmentation, indexing, and retrieval for environmental and natural sounds," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 688–707, 2010.

[5] Yuxin Peng, Xin Huang, and Yunzhen Zhao, "An overview of cross-media retrieval: Concepts, methodologies, benchmarks and challenges," *IEEE Transactions on Circuits and Systems for Video Technology*, 2017.

[6] Gal Chechik, Eugene Ie, Martin Rehn, Samy Bengio, and Dick Lyon, "Large-scale content-based audio retrieval from text queries," in *Proceedings of the 1st ACM international conference on Multimedia information retrieval*. ACM, 2008, pp. 105–112.

[7] Malcolm Slaney and C San Jose, "Semantic-audio retrieval," in *ICASSP*, 2002, vol. 4.

[8] Gregoire Lafay, Mathieu Lagrange, Mathias Rossignol, Emmanouil Benetos, and Axel Roebel, "A morphological model for simulating acoustic scenes and its application to sound event detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

[9] Yusuf Aytar, Carl Vondrick, and Antonio Torralba, "See, hear, and read: Deep aligned representations," *arXiv preprint arXiv:1706.00932*, 2017.

[10] Liwei Wang, Yin Li, and Svetlana Lazebnik, "Learning deep structure-preserving image-text embeddings," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5005–5013.

[11] Raia Hadsell, Sumit Chopra, and Yann LeCun, "Dimensionality reduction by learning an invariant mapping," in *null*. IEEE, 2006, pp. 1735–1742.

[12] Diego Droghini, Fabio Vesperini, Emanuele Principi, Stefano Squartini, and Francesco Piazza, "Few-shot siamese neural networks employing audio features for human-fall detection," in *Proceedings of the International Conference on Pattern Recognition and Artificial Intelligence*. ACM, 2018, pp. 63–69.

[13] Jonathan Masci, Michael M Bronstein, Alexander M Bronstein, and Jürgen Schmidhuber, "Multimodal similarity-preserving hashing," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 4, pp. 824–830, 2014.

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

[15] Jeffrey Pennington, Richard Socher, and Christopher D. Manning, "Glove: Global vectors for word representation," in *Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543.

[16] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen, "DCASE 2017 challenge setup: Tasks, datasets and baseline system," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop*, 2017.

[17] Ankit Shah, Anurag Kumar, Alexander G Hauptmann, and Bhiksha Raj, "A closer look at weak label learning for audio events," *arXiv preprint arXiv:1804.09288*, 2018.

[18] Florian Eyben, Martin Wöllmer, and Björn Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 2010, pp. 1459–1462.

[19] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, and others., "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[20] Filip Radlinski and Nick Craswell, "Comparing the sensitivity of information retrieval metrics," in *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2010, pp. 667–674.