Multi-Modal Zero-Shot Learning for Music

Harshita Seth
Student Number: 190648295
Supervisor: George Fazekas, Elio Quinton
Msc Sound and Music Computing
Queen Mary, University of London
ec21981@qmul.ac.uk

***Abstract*—In traditional supervised classification paradigm, a model learns to output predictions over a set of predefined target classes. Models trained by such a method are unable to predict unseen classes, limiting their use in real-world scenarios. Zero-shot learning aims to overcome this limitation by enabling a model to extend to unseen classes through side information. Recent developments in multi-modal learning have demonstrated that natural language supervision can be leveraged to achieve good Zero-shot performance when transferring a pre-trained multi-modal model to unseen tasks. In this work, we explore a Zero-shot methodology to learn the audio music representation by leveraging the natural language supervision to further solve the music information retrieval downstream task. We are using noisy text descriptions which represent the overall music content of the input audio file and these descriptions are weakly annotated as they are not aligned with the audio at the token level. We are proposing a cross-modal generative framework to learn the robust and generic audio representation. We have demonstrates the performance of our model on different MIR downstream task across different datasets.**

***Index Terms*—Zero-shot, Multi-modal, Music Information Retrieval**

I. Introduction

Audio-based music tagging is a task to identify the different genres, moods, themes of the music, and other song qualities. Automatic music tag (Won et al. 2020) classification is one of the problems in the domain of music information retrieval tasks. Most of the current work is around training the supervised network which classifies the limited labels and tags(Juhan Nam & Yang 2019). In the real world, music tag queries can have a huge number of tags so the music information retrieval system needs to be more flexible and not restricted to a limited number of classes. Zero-shot learning is a methodology to overcome this limitation and enable the network to give results on new unseen labels. Zero-shot learning utilized the side information which includes music genres, instrument annotation, and mood/theme in a semi/supervised way. A model trained by such learning method will be able to predict new genres on an input audio file.

In (Choi & Nam 2019*b*, Sandouk & Chen 2016), author used pairs of music track and corresponding tags to learn representation to develop Zero-shot methodology. Another way to perform supervised learning is to learn the representation from a noisy text description of music audio file (Manco & Fazekas 2022). In this work, we are proposing an approach to investigate whether Zero-shot transfer learning with noisy text supervision can be leveraged in audio-and-language models for music representation learning. Our approach does not require the text annotation of the audio file at the token level, we are utilizing the text data which contains the overall description of the audio file and doesn't have any alignment with the audio file. We are proposing a Cross-modal generative network architecture to learn the audio music representation. There has been some work in this domain around learning audio representation through metric learning (X. Favory & Serra 2020, 2021, A. Ferraro & Bogdanov 2021). Our work is different from previous work in terms of using text-audio pairs. Previous work utilized tags as their text modality while we are using long captions as our text data.

II. Related Work

A. Zero-shot Learning

In Zero-shot learning(Choi & Nam 2019*b*, Sandouk & Chen 2016), the model learns to classify instances of a new unseen class with only training examples of seen classes. Two types of zero-shot learning that are very close to our task are compositional zero-shot learning and multi-label Zero-shot learning. Compositional Zero-shot (Ishan Misra & Hebert 2017) methods aim to recognize the unseen composition from known data using attributes and objects. Ruis et al. (Ruis & Bucur 2021) used Compositional Zero-shot to learn the visual representation to solve the Zero-shot image classification task, (Ruis & Bucur 2021) has built the prototypical representation of the object, and pass it to the graph network to learn the compositional prototype of a novel attribute-object combination. Another type of Zero-shot learning is multi-label Zero-shot where class has more than one label corresponding to each instance. The most important difference between multi-label Zero-shot and single-label Zero-shot is data distribution for training and testing. In single-label Zero-shot learning, data is split into seen and unseen labels whereas in multi-label Zero-shot learning it's not straightforward, (Zhou Ren & Yuille. 2017, Wang & Chen 2017, YongqinXian 2017) has proposed a few ways to split data for multi-label Zero-shot learning.

B. Multi-modal Learning

Multi-modal learning involves learning feature representation using the data from more than one modality (audio-text,

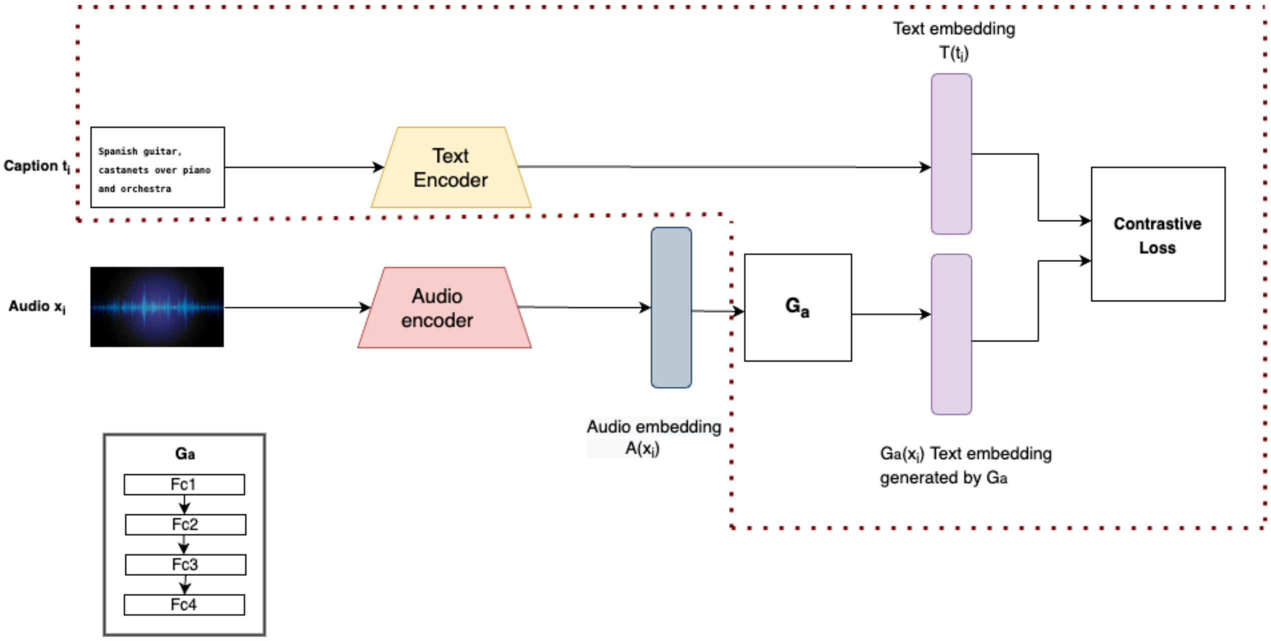


Fig. 1: Architecture of our proposed framework of Cross-modal Generative Network. The framework consist of audio encoder, text encoder and a Generator G_a. Part of framework inside the dotted line represents the networks whose weights gets updated during training

audio-image, audio-text-image etc.). R. Kiros et al.(R. Kiros & Zemel 2014) has propose encoder-decoder pipeline to align visual and semantic information in a joint embedding space. Faghri et al.(F. Faghri & Fidler. 2017) has improved it by adding the triplet loss. Some other example of work (X. Favory & Serra 2020, 2021, A. Ferraro & Bogdanov 2021) which learns audio representation through contrastive approach. (Won & Serra 2021) have used triplet loss for multi-modal metric learning for music audio retrieval task. Data generation is another field where multimodal learning has been used to generate the data of different modality. S. Reed et al.(S. Reed & Lee 2016) utilize GAN for image synthesis based on textual information. Frederik Pahde et al.(Frederik Pahde & Nabi 2021) have used a generative prototypical network and generate new image embedding using the text data to solve the problem of image classification in low data scenario.

III. Multimodal Generative Network

The *Fig* 1 represents the overall proposed framework. The framework consists of a pre-trained audio encoder and text encoder, and a Generator. The audio and text encoder converts the input audio/text data into embeddings. The Generator with the help of a loss function optimized to transform audio embedding extracted from an audio encoder into the text embedding space created by the text encoder.

A. Audio Encoder and Text Encoder

We are using Harmonic CNN (Won & Serrc 2020) as our audio encoder. It exploits the inherent harmonic structure of the audio signal and preserves the spectro-temporal locality. The front-end outputs the harmonic tensors and the back-end process it depending on the task. In our experiments, we are not training the audio encoder, only extracting embedding.

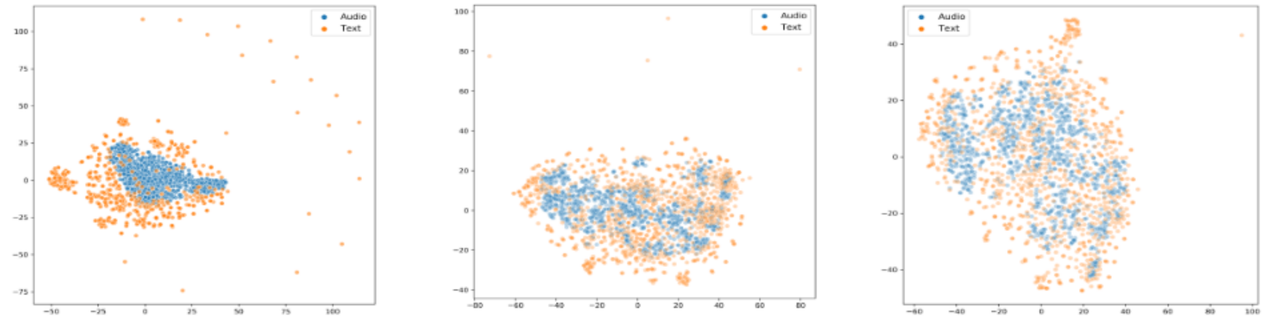
We are using Bidirectional Encoder Representations from Transformers (BERT) (J. Devlin & Toutanova 2019) as text encoder to extract features of text data. We have initialized the BERT with pre-trained weights. The text branch is identical to the standard design of BERT, first, we tokenized input text data and passed it through multi-head attention layers.

B. Cross-modal Feature Generation

Our core idea is to generate new text features using audio data which is provided as training data. The Generator part of the network generates new text embedding by converting the input audio embedding into text embedding space. Since we are training the network to bring both embedding spaces together so instead of generating full textual description, we are only generating text embedding. This results in low computational cost compared to the generation of the full text. The generator network is represented by G_a in *Fig* 1. G_a consists of 4 fully connected layers where the first layer input is a 256-dimensional vector which is the length of the audio embedding from the audio encoder. The G_a converts the 256-dimensional audio embedding into 768-dimensional embedding in text space.

C. Losses

1) Contrastive Loss: We are using a metric learning approach by utilizing contrastive loss. We have written our loss function using the approach mentioned in (Spijkervet & Burgoyne 2021, Saeed & Zeghidour 2021). For a given text



Epoch: 0

Epoch: 10

Epoch: 45

Fig. 2: T-sne plots of embeddings of audio-text pairs of validation data across different epoch of model trained without finetuning text encoder

embedding, we take the softmax of the dot products across all audio embedding and then take cross-entropy loss. Similarly, for a given audio embedding, we do it across all the text embedding. Our final loss is an average of two losses.

Given a set of positive pairs, $G_a(xi)$ and $T(ti)$, the contrastive task to minimize the distance between the positive pairs and increase the distance between the negative pairs, $G_a(x_i)$ and $T(t_k)$ where $k \neq i$. We do not select the negative pairs explicitly, instead for a given N positive pairs, the other (N-1) pairs are negative pairs per positive pair. We have used cosine similarity as our distance metric. For a given embedding u and v, similarity metric $sim(u, v)$ denotes by:

$$sim(u, v) = u^T v / |u||v|$$

The final loss is computed for all pairs of (t_i, x_i) and (x_i, t_i) in a batch represents by $L_(x_i, t_i)$.

$$L_(x_i, t_i) = -\log(\frac{exp(sim(t_i, x_i)}{\sum_{k=1,k!=i}^{N} exp(sim(t_i, x_k))})$$

2) Discriminator Loss: Inspired by GANs (S. Reed & Lee 2016, Frederik Pahde & Nabi 2021) ability to learn better representation in the multimodality domain, we are proposing another loss function, discriminator loss. To introduce loss, we have a separate network called Discriminator network D. D along with generator G formed a standard GAN architecture which is trained on reconstruction loss. Generally, the generator network is initialized with random noise and it is trained to generate the target data with Discriminator loss. In our case, instead of taking initial embedding from random noise space, we are taking it from the pre-trained audio encoder. Detailed experiments with this loss function have shown in Appendix.

D. Training and Evaluation Protocols

We have trained our model on 115k audio-caption pairs from a private production music library. We have randomly split our dataset into train, validation, and test with a 75/15/10 ratio. The caption contains information covering different categories such as genre, instruments, mood, theme, vocals, and time(like 60s or 70s). The caption contains an overall description of the audio and does not have any strong alignment with audio segments. The caption can be considered a weak annotation as there is no strong alignment between the audio and caption at the token level. Our Zero-shot model consists of audio and text branch. The audio branch consists of the audio encoder and a generator G_a whereas the text branch consists of a text encoder. The model is trained to convert the audio embedding space into text embedding space. The two branches are jointly trained using contrastive loss. Due to memory constraints, we have taken the first 30s as input of full audio file for training.

During the Evaluation of test datasets, we generated a 768-dimensional text embedding corresponding to each tag/label using our text branch. Another 768-dimensional audio embedding is extracted with the audio branch. The audio embeddings are extracted from the 30s audio segment taken from the center of the full audio track. We have used the cosine similarity distance metric to calculate the similarity score between the audio embedding and text embedding. All the evaluated metrics are reported using the similarity score.

IV. Experiments

In our baseline network, we have trained only G_a with contrastive loss without finetuning the text encoder. We have fixed text embedding space created by a pre-trained text encoder and trained the G_a network to convert the audio embedding space created by the audio encoder into the fixed text embedding. Fig 2 represents the t-sne plot of audio-text embedding pairs of the validation dataset across different epochs. As the training proceeds, the audio text embedding pairs get scattered. We have further evaluated the model on the GTZAN test dataset. *Fig* 3 shows the t-sne plots of audio embedding with and without training G_a. Note that, *Fig* 3 (b) represents the combine output of network, first audio embedding extracted from pre-trained audio encoder and then pass to trained G_a.

In the second set of experiments, we have trained G_a

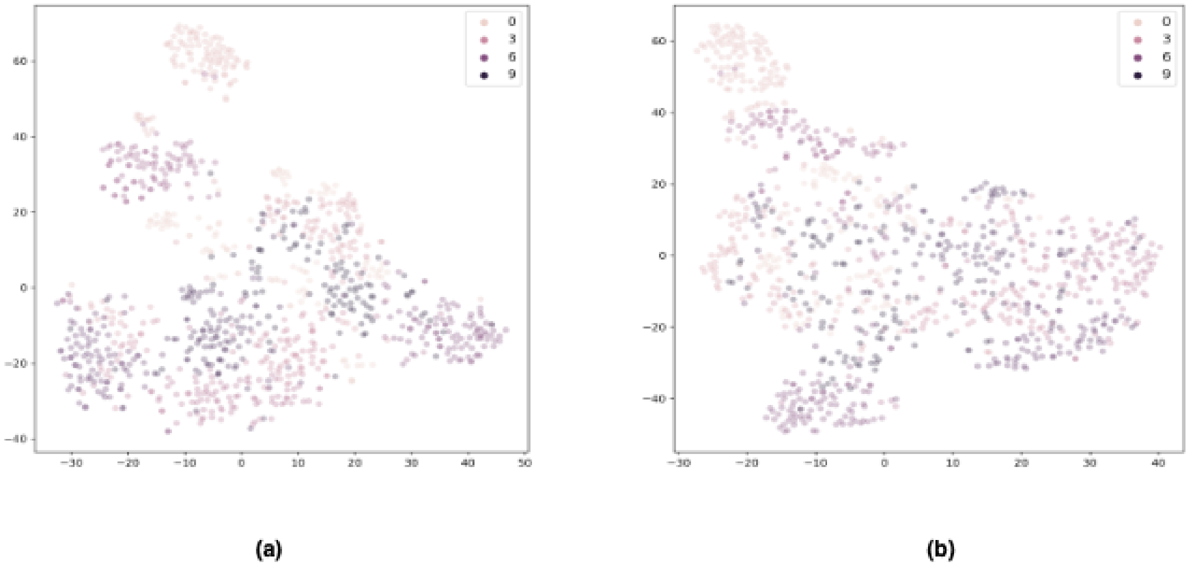


Fig. 3: T-sne plots of audio embeddings of GTZAN test dataset extracted from (a)Pretrained Audio encoder (b) Trained network G_a

by fine-tuning the text encoder. This iteration results in better convergence of loss. To compare the result, we have evaluated the model from both experiments on the GTZAN dataset taking the audio file as input to the audio branch and genre name as text input. The *Fig* 4 represents the t-sne plot of GTZAN audio test files along with genre-name text embeddings using both network. The mapping of genre-name text embedding of fine-tune network is much better as compared to the baseline model. After fine-tuning the text encoder, the model learns better mapping of text and audio The audio-embedding corresponding to the same class genre is better clustered together and genre text embeddings are better matched with the audio-embeddings clusters. The model with fine-tuned text encoder is used to report all the results.

All our experiments are performed with an SGD optimizer and a learning rate of 0.001 with a batch size of 16 audio-caption pairs. Loss is calculated on more number of pairs as mentioned in the Constrative Loss section. To train our model we have taken the first 30s as input of the full audio file.

V. Datasets

We have evaluated our model for auto-tagging and genre classification tasks. Auto-tagging is assigning one or more labels to the audio track and genre classification is classifying a genre for a track. MTG-Jamendo and MagnaTagATune dataset is used for Autotagging and GTZAN for genre classification

A. GTZAN

Dataset (Tzanetakis & Cook 2002) consists of 10 classes of different music genres, and its test dataset consists of 100 audio files corresponding to each genre. The tracks are all 22050Hz Mono 16-bit audio files. Since our training data is consist of caption and music track it motivates us to evaluate the performance of the model on a longer description instead of just only single text. We have reported accuracy on this dataset in two ways, genre name, and genre description. In the genre name category, we have calculated the distance of audio embedding with the text embedding of the genre name, for example: ”rock”. For genre description, We have selected a one-line description corresponding to each genre and generated text embedding corresponding to the genre description. Descriptions corresponds to each genre name is mentioned in the *Table* I.

B. MTG-Jamendo

Dataset (D. Bogdanov & Serra 2019) consists of audio files corresponding to 55,701 full songs created using music publicly available on Jamendo platform under Creative Commons Licenses. The minimum duration of audio files is 30 secs and is available in MP3 format. The audio files are annotated by 195 different tags covering genres, mood/theme, and instruments. The dataset is pre-processed to do tag cleaning before making it available to the public. The dataset is available in audio and mel-spectrogram features correspond to all the audio files. Multiple splits of the dataset into train, validation and test are available. We have evaluated on split-0[1] which contain multiple subsets: mood/theme, instrument, genre, top50. Each subset has a different number of tags and test files. The dataset statistics mentioned in the *Table* II.

[1] https://github.com/MTG/mtg-jamendo-dataset/tree/master/data/splits/

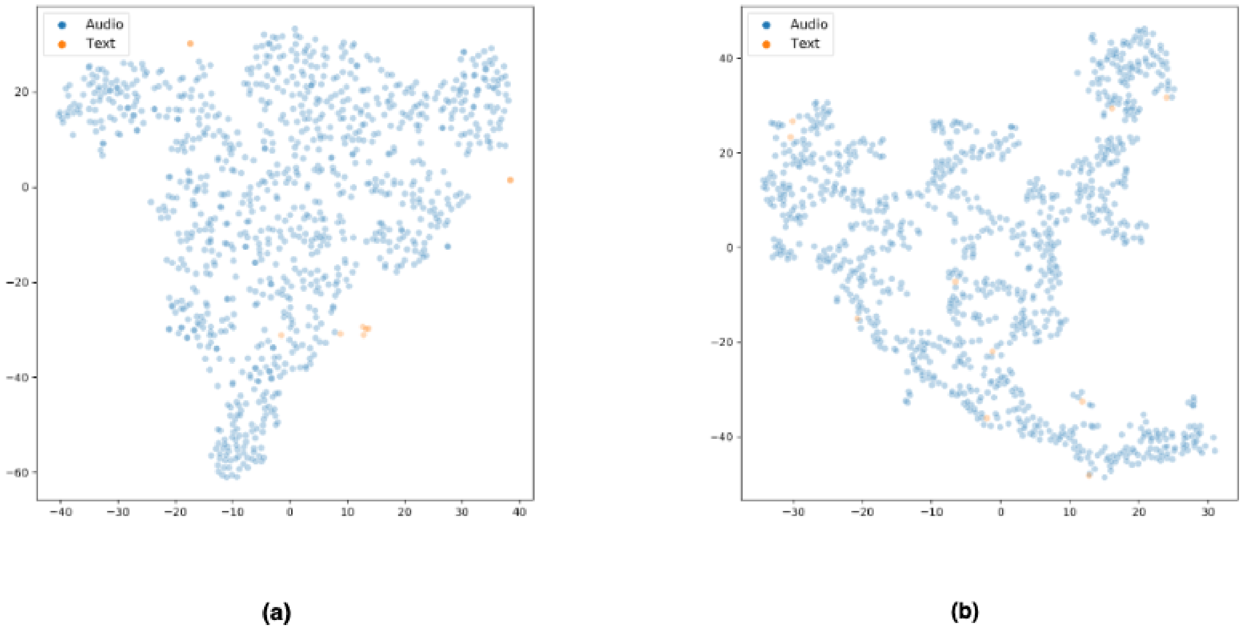


Fig. 4: T-sne plot of audio embedding of 1000 files and text embedding corresponds to 10 genre names of the GTZAN dataset. (a) Baseline (b) Trained with fine-tuning Text encoder

Genre	Description
rock	”Driving, intense, energetic, and almost always loud, rock music”
reggae	”music of Jamaican origin, moderate tempos, guitar/piano offbeats”
pop	”electronic instruments, strong beat and simple tunes”
metal	”a genre of rock music, intense, virtuosic, and powerful”
blues	”dissonant harmonies, flattened ’blue’ notes”
classical	”traditional music, elegance, balance, and homophonic textures”
country	”country choruses, rural music, stringed instruments”
disco	”electronically produced sounds disco”
hiphop	”rhythmic lyrics making use of techniques like assonance, alliteration, and rhyme”
jazz	”complex harmony, distortions of pitch and timbre, and a heavy emphasis on improvisation”

Text

TABLE I: Genre Description

C. *MagnaTagATune(MTAT)*

Dataset (E. Law & Downie 2009) consists of approx 25k music clips each around 30 seconds. One of the most commonly used datasets for music auto-tagging. The audio files are in mp3 format. Dataset with top 50 tags is popularly used for benchmarking and these tags consist of the genre, instrument labels, and vocal information like ”vocal”, ”female voice” etc. We have used standard split [2] and evaluated on 4331 test files. Similar to MTG-Jamendo, this dataset is also multi-label where a single audio file can have more than one tag.

Subset	Tags	Test Files
Genre	95	11479
Mood/theme	41	4231
Instrument	59	5115
Top50	50	11356

TABLE II: The MTG-Jamendo Dataset statistics

Classification Acc	Genre-name	Genre-description
Ours	0.394	0.283
MSD-GLOVE(Choi & Nam 2019*a*)	0.731	-

TABLE III: Evaluation on GTZAN Dataset

VI. Results and Discussion

We have evaluated our model for two tasks, the annotation task, and the retrieval task. We have used ROC-AUC and PR-AUC over a tag as metrics for tag-based retrieval tasks. The annotation task is evaluated on ROC-AUC and PR-AUC over a single instance. The above metrics are used to calculate results on MTG-Jamendo and MagnaTagATune datasets. GTZAN is a single-label class dataset so we have performed simple classification accuracy.

[2] https://github.com/minzwon/sota-music-tagging-models/tree/master/split/mtat

Category	ROC-AUC	PR-AUC
Top50	0.698	0.149
Mood	0.666	0.068
Genre	0.727	0.077
Instrument	0.612	0.100

TABLE IV: Zero-Shot learning results for Retrieval task on MTG-Jamendo Dataset

Category	ROC-AUC	PR-AUC
Top50	0.729	0.217
Mood	0.674	0.175
Genre	0.732	0.173
Instrument	0.679	0.254

TABLE V: Zero-Shot learning results for Annotation task on MTG-Jamendo Dataset

The Zero-shot classification accuracy on the GTZAN dataset with genre name and genre description is shown in *Table* III. To compare our results with prior work on the same dataset we have reported results of MSD-Glove (Choi & Nam 2019*a*). The possible reason for high accuracy by (Choi & Nam 2019*a*) is that they have trained their model on MSD dataset (Thierry Bertin-Mahieux & Lamere 2011) which contains several tag names same as the genre name of GTZAN dataset like ”Reggae”, ”Jazz”.

The task-based retrieval and annotation accuracy of MSD-Jamendo dataset is represented in *Table* IV and *Table* V respectively. The annotation and task-based retrieval accuracy of the MTAT dataset is represented in Table VI. We have reported results of (Choi & Nam 2019*a*) on MTAT dataset. The actual comparison of results on MTAT might vary as (Choi & Nam 2019*a*) doesn’t mention their test dataset split. To the best of our knowledge, no previous work has evaluated Zero-shot accuracy on the MTG-Jamendo dataset.

On further analysis of ROC-AUC and PR-AUC numbers on the MTAT dataset we have found the labels like “male voice”, and “female voice” have the least values as shown in *Table* VII. One possible reason for the poor performance is the absence of such words in the caption dataset. We have performed a similar analysis on the top50 tags of the MTG-Jamendo dataset. Surprisingly, the model didn’t perform well on a few common tags “drums”, ”keyboard” etc. Further analysis needs to be done on the training caption dataset to understand the performance of the network on these tags.

Method	Ours	MSD-Glove
$ROC - AUC_R$	0.735	0.739
$PR - AUC_R$	0.227	-
$ROC - AUC_A$	0.770	-
$PR - AUC_A$	0.360	-

TABLE VI: Zero-Shot learning results for Retrieval task(R) and Annotation task(A) on MagnaTagATune Dataset

Best Performing Tags	’metal’, ’rock’, ’choral’, ’dance’, ,’country’, ’techno’, ’opera’, ’harpsichord’, ’piano’, ’choir’, ’classical’, ’beat’, ’electronic’, ’violin
Worst Performing Tags	’male voice’, ’no vocals’, ’female voice’, ’no vocal’, ’new age’, ’voice’, ’vocals’, ’no voice’, ’vocal’, ’female vocal’, ’female’, ’man’, ’woman’

TABLE VII: MTAT tags performance

Best Performing Tags	’metal’, ’rock’, ’classical’, ’house’, ’orchestral’, ’techno’, ’popfolk’, ’folk’, ’indie’, ’hiphop’, ’dance’, ’reggae’, ’film’
Worst Performing Tags	world’, ’drummachine’, ’instrumentalpop’,’voice’,’newage’,’bass’, ’drums’,’keyboard

TABLE VIII: MTG-Jamendo tags performance

VII. Conclusion and future work

We have presented a cross-modal generative framework to investigate whether natural language supervision can be leveraged to achieve good Zero-shot performance when transferring a pre-trained multimodal to unseen tasks. We have shown that model is able to associate with new tags and labels using the side information. To show the learned representation are generic and robust, we have shown the model performance on several MIR downstream tasks.

In future work, different audio encoder backbones (J. Pons & Serra 2018*a*,*b*, K. Choi & Cho 2017) can be explored. They can capture more music domain information. In our current work, we do not fine-tune the audio encoder which can also be done in future experiments. Another work of direction in terms of experimenting with different loss functions like discriminator loss along with generator network.

VIII. Acknowledgement

I would like to express my deepest appreciation to both my supervisor, Elio Quinton, and George Fazekas for their guidance and support throughout the project. Special gratitude to Ilaria Manco for helping me with my project at all stages with patience and for always being available for feedback and discussions. Furthermore, I would like to thank Universal Music Group for providing data and relevant resources to train my model.

References

A. Ferraro, X. Favory, K. D. Y. K. & Bogdanov, D. (2021), ‘Enriched music representations with multiple cross-modal contrastive learning’, *IEEE Signal Processing Letters, vol. 28, pp. 733–737, Apr. 2021.* .

Choi, J., L. J. P. J. & Nam, J. (2019*a*), ‘Zero-shot learning and knowledge transfer in music classification and tagging’, *arXiv preprint arXiv:1906.08615* .

Choi, J., L. J. P. J. & Nam, J. (2019*b*), 'Zero-shot learning for audio-based music classification and tagging', *arXiv preprint arXiv:1907.02670.* .

D. Bogdanov, M. Won, P. T. A. P. & Serra, X. (2019), 'The mtg-jamendo dataset for automatic music tagging', *International Conference on Machine Learning (ICML), 2019.* .

E. Law, K. West, M. I. M. M. B. & Downie, J. S. (2009), 'Evaluation of algorithms using games: The case of music tagging', *International Society for Music Information Retrieval Conference (ISMIR)* .

F. Faghri, D. J. Fleet, J. R. K. & Fidler., S. (2017), 'Vse++: Improving visual-semantic embeddings with hard negatives', *arXiv:1707.05612 [cs], July 2017. arXiv: 1707.05612.* .

Frederik Pahde, Mihai Puscas, T. K. & Nabi, M. (2021), 'Independent prototype propagation for zero-shot compositionality.', *In IEEE Winter Conference on Applications of Computer Vision (WACV)* .

Ishan Misra, A. G. & Hebert, M. (2017), 'From red wine to red tomato: Composition with context', *IEEE Conference on Computer Vision and Pattern Recognition, pages 1792–1801* .

J. Devlin, M. W. Chang, K. L. & Toutanova, K. (2019), 'Bert:pre-training of deep bidirectional transformers for language understanding', *INAACL-HLT* .

J. Pons, O. Nieto, M. P. E. S. A. E. & Serra, X. (2018*a*), 'End-to-end learning for music audio tagging at scale', *In Proc. of the 19th International Society for Music Information Retrieval Conference (ISMIR), 2018* .

J. Pons, O. Nieto, M. P. E. S. A. E. & Serra, X. (2018*b*), 'End-to-end learning for music audio tagging at scale', *In Proc. of the 19th International Society for Music Information Retrieval Conference (ISMIR), 2018* .

Juhan Nam, Keunwoo Choi, J. L. S.-Y. C. & Yang, Y.-H. (2019), 'Deep learning for audio-based music classification and tagging:teaching computers to distinguish rock from bach', *IEEE Signal Processing Magazine* .

K. Choi, G. Fazekas, M. S. & Cho, K. (2017), 'Convolutional recurrent neural networks for music classification', *In Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017, pp. 2392–2396* .

Manco, I., B. E. Q. E. & Fazekas, G. (2022), 'Learning music audio representations via weak language supervision', *In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 456-460). IEEE* .

R. Kiros, R. S. & Zemel, R. S. (2014), 'Unifying visual-semantic embeddings with multimodal neural language models', *arXiv:1411.2539 [cs], Nov. 2014. arXiv: 1411.2539.* .

Ruis, F., B. G. & Bucur, D. (2021), 'Independent prototype propagation for zero-shot compositionality.', *Advances in Neural Information Processing Systems, 34, pp.10641-10653.* .

S. Reed, Z. Akata, X. Y. L. L. B. S. & Lee, H. (2016), 'Generative adversarial text to image synthesis.', *In ICML, volume 48 of Proceedings of Machine Learning Research, pages 1060–1069. PMLR, 2016.* .

Saeed, A., G. D. & Zeghidour, N. (2021), 'Contrastive learning of general-purpose audio representations', *In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 3875-3879). IEEE.* .

Sandouk, U. & Chen, K. (2016), 'Multi-label zero-shot learning via concept embedding', *arXiv preprint arXiv:1606.00282.* .

Spijkervet, J. & Burgoyne, J. (2021), 'Contrastive learning of musical representations', *arXiv preprint arXiv:2103.09410.* .

Thierry Bertin-Mahieux, Daniel P.W. Ellis, B. W. & Lamere, P. (2011), 'he million song dataset', *In Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR 2011)* .

Tzanetakis, G. & Cook, P. (2002), 'Musical genre classification of audio signals', *IEEE Transactions on speech and audio processing, 10(5):293–302, 2002.* .

Wang, Q. & Chen, K. (2017), 'Multi-label zero-shot human action recognition via joint latent embedding', *arXiv preprint arXiv:1709.05107, 2017.* .

Won, M., C. S. N. O. & Serrc, X. (2020), 'Gdata-driven harmonic filters for audio representation learning', *In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 536-540). IEEE* .

Won, M., Ferraro, A., Bogdanov, D. & Serra, X. (2020), 'Evaluation of cnn-based automatic music tagging models', *arXiv preprint arXiv:2006.00751,2020* .

Won, M., O. S. N. O. G. F. & Serra, X. (2021), 'Multimodal metric learning for tag-based music retrieval', *In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 591-595). IEEE.* .

X. Favory, K. Drossos, T. V. & Serra, X. (2021), 'Learning contextual tag embeddings for cross-modal alignment of audio and tags', *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* .

X. Favory, K. Drossos, V. T. & Serra, X. (2020), 'Coala:co-aligned autoencoders for learning semantically enriched audio representations', *International Conference on Machine Learning (ICML), Workshop on Self-supervised learning in Audio and Speech* .

YongqinXian, BerntSchiele, a. Z. (2017), 'Zero-shot learning - the good, the bad, and the ugly', *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* .

Zhou Ren, Hailin Jin, Z. L. C. F. & Yuille., A. (2017), 'Multiple instance visual-semantic embedding', *In Proc. of the British Machine Vision Conference (BMVC)* .

Appendix

This study is an extension of the experiment with Discriminator loss. Due to time-constraint, we are not able to develop meaningful results with this loss but recent work in GAN has shown the generator-discriminator holds great potential in learning the embedding in multi-modal space. This could be the potential direction to extend the Zero-shot methodology work in multimodality space.

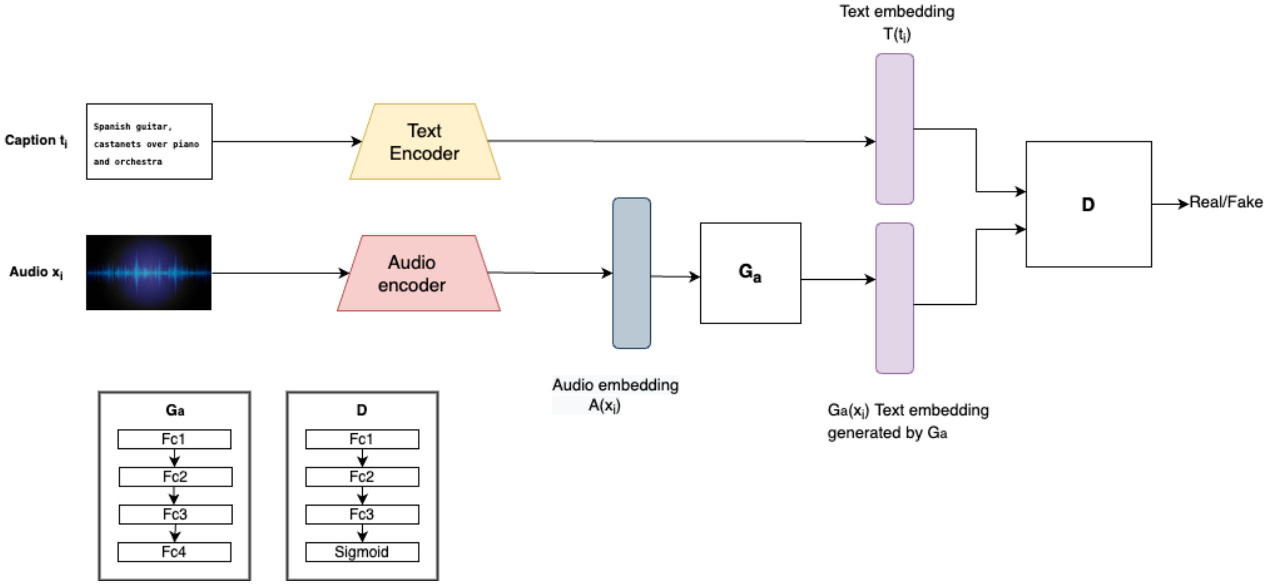


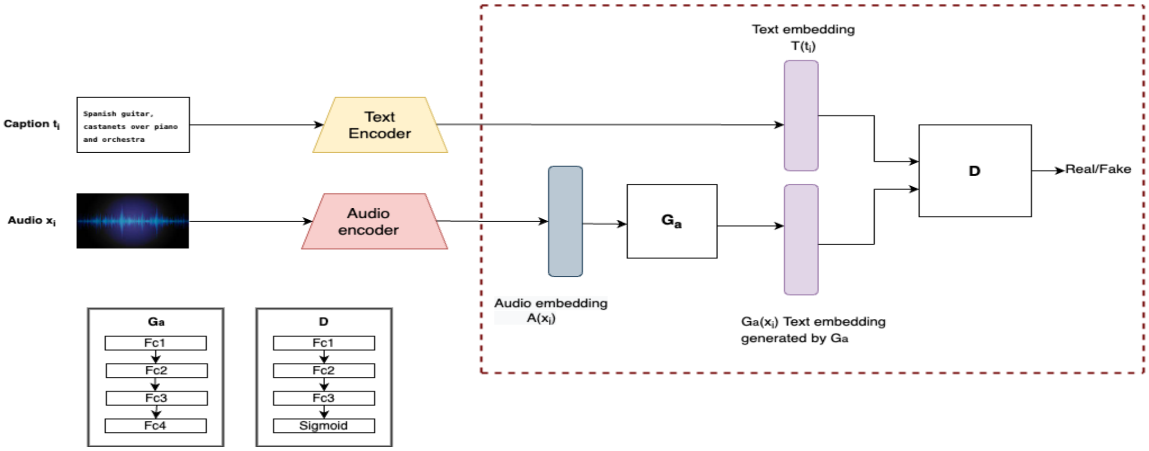
Fig (1) Proposed multimodal network with Discriminator loss

Discriminator Loss:

In this part of the experiments, we have used discriminator loss instead of contrastive loss along with Generator network Ga. Generator part is initialized with random noise and is trained to generate meaningful data with Discriminator loss. In our case, instead of taking initial embedding from random noise space, we are taking it from the pre-trained audio encoder and with the help of discriminator loss, the cross-modal feature generation model learns to bring the audio embedding into the text embedding space. Discriminator network learns to classify the generated and original text embedding as real or fake. The generator and discriminator optimize the transform audio embedding given by the audio encoder into the text embedding space generated by the pre-trained text encoder. As mentioned earlier, Ga consists of 4 fully connected layer and take input 256-dimensional vector from the trained audio encoder. Our discriminator network D consists of 3 fully connected layers and 1 sigmoid layer. Except for the last FC layer, each layer is followed by Relu. Fig 1 has shown the full proposed network architecture.

Ga and D networks combined work to solve the adversarial game of generating the text embedding Ga(xi) that can not be differentiated from text embedding T(ti) and predicting the generated embedding as fake. Since, during evaluation, we have to perform similarity between the embeddings we are not generating the text sample. We are optimizing Ga to generate the feature vector in text embedding space. Our Ga and D network is trained by optimizing the following loss function:

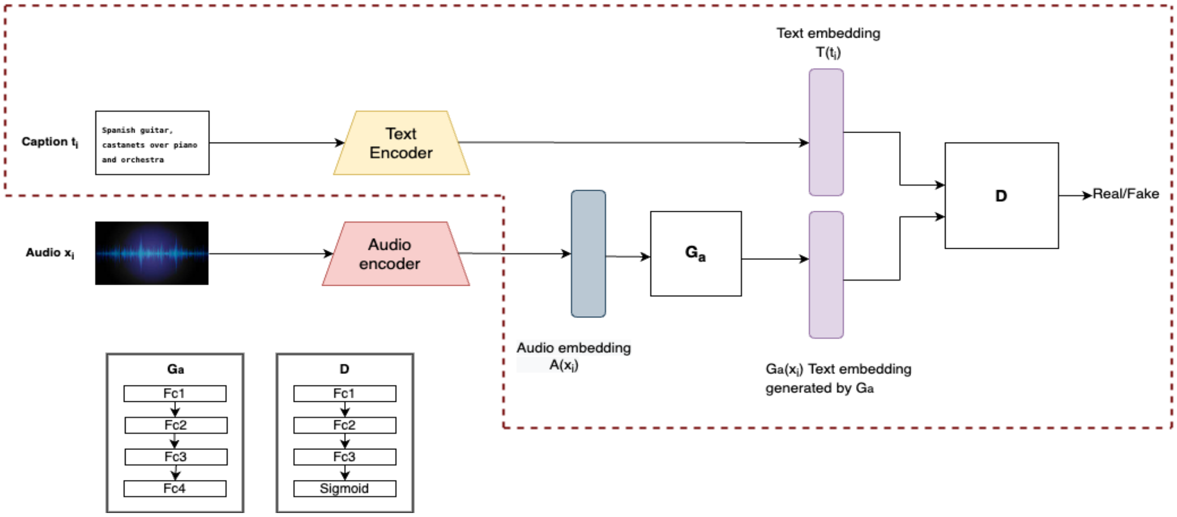
$$L(Ga, D) \;=\; log(D(Ga(xi)) \;+\; log\, D(Ga(xi),\, T(ti)) \text{........................eq(1)}$$



Fig(2) Training network without finetuning Text Encoder

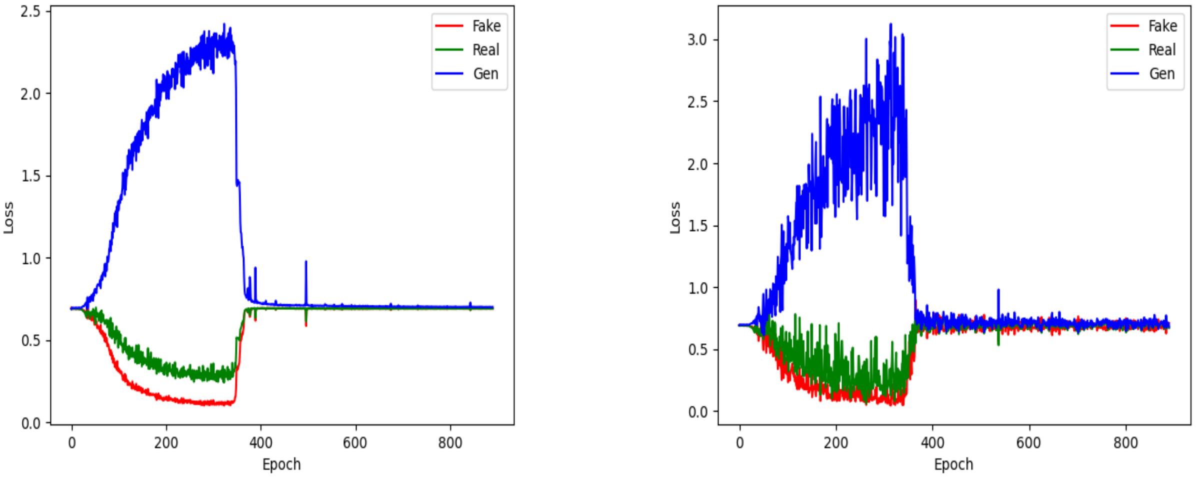
Experiments:

In our first experiment, we trained only Generator and discriminator network as shown in Fig 2. We take text embedding directly from the text encoder. Audio embedding is generated by the audio encoder and passed to the generator. The Generator along discriminator is trained to convert the embedding A(xi) from audio embedding space into text embedding by optimizing loss eq1. The loss values for both the networks was very high and it was hard to stabilize the loss value of both network with this setting.



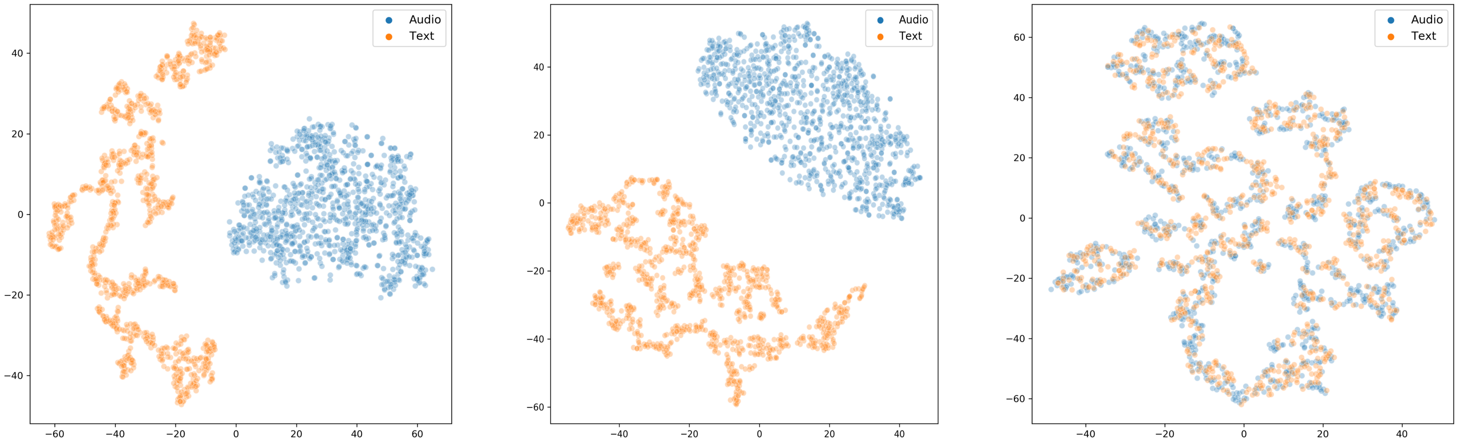
Fig(3) Training network with finetuning Text Encoder

In the second set of experiments, We have fine-tuned the text encoder and trained the generator and discriminator network. To stabilize our training better, instead of finetuning the text encoder along with the generator and discriminator we have pre-trained the text encoder on the caption dataset. During Generator and discriminator training we have kept the text and audio encoder both frozen. This results in more stable training. Figure(4) represents the loss curves of generator and discriminator on both train and validation datasets respectively.



(a) (b)

Fig(4) (a) Train Loss (b) Validation Loss



Epoch:0 **Epoch:50** **Epoch:112**

Fig(5) T-sne plot of embeddings of audio-text pairs of validation data across a different epoch of the trained model

Fig(5) represent the embeddings of audio-text pair across different epoch. On further analysis of validation data at epoch 112, we have observed the embedding has formed the clustered instead of getting close to the corresponding pair. This observation is very different from experiments with contrastive loss. One possible hypothesis for this clustering can be the clustering of similar text features. As we have used a pre-train encoder that is trained on the caption dataset, there is a possibility that it might have already learned some similarities between the similar text and formed cluster. During training, the model changes the audio embedding into text embedding space which already has clusters.

Further Work:

Further work can be done to verify the above hypothesis. To verify it, we can map all embedding pairs from one cluster with ground truth text to understand any similarity between the texts. Or adding a discriminative loss along with GANs loss to further push the text embedding apart.
Another set of experiments can be done by utilizing the model trained with contrastive loss. Text encoder and Generator network trained with contrastive loss can be utilized as the pre-trained weight for the framework with Discriminator loss.