EMOTION EMBEDDING SPACES FOR MATCHING MUSIC TO STORIES

Minz Won[1]   **Justin Salamon**[2]   **Nicholas J. Bryan**[2]   **Gautham J. Mysore**[2]   **Xavier Serra**[1]

[1] Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain
[2] Adobe Research, San Francisco, California, United States

`minz.won@upf.edu`

ABSTRACT

Content creators often use music to enhance their stories, as it can be a powerful tool to convey emotion. In this paper, our goal is to help creators find music to match the emotion of their story. We focus on text-based stories that can be auralized (e.g., books), use multiple sentences as input queries, and automatically retrieve matching music. We formalize this task as a cross-modal text-to-music retrieval problem. Both the music and text domains have existing datasets with emotion labels, but mismatched emotion vocabularies prevent us from using mood or emotion annotations directly for matching. To address this challenge, we propose and investigate several emotion embedding spaces, both manually defined (e.g., valence/arousal) and data-driven (e.g., Word2Vec and metric learning) to bridge this gap. Our experiments show that by leveraging these embedding spaces, we are able to successfully bridge the gap between modalities to facilitate cross modal retrieval. We show that our method can leverage the well established valence-arousal space, but that it can also achieve our goal via data-driven embedding spaces. By leveraging data-driven embeddings, our approach has the potential of being generalized to other retrieval tasks that require broader or completely different vocabularies.

1. INTRODUCTION

Content creators, both amateur and professional alike, often use music to enhance their storytelling due to its powerful ability to elicit emotion [1]. For example, when dissonant music is added to a horror movie, it can amplify the scary mood of the story line. Similarly, cheerful music can emphasize the excited mood in a scene of a birthday party. Matching text and music to create a narrative, typically requires tediously browsing large-scale music collections, significant experience, and musical expertise. In this

---

[1] We use the terms *emotion* and *mood* interchangeably following previous work [1].
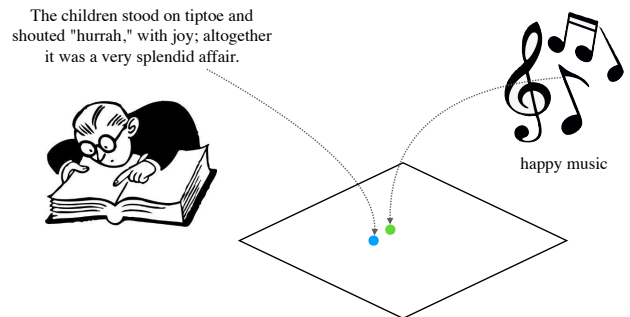
Figure 1: Cross-modal text-to-music retrieval using an aligned, multimodal embedding space.

paper, we therefore address the problem of automatically matching music to text as shown in Figure 1.

We formalize this task as a cross-modal retrieval problem [2] and focus on matching long-form text (multiple sentences, paragraphs) to music. For queried sentences like books and scripts, we seek to retrieve matching music for applications such as podcasts, audio books, movies, and film. To facilitate cross-modal retrieval, a common approach is to first perform feature extraction to convert each data modality into an embedding space. Then, the different embedding spaces must be matched to bridge the *modality gap* by somehow aligning their different distributions [2]. Once aligned, (fast) nearest neighbor search can be used for retrieval.

Various methods have been proposed for cross-modal feature extraction and alignment. For example, canonical correlation analysis has been used to bridge the modality gap [3] as well as modern deep learning techniques that learn common representation spaces [4, 5]. Such methods can be categorized into four groups: unsupervised, pairwise-based, rank-based, and supervised methods [6]. Among these, supervised methods are the most straightforward and in theory can take advantage of existing labeled datasets (e.g., labels of *happy*, *sad*) and themes (e.g., *party*, *wedding* with corresponding text and music). Difficulties, however, immediately arise because of mismatched dataset taxonomies (vocabularies) per modality, making it challenging to use standard techniques directly.

Therefore, in this work we focus on the task of emotion-based text- (e.g. sentences, paragraphs) to-music retrieval, and investigate how we can best perform cross-modal retrieval with heterogeneous dataset taxonomies. To the best of our knowledge, this problem has not been previously ad-

dressed and could be beneficial to media content creation applications. We propose six different deep learning strategies to extract relevant features and bridge the modality gap between text and music including (1) classification (2) multi-head classification (3) valence-arousal regression (4) Word2Vec regression (5) two-branch metric learning and (6) three-branch metric learning. We then evaluate each approach on multiple text and music datasets, report objective results via precision at five and mean reciprocal rank, and conclude with qualitative analysis and discussion. Our results show that our valence-arousal-based method is a powerful baseline for emotion-based cross-modal retrieval, but that our three-branch metric-learning approach is comparable, more general, and does not require manually engineered valence and arousal mappings.

2. RELATED WORK

2.1 Text Emotion Classification

Text emotion classification methods or the task of predicting emotion from text can be divided into three categories: lexicon-based models, traditional machine learning models, and deep learning models. Lexicon-based models take advantage of pre-defined emotion lexicons, such as NRC EmoLex [7] and WordNet-Affect [8] to match keywords. Traditional machine learning approaches recognize emotions using algorithms such as support vector machine (SVM) [9] and Naive Bayes [10]. Finally, deep learning models use deep sequence models such as gated recurrent unit (GRU) [11], bidirectional long-short term memory (BiLSTM) [12], and Transformers [13]. Most recently, Transformer models [14–16] have become quite prevalent. Such models take advantage of transfer learning, are commonly pre-trained to learn language representation with large datasets, and then applied to various downstream tasks including question and answer systems as well as emotion recognition [13].

2.2 Music Emotion Classification

Music emotion classification or the task of predicting emotion from music audio is commonly divided into conventional feature extraction and prediction approaches [17–19], and end-to-end deep learning approaches [20, 21]. Deep learning approaches have become most prevalent and commonly frame emotion recognition as a multi-class or multi-label auto-tagging classification problem [22–26]. Recently, multiple music tagging models were evaluated in a homogeneous evaluation pipeline [27] and found three design recommendations for automatic music tagging models: (1) use mel-spectrogram inputs, (2) use 3×3 convolutional filters, and (3) use short-chunk audio inputs with small hop sizes and max-pooling. Based on this, a model using mel-spectrogram inputs and convolutional neural networks with focal loss [28] won the MediaEval 2020 Emotion-and-Theme-Recognition-in-Music-Task [2] [29].

---

[2] https://multimediaeval.github.io/2020-Emotion-and-Theme-Recognition-in-Music-Task

Tag	GoogleNews	Domain-specific [35]
Chill	chilly, cold, chilled, chills, shivers, shiver, warm, frigid, frosty, balmy	**chill_out**, relax, chilled, kick_back, relaxing, **chill-out**, **chilled_out**, **downtempo**, **down_tempo**, unwind

Table 1: Nearest words in GoogleNews and domain-specific word embeddings [35]. Music-related words are highlighted in bold.

2.3 Valence-arousal Regression & Word Embeddings

Beyond classification, previous works [30, 31] suggest that regression approaches can outperform classification approaches in music emotion recognition. Here, researchers use the well-known valence-arousal emotion space [32, 33] where valence represents positive-to-negative emotions, and arousal indicates the intensity of the emotions. These annotations can be collected by human annotators directly [30] or by mapping existing mood labels into the valence-arousal space using pre-defined lexicons [21, 34].

As an alternative to using the manually annotated valence-arousal space, we can obtain tag (mood) embeddings in a more data-driven fashion. Pre-trained word embeddings, such as Word2Vec [36] and GloVe [37], represent words as vectors by learning word associations from a large corpus. These embedding spaces use the cosine similarity as a measure of semantic similarity. Recent works [35, 38] show the suitability of pre-trained word embedding in music retrieval and that the embedding can include more music related context by training it with music related documents [35, 39]—see Table 1.

2.4 Cross-modal Retrieval

Instead of targeting a pre-defined embedding space, multimodal metric learning models aim at learning a shared embedding space in which semantically similar items are close together while dissimilar items are far apart in the embedding space. Unsupervised approaches leverage co-occurrence information. For example, when we collect user-created video from the web, the video and audio streams are synchronized, and this correspondence can be exploited for representation learning [40, 41]. On the other hand, supervised methods learn discriminative representations by exploiting annotated labels. Here, data from different modalities are used to train models such that data points with the same label should be close while data with different labels should be far apart. Metric learning is also used for bridging the modality gap between text and audio, such as keyword spotting [42], text-based audio retrieval [43, 44], and tag-based music retrieval [35, 38] in both supervised and unsupervised ways.

Two branch metric learning [45] is one of the most prevalent architectures for cross-modal retrieval. It consists of two branches where each branch extracts features from each modality and maps them into a shared embedding space. When optimized with a conventional triplet loss (e.g. anchor text, positive song, negative song), however, the model loses neighborhood structure within

modalities. To alleviate this issue, previous work [46] added structure-preserving constraints by using additional triplet losses within modalities (e.g., anchor text, positive text, negative text).

3. MODELS

Cross-modal retrieval comprises two parts: feature extraction and bridging the modality gap. Our text and music embeddings, E_{text} and E_{music} respectively, are defined as follows:

$$E_{text} = M(P_{text}(x))$$
$$E_{music} = M(P_{music}(x)) \quad (1)$$

where P is a pre-trained model to extract features from each modality and M is a multilayer perceptron (MLP) to map them to a multimodal embedding space.

3.1 Pre-trained Models for Feature Extraction

In our work, we leverage the DistilBERT [16] transformer model for text analysis, which is a compact variant of the popular BERT transformer model [14, 16]. We use a pre-trained model from the Huggingface library [47].

For the music representation model P_{music}, we use a CNN with residual connections that are trained with mel-spectrograms (ResNet) [27]. Due to its simplicity and high performance, it is a broadly used architecture not only in music but also in general audio representation learning. Our ResNet consists of 7 convolutional layers with 3×3 filters followed by 2×2 max-pooling. The model is pre-trained with the MagnaTagATune dataset [3] [48]. Both pre-trained models are updated during the training process so that they can adapt to the data.

3.2 Embedding Models to Bridge the Modality Gap

3.2.1 Classification

As a starting point, we train two separate mood classification models for text and music (Figure 2-(a)). Then the model returns mood predictions and their likelihood with softmax. From the predicted text mood, songs are re-ranked based on their likelihoods of the text mood. However, this classification approach has an inherent limitation- the model cannot bridge the modalities when they have different mood taxonomies.

3.2.2 Multi-head Classification with Shared Weights

Multi-head model is similar to the classification model but it shares a 3-layered MLP for multimodal fusion in it (Figure 2-(b)). Since the model shares the weights across different modalities, it can predict the mood in different taxonomies by switching the classification head. We included this model to see if the shared MLP can generalize across modalities.

---

[3] We use the pre-trained model from this open source repository: https://github.com/minzwon/sota-music-tagging-models

3.2.3 Regression

Following previous work [21], we reformulate the classification task as a regression problem. By using NRC VAD Lexicon [34], emotion labels can be mapped to the valence-arousal space. However, this mapping process is hand-crafted and also they cannot handle bi-grams or tri-grams since the lexicon was created in a word-level. In addition to leveraging the valence-arousal space, we also experiment with a Word2Vec [36] embedding which was pre-trained with music related text [35]. This data-driven space supports a larger vocabulary, including bi-grams and tri-grams, and is thus more flexible.

Regression models are trained separately for each modality (Figure 2-(a)). Then the nearest items are retrieved based on their distance in the embedding space. Note that, distance metrics are Euclidean distance for the valence-arousal space, and cosine distance for the Word2Vec space. However, regression is a one-way optimization, i.e., optimizing text or mood into the pre-defined word embedding space. In this case, neighborhood structure within each modality can be ignored. For example, music with *angry* and *exciting* can share similar acoustic characteristics. However, if two words are far apart in Word2Vec space, this similarity cannot be considered by regression. This obstacle motivates us to learn a shared embedding space in a data-driven fashion using metric learning.

3.2.4 Metric Learning

Finally, we explore metric learning, which is a fully data-driven approach that solves the cross modal text-to-music retrieval in an end-to-end manner. Metric learning is optimized to minimize a triplet loss $\mathcal{T}$:

$$\mathcal{T}(E_a, E_p, E_n) = [D(E_a, E_p) - D(E_a, E_n) + \delta]_+ \quad (2)$$

where D is a cosine distance function, δ is a margin, and E_a, E_p, E_n are embedding of anchor, positive, and negative examples, respectively. $[\cdot]_+$ is rectified linear unit. Following conventional metric learning models for cross-modal retrieval, we implement a two-branch metric learning model [45] (Figure 2-(c)) that optimizes the loss function L,

$$L = \mathcal{T}(E_{text}^a, E_{music}^p, E_{music}^n). \quad (3)$$

However, with the triplet function, neighborhood structure or data distribution within modalities can be lost. Structure-preserving constraints [46] can alleviate the issue but our problem is different from the case, since we have different taxonomies across the modality which includes many non-overlapped moods.

To take advantage of different mood distribution of different modalities, we investigate metric learning model with three branches (Figure 2-(d)) that results in three triplet loss functions. Each loss function is designed to optimize tag-to-text, tag-to-music, and text-to-music triplet losses as following:
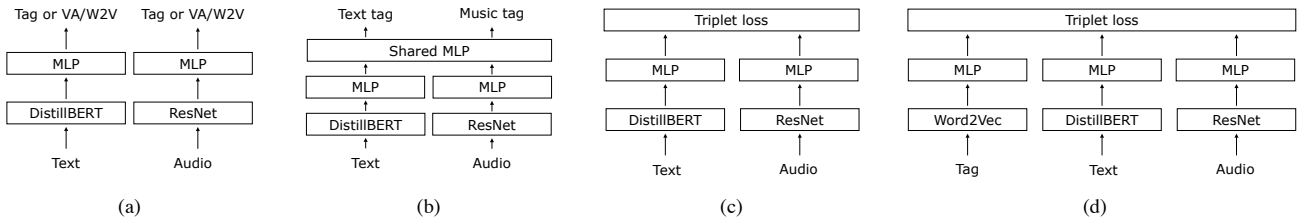
Figure 2: Model architectures. (a) Classification and regression models (b) Multi-head classification model with shared weights (c) Two-branch metric learning (c) Three-branch metric learning.

$$L_{text} = \mathcal{T}(E_{tag}^a, E_{text}^p, E_{text}^n),$$
$$L_{music} = \mathcal{T}(E_{tag}^a, E_{music}^p, E_{music}^n), \quad (4)$$
$$L_{cross} = \mathcal{T}(E_{text}^a, E_{music}^p, E_{music}^n).$$

The model learns a shared mood space between Word2Vec embedding and text embedding with a loss L_{text}, and a shared mood space between Word2Vec embedding and music embedding with a loss L_{music}. Finally, they are bridged together with a cross-modal triplet loss L_{cross}. We refer to this model as three-branch metric learning.

Since text and music have different vocabularies in our scenario, for both two-branch and three-branch metric learning, we regard the nearest tags in pre-trained Word2Vec space as positive pairs in cross-modal triplet sampling (Table 2). We used distance-weighted sampling [49] for more efficient negative mining following previous work [35].

4. EXPERIMENTAL DESIGN

4.1 Text Datasets

Alm's affect dataset [50] includes 1,383 sentences collected from books written by three different authors: B. Potter, H.C. Andersen, and the Brothers Grimm. 1,207 sentences in the dataset are annotated with one representative emotion among five: *angry*, *fearful*, *happy*, *sad*, and *surprised*. To avoid unintended information leakage, we decided to split data in an author-level. 1,040 sentences by the Brothers Grimm and H.C. Andersen were used for training and 167 sentences by B. Potter were used for validation and test.

ISEAR dataset [51] is a corpus with 7,666 sentences that are categorized into one of seven emotion: *anger*, *disgust*, *fear*, *joy*, *sadness*, *shame*, and *guilt*. Each sentence describes certain antecedents and those are associated with according reactions (emotions). We split the dataset in a stratified manner with ratio of 70% train, 15% validation, and 15% test set.

4.2 Music Dataset

There are multiple datasets for music emotion recognition such as the Million Song Dataset (MSD) subset [52], the MTG-Jamendo mood subset [53], and the AudioSet mood subset [54]. Before we choose our dataset, we run classification experiments for each subset. AudioSet subset returned the highest accuracy, which means the labeled emo-

Original	VA	W2V	Manual
anger	angry	angry	angry
fearful	sad	scary	scary
happy	happy	happy	exciting, funny, happy
sad	sad	sad	sad
surprised	exciting	happy	exciting
anger	angry	angry	angry
disgust	angry	angry	angry, scary
fear	angry	angry	scary
guilt	sad	angry	angry, sad
joy	exciting	tender	exciting, funny, happy
sadness	sad	tender	sad
shame	angry	sad	angry, sad

Table 2: Similar moods from Alm's dataset (upper) and ISEAR dataset (lower). Original is from text mood taxonomy and mapped tags are from music dataset.

tions are predictable with our ResNet model. One possible reason for this result is that unlike other datasets, emotion labels of AudioSet subset are exclusive, having a single emotion label per song. This is also beneficial since we can map each song directly to the valence-arousal space or word embedding space using emotion lexicons or Word2Vec model, respectively. Otherwise, to handle multiple tags, we need to average their embedding vectors as previous researchers did [21]. For these simplicity and reliability reasons, we use AudioSet mood subset.

AudioSet [54] mood subset consists of 16,995 music clips collected from YouTube and each audio clip is 10-second long. The dataset is categorized into 7 mood categories: *happy*, *funny*, *sad*, *tender*, *exciting*, *angry*, and *scary*. The dataset is provided with a training set of 16,104 clips and an evaluation set of 540 clips.

4.3 Evaluation

We use two evaluation metrics: Precision at 5 (P@5) and Mean Reciprocal Rank (MRR). However, since our text and audio datasets use different taxonomies, we need a mapping between the different vocabularies in order to compute the metrics directly. Thus, we map the text emotion taxonomy to the music emotion taxonomy — see Table 2. We introduce three possible mappings: (1) mapping based on the Euclidean distance between emotion labels in the valence-arousal space (VA), (2) the cosine distance between emotion labels in Word2Vec space (W2V), or (3) direct manual mapping of emotion labels. Given these mappings, we compute P@5 and MRR. Another challenge is the label distribution in our datasets, which is unbalanced.

Methods	Alm's dataset						ISEAR dataset					
	VA		W2V		Manual		VA		W2V		Manual	
	P@5	MRR	P@5	MRR	P@5	MRR	P@5	MRR	P@5	MRR	P@5	MRR
Classification	0.2161	0.2436	0.1861	0.2157	0.2161	0.2436	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Multi-head Classification	0.2819	0.4181	0.1271	0.1381	0.3446	0.5304	**0.3440**	0.5084	0.3325	0.3625	0.3551	0.4803
V-A Regression	**0.4325**	**0.6282**	0.4125	0.5749	**0.6100**	**0.7398**	0.3018	**0.5247**	0.1866	0.3709	**0.6218**	0.7075
W2V Regression	0.3960	0.5010	0.4613	0.5591	0.5413	0.6363	0.3008	0.3829	0.4164	0.4908	0.5527	**0.7668**
Metric Learning (2-branch)	0.3399	0.3778	0.4897	0.5239	0.5374	0.5579	0.2695	0.3287	0.3951	0.4336	0.4438	0.6175
Metric Learning (3-branch)	0.3574	0.4348	**0.5095**	**0.5863**	0.5156	0.5880	0.2591	0.3445	**0.4317**	**0.4953**	0.6019	0.6675

Table 3: Retrieval scores

This can lead to over-optimistic results if the model performs well on the majority class, even if it performs very poorly on less common labels in the test dataset. To alleviate this problem, we compute the macro-P@5 and macro-MRR, i.e., we compute the metrics per class (emotion label) then average the per-class results. Henceforth we will use P@5 and MRR to denote *macro* P@5 and MRR, respectively.

Regression models are optimized to reduce mean squared error and metric learning models are optimized with the triplet losses detailed in Section 3.2.4. We use the Adam optimizer with learning rate 0.0001 for all models. Audio inputs are resampled into 16 kHz and then converted to 128-bin mel-spectrograms via a 512-point FFT with 50% frame overlap. Implementation details are available online [4].

5. RESULTS

5.1 Quantitative Results

The retrieval results for the different proposed models, using our three different proposed vocabulary mappings (VA, W2V, Manual), for our two text datasets, are presented in Table 3. First, we see that the classification model fails in cross-modal retrieval. Since there are only two emotions in common between Alm's dataset and AudioSet (i.e., *happy* and *sad*), text inputs with other emotions will not have any retrieval result. Furthermore, there's no common emotion between ISEAR dataset and AudioSet, hence P@5 and MRR are zero in this case. Classification models can be powerful when there are exactly identical or partially overlapped vocabularies, but since it is less likely in real-world data, classification approach is less desirable for cross-modal retrieval.

The multi-head classification model also performs worse than other regression and metric learning models. Some metrics look optimistic but when we check the confusion matrix of the multi-head classification model, it constantly predicts one or two specific emotions (e.g., predict *angry* for any type of input) no matter what the input is. This means the shared MLP cannot generalize across different modality heads.

The regression model using valence-arousal consistently shows the best metrics as already proven in previous single-modality emotion recognition works [30, 31]. Since

---
[4] https://github.com/minzwon/text2music-emotion-embedding.git

the space is carefully designed and the tag-to-space mapping process has been done manually [34], the valence-arousal regression suits our cross-modal retrieval task. However, this method cannot generalize to other datasets that possibly have some tags that do not have manual tag-to-space mapping. Word2Vec regression is suitable in that case. It shows slightly lower but comparable retrieval performance and it can handle abundant vocabulary, even bigrams and tri-grams, without a manual mapping process.

Finally, we assess the performance of metric learning. Instead of predicting manually defined or pre-trained embeddings, metric learning aims at learning a shared embedding space across different modalities. Both two-branch and three-branch approaches claim their suitability for cross-modal retrieval, and the three-branch metric learning model consistently outperforms the two-branch model by leveraging the relationship of tag-to-text and tag-to-music within each modality.

5.2 Qualitative Results

To further investigate the characteristics of various embedding spaces, we visualize them with 2D projection—Figure 3. Due to limited space, we only visualize embedding spaces with Alm's dataset and AudioSet mood subset. Note that they are all predicted embeddings using the test set. Except valence-arousal space (first row), which is already 2D, high dimensional embedding spaces are projected to a 2D space using the uniform manifold approximation and projection (UMAP) [55]. We use UMAP since it preserves more of the global structure compared to tSNE [56]. In the projection process, we first fit the UMAP with one modality (in our figure: music), then projected other embeddings (in our figure: tag and text) into the fitted 2D space.

First of all, for both the Word2Vec embedding space and the metric learning space, relevant moods from different taxonomies are neighboring together in the embedding space. This is natural for the Word2Vec space because each modality is fitted to optimize the pre-defined word embeddings. But this neighboring also can be found in metric learning space. In Figure 3-(g) and (h) for example, *anger* from text and *angry* from music are together, and *fearful* from text and *scary* from music are together. Note that Figure 3-(e) and (f) do not have word embeddings since the two-branch metric learning model does not have a branch to map the mood tags into the embedding space.

Model	Retrieval	Distribution	Mapping
Classification	fail	.	.
Multi-head classification	fail	.	.
V-A regression	success	continuous	manual
W2V regression	success	discriminative	data-driven
Metric learning (2 branch)	success	discriminative	data-driven
Metric learning (3 branch)	success	continuous	data-driven

Table 4: Characteristics of different models

One of our main motivations to use metric learning with three branches is to preserve neighborhood structure within modalities. Since Word2Vec regression is a one-way optimization, their embeddings are very discriminative (Figure 3-(c)). Also, the two-branch neural network does not have any means to learn the neighborhood structure of each modality. Especially, as shown in Table 2, when two-branch metric learning uses the mapping of Alm's mood into AudioSet mood with Word2Vec similarity, *exciting* and *tender* from music are not being used in training. If we compare Figure 3-(f) and (h), *exciting* music in (h) are more continuously distributed between *angry* and *happy* while they are simply with *happy* in (f). Also, when we compare text embeddings (see (e) and (g)), *surprised* is continuously distributed between *anger* and *happy* in (g) but not in (e). This continuity between music and text can be found in the manually annotated valence-arousal space (see (b) and (a), respectively), which means the proposed three-branch metric learning model preserves neighborhood structure within modalities in the learned multimodal embedding space. We summarize all the introduced characteristics in Table 4.

6. CONCLUSION

In this work we tackled the task of matching music to text with the goal of allowing users to enhance their text-based stories with music that matches the mood of the text. We formulated the problem as a cross-modal text-to-music retrieval problem, and identified the lack of a shared vocabulary as a key challenge for bridging the gap between modalities. To address this challenge, we proposed and investigated several emotion embedding spaces, both manually defined (valence/arousal) and data-driven (Word2Vec and metric learning), to bridge between the text and music modalities. Our experiments showed that by leveraging these embedding spaces, we were able to facilitate cross modal retrieval successfully. We showed that the carefully designed valence-arousal space can bridge different modalities, but this can be also achieved via data-driven embedding spaces. Especially, our proposed three-branch metric learning model preserves the neighborhood structure of emotions within modalities. By leveraging data-driven embeddings, our approach has the potential of being generalized to other cross-modal retrieval tasks that require broader or completely different vocabularies.
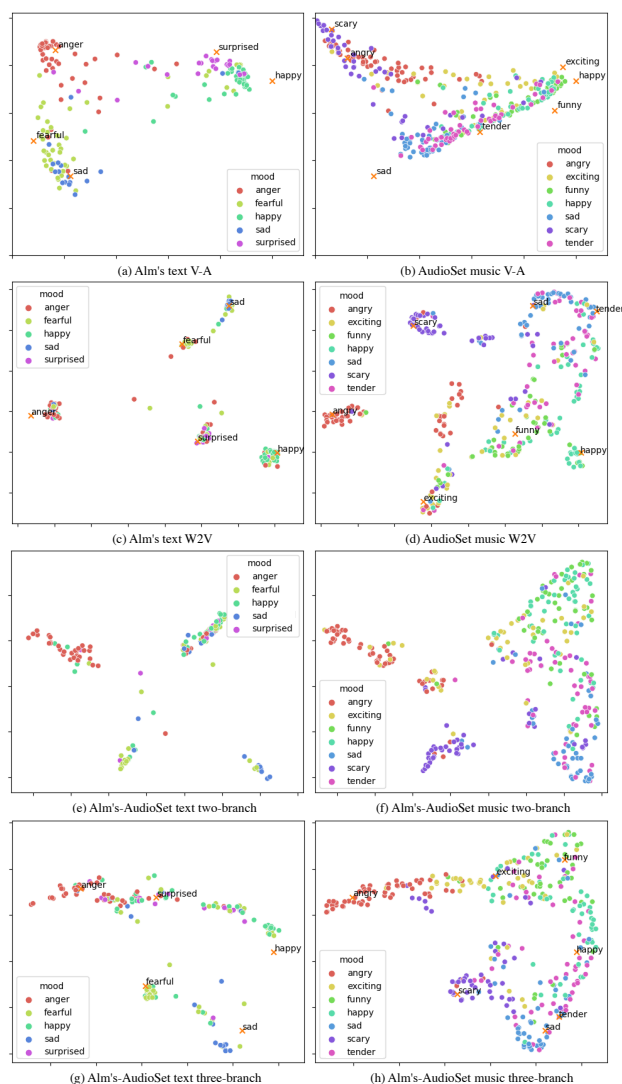

(a) Alm's text V-A

(b) AudioSet music V-A

(c) Alm's text W2V

(d) AudioSet music W2V

(e) Alm's-AudioSet text two-branch

(f) Alm's-AudioSet music two-branch

(g) Alm's-AudioSet text three-branch

(h) Alm's-AudioSet music three-branch

Figure 3: Valence-arousal embedding (first row), UMAP of Word2Vec embedding (second row), UMAP of shared embedding space from two-branch metric learning (third row), and UMAP of shared embedding space from three-branch metric learning (fourth row).

7. ACKNOWLEDGEMENT

8. REFERENCES

[1] Y. E. Kim, E. M. Schmidt, R. Migneco, B. G. Morton, P. Richardson, J. Scott, J. A. Speck, and D. Turnbull, "Music emotion recognition: A state of the art review," in *Proc. of International Society for Music Information Retrieval Conference (ISMIR)*, 2010.

[2] B. Wang, Y. Yang, X. Xu, A. Hanjalic, and H. T. Shen, "Adversarial cross-modal retrieval," in *Proc. of the 25th ACM International Conference on Multimedia*, 2017.

[3] T. Yao, T. Mei, and C.-W. Ngo, "Learning query and image similarities with ranking canonical correlation analysis," in *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

[4] L. Zhen, P. Hu, X. Wang, and D. Peng, "Deep supervised cross-modal retrieval," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[5] Y. Wei, Y. Zhao, C. Lu, S. Wei, L. Liu, Z. Zhu, and S. Yan, "Cross-modal retrieval with cnn visual features: A new baseline," *IEEE Transactions on Cybernetics*, vol. 47, no. 2, 2016.

[6] K. Wang, Q. Yin, W. Wang, S. Wu, and L. Wang, "A comprehensive survey on cross-modal retrieval," *arXiv preprint arXiv:1607.06215*, 2016.

[7] S. M. Mohammad and P. D. Turney, "Crowdsourcing a word–emotion association lexicon," *Computational Intelligence*, vol. 29, no. 3, 2013.

[8] C. Strapparava, A. Valitutti *et al.*, "Wordnet affect: an affective extension of wordnet." in *Proc. of International Conference on Language Resources and Evaluation*, 2004.

[9] T. Danisman and A. Alpkocak, "Feeler: Emotion classification of text using vector space model," in *AISB Convention: Communication, Interaction and Social Intelligence*, vol. 1, 2008.

[10] M. Hasan, E. Rundensteiner, and E. Agu, "Emotex: Detecting emotions in twitter messages," 2014.

[11] M. Abdul-Mageed and L. Ungar, "Emonet: Fine-grained emotion detection with gated recurrent neural networks," in *Proc. of annual meeting of the association for computational linguistics*, 2017.

[12] E. Batbaatar, M. Li, and K. H. Ryu, "Semantic-emotion neural network for emotion recognition from text," *IEEE Access*, vol. 7, 2019.

[13] D. Cortiz, "Exploring transformers in emotion recognition: a comparison of bert, distillbert, roberta, xlnet and electra," *arXiv preprint arXiv:2104.02041*, 2021.

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *Proc. of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, 2019.

[15] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[16] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *Neural Information Processing Systems Workshop on Energy Efficient Machine Learning and Cognitive Computing*, 2019.

[17] G. Tzanetakis, "Marsyas submissions to mirex 2007," *Music Information Retrieval Evaluation eXchange*, 2007.

[18] G. Peeters, "A generic training and classification system for mirex08 classification tasks: audio music mood, audio genre, audio artist and audio tag," in *Proc. of the International Symposium on Music Information Retrieval (ISMIR)*, 2008.

[19] C. Cao and M. Li, "Thinkit's submissions for mirex2009 audio music classification and similarity tasks," *Music Information Retrieval Evaluation eXchange*, 2009.

[20] T. Lidy, A. Schindler *et al.*, "Parallel convolutional neural networks for music genre and mood classification," *Music Information Retrieval Evaluation eXchange*, 2016.

[21] R. Delbouys, R. Hennequin, F. Piccoli, J. Royo-Letelier, and M. Moussallam, "Music mood detection based on audio and lyrics with deep neural net," *In Proc. of International Society for Music Information Retrieval Conference (ISMIR)*, 2018.

[22] K. Choi, G. Fazekas, and M. Sandler, "Automatic tagging using deep convolutional neural networks," *In Proc. of International Society for Music Information Retrieval Conference (ISMIR)*, 2016.

[23] J. Lee, J. Park, K. L. Kim, and J. Nam, "Sample-level deep convolutional neural networks for music auto-tagging using raw waveforms," *In Proc. of Sound and music computing (SMC)*, 2017.

[24] J. Pons, O. Nieto, M. Prockup, E. Schmidt, A. Ehmann, and X. Serra, "End-to-end learning for music audio tagging at scale," *In Proc. of International Society for Music Information Retrieval Conference (ISMIR)*, 2018.

[25] M. Won, S. Chun, , O. Nieto, and X. Serra, "Data-driven harmonic filters for audio representation learning," *In Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[26] J. Lee, N. J. Bryan, J. Salamon, Z. Jin, and J. Nam, "Metric learning vs classification for disentangled music representation learning," in *In Proc. of International Society for Music Information Retrieval Conference (ISMIR)*, 2020.

[27] M. Won, A. Ferraro, D. Bogdanov, and X. Serra, "Evaluation of cnn-based automatic music tagging models," *In Proc. of Sound and Music Computing (SMC)*, 2020.

[28] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[29] "Mediaeval 2020 emotion and theme recognition in music task: Loss function approaches for multi-label music tagging," *MediaEval2020*, 2020.

[30] E. M. Schmidt, D. Turnbull, and Y. E. Kim, "Feature selection for content-based, time-varying musical emotion regression," in *Proc. of the international conference on Multimedia information retrieval*, 2010, pp. 267–274.

[31] B.-j. Han, S. Rho, R. B. Dannenberg, and E. Hwang, "Smers: Music emotion recognition using support vector regression." in *In Proc. of International Society for Music Information Retrieval Conference (ISMIR)*, 2009.

[32] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[33] R. E. Thayer, *The biopsychology of mood and arousal.* Oxford University Press, 1990.

[34] S. Mohammad, "Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words," in *Proc. of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 174–184.

[35] M. Won, S. Oramas, O. Nieto, F. Gouyon, and X. Serra, "Multimodal metric learning for tag-based music retrieval," *In Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

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

[37] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proc. of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[38] J. Choi, J. Lee, J. Park, and J. Nam, "Zero-shot learning for audio-based music classification and tagging," *In Proc. of International Society for Music Information Retrieval Conference (ISMIR)*, 2019.

[39] S. Doh, J. Lee, T. H. Park, and J. Nam, "Musical word embedding: Bridging the gap between listening contexts and music," *arXiv preprint arXiv:2008.01190*, 2020.

[40] R. Arandjelovic and A. Zisserman, "Look, listen and learn," in *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 609–617.

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

[42] J. Huh, M. Lee, H. Heo, S. Mun, and J. S. Chung, "Metric learning for keyword spotting," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 133–140.

[43] B. Elizalde, S. Zarar, and B. Raj, "Cross modal audio search and retrieval with joint embeddings based on text and audio," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[44] A.-M. Oncescu, A. Koepke, J. F. Henriques, Z. Akata, and S. Albanie, "Audio retrieval with natural language queries," *arXiv e-prints*, pp. arXiv–2105, 2021.

[45] L. Wang, Y. Li, J. Huang, and S. Lazebnik, "Learning two-branch neural networks for image-text matching tasks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 394–407, 2018.

[46] L. Wang, Y. Li, and S. Lazebnik, "Learning deep structure-preserving image-text embeddings," in *Proc. of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2016.

[47] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "Transformers: State-of-the-art natural language processing," in *Proc. of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2020.

[48] E. Law, K. West, M. I. Mandel, M. Bay, and J. S. Downie, "Evaluation of algorithms using games: The case of music tagging." in *In Proc. of International Society for Music Information Retrieval Conference (ISMIR)*, 2009.

[49] C.-Y. Wu, R. Manmatha, A. J. Smola, and P. Krahenbuhl, "Sampling matters in deep embedding learning," in *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[50] E. C. O. Alm, *Affect in* text and speech.* Citeseer, 2008.

[51] K. R. Scherer and H. G. Wallbott, "Evidence for universality and cultural variation of differential emotion response patterning." *Journal of personality and social psychology*, vol. 66, no. 2, p. 310, 1994.

[52] X. Hu, J. S. Downie, and A. F. Ehmann, "Lyric text mining in music mood classification," *American music*, vol. 183, no. 5,049, pp. 2–209, 2009.

[53] D. Bogdanov, M. Won, P. Tovstogan, A. Porter, and X. Serra, "The mtg-jamendo dataset for automatic music tagging," *Machine Learning for Music Discovery Workshop, International Conference on Machine Learning*, 2019.

[54] J. F. Gemmeke, D. P. 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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[55] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.

[56] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.