Comparison and Analysis of Deep Audio Embeddings for Music Emotion Recognition

Eunjeong Koh, Shlomo Dubnov
Music Department
University of California, San Diego
La Jolla, CA 92039
{eko, sdubnov}@ucsd.edu

Abstract

Emotion is a complicated notion present in music that is hard to capture even with fine-tuned feature engineering. In this paper, we investigate the utility of state-of-the-art pre-trained deep audio embedding methods to be used in the Music Emotion Recognition (MER) task. Deep audio embedding methods allow us to efficiently capture the high dimensional features into a compact representation. We implement several multi-class classifiers with deep audio embeddings to predict emotion semantics in music. We investigate the effectiveness of L^3-Net and VGGish deep audio embedding methods for music emotion inference over four music datasets. The experiments with several classifiers on the task show that the deep audio embedding solutions can improve the performances of the previous baseline MER models. We conclude that deep audio embeddings represent musical emotion semantics for the MER task without expert human engineering.

Introduction

It is an essential step for music indexing and recommendation tasks to understand emotional information in music. Previous Music Emotion Recognition (MER) studies explore sound components that can be used to analyze emotions such as duration, pitch, velocity, and melodic interval. Those representations are high-level acoustic features based on domain knowledge (Wang, Wang, and Lanckriet 2015; Madhok, Goel, and Garg 2018; Chen et al. 2016; Lin, Chen, and Yang 2013).

Relying on human expertise to design the acoustic features for pre-processing large amounts of new data is not always feasible. Furthermore, existing emotion-related features are often fine-tuned for the target dataset based on music domain expertise and are not generalizable across different datasets (Panda, Malheiro, and Paiva 2018b).

Advancement in deep neural networks now allows us to learn useful domain-agnostic representations, known as deep audio embeddings, from raw audio input data with no human intervention. Furthermore, it has been reported that deep audio embeddings frequently outperform hand-crafted feature representations in other signal processing problems such as Sound Event Detection (SED) and video tagging task (Wilkinghoff 2020; DCASE 2019).

The power of deep audio embeddings is to automatically identify predominant aspects in the data at scale. Specifically, the Mel-based Look, Listen, and Learn network (L^3-Net) embedding method recently matched state-of-the-art performance on the SED task (Cramer et al. 2019). Using a sufficient amount of training data (around 60M training samples) and carefully designed training choices, Cramer et al. were able to detect novel sound features in each audio clip using the L^3-Net audio embeddings (Cramer et al. 2019). Cramer et al. released their optimal pre-trained L^3-Net model which can now be extended to new tasks.

In this paper, we compare and analyze the deep audio embeddings, L^3-Net and VGGish, for representing musical emotion semantics. VGGish is also a type of deep audio embedding method based on a VGG-like structure trained to predict video tags from the Youtube-8M dataset (Abu-El-Haija et al. 2016; Hershey et al. 2017; Jansen et al. 2017, 2018). We repurpose the two deep audio embeddings, originally designed for the SED task, to the task of MER. In evaluating the performance of the embedding methods over four different music emotion datasets, we use the embeddings in several classification models and evaluate their efficacy for the MER task on each dataset.

Our results show that the embedding methods provide an effective knowledge transfer mechanism between SED and MER domains without any additional training samples. More importantly, the deep audio embedding does not require expert human engineering of the sound features for the emotion prediction task. Our study reveals that audio-domain knowledge from the SED task can be extended to the MER task.

Related Work

One of the goals of the MER task is to automatically recognize the emotional information conveyed in music (Kim et al. 2010). Although there are many studies in the MER field (Soleymani et al. 2013; Yang and Chen 2012; Yang, Dong, and Li 2018), it is a complex process to compare features and performances of the studies because of the technical differences in data representation, emotion labeling, and feature selection algorithm. In addition, different studies are

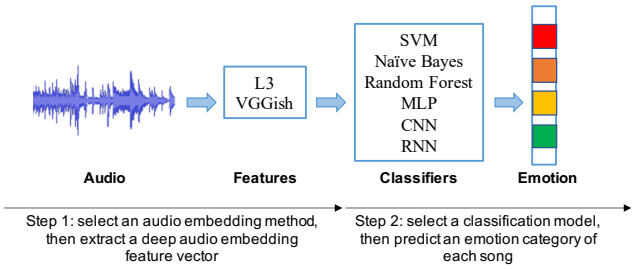

Figure 1: **The Proposed Workflow.** The figure shows the proposed approach using deep audio embeddings for the MER task.

difficult to reproduce as many of them use different public datasets or private datasets with small amounts of music clips and different levels of features.

Previous studies have utilized neural networks to efficiently extract emotional information and analyze the salient semantics of the acoustic features. Recent works explore neural networks given the significant improvements over hand-crafted feature-based methods (Piczak 2015; Salamon and Bello 2017; Pons and Serra 2019; Simonyan and Zisserman 2014). Specifically, using Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) based models, several studies attempt to extract necessary parameters for emotion prediction and reduce the dimensionality of the corresponding emotional features (Cheuk et al. 2020; Thao, Herremans, and Roig 2019; Dong et al. 2019; Liu, Fang, and Huang 2019). After careful feature engineering, these methods are suitable for a target data set for emotion prediction, however, a considerable amount of training and optimization process is still required.

Deep audio embeddings are a type of audio features extracted by a neural network that take audio data as an input and compute features of the input audio. The advantages of deep audio embedding representations are that they summarize the high dimensional spectrograms into a compact representation. Using deep audio embedding representation, 1) information can be extracted without being limited to specific kinds of data, and 2) it can save time and resources.

Several studies have used deep audio embedding methods in music classification tasks. For example, Choi et al. implemented a convnet feature-based deep audio embedding and showed how it can be used in six different music tagging tasks such as dance genre classification, genre classification, speech/music classification, emotion prediction, vocal/non-vocal classification, and audio event classification (Choi et al. 2017). Kim et al. proposed several statistical methods to understand deep audio embeddings for usage in learning tasks (Kim et al. 2019). However, there are currently no studies analyzing the use of deep audio embeddings in the MER task across multiple datasets.

Knowledge transfer is getting increased attention in the Music Information Retrieval (MIR) research as a method to enhance sound features. Recent MIR studies report considerable performance improvements in music analysis, indexing, and classification tasks by using cross-domain knowledge transfer (Hamel and Eck 2010; Van den Oord, Dieleman, and Schrauwen 2013). For automatic emotion recognition in speech data, Feng and Chaspari used a Siamese neural network for optimizing pairwise differences between source and target data (Feng and Chaspari 2020). In the context of SED, where the goal is to detect different sound events in audio streams, Cramer et al. (Cramer et al. 2019) propose a new audio analysis method, using deep audio embeddings, based on computer vision techniques. It remains to be seen if knowledge transfer can be successfully applied on deep audio embeddings from the SED domain to the MIR domain for the task of MER.

In this study, we use deep audio embedding methods designed for the SED task and apply it over four music emotion datasets for learning emotion features in music.

Methods

Downstream Task: Music Emotion Recognition

We employ a two-step experimental approach (see Figure 1).

Step 1. Given a song as an input, a deep audio embedding model extracts the deep audio embeddings that indicate the acoustic features of the song.

Step 2. After extracting deep audio embeddings, the selected classification model predicts the corresponding emotion category that indicates the emotion label of the song.

Deep Audio Embeddings

We choose two deep audio embedding methods, L^3-Net and VGGish, which are state-of-the-art audio representations pre-trained on 60M AudioSet (Gemmeke et al. 2017) and Youtube-8M data (Abu-El-Haija et al. 2016). AudioSet and Youtube-8M are large labeled training datasets that are widely used in audio and video learning with deep neural networks.

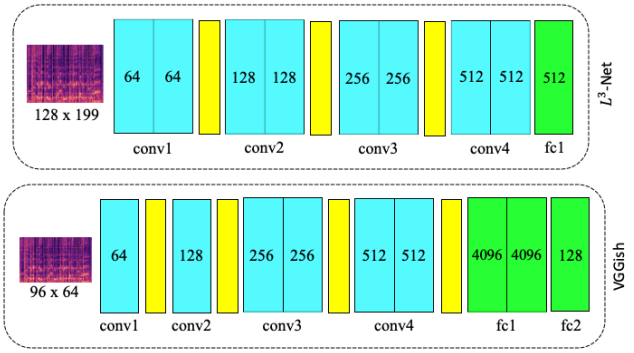

Figure 2: **Network Architecture of L^3-Net and VGGish.** The input spectrogram representations are 128x199 for L^3-Net and 96x64 for VGGish. Blue boxes, yellow boxes, and green boxes denote the 2D convolutional layers, max-pooling layers, and fully-connected layers, respectively. The number inside of the blue box is the size of filters and the number inside of the green box is the number of neurons.

Look, Listen, and Learn network (L^3-Net) L^3-Net is an audio embedding method (Cramer et al. 2019) motivated by the original work of Look, Listen, and Learn (L^3) (Arandjelovic and Zisserman 2017) that processes Audio-Visual Correspondence learning task in computer vision research. The key differences between the original L^3 (by Arandjelović and Zisserman) and L^3-Net (by Cramer et al.) are (1) input data format (video vs. audio), (2) final embedding dimensionality, and (3) training sample size.

The L^3-Net audio embedding method consists of 2D convolutional layers and 2D max-pooling layers, and each convolution layer is followed by batch normalization and a ReLU nonlinearity (see Figure 2). For the last layer, a max-pooling layer is performed to produce a single 512 dimension feature vector (L^3-Net serves as an option for output embedding size such as 6144 or 512, and we choose 512 as our embedding size). The L^3-Net method is pre-trained on Google AudioSet 60M training samples containing mostly musical performances (Gemmeke et al. 2017).

We follow the design choices of the L^3-Net study which result in the best performance in their SED task. We use Mel spectrograms with 256 Mel bins spanning the entire audible frequency range, resulting in a 512 dimension feature vector. We revise OpenL3 open-source implementation[1] for our experiments.

VGGish We also verify another deep audio embedding method, VGGish (Simonyan and Zisserman 2014), VGG-structure (VGGNet) based deep audio embedding model. VGGish is a 128-dimensional audio embedding method, motivated by VGGNet (Simonyan and Zisserman 2014), and pre-trained on a large YouTube-8M dataset (Abu-El-Haija et al. 2016). Original VGGNet is targeting large scale image classification tasks, and VGGish is targeting extracting acoustic features from audio waveforms. The VGGish audio embedding method consists of 2D convolutional layers and 2D max-pooling layers to produce a single 128 dimension feature vector (see Figure 2). We modify a VGGish open-source implementation[2] for our experiments.

Music Emotion Classifiers

From the computed deep audio embeddings, we predict an emotion category corresponding to each audio vector as a multi-class classification problem. We employ six different classification models, Support Vector Machine (SVM), Naive Bayes (NB), Random Forest (RF), Multilayer Perceptron (MLP), Convolution Neural Network (CNN), and Recurrent Neural Network (RNN).

For each classification task, we use 80% of the data for training, 10% for testing, and 10% for validation. All six classification models are implemented in Scikit-learn (Pedregosa et al. 2011), Keras (Chollet et al. 2015), and Tensorflow (Abadi et al. 2016). In the case of MLP, CNN, and RNN classification models, we share some implementation details below.

[1] OpenL3 open-source library:https://openl3.readthedocs.io/en/latest/index.html

[2] VGGish:https://github.com/tensorflow/models/tree/master/research/audioset/vggish

• MLP: We implement the MLP model with two of a single hidden layer with 512 nodes, a ReLU activation function, an output layer with a number of emotion categories, and a softmax activation function. The model is processed using the categorical cross-entropy loss function and we use Adam stochastic gradient descent (Kingma and Ba 2014). We fit the model for 1000 training epochs with the default batch size of 32 samples and evaluate the performance at the end of each training epoch on the test dataset.

• CNN: For CNN classification model, we revise the convolutional filter design proposed by Abdoli et al. (Abdoli, Cardinal, and Koerich 2019), which includes four 1D convolution layers and a 1D max-pooling operation layer. Each layer processes 64 convolutional filters. The input to the network is a Mel spectrogram, size of 512 feature vector extracted from a deep audio embedding method. This input size is varied depending on the type of embedding methods. For example, in the case of L^3-Net, the embedding size is 512, VGGish embedding size is 128. ReLU activation functions are applied to the convolutional layers to reduce the backpropagation errors and accelerate the learning process (Goodfellow, Bengio, and Courville 2016). The softmax function is used as the output activation function with a number of emotion categories. Adam optimizer, categorical cross-entropy loss function, and the batch size of 32 samples are used. The stopping criterion is set as 1000 epochs with an early-stopping rule if there is no improvement to the score during the last 100 learning epochs.

• RNN: Weninger et al. (Weninger, Eyben, and Schuller 2014) propose LSTM-RNN design as an automaton-like structure mapping from an observation sequence to an output feature sequence. We use LSTM networks with a pointwise softmax function based on a number of emotion categories. Adam optimizer, the categorical cross-entropy loss function, and the batch size of 32 samples are used. The same stopping criterion is set as CNNs.

Evaluation

Dataset

Four different datasets are selected for computing the emotional features in music data. In Table 1, we show the number of music files of each dataset by emotion category.

• 4Q Audio Emotion Dataset: This dataset is introduced by Panda et al. (Panda, Malheiro, and Paiva 2018a), annotated each music clip into four Arousal-Valence (A-V) quadrants based on Rusell's model (Russell 2003): Q1 (A+V+), Q2 (A+V-), Q3 (A-V-), Q4 (A-V+). Each emotion category has 225 music clips, and each music clip is 30 seconds long. The total music clips for the dataset are 900 files.

• Bi-modal Emotion Dataset: This dataset is introduced by Malheiro et al. (Malheiro et al. 2016) in a context of bi-modal analysis in the emotion recognition with audio and lyric information. The emotion category is also annotated into four A-V quadrants by Russell's model. In this dataset, each emotion category has a different number of music clips, Q1: 52 clips; Q2: 45 clips; Q3: 31 clips, and Q4: 34 clips, and each music clip is 30 seconds long. The total music clips for the dataset are 162 files. The size of this dataset is the

Table 1: **Dataset Details.** The number of emotion categories in each dataset and the number of clips in each emotion category are described. Q1, Q2, Q3, Q4 means the emotion categories of the four Arousal-Valence (A-V) quadrants based on Russell's model (Russell 2003): Q1 (A+V+), Q2 (A+V-), Q3 (A-V-), Q4 (A-V+). For RAVDESS singing data, it has been classified into six emotion categories, N:Neutral, C:Calm, H:Happy, S:Sad, A:Angry, F:Fearful

	Emotion Category				
Dataset	Q1	Q2	Q3	Q4	Total
4Q Audio Emotion	225	225	225	225	900
Bi-modal Emotion	52	45	31	34	162
Emotion in Music	305	87	241	111	744

	Emotion Category						
Dataset	N	C	H	S	A	F	Total
RAVDESS	92	184	184	184	184	20	848

smallest for our experiments.

• Emotion in Music: Using a crowdsourcing platform, Soleymani et al. (Soleymani et al. 2013) release a music emotion dataset with 20,000 arousal and valence annotations on 1,000 music clips. For our experiments, we map the arousal and valence annotation into four A-V quadrants followed by previous Russell's model settings. Each emotion category has a different number of music clips, Q1: 305 clips; Q2: 87 clips; Q3: 241 clips, and Q4: 111 clips, and each music clip is 45 seconds long. We use 744 music clips of the dataset in our experiments. This dataset is one of the most frequently used datasets for the MER task.

• Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): This dataset is introduced by Livingstone et al. (Livingstone and Russo 2018) for understanding the emotional context in speech and singing data. In singing data, it includes the recording clips of human singing with different emotional contexts. 24 different actors were asked to sing in six different emotional states: neutral, calm, happy, sad, angry, fearful. We choose singing data only for our experiments. Each emotion category has a different number of music clips, neutral: 92 clips; calm: 184 clips; happy: 184 clips, sad: 184 clips, angry: 184 clips, and fearful: 20 clips, and each music clip is 5 seconds long. The total music clips for the dataset are 848 files.

Baseline Audio Features

As a baseline feature, we use Mel-Frequency Cepstral Coefficients (MFCCs), which are known to be efficient low-level descriptors for timbre analysis, used as features of music tagging tasks (Choi et al. 2017; Kim, Lee, and Nam 2018). MFCCs describe the overall shape of a spectral envelope. We first calculate the time derivatives of the given MFCCs and then take the mean and standard deviation over the time axis. Finally, we concatenate all statistics into one vector. We generate the MFCC features of each music clip into a matrix of 20 x 1500. Librosa is used for MFCCs extraction and audio processing (McFee et al. 2015).

Performance Measures

For classification problems, classifier performance is typically defined according to the confusion matrix associated with the classifier. We use accuracy measure as a primary evaluation criterion. We also calculate F1-score and r^2 score for comparison with other baseline models.

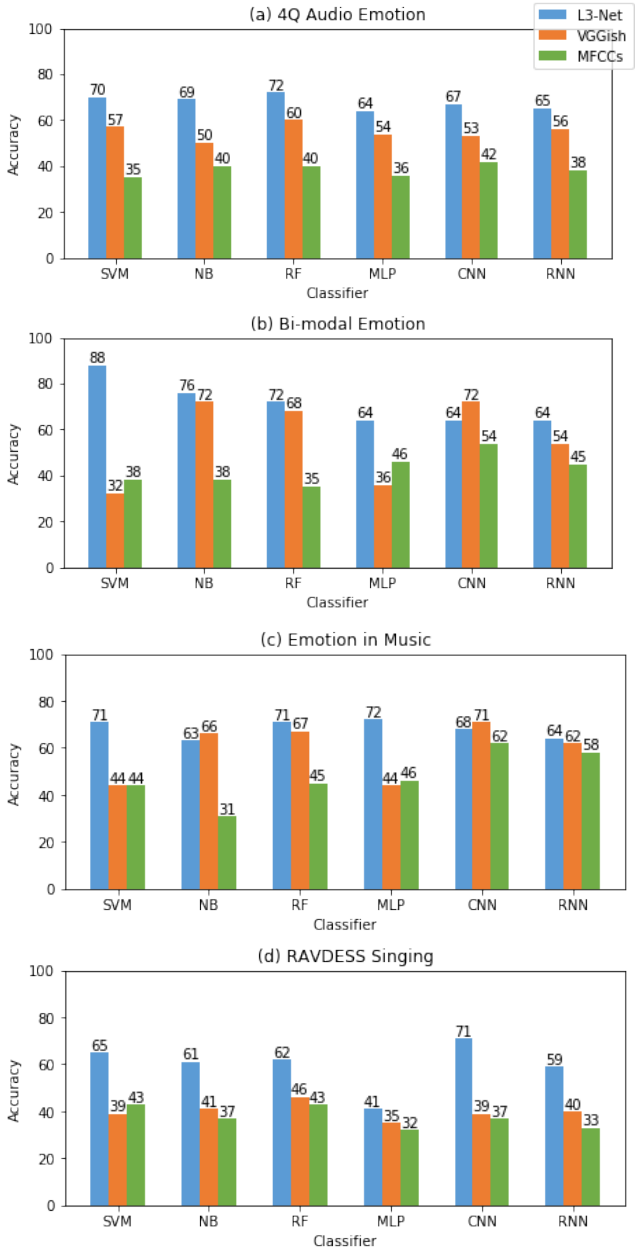

Figure 3: **Performance of Emotion Recognition on the Music Emotion Datasets.** Blue bar means the performance of L^3-Net, orange bar for VGGish, and green bar for MFCCs. X-axis indicates the type of classifiers we used, and Y-axis indicates the classification accuracies of the emotion category recognition.

Evaluation of Music Emotion Recognition

In Figure 3, we show the performance of deep audio embeddings over four music emotion datasets. We empirically analyze deep audio embeddings in several settings against baseline MFCC features. The experiments are validated with 20 repetitions of cross-validation where we report the average results. We share key observations in the next sections.

Performance Analyzed by Features The L^3-Net embedding has the best performance in all considered cases except for two, CNN classifier accuracy both in Bi-modal Emotion and Emotion in Music dataset (see Figure 3). Even though the L^3-Net embedding is generally not a descriptor for any music-related tasks before, the performance convinces us to use a pre-trained L^3-Net audio embedding model for the MER task.

Since the direct use of the L^3-Net embedding shows the better performance, we also investigate more about the different embedding dimension of the L^3-Net and compare the performance between 512 and 6144. Interestingly, we observe decreasing results with the dimension of 6144 L^3 embeddings. This indicates that those extra features might not be relevant but introducing noise. While the 512 L^3 embeddings show consistent higher performance in many cases, based on our observations, even we increase the depth and number of parameters, 6144 L^3-Net embeddings perform slightly lower on this MER task. Thus, we have not included the performance in the figure. Note that reported results in Figure 3 are only considered the performance of 512.

Comparing between L^3-Net and VGGish, L^3-Net outperforms VGGish across the dataset. This could be because L^3-Net was pre-trained on both visual and audio onto the same embedded space which can include more features. The performance of VGGish is better than MFCC baseline features with the rest of the classification models, even though it has fewer parameters, 128. This justifies our use of L^3-Net as a main deep audio embedding choice for MER task and VGGish is for some cases.

It is generally known that decision trees and in our case, RF is better than other neural network based classifiers where the data comprises a large set of categories (Pons and Serra 2019). It also deals better with dependence between variables, which might increase the error and cause some significant features to become insignificant during training. SVM uses kernel trick to solve non-linear problems whereas decision trees derive hyper-rectangles in input space to solve the problem. This is why decision trees are better for categorical data and it deals with co-linearity better than SVM. We still find that SVM outperforms RF in some cases. The reason can be that SVM deals better with margins and thus better handles outliers. Although these are tangential considerations, it seems to support the overall notion that MER is a higher level recognition problem that first needs to address the division of the data into multiple acoustic categories, also requiring the learning of a rather non-trivial partition structure within these sub-categories.

Performance Analyzed by Datasets For comparison with prior works studying emotions in audio signals, we analyze the performance of previous studies on each dataset we used. We choose four baseline MER models for our experiments: 1) Panda et al. (Panda, Malheiro, and Paiva 2018a) release the 4Q Music Emotion dataset and present the study of musical texture and expressivity features, 2) Malheiro et al. (Malheiro et al. 2016) present novel lyrical features for MER task and release the Bi-modal Emotion dataset, 3) Choi et al. (Choi et al. 2017) present a pre-trained convnet feature for music classification and regression tasks and evaluate the model using Emotion in Music dataset, 4) Arora and Chaspari (Arora and Chaspari 2018) present the method of a siamese network for speech emotion classification and evaluate the method using RAVDESS dataset. We compare those baseline models to the performance of our proposed method (see Table 2).

Table 2: **Performance Comparison with Baseline MER models.** This table shows the data and feature information used in previous baseline models. Data column indicates each dataset for the experiment. Feature column indicates the set of feature vectors extracted by the baseline model. Metric column indicates the metric used for the performance analysis. Baseline column includes the performance of the baseline models. Proposed L^3-Net column includes the best performance of L^3-Net embeddings on each music dataset.

DATA	FEATURE	METRIC	BASELINE	PROPOSED L^3-NET
4Q AUDIO	DOMAIN KNOWLEDGE	*F*	73.5%	72.0%
BI-MODAL	DOMAIN KNOWLEDGE	*F*	72.6%	**88.0%**
EMOMUSIC	CONVNET	r^2	A: 0.656 V: 0.462	A: **0.671** V: **0.556**
RAVDESS	SIAMESE	*Acc*	63.8%	**71.0%**

In the case of the 4Q Audio Emotion dataset, the previous study by Panda et al. obtained its best result of 73.5% F1-score with a high number of 800 features. In Table 3, Domain Knowledge means a feature set defined by domain knowledge in the study. For achieving the performance of the previous study, the following steps are needed. First, we need to pre-process standard or baseline audio features of each audio clip. The study used Marsyas, MIR Toolbox, and PsySound3 audio frameworks to extract a total of 1702 features. Second, we need to calculate the correlation between the pair of features for normalization. After the preprocessing, the number of features can be decreased to 898 features. Third, after computing these baseline audio features, we also need to compute novel features of each audio clip proposed by the study. Those features were carefully designed based on domain expertise, such as glissando features, vibrato, and tremolo features. Finally, baseline features and extracted novel features are combined for the MER task. For the evaluation, the study conducted post-processing of the features with the ReliefF feature selection algorithm (Robnik-Šikonja and Kononenko 2003), ranked the features and evaluated its best-suited features. Since the performance

has been evaluated by hyperparameter tuning and feature selection algorithms, these factors may influence the performance of the MER task significantly. Note that in our proposed approach, we show the performance without any post-processing.

In the case of the Bi-modal Emotion dataset, the previous study by Malheiro et al. (Malheiro et al. 2016) presented its best classification result of 72.6% F1-score on the dataset which is lower than the performance we have, 88% F1-score from the result of L^3-Net embedding with SVM classifier.

In the case of the Emotion in Music dataset, previous studies predicted the time-varying arousal and valence annotation and calculated r^2 score as a performance measure (Weninger, Eyben, and Schuller 2014; Lee et al. 2019; Kim, Lee, and Nam 2018; Choi et al. 2017). We previously map these time-varying annotations into four A-V quadrants based on Rusell's model and show our prediction performance with four emotion categories (see Figure 3-(c)). For a fair comparison, we also verify the original time-varying dynamic annotations from the dataset (Soleymani et al. 2013) and compare the result with the baseline model. Using the Emotion in Music dataset, Choi et al. reported its r^2 scores of arousal annotation, 0.656 and valence annotation, 0.462 (Choi et al. 2017). The best performance of L^3-Net embeddings achieves 0.671 r^2 score on arousal and 0.556 r^2 score on valence annotation. The result shows that we have a considerable and higher performance on arousal and valence annotation. The result confirms that L^3-Net embedding method shows favorable performance than the previous embedding features over Emotion in Music data.

In the case of RAVDESS data, the study by Arora and Chaspari (Arora and Chaspari 2018) reported its best classification accuracy of 63.8% over the dataset which is lower than our accuracy, 71.0%, from the result of L^3-Net embedding with CNN classifier (see Figure 3-(d)).

Table 3: **Classification Results of Each Quadrant.** The top table indicates the classification report of L^3-Net embedding with Random Forest classifier on 4Q Audio Emotion Dataset. The bottom table indicates the classification report of L^3-Net embedding with SVM classifier on Bi-modal Emotion Dataset.

4Q Audio Emotion	Precision	Recall	F1-score
Q1	0.64	0.85	0.73
Q2	0.85	0.80	0.83
Q3	0.73	0.60	0.66
Q4	0.64	0.61	0.62
Accuracy			0.72
Weighted avg	0.73	0.72	0.72

Bi-modal Emotion	Precision	Recall	F1-score
Q1	0.80	1.00	0.89
Q2	1.00	0.89	0.94
Q3	1.00	0.67	0.80
Q4	0.80	0.80	0.80
Accuracy			0.88
Weighted avg	0.90	0.88	0.88

Performance Analyzed by A-V Quadrants In Table 3, we show the results analyzed by each quadrant. This classification report gives us a further understanding of the characteristic of each emotion category in music. The meaning of each quadrant (Q1, Q2, Q3, Q4) information is described in Table 1.

In the case of the 4Q Audio Emotion dataset, Q2 and Q3 categories obtain a higher score compared to the Q1 and Q4. This indicates that emotional features in music clips with lower valence components are easier to recognize. Specifically, the Q2 category shows higher performance which is distinctive than others. Based on the dataset (Panda, Malheiro, and Paiva 2018b), the study describes music clips of the Q2 category belong to specific genres, such as heavy metal, which have recognizable acoustic features than others.

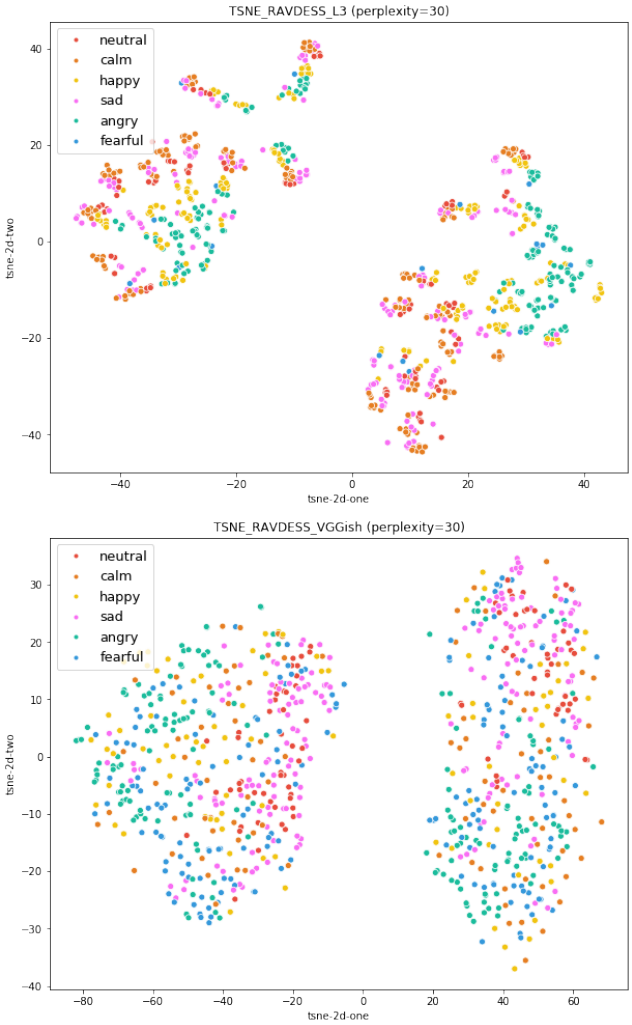

Figure 4: **T-SNE Visualization on RAVDESS dataset.** Different colors of dots indicate the type of emotion in the dataset. Both visualizations use a perplexity value of 30. Top: T-SNE Visualization of L^3-Net embeddings, bottom: T-SNE Visualization of VGGish embeddings.

Lower results in Q1 and Q4 categories may also reflect the characteristics of music clips. For instance, the Q1 category indicates happy emotions, which are typically energetic based on positive arousal and positive valence components. Since Q1 and Q4 categories share the same valence axis based on Rusell's model, if the intensity of the song is not intense, the difference between the two quadrants (Q1&Q4 or Q2&Q3) may not be apparent. This aspect results in similar behaviors on the Q2 and Q3 categories' performances as well.

Discussion and Conclusion

In this paper, we evaluate L^3-Net and VGGish pre-trained deep audio embedding methods for MER task over 4Q Audio Emotion, Bi-modal Emotion, Emotion in Music, and RAVDESS datasets. Even though L^3-Net has not been intended for emotion recognition, we find that L^3-Net is the best representation for the MER task. Note that we achieve this performance without any additional domain knowledge feature selection method, feature training process, and fine-tuning process. Comparing to MFCC baseline features, the empirical analysis shows that L^3-Net is robust across multiple datasets with favorable performance. Overall, the result using L^3-Net shows improvement compared to baseline models for Bi-modal Emotion, Emotion in Music, and RAVDESS dataset. In the case of the 4Q Audio Emotion dataset, complex hand-crafted features (over 100 features) still seem to perform better. Specifically, our work does not consider rhythm or specific musical parameters over the time axis that 4Q Audio Emotion had, looking into time-based aspects could be the next step for future research.

In order to gain deeper insight into the meaning of acoustic features for emotional recognition, we use T-SNE visualization (see Figure 4). In both cases of L^3-Net and VGGish, two main clusters on the left and right side of the figure mean male/female singer groups. We can also see a relatively smooth grouping of samples by emotions with different colors. In the case of L^3-Net embeddings (top figure of Figure 4), multiple small groups in each cluster indicate individual singer which has audio recordings in different emotions. L^3-Net data seems to cluster into multiple smaller groups according to gender and individual categories, and this shows L^3-Net outperforms for detecting different timbre information than VGGish. This pattern seems to be consistent in the wild range of T-SNE perplexity parameters. This also shows that our study provides an empirical justification that L^3-Net outperforms VGGish, with the intuition discussed in the paper based on the clustering shown in Figure 4.

Accordingly, for the next step, a possible direction to validate different classifiers is to explore a combination of discrete neural learning methods, such as VQ-VAE, to first solve the categorical problem, and only later learn a more smooth decision surface. VQ-VAE has been recently explored for spectrogram-based music inpainting (Bazin et al. 2020). It would be interesting to explore similar high-level parameterization using L^3-Net embeddings.

References

Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M.; et al. 2016. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, 265–283.

Abdoli, S.; Cardinal, P.; and Koerich, A. L. 2019. End-to-end environmental sound classification using a 1d convolutional neural network. *Expert Systems with Applications* 136: 252–263.

Abu-El-Haija, S.; Kothari, N.; Lee, J.; Natsev, P.; Toderici, G.; Varadarajan, B.; and Vijayanarasimhan, S. 2016. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675* .

Arandjelovic, R.; and Zisserman, A. 2017. Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision*, 609–617.

Arora, P.; and Chaspari, T. 2018. Exploring siamese neural network architectures for preserving speaker identity in speech emotion classification. In *Proceedings of the 4th International Workshop on Multimodal Analyses Enabling Artificial Agents in Human-Machine Interaction*, 15–18.

Bazin, T.; Hadjeres, G.; Esling, P.; and Malt, M. 2020. Spectrogram Inpainting for Interactive Generation of Instrument Sounds. URL https://boblsturm.github.io/aimusic2020/papers/CSMC__MuMe_2020_paper_49.pdf.

Chen, P.; Zhao, L.; Xin, Z.; Qiang, Y.; Zhang, M.; and Li, T. 2016. A scheme of MIDI music emotion classification based on fuzzy theme extraction and neural network. In *2016 12th International Conference on Computational Intelligence and Security (CIS)*, 323–326. IEEE.

Cheuk, K. W.; Luo, Y.-J.; Balamurali, B.; Roig, G.; and Herremans, D. 2020. Regression-based music emotion prediction using triplet neural networks. In *2020 International Joint Conference on Neural Networks (IJCNN)*, 1–7. IEEE.

Choi, K.; Fazekas, G.; Sandler, M.; and Cho, K. 2017. Transfer learning for music classification and regression tasks. *arXiv preprint arXiv:1703.09179* .

Chollet, F.; et al. 2015. keras.

Cramer, J.; Wu, H.-H.; Salamon, J.; and Bello, J. P. 2019. Look, Listen, and Learn More: Design Choices for Deep Audio Embeddings. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 3852–3856. IEEE.

DCASE. 2019. Detection and Classification of Acoustic Scenes and Events. Task 4: Sound Event Detection in Domestic Environments. URL http://dcase.community/challenge2019/task-sound-event-detection-in-domestic-environments.

Dong, Y.; Yang, X.; Zhao, X.; and Li, J. 2019. Bidirectional Convolutional Recurrent Sparse Network (BCRSN): An Efficient Model for Music Emotion Recognition. *IEEE Transactions on Multimedia* 21(12): 3150–3163.

Feng, K.; and Chaspari, T. 2020. A Siamese Neural Network with Modified Distance Loss For Transfer Learning in Speech Emotion Recognition. *arXiv preprint arXiv:2006.03001* .

Gemmeke, J. F.; Ellis, D. P.; Freedman, D.; Jansen, A.; Lawrence, W.; Moore, R. C.; Plakal, M.; and Ritter, M. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 776–780. IEEE.

Goodfellow, I.; Bengio, Y.; and Courville, A. 2016. *Deep learning*. MIT press.

Hamel, P.; and Eck, D. 2010. Learning features from music audio with deep belief networks. In *ISMIR*, volume 10, 339–344. Utrecht, The Netherlands.

Hershey, S.; Chaudhuri, S.; Ellis, D. P.; Gemmeke, J. F.; Jansen, A.; Moore, R. C.; Plakal, M.; Platt, D.; Saurous, R. A.; Seybold, B.; et al. 2017. CNN architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, 131–135. IEEE.

Jansen, A.; Gemmeke, J. F.; Ellis, D. P.; Liu, X.; Lawrence, W.; and Freedman, D. 2017. Large-scale audio event discovery in one million youtube videos. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 786–790. IEEE.

Jansen, A.; Plakal, M.; Pandya, R.; Ellis, D. P.; Hershey, S.; Liu, J.; Moore, R. C.; and Saurous, R. A. 2018. Unsupervised learning of semantic audio representations. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 126–130. IEEE.

Kim, J.; Urbano, J.; Liem, C.; and Hanjalic, A. 2019. Are Nearby Neighbors Relatives?: Are Nearby Neighbors Relatives?: Testing Deep Music Embeddings. *Frontiers in Applied Mathematics and Statistics* 5: 53.

Kim, T.; Lee, J.; and Nam, J. 2018. Sample-level CNN architectures for music auto-tagging using raw waveforms. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 366–370. IEEE.

Kim, Y. E.; Schmidt, E. M.; Migneco, R.; Morton, B. G.; Richardson, P.; Scott, J.; Speck, J. A.; and Turnbull, D. 2010. Music emotion recognition: A state of the art review. In *Proc. ISMIR*, volume 86, 937–952.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

Lee, D.; Lee, J.; Park, J.; and Lee, K. 2019. Enhancing music features by knowledge transfer from user-item log data. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 386–390. IEEE.

Lin, Y.; Chen, X.; and Yang, D. 2013. Exploration of Music Emotion Recognition Based on MIDI. In *ISMIR*, 221–226.

Liu, H.; Fang, Y.; and Huang, Q. 2019. Music emotion recognition using a variant of recurrent neural network. In *2018 International Conference on Mathematics, Modeling, Simulation and Statistics Application (MMSSA 2018)*. Atlantis Press.

Livingstone, S. R.; and Russo, F. A. 2018. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PloS one* 13(5): e0196391.

Madhok, R.; Goel, S.; and Garg, S. 2018. SentiMozart: Music Generation based on Emotions. In *ICAART (2)*, 501–506.

Malheiro, R.; Panda, R.; Gomes, P.; and Paiva, R. 2016. Bi-modal music emotion recognition: Novel lyrical features and dataset. 9th International Workshop on Music and Machine Learning–MML'2016–in ....

McFee, B.; Raffel, C.; Liang, D.; Ellis, D. P.; McVicar, M.; Battenberg, E.; and Nieto, O. 2015. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, volume 8.

Panda, R.; Malheiro, R.; and Paiva, R. P. 2018a. Musical Texture and Expressivity Features for Music Emotion Recognition. In *ISMIR*, 383–391.

Panda, R.; Malheiro, R. M.; and Paiva, R. P. 2018b. Novel audio features for music emotion recognition. *IEEE Transactions on Affective Computing* .

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* 12: 2825–2830.

Piczak, K. J. 2015. Environmental sound classification with convolutional neural networks. In *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*, 1–6. IEEE.

Pons, J.; and Serra, X. 2019. Randomly weighted CNNs for (music) audio classification. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 336–340. IEEE.

Robnik-Šikonja, M.; and Kononenko, I. 2003. Theoretical and empirical analysis of ReliefF and RReliefF. *Machine learning* 53(1-2): 23–69.

Russell, J. A. 2003. Core affect and the psychological construction of emotion. *Psychological review* 110(1): 145.

Salamon, J.; and Bello, J. P. 2017. Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Processing Letters* 24(3): 279–283.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* .

Soleymani, M.; Caro, M. N.; Schmidt, E. M.; Sha, C.-Y.; and Yang, Y.-H. 2013. 1000 songs for emotional analysis of music. In *Proceedings of the 2nd ACM international workshop on Crowdsourcing for multimedia*, 1–6. ACM.

Thao, H. T. P.; Herremans, D.; and Roig, G. 2019. Multimodal Deep Models for Predicting Affective Responses Evoked by Movies. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 1618–1627. IEEE.

Van den Oord, A.; Dieleman, S.; and Schrauwen, B. 2013. Deep content-based music recommendation. In *Advances in neural information processing systems*, 2643–2651.

Wang, J.-C.; Wang, H.-M.; and Lanckriet, G. 2015. A histogram density modeling approach to music emotion recognition. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 698–702. IEEE.

Weninger, F.; Eyben, F.; and Schuller, B. 2014. On-line continuous-time music mood regression with deep recurrent neural networks. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5412–5416. IEEE.

Wilkinghoff, K. 2020. On open-set classification with L3-Net embeddings for machine listening applications. In *28th European Signal Processing Conference (EUSIPCO)*.

Yang, X.; Dong, Y.; and Li, J. 2018. Review of data features-based music emotion recognition methods. *Multimedia Systems* 24(4): 365–389.

Yang, Y.-H.; and Chen, H. H. 2012. Machine recognition of music emotion: A review. *ACM Transactions on Intelligent Systems and Technology (TIST)* 3(3): 1–30.