HETEROGENEOUS GRAPH NEURAL NETWORK FOR MUSIC EMOTION RECOGNITION

Angelo Cesar Mendes da Silva
Universidade de São Paulo, Brazil
angelo.mendes@usp.br

Diego Furtado Silva
Universidade Federal de São Carlos, Brazil
diegofs@ufscar.br

Ricardo Marcondes Marcacini
Universidade de São Paulo, Brazil
ricardo.marcacini@usp.br

ABSTRACT

Music emotion recognition has been a growing field of research motivated by the wealth of information that these labels express. Recognition of emotions highlights music's social and psychological functions, extending traditional applications such as style recognition or content similarity. Once musical data are intrinsically multi-modal, exploring this characteristic is usually beneficial. However, building a structure that incorporates different modalities in a unique space to represent the songs is challenging. Integrating information from related instances by learning heterogeneous graph-based representations has achieved state-of-the-art results in multiple tasks. This paper proposes structuring musical features over a heterogeneous network and learning a multi-modal representation using Graph Convolutional Networks with features extracted from audio and lyrics as inputs to handle the music emotion recognition tasks. We show that the proposed learning approach resulted in a representation with greater power to discriminate emotion labels. Moreover, our heterogeneous graph neural network classifier outperforms related works for music emotion recognition.

1. INTRODUCTION

Music is intrinsically connected to human emotions. At the same time that a songwriter expresses emotions in the composed songs, we can use music's emotional information to characterize the listener's moments and feelings. Music emotion recognition (MER) is a task that highlights social and psychological functions comprised by recordings, extending traditional applications in the area of music information retrieval [1]. According to the literature, mapping the spectrum of emotions can be done by analyzing the values in arousal and valence domains [2]. Besides, the emotions can be represented by labels that indicate negative, positive, and neutral values of arousal and valence and labels that point to emotional expressions, such as happiness, anger, or satisfaction [3]. Each label may be defined according to the significance of valence and arousal simultaneously, as illustrated in Figure 1. The MER task aims to use a musical representation to estimate these significances or predict discretized labels.

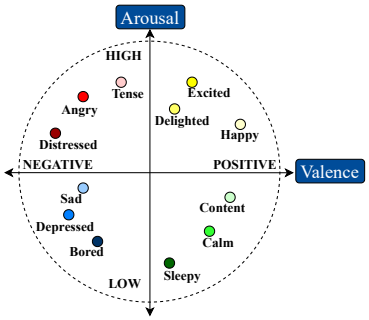

Figure 1: Emotions mapped in the arousal and valence domains.

Textual information is the most used data type in tasks involving emotion labels [4]. When emotion recognition is performed on music data, lyrics are commonly used as input [5, 6]. However, some studies emphasize the importance of acoustic features for the same task [7]. The fact that both modalities complement musical perception is well-known [8–10]. However, there is no consensual approach for incorporating multiple modalities into a unified representation.

Defining the structure to deal with different data types, such as text and audio, is a preliminary step to generate a unified multi-modal representation of musical data. Building this structure is an open problem in the literature. Techniques based on feature fusion are the most explored approach in the literature [11, 12]. Fusion of features can be performed by simply concatenating text and audio features or even learning embeddings via multimodal deep learning. However, these existing methods have limitations regarding incomplete modalities, deal with modalities of different dimensionality, and the interpretability of the generated representations [13].

Heterogeneous networks are a well-known representation for manipulating multiple modalities of unstructured data [14, 15]. This resource has been explored due to the ability to map data on graphs, representing connections between objects (nodes) through edges. Therefore, this process produces a graph that can be used as input in Graph

Neural Network models [16]. Structuring data on heterogeneous networks has been widely studied in varied data such as text and images [17,18], but there are still few studies for musical data.

This work introduces a novel model for music emotion recognition that uses a heterogeneous network to structure audio and lyrics features and build a new multi-modal graph-based representation of musical data using the graph convolutional network (GCN). Our proposed model, described in detail in Sec. 3, uses content-based features (audio descriptors) together with features extracted from lyrics (embeddings from the language model). The music graph topology is based on relations defined by cluster information from audio e textual metadata. Some music nodes have labels with emotional information. Therefore, the proposed structure is composed of a heterogeneous network, and GCN can predict the emotion of unlabeled music nodes.

We apply a network regularization framework to propagate and refine information between neighboring nodes from different modalities. The matrix resulting from this regularization, along with the graph connections, is used as input for training our GCN. In this embedding space, music with similar emotions are close to each other, while dissimilar ones are further apart.

To evaluate our approach, we used the publicly available dataset PMEmo [19], which has 794 songs represented by audio descriptors and lyrics with annotations for arousal and valence domains. We measured the relevance of the learning process of a multi-modal graph-based representation for MER. Our proposal was compared with two different methods and other works from the literature that reproduced a similar experiment.

Our method was able to obtain a meaningful embedding representation that unifies different audio and text music features across the heterogeneous network. In addition to the heterogeneous network enabling interpretability of the relationships between text and audio features, our method outperforms other existing methods for music emotion recognition tasks.

2. RELATED WORK

The construction of methods to learn representations for unstructured data aims to reduce the work for features extraction and selection to describe the data and make learning algorithms less dependent on these processes [20]. Learning a representation for musical data involves manipulating features obtained by multiple modalities and projected in different spaces [21, 22].

Learning to represent music to recognize the comprised emotion involves exploring different features [1, 23]. We can observe most applications using the lyrics to represent the music [5, 6, 24] justified by the direct association we can make with the meaning of words and emotion labels. However, we have evidence that acoustic features also contain information that characterizes emotions [11,12,25]. In summary, there is no consensus on representing music using multiple modalities for the MER task, although we note that audio and lyrics features are relevant.

Recent work about music representation learning has explored strategies based on deep neural networks, such as early and late fusion [26, 27] and evaluating combinations between different input features [13]. This strategy advances and presents successful results that justify the representation learning process. However, there is a demand for approaches that explain the semantic complementarity existing between the features that justify the obtained results [28]. We explore the heterogeneous network resources to embed multiples music features and explicit the latent relationships among nodes, similar to [29, 30]. Moreover, we learn a graph-based representation that concentrates information from related music.

The complementarity between distinct features in musical perception justifies the motivation to build a multimodal representation [31]. For the emotion recognition problem, [7] describes what semantic information exists in various acoustic features and which emotions each feature can emphasize. We note that lyrics are also associated with emotions and must be used for the recognition task [32]. Therefore, we aim to build a multi-modal representation that integrates features obtained through the audio and lyrics and compares its performance against unimodal scenarios.

The construction of multi-modal embeddings requires that distinct features be mapped in a unified space. Graph-based methods have been applied in recent years when nodes and edges have different types [33]. We can exploit the relationships between neighboring nodes to build an embedding space that aggregates information from the node features even in different spaces. Significant advances are observed in multi-modal and unstructured data applications [34].

For musical data, [35] explores the task of similarity between artists. Authors use data previously structured to create the graph and apply it in GCN to build music embeddings with acoustic characteristics and tags. Finally, the embeddings are used to predict links to new artists. In contrast, [36] introduces a proposal of graph-based representation learning with the graph being constructed from a co-occurrence matrix obtained by the presence of pairs of songs in playlists.

Our proposal aims to create a heterogeneous network with multiple musical features and connect nodes that share cluster information. This network produces the features input for GCN to build a multi-modal representation and handle the music emotion recognition task. Our learned representation does not depend on previous information to structure the graph. Although we do not extend the approach to the multi-task scenario, it can be easily adapted for different tasks, maintaining the same heterogeneous network.

3. MODELLING

Our goal is to recognize emotions present in songs, such as a classification problem $Y = f(X)$, where Y represents the set of emotion labels, X represents the songs, and $f(.)$

denotes the function responsible for predicting emotions from a multi-modal musical representation given as input.

The proposed method for handling the MER is divided into three steps. The first step is to construct a heterogeneous network to structure songs in nodes and organize the acoustic and lyrics features available in the PMEmo dataset into different layers and build the relationships. Then, in sequence, we regulate the network to propagate information between nodes of different layers and explore the semantic complementarity between features of both modalities. Finally, we use the regularized network as input our multi-modal network embedding method using Graph Convolutional Network (GCN). The generated embeddings are also used by GCN to classify the emotion of the songs.

3.1 Heterogeneous network to structure music data

Musical data are intrinsically multi-modal, formed by features semantically complementary. Therefore, building a representation that incorporates multiple features demands manipulating modalities organized in different spaces. In this sense, heterogeneous networks offer resources for structuring data as nodes, with each modality projected on a layer and forming relationships between nodes for information sharing.

Mapping unstructured data on graphs has been gaining attention in the literature, motivated by the success of GNN-based learning models. For musical data, [37, 38] proposes to build a graph relating nodes from a distance function where nearby nodes are connected, while [39] assumes that all nodes are connected by adding weights in the edges, and as the works aforementioned [35, 36] that employing existing information to create a graph structure. We propose constructing the graph topology by using clusters for each modality as network nodes. Thus, we model the relationships between music, audio and texts from the cluster labels for each modality. Figure 2 illustrates this process, where songs are initially clustered according to their audio features. In this work, we use the k-means algorithm; however, the modeling is not dependent on a specific clustering algorithm. Then both clusters and songs are considered network nodes. The links indicate association between songs and clusters formed with audio features. This process is repeated for textual features, thereby allowing to obtain a heterogeneous network with three layers as shown in Figure 3.

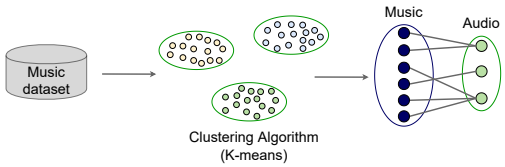

Figure 2: Illustration of mapping songs and respective audio features in a bipartite graph using clustering. This process is repeated for text features to generate the proposed heterogeneous network.

The heterogeneous network is formally defined as $N = (O, R, W)$, where O represents the set of objects, R represents the relationships and W represents the weights. Let $o_i \in O$ as the notation for an object, $r\_o\_i, o\_j = (o_i, o_j) \in R$ indicates whether there is a connection between the objects o_i and o_j, where the weight of r_{o_i,o_j} is given by w_{o_i,o_j} with $w \in W$. Currently, our propose use binaries relationships, $w_{o_i,o_j} \in \{0, 1\}$, for example, if a music node o_i is associated to a cluster with textual information, or audio information, o_j.

The benchmark dataset used has textual and acoustic features so that the set of objects $O = \{O_M \cup O_A \cup O_T\}$ is organizing each feature modality in the heterogeneous network proposed for musical data. O_M are objects representing each music in the dataset, O_A are objects representing acoustic features formed by audio descriptors, and O_T are objects representing textual features extracted from the lyrics.

3.2 Network regularization

In the proposed heterogeneous network, the objects $o \in O$ have one or more initial features vectors. For example, an object $o_m \in O_M$ that represents a music might contain an acoustic feature vector $x_{o_m}^{(A)} \in \mathbb{R}^{d_A}$, as well as a textual feature vector $x_{o_m}^{(T)} \in \mathbb{R}^{d_T}$. The vectors are in their respective spaces $\mathbb{R}^{d_A}$ (e.g. melspectrogram or chromagram) and $\mathbb{R}^{d_T}$ (e.g. word2vec or BERT). However, objects in sets O_A and O_T exclusively have the initial features of their modalities. Still, we highlight that objects in O_M may contain absent modalities, such as missing lyrics for some musics. Thus, inspired by the concept of embedding propagation from networks [40], we integrate a regularization framework capable of propagating information between objects of different modalities according to the network topology.

The regularization process exploits the network topology to propagate information between objects to complement missing information and adjust existing features according to the object label. The process is represented in Figure 3. The music is the input data, where each feature modality composes a layer in the heterogeneous network. Objects in the central layer are in O_M and are connected to objects of other types, having their feature vectors. The objects in O_A contain audio descriptors, and objects in O_T contain lyrics features. When finish regularization process, all objects will have feature vectors from objects of different modalities.

The proposed method to perform the network regularization is an instance of label propagation methods [41]. This step aims to propagate the object's features, assuming two constraints: neighboring objects must have similar features; the final feature vector must be similar to the initial features for objects. Formally, network regularization can be associated with a representation learning problem. We define as learning a mapping function $f : o_i \rightarrow \mathbf{z}_{o_i} \in \mathbb{R}^d$, where $\mathbf{z}_{o_i}$ is the learned vector of object $o_i \in O$ in network $N(O, R, W)$. The equation 1 defines the function to be minimized to learn the new space $\mathbf{Z} \in \mathbb{R}^d$, in which all

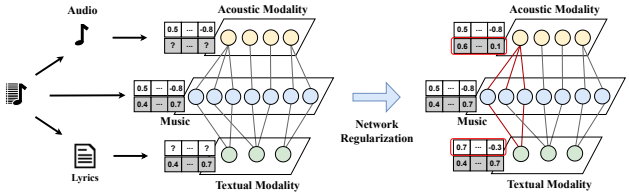

Figure 3: The heterogeneous network regularization aims to fill in missing modalities in objects. Objects on the central layer have relationships with objects on other layers. After regularization, all nodes receive information from neighbors nodes (highlights in red) with different modalities from the network topology.

heterogeneous network objects are mapped.

$$Q(\mathbf{Z}) = \frac{1}{2} \sum_{o_m \in O_M} \sum_{o_a \in O_A} w_{o_m,o_a} (\mathbf{z}_{o_m} - \mathbf{z}_{o_a})^2 + \frac{1}{2} \sum_{o_m \in O_M} \sum_{o_t \in O_T} w_{o_m,o_t} (\mathbf{z}_{o_m} - \mathbf{z}_{o_t})^2 + \mu \sum_{o_i \in O_X} (\mathbf{z}_{o_i} - \mathbf{x}_{o_i})^2 \quad (1)$$

The two initials terms of the regularization function are responsible for computing the proximity between the feature vectors for each pair of related objects in the network, indicating the weight of the relationship between the objects. The last term is responsible for computing the distance between the features of objects that had initial features, O_X, and the features learned in the space $\mathbf{Z}$. The parameter μ determines the level of preservation in the initial features. The high value indicates the more preservation of the features, while low values allow adjusting the features according to the network topology. The Equation (1) is applied for each modality, textual and acoustic. Therefore, at the end of the regularization process, we obtained Z_{audio} and Z_{text} for all nodes in the network. Finally, we concatenate both spaces, $Z_{\oplus} = Z_{audio} \oplus Z_{text}$.

3.3 Graph Convolutional Networks

The general idea of graph learning methods is to iteratively update object representations, combining them with representations of their neighbors. In this context, initially, the objects are represented by $Z_{\oplus}$ resulting from the regularization. The neighboring nodes are obtained from the heterogeneous network N, indicated by an adjacency matrix A. In particular, we used a GCN to learn the final representation and handle the MER task. The representation obtained by the GCN is represented by the matrix $F \in R^{T \times D}$, where T is the number of nodes and D is the dimension of the new unified space learned. In general, a GCN can be described according to Equation 2,

$$H^{(l+1)} = f(H^{(l)}, A) = \alpha(AH^{(l)}W^{(l)}) \quad (2)$$

where $H^{(0)} = X$, l, represents the current layer, A represents the adjacency matrix, $W^{(l)}$ represents the weight in l-th layer in a neural network, and $\alpha(.,.)$ defines the activation function. The layer $H^{(l+1)} = H^{(L)} = U$ contains the new learned space by GCN.

We pre-process the adjacency matrix A in the GCN. First, we need to add self-loops on each node to consider its features in the learning process. Then, we add A to identity matrix I, which result in $\hat{A}$. In addition, we normalize A to maintain the representation vectors scale in each layer. We did the matrix normalization by multiplying A with D^{-1}, where D indicates a diagonal matrix formed by the degrees of each node in the network, or A by $D^{-\frac{1}{2}}$, for symmetric normalization. Thus, the adjacency matrix used in the GCN is defined by Equation 3,

$$\hat{A} = D^{-\frac{1}{2}} S D^{-\frac{1}{2}} \quad (3)$$

where $S = A + I$, I represents the identity matrix, and $D_{ii} = \sum_j S_{ij}$ indicates the node degree.

The proposed GCN architecture comprises three graph convolutional layers combined with a hyperbolic tangent activation function. The dimensionality of vector input for each layer is indicated by the previous layer's output. The first layer receives the learned matrix $Z_{\oplus}$ to produce vectors with 512, 256, and 128 dimensions in next layers. Furthermore, we computed a softmax cross-entropy as the last layer to indicate the probability that each music is associated with emotion labels. Finally, we train the networks using the ADAM optimizer with a learning rate of 0.0001 and 3000 epochs.

4. EXPERIMENTS

Our experiments aim to evaluate the construction of a heterogeneous network to structure musical data and learn and evaluate a multi-modal graph-based representation in the music emotion recognition task. The k parameter defines the number of clusters used to connect objects in the heterogeneous network. This parameter is the main target of the analysis during network construction. The number of clusters allows us to create new connections between objects that initially have different labels, exploit regularization to propagate information and refine their representations.

We evaluated the representation learned with the GCN by comparing using two classifiers that utilize as input the representations formed by features provided in the dataset, including combinations between them. Finally, our results are relationships with similar works of literature. In all evaluation scenarios, our proposal presented the best results.

4.1 Dataset and features

The PMEmo is a popular music dataset with emotional annotations as a benchmark in music emotion retrieval and recognition [19]. PMEmo dataset containing emotion annotations of 794 songs in arousal and valence domains with audio and lyrics features. For acoustic features, PMEmo has a pre-computed vector composed of audio descriptors, which represent the music chorus information. In addition,

there are the lyrics and a source code for textual features to pre-process it and build the feature vector based on Bag-of-Words (BoW).

In addition to the representations of each modality, the authors of the PMEmo dataset [19] also presented competitive approaches that explore emotion classification based on fusion features, which were explored for comparison with our proposed method.

Besides these two representations, we build another textual feature on the lyrics using a pre-trained language model based on BERT models [1] fine-tuned to sentence similarity tasks, to have a more robust representation than the BoW. The model uses a triplet network structure to produce lyrics embeddings. Thus, we explored five strategies to represent the music that varies between isolated features and concatenation between them, as indicated in the PMEmo baseline scenario: `Audio`, `Bert`, `BoW`, `Audio+BoW`, `Audio+Bert`. Therefore, we have resources to build uni-modal and multi-modal scenarios to evaluate the emotion recognition task.

We have discretized the annotations in arousal and valence domains to transform the values in the labels, according to the works [42, 43]. First, we remove instances with missing information, and normalize the values of both domains with a range between 1 and 9. So, we assign the labels according to the values. An instance is negative if the normalized value is < 4.5; neutral if value $>= 4.5$ and < 5.5; or positive if value $>= 5.5$. The table 1 summarize the number of instances distributed for each label, as well as the dimensionality for all feature vectors.

Feature size	
Audio	6373
Bert	512
Bert	7670

Instances in each label		
	Arousal	Valence
Negative	151	145
Neutral	100	96
Positive	354	364
Total	605	

Table 1: Summary of dataset properties

4.2 Experimental setup

The representation learning with a GCN is done through a transductive process, where we need to know the complete dataset. As our goal is to recognize the emotions present in the music, we divided the dataset into stratified ten folds and masked the labels in test folds. As comparison approaches, we reproduce another transductive classification method based on label propagation (LPA). The instance labels in the test set are changed during the learning process and must converge to the original labels. We also reproduce an inductive scenario using a multilayer perceptron classifier (MLP). In this scenario, the test set is unknown, and the labels are predicted according to patterns learned during training [2]. In both scenarios, the features were normalized for each split according to train and test folds.

We evaluated a variation in the number of clusters that structure the objects represented by acoustic and textual features such that $k \in \{3, 7, 13, 20, 30\}$. The smallest value represents the original number of labels, while the largest value is defined by $\sqrt{N}$, where N indicates the samples in the dataset. We assume the hypothesis that by increasing the number of clusters, we can relate objects in the network with similar content but do not share the same label or objects that share a label and can be associated with objects with more similar content.

4.3 Results and discussions

We reported the mean and standard deviation for the F1-score and Accuracy metrics for each musical data representation strategy for the three evaluated methods. We refer to our representation as MRLGCN, an acronym for Music Representation Learning using GCN. Emotion labels are commonly explored in conjunction with text-based representations, so acoustic features are proposed to complement the representation. We can notice in Figure 4 that representations that concatenate both features do not lead to performance improvement. Observing the performance obtained by our approach, we can highlight the relevance of a learning process to incorporate the information of the different features.

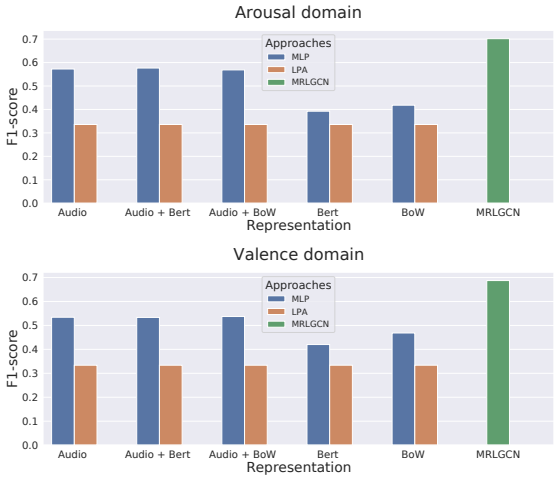

Figure 4: In the uni-modal scenario, textual features are more associated with emotion labels than acoustic features. In the multi-modal scenario, concatenating acoustic and lyrics features did not result in a more efficient representation, showing that exploring the semantic complementarity demands more robust strategies. Our representation presented the best result in both domains, evidencing the presence of relevant information in all the features used.

Figure 5 shows the results obtained by our method when the number k of clusters varies. We can notice that the performance of MRLGCN increases when k increases. It validates the hypothesis that we have objects with similar content that must be related, even those that do not share labels. Although it shows an upward trend, it is expected

[1] https://huggingface.co/sentence-transformers/distiluse-base-multilingual-cased-v1

[2] The LPA and MLP algorithms used are implementations available in the sklearn library using default parameters.

that the performance stabilizes and decreases after a certain k. This behavior is explained by the appearance of disconnected graphs that influence the regularization of the network, affecting the propagation of information between objects.

Clusters	F1-score		Accuracy	
	μ	σ	μ	σ
3	0.4530	0.096	0.6421	0.050
7	0.5231	0.087	0.6701	0.051
13	0.6010	0.107	0.7224	0.059
20	0.6612	0.120	0.7589	0.073
30	0.7018	0.131	0.7833	0.084

(a) Arousal domain

Clusters	F1-score		Accuracy	
	μ	σ	μ	σ
3	0.4820	0.089	0.6659	0.055
7	0.5300	0.099	0.6945	0.062
13	0.6095	0.127	0.7386	0.079
20	0.6399	0.137	0.7564	0.078
30	0.6873	0.151	0.7842	0.095

(b) Valence domain

Table 2: Results obtained with average (μ) and standard deviation (σ) for F1-score and accuracy metrics for each k value.

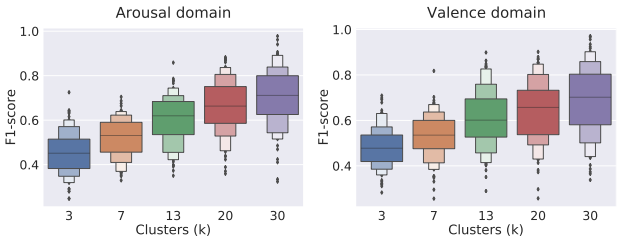

Figure 5: Performance in arousal and valence domains according to increase k value. The variation in the number of clusters highlights that the contents of the objects have similar relationships that may not be presented only by the label information.

Due to the unbalanced data distribution, we report a confusion matrix for each k to validate that the values presented for both metrics result from a learning process without bias towards the majority label. We can see that the increase in the metrics shown in Table 2 is accompanied by a proportional increase in the percentage of correct predictions for three labels, as seen in Figure 6.

5. CONCLUDING REMARKS

This work presents a new proposal to structure musical data using heterogeneous networks and build a graph-based multi-modal representation using Graph Convolutional Network to handle music emotion recognition tasks. We map acoustic and textual features as objects in different layers on a heterogeneous network, using cluster information to build relationships among objects, and insert a network regularization framework to propagate information between objects of other modalities. The embeddings

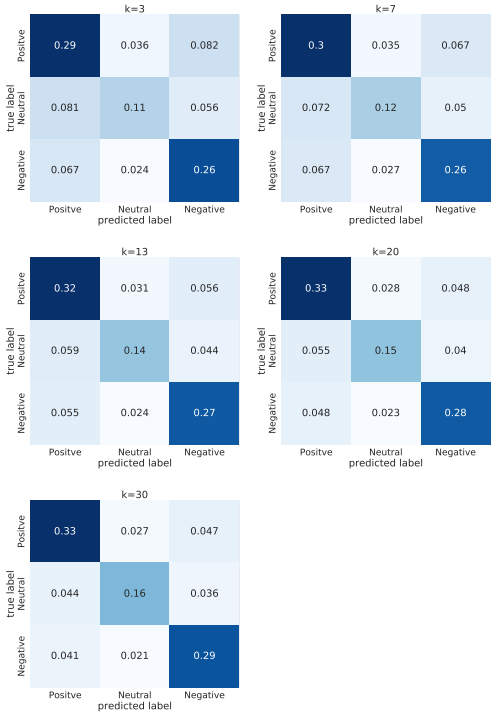

Figure 6: We highlighted the behavior of balanced growth of correct predictions across all labels, without bias for the majority label. This matrix evidences the learning process during the increase in the number of clusters.

resulting are used as input for GCN to learn a new music representation, with application in music emotion recognition task. Our proposal presented superior results in all evaluation scenarios and an significant improvement ratio concerning works in the literature. The results reinforce the applicability of graph-based representations in unstructured and multi-modal data.

Our proposal can be seen as a strategy that exploits a welcome trade-off for musical data representation. First, heterogeneous networks show explicit relationships between audio and text features and facilitate interpretability. Second, a GCN-based deep neural network learns embeddings that map the heterogeneous network into a promising representation for music emotion recognition.

In future work, we want to add other features as layers in the heterogeneous network and propose the inclusion of an importance matrix during representation learning. We will define criteria for determining optimal parameters for building the heterogeneous network and GCN architecture. In addition, we want to identify semantic relevance for each feature, measure the discriminative capacity's impact on learned representation, and evaluate the effect of the regularization step, like an ablation study. Finally, we can integrate the regularization process and GCN in an end-to-end framework for emotion music classification. The source code, datasets, and heterogeneous networks used in this paper are publicly available at `https://github.com/AngeloMendes/Heterogeneous-Graph-Neural-Network-for-Music-Emotion-Recognition`.

6. ACKNOWLEDGEMENTS

The authors of this work would like to thank the Center for Artificial Intelligence (C4AI-USP) and the support from the São Paulo Research Foundation (FAPESP) grant #2019/07665-4) and from the IBM Corporation, Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brazil (CAPES) – grant #PROEX-12049601/D, Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPQ) grant #426663/2018-7.

7. REFERENCES

[1] G. Rotter and I. Vatolkin, "Emotions," in *MUSIC DATA ANALYSIS*, C. Weihs, D. Jannach, I. Vatolkin, and G. Rudolph, Eds. CRC Press, 2017, ch. 21, pp. 511–535.

[2] P. B. Posner J, Russell JA, "The circumplex model of affect: an integrative approach to affective neuroscience, cognitive development, and psychopathology." *Dev Psychopathol*, vol. 17(3), pp. 715–734, 2005.

[3] M. Ogihara and Y. Kim, "Mood and emotional classification," in *Music data mining*, T. Li, M. Ogihara, and G. Tzanetakis, Eds. CRC Press, 2012, ch. 5, pp. 135–160.

[4] S. Zad, M. Heidari, H. James Jr, and O. Uzuner, "Emotion detection of textual data: An interdisciplinary survey," in *2021 IEEE World AI IoT Congress (AIIoT)*. IEEE, 2021, pp. 0255–0261.

[5] J. Choi, J. Song, and Y. Kim, "An analysis of music lyrics by measuring the distance of emotion and sentiment," in *2018 19th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, 2018, pp. 176–181.

[6] J. Abdillah, I. Asror, and Y. F. A. Wibowo, "Emotion classification of song lyrics using bidirectional lstm method with glove word representation weighting," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 4, no. 4, pp. 723 – 729, Aug. 2020.

[7] R. Panda, R. M. Malheiro, and R. P. Paiva, "Audio features for music emotion recognition: a survey," *IEEE Transactions on Affective Computing*, pp. 1–1, 2020.

[8] 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 *International Society for Music Information Retrieval Conference*, vol. 86, 2010, pp. 937–952.

[9] H. Xue, L. Xue, and F. Su, "Multimodal music mood classification by fusion of audio and lyrics," in *MultiMedia Modeling*, X. He, S. Luo, D. Tao, C. Xu, J. Yang, and M. A. Hasan, Eds. Cham: Springer International Publishing, 2015, pp. 26–37.

[10] R. Rajan, J. Antony, R. A. Joseph, J. M. Thomas *et al.*, "Audio-mood classification using acoustic-textual feature fusion," in *2021 Fourth International Conference on Microelectronics, Signals & Systems (ICMSS)*. IEEE, 2021, pp. 1–6.

[11] Y. Pandeya and J. Lee, "Deep learning-based late fusion of multimodal information for emotion classification of music video," *Multimed Tools Appl*, 2020.

[12] C. Chen and Q. Li, "A multimodal music emotion classification method based on multifeature combined network classifier," *Mathematical Problems in Engineering*, 2020.

[13] H. Wu, C. Kao, Q. Tang, M. Sun, B. McFee, J. Bello, and C. Wang, "Multi-task self-supervised pre-training for music classification," *Proceedings - ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2021-June, pp. 556–560, 2021.

[14] C. Shi and P. Yu, *Heterogeneous Information Network Analysis and Applications*, 01 2017.

[15] C. Yang, Y. Xiao, Y. Zhang, Y. Sun, and J. Han, "Heterogeneous network representation learning: A unified framework with survey and benchmark," *IEEE Transactions on Knowledge and Data Engineering*, vol. PP, pp. 1–1, 12 2020.

[16] C. Shi, *Heterogeneous Graph Neural Networks.* Singapore: Springer Singapore, 2022, ch. 16, pp. 351–370.

[17] H. Cai, V. W. Zheng, and K. Chang, "A comprehensive survey of graph embedding: Problems, techniques, and applications," *IEEE Transactions on Knowledge & Data Engineering*, vol. 30, no. 09, pp. 1616–1637, sep 2018.

[18] F. Chen, Y.-C. Wang, B. Wang, and C.-C. J. Kuo, "Graph representation learning: a survey," *APSIPA Transactions on Signal and Information Processing*, vol. 9, p. e15, 2020.

[19] K. Zhang, H. Zhang, S. Li, C. Yang, and L. Sun, "The pmemo dataset for music emotion recognition," in *International Conference on Multimedia Retrieval*. New York, NY, USA: Association for Computing Machinery, 2018, p. 135–142. [Online]. Available: https://doi.org/10.1145/3206025.3206037

[20] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, Aug 2013.

[21] W. W. Y. Ng, W. Zeng, and T. Wang, "Multi-level local feature coding fusion for music genre recognition," *IEEE Access*, vol. 8, pp. 152 713–152 727, 2020.

[22] H. Phan, H. L. Nguyen, O. Y. Chén, L. Pham, P. Koch, I. McLoughlin, and A. Mertins, "Multi-view audio and music classification," 2021.

[23] J. S. Gómez-Cañón, E. Cano, T. Eerola, P. Herrera, X. Hu, Y.-H. Yang, and E. Gómez, "Music emotion recognition: Toward new, robust standards in personalized and context-sensitive applications," *IEEE Signal Processing Magazine*, vol. 38, no. 6, pp. 106–114, 2021.

[24] Y. Agrawal, R. Shanker, and V. Alluri, "Transformer-based approach towards music emotion recognition from lyrics." in *European Conference on Information Retrieval*, 2021, pp. 167–175.

[25] F. H. Rachman, R. Sarno, and C. Fatichah, "Music emotion detection using weighted of audio and lyric features," in *Information Technology International Seminar*, 2020, pp. 229–233.

[26] H. C. Wang, S. W. Syu, and P. Wongchaisuwat, "A method of music autotagging based on audio and lyrics." *Multimedia Tools and Applications*, vol. 80, p. 15511–15539, 2021.

[27] S. Grollmisch, E. Cano, C. Kehling, and M. Taenzer, "Analyzing the potential of pre-trained embeddings for audio classification tasks," in *European Signal Processing Conference (EUSIPCO)*, 2021, pp. 790–794.

[28] J. Kim, J. Urbano, C. C. S. Liem, and A. Hanjalic, "One deep music representation to rule them all? a comparative analysis of different representation learning strategies," *Neural Computing and Applications*, Mar 2019. [Online]. Available: https://doi.org/10.1007/s00521-019-04076-1

[29] A. C. M. da Silva, P. R. V. do Carmo, R. M. Marcacini, and D. F. Silva, "Instance selection for music genre classification using heterogeneous networks," *Simpósio Brasileiro de Computação Musical*, vol. 18, pp. 11–18, 2021.

[30] A. C. M. da Silva, D. F. Silva, and R. M. Marcacini, "Multimodal representation learning over heterogeneous networks for tag-based music retrieval," *Expert Systems with Applications*, vol. 207, pp. 1–9, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417422012003

[31] P. Knees and M. Schedl, *Music similarity and retrieval: an introduction to audio-and web-based strategies*. Springer, 2016, vol. 36.

[32] D. Edmonds and J. Sedoc, "Multi-emotion classification for song lyrics," in *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 2021, pp. 221–235.

[33] F. Xia, K. Sun, S. Yu, A. Aziz, L. Wan, S. Pan, and H. Liu, "Graph learning: A survey," *IEEE Transactions on Artificial Intelligence*, vol. 2, no. 02, pp. 109–127, apr 2021.

[34] W. L. Hamilton, R. Ying, and J. Leskovec, "Representation learning on graphs: Methods and applications," *IEEE Data Eng. Bull.*, vol. 40, no. 3, pp. 52–74, 2017. [Online]. Available: http://sites.computer.org/debull/A17sept/p52.pdf

[35] F. Korzeniowski, S. Oramas, and F. Gouyon, "Artist similarity with graph neural networks." in *International Conference on Music Information Retrieval*, 2021, pp. 350–357.

[36] A. Saravanou, F. Tomasi, R. Mehrotra, and M. Lalmas, "Multi-task learning of graph-based inductive representations of music content," in *International Conference on Music Information Retrieval*, 2021, pp. 602–609.

[37] F. Simonetta, F. Carnovalini, N. Orio, and A. Roda, "Symbolic music similarity through a graph-based representation," in *Proceedings of the Audio Mostly 2018 on Sound in Immersion and Emotion*, 09 2018, pp. 1–7.

[38] D. d. F. P. Melo, I. d. S. Fadigas, and H. B. d. B. Pereira, "Graph-based feature extraction: A new proposal to study the classification of music signals outside the time-frequency domain," *PLOS ONE*, vol. 15, no. 11, pp. 1–26, 11 2020. [Online]. Available: https://doi.org/10.1371/journal.pone.0240915

[39] S. Dokania and V. Singh, "Graph representation learning for audio & music genre classification," 2019.

[40] Z. Zhang, P. Cui, and W. Zhu, "Deep learning on graphs: A survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. PP, pp. 1–20, 03 2020.

[41] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," in *Advances in neural information processing systems*, 2004, pp. 321–328.

[42] Z. Yin, Y. Wang, L. Liu, W. Zhang, and J. Zhang, "Cross-subject eeg feature selection for emotion recognition using transfer recursive feature elimination," *Front Neurorobot*, pp. 11–19, 2017.

[43] D. Garg and G. K. Verma, "Emotion recognition in valence-arousal space from multi-channel eeg data and wavelet based deep learning framework," *Procedia Computer Science*, vol. 171, pp. 857–867, 2020, third International Conference on Computing and Network Communications (CoCoNet'19). [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050920310644