INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Graph Isomorphism Network for Speech Emotion Recognition

Jiawang Liu, Haoxiang Wang

School of Computer Science and Engineering, South China University of Technology, China
202021044688@mail.scut.edu.cn, hxwang@scut.edu.cn

Abstract

Previous deep learning approaches such as Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) have been broadly used in speech emotion recognition (SER). In these approaches, speech signals are generally modeled in the Euclidean space. In this paper, a novel SER model (LSTM-GIN) is proposed, which applies Graph Isomorphism Network (GIN) on LSTM outputs for global emotion modeling in the non-Euclidean space. In our LSTM-GIN model, speech signals are represented as graph-structured data so that we can better extract global feature representation. The deep frame-level features generated from the bidirectional LSTM are converted into an undirected graph with nodes represented by frame-level features and connections defined according to temporal relations between speech frames. GIN is adopted to classify the graph representations of utterances, as it is proved of excellent discriminative power in comparative experiments. We conduct experiments on the IEMOCAP dataset, and the results show that our proposed LSTM-GIN model surpasses other recent graph-based models and deep learning models by achieving 64.65% of weighted accuracy (WA) and 65.53% of unweighted accuracy (UA).

Index Terms: speech emotion recognition, graph convolutional network, graph isomorphism network

1. Introduction

Speech emotion recognition (SER) has gained extensive attention over the past two decades. It can be applied to many fields, such as diagnosis of depression [1] and human-computer interaction [2, 3]. Recognizing emotion from speech signals is a challenging task due to the fact that acoustic features must be robust enough for various styles of speakers and environments. Besides, speech emotion always lies in parts across an entire utterance, so the outstanding capacity of global emotion modeling is important for an SER system.

Recent SER methods focus on deep neural networks (DNNs) which can extract hierarchical and discriminative feature representations by supervised learning [4, 5, 6, 7, 8]. In [4], the authors propose to use Extreme Learning Machine (ELM) for feature representations learning and emotion classification. Accuracy is further improved with the application of CNN and LSTM [9, 10]. Most recently, other artificial intelligence (AI) technologies, such as self-attention mechanism [7] and multimodal models [11] have been proposed for SER task.

Graph Neural Network (GNN) has attracted more researchers' attention in recent years because of its excellent capacity in the graph-structured data processing. GNN-based models can be used in computer vision such as action recognition [12, 13] and semantic segmentation [14]. Besides, some scholars address the traffic prediction problem by treating the traffic network as a spatial-temporal graph [15]. For SER, some scholars utilize GNN for conversational emotion analysis [16]. In [16], two types of nodes are included in a graph, one is utterance node and the other is speaker node. Each utterance node has edges according to its speech order and another edge is connected with its speaker. Moreover, there are also some scholars who build a graph directly on individual utterances. For example, in [17], the authors apply a graph attention module on Gated Recurrent Unit (GRU) network. In [18], the authors transform extracted acoustic features into a line graph. The SER task is taken as a graph classification task and spectral-based graph convolutional network (GCN) is performed on each graph.

But these methods based on spectral-based GCNs focus on spectral information from the perspective of graph signal processing ignoring spatial relation between nodes and can not propagate node information along edges. Besides, the spectral-based approach processes the entire graph simultaneously, which can not take advantage of parallel processing. To address these problems, we propose to utilize the spatial-based GCN for SER. Spatial-based GCNs update node representation with itself and localized neighbors' information through the aggregation function. Various aggregation functions and graph-level pooling methods have been proposed, such as GAT [19], GraphSAGE [20] and PATCHY-SAN [21]. These spatial-based GCNs achieve great success in link prediction [22], node classification [23] and graph classification [24]. In [25], Xu et al. propose the Graph Isomorphism Network (GIN) which possesses discriminative power over other GCNs. In this paper, we propose to utilize the discriminative power of GIN for SER task by encoding the high-level acoustic feature into a graph. The results show the effectiveness of the proposed LSTM-GIN model.

The contributions of this article are summarized as follows:

- To the best of our knowledge, this is the first try of using spatial-based GCN for SER. The SER task is turned into a graph classification problem by transforming an utterance into a graph, in which nodes are represented by frame-level features and connections are defined according to temporal relations between frames.
- We compare the proposed LSTM-GIN model with different implementations using both spatial-based GCNs and spectral-based GCNs. Experimental results reveal that spatial-based GCNs, especially GIN, can conduct very competitive accuracy in SER task.
- Our proposed LSTM-GIN model achieves 64.65% of WA and 65.53% of UA on the IEMOCAP dataset, which surpasses other state-of-the-art SER models.

2. Proposed Method

Figure 1 shows the proposed architecture of LSTM-GIN. First, the low-level descriptors (LLDs) features are extracted from each utterance as model inputs. To reduce the network size and the complexity of computation, every five frames are averaged to one. Next, a bidirectional LSTM is used for temporal modeling and the following five-layered GINs are used to aggregate

http://dx.doi.org/10.21437/Interspeech.2021-1154

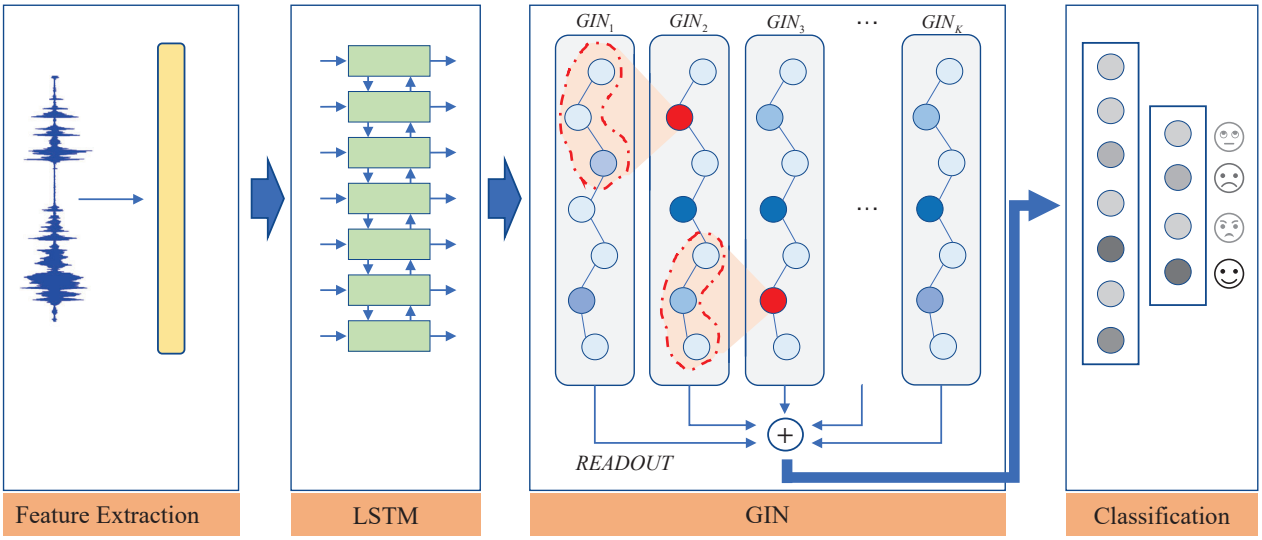

Figure 1: *Illustration of the proposed LSTM-GIN model for SER to aggregate global emotional information.*

global salient information. Finally, a readout function is applied on every GIN layer to get graph-level emotion representations.

2.1. Graph Construction

Speech signal does not have a graph structure similar to the molecular structures and citation relationships. But we can transform each utterance into a line graph where each frame of speech signal form a node in the graph and the frame-level acoustic features are treated as node feature vectors [18]. For each utterance, 78 dimensional frame-level LLDs from the INTERSPEECH 2010 paralinguistic challenge [26] are extracted using the openSMILE toolkit [27]. The LLDs feature set contains PCM loudness, F0 Envelope, LSP Frequency, Mel-frequency cepstral coefficients (MFCCs), jitter, etc. Every five frames are averagely down-sampled into one to decrease the complexity of computation. Finally, we get a feature size of 120×78 with longer utterances are cut and shorter ones are padded with zeros.

Different from [18], instead of using the line graph directly, we connect a node with its several nearby nodes, for learning more emotional information from the context of a frame. The frames in a certain period of time are considered as connected nodes in our graph. Figure 2 illustrates that for a given node in the line graph, it can be modeled with 2 or 4 neighboring nodes respectively.

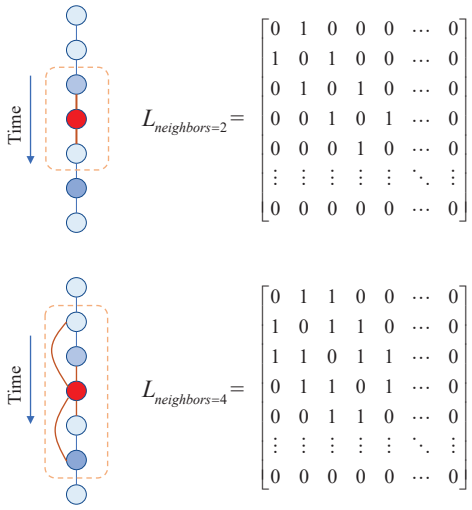

Figure 2: *Based on the line graph with nodes corresponding to each frame in an utterance, we can model the nodes in a certain time period as connected neighbors to enlarge the receptive context. The top half of the picture demonstrates the modeling of a node with two nearby neighbors and the corresponding adjacent matrix. The bottom half displays the modeling of a node with four nearby neighbors and its adjacent matrix.*

2.2. Graph Isomorphism Network for Graph Classification

In this subsection, we introduce a GIN-based graph classification network named LSTM-GIN. We use the bidirectional LSTM as the deep features extraction block given the excellent capacity of long-range dynamic dependencies modeling. 128 memory cells are included in each direction to encodes the left and right sequence contexts. Then GIN is used to further integrate global emotional information.

GIN is a spatial-based GCN network proposed in [25] achieving state-of-the-art result on the graph classification task. It has been proved that GIN is a maximal power GNN for neighbor aggregation task the same effect as Weisfeiler-Lehman test [28]. Unlike spectral convolution which requires the entire graph to be computed simultaneously, spatial convolution aggregates information from its local neighborhood and can be computed in a mini-batch of nodes. The node representation in GIN updates as

$$h_i^{(k)} = \mathrm{MLP}^{(k)}\left(\left(1+\epsilon^{(k)}\right)\cdot h_i^{(k-1)} + \sum_{u\in\mathcal{N}(i)} h_u^{(k-1)}\right) \quad (1)$$

where MLP denotes Multilayer Perceptron, $h_i^{(k)}$ denotes feature representation of node i in the kth hidden layer. We initialize $h_i^{(0)} = \boldsymbol{s}_i$ for $i \in \{1, \ldots, N\}$ where $\boldsymbol{s}_i$ is the LSTM output of time step i. The number of nodes in the built graph is equal to the number of segmented frames N and $N = 120$. The set of neighborhoods of a vertex i is denoted by $\mathcal{N}(i)$ and ϵ is a learnable parameter.

Node embeddings can be directly used for link prediction and node classification. But for graph classification tasks, we usually get the embedding of the entire graph through a readout function that operates on given embeddings of individual nodes. Then, we concatenate graph embeddings of different GIN layers

$$h_G = \underset{k=0,1,\ldots,K}{\mathrm{CONCAT}} \left(\mathrm{READOUT} \left(\left\{ h_i^{(k)} \mid i \in G \right\} \right) \right) \tag{2}$$

where h_G is the aggregated graph-level representations used for final emotion classification. Readout function can be chosen from sum, mean and max. K denotes the number of GIN layers and is set to 5 in our experiment.

3. Experimental Setup and Results

3.1. Dataset

For experiments, we choose the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset, which is widely used for speech emotion recognition. The IEMOCAP dataset includes five sessions, in each of five sessions, an actor and an actress are recorded in emotional scripts and improvised hypothetical scenarios respectively, a total of 12 hours. To be consistent with previous studies, four major emotions are chosen including natural, happy(including excitement), sad and angry. Finally, 5531 utterances are chosen for speech emotion recognition.

3.2. Experimental Settings

To be consistent in comparing with past works[17], we adopt the leave-one-speaker-out cross-validation in our experiment to ensure that there is no speaker overlap between the training and testing set. The results from Both weighted accuracy (WA) and unweighted accuracy (UA) are computed where WA presents the overall accuracy of the entire test data and UA is an average of each emotion class. We train LSTM-GIN for a maximum of 200 epochs with early stopping. Adam [29] optimizer is used with the initial learning rate is set to 1e-4 and weight decay is set to 1e-8. All models are implemented in PyTorch [30].

3.3. Comparison with GCNs

In order to prove the effectiveness of the proposed LSTM-GIN model, we implement the popular spectral-based and spatial-based GCNs including standard GCN [31], GAT [19] and GraphSAGE [20]. To be consistent with the settings used in the LSTM-GIN model, five GCN layers are used in all the graph-based experiments. We also compare the performance against the graph-based GA-GRU SER model [17] which was proposed in INTERSPEECH 2020. The details of each model are as follows:

- **GCN[31]**. A spectral-based GCN applies a first-order approximation of spectral convolution on the graph by using a layer-wise propagation rule.
- **GAT[19]**. This network is proposed by Petar Veličković in 2017. The graph attentional layer allows for assigning different importance to different neighbor nodes, so that more important information can be aggregated from neighbor nodes.
- **LSTM-GAT**. On the basis of GAT, we insert a bidirectional LSTM layer before five-layered GATs to model long-range dynamic emotion dependencies first. The memory cells are set to 128 to be consistent with the proposed LSTM-GIN model.
- **GraphSAGE[20]**. An inductive approach is utilized to generate representations for unseen nodes. Different aggregator architectures are applied to explore the best performance. We compare the LSTM and mean aggregator in our experiments.
- **GA-GRU[17]**. This model was proposed by Su et al. in INTERSPEECH 2020. A graph attention mechanism is imposed on the GRU network to integrate emotion salient parts.

Table 1: *Comparison with graph-based methods on the IEMOCAP dataset.*

Models	WA(%)	UA(%)
GCN	61.16	62.21
GAT	60.93	62.09
LSTM-GAT	64.28	65.19
GraphSAGE-mean	63.39	63.97
GraphSAGE-lstm	63.64	64.14
GA-GRU[17]	62.27	63.8
LSTM-GIN(our)	**64.65**	**65.53**

As shown in Table 1, our proposed LSTM-GIN model obtains the best performance among graph-based methods. The performance of our LSTM-GIN is 2.5% and 3.3% better than the spectral-based GCN model on WA and UA respectively. We can find that spatial-based models tend to perform better than spectral-based models attributing to more efficient information aggregation methods used in spatial-based GCNs. With further observation on the GAT model, we can find that the use of LSTM layer improves the WA and UA by about 3% showing the effectiveness of LSTM layer. For the GraphSAGE model, LSTM aggregator gets competitive results with the mean aggregator, but LSTM aggregator is time-consuming. The confusion matrix of LSTM-GIN is shown in Figure 3.

3.4. Comparison with Deep Learning Methods

Previous deep learning models on SER mainly focus on CNN and LSTM based architecture. We compare the following recent per-utterance deep learning methods with our proposed LSTM-GIN. The details of each model are as follows:

- **SVM[17]**. Utterance-level 88 dimensional eGemaps features are extracted and classified using support vector machine (SVM).
- **3D-ACRNN[6]**. The authors append an attention layer on the CNN and LSTM block to focus on emotion salient parts.
- **CNN-LSTM-DNN[32]**. An architecture that models the SER task directly on raw speech. CNN is used to extract

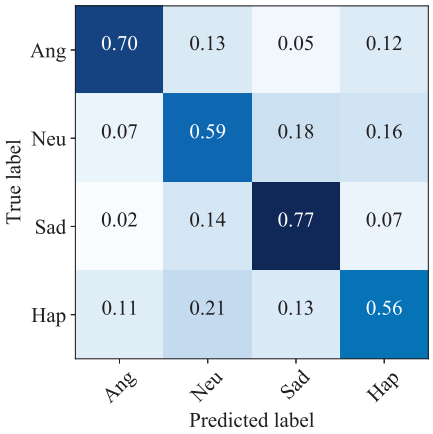

Figure 3: *Confusion Matrix of our proposed LSTM-GIN model on four emotions classification.*

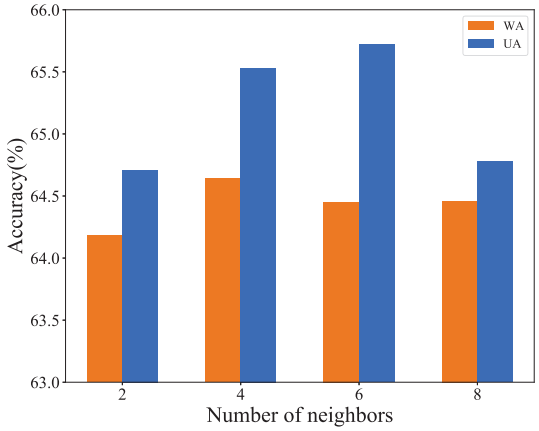

Figure 4: *Impacts of Neighbor Number.*

high-level features and an LSTM layer is used to capture the temporal information.

- **BiGRU+Attn[17]**. A framework which is used as a baseline in [17] in which attention module is applied on a bidirectional GRU network.
- **SegCNN[33]**. Feature representation learning is combined with multiple instance learning for the SER task. Segment-level emotional state decisions are aggregated by utterance-level classification to get the final emotion prediction.

Table 2: *UA and WA performances of deep learning methods and our proposed LSTM-GIN model on the IEMOCAP dataset.*

Models	WA(%)	UA(%)
SVM [17]	56.88	59.38
3D-ACRNN 2018[6]	-	64.74
CNN-LSTM-DNN 2019[32]	-	60.23
BiGRU+Attn 2020[17]	61.51	62.44
SegCNN 2019[33]	64.53	62.34
LSTM-GIN(max)	62.79	64.07
LSTM-GIN(mean)	63.99	65.30
LSTM-GIN(sum)	**64.65**	**65.53**

Table 2 shows the results compared with recent DNN-based SER models. We see that our proposed LSTM-GIN model outperforms all the deep learning methods listed. According to [25], readout function acts on an entire graph to obtain graph-level embeddings. Different readout functions may produce different results. We test the accuracies under different readout functions including mean, max and sum. As shown in Table 2, we get the best WA and UA when we choose sum as the readout function for the reason that max and mean readout may fail to distinguish graph structures and lead to underlying information loss, which is consistent with the past work [25].

3.5. Impacts of Neighbor Number

Different from the line graph proposed in [18], we suppose that different neighbor nodes setting may affect the accuracy of emotion recognition. In theory, the more neighbors, the more information can be aggregated from its context, but when the neighbor number reaches a certain amount, distracting and useless information will also increase. We conduct comparative experiments with the number of neighbors set as 2, 4, 6 and 8. As can be seen from Figure 4, with the increase of neighbor nodes, the accuracy increases first. WA reaches up to 64.65 when the number of neighbors is set to 4 and the UA reaches the maximum of 65.72 when the number of neighbors is 6. Then accuracies began to decline when the number of neighbors exceeds 6 as we predict.

4. Conclusion

In this paper, we first propose to utilize spatial-based GIN for SER modeling. Frame-level LLDs features are first fed into a bidirectional LSTM for time series modeling. The deep feature outputs from the bidirectional LSTM at every time step are taken as node features and nodes are connected according to their temporal relations. The proposed LSTM-GIN model achieves 64.65% of WA and 65.53% of UA on the IEMOCAP dataset, which surpasses other graph-based GCN models and common deep neural networks such as CNN and LSTM. We implement the most popular GNN models used for the SER task, experimental results show that the spatial-based GCNs also possess excellent ability for global emotion recognition. In the future, we will explore more efficient aggregation methods, which can learn more useful information from neighbor nodes.

5. Acknowledgements

This work was supported by Guangdong Basic and Applied Basic Research Foundation, 2021A1515011852; The Fundamental Research Funds for the Central Universities, x2js-D2190680.

6. References

[1] Y. Zhu, Y. Shang, Z. Shao, and G. Guo, "Automated depression diagnosis based on deep networks to encode facial appearance and dynamics," *IEEE Transactions on Affective Computing*, vol. 9, no. 4, pp. 578–584, 2018.

[2] T. Balomenos, A. Raouzaiou, S. Ioannou, A. Drosopoulos, K. Karpouzis, and S. Kollias, "Emotion analysis in man-machine

interaction systems," in *Machine Learning for Multimodal Interaction*, 2005, pp. 318–328.

[3] J. G. R´azuri, D. Sundgren, R. Rahmani, A. Larsson, A. M. Cardenas, and I. Bonet, "Speech emotion recognition in emotional feedback for human-robot interaction," *International Journal of Advanced Research in Artificial Intelligence*, vol. 4, no. 2, pp. 20–17, 2015.

[4] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Proc. Interspeech*, 2014, pp. 223–227.

[5] J. Lee and I. Tashev, "High-level feature representation using recurrent neural network for speech emotion recognition," in *Proc. Interspeech*, 2015, pp. 1537–1540.

[6] M. Chen, X. He, J. Yang, and H. Zhang, "3-D convolutional recurrent neural networks with attention model for speech emotion recognition," *IEEE Signal Processing Letters*, vol. 25, no. 10, pp. 1440–1444, 2018.

[7] Y. Li, T. Zhao, and T. Kawahara, "Improved end-to-end speech emotion recognition using self attention mechanism and multitask learning," in *Proc. Interspeech*, 2019, pp. 2803–2807.

[8] Z. Zhao, Z. Bao, Z. Zhang, N. Cummins, H. Wang, and B. W. Schuller, "Attention-enhanced connectionist temporal classification for discrete speech emotion recognition," in *Proc. Interspeech*, 2019, pp. 206–210.

[9] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 2227–2231.

[10] D. Wang, J. Dong, D. Zhou, X. Wei, and Q. Zhang, "Speech emotion recognition based on image enhancement," in *2019 IEEE 14th International Conference on Intelligent Systems and Knowledge Engineering (ISKE)*, 2019, pp. 1126–1132.

[11] Z. Pan, Z. Luo, J. Yang, and H. Li, "Multi-modal attention for speech emotion recognition," in *Proc. Interspeech*, 2020, pp. 364–368.

[12] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[13] A. Jain, A. R. Zamir, S. Savarese, and A. Saxena, "Structural-RNN: Deep learning on spatio-temporal graphs," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 5308–5317.

[14] X. Qi, R. Liao, J. Jia, S. Fidler, and R. Urtasun, "3D graph neural networks for rgbd semantic segmentation," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 5209–5218.

[15] J. Zhang, X. Shi, J. Xie, H. Ma, I. King, and D. Y. Yeung, "GaAN: Gated attention networks for learning on large and spatiotemporal graphs," in *34th Conference on Uncertainty in Artificial Intelligence 2018, UAI 2018*, 2018.

[16] Z. Lian, J. Tao, B. Liu, J. Huang, Z. Yang, and R. Li, "Conversational emotion recognition using self-attention mechanisms and graph neural networks," in *Proc. Interspeech*, 2020, pp. 2347–2351.

[17] B.-H. Su, C.-M. Chang, Y.-S. Lin, and C.-C. Lee, "Improving speech emotion recognition using graph attentive bi-directional gated recurrent unit network," in *Proc. Interspeech*, 2020, pp. 506–510.

[18] A. Shirian and T. Guha, "Compact graph architecture for speech emotion recognition," in *2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6284–6288.

[19] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *International Conference on Learning Representations*, 2018.

[20] W. L. Hamilton, R. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *International Conference on Neural Information Processing Systems*, 2017, p. 1025–1035.

[21] M. Niepert, M. Ahmed, and K. Kutzkov, "Learning convolutional neural networks for graphs," in *International Conference on Machine Learning*, 2016, pp. 2014–2023.

[22] M. Zhang and Y. Chen, "Link prediction based on graph neural networks," in *International Conference on Neural Information Processing Systems (NIPS)*, 2018, pp. 5171–5181.

[23] A. Grover and J. Leskovec, "Node2vec: Scalable feature learning for networks," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 855–864.

[24] M. Zhang, Z. Cui, M. Neumann, and Y. Chen, "An end-to-end deep learning architecture for graph classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, 2018.

[25] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, "How powerful are graph neural networks?" in *International Conference on Learning Representations*, 2019.

[26] C. Müller, "The INTERSPEECH 2010 paralinguistic challenge," in *Proc. Interspeech*, 2010, pp. 2794–2797.

[27] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the Munich versatile and fast open-source audio feature extractor," in *Proceedings of the International Conference on Multimedia*, 2010, pp. 1459–1462.

[28] B. Weisfeiler and L. Andrei, "The reduction of a graph to canonical form and the algebra which appears therein," *NTI*, vol. 2, pp. 12–16, 1968.

[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2015.

[30] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch," in *International Conference on Neural Information Processing Systems Workshop (NIPS Workshop)*, 2017.

[31] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *International Conference on Learning Representations*, 2017.

[32] S. Latif, R. Rana, S. Khalifa, R. Jurdak, and J. Epps, "Direct modelling of speech emotion from raw speech," in *Proc. Interspeech*, 2019, pp. 3920–3924.

[33] S. Mao, P. Ching, and T. Lee, "Deep learning of segment-level feature representation with multiple instance learning for utterance-level speech emotion recognition," in *Proc. Interspeech*, 2019, pp. 1686–1690.