[Re] COGMEN: COntextualized GNN based Multimodal Emotion recognitioN

Srijan,
IIIT-Hyderabad (Healthcare and AI Center)
Hyderabad, 500032, India
srijan.dokania@ihub-data.iiit.ac.in

Abstract

For development of AI system that understand human emotions accurately, a multi-modal model which takes context of the conversation, as well as neighbour utterances into account is necessary. Different modalities like text, audio and video should be taken into account for better predictions. There is not much attention given towards the multi-modal approach of emotions to model the complex dependencies between emotions. To mitigate this issue, the authors in the paper has proposed a model which uses Graph Neural Network for better multi-modal predictions. IEMOCAP dataset is used. This is the report of an attempt to reproduce the results of the mentioned paper [5] and compare the experimental results of the original and reproduced models. Re-implemented Code can be found at: https://github.com/Srijan221/COGMEN.git

1. Introduction

In the modern era, emotion recognition have become critical for the advancement of artificial intelligence technologies. Each person's emotions feel and express themselves in a continual ebb and flow. There is a role of overall context as well as inter and intra-personal dependency in predicting emotion in a conversation. For example, if the overall context is about having a holiday or going on a vacation, then the emotions in the conversation would be inclined more towards happy and excited. So, the emotional responses represented in utterances are intimately related to the context. Not much work is done in the multi-modal approach for emotion recognition as compared to unimodal recognition (text only). A multi-modal approach is required as any real person normally takes different modalities into account to understand emotions of others in a conversation like speech (audio) and expressions (video).

To address this gap, the authors propose a Contextualized Graph neural network based multimodal emotion recognition architecture (COGMEN) for per-utterance emotion detection that leverages context and inter and intra dependency of utterances to give better state of the art performance on the IEMOCAP dataset in compared to existing algorithms.

2. Methodology

2.1. Dataset

The dataset used is the IEMOCAP dataset [2]. This is one of the largest dataset for multimodal emotion prediction. Here, each utterance in a conversation is assigned a label from one of the six emotion categories— angry, thrilled, sad, happy, frustrated, and neutral. Two IEMOCAP settings are utilised for experiments: having four emotions (anger, sorrow, happiness, and neutral) and six emotions.

2.2. Model Architecture

To get the context as well as effect of nearby utterances, author's uses two type of sources of information:

- **Global Information:** To capture overall context and deal with its impact on each utterance, authors leverage a transformer encoder architecture [12].
- **Local Information:** To construct a relationship between neighbouring utterances that is capable of obtaining intra-speaker and inter-speaker impacts of stimuli over an utterance's affective state, authors leverage Relational GCN [10] and a GraphTransformer [11]. This approach is very similar to already proposed DialogueGCN [4].

The authors divided proposed model into four parts as shown in Figure 1.

1. **Context Extractor:** Context Extractor uses a transformer encoder to extract the context of each conversation utterance utilising concatenated features from several modalities (audio, video, text).
2. **Graph Formation:** The inter and intra-speaker dependencies between utterances form a graph. Each statement functions as a node in a graph that is connected by directed relations (past and future relations). In Figure 1, R_{intra} is the self-dependent relations between utterances said by the same speaker and R_{inter} is the inter-relations said by different speakers.
3. **RGCN + GraphTransformer:** Relational Graph Convolutional Network [10] captures the inter and intra-speaker dependency on the connected utterances

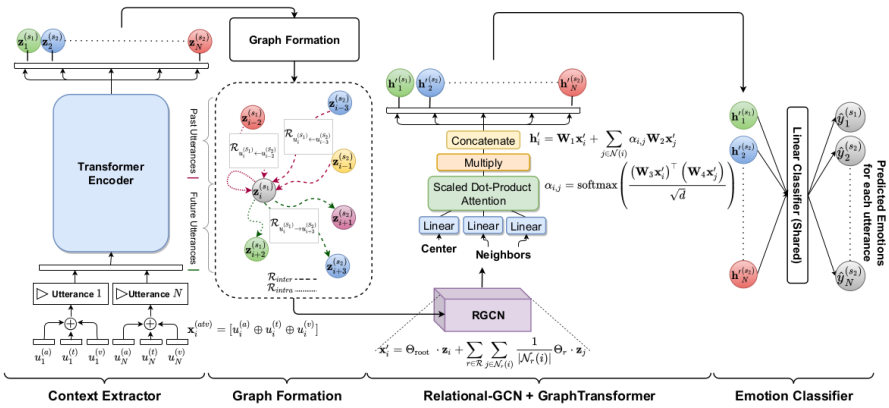

Figure 1. Model Architecture

from the graph obtained and GraphTransformer [11] then, extract rich representations from these node features by multi-head-attention into graph-learning.

4. **Emotion Classifier:** Finally a linear layer like ReLU acts as an emotion classifier and use the features to predict emotion class.

3. Experimental Work

To preprocess the IEMOCAP dataset, audio features (size 100) are extracted using OpenSmile [3], video features (size 512) are dervied from OpenFace 2.0 [1] and text features (size 768) are obtained using sBERT [9].

Different Hyperparameters setting are used to train the model to obtain better accuracy over IEMOCAP (4-way and 6-way) dataset. The Hyperparameter values used by the original model and the reimplemented model is shown in Table 2. These Hyperparameters were tuned using a Bayesian Optimizer.

The authors also compared their proposed COGMEN model with some existing multimodal frameworks shown in Table 1, like DialogueRNN [7], bc-LSTM [8] and CHFusion [6].

Table 2. Hyperparameter Setting

Model	Optimizer	Dropout	GNNHead	SeqContext
Original	Adam	0.1	7	4
Reproduced	RMSProp	0.5	1	6

4. Results

Table 1 shows the result for IEMOCAP (6-way setting), that the proposed model achieved the best performance across the accuracy and F1 score as compared to the other models. There is also an improvement in class-wise F1 scores. This is because their model, uses GNN architecture which no other model incorporates. Results for IEMOCAP (4-way setting) are shown in Table 3, that shows how the change in hyperparameters in the re-implemented model increases the weighted F1 score from the original COGMEN model. This improvement can be seen again at class level in Table 1.

In Table 4, a-audio, t-text, v-video are different modalities for which the proposed and reproduced COGMEN model are compared with IEMOCAP 4-way and 6-way settings. The maximum F1 score obtained is of A+T+V modality which further asserts the fact that a multi-modal architecture is better for emotion prediction. Confusion matrix for IEMOCAP (4-way) dataset is shown in Figure 2.

Table 3. Comparison of Proposed / Reproduced Model on IEMOCAP (4-way) emotion setting (weighted F1-score).

Model	F1 Score (%)	
bc-LSTM [8]	75.13	
CHFusion [6]	76.8	
COGMEN [5]	**Proposed**	**Reproduced**
	84.50	**85.44**

Table 1. Overall Performance Comparison of Proposed / Reproduced Model on IEMOCAP (6-way A+T+V) emotion setting.

Model	Happy	Sad	Neutral	Angry	Excited	Frustrated	Avg. Acc. (%)	Avg F1 (%)
bc-LSTM	35.6	69.2	53.5	**66.3**	61.1	**62.4**	59.8	59
DialogueRNN	32.8	78	59.1	63.3	73.6	59.4	63.3	62.8
COGMEN	51.9 / **57.9**	**81.7** / 80.16	**68.6** / 60.06	66 / 63.02	75.3 / **77.61**	58.2 / 61.4	68.2 / **69.6**	67.6 / **69.4**

Table 4. Comparison of modalities in Proposed / Reproduced Model on IEMOCAP (4-way, 6-way) dataset.

Model	IEMOCAP 4-way F1 Score (%)		IEMOCAP 6-way F1 Score (%)	
	Proposed	Reproduced	Proposed	Reproduced
a	63.58	58.29	47.57	47.30
t	81.55	82.45	66.00	68.11
v	43.85	39.93	37.58	40.12
at	81.59	81.83	65.42	63.18
av	64.48	60.20	52.20	55.50
tv	81.52	82.30	62.19	64.76
atv	84.50	**85.44**	67.63	**69.42**

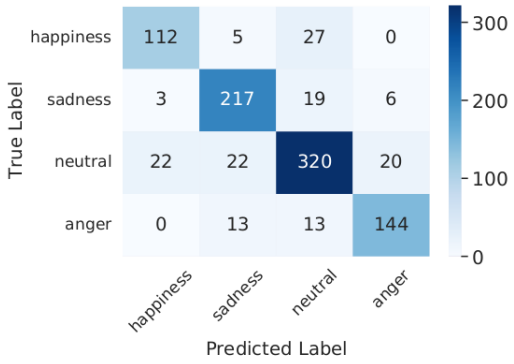

Figure 2. Confusion Matrix for IEMOCAP 4-Way classification

5. Conclusion

In this paper, the model for Contextualized GNN based multimodal Emotion recognition is re-implemented and results of comparison between original and reproduced scores are shown. The reproduced model is trained on a different hyperparameter setting and tested on IEMOCAP dataset that outperforms the other existing models and gives better results than the original model. The accuracy scores slightly deviate from the original reported values because of the limited GPU memory and overall errors due to different hyperparameter tuning.

In future scope, the model can be trained on not just 4-way or 6-way class but more. A larger dataset can also be used for better training of model. As observed in Figure 2, the model sometimes fails to capture emotional shifts when the emotions are similar like happiness and neutral. A new component can be integrated which can capture these emotional shifts and increase overall accuracy. Also the model can be evaluated on wider range of emotions and experimented with different settings of modalities to get better understanding on its interpretability.

References

[1] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, pages 59–66, 2018. 2

[2] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette Chang, Sungbok Lee, and Shrikanth S. Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Journal of Language Resources and Evaluation*, 42(4):335–359, dec 2008. 1

[3] Florian Eyben, Martin Wöllmer, and Björn Schuller. Opensmile: The munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, page 1459–1462, New York, NY, USA, 2010. Association for Computing Machinery. 2

[4] Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. DialogueGCN: A graph convolutional neural network for emotion recognition in conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 154–164, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. 1

[5] Abhinav Joshi, Ashwani Bhat, Ayush Jain, Atin Vikram Singh, and Ashutosh Modi. Cogmen: Contextualized gnn based multimodal emotion recognition, 2022. 1, 2

[6] N. Majumder, D. Hazarika, A. Gelbukh, E. Cambria, and S. Poria. Multimodal sentiment analysis using hierarchical fusion with context modeling, 2018. 2

[7] Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. Dialoguernn: An attentive rnn for emotion detection in conversations, 2018. 2

[8] Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. Emotion recognition in conversation: Research challenges, datasets, and recent advances, 2019. 2

[9] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019. 2

[10] Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks, 2017. 1

[11] Yunsheng Shi, Zhengjie Huang, Shikun Feng, Hui Zhong, Wenjin Wang, and Yu Sun. Masked label prediction: Unified message passing model for semi-supervised classification, 2020. 1, 2

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017. 1