PREDICTING PERFORMANCE OUTCOME WITH A CONVERSATIONAL GRAPH CONVOLUTIONAL NETWORK FOR SMALL GROUP INTERACTIONS

Yun-Shao Lin[1,2], *Chi-Chun Lee*[1,2]

[1]Department of Electrical Engineering, National Tsing Hua University, Taiwan
[2]MOST Joint Research Center for AI Technology and All Vista Healthcare, Taiwan

ABSTRACT

Studying behaviors of members during small group interaction provides objective insights in improving the efficiency of the decision making process in our daily working life. By introducing the use of the graph structure in modeling the natural inter-member conversational ties during such an interaction, we aim to advance the state-of-art computational approach in predicting group performance scores. Specifically, we proposed a Conversational Graph Convolutional Network (CGCN) that utilizes conversation dynamic as the graph to aggregate group member's speech and lexical behaviors in predicting the group performance. Our result shows that Speech CGCN achieves the state-of-the-art performance at MSE 3.896 (0.323 Pearson correlation) outperform the current best method in ELEA dataset. Our model additionally reveals that an *imbalance* conversational graph structure is positively correlated to group performances.

Index Terms— small group interaction, group performance, graph convolutional network, conversation

1. INTRODUCTION

Small group is a highly structured unit of human face-to-face interaction defined by a strict number of interacting members: three to six people [1]; it is one of the most common forms of interaction styles especially in professional settings or working environments. This particular type of human interaction provides critical advantages for improved mechanisms of human communication, such as sharing knowledge, stimulating ideas, dividing specialized works, that often lead to a more effective and a higher quality of decision making process [2, 3]. During such an interaction, each member of the group communicates with each other through both verbal and nonverbal behaviors in order to convey ideas and jointly complete a given task. The performance outcome of each group on their collaborative talk and the emergent leadership within each group has commonly been considered as result of the unique behavioral interaction dynamics between members [4, 5].

The complex and intricate behavioral interaction dynamics between members have sparked a growing interest for computational researchers to analyze behaviors of the members during small group interactions recently. In fact, several public releases of the multimodal corpus, e.g., ELEA [6], GAP [7] and UGI [8], provide well-thought-out scenarios and conversational behavior data to enable computational studies on small group interactions. There are a couple of recent works in developing advanced deep learning frameworks in joint modeling vocal behaviors and personality [9, 10]. In terms of predicting group performances, most of the previous works have focused on designing features on member's behaviors, e.g., turn-taking dynamics, lexical coherence, or non-verbal dynamics [11, 12].

While these approaches have established the basic foundation on understanding the relationship between behaviors of individual members during small group interaction and task performances, these works do not model the structure between members during the interaction, and the technical approaches are often sub-optimal compared to the current state-of-the-art machine learning. In group studies, researchers have argued the existence of a static network structure that can be used to characterize team efficiency in small group communication [13]. As an extension of the static structure, using the network as an aggregation of social interaction patterns over time also lead to applications in analyzing other virtual interaction data, such as e-mail streams in a team [14, 15]. In this work, we propose to model the real-world small group face-to-face behavioral interaction using a graph structure with conversational dynamics as the graph building block in order to perform task performance prediction.

Specifically, we propose a Conversational Graph Convolutional Network(CGCN) to predict AGS (absolute group score) in ELEA corpus based on members' speech and lexical features. This framework is inspired by the successful usage of graph convolutional network (with its ability in emphasizing the structural relationship on the data with non-grid structure [16]) for applications such as edge prediction in social network [16] or even chemical structure classification [17]. Our result shows that Speech CGCN can achieve the performance at MSE 3.896 and 0.323 Pearson correlation. It outperforms the current state-of-the-art tree-based method in the same dataset by 0.211 MSE and 0.115 Pearson correlation. Our further analysis shows that a more *imbalance* graph structure of the group conversation is positively correlated with higher group performances.

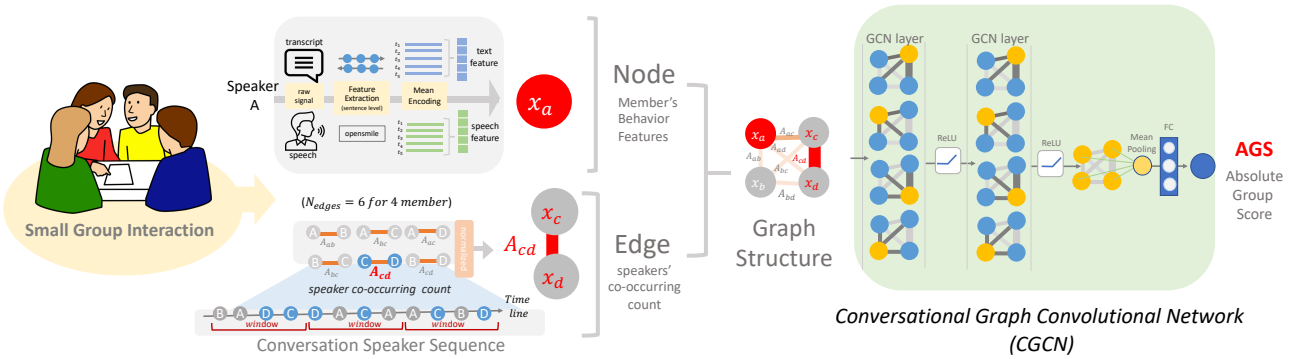

Fig. 1. Our Proposed Conversation Graph Convolutional Network: three main parts are included in the figure, the speakers' behavior feature X_i where $i = a, b, c, d$, the edge weights of conversational graph A_{ij} and the architecture of our CGCN model.

2. METHODOLOGY

2.1. Databases

Three datasets, i.e., ELEA[6], GAP[7] and UGI[8], are used in this paper. The Emergent LEAder (ELEA) corpus is one of the largest multimodal group interaction databases that has facilitated many small group interaction studies. In this work, our algorithm is evaluated in ELEA dataset. In ELEA dataset, group members are asked to imagine themselves as the survivors of an airplane crash, and they need to rank 12 objects according to the order of importance in order to survive through winter. The absolute group scores(AGS) on the task are calculated by summing the absolute differences between the group ranking and the expert ranking for each item. The ELEA dataset includes 29 groups with English speaking are used in our paper, GAP(28 groups) and UGI(22 groups) datasets are used in this work as data augmentation for ELEA dataset to stabilize and improve the prediction accuracy. GAP and UGI datasets are two newly collected small group datasets also designed with a similar survival scenarios.

2.2. Behavior Feature Extraction

2.2.1. *Speaker's Acoustic Feature*

ELEA includes manually segmented utterances for each speaker in the group. We first extract the INTERSPEECH 2010 ComparE Challenge feature set using the openSMILE toolkit [18] on each sentence. It contains statistical functionals operated on acoustic low-level descriptors (LLDs), including jitter, shimmer, MFCC, associated delta features, PCM loudness, F0 envelope, F0 contour, and voicing probability. We follow the work of Murray and Oertel [12] to select standard deviation-related features only from the original set. This results in a final set of 76 dimensions for each sentence. Finally, for each speaker of an interaction, we average all his/her sentence-level features to obtain the final 76 dimensions of session-level acoustic features.

2.2.2. *Speaker's Lexical Feature*

The lexical features are obtained using a pre-trained bidirectional skip-thought [19] sentence embedding for each sentence from the transcript given by ELEA. The skip-thought vectors are learned by using an unsupervised learning framework of an encoder-decoder structure composed by GRU-RNN. In our task, the sentence embedding of 2400 dimensions is obtained for each utterance of a given speaker. We then obtain the final session-level speaker's 2400 dimensional lexical feature by averaging all of his/her sentences.

2.3. Conversational Graph Convolutional Network

For each of the k_{th} group, we first form an undirected graph with a set of conversational weight $E_{(k)}$ and the speaker' s behavior feature $X_{(k)} = \{x_j\}$ where $j = 1, ..., N$ (with a group size of N), i.e., $G_{(k)} = (E_{(k)}, X_{(k)})$. For each graph $G_{(k)}$, there is a corresponding absolute group score $y_{(k)} \in Y_{label}$. In essence, our proposed CGCN framework learn the mapping between the pair of $(\mathbf{G}, \mathbf{Y_{label}})$.

2.3.1. *Conversational Graph*

The adjacency matrix A is constructed to represent the structure of the group G_k. First, we take all of the utterances in the interaction and form a conversation speaker sequence (a sequence indicating the order of speaking for each member in the conversation). We further use a sliding window win with a fixed window size $L = 4$ and step size $S = 4$ to count the total number of times over the entire conversation sequence that each pair of the speakers co-occurs within the sliding window. Assume that in the k_{th} group, it has N group members, we would have the symmetric matrix A with size of $N * N$ as our undirected graph. In other word, the value of the element $A_{ij} = A_{ji}$ is defined as follow,

$$A_{ij} = \begin{cases} normalize(\#win(i,j)) & , i \neq j \\ 1 & , i = j \end{cases}$$

where $\#win(i, j)$ is the number of sliding windows in the conversation speaker sequence that contain both member i and member j. We further define the edges set $\mathbf{e}$ includes N_{edges} non-self connected edges in a graph, where N_{edges} equals to combination of choosing 2 member from the group

size N, i.e., $N_{edges} = \frac{N!}{2!(N-2)!}$. The normalize method of the graph is then defined as follows:

$$normalize(e_n) = softmax(\frac{e_n}{max(\mathbf{e})}) \quad (1)$$

Hence, the summation of the normalized non-self-connected edges set $normalized(\mathbf{e})$ equals to 1. Intuitively, this conversational graph connects two members of the group with a higher edge weights when they tend to converse more often in vicinity of each other.

2.3.2. Conversational Graph Convolutional Network

Our proposed CGCN model contains 2 GCN layers and 1 feedforward linear layer (complete architecture is demonstrated in Figure 1). We first calculate

$$\widehat{A} = \widetilde{D}^{-\frac{1}{2}} A \widetilde{D}^{-\frac{1}{2}}$$

where D is a diagonal degree matrix, $\widetilde{D}_{ii} = \sum_{j=0} A_{ij}$. Then, we can pass our behavior features X through the following 2 layers of GCN to obtain the group structural representation Z,

$$Z = f(X, A) = mean(\widehat{A} ReLU(\widehat{A} X W^{[0]}) W^{[1]})$$

where $W^{[n]}$ is a layer-specific trainable weight matrix. Finally, we perform the prediction by mapping the group representation Z to 1-dim group performance score $Y_{predict}$ by

$$Y_{predict} = W^{[2]} Z + b^{[2]}$$

Since the dimension of $W^{[n]}$ only concerns the feature dimension from feature size of previous layer to next layer, it is able to be trained and updated across different size of groups. All of the parameters of our CGCN are updated by using the following loss function,

$$Loss = \sum_{batch} MSE(Y_{label}, Y_{predict}) + \lambda \sum_{n=0}^{2} (W^{[n]})^2$$

where we use the mean square error as loss function and add the l_2 regularization term with parameter λ to prevent overfitting. All the parameters are updated batch-wised.

3. EXPERIMENT SETUP AND RESULT

3.1. Experimental Setup

We evaluate our proposed method in improving the Absolute Group Score(AGS) prediction performance on ELEA corpus using the metric of MSE (Mean Square Error) and Pearson correlation. The Random Forests model is used to be our baseline since previous work [12] has showed that Random Forests is a well performing model on this prediction task. The number of estimators for tree is set at 20, which is the same setting as in [12]. In order to be consistent with recent works on prediction performance in this database [11], we transform the original score to the normalized AGS to the range from 1 and 10 using the MinMaxScaler.

3.1.1. Data Augmentation

We further include two external datasets of similar setting, UGI and GAP, as augmented data to be used inside each training fold when learning our proposed CGCN. We first provide a weak label for each sample of UGI and GAP by using a simple Random Forest trained on the training folds of ELEA dataset. Hence, in specifics, we have three different conditions of data augmentations: 1) using GAP to include extra 28 samples, 2) using UGI to include 22 data samples, and 3) using both GAP and UGI to include 50 samples. UGI dataset only has text modality, hence, it is only used for data augmentation when training text CGCN.

3.1.2. Detail Model Parameters

Our proposed CGCN is trained with 2 different parameters separately on speech and textual feature. For text CGCN model, we have our model parameters $W^{[0]}$, $W^{[1]}$ and $W^{[2]}$ with the matrix size as $[2400 * 32]$, $[32 * 32]$ and $[32 * 1]$. For the speech CGCN model, we have our model parameters $W^{[0]}$, $W^{[1]}$ and $W^{[2]}$ with size of $[76 * 32]$, $[32 * 16]$ and $[16 * 1]$. The parameter of CGCN on 2 different modalities are both trained with the ADAM [20] optimizer with learning rate equals to 0.01, batch size equals to 5 and the λ equals to 0.0005. We used leave-one-out cross validation with same parameter setting to evaluate our model. Each of the models are trained with 20 different random seeds, and the mean, std and the best results of models are presented.

3.2. Experimental Result and Analysis

Our proposed Speech CGCN method achieves an overall best performance at MSE 3.896 and 0.323 Pearson corr. on ELEA absolute group score prediction task. It outperforms random guessing baseline by 0.663 MSE and 1.323 Pearson corr. improvement. Our method also outperforms baseline tree-based method [12] on lexical modality by 0.211 MSE and 0.115 Pearson corr.. As the result shown in Table 1, our CGCN model generally improves by including the augmented data from UGI and GAP. This improvement can be observed on both speech and text modality. In contrast, although using Random Forest model shows a better performance when training directly only with the ELEA data sampels, it does not improve further when using augmented data.

For text modality, without data augmentation, Text CGCN achieves a moderate 4.703 MSE and 0.213 Pearson corr., which performs worse than the best performance 4.107 MSE and 0.208 corr. using Random Forest. However, when using GAP as augmented data, the best performance improves 0.7 MSE, and the average performance also improves 0.4 MSE. Similarly, when using UGI as the augmented data, the best performance improves 0.671 MSE, and the average performance also improves 0.4 MSE. When augmented with both datasets, we obtain improvement on average of 0.534 and decrease prediction variance (std goes from 0.435 to 0.239).

For audio modality, with only ELEA corpus, Speech CGCN can achieve 4.259 MSE and 0.198 Pearson corr.,

Model Type	Augmented Data	Text			Speech		
		MSE Range	Best MSE	Best Pearson	MSE Range	Best MSE	Best Pearson
Mean Guessing	-	4.559		-1.00	4.559		-1.00
Random Forests	-	4.896 ± 0.342	4.107	0.208	6.4 ± 0.383	5.677	-0.039
	GAP	5.131 ± 0.435	4.139	0.100	6.045 ± 0.362	5.339	-0.045
	UGI	4.758 ± 0.305	4.31	0.187	-	-	-
	GAP+UGI	5.301 ± 0.325	4.709	0.003	-	-	-
CGCN	-	5.33 ± 0.435	4.703	0.213	5.16 ± 0.448	4.259	0.198
	GAP	4.934 ± 0.342	4.07	0.229	4.726 ± 0.499	**3.896***	**0.323***
	UGI	4.886 ± 0.467	**4.034***	**0.282***	-	-	-
	GAP+UGI	4.8 ± 0.239	4.172	0.203	-	-	-

Table 1. Comparison between the performance of our proposed CGCN and the Random Forests on sentence level verbal and nonverbal feature. Noted that the MSE here is calculated based on the normalized AGS score (1 to 10) mentioned in section 3.1
.

Type	#	Feature	Best MSE
Avci & Aran[21]	40	sp. + turn-taking	71.3
Murray & Oertel[12]	28	sp.+text.	64.4
Our Method	29	sp.	66.7 (3.90)
Our Method	29	text	71.9 (4.03)
Our Method	29	sp. + text.	78.5(4.40)

Table 2. Comparison to the related work on ELEA using text and speech feature. We provide the number of data sample and the corresponding feature types for proper comparison. For our method, the equivalent AGS score in both normalized or non-normalized term is presented.

which is already better than the best performance 5.677 MSE and -0.039 corr. using Random Forest. By using augmented GAP dataset, the best performance achieved is 0.389, which is 0.363 improvement on MSE and the overall average performance also improves 0.424 on MSE. We also present the equivalent MSE score (without min-max scaling) in order to compare with two other methods on predicting group performance score on the same dataset (table 2). Our proposed model consistently outperforms previous model proposed by Avci et al. [21]. It also achieves similar performance compared to Murray and Oertel [12] though their textual features are extremely complex and provides a non-intuitive approach in analyzing small group structure.

3.3. Analysis of Graph Structure

Since our CGCN encodes the conversational structure between member behaviors within the group, we further analyze the relationship between graph structure and the group score. We evaluate the *imbalance* level of each group's structure by calculating the edges differences between balance structure and imbalance structure. In specific, the balance structure is defined with the edges weight $\frac{1}{N_{edges}}$. The structure imbalance level (SIL) measure for the group k are defined as follows:

$$SIL_k = \frac{\sum_{n=1}^{N_{edges}} (e_n - \frac{1}{N_{edges}})^2}{N_{edges}}$$

We observe an interesting insight that the SIL value has a Pearson corr. -0.391 with the AGS score in ELEA corpus. It indicates that the more imbalanced structure leads to a better group performance. In plain words, our analysis demonstrates that for those groups that have *key* members talk and/or facilitate the interaction would lead to a better overall group performance score; this results corroborate with past findings on the relationship of centralized structures and group performances [13, 22, 23] and further imply the connection to the emergent leadership [24].

4. CONCLUSION AND FUTURE WORK

Recently, computationally studying small group dynamics by using the member's behavior data has gradually become more important in understanding key factors in an efficient and effective decision-making process. In this work, we proposed an automatic prediction framework of Conversational Graph Convolutional Network that jointly models the explicit conversation structure as the graph with speech and language behavior features, we obtain the state-of-the-art prediction of group performance score on ELEA dataset. The usage of the graph-based deep learning network provides an intuitive mechanism in studying the dynamics between team members. In the future, we would like to include a recently collected larger corpus to evaluate the robustness of our results and further advance the framework with the multimodal fusion of behavior data.

5. REFERENCES

[1] Daniel Gatica-Perez, Oya Aran, and Dinesh Babu Jayagopi, "Analysis of small groups," *Social Signal Processing*, 2017.

[2] Tamar Gilad and Benjamin Gilad, "Smr forum: business intelligence-the quiet revolution," *Sloan Management Review (1986-1998)*, vol. 27, no. 4, pp. 53, 1986.

[3] Frances J Milliken and David A Vollrath, "Strategic decision-making tasks and group effectiveness: Insights from theory and research on small group performance," *Human Relations*, vol. 44, no. 12, pp. 1229–1253, 1991.

[4] Joseph Edward McGrath, *Social psychology: A brief introduction*, Holt, Rinehart and Winston, 1964.

[5] J Richard Hackman and Charles G Morris, "Group tasks, group interaction process, and group performance effectiveness: A review and proposed integration1," in *Advances in experimental social psychology*, vol. 8, pp. 45–99. Elsevier, 1975.

[6] Dairazalia Sanchez-Cortes, Oya Aran, Marianne Schmid Mast, and Daniel Gatica-Perez, "A nonverbal behavior approach to identify emergent leaders in small groups," *IEEE Transactions on Multimedia*, vol. 14, no. 3, pp. 816–832, 2012.

[7] McKenzie Braley and Gabriel Murray, "The group affect and performance (gap) corpus," in *Proceedings of the ICMI 2018 Workshop on Group Interaction Frontiers in Technology (GIFT)*, 2018.

[8] Indrani Bhattacharya, Michael Foley, Christine Ku, Ni Zhang, Tongtao Zhang, Cameron Mine, Manling Li, Heng Ji, Christoph Riedl, Brooke Foucault Welles, et al., "The unobtrusive group interaction (ugi) corpus," in *Proceedings of the 10th ACM Multimedia Systems Conference*. ACM, 2019, pp. 249–254.

[9] Yun-Shao Lin and Chi-Chun Lee, "Using interlocutor-modulated attention blstm to predict personality traits in small group interaction," in *Proceedings of the 2018 on International Conference on Multimodal Interaction*. ACM, 2018, pp. 163–169.

[10] Shun-Chang Zhong, Yun-Shao Lin, Chun-Min Chang, Yi-Ching Liu, and Chi-Chun Lee, "Predicting group performances using a personality composite-network architecture during collaborative task," *Proc. Interspeech 2019*, pp. 1676–1680, 2019.

[11] Uliyana Kubasova, Gabriel Murray, and McKenzie Braley, "Analyzing verbal and nonverbal features for predicting group performance," *arXiv preprint arXiv:1907.01369*, 2019.

[12] Gabriel Murray and Catharine Oertel, "Predicting group performance in task-based interaction," in *Proceedings of the 2018 on International Conference on Multimodal Interaction*. ACM, 2018, pp. 14–20.

[13] Alex Bavelas, "Communication patterns in task-oriented groups," *The Journal of the Acoustical Society of America*, vol. 22, no. 6, pp. 725–730, 1950.

[14] Lukas Zenk, Christoph Stadtfeld, and Florian Windhager, "How to analyze dynamic network patterns of high performing teams," *Procedia-Social and Behavioral Sciences*, vol. 2, no. 4, pp. 6418–6422, 2010.

[15] Aaron Schecter, Andrew Pilny, Alice Leung, Marshall Scott Poole, and Noshir Contractor, "Step by step: Capturing the dynamics of work team process through relational event sequences," *Journal of Organizational Behavior*, vol. 39, no. 9, pp. 1163–1181, 2018.

[16] Thomas N Kipf and Max Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[17] John Boaz Lee, Ryan Rossi, and Xiangnan Kong, "Graph classification using structural attention," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2018, pp. 1666–1674.

[18] 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*. ACM, 2010, pp. 1459–1462.

[19] Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler, "Skip-thought vectors," in *Advances in neural information processing systems*, 2015, pp. 3294–3302.

[20] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[21] Umut Avci and Oya Aran, "Predicting the performance in decision-making tasks: From individual cues to group interaction," *IEEE Transactions on Multimedia*, vol. 18, no. 4, pp. 643–658, 2016.

[22] Raymond T Sparrowe, Robert C Liden, Sandy J Wayne, and Maria L Kraimer, "Social networks and the performance of individuals and groups," *Academy of management journal*, vol. 44, no. 2, pp. 316–325, 2001.

[23] Thomas M Brown and Charles E Miller, "Communication networks in task-performing groups: Effects of task complexity, time pressure, and interpersonal dominance," *Small Group Research*, vol. 31, no. 2, pp. 131–157, 2000.

[24] Gita De Souza and Howard J Klein, "Emergent leadership in the group goal-setting process," *Small group research*, vol. 26, no. 4, pp. 475–496, 1995.