MAG+: AN EXTENDED MULTIMODAL ADAPTATION GATE FOR MULTIMODAL SENTIMENT ANALYSIS

Xianbing Zhao[1⋆], *Yixin Chen*[1⋆], *Wanting Li*[1], *Lei Gao*[2], *Buzhou Tang*[1,3]

[1] Intelligent Computing Research Center, Harbin Institute of Technology (Shenzhen)
[2] UCL Interaction Center, University College London
[3] Peng Cheng Laboratory, Shenzhen, China

ABSTRACT

Human multimodal sentiment analysis is a challenging task that devotes to extract and integrate information from multiple resources, such as language, acoustic and visual information. Recently, multimodal adaptation gate (MAG), an attachment to transformer-based pre-trained language representation models, such as BERT and XLNet, has shown state-of-the-art performance on multimodal sentiment analysis. MAG only uses a 1-layer network to fuse multimodal information directly, and does not pay attention to relationships among different modalities. In this paper, we propose an extended MAG, called MAG+, to reinforce multimodal fusion. MAG+ contains two modules: multi-layer MAGs with modality reinforcement (M3R) and Adaptive Layer Aggregation (ALA). In the MAG with modality reinforcement of M3R, each modality is reinforced by all other modalities via crossmodal attention at first, and then all modalities are fused via MAG. The ALA module leverages the multimodal representations at low and high levels as the final multimodal representation. Similar to MAG, MAG+ is also attached to BERT and XLNet. Experimental results on the two widely used datasets demonstrate the efficacy of our proposed MAG+.

Index Terms— Multimodal Sentiment Analysis, Multimodal Fusion, BERT

1. INTRODUCTION

Multimodal Sentiment Analysis(MSA) has become a significant research topic that aims to enable machines to recognize, interpret, and express emotion using time-series data from multiple resources, such as language, acoustic and visual information [1, 2, 3, 4, 5]. Multimodal fusion that extracts and integrates information from different modalities is one of the most important aspects of MSA, and a number of methods have been proposed to effectively leverage multimodal information [6, 7, 8]. Crossmodal attention and multimodal gating are two popular and effective mechanisms used for multimodal fusion, where crossmodal attention can reinforce a modality from other modalities by modeling their relationships, while multimodal gating can combine multimodal information [9, 10].

In recent years, some researchers have attempted to integrate a multimodal fusion module into pre-trained language representation models and achieved promising results on MSA. For example, Rahman et al.[4] proposed MAG-BERT by introducing a multimodal adaptation gate (MAG) as an attachment to BERT, where MAG is used to control how much information should be taken from each modality for fusion. Although MAG is effective, it only uses a 1-layer network to fuse multimodal information directly without considering the relationships among different modalities. Moreover, the representation capacity of 1-layer network may not be sufficient to model multimodal fusion [11].

Inspired by MAG-BERT [4], crossmodal attention [12] and fine-grained layer attention [11], we propose an extended MAG model (MAG+) for multimodal fusion. It uses Multi-layer MAGs with Modality Reinforcement (M3R) to reinforce each modality via crossmodal attention and fuses all modalities via dynamic gating mechanism at first, and then adopts an Adaptive Layer Aggregation module (ALA) to leverage the fused multimodal representations at multiple layers. Similar to MAG-BERT, we also integrate MAG+ into popular pre-trained language representation models, such as BERT [13] and XLNet [14], for MSA.

The main contributions of our work are:**1)** Proposing M3R to capture content-aware multimodal relationships at multiple levels. **2)** Introducing ALA to leverage multimodal fusion representations at multiple layers of M3R effectively. **3)** Taking MAG+ as an attachment to pre-trained language representation models, such as BERT and XLNet, for MSA with state-of-the-art performance.

2. PROPOSED APPROACH

2.1. Overview Architecture

Following MAG-BERT [4], in the aligned data, we attach the visual and acoustic information corresponding to each word embedding as input of the BERT next encoder

⋆ Equal Contribution.

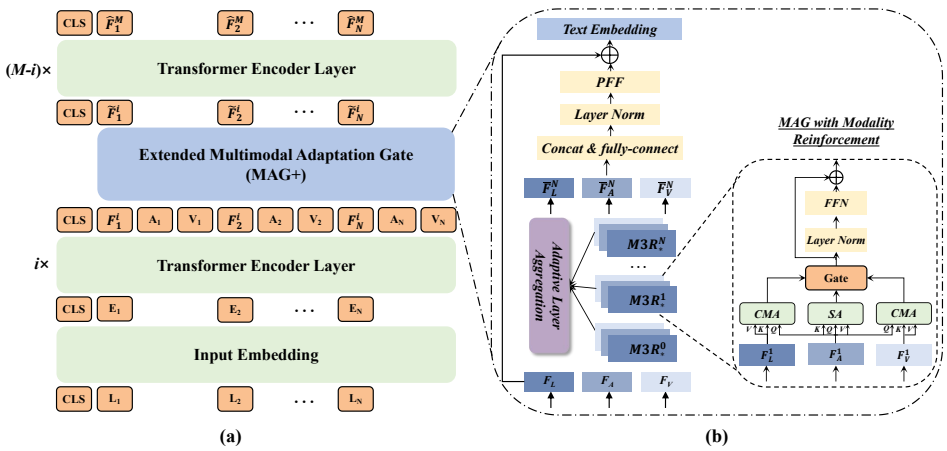

Fig. 1. (a) Pre-trained language model structure based on Transformer. BERT/XLnet has M=12 encoder layers, and integrates multimodal information on the i-th layer; (b) The network structure of MAG+ in detail.

layer. Figure1(a) illustrates the architecture of BERT/XLNet with MAG+ applied at i-th layer. Given a sentence with N words $L = \{L_1, L_2, ..., L_N\}$, where each word is represented by word embeddings, position embeddings and segmentation embeddings, we acquire an input representation $E = \{E_1, E_2, ..., E_N\}$. BERT has M=12 encoder layers, and the output of the i-th encoder layer is denoted as $F_L^i = \{F_{L1}^i, F_{L2}^i, ..., F_{LN}^i\}$, where the acoustical and visual information is defined as $F_A = \{F_{A1}, F_{A2}, ..., F_{AN}\}$, $F_V = \{F_{V1}, F_{V2}, ..., F_{VN}\}$ respectively.

The triple $(F_{Lj}^i, F_{Aj}, F_{Vj})$ is fed into the MAG+ to generate the updated language representation $\tilde{F}_{Aj}^i$. Figure1(b) shows the network architecture of MAG+. In the process of M3R, each modality passes information through the modal update layer and updates itself through interaction with other modalities. After stacking multiple M3R layers , we introduce the ALA mechanism at each layers to exploit low-level and high-level features. After M-i encoder layers, we obtain the final text representation $\hat{F}_L^M$.

2.2. Multi-layer MAGs with Modality Reinforcement(M3R)

Crossmodal Attention Based on crossmodal attention [12], given an input sequence $X_s \in \mathbf{R}^{T_s \times d_s}$ from the source modality and the sequence $X_t \in \mathbf{R}^{T_t \times d_t}$ from the target modality, where $s, t \in \{L, V, A\}$. We perform Queries from the target features while Keys and Values from the source features, $Q_t = X_t W_q$, $K_s = X_s W_k$ and $V_s = X_s W_v$ with $W_q \in \mathbf{R}^{d_t \times d_t}$, $W_k \in \mathbf{R}^{d_s \times d_t}$ and $W_v \in \mathbf{R}^{d_s \times d_t}$, respectively. Formally, cross-modal attention is defined as:

$$\begin{aligned} CMA_{s\to t}^{MH}(X_s, X_t) &= softmax(\frac{Q_t K_s^T}{\sqrt{d_k}} V_s) \\ &= softmax(\frac{X_t W_q W_k^T X_s^T}{\sqrt{d_k} X_s}) X_s W_v \end{aligned} \tag{1}$$

The target modality is reinforced by encouraging the model to attend to crossmodal interaction between elements.

Multi-layer MAGs with Modality Reinforcement The M3R module contains three modality update blocks, namely $M3R_L$, $M3R_V$ and $M3R_A$. Compared with the MAG, our M3R mainly relies on cross-modal attention and self-attention to update modality information. The modality of j-th M3R layer that needs to be updated is defined as F_*^j:

$$F_*^{j+1} = M3R_*^j(f_\alpha^j, f_*^j, f_\beta^j) \tag{2}$$

where $*, \alpha, \beta \in \{L, V, A\}$ and $* \neq \alpha \neq \beta$. f_α^j presents modality α in j-th modality reinforcement layer. When $j = 0$, we use the word embeddings F_L^i as f_L^0, similarity, $f_V = F_v, f_A = F_A$. When we update the modality $*$, the α, β denote the other two modalities. Concretely, $M3R_*^j$ reinforces F_*^{j+1} using three branches, one self-attention and two cross attentions:

$$f_{\alpha\to *}^j = CMA_{\alpha\to *}^{MH}(LN(f_\alpha^j), LN(f_*^j)) \tag{3}$$

$$f_{\beta\to *}^j = CMA_{\beta\to *}^{MH}(LN(f_\beta^j), LN(f_*^j)) \tag{4}$$

$$f_*^j = SA_{\beta\to *}^{MH}(LN(f_*^j)) \tag{5}$$

where $f_{\alpha\to *}^j$, $f_{\beta\to *}^j$, $f_*^j \in \mathbf{R}^{T_* \cdot d}$, CMA^{MH}, SA^{MH} and LN represent the multi-head cross-modal attention, multi-head self-attention operation and layer normalization respectively. Immediately, the three modalities are fused into F_*^j via an adaptive gated layer:

$$g_*^j = sigmoid(f_{\alpha\to *}^j \cdot W_{\alpha\to *}^j + f_{\beta\to *}^j \cdot W_{\beta\to *}^j + f_*^j \cdot W_*^j) \tag{6}$$

$$F_*^{j+1} = \sum_{c \in \{\alpha\to *, \beta\to *, *\}} g_c^j \odot f_c^j \tag{7}$$

where $W_{\alpha\to *}^j$, $W_{\beta\to *}^j$ and $W_*^j \in \mathbf{R}^{d \times d}$ are learnable parameters. To this end, we adopt a position-wise feedforward layer and layer normalization with skip connection to generate the final representations.

2.3. Adaptive Layer Aggregation

Further, different modal representations are encoded in a multi-level fashion, in which low-level and high-level features are taken into account. When modeling the hierarchical features, it is unclear whether the high-level features of the source modality are better than the low-level features. Instead, the increased modal complexity can reduce the performance. To overcome the limitation of the multi-level architecture, we propose an ALA to exploit both low- and high-level multimodal relationships instead of having just a single representation from the uppermost fusion layer. This is achieved through a learned Adaptive Aggregation, which weights multi-level contributions at each layer. Given a sequence from all M3R layers $\{F_*^0, F_*^1, ..., F_*^N\}$, we generate the final representation $\bar{F}_*$ with the layer-aware representation, which is formulated as:

$$\bar{F}_* = \sum_{n=0}^{N} \hat{w}_*^n \cdot F_*^n, * \in \{L, V, A\} \tag{8}$$

Where $\hat{w}_*^n$ denotes the element of $n-$th layer in the learnable attention weights:

$$\hat{w}_*^n = \frac{exp(w_*^n)}{\sum_{n=0}^{N} exp(w_*^n)}, * \in \{L, V, A\} \tag{9}$$

When $n = 0$, we use the word embeddings F_{emb} without position embeddings as F^0, which has been empirically proved effective.

2.4. Deep Multimodal Fusion

Finally, multimodal feature for each modality after N rounds of iteration update are concatenated and fed to the self-attention layer. Next, we use the fully connected layer to generate the final representation:

$$\tilde{F} = FC(SA^{MH}(CONCAT(\bar{F}_L^N, \bar{F}_A^N, \bar{F}_V^N))) \tag{10}$$

where FC denotes fully connected layer, SA^{MH}means self-attention with multi-head and CONCAT represents the vector in the combined feature dimension.

3. EXPERIMENTS

3.1. Datasets, Metrics and Settings

Datasets We evaluate model on the CMU-MOSI and CMU-MOSEI datasets. 1)*CMU-MOSI* [15]: contains a total of 93 videos and 89 speakers. 2)*CMU-MOSEI* [16]: contains 22,856 samples of movie review video clips. The dataset has real-valued high agreement sentiment intensity annotations from -3(strongly negative) to 3(strongly positive).

Metrics Following the previous works [4, 5], We perform MAG+ on two different tasks, classification and regression. In the regression problem, we will report the mean square error (MAE) and the correlation coefficient (CC). Besides, the predicted score on the regression task is converted into a binary classification result to obtain the model's binary classification accuracy (Acc-2) and F1 score.

Settings MAG+ is integrated into the pre-trained language model BERT/XLNet. For fair comparison, our model has the same configuration as the three most advanced models(MISA [7], MAG-BERT [4] and Self-MM [5]). That is, the dropout rate is set to 0.5, the ADAM optimizer with a learning rate of 0.0005 and a batch size of 48 is adopted, and the number of epochs is set to 60. The learning rate is initialized to 1e-5, and the encoder layer of BERT i is set to 1. Besides, we run experiments five times and report the average and some best performances.

3.2. Results and Analysis

We compare our approach with a wide range of state-of-the-art multimodal sentiment analysis models, i.e.,TFN [1], LMF [2], LMF [2], MFM [6], MulT [3], MISA [7], MTAG [19]. As well as the BERT-based models, MAG-BERT [4] and Self-MM [5]. As shown in Table 1, our MAG+ outperforms the state-of-the-art models across all metrics on both CMU-MOSI and CMU-MOSEI datasets. The improvement of MAG+ demonstrates the validity of multimodal reinforcement with an adaptive layer aggregation. More encouragingly, MAG+ based on XLNet sets achieve an accuracy of 86.30% and making the absolute improvement over the baseline MAG-XLNet by 0.7% on the CMU-MOSEI dataset, demonstrating the effectiveness and the compatibility of our proposed approach.

3.3. Quantitative Analysis

We conduct the quantitative analysis to investigate the contribution of each component in our MAG+.

Effect of Fine-grained Multimodal Reinforcement. Comparing the results of baselines, we can find that incorporating the M3R module substantially boosts the performance of BERT-based models, e.g., 83.90%→86.13% in accuracy score and 83.90%→86.09% in F1 score on CMU-MOSEI dataset(see Table1). We hypothesize that this performance gain may due to that M3R can reinforce modalities interaction, which alleviate the unaligned problem.

Effect of Adaptive Layer Aggregation. As shown in the last two parts of Table2, it is clear that our ALA successfully boosts the performance. As the number of layers increases, the "BERT+M3R" model performs worse, whose accuracy decreases from 86.13% to 85.69%. We apply ALA to the "BERT+M3R", which leads to the excellent performance. More encouragingly, the "MAG+" with 6 M3R layers even achieve an accuracy of 86.55% on CMU-MOSEI dataset.

Model	CMU-MOSI				CMU-MOSEI			
	Acc-2 ↑	F1-Score ↑	MAE ↓	CC ↑	Acc-2 ↑	F1-Score ↑	MAE ↓	CC ↑
TFN#[1]	-/80.8	-/80.7	0.901	0.698	-/82.5	-/82.1	0.593	0.700
LMF#[2]	-/82.4	-/82.4	0.917	0.695	-/82.0	-/82.1	0.623	0.677
MFN[17]	77.4/-	77.3/-	0.965	0.632	76.0/-	76.0/-	-	-
RAVEN[18]	78.0/-	76.6/-	0.915	0.691	79.1/-	79.5/-	0.614	0.662
MFM[6]	-/81.7	-/81.6	0.877	0.706	-/84.4	-/84.3	0.568	0.717
MulT[3]	81.5/84.1	80.6/83.9	0.861	0.711	-/82.5	-/82.3	0.580	0.703
MISA[7]	81.8/83.4	81.7/83.6	0.783	0.761	83.6/85.5	83.8/85.3	0.555	0.756
MTAG#[19]	-/82.3	-/82.1	0.866	0.722	-	-	-	-
PMR[20]	83.6	83.4	-	-	-/83.3	-/82.6	-	-
Self-MM#[5]	84.0/86.0	84.4/85.9	0.713	0.798	82.8/85.2	82.5/85.3	**0.530**	0.765
MAG-BERT[4]	84.2/86.1	84.1/86.0	0.712	0.796	84.7/-	84.5/-	-	-
MAG-XLNet[4]	85.7/87.9	85.6/**87.9**	**0.675**	0.821	85.6/-	85.7/-	-	-
LSDR (BERT)	85.9/86.4	85.2/86.3	0.702	0.813	85.8/86.6	85.8/86.5	0.583	0.797
LSDR (XLNet)	**87.6/87.9**	**86.8**/87.4	0.684	**0.832**	**86.3/86.7**	**86.2/86.6**	0.579	**0.800**

Table 1. Performance of the proposed MAG+ and other state-of-the-art methods on the CMU-MOSI and CMU-MOSEI datasets. The average '/' best results of 5 runs are reported. # indicates the results on unaligned data.

Model	Acc-2	F1-Score
BERT	83.90	83.90
BERT+M3R	86.13	86.09
BERT+M3R+ALA(MAG+)	**86.55**	**86.54**
BERT+M3R (w/ 1 layer)	86.13	86.09
BERT+M3R (w/ 3 layers)	85.99	85.90
BERT+M3R (w/ 6 layers)	85.69	85.61
MAG+ (w/ 1 layer)	85.87	85.82
MAG+ (w/ 3 layers)	86.33	86.27
MAG+ (w/ 6 layers)	86.55	86.54

Table 2. Ablation analysis of MAG+(BERT) on the CMU-MOSEI dataset.

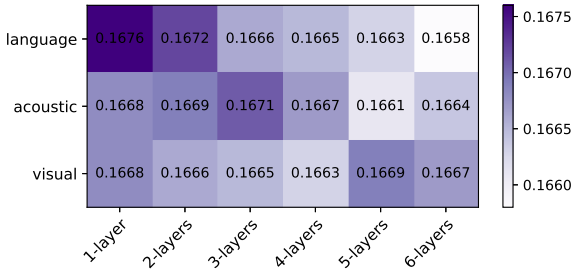

Fig. 2. Attention distribution that each M3R layers(x-axis) attending to different modalities(y-axis).

3.4. Qualitative results and visualization

In this section, we investigate which level of features at M3R layers contribute to the model performance via carefully designed experiment. We first visualize the learnable layer attention distribution in Figure 2. Generally, a higher weight means more contribution of an modality of a M3R layer to the final multimodal representations. Among three modalities, almost all modality channels focus on the low-level M3R layers, especially the language representations. In contrast, the acoustic representation pay more attention to the intermediate layers while visual information require more multimodal interaction, which are generally embedded in the high-level layers. Intuitively, multimodal features generation based on the source linguistic representations, supported with visual and acoustic representations.

4. CONCLUSIONS AND FUTURE WORK

In this paper, to reinforce multimodal fusion, we propose an extended multimodal adaption gate model MAG+ for human sentiment analysis, which contain two modules M3R and ALA. The M3R module can encourage a more efficient multimodal fusion via crossmodal attention and dynamic gating mechanism. The ALA module is proposed to aggregate fine-grained modality reinforcement layers, which exploits low-level and high-level features. This allows us to qualitatively and quantitatively evaluate the contribution of each M3R layers. The experimental results over different benchmarks clearly demonstrate that our approach obtains better results than the existing state-of-the-art works.

5. REFERENCES

[1] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency, "Tensor fusion network for multimodal sentiment analysis," *arXiv preprint arXiv:1707.07250*, 2017.

[2] Zhun Liu, Ying Shen, and Varun Bharadhwaj et al. Lakshminarasimhan, "Efficient low-rank multimodal fusion with modality-specific factors," *arXiv preprint arXiv:1806.00064*, 2018.

[3] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, and et al. Kolter, "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the conference. Association for Computational Linguistics. Meeting*. NIH Public Access, 2019, vol. 2019, p. 6558.

[4] Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, and et al. Zadeh, "Integrating multimodal information in large pretrained transformers," in *Proceedings of the conference. Association for Computational Linguistics. Meeting*. NIH Public Access, 2020, vol. 2020, p. 2359.

[5] Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu, "Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis," *arXiv preprint arXiv:2102.04830*, 2021.

[6] Yao-Hung Hubert Tsai, Paul Pu Liang, and Amir et al. Zadeh, "Learning factorized multimodal representations," *arXiv preprint arXiv:1806.06176*, 2018.

[7] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria, "Misa: Modality-invariant and-specific representations for multimodal sentiment analysis," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 1122–1131.

[8] Vishaal Udandarao, Abhishek Maiti, and Srivatsav et al., "Cobra: Contrastive bi-modal representation algorithm," *arXiv preprint arXiv:2005.03687*, 2020.

[9] Wei Han, Hui Chen, Alexander Gelbukh, and Zadeh et al., "Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis," *arXiv preprint arXiv:2107.13669*, 2021.

[10] Ayush Kumar and Jithendra Vepa, "Gated mechanism for attention based multi modal sentiment analysis," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 4477–4481.

[11] Xuebo Liu, Longyue Wang, Derek F Wong, and Ding et al., "Understanding and improving encoder layer fusion in sequence-to-sequence learning," *arXiv preprint arXiv:2012.14768*, 2020.

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, and Uszkoreit et al., "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[14] Zhilin Yang, Zihang Dai, Yiming Yang, and Carbonell et al., "Xlnet: Generalized autoregressive pretraining for language understanding," *Advances in neural information processing systems*, vol. 32, 2019.

[15] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency, "Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos," *arXiv preprint arXiv:1606.06259*, 2016.

[16] Amir Zadeh, Paul Pu Liang, Navonil Mazumder, and et al. Poria, "Memory fusion network for multi-view sequential learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018, vol. 32.

[17] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency, "Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018a, pp. 2236–2246.

[18] Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency, "Words can shift: Dynamically adjusting word representations using nonverbal behaviors," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, vol. 33, pp. 7216–7223.

[19] Jianing Yang, Yongxin Wang, Ruitao Yi, and Zhu et al., "Mtag: Modal-temporal attention graph for unaligned human multimodal language sequences," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021, pp. 1009–1021.

[20] Fengmao Lv, Xiang Chen, Yanyong Huang, Lixin Duan, and Guosheng Lin, "Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2554–2562.