GATED MECHANISM FOR ATTENTION BASED MULTIMODAL SENTIMENT ANALYSIS

Ayush Kumar and Jithendra Vepa

Observe.AI

ABSTRACT

Multimodal sentiment analysis has recently gained popularity because of its relevance to social media posts, customer service calls and video blogs. In this paper, we address three aspects of multimodal sentiment analysis; 1. Cross modal interaction learning, i.e. how multiple modalities contribute to the sentiment, 2. Learning long-term dependencies in multimodal interactions and 3. Fusion of unimodal and cross modal cues. Out of these three, we find that learning cross modal interactions is beneficial for this problem. We perform experiments on two benchmark datasets, CMU Multimodal Opinion level Sentiment Intensity (CMU-MOSI) and CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) corpus. Our approach on both these tasks yields accuracies of 83.9% and 81.1% respectively, which is 1.6% and 1.34% absolute improvement over current state-of-the-art.

Index Terms— sentiment analysis, multimodal fusion, gated mechanism

1. INTRODUCTION

Sentiment analysis has been one of the widely studied problems in spoken language understanding that aims to determine the opinion of the speaker towards a product, topic or event. With the proliferation of social media platforms such as Facebook, Whatsapp, Instagram and YouTube, huge volume of data is being generated in the forms of podcasts, vlogs, interviews, commentary etc. Multimodal data offer parallel acoustic (vocal expressions like intensity, pitch) and visual cues (facial expressions, gestures) along with the textual information (spoken words), which in particular, provides advanced understanding of affective behavior.

Several approaches have been proposed for multimodal sentiment analysis that attempt to effectively leverage multimodal information. These are categorised into three types, 1. Methods that learn the modalities independently and fuse the output of modality specific representations [1, 2], 2. Methods that jointly learn the interactions between two or three modalities [3, 4], and 3. Methods that explicitly learn contributions from these unimodal and cross modal cues, typically using attention based techniques [5, 6, 7, 8, 9, 10].

Most of the existing approaches propose either fusion at different granularities [3, 9] or use a cross interaction block that couple the features from different modalities [10, 6]. Combining features from different modalities is necessary as they offer parallel information for same source and help in disambiguation of affective behavior. For example, while uttering sarcastic statements, the speaker shows a distinct intonation which aids in determining the correct sentiment of the speaker. It is imperative that all modalities in multimodal sentiment analysis do not contribute equally, rather act as cues to reinforce or rectify the information from the other modalities. This is more evident in the case of imperfect modalities, for example; errors in automatic speech recognition might corrupt the textual information, or poor recording distort the acoustic information, or improper lighting might negatively impact visual features.

Therefore, to learn better cross modal information, we introduce novel *conditional gating mechanism* to modulate the information during cross interactions. Proposed gating mechanism selectively learns the relative importance of different modalities based on the linguistic information, tone of the speaker and facial expressions of an utterance.

Furthermore, to capture long term dependencies across the utterances in the video, we apply a self attention layer on unimodal contextual representations. The major advantage of self attention is that it induces direct interaction between any two utterances and hence offers unrestricted information flow in the network. Finally, we feed the self attended unimodal contextual representations and the gated cross interaction representations to a recurrent layer to obtain *deep multimodal contextual feature vectors* for each utterance.

The main contributions of our proposed approach are: **1)** Learnable gating mechanism to control information flow during cross interaction, **2)** Self attended contextual representation to capture long term dependencies, and **3)** Recurrent layer based fusion of self and gated cross fusion feature vectors to obtain modality specific deep multimodal feature vectors.

2. PROPOSED APPROACH

In our proposed model, we aim to learn the interaction between different modalities controlled by learnable gates. Figure 1 shows the overall architecture of the system outlining the main components in the model: *contextual utterance representation, self attention, cross attention, gating mechanism*

Authorized licensed use limited to: SHANGHAI UNIVERSITY. Downloaded on October 08,2021 at 06:39:05 UTC from IEEE Xplore. Restrictions apply.

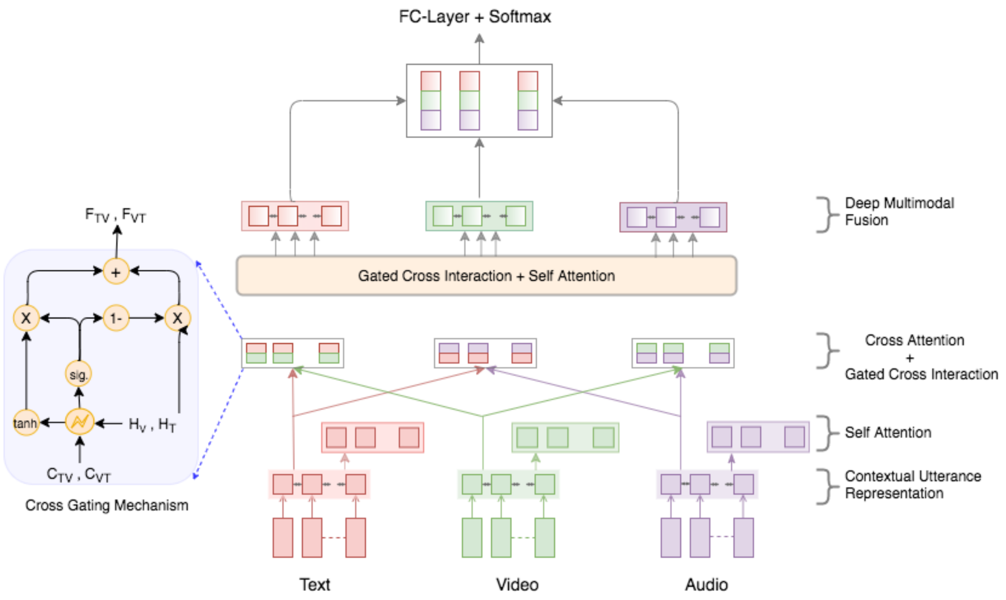

Fig. 1. Architectural diagram of the proposed approach.

for cross interaction and, *deep multimodal fusion*.

2.1. Contextual Utterance Representation 语境话语表征

We feed a sequence of utterance level features for each modality to a separate Bi-GRU [11] and obtain modality specific contextual utterance representation, H. Formally 形式上, contextual utterance representations ($H_T \in R^{u\times d}$) for a sequence of utterances ($U_1, U_2, ..., U_u$) for a *Text* modality can be defined as:

$$H_T = Bi\text{-}GRU(U_1, U_2, ..., U_u) \quad (1)$$

下标 Subscript T denotes *Text* modality, A and V represent *Audio* and *Video* modalities respectively.

2.2. Self Attention

In order to capture long term dependencies, a bilinear 双线性 attention [12] based self matching layer 自匹配层 on contextual utterance representations is employed. Since we have sequences of up to 100 utterances in a video, self attention allows us to capture the long context. For a *Text* modality, self attention can be represented as:

$$M_T = H_T W H_T^T,\ M_T \in R^{u\times u} \quad (2a)$$

$$A_T(i,) = softmax(M_{T_{i,}}) \quad (2b)$$

$$S_T = A_T.H_T,\ S_T \in R^{u\times d} \quad (2c)$$

Equation *2a* computes the self matching matrix; $W \in R^{d\times d}$ being a trainable matrix, Equation *2b* computes self-attention scores for utterance, U_i and finally Equation *2c* generates the self attended utterance representations.

2.3. Cross Attention

Multimodal sentiment analysis provides an opportunity to learn interactions between different modalities. Similar to approaches mentioned for intermodal attention in Ghosal et al [10], we propose a method to learn cross-interaction vectors. For a pair of Text (H_T) and Video (H_V) modalities, co-attention matrix ($M_{TV} \in R^{u\times u}$) can be defined as:

$$M_{TV} = H_T W H_V^T;\ W \in R^{d\times d} \quad (3)$$

Cross attentive representations of Text ($C_{VT} \in R^{u\times d}$) and Video ($C_{TV} \in R^{u\times d}$) can be represented as:

$$A_{TV}(i:) = softmax(M_{TV_{i:}}) \quad (4a)$$

$$A_{VT}(:j) = softmax(M_{TV_{:j}}) \quad (4b)$$

$$C_{VT} = A_{VT}.H_T, C_{TV} = A_{TV}.H_V \quad (4c)$$

2.4. Gating Mechanism for Cross Interaction

As much as there is an opportunity to leverage cross modal interactions, it brings in challenges of fusing imperfect modalities. To overcome the noise present in individual modalities, we propose a gating mechanism to selectively learn the cross fused vector [13, 14]. The gated cross fused vector ($F_{PQ} \in R^{u\times d}$) for a pair of Text-Video modalities can be obtained as:

$$F_{VT} = fusion(C_{VT}, H_T) \quad (5a)$$

$$F_{TV} = fusion(C_{TV}, H_V) \quad (5b)$$

We define fusion kernel 核心 $fusion(\cdot,\cdot)$ to be gated combination of cross interaction and contextual representation. Cross interaction, $X(P, Q)$, is a non-linear transformation 变换 on cross attended vector (P) and contextual representation (Q). Gating

Authorized licensed use limited to: SHANGHAI UNIVERSITY. Downloaded on October 08,2021 at 06:39:05 UTC from IEEE Xplore. Restrictions apply.

Sl. No.	Model	CMU-MOSI	CMU-MOSEI
B1	Contextual Unimodal (**Unimodal Baseline**)	80.57	78.58
B2	B1 + Self Attention	81.11	79.12
B3	Cross Interaction w/o gating (**Bimodal Baseline**)	81.91	80.00
B4	Cross Interaction w/ gating	82.91	80.59
B5	B2 + B4 w/o deep multimodal fusion	83.37	80.88
B6	**Proposed: B2 + B4 w/ multimodal fusion**	**83.91**	**81.14**

Table 1. Comparison of performance of each step in the proposed model. Accuracy values are mentioned in the table

function, $G(P, Q)$, modulates 调节 the information to be passed from cross interaction to next layer.

$$X(P, Q) = tanh(W_F.[P, Q, P\text{-}Q, P \circ Q] + b_F) \quad (6a)$$

$$G(P, Q) = \sigma(W_G^T.[P, Q, P\text{-}Q, P \circ Q] + b_G) \quad (6b)$$

$$F_{PQ} = G(P, Q).X(P, Q) + (1 - G(P, Q)).Q \quad (6c)$$

where, W_F, b_F, W_G^T, b_G are trainable parameters and $\circ$ represents element wise product.

If features from participating modalities are complementary 补充的, gating function favours 有利于 cross interaction and hence would have higher value. On the other hand, if the features from participating modalities is not rich enough or unimodal representation is self-sufficient 自给自足, the gating function would favor contextual representation and hence would have lower value.

2.5. Deep Multimodal Fusion

To aggregate 聚合 the information from the self and gated cross interactions, we use a Bi-GRU layer to learn deep multimodal feature vector for each modality.

$$Deep_T = Bi\text{-}GRU(S_T, F_{VT}, F_{AT}) \quad (7)$$

Finally, deep multimodal feature vector for each modality for an utterance is concatenated 连接 and fed to the prediction layer containing a fully connected layer followed 然后 by softmax layer for final classification.

3. EXPERIMENTS

3.1. Dataset

We evaluated our system on two standard multimodal sentiment analysis datasets from CMU multimodal SDK[1] [6], 1) *CMU-MOSI*: CMU Multimodal Opinion level Sentiment Intensity [15] and; 2) *CMU-MOSEI*: CMU Multimodal Opinion Sentiment and Emotion Intensity [7]. To compare with the existing approaches, we report results on the binary sentiment classification setup, where *values* ≥ 0 signify positive sentiments and *values* < 0 signify negative sentiments. There are 1284, 229 and 686 utterances in the training, validation and the test set for CMU-MOSI dataset while CMU-MOSEI has 16216, 1835 & 4625 utterances in training, validation & test set respectively.

3.2. Implementation Details

In our experiments, we used same features mentioned in Ghosal et al [10]. Specifically, for CMU-MOSEI dataset, we used Glove embeddings for word features, Facets [2] for visual features and CovaRep [16] for acoustic features. For MOSI dataset, we used output of a CNN network for utterance level features, 3D CNN features for visual and openSMILE [17] for acoustic features.

We trained Bi-GRUs with hidden size of 100 for CMU-MOSI dataset and 200 for CMU-MOSEI dataset, also used a dropout of 0.4 for regularization and ReLU activation [18] in dense layers. We used Adam optimizer [19] with a learning rate 0.0005 and a batch size 16 for CMU-MOSI and 32 for CMU-MOSEI dataset and, finally train the network for 75 epochs.

3.3. Results and Analysis

3.3.1. Baselines and Ablation Study

We carried out several experiments to analyze the contribution of the proposed approach (Table 1). We frame a unimodal (**B1**) and a bimodal baseline (**B3**) to compare the impacts of self attention (**B2**) and gating mechanism (**B4**). Further we also evaluate the model with deep multimodal fusion (**B6**). We see that by using self attention, the performance of model improves by 0.54% on MOSI and MOSEI corpora. Gating mechanism improves the accuracy by absolute 1% on MOSI while multimodal fusion adds additional 0.54% and 0.26% accuracy on two corpora.

The gains in performance over these baselines clearly validates our main hypotheses that attention focused gating selectively learns the noise-robust interactions between different modalities and self attention is required to exploit long term context dependencies present in the video. Finally, deep multimodal feature representations learned using self attended representations and gated cross interactions provides additional gains in the accuracies.

[1] https://github.com/A2Zadeh/CMU-MultimodalSDK

[2] https://https://pair-code.github.io/facets/

Authorized licensed use limited to: SHANGHAI UNIVERSITY. Downloaded on October 08,2021 at 06:39:05 UTC from IEEE Xplore. Restrictions apply.

CMU-MOSI			CMU-MOSEI		
Approach	**Accuracy**	**F1-Score**	**Approach**	**Accuracy**	**F1-Score**
Zadeh et al [3]	77.1	79.1	Zadeh et al [8]*	76.0	76.0
Chen et al [5]	76.5	73.4	Zadeh et al [7]	76.9	77.0
Georgiou et al [9]	76.9	76.9	Poria et al [2]	77.64	-
Ghosal et al [10]	82.31	80.69	Ghosal et al [10]	79.80	-
Sun et al [4]	80.6	80.57	Sun et al [4]†	(83.62)†	(83.75)†
Proposed Approach	**83.91**	**81.17**	**Proposed Approach**	**81.14 / (85.27)**†	**78.53 / (84.08)**†

Table 2. Comparative results on CMU-MOSI and CMU-MOSEI multimodal sentiment analysis. (*) results are taken from Zadeh et al [7], (†) results are obtained on CMU-MOSEI dataset after excluding the utterances with sentiment score of 0. We mention the results of proposed model with this setup in the parenthesis.

Utterance	Gold Label	Predicted Label	Remark
I really really loved it	Pos.	Pos.	S_{T_u} = 0.91, which justifies that text is self sufficient cross-interaction score for this utterance is 0.43
i was just thinking about um how its the performances in it were sort of over overlooked at the academy awards	Pos.	Pos.	Text modality suggests it to be a negative sentiment. Contribution of *T-A* and *T-V* cross-interaction is less (0.12 and 0.05). S_{V_u} = 0.75 and S_{A_u} = 0.67 suggests that V, A modality drives the prediction.
maybe only 5 jokes made me laugh	Neg.	Neg.	All three modalities are correlated in this utterance of the video, evident by cross-interaction contributions of *T-A*, *A-V* and *T-V* to be 0.78, 0.69 and 0.83 respectively.
oh oh my gosh i was blown away	Pos.	Pos.	Audio (S_{A_u} = 0.62) and video (S_{V_u} = 0.49) contributes in all cross-interactions (0.74) to reinforce their learning.

Table 3. Qualitative analysis of the proposed model. *T, A, V* refers to text, audio and video respectively. S_{M_u} denotes self attention score for utterance *u* in modality *M*. Cross-interaction score are average values of gate $G(P, Q)$ for a pair of modalities *P, Q*.

3.3.2. Benchmarking

To comprehensively compare our method, we list several baselines for multimodal sentiment analysis. Tensor fusion network [3] uses a 3-fold cartesian product on unimodal embeddings; Context-dependent sentiment analysis [2] learns context dependent multimodal feature representations; Memory fusion network (MFN) [8] proposes a 3-step architecture for multi-view sequential learning using attention network and gated memory; Graph-MFN [7] replaces attention network in MFN with dynamic fusion graph to learn modal dynamics; Gated multimodal-embedding with temporal attention [5] performs word level modality fusion using gating; Hierarchical fusion [9] performs 3-step fusion at word, sentence and high level for sentiment classification; Deep canonical correlation analysis (DCCA) based multi-modal embeddings [4]; and Contextual inter-modal attention based network [10] that proposes a multi-modal attention framework to learn joint-association between multiple modalities & utterances.

In Table 2, we present the comparison of our proposed method with other state-of-the-art approaches. Our proposed method outperforms the state-of the-art by 1.6% (absolute) points for CMU-MOSI corpus and 1.34% points for CMU-MOSEI corpus. Qualitative analysis of our results is presented in Table 3 with a few examples. The analysis demonstrates the effectiveness of the model in selectively attending to the relevant modalities by adjusting the modality specific scores (self attention) as well as cross interactions.

4. CONCLUSIONS AND FUTURE WORK

In this paper, we propose an approach to improve the multimodal sentiment analysis using self attention to capture long term context and gating mechanism to selectively learn cross attended features. The gating function emphasize on cross interactions when unimodal information is insufficient to decide the sentiment while it assigns lower weightage to cross modal information when unimodal information is sufficient to predict the sentiment. Evaluations on two well known benchmark datasets (CMU-MOSI and CMU-MOSEI) show that our proposed method is significantly better than the state-of-the-art. In future, we will extend the proposed techniques for real world data, e.g. call center customer conversations, where noise in both Text and Audio modalities is high due to poor audio quality, thus resulting in lower speech recognition accuracies.

Authorized licensed use limited to: SHANGHAI UNIVERSITY. Downloaded on October 08,2021 at 06:39:05 UTC from IEEE Xplore. Restrictions apply.

5. REFERENCES

[1] M. Wöllmer, F. Weninger, T. Knaup, B.W. Schuller, C. Sun, K. Sagae, and L.P. Morency, "Youtube movie reviews: Sentiment analysis in an audio-visual context," *IEEE Intelligent Systems*, vol. 28, no. 3, pp. 46–53, 2013.

[2] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency, "Context-dependent sentiment analysis in user-generated videos," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL*, 2017, pp. 873–883.

[3] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP*, 2017, pp. 1103–1114.

[4] Z. Sun, P.K. Sarma, W. Sethares, and E.P. Bucy, "Multimodal sentiment analysis using deep canonical correlation analysis," *Proc. Interspeech 2019*, pp. 1323–1327, 2019.

[5] M. Chen, S. Wang, P.P. Liang, T. Baltrusaitis, A. Zadeh, and L.P. Morency, "Multimodal sentiment analysis with word-level fusion and reinforcement learning," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction, ICMI*, 2017, pp. 163–171.

[6] A. Zadeh, P.P. Liang, S. Poria, P. Vij, E. Cambria, and L.P. Morency, "Multi-attention recurrent network for human communication comprehension," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[7] A.B. Zadeh, P.P. Liang, S. Poria, E. Cambria, and L.P. 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, ACL*, 2018b, pp. 2236–2246.

[8] A. Zadeh, P.P. Liang, N. Mazumder, S. Poria, E. Cambria, and L.P. Morency, "Memory fusion network for multi-view sequential learning," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[9] E. Georgiou, C. Papaioannou, and A. Potamianos, "Deep hierarchical fusion with application in sentiment analysis," *Proc. Interspeech 2019*, pp. 1646–1650, 2019.

[10] D. Ghosal, M.S. Akhtar, D. Chauhan, S. Poria, A. Ekbal, and P. Bhattacharyya, "Contextual inter-modal attention for multi-modal sentiment analysis," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 3454–3466.

[11] K. Cho, B. Merrienboer, Ç Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP*, 2014, pp. 1724–1734.

[12] T. Luong, H. Pham, and C.D. Manning, "Effective approaches to attention-based neural machine translation," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 1412–1421.

[13] W. Wang, C. Wu, and M. Yan, "Multi-granularity hierarchical attention fusion networks for reading comprehension and question answering," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL*, 2018b, pp. 1705–1714.

[14] Y. Gong and S.R. Bowman, "Ruminating reader: Reasoning with gated multi-hop attention," in *Proceedings of the Workshop on Machine Reading for Question Answering@ACL*, 2018, pp. 1–11.

[15] A. Zadeh, R. Zellers, E. Pincus, and L.P. Morency, "MOSI: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos," *CoRR*, vol. abs/1606.06259, 2016.

[16] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "COVAREP - A collaborative voice analysis repository for speech technologies," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2014, pp. 960–964.

[17] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 835–838.

[18] V. Nair and G.E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning (ICML)*, 2010, pp. 807–814.

[19] D.P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR*, 2015.

Authorized licensed use limited to: SHANGHAI UNIVERSITY. Downloaded on October 08,2021 at 06:39:05 UTC from IEEE Xplore. Restrictions apply.