Multi-Attention Multimodal Sentiment Analysis

Taeyong Kim
Hyundai Robotics
Yongin, South Korea
taeyong.kim@hyundai-robotics.com

Bowon Lee
Department of Electronic Engineering, Inha University
Incheon, South Korea
bowon.lee@inha.ac.kr

ABSTRACT

Sentiment analysis plays an important role in natural-language processing. It has been performed on multimodal data including text, audio, and video. Previously conducted research does not make full utilization of such heterogeneous data. In this study , we propose a model of Multi-Attention Recurrent Neural Network (MA-RNN) for performing sentiment analysis on multimodal data. The proposed network consists of two attention layers and a Bidirectional Gated Recurrent Neural Network (BiGRU). The first attention layer is used for data fusion and dimensionality reduction, and the second attention layer is used for the augmentation of BiGRU to capture key parts of the contextual information among utterances. Experiments on multimodal sentiment analysis indicate that our proposed model achieves the state-of-the-art performance of 84.31 % accuracy on the Multimodal Corpus of Sentiment Intensity and Subjectivity Analysis (CMU-MOSI) dataset. Furthermore, an ablation study is conducted to evaluate the contributions of different components of the network. We believe that our findings of this study may also offer helpful insights into the design of models using multimodal data.

CCS CONCEPTS

• **Information systems → Multimedia and multimodal retrieval**; *Multimedia information systems.*

KEYWORDS

Multimodal Machine Learning, Deep Learning, Sentimental Analysis on Multimedia

ACM Reference Format:
Taeyong Kim and Bowon Lee. 2020. Multi-Attention Multimodal Sentiment Analysis. In *Dublin, Ireland '20: ACM ICMR, June 08–11, 2020, Dublin, Ireland.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/1122445.1122456

ICMR '20, June 08–11, 2020, Dublin, Ireland

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

For understanding inherent meanings in conversations, it is important to recognize *what* was said and *how* it was said [26]. Thus, studies have been conducted in the field of natural-language understanding to recognize the emotions and sentiments of a speaker.

Many models have been developed based on unimodal data, such as text [1, 18] or speech [15, 27]. However, in real world situations, humans recognize sentiments based on multimodal ques, such as facial expression and the tone of the voice.

Recently, sentiment recognition has been explored with multimodal data, such as audio and video. To tackle this task, there are numerous challenges that need to be overcome. For example, the sentiment expression of a person can vary extensively. Some people express their feelings solely through their voice, but others may convey them more visually. Furthermore, fusing multimodal data from audio and video requires careful consideration. Attempts have been made to accomplish this task using a Convolutional Neural Network (CNN) and/or Recurrent Neural Network (RNN) [21, 24], a fuzzy logic classifier [3] and achieved considerable improvements . However, these models still lack the ability to utilize long-term dependencies and fuse multimodal data effectively.

The concept of attention has recently been applied for a variety of tasks as an integral part of sequence and transduction models [30]. In this study we propose a model of Multi-Attention Recurrent Neural Network (MA-RNN) and show that the attention mechanism can be used for data fusion and dimensionality reduction while maintaining long-term dependencies. Our model consists of two attention layers; one for multimodal data and the other for BiGRU. The proposed model allows for effective multimodal data fusion, and it achieved a new state-of-the-art performance in sentiment analysis on the recently introduced CMU-MOSI dataset [33].

Performing sentiment analysis is an arduous task that requires an understanding of the context of a sentence in sequential order [24]. Studies have been conducted that perform this task by using a CNN [23] or Hidden Markov

Model [27]. Because these models were implemented based on unimodal data, such as audio or video only, their capability of being applied to multimodal data is limited .

In one of the earliest studies based on the fusion of different modalities, De et al. [7] implemented a model by fusing audio and visual data with weighting functions, and the accuracy that they achieved was higher than those delivered by any other unimodal system. After this work, some bimodal systems were proposed [8, 19] ; however, sentiment analysis using multimodal data, such as a combination of text, video, and audio modalities was not explored.

Zadeh et al. [33] considered the lack of proper datasets and baselines as one of challenges presented by mulimodal sentiment analysis; they introduced the first multimodal sentiment analysis dataset (CMU-MOSI) labeled with sentiment intensity and subjectivity annotations at the opinion level. It consists of manual and automatic annotations of text and visual and audio features. With the dataset, multimodal-sentiment analysis research has made considerable progress [16].

Prior research has not made full utilization of multimodal data for sentimental analysis in terms of data fusion and modeling the long-term dependency. To address these challenges, we propose an MA-RNN model in this study using two attention layers and a BiGRU. The proposed approach achieved the state-of-the-art accuracy of 84.31 % on the CMU-MOSI dataset. The following sections describe the proposed model and its advantages over previous models.

2 DATASET (CMU-MOSI)

The CMU-MOSI dataset[33] consists of 2199 opinionated utterances by 89 speakers from 93 videos and each of them is labeled with its sentiment label, either positive or negative. Video clips covering a variety of topics, such as movies and books, from YouTube were used.

We intend to emphasize that the objective of this work was to evaluate the performance of sentimental analysis using multimodal data rather than emotion recognition. Sentiment analysis classifies the attitude or opinion of a user regarding; the content it can be positive or negative, whereas emotion recognition classifies specific emotions, such as happiness, sadness, or anger. As the other datasets, namely IEMOCAP, and AFEW for the EmotiW challenge are designed for emotion recognition with no sentiment labelling, they are not suitable for sentimental analysis. To the best of our knowledge, the CMU-MOSI dataset is the most common dataset used as the baseline for multimodal sentiment analysis ; therefore, we used it in our work.

The training /validation set of the CMU-MOSI dataset contained the first 62 video clips and the remaining 31 videos were used as the test set. Specifically, 1447 and 752 utterances were used in training and testing, respectively. For data pre-processing, feature extraction techniques suggested by Poria et al. [22] were applied in our work. The feature extraction details for each modality and multimodal data fusion are described in the following subsections.

Textual Feature Extraction. A CNN having two convolutional layers was selected to extract features from utterances represented as a matrix of Google word2vec vectors [20]. The first layer has two kernels of size 3×3 and 4×4, with 50 feature maps each, and the second layer has a kernel of size 2×2 with 100 feature maps, where the convolution layers are interleaved with max-pooling layers of size 2×2. A fully connected layer of size d and softmax output followed the convolutional layers, and the ReLU was used for the activation function. The activation values of the fully-connected layer were used as the features of the utterances for the text modality.

Audio Feature Extraction. An open-source software program , openSMILE toolkit [9] , was used to extract audio features. Audio features were provided with a 30 Hz frame rate and sliding window of 100 ms; then, to identify samples with and without voice, voice normalization was performed using Z-standardization. The features provided by openSMILE contained several low-level descriptions (LLD), such as MFCC, pitch, voice intensity, and their statistics, e.g., mean and the root mean square values .

Visual Feature Extraction. To learn the spatiotemporal features in a video, a 3D-CNN [29] was used in this work. Let $V \in \mathbb{R}^{c \times f \times h \times w}$ represent each video in training/testing dataset, where c = number of channels, f = number of frames, h = height of each frame, and w = width of each frame. A 3D convolutional filter, $F \in \mathbb{R}^{f_m \times c \times f_d \times f_h \times f_w}$ where f_m = number of feature maps, c = number of channels, f_d = number of frames, f_h = height of the filter, and f_w = width of the filter, was applied to each video V and produced $V_{\text{out}} \in \mathbb{R}^{f_m \times c \times (f - f_d + 1) \times (h - f_h + 1) \times (w - f_w + 1)}$. Max-pooling of $3 \times 3 \times 3$ on V_{out} was performed, and the output of the pooling layer was used as the input for the fully-connected layer of size d and a softmax layer. The activation values of the fully-connected layer were used as the features of utterances for the visual modality.

Multimodal Data Fusion. The dimensions of the extracted unimodal features of utterances (d_t,d_a,d_v) where d_t is for text, d_a is for audio, and d_v is for video, were equalized using a fully-connected layer of size d. Then, by concatenating the unimodal features, a multimodal feature $D \in \mathbb{R}^{d \times 3}$ was generated as given below:

$$D = [d_t, d_a, d_v]. \tag{1}$$

According to our experiments, the proposed model achieved the best performance when the value of d was set to 100.

3 MODEL ARCHITECTURE

Our proposed model consists of two attention layers and a BiGRU. The first attention layer is used to fuse multimodal data and reduce dimensionality, as will be described in more detail in Section 3. To capture the key parts of the contextual information among utterances, an attention-based BiGRU was implemented as will be described in Section 3. Figures 1 and 2 show the input layer with the first attention model and the second attention model with the BiGRU, respectively. The combination of these two networks constitute the overall structure of our proposed MA-RNN model.

Scaled Dot-Product Attention for Multimodal Data

The proposed network uses multimodal data consisting of text, audio, and video as its input. The first attention layer fuses these heterogeneous data and reduces dimensionality. It is applied to the multimodal feature vector D, the concatenation of the individual unimodal features, as described in Section 2.

The concept of the attention-based multimodal data fusion mechanism was originally suggested by Poria et al. [22]. However, we replaced Poria's mechanism with the scaled dot-product attention proposed by Vaswani et al. [30]. The equation for the scaled dot-product in our proposed network is as follows:

$$A_D = D \ \text{softmax}\left(\frac{\tanh(W_F D)^T w_F}{\sqrt{d}}\right), \tag{2}$$

where $W_F \in \mathbb{R}^{d' \times d}$, $D \in \mathbb{R}^{d \times 3}$, $w_F \in \mathbb{R}^{d' \times 1}$, and $A_D \in \mathbb{R}^{d \times 1}$. Here, W_F and w_F are parameters to be learned during training, d is the size of the fully-connected layer used in a process of the feature extraction as described in Section 2, and d' is a dimension of an attention weight vector. The output, A_D, produces an attention score for each modality.

The dimensions of d' and d, it achieved the best performance when d' was set to $4d$ and we used $d = 100$ and $d' = 400$ in our model . A further description is provided in Section 4.

Multi-head Attention. Vaswani et al. [30] demonstrated that the multi-head attention can be beneficial in terms of helping a model to use the data from different representation subspaces at different positions. The result of multi-head attention $w_a \in \mathbb{R}^{d' \times 1}$ was derived by summing the matrix multiplication of w_F and w_F^T as follows:

$$w_a = \text{sum}\left(w_F w_F^T\right), \tag{3}$$

where $w_F \in \mathbb{R}^{d' \times r}$ is the concatenated attention and r is the number of multi-heads. Therefore, to enable our attention layer to have multiple representation subspaces at different positions, the multi-head attention [30] was applied. However, we use $w_a \in \mathbb{R}^{d' \times 1}$ as calculated using Eq. (3) instead of $w_F \in \mathbb{R}^{d' \times 1}$ in Eq. (2). We achieved the best performance when r is set to 150.

Attention-based BiGRU

To address the issue of learning long-term dependencies when training a Recurrent Neural Network (RNN), such as gradient vanishing or exploding problem [2, 11, 12], Long Short- Term Memory network (LSTM) and Gated Recurrent Unit (GRU) were introduced by Hochreiter & Schmidhuber [13] and Cho et al., [5] respectively.

Since the sequential utterances in a video are temporally and contextually dependent, understanding inter-utterance relationships and finding contextual evidences for sentiment classification of target utterance are crucial [22]. To capture the key parts of the sequential data, recent works [17, 31, 34] have been augmenting LSTM/GRU models with the attention mechanism. Although Chung et al. [6] empirically evaluated the performance between LSTM and GRU, they could not confirm which model is better. Nevertheless, we observed that our proposed model achieved a 6.2 % improvement with GRU compared with LSTM.

Gated Recurrent Unit(GRU) network. When M utterances are given in a video, the input $X \in \mathbb{R}^{d \times M}$ can be represented as $[x_1, x_2, \ldots, x_M]$ where $x_t \in \mathbb{R}^d$ for $t = 0$ to $M - 1$. Each cell in GRU can be computed as follows:

$$h_t^j = \left(1 - z_t^j\right) h_{t-1}^j + z_t^j \tilde{h}_t^j \tag{4}$$

$$z_t = \sigma\left(W_z x_t + U_z h_{t-1}\right) \tag{5}$$

$$\tilde{h}_t = \tanh\left(W x_t + r_t \odot U H_{t-1}\right) \tag{6}$$

$$r_t = \sigma\left(W_r x_t + U_r h_{t-1}\right), \tag{7}$$

where W_z, W_r, U_z, $U_r \in \mathbb{R}^{d \times d}$ are the parameters to be learned during the training, σ is the sigmoid function and $\odot$ is element-wise multiplication.

The output of Eq. (4) is connected to the BiGRU layer that obtains temporally and contextually-aware utterance H where $H = [\overrightarrow{h_t}\ \overleftarrow{h_t}]$ and $H \in \mathbb{R}^{2d \times M}$.

Attention network. To augment the BiGRU such that it captures the key parts of the contextual information in the utterances, an attention network is implemented as shown in Figure 2.

An attention weight vector α_t for utterance information represented by h_t at time t is calculated as follows:

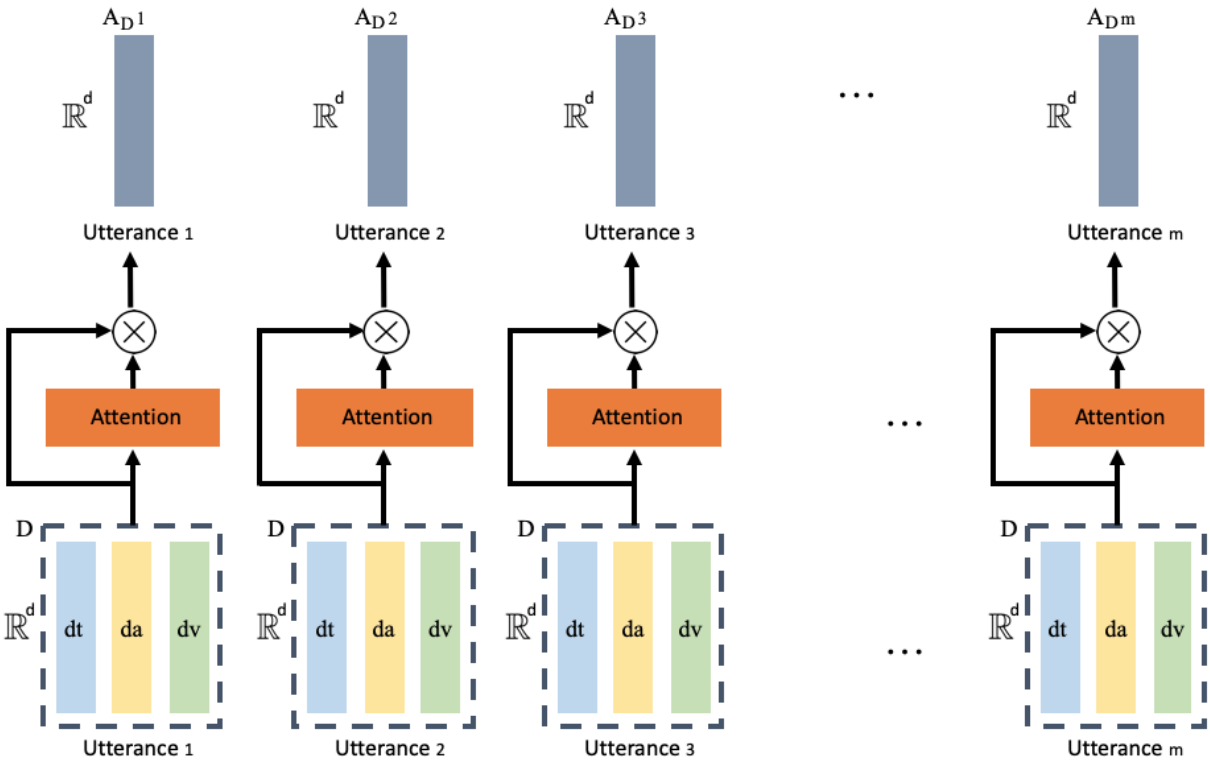

Figure 1: Scaled Dot-Product Attention for Multimodal Input

$$p_t = \tanh\left(W_h[t]H + b_t\right) \tag{8}$$

$$\alpha_t = \text{softmax}\left(p_t^T w_t\right) \tag{9}$$

$$r_t = H\alpha_t, \tag{10}$$

where $W_h[t] \in \mathbb{R}^{2d\times 2d}$, $b_t \in \mathbb{R}^{2d}$, $p_t \in \mathbb{R}^{2d\times M}$, $w_t \in \mathbb{R}^{2d\times 1}$, $\alpha_t \in \mathbb{R}^{M\times 1}$, and $r_t \in \mathbb{R}^{2d\times 1}$.

As in prior studies [22, 31], a projected vector, p_t, was multiplied with a randomly initialized weighting vector w_t to obtain a weighted hidden representation r_t. Our final BiGRU representation with the attention mechanism for t^{th} utterance was obtained as follows:

$$h_t^* = \tanh\left(r_t + h_t\right), \tag{11}$$

where $h_t^* \in \mathbb{R}^{2d}$.

Classification. For sentiment classification, the output of the BiGRU cell, h_t^*, was fed into a softmax layer.

$$z_t = \text{softmax}\left(W_{soft}[t]h_t^* + b_{soft}[t]\right) \tag{12}$$

$$\hat{y}_t = \underset{j}{\text{argmax}}\left(z_t[j]\right) \quad \forall j \in class, \tag{13}$$

where $W_{soft}[t] \in \mathbb{R}^{ydim\times 2d}$, $b_{soft}[t] \in \mathbb{R}^{ydim}$, $z_t \in \mathbb{R}^{ydim}$, $ydim$ = number of classes, and $\hat{y}_t$ is the predicted class.

4 EXPERIMENTAL RESULTS

The proposed model was trained with a workstation equipped with 4 Titan Xp NVIDIA GPUs with 12 GB Memory, 64 GB RAM, and an Intel Xeon E5-1650 CPU operating at 3.6 GHz. We used the Adam optimizer [14] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and applied dropout [28] to the output of each sub-layer with the rate of 0.2. The learning rate of 0.0001 was selected with the batch size of 20. The training was completed after 100 epochs.

Table 1: Comparison of the state-of-the-art models on the CMU-MOSI dataset

Models	Accuracy
Temporally Selective Attention Model [32]	75.10%
GME-LSTM with Attention [4]	76.50%
Character-level RNN [16]	80.40%
Contextual Attention BiLSTM [22]	81.30%
MMMU-BA [10]	82.31%
Multi-Attention with BiGRU	**84.31%**

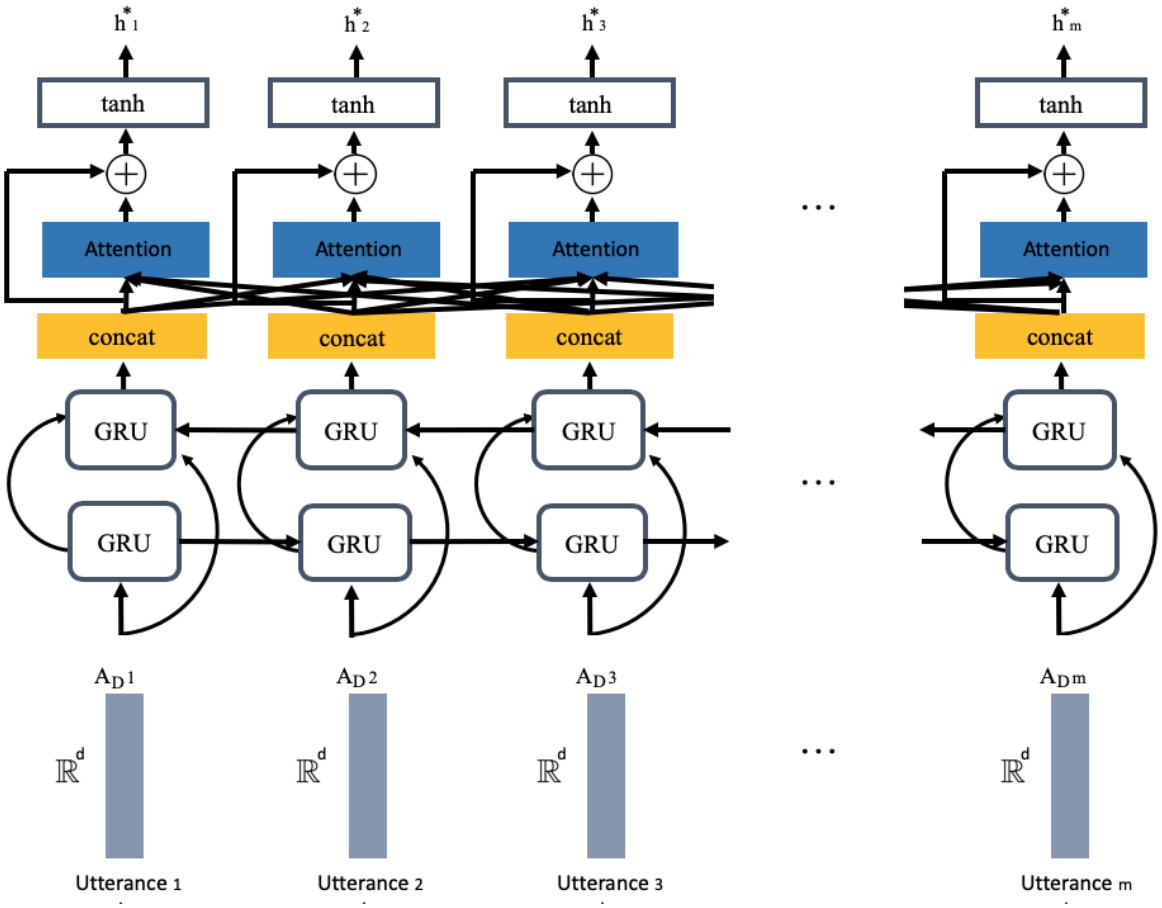

Figure 2: Attention-based Bidirectional GRU

As shown in Table 1, we demonstrated that our multi-attention mechanism with BiGRU outperformed the best existing models on the CMU-MOSI dataset, with an accuracy of 84.31 %.

Various attention mechanism were analyzed to improve the performance of our model. The influence of different components on the improvements in accuracy were investigated . The results of our ablation study are summarized in Table 2. Overall, the attention mechanism used in our model helped improve the accuracy of multimodal sentiment analysis by 17 %.

Table 2: Adapted attention techniques and improvement

Attention techniques	Improvement
Scaled Dot-Product	1.8 %
Dimension expansion of an attention layer	0.5 %
Multi-head Attention	0.7 %
Attention with LSTM vs BiGRU	6.2 %
The addition of a hidden vector	3.3 %
None vs 1^{st} attention layer	1.1 %
None vs 2^{nd} attention layer	3.4 %
Total	**17 %**

Scaled Dot-Product

We can clearly see the advantage of the scaled dot-product [30] in the attention mechanism. By applying only the scaled dot product with the same settings, the performance showed an improvement of 1.8 %.Vaswani et al. [30] reported that the scaled dot-product outperformed the dot product attention by preventing the softmax function from having extremely small gradients. Moreover, the dot-product attention is faster and more space-efficient owing to the highly optimized matrix multiplication code. To scale the dot product, d was used instead of $\sqrt{d}$ as in Eq. (2). This provided higher accuracy than the dot product without the scaling, but less than $\sqrt{d}$.

Dimension expansion of an attention layer

In our attention mechanism, we expanded the dimension of the attention weights from d to d' using the dot-product of W_F with D. Then, the expanded dimension (d') was contracted to the original dimension (d) as Eq. (2). This provided an accuracy improvement of approximately 0.5 %. We believe that through dimension expansion, the attention weight can be expanded in a larger dimension and reorganized to focus on the important features in dimension reduction. When the number of dimensions used for d' are smaller than d , the improvement in accuracy was lower.

Multi-head Attention

The concept of multi-head attention suggested by Vaswani et al. [30] was applied. As they described, it helped a model to have multiple representation subspaces at different positions. In our model, the multi-head attention technique exhibited improvement of approximately 0.7 % over single attention. It achieved the best performance when the dimension of the multi-head attention was set to 150. The dimension of the matrix multiplication of w_F and $w_F^T \in \mathbb{R}^{d' \times d'}$ in Eq. (3) was reduced to $\mathbb{R}^{d' \times 1}$ by summing the rows. It showed greater accuracy improvement than when a max value was used in rows (e.g., *reduce_max* function in tensorflow).

LSTM vs BiGRU

Although Chung et al. [6] empirically compared the performance of LSTM with GRU, they could not determine which model was better. However, in the present work, BiGRU achieved 6.2 % improvement over the BiLSTM. We believe this is because the GRU can be trained relatively faster with low number of training data owing to its simpler model architecture.

Additional Hidden Vectors

Inspired by Rocktaschel et al. [25], h_t^* was calculated by the addition of a weighted hidden representation r_t and utterance information h_t as in Eq. (11). Compared with their approach, the projection parameters to be learned during training, such as weight vectors, were not used. In our experiments, it achieved better results, accuracy improvement of 3.3 %, than when only r_t was used . We believe that by adding the output of a BiGRU h_t with a weighted hidden representation (r_t), h_t^* can capture more temporal and contextual information than only with r_t.

The first and second attention layer

In our model, the first and second attention layer achieved accuracy improvements of 1.1 % and 3.4 %, respectively. Although the second attention layer achieved better improvement, the first attention layer allowed multimodal data fusion and dimensionality reduction. We believe that such multi-attention architecture may be beneficial to models that require heterogeneous data fusion and long-term dependency modeling.

5 CONCLUSION

In this work, we presented a model of MA-RNN for the sentiment analysis on multimodal data and achieved the state-of-the-art performance on CMU-MOSI dataset. The first attention layer in our model was applied for data fusion and dimensionality reduction, and the second attention layer was implemented to augment BiGRU for capturing the key parts of the contextual information among utterances. Based on our results, we believe that the proposed attention demonstrated its strengths in data fusion and dimensionality reduction, while keeping long-term dependencies. In addition, the attention mechanism used in our experiments helped achieve an accuracy improvement of 17 % for the multimodal sentiment analysis. Further, attention mechanism and model architecture may offer helpful insights for future attention-based models.

REFERENCES

[1] Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 2005. Emotions from text: machine learning for text-based emotion prediction. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*. Association for Computational Linguistics, 579–586.

[2] Yoshua Bengio, Patrice Simard, Paolo Frasconi, et al. 1994. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks* 5, 2 (1994), 157–166.

[3] Iti Chaturvedi, Ranjan Satapathy, Sandro Cavallari, and Erik Cambria. 2019. Fuzzy commonsense reasoning for multimodal sentiment analysis. *Pattern Recognition Letters* 125 (2019), 264–270.

[4] Minghai Chen, Sen Wang, Paul Pu Liang, Tadas Baltrušaitis, Amir Zadeh, and Louis-Philippe Morency. 2017. Multimodal sentiment analysis with word-level fusion and reinforcement learning. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. ACM, 163–171.

[5] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2014), 1724–1734.

[6] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. [n.d.]. Empirical evaluation of gated recurrent neural networks on sequence modeling. *NIPS 2014 Workshop on Deep Learning* ([n. d.]).

[7] Liyanage C De Silva, Tsutomu Miyasato, and Ryohei Nakatsu. 1997. Facial emotion recognition using multi-modal information. In *Proceedings of ICICS, 1997 International Conference on Information, Communications and Signal Processing. Theme: Trends in Information Systems*

Engineering and Wireless Multimedia Communications (Cat., Vol. 1. IEEE, 397–401.

[8] Florian Eyben, Martin Wöllmer, Alex Graves, Björn Schuller, Ellen Douglas-Cowie, and Roddy Cowie. 2010. On-line emotion recognition in a 3-D activation-valence-time continuum using acoustic and linguistic cues. *Journal on Multimodal User Interfaces* 3, 1-2 (2010), 7–19.

[9] Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 1459–1462.

[10] Deepanway Ghosal, Md Shad Akhtar, Dushyant Chauhan, Soujanya Poria, Asif Ekbal, and Pushpak Bhattacharyya. 2018. Contextual inter-modal attention for multi-modal sentiment analysis. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 3454–3466.

[11] Sepp Hochreiter. 1991. Untersuchungen zu dynamischen neuronalen Netzen. *Diploma, Technische Universität München* 91, 1 (1991).

[12] Sepp Hochreiter. 1998. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 6, 02 (1998), 107–116.

[13] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[14] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *Proceedings of the International Conference on Learning Representations (ICLR)* (2014).

[15] Shashidhar G Koolagudi and K Sreenivasa Rao. 2012. Emotion recognition from speech: a review. *International journal of speech technology* 15, 2 (2012), 99–117.

[16] Egor Lakomkin, Mohammad Ali Zamani, Cornelius Weber, Sven Magg, and Stefan Wermter. 2019. Incorporating End-to-End Speech Recognition Models for Sentiment Analysis. *International Conference on Robotics and Automation (ICRA)* (2019), 7976–7982.

[17] Yukun Ma, Haiyun Peng, and Erik Cambria. 2018. Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive LSTM. In *Thirty-Second AAAI Conference on Artificial Intelligence*. 5876–5883.

[18] Andrew L Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1*. Association for Computational Linguistics, 142–150.

[19] Angeliki Metallinou, Sungbok Lee, and Shrikanth Narayanan. 2008. Audio-visual emotion recognition using gaussian mixture models for face and voice. In *2008 Tenth IEEE International Symposium on Multimedia*. IEEE, 250–257.

[20] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[21] Soujanya Poria, Erik Cambria, and Alexander Gelbukh. 2015. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In *Proceedings of the 2015 conference on empirical methods in natural language processing*. 2539–2544.

[22] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Mazumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Multi-level multiple attentions for contextual multimodal sentiment analysis. In *2017 IEEE International Conference on Data Mining (ICDM)*. IEEE, 1033–1038.

[23] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, and Prateek Vij. 2016. A Deeper Look into Sarcastic Tweets Using Deep Convolutional Neural Networks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. 1601–1612.

[24] Soujanya Poria, Iti Chaturvedi, Erik Cambria, and Amir Hussain. 2016. Convolutional MKL based multimodal emotion recognition and sentiment analysis. In *2016 IEEE 16th international conference on data mining (ICDM)*. IEEE, 439–448.

[25] Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomáš Kočiskỳ, and Phil Blunsom. 2015. Reasoning about entailment with neural attention. *Proceedings of the International Conference on Learning Representations (ICLR)* (2015).

[26] Marc Schröder. 2001. Emotional speech synthesis: A review. In *Seventh European Conference on Speech Communication and Technology*. EUROSPEECH, 561–564.

[27] Björn Schuller, Gerhard Rigoll, and Manfred Lang. 2003. Hidden Markov model-based speech emotion recognition. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, Vol. 2. IEEE, II–1.

[28] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research* 15, 1 (2014), 1929–1958.

[29] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*. 4489–4497.

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[31] Yequan Wang, Minlie Huang, Li Zhao, et al. 2016. Attention-based LSTM for aspect-level sentiment classification. In *Proceedings of the 2016 conference on empirical methods in natural language processing*. 606–615.

[32] Hongliang Yu, Liangke Gui, Michael Madaio, Amy Ogan, Justine Cassell, and Louis-Philippe Morency. 2017. Temporally selective attention model for social and affective state recognition in multimedia content. In *Proceedings of the 25th ACM international conference on Multimedia*. ACM, 1743–1751.

[33] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. MOSI: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *IEEE Intelligent Systems* (2016).

[34] Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Vol. 2. 207–212.