Trimodal Attention Module for Multimodal Sentiment Analysis (Student Abstract)

Anirudh Bindiganavale Harish
Department of Electronics and Communication Engineering
National Institute of Technology Karnataka, India
anirudhbh@ieee.org

Fatiha Sadat
Department of Computer Science
Université du Québec à Montréal(UQÀM)
sadat.fatiha@uqam.ca

Abstract

In our research, we propose a new multimodal fusion architecture for the task of sentiment analysis. The 3 modalities used in this paper are text, audio and video. Most of the current methods deal with either a feature level or a decision level fusion. In contrast, we propose an attention-based deep neural network and a training approach to facilitate both feature and decision level fusion. Our network effectively leverages information across all three modalities using a 2 stage fusion process. We test our network on the individual utterance based contextual information extracted from the CMU-MOSI Dataset. A comparison is drawn between the state-of-the-art and our network.

Introduction

Multimodal sentiment analysis is an upcoming field in natural language processing(NLP). Owing to the success of multi-head attention mechanisms in machine translation(Vaswani et al. 2017), we have utilized a customized version of the attention mechanism for our task of sentiment classification. Prior research conducted has been focused on 2 steps, namely feature extraction and feature fusion.

In this paper, we introduce a new method for fusion of features as a 2 stage process. The first takes place early, on a feature level. This early fusion helps efficiently model the complex inter-dependencies of the 3 modalities. This is followed by a secondary decision level fusion. The second stage takes a naive, emphasising on the final contribution of each of the modalities. This novel two stage fusion facilitates better calibration, for a more optimal performance. Our approach additionally provides a means of masking the input sequence, which we use to induce causality and help effectively implementation our network.

Dataset

We will utilize the Multimodal Opinion Sentiment Intensity (CMU-MOSI) dataset(Zadeh et al. 2016). Following in line with most previous research, we will be focusing on the binary classification of sentiments. The raw dataset and pre-processed features can be obtained from the CMU-Multimodal SDK[1].

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

CMU-MOSI dataset consists of 93 videos spanning over 2199 utterances. Each utterance has a sentiment label associated with it. It has 62 & 31 videos in training(train+val) & test set accounting for 1447 & 752 utterances. The videos are segmented with each segments sentiment label scored between +3 (strong positive) to -3 (strong negative). Due to our experiments involving only binary classification, we group the data into positive and negative opinion segments.

Methodology

Feature Extraction

For the analysis conducted in this study, we utilize pre-extracted utterance level features[2] extracted from (Poria et al. 2017). These features are calculated from the first layer of the hierarchical Long Short Term Memory(LSTM) using the features from the convolutional neural network(CNN), 3D CNN and openSmile for the textual, video and audio data respectively. The dimension of the data provided are 100, 73 and 100 for the text, audio and video respectively.

Feature and Decision Fusion

In our approach, we aim to utilize both the multimodal information as well as the contextual information present in the data. Utterances-level features are characterised by the features extracted in a 5-20 second interval of a user video. Hence, it would be logical to conclude that the sentiments of the previous utterances samples reveal information about future samples, which we define as contextual data. In line with this conclusion, we deal with a single utterance as a single data-point in time. We exploit this information using our network developed for the purpose of multimodal fusion.

Feature Level Fusion We first pass the utterance features, each through a separate dense layer with 256 ReLU units. The output of the dense layer are $l_t, a_t, v_t \in \mathbb{R}^{256}$.

These vectors are then passed through a Trimodal attention module(TAM). This module is an adaptation based on the multi-head attention(MHA) layers proposed in (Vaswani et al. 2017). The hyper parameters for the MHA is kept at 8 heads, 6 layers and dmodel as 256.

[2]https://github.com/soujanyaporia/multimodal-sentiment-analysis

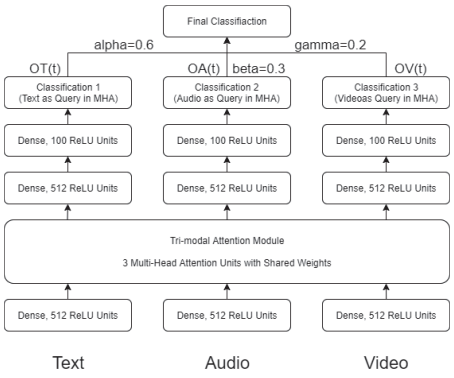

Figure 1: The Trimodal Attention Module(TAM) used in our network

The attention module utilizes 3 MHA units, one for each modality. The 3 MHA units share their weights to reduce the overall storage and computation cost. We employ the use of a look ahead mask to ignore future time instances, there-by inducing causality. The input Key and Value for the MHA units are kept the same, changing only the Query vector, across the 3 units. The inputs to these are given by

$$Q1 = l_t;\ Q2 = a_t;\ Q3 = v_t \tag{1}$$

$$K = V = concat(l_t, a_t, v_t) \tag{2}$$

The obtained output sequences are passed though 2 dense layers of 256 ReLU and 100 ReLU units respectively followed by a classification layer. Fig.1 illustrates the architecture used in the trimodal attention module. The three binary classified output sequences OT_t, OA_t, OV_t, are optimized w.r.t the ground truth.

Decision Level Fusion Following the optimization, we take the weighted sum of the 3 sequences. We theorize that the fusion of 3 optimized sequences will further increase the accuracy and provide better fidelity. The fusion is given by

$$output(t) = \alpha \times OT_t + \beta \times OA_t + \gamma OV_t \tag{3}$$

$$\alpha + \beta + \gamma = 1 \tag{4}$$

Under the constraints of Eq.4, a search was conducted for the best coefficients. It was found that 0.5, 0.3 and 0.2 as α, β, γ gave superior results.

Training Configuration The Adam optimizer with the learning rate described in (Vaswani et al. 2017) has been used for training the network. The output labels and inputs to the MHA units were masked for padding. The network was trained for a total of 200 Epochs.

Results

The proposed network is tested against the state-of-the-art(SOTA) networks. Accuracy and the F1-score are the two metrics used the comparison. Table(1) show that our proposed contextual network outperforms all the SOTA networks in terms of the F1-Score by a considerable margin[3].

[3]The accuracy for VAE+bc-LSTM is missing. The F1-Score of the contextual bc-LSTM is presented in (Majumder et al. 2019)

Method	Acc	F1-Score
TFN (Zadeh et al. 2017)	77.1	77.9
MFN (Zadeh et al. 2018)	77.4	77.3
bc-LSTM (Poria et al. 2017)	**80.3**	75.1
VAE+bc-LSTM (Majumder et al. 2019)	-	80.4
Our Network	79.3	**83.8**

Table 1: Results of Trimodal Attention Module on MOSI.

An overall increase of 3% can be seen compared to the SOTA algorithms. This showcases a well balanced in terms of precision and recall. In our case, we not only need precise classification of classes, bit also the ability to recall properties relevant to each class, making our network robust. In terms of accuracy, our network outperforms most of the SOTA networks, nearly achieving the highest results.

α, β, γ values suggest a higher contribution from the text. We attribute this result to the transcripts containing strong discerning features. For example, a high pitch can imply a positive or negative sentiment. Videos of sarcastic and excited reactions could be mistook for the other. Yet another case would be a stoic response irrespective of the content. In such cases, the text is the only reliable discerning factor.

Future Work

In the coming future, we envision the use of an end-to-end trainable network, where the decision fusion weights can also be trainable. Further, we envision the use of this network to achieve multi-task for emotion and sentiment analysis learning by utilizing the MOSEI dataset.

Conclusion

In this paper, we propose a novel architecture that facilitates the use of both feature level and decision level fusion of features simultaneously. Our proposed network is shown to outperform the state-of-the art algorithms.

References

Majumder, N.; Poria, S.; Krishnamurthy, G.; Chhaya, N.; Mihalcea, R.; and Gelbukh, A. 2019. Variational fusion for multimodal sentiment analysis. *ArXiv* abs/1908.06008.

Poria, S.; Cambria, E.; Hazarika, D.; Majumder, N.; Zadeh, A.; and Morency, L.-P. 2017. Context-dependent sentiment analysis in user-generated videos. In *ACL*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. In *NIPS*.

Zadeh, A.; Zellers, R.; Pincus, E.; and Morency, L.-P. 2016. Mosi: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *CoRR* abs/1606.06259.

Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; and Morency, L.-P. 2017. Tensor fusion network for multimodal sentiment analysis. In *EMNLP*.

Zadeh, A.; Liang, P. P.; Mazumder, N.; Poria, S.; Cambria, E.; and Morency, L.-P. 2018. Memory fusion network for multi-view sequential learning. In *AAAI*.