Emotion Recognition and sentiment analysis*

Final Report

Qirui Sun
Ming Hsieh Department of Electrical and Computer Engineering
University of Southern California
qiruisun@usc.edu

Abstract

With the huge improvement of hardwares as well as Internet technology, people are easier to generate multimedia data on the Internet everyday, which makes the multimodal based machine learning pipelines to solve problems come true. In this report, I will introduce my project, which is based on the existing method emotion recognition and sentiment analysis Transformer based multimodal model.[5] I used a cascade structure co-attentions maps and embedded them to the transformer model. The pipeline improved the accuracy of Sentiment analysis for 7 levels for sure while for emotion recognition and sentiment analysis for 2 labels, there are some better than baseline scenarios but not stable.

Introduction

Sentiment is a long-term reaction when people confront various scenarios, persons, etc.[1] Understanding people's sentiment is important in business, politics and other fields that need to communicate with people.[2] Emotion is a complicated and personal state that reacts people's psychophysiological reactions to the outside world.[3] Nowadays, with increase of Internet footprints, people spent more time on the Internet and expressed their suggestions opinions on social media, like Twitter, Facebook, Youtube, etc. These resources provide a large amount of good material for multimodal sentiment analysis and emotion recognition which make that use data to analyze and predict humans sentiments and emotions come ture.

Sentiment analysis and emotion recognition are promising fields that would help humans in multiple areas. Recently, emotion-aware systems have been implemented in health, online learning, chatbot, online gaming, etc.[4] In mental health, Chatting with the chatbot is easier for patients to share their feelings because they don't need to camouflage themselves and worry about privacy. For online business, using robots instead of people to work as assistants to communicate with people is also an important part, this can release people from simple boring work as well as saving the budget for companies.[13] In elder caring, robots have been implemented at industry level to accompany the elder, including guidance, communication, etc. [13] All areas mentioned above need a high ability to understand targets sentiment and make appropriate communications. Hence, Emotion recognition and sentiment analysis are important and have promising future in machine learning or phycology.

Predicting personal states is really hard because of the difficulty to define personal states. Humans can use different ways to express themselves, facial expressions, audios, ect.[3]

Related work

Human effect recognition problems can be tackled as NLP problems. There are some basic and classical problems that are related to sentiment analysis, like IMDB sentiment analysis.[8] Some traditional approaches like focusing on important words, n-gram, etc have been proposed in this area. Most recently, People also used LSTM and Transformer language models to do sentiment analysis on text for sentiment analysis. [6,7]

Multimodal Sentiment Analysis is an emerging research area with the emergence of social media and the Internets, which gave people different ways to express their feelings. Previous works in multimodal sentiment analysis directly fuse the modalities with either early fusion or late fusion. The simplest early fusion is just to concatenate multimodal features, and simple late fusion is majority vote which can help to combine different unimodal classifiers.[8]

Emotion recognition is a well-studied area in facial expressions.[9] Researchers can only use facial expressions to predict human emotions. During the audio-visual fields, emotion recognition is quite similar to sentiment analysis because they may share the same dataset with different labels.

There are multiple advanced multimodal machine learning algorithms. TFN (Tensor Fusion Network) introduced modeling the inter and intra modalities instead of simple early fusion.[9] They proposed an Outer Product method to fuse various modalities. MFN (Memory Fusion Network) researches the inner interactions of different modalities, which regards as view-specific interactions and cross-view interactions. Meanwhile, they designed two networks to tackle the interactions mentioned above.[10]

In this project, we focus on a transformer based network, which used guided attentions to fuse the two modalities.[5]

Method

1. Dataset

In this project, I used CMU-MOSEI dataset. Multimodal Opinion Sentiment and Emotion Intensity [11] is one of the largest multimodal datasets. All data was collected from Youtube self-presentation videos. They have three types of data in parallel, videos, audios and text. The dataset contains 3228 videos, 23453 sentences from 1000 presentators. Each video clip is around 2 minutes. Besides, there are multiple label choices of the dataset, 2,6 and 7. 2 is for negative and positive. 6 is for anger, disgust, fear, gender, happiness and sadness. 7 is for the degree of sentiment level and from -3 to 3 which means from highly negative to highly positive. CMU-MOSEI dataset is one of the largest and most comprehensive datasets in multiple views, shown in Table1 as well as Figure 1.

2. Transformer-based multimodal model [5]

Ronghang et al. [12] proposed transformer architecture which is based on attention and seq2seq. The transformer is composed of encoder and decoder. In every layer, there are two basic structures, multi-head attention and feedforward network.

Dataset	# S	# P	M	Sent	Emo
CMU-MOSEI	23453	1000	{l, v, a}	Y	Y
CMU-MOSI	2199	98	{l, v, a}	Y	N
ICT-MMO	340	98	{l, v, a}	Y	N
YouTube	300	50	{l, v, a}	Y	N
MOUD	400	101	{l, v, a}	Y	N

Table 1. Comparison of different multimodal dataset #S is the number of data point and #P is the number of presentators; M collected modality [7]

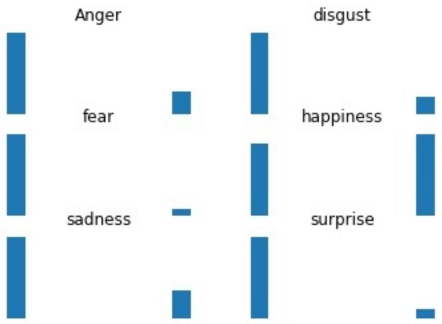

Figure 1. The distribution of Emotion labels for CMU-MOSEI dataset

Self-attention in a transformer is an interaction of Q, Query and K, key with value V. Multi-head self-attention means parallel processing the V with different (Q,K) parts and concatenating the result to output.

$$Attention(Q,K,C) = Softmax(\frac{QK^T}{\sqrt{k}}) \times V$$

In traditional late fusion multimodal transformer encoders, the only used two bunch of separate encoders to tackle the different modalities in parallel. After the multihead attention unit and feedforward network, it is common to use concatenation to concatenate the two modalities together and feed the output to a fully-connected network, shows like Figure 2.

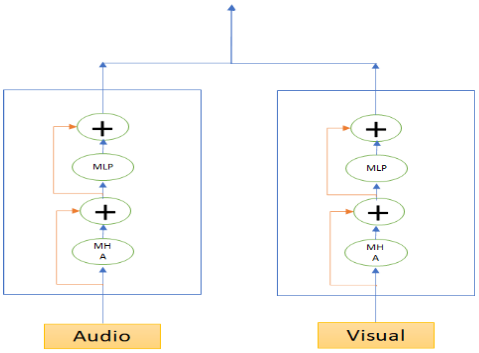

Figure 2. Traditional Fusion model of multimodal learning based on transformer

In the transformer-based multimodal model for CMU-MOSEI dataset, [5] they proposed a co-attention map, which contains self-attention and guided attention for each modality. Guided attention is to fuse the two modality early by getting input from the other modality. The used one modality for calculating Key and Value while another modality to calculate the Query, shows in Figure 3.

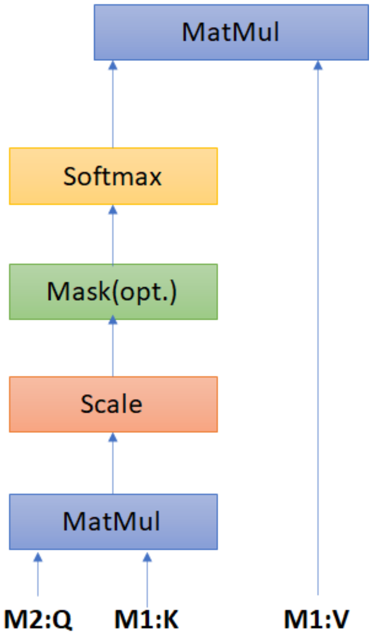

Figure 3. Guided attention in the paper as well as our project. M2 is modality2, M1 is the Modality1

Besides, transformer-based multimodal models also introduced a Glimpse layer, which is a linear layer to project the output from the feedforward network to a high dimensional space, making the output suitable for concatenation. Hence, the whole pipeline of the models, shown as Figure 4.

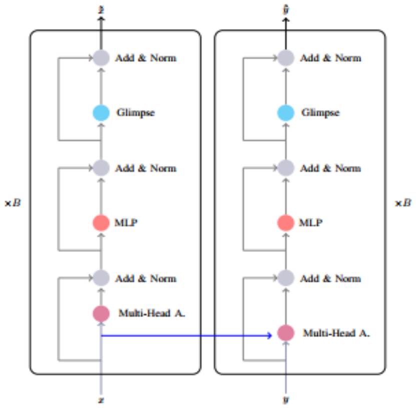

Figure 4. The architecture of transformer-based multimodal model for CMU-MOSEI dataset[5]

3. Multimodal transformer with co-attention map in cascade

Inspired by Yu et al. [14], they proposed a co-attention unit for the first time and stated that using multiple co-attention units cascade together may improve the performance. Hence, here we proposed two cascade co-attention pipelines with different architecture and embedding them to the original transformer-based model.

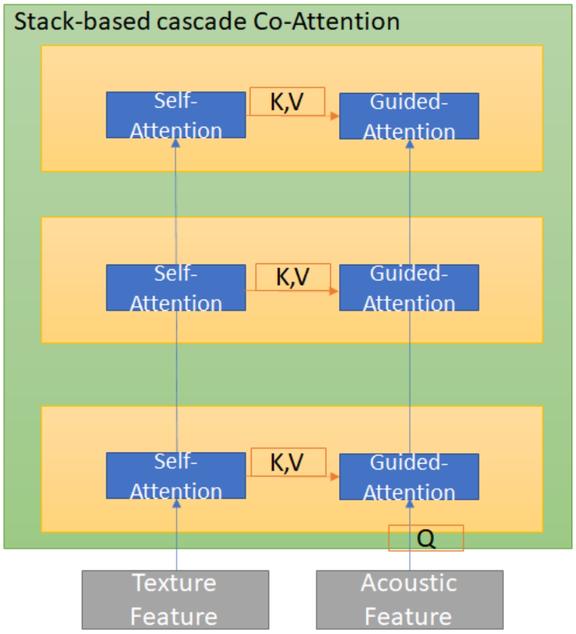

Figure 5. Stack-based co-attention map architecture

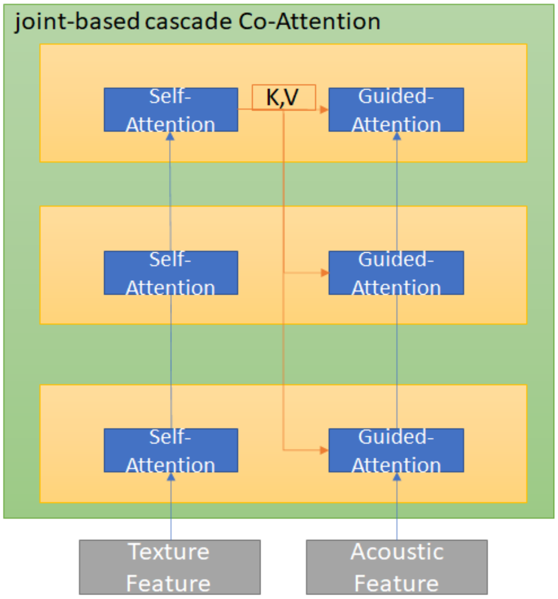

Figure 5. Joint based co-attention map architecture

Figure 4 and Figure 5 illustrate the two different cascade methods of co-attention map. The stack-based method is just like stacking every co-attention unit in cascade. Every self-attention unit passes its value to the guided attention, which looks like tackling the information from two modalities at the same time. The jointed based model is like an encoder-decoder pipeline that self-attention can be regarded as encoder and guided attention can be regarded as decoder. Text features pass through all self-attentions and transfer their final output to every guided attention, which means that every guided attention received the same values from self-attention. In this report, I referred to the joint based model as a joint model, stack based model as stack model, the both of them as cascade model and original model as basic.

Experiments

1. Feature extraction

CMU-MOSEI dataset are raw texture, audio and video data, which cannot be used for training purposes. In the paper transformer based multihead attention peper [7], the author used GloVe embedding to handle the texture feature and extract Mel-spectrum from audios.
Here, I extracted the text feature by word2vec embedding with the same dimensions as the author's feature dimension, 300. However, I still confronted a lot of problems, the author didn't mention some details about their data preprocessing features and I don't have enough time for preprocess video features and audio features, so I decided to use their preprocessed feature.

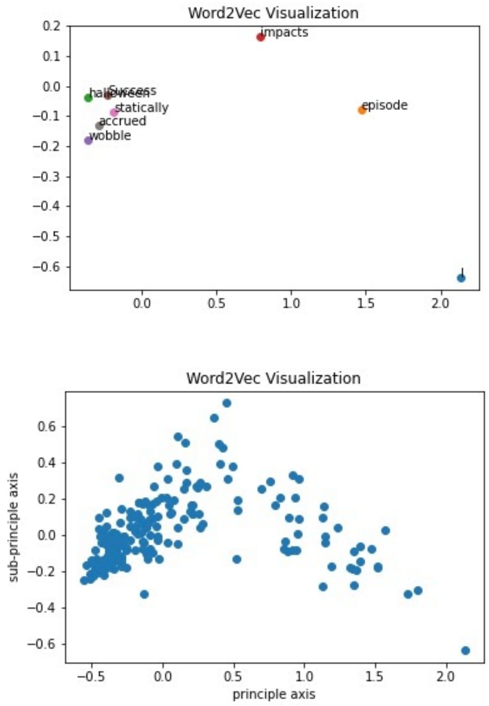

Figure6&7. Visualization of Word2Vec embedding

2. Experiments settings

I trained the model with different cascade depth in three individual tasks, emotion recognition, sentiment analysis(7 and 2). I followed the paper hyperparameters setting and chose Adam optimizer with batch size 32 and learning rate 1e-4. Besides, I also used learning rate decay factor 0.2 and early shop criterion. For modality, here I used acoustic feature and text feature. All experiments are under the same environment, GeForce GTX 2080Ti with seed 555.

3. Results

To compare the performance, I referred to the basic model as our baseline and to see if the cascade models improve the accuracy. By testing three tasks, (emotion recognition, sentiment analysis 2 and sentiment analysis 7), I got 81.03%, 81.61% and 42.33% for three tasks baseline accuracy respectively. Then I vary the depth of two proposed co-attention maps and compare to find the relation with depth, for sure. In the experiments, we found that there do have some improvement on my proposed methods. However, for Sentiment analysis (2) and emotion recognition, we cannot find obvious and steady improvement.

From the test accuracy versus epochs, shown in Figure7, we can observe that stack models have higher accuracy than joint models.

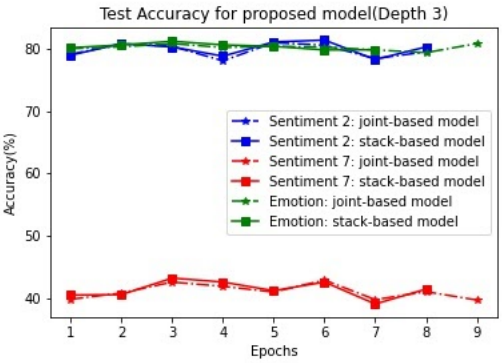

Figure7. Test accuracy Versus epochs, the same color means the same task, different line type means different models.

Besides, I also tried to find the relationship between the accuracy versus depth of the cascade co-attention map. From the results, shown in figure 8 and figure 9, we can find that there are some scenarios when proposed architecture gets a higher accuracy than baseline. However, The improvement can not be observed steadily or always higher than baseline in Figure 7.

For sentiment analysis for 7 labels, shown in Figure 8, we can get the steady improvement which proves that the co-attention units in cascade architecture have better test accuracy than baseline models.

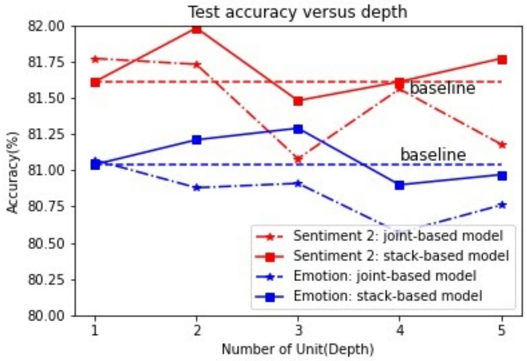

Figure 7 Test accuracy versus depth of cascade depth on sentiment analysis and emotion recognition

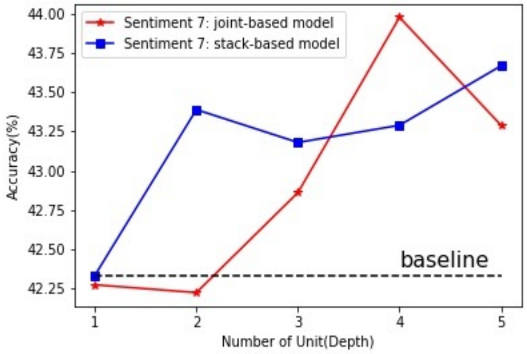

Figure 8 Test accuracy versus depth of cascade depth on sentiment analysis for 7 labels

Besides, I also compared the F1 scores and Accuracy between the proposed model and the baseline. The results are task based. We can find some improvement on accuracy on sentiment analysis for 7 labels, but we cannot get the same increase on F1 scores. For the other two tasks, the two metrics scores are quite similar.

Model	Metrics	Joint (3)	Joint (5)	Baseline
Sentiment 2	Macro F1	0.74	0.74	0.74
	Accuracy	81.08%	81.18%	81.61%
Sentiment 7	Macro F1	0.33	0.35	0.37
	Accuracy	42.86%	43.29%	42.33%
Emotion	Macro F1	0.67	0.66	0.69
	Accuracy	80.91%	80.76%	81.03%

Model	Metrics	stack (3)	Stack (5)	Baseline
Sentiment 2	Macro F1	0.77	0.76	0.74
	Accuracy	81.19%	81.77%	81.61%
Sentiment 7	Macro F1	0.32	0.34	0.37
	Accuracy	43.18%	43.67%	42.33%
Emotion	Macro F1	0.68	0.67	0.69
	Accuracy	81.29%	80.97%	81.03%

Table2 Accuracy and F1 score of experiments

Discussions

In this project, I tried two kinds of cascade co-attention map on the base of a transformer based multimodal model on CMU-MOSEI dataset.[5] For the two proposed cascade methods, stack models perform better than joint models. The difference of the two methods is that time to apply modalities fusion. Joint models pass the value for guidance after the one modality passes through all self-attention maps, which stack models do the joint attention parallel. The results show that stack base models have better accuracy, which means that to keep the information parallel during co-attention is important and can observe better results.

From the task perspective, the cascade models have a steady increase in sentiment analysis for 7 labels, which have the baseline accuracy for 42.33%. However for the other two tasks, which both have baseline accuracy around 81%, I cannot observe the steady improvement. Hence, I think that for the hard task, the cascade models work better because they have the larger model size and more capacity to learn from the multimodal features. However, for some easy tasks, there is no need to use deep cascade structure and it may confront gradient vanish or over fitting problems.

Even though there are some benefits or improvements on the proposed architecture, the baseline is still important because of the computational efficiency. Table 3 shows that with the depth of cascade models increases, the number of parameters increase linearly, which sacrifice training time and task accuracy.

Depth	1	2	3	4	5
Parameters	32.63 M	53.67 M	74.68 M	95.72 M	116.75 M

Table 3. The Number of parameters on different cascade depth's model

Conclusions

In this project, I applied two cascade co-attention maps on the base of a transformer based multimodal model to tackle the three different tasks, sentiment analysis for 7 levels, for 2 levels and emotion recognition.

- Stack models can reach better accuracy than joint models. It may result from that stack cascade model tackle the two modalities in parallel which can preserve more information
- The two proposed cascade models work better for hard problems, that is the baseline accuracy is low, like sentiment analysis for 7 levels. For easy problems, the deep architectures may cause gradient vanish and overfitting, which is not suitable for some large model size.
- There are still some limitations that I didn't add video features because videos don't have any improvement in the baseline model. However, if we use three modalities, there may be multiple different ways for cascade, which may have some improved scenarios.

Reference

[1] CARLSON, A.B. 1929. DCM 0861: Central American Indian / Whistle (Duct Flute). The Journal of Central American Indian Whistles Volume: 5, 11-21. .

[2] JANSEN, B.J., ZHANG, M., SOBEL, K. AND CHOWDURY, A. 2009. Twitter power: Tweets as electronic word of mouth. Journal of the American Society for Information Science and Technology 60, 2169-2188. .

[3] LIU, B. 2012. Sentiment Analysis and Opinion Mining. .

[4] TUMASJAN, A., SPRENGER, T., SANDNER, P. AND WELPE, I. 2010. Predicting elections with twitter: What 140 characters reveal about political sentiment. In Proceedings of the International AAAI Conference on Web and Social Media, Anonymous .

[5] DELBROUCK, J., TITS, N., BROUSMICHE, M. AND DUPONT, S. 2020. A Transformer-based joint-encoding for Emotion Recognition and Sentiment Analysis. arXiv preprint arXiv:2006.15955 .

[6] MYAGMAR, B., LI, J. AND KIMURA, S. 2019. Cross-domain sentiment classification with bidirectional contextualized transformer language models. IEEE Access 7, 163219-163230. ..

[7] LI, D. AND QIAN, J. 2016. Text sentiment analysis based on long short-term memory. In 2016 First IEEE International Conference on Computer Communication and the Internet (ICCCI), Anonymous IEEE, , 471-475.

[8] ZADEH, A., CHEN, M., PORIA, S., CAMBRIA, E. AND MORENCY, L. 2017. Tensor fusion network for multimodal sentiment analysis. arXiv preprint arXiv:1707.07250 .

[9] WANG, W., XU, K., NIU, H. AND MIAO, X. 2020. Emotion Recognition of Students Based on Facial Expressions in Online Education Based on the Perspective of Computer Simulation. Complexity 2020, .

[10] ZADEH, A., LIANG, P.P., MAZUMDER, N., PORIA, S., CAMBRIA, E. AND MORENCY, L. 2018. Memory fusion network for multi-view sequential learning. In Proceedings of the AAAI Conference on Artificial Intelligence, Anonymous .

[11] ZADEH, A.B., LIANG, P.P., PORIA, S., CAMBRIA, E. AND MORENCY, L. 2018. 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), Anonymous , 2236-2246.

[12] HU, R. AND SINGH, A. 2021. Transformer is all you need: Multimodal multitask learning with a unified transformer. arXiv preprint arXiv:2102.10772 .

[13] LOHSE, M., HEGEL, F., SWADZBA, A., ROHLFING, K., WACHSMUTH, S. AND WREDE, B. 2007. What can I do for you? Appearance and application of robots. In Proceedings of AISB, Anonymous , 121-126.

[14] YU, Z., YU, J., CUI, Y., TAO, D. AND TIAN, Q. 2019. Deep modular co-attention networks for visual question answering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Anonymous , 6281-6290.