Multimodal Emotion Recognition based on Self-Attention Strategy with Word-Level Alignment

Department of Electrical and Computer Engineering
Rutgers University
xl422@scarletmail.rutgers.edu

Abstract

Multimodal Emotion recognition is still challenging, because: it is difficult to extract the proper features revealing the actual human state of mind from the Multimodal dataset. Also, its hard to reasonably fuse the extracted features from different kinds of data branches, which should make sense to determine emotion recognition. Therefore, we attempt to build up a deep multimodal network with self-attention architecture to help extract and fuse features.

1 Introduction

Human-computer interaction is a fundamental field for AI. And before the communication, the computer should understand what human means during the conversations. However, teaching machines to speculate human emotions from humans conversations is by no means an easy thing. And that is because: 1. For us, it is still hard to precisely figure out the principles for emotion analysis between humans. 2. It is difficult to teach computers how to speculate human emotions from humans conversations. Because it is difficult to extract the proper features revealing the actual human state of mind. 3. We, humans, make emotion analysis based on a fusion of facial expressions, gestures linguistic contents, and vocal signals. So, it is necessary for us to find out how to integrate multiple resources to make proper emotion analysis.

There is a variety of data from the different source, like linguistic content and vocal signal, facial and body movements of from the conversation. And for our project, we choose to use linguistic content and the vocal signal of the communication to build the baseline model for multimodal data fusion. That is because these two branches of data are the most potent source of information in human conversations.

Recent work has done too much on extracting the features based on the sentence level from audio data and text data. And it doesnt make sense to me, because the basic unit for us to understand the meaning of the conversations is the word-level. For example, when learning English, it is common for us to find the corresponding meaning from our native spoken language. So, we make a word-level alignment between text data and the audio data before inputting them into our model.

Another challenge is that we need to extract the associated features from the audio data and text data. Recent works use traditional methods to obtain the handcrafted features from audio and text data with the help of LSTM and CNN. However, with the inspired from the Attention is all you need published by google research [1], we want to introduce an architecture only with the multi-head self-attention strategy to extract the features individually from each branch.

The last challenge is that since we have extracted the associated features individually from the audio data and text data, how can we make a fusion on it to decide on emotion classification. In this part, we introduce a weighted fusion model with self-attention and CNN's. That makes the model learn the weights of audio data and text data on the final decision of emotion classification.

For the experiments, we used the published multimodal emotion analysis dataset IEmocap to evaluate our model.

The report is consisting of 6 parts: related work has been described in section 2, we introduce the methods for our model in section 3, experiments presented in section 4, we provide a result from analysis and conclusions in section 5 & section 6.

2 Related Work

There are a lot of researchers have focused on audio-video emotion recognition, because it seems to be more intuitive. However, there are a few

numbers of research on text-audio one. As we know, much works are focusing on extracting the features from multimedia data, audio data like openSMILE [2], visual features with CNN attention from [3], texture features like BoW in [4] and PoS in [5]. Also, many researchers focused on the study on word-level multimodal features fusion such as [6].Additionally, there are many deep learning approaches introduced to work on the multimodal emotion recognition topics, such as CNN and LSTM from [7], [8]. Moreover, in [9], they put forward a hybrid deep multimodal structure to extract and fuse the textual and acoustic features on the multimodal dataset.

3 Methods

3.1 Data Preprocessing and Word-level alignment

As mentioned before, we made the word-level alignment between the audio data and the text data for the reason that, words are the basic units to comprehend during the conversations.

For the texture data, we first remove the punctuations, because they dont match any audio data. And with the help from pre-trained word embedding dictionary from Google News [10], we extract 200-dimension features by word2vec. And the word2vec features for each sentence have been stored in a 2D feature matrix as W*200. W stands for the number of words in a single sentence. And we zero-pad W dimension to its max length 98 for each sentence.

For the audio data, we extract MFSC features instead of MFCC features mentioned in the proposal [11]. That is because: 1. MFSC is correlated and appears to be smoother over the spectrum. 2. MFSC has a higher dimension for the audio data, which performs better in training deep neural networks. And since we apply 64 filters to extract the MFSC features, so the third dimension of 64 denotes the frequency of the audio data. Therefore, the feature matrix for each word in a single sentence represents as N*T*64. To make word-level alignment, we set the N to be 98. T stands for the time interval between adjacent word frame. Then for each word frame, we also zero-pad the time interval to the same length 513. Finally, the feature matrix for each sentence of audio data represents as 98*513*64.

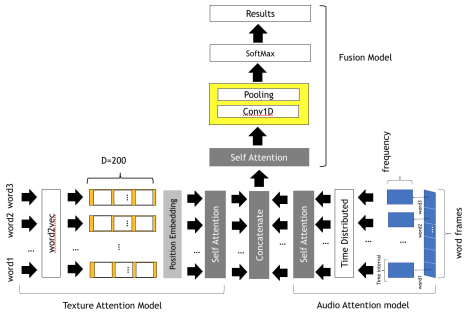

Figure 1: Whole model architecture

3.2 Self-attention

Before describing the application of self-attention layer in texture branch, audio branch, and fusion model, I want to briefly explain why we want to use the self-attention.

Many researches have applied LSTM & Attention architecture to enhance the performances of their model in NLP because LSTM structure helps to generate a context-hidden state for each word vector. And attention layer allows the model to focus on the relevant parts of the input vectors as needed by amplifying the weight on the representative vectors.

Since reading the paper, Attention Is All You Need publishing from Google search, I was a little shocked about their results, which is elegant. They introduce the self-attention layer which can perform attention operation to the sequence itself without extra information. It is different from the previous mainstream machine translation using an RNN-based seq2seq model framework. And the paper replaced the RNN with the attention mechanism to build the entire model framework. Self-Attention has some advantages: 1. It can directly capture the inner relationships of the word vectors regardless of the distance.2. It can learn the internal structure of sentences. 3. It is simpler, faster and can be generalized well to tasks in NLP. So, I want to find out whether the performances are outstanding like what they said in the paper.

3.3 Texture branch model with self-attention

The texture branch model is consisting of three parts, the input layer, attention layer and predicts layer as shown in Figure 1.

The input layer takes in (W,200) word2vec feature matrix. Before applying the self-attention layer, we first use the positionembedding layer to help improve the performance of the selfattention.

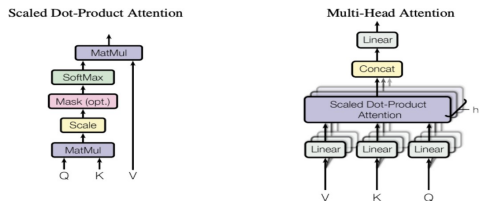

Figure 2: Whole model architecture

That is because selfattention model does not capture the order of the sequences. In other words, the word order in the sentence may be upset. Therefore, according to [1], we use positionembedding layer to attach the position information vector onto the word vectors.

As described in [1] and shown in Figure 2, the self-attention can be representing as below:

$$MultiHead(Q, K, V) = Concat(head_i, \cdots, head_h)W^O \quad (1)$$

Where

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V).$$

The Q is the query vector, K is the key vector, V is the value vector for each word in the texture data. More specifically, for self-attention, the inputs of Q, K, V are all the embedding word-level features we extracted by word2vec. And we can represent the text attention extracted from the self-attention as follows:

$$A_i^t = tanh(W_t[em_t, em_t, em_t] + b_t), is[1, N] \quad (2)$$

Where W_t, b_t are learnable parameters from self-attention. And em_t denotes the word2vec features after position embedding operation.

And between the self-attention layers, we apply Batch Normalization layers to solve Internal Covariate Shift problem [12] to make the model easier to train and converges more quickly. For Mini-Batch SGD, there are m training instances in one training process. The specific Batch Normalization operation is to transform the activation value of each neuron in the hidden layer as follows:

$$x(k) = \frac{x(k) - E[x(k)]}{\sqrt{Var[x(k)]}} \quad (3)$$

Then we use GlobalAveragePooling1D to reduce the dimension of the feature matrix to one dimension which will be convenient in the word-level fusion model. In GlobalAveragePooling1D, the input word vector sequences are added to the averaging and integrated into a vector [13].

With the text self-attention, we apply softmax activation function to calculate the distribution of the text attention.

$$a_i^t = \frac{exp(A_i^t)}{\sum_{k=1}^{N}(A_k^t)} \quad (4)$$

Where a_i^t represents the score of each attention. Finally, we use fully connected layers to help us make the prediction of the result in the audio part.

3.4 Audio branch model with self-attention

The audio branch model is almost similar to the text branch model shown in Figure 1. We also use the self-attention layers to find out the sequential relationship of the data.

$$A_i^t = tanh(W_t[au_t, au_t, au_t] + b_t), is[1, N] \quad (5)$$

Where W_t, b_t are learnable parameters from self-attention. And aut denotes the MFSC features from the acoustic data.

Then we apply GlobalAveragePooling1D to reduce the dimension of the features matrix to predict with softmax activation functions.

$$a_i^t = \frac{exp(A_i^t)}{\sum_{k=1}^{N}(A_k^t)} \quad (6)$$

Where a_i^t represents the score of each attention.

However, as we mentioned above, the MFSC feature maps represents as (N*T*64), in which N is the number of words in a single sentence and T represents the time interval between neighboring time frame. It is different from the features of the texture data with only two dimensions, and the features channels for the MFSC feature maps is 3.

Thus, we use Time Distributed layer [14] to apply fully connected dense on each time step and get output separately by timesteps.In this way, we reduce the second dimension of the audio feature matrix, and the output of the Time Distributed layer will be (N*64). In this way, we unify the texture feature together with the audio features to (N*64) and (N*200), which will be convenient for feature fusion in the fusion model.

3.5 Word-level Fusion model with self-attention

Many types of research only concatenate the text features and audio features together to make the fusion for decision making. However, it doesnt make sense to me, because we cant take it for granted that the all modality data equally contribute towards the sentiment prediction. For example, we can use voice and intonation information to distinguish the neural and happy emotion, but we cant discriminate the happy and angry emotion without the text information. Therefore, we should attach different weights to different modality features to achieve the performance of the model. So, we separate the fusion model into two parts, the attention part, and the decision part.

In the attention part, we first concentrate the text features and audio features, then we use the self-attention layer to help find the inner relationship between audio and texture features. Meanwhile, it helps to attach the weights to audio and texture features.

$$A_i^t = tanh(W_t[fu_t, fu_t, fu_t] + b_t),\ is[1, N] \quad (7)$$

Where W_t, b_t are learnable parameters from self-attention. And fut denotes the fusion features.

We use four 1D convolution layers F with different kernel size and maximum pooling layer with Batch Normalization to furtherly extract the representations of the word-level vectors [15]. And the output features map Ckj is shown as below.

$$C_{kj} = max f_{1j}, f_{2j}, \cdots, f_{Nj} \quad (8)$$

Where f_{ij} represents the features of word vector i extracted from CNN layer j.

Then, we merge the output features maps of these CNNs together and get the final representations. Finally, we attach two fully connected layer with softmax as the activation functions to make the final emotion classifications.

4 Experiments

4.1 Hardware

Our model was trained on a GTX 1080 GPU with 32GB RAM.

4.2 Datasets

For the project, we used the Interactive Emotional Dyadic Motion Capture (IEMOCAP) database [16], an acted multimodal and multispeaker dataset, which is collected at SAIL lab in USC recently. It contains approximately 12 hours of audiovisual data, including video, speech, motion capture of the face, text transcriptions from the actors when answering impromptu questions during scripted scenarios. And the text transcriptions have been attached to the emotion labels such as anger, happiness, sadness, based on facial expressions. IEMOCAP database has shown up on many conferences and journals.

4.3 Model training

We use Keras 2.0.9 and Tensorflow 1.3.0 as the backend to implement our model.

We separate both the audio and text data, label data in the whole dataset to training dataset and testing dataset with the scale as 4:1. For instance, there are 1005 excited data. And we separate them into a training dataset and test dataset with the scale as 4:1. In this way, both the test and train dataset are balanced with all kinds of emotion data.

When training the model, we first train the text branch model with 100 epochs, then we train the audio branch model with 50 epochs. At last, we train the fusion model with both text and audio features as input about 100 epochs. For the input of the fusion branch, we set 200 and 64 as the dimension of the features for the texture data input and the audio data input. We use batch normalization between layers to normalize the batch data. And we also applied dropout layer with the rate as 0.6 to avoid overfitting. Then, for the fusion branch on the word-level, we use relu as activation functions. And for all the training process, we set Adam as the optimizer with learning rate as 0.001 and categorical cross-entropy as the loss function. For the test part, we apply 5-fold cross-validation.

5 Result Analysis

5.1 Baselines

H-DMS: A hybrid deep multimodal structure to extract and fuse the textual and acoustic features on the IEMOCAP dataset [9].

Uni-SVM: The SVM is concatenating the unimodal features for the final classification on the IEMOCAP dataset [8].

Approach	Total Accuracy(%)
H-DMS	68.4
Uni-SVM	70.1
h-LSTM	74.1
sc-LSTM	75.2
bc-LSTM	75.6
Ours	**80.2**

Table 1: Total Accuracy for 5 emotions Recognization with 5-class (IEmocap)

	Accuracy (%) for each emotion from IEmocap with 5-class				
Approach	Ang	Hap	Sad	Neu	Fru
BoW+SVM	40.6	45.0	42.·2	31.7	44.2
CNN_{word} [17]	42.9	54.2	50.3	39.7	49.2
$LHAF_{wo} + SVM$[18]	41.2	36.6	38.3	39.2	49.2
$LHAF_{w} + SVM$[18]	40.2	37.1	40.2	40.1	41.8
CNN_{mel}[19]	39.7	41.2	43.5	39.1	41.4
$CNN_{word} + LHAF_{wo} + MKL$[20]	50.3	52.5	53.2	49.2	52.2
$CNN_{word} + CNN_{mfsc}$[21]	50.1	52.3	56.3	51.2	50.4
$CNN_{word} + CNN_{mfsc} + SVM$	51.2	50.8	55.3	51.7	51.4
H-DMS [9]	57.2	65.8	60.2	56.3	61.6
Ours	**79.4**	**85.3**	**90.2**	**71.9**	**77.6**

Figure 3: Accuracy for each emotion from IEmocap with 5-class

h-LSTM: Omit the dense layer after the LSTM cell to provide context-dependent features and the softmax layer providing the recognition on the IEMOCAP dataset.
sc-LSTM: Simple contextual LSTM consists of unidirectional LSTM cells on the IEMOCAP dataset [8].
bc-LSTM: A bidirectional LSTM structure to learn contextual information among utterances on the IEMOCAP dataset [8].

5.2 Comparisons

In this part, we show off the results of our model and make the comparison to other models presented in other paper. Since the limited time, we cant implement the models shows in the baseline by ourselves. Therefore, we directly use the results posted in their paper [8] and [9]. Table 1 and Figure 3 show the results of sentiment recognition from different kinds of models using IEmocap dataset.

5.2.1 Accuracy

In Figure 4 we show our test result of the trained model. As we mentioned, we separate dataset to 5 categories angry (0), happy+excited(1), sad(2), frustration(3), neutral(4). Take the first line under the final_acc as example. When testing with angry dataset with the amount of 232, the model recognize 181 of them as ang, 8 of them as happy+ excited, 3 of them as sad, 34 of them as frustration and 2 of them as neutral.

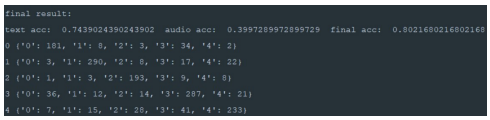

Figure 4: Test results

5.2.2 Training time

We also test the training time between the traditional method with LSTM & Attention and our way only with self-attention with the same dataset IEmocap. We use IEmocap dataset and separate the data to training data and test data with the scale as 4:1. For text training, the time for an epoch in LSTM & Attention is nearly one minute. For our method, it only takes 10 seconds. For an epoch in training audio branch, the time for LSTM & Attention is almost 1 hour and 20 minutes, and only 50 minutes for our model only with self-attention. And this is also a significant improvement in our model from the current works.

5.3 Analysis

From the result, we can see that our model performs best in comparison to other methods from the accuracy and time. It means we can use the self-attention to replace the traditional techniques with LSTM in multimodal sentiment analysis. I think self-attention will be more and more potent in multimodal researches. This project is just a start! And we will make more effort to optimize our model. Then we will do more rigorous test experiments against other methods to make our performance more outstanding in the later study.

6 Discussion and Future work

There may be several potential advancements to our project. Actually, because of the limit of the time, we dont attach the video branch onto our model, which is also very important to some extent. For example, for some clips of the conversations, the actors didnt say anything, so there will not be any text data at all. In such occasion, there are also audio data and video data. Therefore, we can solve such problems by adding video branches of data. However, it will be tough to think of a strategy to merge these three branches.

Also, another critical limitation for our project

is the amount of the multimodal emotion analysis dataset. We only try to use the IEmocap dataset. There are many other excellent datasets for us to use, such as MOSI [22], YouTube [23], EmotiW2 and MOSEI. We will make try more dataset, and I believe the project will be a much more interesting topic!

7 Acknowledgement

The libraries we used include:

- Python 3.6
- Keras 2.0.9
- TensorFlow 1.3.0
- sklearn 0.19.1
- numpy 1.15.3
- scipy 1.1.0

8 Conclusion

In our project, we attempt to introduce a multimodal emotion analysis model with self-attention and word-level feature fusion. Our model makes the forced word-level alignment between the texture and acoustic data, and extract the features with self-attention strategy. We introduce a weighted fusion model with self-attention and CNN's to merge texture and acoustic features to make the classification decision on emotions.

References

[1] Attention Is All You Need. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin (Submitted on 12 Jun 2017 (v1), last revised 6 Dec 2017 (this version, v5))

[2] openSMILE The Munich Versatile and Fast Open-Source Audio Feature Extractor, Florian Eyben, Martin Wllmer, Bjrn Schuller

[3] A model of saliency-based visual attention for rapid scene analysis. L. Itti ; C. Koch ; E. Niebur, IEEE Transactions on Pattern Analysis and Machine Intelligence (Volume: 20 , Issue: 11 , Nov 1998)

[4] Topic Modeling: Beyond Bag-of-Words. Hanna M. Wallach Cavendish Laboratory, University of Cambridge, Cambridge CB3 0HE, UK

[5] Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. Kristina Toutanova, Dan Klein, Christopher D. Manning, Yoram Singer

[6] Multimodal sentiment analysis with word-level fusion and reinforcement learning. Minghai Chen,Sen Wang, Paul Pu Liang, Tadas Baltruaitis, Amir Zadeh, Louis-Philippe Morency from Carnegie Mellon University, USA, ICMI 2017 Proceedings of the 19th ACM International Conference on Multimodal Interaction Pages 163-171.

[7] Convolutional Neural Networks for Speech Recognition. Ossama Abdel-Hamid Department of Electrical Engineering and Computer Science, Lassonde School of Engineering, York University, Toronto, Canada; Abdel-rahman Mohamed ; Hui Jiang ; Li Deng ; Gerald Penn ; Dong Yu

[8] In Context-Dependent Sentiment Analysis in User-Generated. Poria, S., Cambria, E., Hazarika, D., Majumder, N., Zadeh, A., Morency, L. P. 2017. Context-dependent emotion analysis in user-generated videos. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (Vol. 1, pp. 873-883).)

[9] DEEP MULTIMODAL LEARNING FOR EMOTION RECOGNITION IN SPOKEN LANGUAGE. Yue Gu, Shuhong Chen, Ivan Marsic, Department of Electrical and Computer Engineering

[10] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems

[11] http://practicalcryptography.com/miscellaneous /machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/

[12] Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167

[13] Min Lin1,2, Qiang Chen2, Shuicheng Yan21Graduate School for Integrative Sciences and Engineering2Department of Electronic & Computer Engineering National University of Singapore, Singapore

[14] https://machinelearningmastery.com /timedistributed-layer-for-long-short-term-memory-networks-in-python/

[15] Soujanya Poria, Iti Chaturvedi, Erik Cambria, and Amir Hussain. 2016. Convolutional mkl based multimodal emotion recognition and sentiment analysis. In Data Mining (ICDM), 2016 IEEE 16th International Conference on, pages 439448. IEEE.,S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. 2015.

[16] https://sail.usc.edu/iemocap/iemocap _publication.htm

[17] Kim, Y.: Convolutional Neural Networks for Sentence Classification. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). (2014)

[18] Poria, Soujanya, Erik Cambria, and Alexander Gelbukh. "Deep Convolutional Neural Network Textual Features and Multiple Kernel Learning for Utterance-level Multimodal Sentiment Analysis." Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015. [2] S. Poria, I. Chaturvedi, E. Cambria, A. Hussain. Convolutional MKL based multimodal emotion recognition and sentiment analysis. Proceedings of ICDM, Barcelona (2016)

[19] Zheng, W. Q., J. S. Yu, and Y. X. Zou. "An experimental study of speech emotion recognition based on deep convolutional neural networks." In Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on, pp. 827-831. IEEE, 2015.

[20] S. Poria, I. Chaturvedi, E. Cambria, A. Hus- sain. Convolutional MKL based multimodal emo- tion recognition and sentiment analysis. Proceedings of ICDM, Barcelona (2016)

[21] Y. Gu, X. Li, S. Chen, J. Zhang, and I. Mar- sic, Speech Intention Classification with Multimodal Deep Learning, Advances in Artificial Intelligence Lecture Notes in Computer Science, pp. 260271. (2017)

[22] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis- Philippe Morency. 2016. Mosi: multimodal cor- pus of sentiment intensity and subjectivity anal- ysis in online opinion videos. arXiv preprint arXiv:1606.06259.

[23] Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. 2011. Towards multimodal sentiment analy- sis: Harvesting opinions from the web. In Proceed- ings of the 13th international conference on multi- modal interfaces, pages 169176. ACM.