Internet Meme Emotion Analysis

CMPUT 651 Course Project

Shan Lu*, Shengyao Lu*, Xutong Zhao*

1 Introduction

Nowadays massive users of online social media tend to communicate via multimodal content beyond pure textual information. Increasingly growing amounts of emojis, emoticons, memes, audios, and videos are being created/used on major social media platforms (e.g., WeChat, Facebook, Twitter, etc) every day. Memes are a typical combination of textual and visual contents. They are created based on influential societal and/or cultural contexts, such as adventure movies, sci-fi literature, and sports. For instance, the popular One Does Not Simply is taken from the movie Lord of the Rings. People with such social and cultural backgrounds use memes to convey ideas in a smart and humorous way. Researchers in Natural Language Processing (NLP) and Computer Vision (CV) have achieved distinguished successes on social media study. However, approaches merely focusing on one single modality are developed with the inherent limitation that their performance often fails to generalize under compound information settings. Notwithstanding extensive usage on social media, emotion analysis on Internet memes have not drawn much attention from the NLP/CV research communities. A common approach to multimodal fusion is to concatenate all models and feed the concatenated multimodal feature to fully connected layers to perform the sentiment prediction task, but merely concatenating different models ignores the different contributions of modals to the target sentiment label. Since different models should have different degrees of sentiment effects, inspired by the self-attention mechanism(Lin et al., 2017), we propose an approach using the self-attention to compute the weights of image and text modals affecting the target sentiment, and we also use transfer learning to fine-tune the pre-trained models to get the image and text embeddings which leverages our model. To evaluate our model, we are going to use the meme dataset which has an image and text for each meme data. Hence, we are motivated to work on the SemEval task Memotion Analysis proposed by Dr. Amitava Das et al, which includes a multiclass sentiment prediction task for the meme data.

2 Related Work

As multimodal data has become viral on popular social media platforms, sentiment analysis on such complex information has drawn researchers' attention in the past few years.

The Deep Sentiment proposed by Hu and Flaxman (2018) demonstrates satisfactory performance in predicting emotion word tags attached to Tumblr posts. Deep Sentiment combines features learned from pretrained Inception (Szegedy et al., 2014) and GloVe word embeddings (Pennington et al., 2014), and it outperforms models that based merely on textual or visual information.

The Low-rank Multimodal Fusion method proposed by Liu et al. (2018) also focuses on learning multimodal representation. Moreover, they perform feature fusion with low-rank tensors, which help to prevent an increase in feature dimension and computational complexity it causes. They adopt GloVe word embeddings (Pennington et al., 2014) for textual features, FACET toolkit (Zhu et al., 2006), for visual features, and COVAREP framework (Degottex et al. 2014) for acoustic features. Experiments conducted on video datasets CMU-MOSI (Zadeh et al., 2016a), POM (Park et al., 2014b), and IEMOCAP (Busso et al., 2008) have shown competitive performance compared to the previous state-of-the-art models.

In addition, there are other previous studies on sentiment analysis on videos, such as the Recurrent Attended Variation Embedding Network (RAVEN) proposed by Wang et al., and the Interactive COnversational memory Network (ICON) proposed by Hazarika et al. The majority of research communities build multimodal sentiment analysis models for videos, where acoustic information plays a significant role. How-ever, there still have been few works developed for the analysis of Internet memes. We adopt the general framework for multimodal tasks, and provide a reliable sentiment analysis model for memes.

3 Approach

3.1 Image Model

We use transfer learning to obtain the image embeddings. The input image is passed into a pre-trained Inception v3 (Szegedy et al., 2015) that has been pretrained on 1000-class ImageNet dataset, and the last fully connected layer of the pre-trained model is reshaped to have the same number of outputs as the number of target classes in our dataset, then we apply the softmax function on the last fully connected layer to obtain a probability vector of the image where each entry corresponds to the probability of this image belongs to the corresponding target class.

Given an image x_{img}, it is reshaped to size 299×299 and normalized before feeding into the pre-trained Inception v3 model. The input to the last fully connected layer of the pre-trained Inception v3 is $h_{img} \in \mathbb{R}^{2048}$, and the output is a vector $\boldsymbol{emb_{img}} \in \mathbb{R}^{number\ of\ target\ classes}$:

$$emb_{img} = softmax(W_1 h_{img} + b_1)$$

Where W_1 is a weight parameter with the dimension $number\ of\ target\ classes - by - 2048$, and b_1 is a bias parameter which is a scalar.

Fine-tuning a pre-trained Inception v3 allows us to get a better representation of the image for our predicting task, and this alleviates our limited training data problem. The alternative approach to obtain the image embeddings is to build a deep CNN model and learn the parameters from scratch. But this alternative approach is not plausible, since we cannot learn a good model to capture the characteristics of image with our limited training data.

The reason why we reshape the last fully connected layer of the pre-trained model to have the same number of outputs as the number of target classes in our dataset and apply the softmax funct -ion on it is that we want to get a probability distribution of the image belonging to the target classes. This probability distribution of the image is used as the image embedding, and can be fusioned with the text modal to get a good result on our prediction task.

3.2 Text Model

We also use transfer learning to obtain the text embeddings. The input text is passed into a pre-trained RoBERTa (Liu et al., 2019) that has been pre-trained over 160GB of text, similar to our approach in the image modal, we add a fully connected layer with the number of outputs the same as the number of target classes in our dataset to the pre-trained RoBERTa, and apply the softmax function on this last fully connected layer to get a probability vector for the text where each entry corresponds to the probability of this text belongs to the corresponding target class.

Given a text x_{text}, the input feeding into the pre-trained RoBERTa model is a text string, and the pre-trained RoBERTa applies Byte-Pair Encoding (Sennrich et al., 2016) to the input text and it outputs a probability distribution for the text $\boldsymbol{emb_{text}} \in \mathbb{R}^{number\ of\ target\ classes}$.

Since the output is a probability distribution for the text belong to the corresponding target classes, it has the same range and dimension as the image embedding $\boldsymbol{emb_{img}}$ from the image modal, it is natural and easier to fusion these two embeddings together as the probability distribution of image-and-text for the target classes.

An alternative approach to obtain the text embedding is to use pre-trained word vectors such as GloVe. The text is tokenized and the corresponding word embeddings are fed into a biLSTM, the last hidden state of the biLSTM can be used as the text embedding. Although this alternative approach is faster than fine-tuning the pre-trained RoBERTa model, from our experiments, the text embedding from the pre-trained RoBERTa gives a much better result. We guess the reason why the pre-trained RoBERTa is much better for extracting the text features is that it is a deeper model and has been pre-trained on much larger number of data, so it can capture more characteristic and meaning of the text.

3.3 Attention-Based Image and Text Modals Fusion

Inspired by the self-attention mechanism (Lin

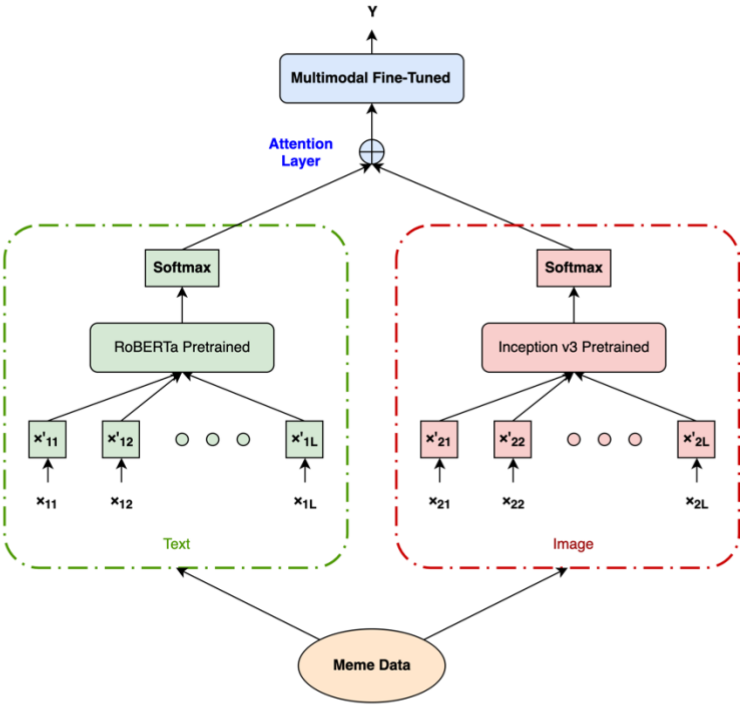

Figure 1: Attention-Based Image and Text Modals Fusion

et al., 2017), our approach to fusion the image and text modals as illustrated in Figure 1 is to use the self-attention mechanism, and the weighted sum of the image embedding $\boldsymbol{emb_{img}}$ and the text embedding $\boldsymbol{emb_{text}}$ is considered as the probability distribution of our prediction task.

Given the image embedding $\boldsymbol{emb_{img}} \in \mathbb{R}^{number\ of\ target\ classes}$ from the image modal and the text embedding $\boldsymbol{emb_{text}} \in \mathbb{R}^{number\ of\ target\ classes}$ from the text modal, the input to the attention layer is $\boldsymbol{E} = [emb_{img}, emb_{text}]$ with dimension $\boldsymbol{2} - by - number\ of\ target\ classes$, and the output of the attention layer is a vector of weights $\boldsymbol{a}$:

$$\boldsymbol{a} = softmax(\boldsymbol{W_{att}E^{T}})$$

Where $\boldsymbol{W_{att}}$ is a weight parameter with the dimension $1 - by - \#\ \boldsymbol{of\ target\ classes}$, and $\boldsymbol{a}$ is a vector of weights with the dimension $1 - by - 2$.

$$a_1 = exp(W_{att}emb_{img})/(exp(W_{att}emb_{img}) + exp(W_{att}emb_{text}))$$

$$a_2 = exp(W_{att}emb_{text})/(exp(W_{att}emb_{img}) + exp(W_{att}emb_{text}))$$

Therefore, the weighted sum of the image embedding $\boldsymbol{emb_{img}}$ and the text embedding $\boldsymbol{emb_{text}}$ is $\boldsymbol{emb_{img-text}}$:

$$emb_{img-text} = a_1 emb_{img} + a_2 emb_{text} = aE$$

Where $\boldsymbol{emb_{img-text}} \in \mathbb{R}^{1\times\#\ of\ target\ classes}$.

Since $\boldsymbol{emb_{img-text}}$ is a weighted sum of probability distributions, the probability of the target classes is

$$P(y|x_{img}, x_{text}) = \boldsymbol{emb_{img-text}}$$

And the predicted target class for this image and text pair (meme) is

$$y_{predicted} = argmax_{y_i \in target\ classes} P(\boldsymbol{y}|\boldsymbol{x_{img}}, \boldsymbol{x_{text}})$$

Using the attention mechanism, we can get the weights that the image embedding and the text embedding should be assigned to, and the weighted sum of the image and text embeddings gives a better combined probability distribution for our prediction task.

An alternative approach to fusion the image and text embeddings is to concatenate the image embedding and the text embedding, and then feed this concatenated vector into a fully connected layer with the number of outputs the same as the number of target classes in our dataset. But this alternative approach ignores the different effects of

	Overall_Sentiment ternary (3 classes)	**Overall_Sentiment (5 classes)**
Random Guessing	0.357	0.198
Logistic Regression	0.464	0.310
SVM - Linear Kernel	0.503	0.394
SVM - Gaussian Kernel	0.522	0.423
GloVe with Only Text Info	0.10	0.06
RoBERTa Finetuned with only Text Info	0.595	0.463
Inception-v3 - only Image Info	0.601	0.411
Deep Sentiment, 2018	0.575	0.425
Deep Sentiment Attention	**0.620**	**0.491**

Table 1: Comparison of the label prediction accuracy on Trial dataset with previous sentiment analysis and emotion recognition datasets

the image and text embeddings on predicting the target class.

4 Experiments

4.1 Datasets

The given training dataset contains 7K human-annotated Internet memes, each of which is labeled with sentiment, and types of emotions. The categories of emotions are humorous, sarcastic, and offensive. If a meme does not fall into any of these categories, then it is marked as a motivational meme. The emotion types are further quantified on a scale. For example, the Humour semantic class is quantified on the scale of not_funny, funny, very_funny, and hilarious. Other emotion types are quantified in a similar manner. The dataset will also contain the extracted captions/texts from the memes.

4.2 Experimental Setup

In our experiments, we seek to evaluate how the models perform on predicting the labels with the text and image information of the internet memes. Two kinds of the labels are evaluated. One is the ‘Overall_Sentiment ternary’ that contains three classes, ‘positive’, ‘neutral’ and ‘negative’. The other is the ‘Overall_Sentiment’ that provided by the competition organizer, which has five classes, ‘very positive’, ‘positive’, ‘neutral’ , ‘negative’, and ‘very negative’.

For the baseline purposes, we are going to use Random Guessing; Logistic Regression with the mean squared error loss function; Linear Kernel SVM and Gaussian Kernel SVM. The features for our input data are going to be the simple concatenation of the GloVe word embedding pre-trained on Twitter data and image pixels of three channels. And for the evaluation metrics, we are going to use test accuracy over ten runs to evaluate the baseline models, some previous approaches that other people purposed, and our approach. The performance on predicting the label with only the text information with GloVe or RoBERTa is also evaluated separately. Besides, the Deep Sentiment proposed by Hu and Flaxman (2018) is implemented and applied on our dataset, which concatenate the image features extracted using the pretrained Inception-v1 model and the text embedding extracted using GloVe.

For the Deep Sentiment Attention, which is the approach that we proposed, we combined the pretrained models RoBERTa large and Inception-v3 with an additional attention layer. And then the entire model is fine-tuned over five epochs before the prediction on the test dataset, which is held out by a fraction of 0.2 from the original dataset. The overall accuracy is calculated by the average over ten runs.

4.3 Results and Discussion

Since we are currently in the midterm phase of the project, we haven’t run the experiments on the whole dataset, but on the small version dataset, which contains 1k data instead of 7k, so called ‘Trial Data’.

The experimental results on the ‘Trial Data’ are presented in Table 1. Deep Sentiment Attention outperforms the baseline SVM with Gaussian kernel by 9.8% and Deep Sentiment by 4.5% in predicting the ‘Overall_Sentiment ternary’ label. We demonstrate that neither a single pretrained RoBERTa large model nor a single pretrained Inception-v3 model can outperform the Deep Sentiment Attention model. Besides, we presents our attention-based approach is more effective than the Deep Sentiment (Liu et al. 2018) that is based on concatenation. The Deep Sentiment Attention contributes a significant improvements on internet meme emotion analysis.

Acknowledgements

This document is the midterm report of the CMPUT 651 course project, which does analysis on the internet meme emotion. The format and the full experimental results will be will be modified in the final report. The repo of this project can be found at:
https://github.com/TongTongX/CMPUT651Project
which can be made public upon request.

References

C. Busso, M. Bulut, C.C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, and S.S. Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. Journal of Language Resources and Evaluation, 42(4):335–359.

Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. 2014. COVAREP – A collaborative voice analysis repository for speech technologies. In IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014, Florence, Italy, May 4-9, 2014, pages 960–964.

Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann. 2018. ICON: Interactive conversational memory network for multimodal emotion detection. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2594–2604, Brussels, Belgium. Association for Computational Linguistics.

Anthony Hu and Seth Flaxman. 2018. Multimodal sentiment analysis to explore the structure of emotions. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery &#38; Data Mining, KDD ’18, pages 350–358, New York, NY, USA. ACM.

Zhouhan Lin, Minwei Feng, C´ıcero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. CoRR, abs/1703.03130.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. CoRR, abs/1907.11692.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2018. Efficient low-rank multimodal fusion with modality-specific factors. CoRR, abs/1806.00064.

Sunghyun Park, Han Suk Shim, Moitreya Chatterjee, Kenji Sagae, and Louis-Philippe Morency. 2014. Computational analysis of persuasiveness in social multimedia: A novel dataset and multimodal prediction approach. In Proceedings of the 16th International Conference on Multimodal Interaction, ICMI ’14, pages 50–57, New York, NY, USA. ACM.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2014. Going deeper with convolutions. CoRR, abs/1409.4842.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2015. Rethinking the inception architecture for computer vision. CoRR, abs/1512.00567.

Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2018. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. CoRR, abs/1811.09362.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. MOSI: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. CoRR, abs/1606.06259. Qiang Zhu, Mei-Chen Yeh, Kwang-Ting Cheng, and Shai Avidan. 2006. Fast human detection using a cascade of histograms of oriented gradients. In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2, CVPR ’06, pages 1491–1498, Washington, DC, USA. IEEE Computer Society.