Iran University of Science and Technology
Deep Learning (Fall 2019)

Danial Kamali
Department of Computer Engineering
Iran University of Science
and Technology
daniel.kamali.dk@gmail.com

Abstract

Information on social media comprises of various modalities such as textual, visual and audio. NLP and Computer Vision communities often leverage only one prominent modality in isolation to study social media. However, computational processing of Internet memes needs a hybrid approach. The growing ubiquity of Internet memes on social media platforms such as Facebook, Instagram, and Twitter further suggests that we can not ignore such multi-modal content anymore. To the best of our knowledge, there is not much attention towards meme emotion analysis.The objective of this paper is to purpose a solution for the task of Memotion analysis to classify them by their sentiment,Data is provided by SemEval in CodaLab Competition

1 Introduction

In the last few years, the growing ubiquity of Internet memes on social media platforms such as Facebook, Instagram, and Twitter has become a topic of immense interest. Memes, one of the most typed English words (Sonnad, 2018) in recent times. Memes are often derived from our prior social and cultural experiences such as TV series or a popular cartoon character (think: One Does Not Simply - a now immensely popular meme taken from the movie Lord of the Rings). These digital constructs are so deeply ingrained in our Internet culture that to understand the opinion of a community, we need to understand the type of memes it shares. (Gal et al., 2016) aptly describes them as performative acts, which involve a conscious decision to either support or reject an ongoing social discourse. Online Hate - A brutal Job: The prevalence of hate speech in online social media is a nightmare and a great societal responsiblity for many social media companies. However, the latest entrant Internet memes (Williams et al., 2016) has doubled the challenge. When malicious users upload something offensive to torment or disturb people, it traditionally has to be seen and flagged by at least one human, either an user or a paid worker. Even today, companies like Facebook and Twitter rely extensively on outside human contractors from start-ups like CrowdFlower, or companies in the Philippines. But with the growing volume of multimodal social media it is becoming impossible to scale. The detection of offensive content on online social media is an ongoing struggle. OffenseEval (Zampieri et al., 2019) is a shared task which is being organized since the last two years at SemEval. But, detecting an offensive meme is more complex than detecting an offensive text – it involves visual cue and language understanding. This is one of the motivating aspects which encourages us to propose this task. Multimodal Social Media Analysis - The Necessity: Analogous to textual content on social media, memes also need to be analysed and processed to extract the conveyed message. A few researchers have tried to automate the meme generation (Peirson et al., 2018; Oliveira et al., 2016) process, while a few others tried to extract its inherent sentiment (French, 2017) in the recent past. Nevertheless, a lot more needs to be done to distinguish their finer aspects such as type of humor or offense.

(LateX template borrowed from NeurIPS 2019)

The Memotion Analyis Task A Sentiment Classification: Given an Internet meme, the first task is to classify it as positive, negative or neutral meme.

The most Challenging problems that faced during in this paper were cleaning some of data and dealing with such an unbalanced dataset.

2 Related work/Background

Traditional methods for sentiment analysis are mainly applied to text mining, which do not consider the presence of multimodal data, e.g., videos or images. As one of the popular data format, images present more information but are more complex in comparison to text. In 2012, Siersdorfer et al. predicted sentiment of images using color histograms and Scale-Invariant Feature Transform (SIFT) techniques dataset with more than half a million Flickr images. They used SentiWordNet as query terms to gather images with sentiment orientations. The bag-of-visual words representation and the color distribution of images are used to learn the image features. Through studying the connection between the sentiment of images expressed in metadata and their visual content, Siersdorfer et al. achieved the precision values of up to 70% but with low recall values. Zhang et al. processed Sentiment Analysis on Microblogging by integrating text and image features [3]. In 2016, Katsurai et al. proposed a method mapping visual, textual and sentiment views into the latent embedding space and using correlations among these features [4]. The visual features are learned from a color histogram of images and this method achieved an accuracy of 74.77% on Flickr dataset and 73.60% on the Instagram dataset.

Text-Image Sentiment Analysis[2]considers the application of Twitter images with captions for the prediction of sentiments applying fine-tune techniques. The Twitter images have corresponding labels or tweets, hence the merging of features from images and text is proposed. In this way, we can predict image sentiment as positive or negative with better performance. We see that the accuracy after fusing text and image features is higher than using a single modality. To extract the image features they consider AlexNet, which is a previously trained deep convolutional architecture. For text features, we extract the significant concepts and project them on the AffectiveSpace of emotions. Lastly, they propose a novel sentiment score to combine the prediction from image and text features. A Multimodal Approach to Image Sentiment Analysis [1] explores the sentiment analysis of tweets that contain both texts and images, focusing on images and their content. they achieve a result on the isolated method image that exceeds the baseline method for the same theme in the paper Cross-media learning for image sentiment analysis in the wild[5] they built a probability distribution table, that is based on 1000 classes of the ImageNet, that summarise a probability of a given image being negative, neutral or positive according to its content. Finally, they built a method that can classify multimedia content with text and image and generate a sentiment classification based on the image content.

3 Proposed method

In this paper we purpose the new approach to text-image sentiment classification by merging the data from input image and text using deep neural network

3.1 Image

Most common technique to extract information from image is ConvNets therefore we Used Deep-CNN network to extract informations from images in dataset multiple approaches have been tried like pure multi-layer ConvNet and VGG base model with trainable weights and non-trainable weights. The ConvNet is sollowed by two dense layer to extract info from features in last flatten layer.

3.2 Text

To extract information from text inputs we use Glove 300d Word2Vec Tokenization and vocabaulary indexing as pre-processing and Embedding weights as inital weights of non-trainble Embedding layer.
Before embedding layer we put Maksing layer to mask zeros from preprocessing.After The Embedding layer There is three Conv1D layers to extract patters from embedding and it is following by

Table 1: Classifications results.

	Accuracy
image-convnet	55
baseline	59
image& w2v	64
Glove 100d& bi-lstm	68
conv1D & Glove 100d	72.13
VGG & part-layer	87.7

Three Layers of Bidirectional LSTM to cover Time Series Pattern in any possible way.
For next layer we use dense layer for two reasons:

1. covering different in output of text and image network
2. extracting text-specified feature by dense layer

3.3 Final Dense Layers

Image and Text network has been connected to multi-layer dense part to extract the sentiment from image and text results

At primitive experiments the dense layers in Text and Image parts wasn't exists therfore and those layers was part of final dense layers that won't let network to learn more 68% accuracy

- Any kind of preprocessing or normalization (if you have used).
- The architecture of the neural network that you have used. It is better to demonstrate this with a graph.

4 Results

In the process of Doing this task serveal approaches has been choosen but the we will show some of the most intresting results of approaches on data trial dataset that is provided by Semeval for evaluating our processes,The first approach was just using images and trainable convNet to classify the sentiment of memes that just couldn't even beat the base line of 59% accuracy in average next approach was to add text with embedding layer of Google Word2Vec and single layer of lstm which concatenated to images and followed by dense layer and it could finally beat the accuracy to 64% ,for the next level i use Glove embedding with 100d with multiple lstm layer that made 68% accuracy. for next step i changed 100d to 300d and add multiple conv1D layers and replaced lstms with bidirectional lstm that make network to learn until 72% , at next step replaced convNet with pretrained VGG16 with imagenet weights and add remove some inital layer of final dense layer and add them to end of lstms and vgg network that make my network to boost alot and could beat 87.5% accuracy on trial data.

5 Discussion

As you can see in figure 1 the training and validation correlated which demonstrate the truth and validation of model. the model perposed in previous paper needs attenion layers but this model is attention-less those specific layer after conv and lstm layers make such a difference from other approaches.This model can mostly understand the meaning of the sentences of Meme and extract sentiment by associating of the CNN.

6 Feuture Work

I really liked to use BERT models for word verctotization and see the results also we could imporve the performance of model by fine tuning the VGG and word embedding model but due to lack of time these modifications didn't evaluated.

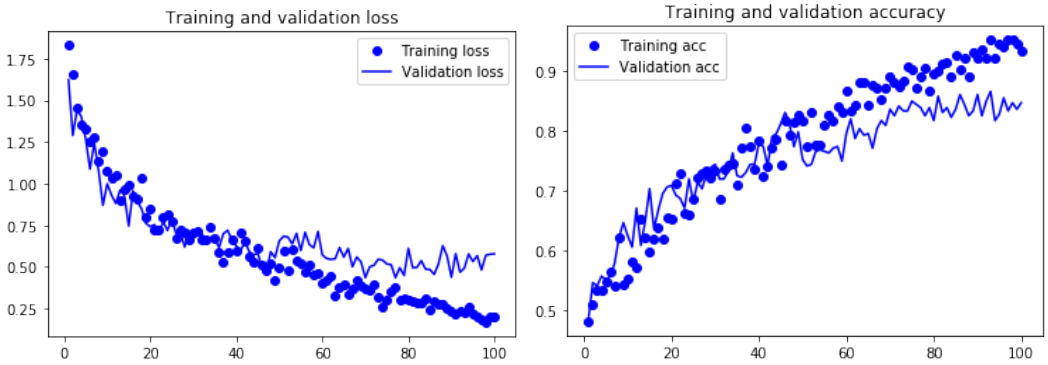

Figure 1: loss and acc

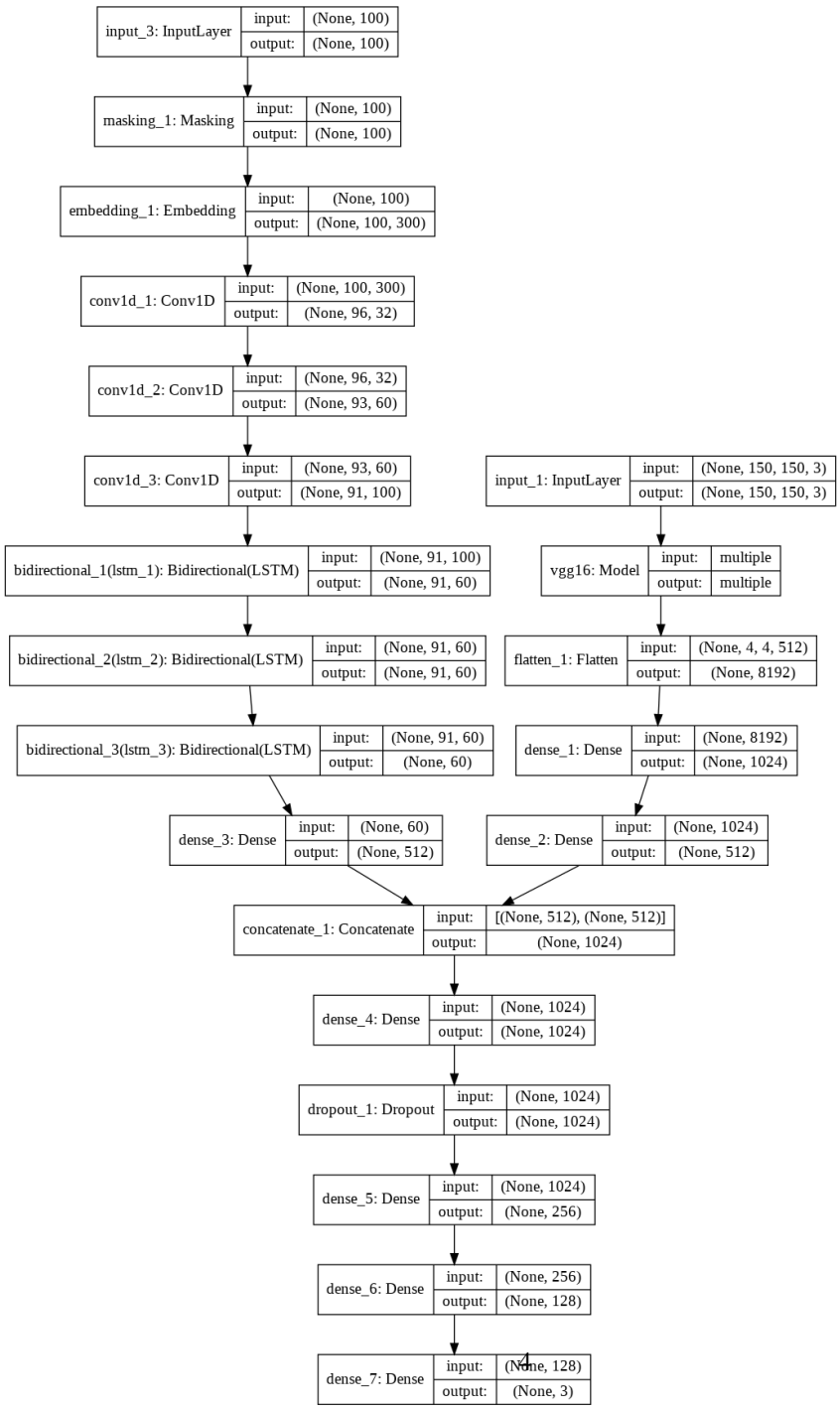

Figure 2: final model schematic

References

[1] A Multimodal Approach to Image Sentiment Analysis Antonio Gaspar and Luıs A. Alexandre Universidade da Beira Interior

[2] Text-Image Sentiment Analysis Chen Qian, Edoardo Ragusa, Iti Chaturvedi ,Erik Cambria1, and Rodolfo Zunino

[3] Zhang, Y., Shang, L., Jia, X.: Sentiment analysis on microblogging by integrating text and image features. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer (2015) 52–63 [4]Katsurai, M., Satoh, S.: Image sentiment analysis using latent correlations among visual, textual, and sentiment views. In: Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on, IEEE (2016) 2837–2841 [5] [Vadicamo, L., Carrara, F., Cimino, A., Cresci, S., Dell'Orletta, F., Falchi, F., Tesconi, M.: Cross-media learning for image sentiment analysis in the wild. In: 2017 IEEE International Conference on Computer Vision Workshops (ICCVW). pp. 308–317 (Oct 2017). https://doi.org/10.1109/ICCVW.2017.45]