Co-attention based Multimodal Factorized Bilinear Pooling for Internet Memes Analysis

Gitanjali Kumari[1] **Amitava Das**[2] **Asif Ekbal**[1]
[1]Department of Computer Science and Engineering,
[1]Indian Institute of Technology Patna, India
[2]Wipro AI Labs, Bangalore, India
{gitanjali_2021cs03,asif}@iitp.ac.in
amitava.das2@wipro.com

Abstract

Social media platforms like Facebook, Twitter, and Instagram have a significant impact on several aspects of society. Memes are a new type of social media communication found on social platforms. Even though memes are primarily used to distribute humorous content, certain memes propagate hate speech through dark humor. It is critical to properly analyze and filter out these toxic memes from social media. But the presence of sarcasm and humor in an implicit way makes analyzing memes more challenging. This paper proposes an end-to-end neural network architecture that learns the complex association between text and image of a meme. For this purpose, we use a recent SemEval-2020 Task-8 multimodal dataset. We proposed an end-to-end CNN-based deep neural network architecture with two sub-modules viz.(i) Coattention based submodule and (ii) Multimodal Factorized Bilinear Pooling(MFB) submodule to represent the textual and visual features of a meme in a more fine-grained way. We demonstrated the effectiveness of our proposed work through extensive experiments. The experimental results show that our proposed model achieves a 36.81% macro F1-score, outperforming all the baseline models.

1 Introduction

Social media such as Facebook, Twitter, Instagram, etc., are interactive platforms that accelerate the idea of creating and sharing information. This information made an enormous impact in different fields of society more powerfully and effectively. But on the other hand, we observe a significantly large amount of offensive content in the various social networking sites, which spread hatred, rumors, etc., between the different communities, groups, or individuals. Meme (Dawkins, 2016) is the form of multimodal media that has been initially created to spread humorous content, but due to its multimodal nature, some memes help users to spread hate speech in the form of dark humor. On social media, posting such memes to troll, cyberbully, or targeting someone is increasing rapidly. Unlike other multimodal tasks (e.g., Visual Question Answering, Image Captioning, etc.), in sentiment analysis for memes, textual and visual information are very weakly semantically aligned. In such a situation, we cannot uncover the complex meaning of hateful content until we get to know both the modalities and their contributions in any content hateful. Analysis of such memes can bring valuable insights that are not explored yet. For example, if there is a meme with text containing "Look how many people love you." The sentiment of this meme can, itself, be positive, negative, or neutral. The sentiment can only be found if and only if we add an image to it (c.f. Figure 1). Some memes are purely humorous, while others spread offensive content in the form of dark humor, sarcasm, mockery, etc. Sentiment analysis of memes in a more effective way will facilitate combating such social media issues.

Figure 1: A meme where only after focusing on both text and image, negative sentiment can be identified.

With the phenomenal growth of social media

Proceedings of the 18th International Conference on Natural Language Processing, pages 261–270
Silchar, India. December 16 - 19, 2021.

networks, sentiment analysis plays a significant role in handling various aspects of political and religious views of society. Sentiment analysis research has been a progressive area in Natural Language Processing (NLP). It is ranging from document-level classification (Lin and He, 2009; Mouthami et al., 2013) to learning the word and phrase polarity(Hatzivassiloglou and McKeown, 1997; Esuli and Sebastiani, 2006). Several supervised machine learning and feature-based techniques have been used to tackle this problem (Lai et al., 2015; Kouloumpis et al., 2011). However, deep learning-based techniques have gained a lot of popularity in recent years (Truong and Lauw, 2019). A significant amount of works have been done which includes the analysis of opinions about hotel reviews ((Kasper and Vela, 2011; Shi and Li, 2011)), product reviews ((Cernian et al., 2015; Wei and Gulla, 2010; Fang and Zhan, 2015)) etc. Initially, sentiment analysis has been carried out mostly using text (Badjatiya et al., 2017; Davidson et al., 2017; Fortuna et al., 2019). Recently due to the growth of multimedia contents in social media, we also need to develop robust models that would deal with multimodal content. There have been very few attempts towards analyzing the sentiment of memes by researchers. However, research shows that there is a far way to go if we compare the system-generated output to the human evaluation. There can be several reasons for this, such as hateful meaning hidden behind humor, sarcasm, the use of very twisted words, or image to spread hate (Sharma et al., 2020). Lack of annotated datasets can also be one of the reasons for this kind of failure.

The key attributes of our current work can be summarized as follows: (i) We develop a deep neural network-based architecture to explore the idea of co-attention to identify the impact of text and image simultaneously for predicting the correct sentiment of a given meme.(ii) We also explore the concept of Multimodal Factorized Bilinear pooling to represent the textual and visual features in a internet meme analysis dataset. We test the significance of our proposed method on SemEval2020 (Sharma et al., 2020) dataset. Evaluation results the accuracy and macro-F1 of 54% and 36.81%, respectively, which are higher than the baseline model for the given task.

2 Related Work

This section briefly discusses the review related to two aspects: $a)$ Sentiment analysis for unimodal data, $b)$ Sentiment analysis for multimodal data.

2.1 Sentiment analysis in unimodal data

With the emergence of social media and vast internet content, Sentiment analysis has received much attention in the Natural Language Processing (NLP) community. It is very useful on topical categorization task to sort documents according to their subjects such as economics or politics ((Ali et al., 2019; Ilyas et al., 2020)). The work reported in (Kouloumpis et al., 2011) investigated the usefulness of some linguistic features and other features to get an idea about the informal and creative language used in microblogging. Similarly, authors in (Agarwal et al., 2011) proposed to use Part-of-Speech (PoS)-specific prior polarity features to examine sentiment analysis on Twitter data. (Hu and Flaxman, 2018) pioneered HEMOS (Humor-EMOji-Slang-based), a kind of fine-grained sentiment analysis system for the Chinese language to investigate the significance of perceiving the impact of humor, pictograms, and slang to affect users on social media. To address the challenge of sentiment reflection prediction in visual content,(Borth et al., 2013) initiated a data-driven systematic approach by using psychology theories to construct a Visual Sentiment Ontology(VSO) which is a collection of 3,000 Adjective Noun Pairs (ANP) to construct SentiBank, a mid-level concept representation of each image to characterize the sentiment reflected in any visual content. Similarly, we also see a few works on aggression detection from the given textual data (Kumar et al., 2018; Xu et al., 2012).

2.2 Sentiment analysis in multimodal data

Although multimedia content is significantly growing on social media, it is a great challenge to uncover the underlying sentiment mentioned in these. Multimodal sentiment analysis for detecting the polarity of image and text is similar to finding out the hateful content in internet memes. It is also observed that deep learning techniques significantly outperform when it is compared to the traditional machine learning approaches on multimodal data (Kumar et al., 2020; Tran and Cambria, 2018). VistaNet (Lecun et al., 2015) shows the significant importance of visual knowledge in the visual and

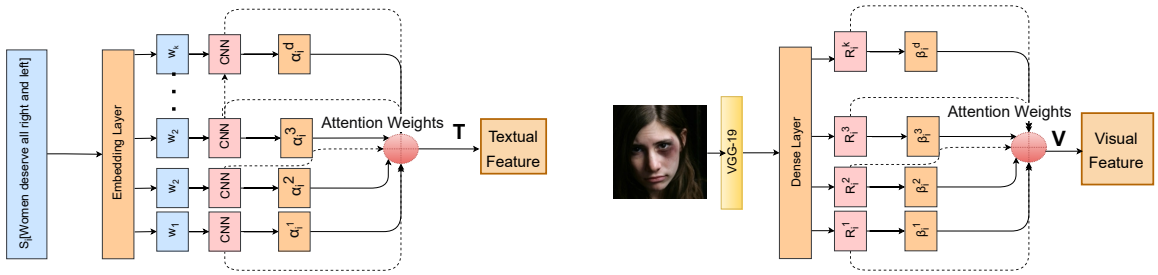


Figure 2: Textual and image feature after applying self-attention

textual content for detecting a sentiment of a document. The research reported in (Yang et al., 2019) tried to explore several deep learning techniques to integrate textual and visual parts of a meme.

The authors in (Yu et al., 2019) pioneered a cascade of Modular Co-Attention Network (MCAN) with a cascade of modular co-attention (MCA) layers, each of which consists of the self-attention and guided-attention units to model the intra and inter modal interactions synergistically. Furthermore, a Dynamic Co-attention Network(DCN) for VQA was introduced in (Xiong et al., 2018). In a paper, (Sabat et al., 2019) reported how visual modality can bring more information than linguistic information for the hateful memes classification task. A hierarchical method for multimodal processing of features using deep learning techniques has shown good result(Majumder et al., 2018). Authors in a paper (**?**) introduced the idea of multimodal factorized bilinear pooling with co-attention to demonstrate that MFB with co-attention on the real-world VQA dataset achieves new state-of-the-art performance.

Based on the above literature survey, we understood the need to develop such a robust model that can quickly identify the sentiment of a given meme. In our work, we explored the significance of the co-attention and MFB mechanism on the multimodal sentiment analysis task.

3 Methodology

Our current task aims at determining the sentiment of a given meme in a multimodal dataset. The problem can be defined as follows: Given every meme M_i in the dataset which is a combination of text $T_i = (t_{i1}, t_{i2}, ..., t_{ik})$ and image I_i with the shape (224,224,3) in RGB pattern, our task is to create one classifier that should predict one correct label $Y \subseteq$ {neg,neu,pos} for M_i i.e. predict the correct sentiment whether a given meme is negative, neutral or positive. The respective optimizing goal is then to learn the parameter θ and get the optimum loss function $L(Y|M, \theta)$.

At first, we develop a unimodal baseline system for text and image each. Finally, different multimodal approaches for the fusion of both modalities, i.e., textual and visual, have been described in the following sections:

3.1 Embedding Layer

At first, the pre-processing is performed on the texts, which include stop-word removal and lowercasing of tokens.After pre-processing, every word of each sentence $S_i = (w_{i1}, w_{i2}, ..., w_{ik})$ where k is the max_length of the sentence, is represented using its semantic representation. Each word w_{ij} is transformed into a pre-defined size of the vector, which contains the semantic meaning of that word, known as the word embedding vector. In our experiment, we use FastText (Bojanowski et al., 2017) word embedding for the same purpose. This embedding vector is passed through a deep neural network-based classifier for the sentiment analysis task further.

3.2 Textual Features

We use the convolutional neural network(CNN) (Simonyan and Zisserman, 2015) architecture for identifying the sentiment of a given textual part of the meme. The CNN model consists of three layers, namely convolutional, pooling, and fully connected layers. Textual features are extracted from the fully connected layer. For our experiment, we use three convolution layers with filter sizes 2, 3, and 4. Each convolution layer consists of 128 filters. Equation 1 shows the textual feature vector T_i of a sentence S_i after passing it through

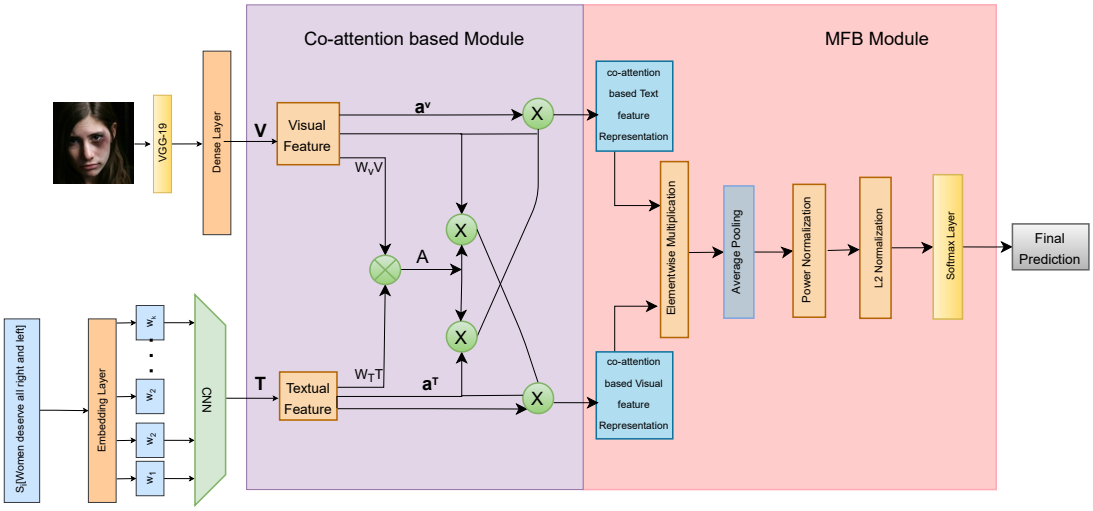


Figure 3: Our proposed co-attention with Multimodal Factorrized Bilinear Pooling based Model

convolution neural network.

$$T_i = (t_i^1, t_i^2, ...t_i^d) \tag{1}$$

3.3 Textual Features with self-attention

On top of this textual feature, we use the attention mechanism. For a given sentence S, the attention model finds out the most important words with the help of attention weights, which is beneficial for the decision-making purposes.

The text feature after applying attention is the weighted sum of all the words present in the sentence S_i. It uses attention weights of each source word, as given in equation 2.

$$c_i^k = \sum_{j=1}^{d} \alpha_{ij} h_j \tag{2}$$

We find out the attention score α_i^j for every feature representation w_{ij} of each word t_i^j in the sentence S_i which is given in equation 3.

$$\alpha_i^j = \frac{exp(e_i^j)}{\sum_{j=1}^{d} exp(e_i^j)} \tag{3}$$

where,

$$e_i^j = \theta(W t_i^j + b) \tag{4}$$

3.4 Visual Features

For extracting the visual features, we use the pre-trained VGG-19 model, which is trained on Imagenet (Simonyan and Zisserman, 2015) dataset. Image with the input shape (224*224) is given to the VGG19 architecture. In VGG19, we kept all the lower layers frozen and extracted the output of the $block5-conv4$ layer. The extracted output has 196 regions, and 512 dimensions represent each region. So, finally, we obtain a region feature with (196*512) dimensions passed to one dense layer with 250 neurons.

3.5 Visual Features with self-attention

The output from the dense layer mentioned in Section 3.4 is passed through the attention layer to obtain the most important regions that play a vital role in image classification.

Equation 5 shows the region feature of an image I_i

$$R_i = (R_i^1, R_i^2, ..., R_i^k) \tag{5}$$

We apply attention on top of the textual features in Section (3.2). Similarly, we use attention on the top of the region feature R_i of an image. After attention, the visual feature is the weighted sum of all the regions present in the Image V_i. It uses attention weights of each region, as given in equation 6.

$$c_i^k = \sum_{j=1}^{d} \beta_{ij} R_j \tag{6}$$

We find out the attention score β_i^j for every region R_i^j in R_i which is given in equation 7.

$$\beta_i^j = \frac{exp(h_i^j)}{\sum_{j=1}^{d} exp(h_i^j)} \tag{7}$$

where,

$$h_i^j = \theta(W R_i^j + b) \tag{8}$$

Table 1: Result of models

Model	Modality	Fusion	F1-Score	Accuracy
Model 1	SemEval2020 baseline	-	0.2176	–
Model 2	Only Text	-	0.3278	0.40
Model 3	Text+Self-Attention	-	0.3397	0.49
Model 4	Only Image	-	0.2936	0.34
Model 5	Image+Self-Attention	-	0.3166	0.37
Model 6	Text+ Image	early fusion with concatenation	0.3222	0.51
Model 7	Proposed Model1	Co-attention model Bilinear Pooling	**0.3532**	**0.52**
Model 8	Proposed Model2	Co-attention with MFB	**0.3681**	**0.54**
(Keswani et al., 2020)	SemEval2020(SOTA)	-	0.3546	–

$$\sum_{t=1}^{T} P(\hat{y}_0(x,t)|x).s(\hat{y}_0(x,t)) \tag{9}$$

3.6 Fusion of textual and visual features for baseline model

After extracting the textual and visual features separately, we use a fusion technique for our baseline model where we merely concatenate both textual T_i and visual feature V_i. This concatenated feature vector is passed through one dense layer which follows one softmax layer. The softmax layer gives the probability distribution for each class to classify the given meme into pre-defined categories.

3.7 Proposed model

Co-attention: Along with self-attention for textual and visual features, we also explore the concept of co-attention to introduce a natural symmetry between text and image where image representations guide the textual attention and textual representation guide the visual representation(Lu et al., 2016).

For a given textual feature T $\in \mathrm{R}^{(d\times T)}$ and given a visual feature T $\in \mathrm{R}^{(d\times V)}$, we calculate a similarity matrix representation called as affinity matrix A $\in \mathrm{R}^{(T\times V)}$ as follows:

$$A = \tanh\left(T^T W_b V\right) \tag{10}$$

Using the affinity matrix A in equation 10, we calculate the textual and visual attention maps in the following way:

$$\begin{aligned} H_V &= \tanh\left((W_t T)A + W_v V\right) \\ a^V &= \mathit{softmax}(w_{hv}^T H_V) \end{aligned} \tag{11}$$

$$\begin{aligned} H_T &= \tanh\left((W_t T + (W_v V)A_T\right) \\ a^T &= \mathit{softmax}(w_{ht}^T H_T) \end{aligned} \tag{12}$$

Here, $W_t, W_v \in \mathrm{R}^{(k\times d)}$ and $\mathrm{w}_{ht}^T, w_{hv}^T$ are weight matrix. a^V and a^T are the attention probabilities of image and textual part, respectively.

After that, we calculate the textual (T_V) and visual (I_V) attention vector, which is the weighted sum of textual and visual features.

$$T_V = \sum_{t=1}^{T} a_i^T T_i \tag{13}$$

$$I_V = \sum_{i=1}^{N} a_i^V V_i \tag{14}$$

3.7.1 Fusion of textual and visual features with bilinear Pooling

In the earlier fusion techniques (e.g.concatenation of both feature vectors, element-wise multiplication), the system could not fully interact with multimodal features. In the case of bilinear pooling, the system fully captures the complex association between image and textual features to get a more fine-grained classification decision. Each element of the textual feature interacts with every element of the visual feature using an outer product. The outer product of two vectors $T_i = (t_1, t_2, ...t_m)$ and $V_i = (v_1, v_2, ...v_n)$ can be defined as $\otimes$ which results in a matrix P $\in \mathrm{R}^{(m\times n)}$. Here, T_i in $\mathbb{R}^m$ is the textual feature vector, and V_i in $\mathbb{R}^n$ is the visual feature vector.

$$M = T_i \otimes V_i = T_i \times V_i^T \tag{15}$$

where M_{m*n} is the output of the bilinear model.

3.7.2 Fusion of textual and visual features with Multimodal Factorized Bilinear(MFB)

Although bilinear pooling adequately captures element-wise interactions between feature dimen-

sions, it does so at the expense of a set of parameters, which can result in significant computing costs and the risk of over-fitting. The Multimodal Factorized Bilinear(MFB) module may be used to fix this problem effectively. The association between textual and visual feature representations is maximised using this fusion mechanism. Given two feature vectors $T_i = (t_1, t_2, ...t_m)$ and $V_i = (v_1, v_2, ...v_n)$, where T_i is textual feature and V_i is visual feature of a meme. We can easily compute the bilinear pooling of these two vectors as follows:

$$M = T_i^T W_i V_i \tag{16}$$

where $W_i \in \mathrm{R}^{(m \times n)}$ is a projection matrix and M_i is the output of the bilinear model.

Furthermore, the projection matrix W_i in Eq.16 can be easily factorized into low-rank matrices.

$$\begin{aligned} M = T_i^T X_i Y_i^T V_i &= \sum_{d=1}^{k} T_i^T x_d y_d^T V_i \\ &= 1^T (X_i^T T_i \circ Y_i^T V_i) \end{aligned} \tag{17}$$

where k is the latent dimensionality of the factorized matrices X_i=$[x_1, x_2, ..., x_k]$ $Y_i = [y_1, y_2, ..., y_k]$, $\circ$ is the Hadamard product of two vectors, $1 \in \mathrm{R}^k$ an all-one vector. In order to obtain the output feature $M_{TV} \in \mathrm{R}^o$ by Eq.17, weights to be learned are two three-order tensors $X = [X_1, X_2, ..., X_o] \in \mathrm{R}^{(m \times k \times o)}$ and $Y = [Y_1, Y_2, ..., Y_o] \in \mathrm{R}^{(n \times k \times o)}$. We can easily reformulate X and Y vectors in $2-D$ matrices X'$\in \mathrm{R}^{m \times ko}$ and Y'$\in \mathrm{R}^{n \times ko}$ easily with simple reshape operation. We can then write Eq.17 as the following:

$$M_{TV} = \textit{AvgPool}(X_i^T T_i \circ Y_i^T V_i, k) \tag{18}$$

$$M_{TV} = sign(M_{TV}) |M_{TV}|^{0.5} \tag{19}$$

$$M_{TV} = (M_{TV})^T / ||M_{TV}|| \tag{20}$$

where AvgPool in Eq.18 is the average pooling over M_{TV}. Furthermore, to reduce the cost of variation in the magnitude of output neurons due to element-wise multiplication, Power-Normalization in Eq.19 and l_2-Normalization in Eq.20 is introduced to the MFB module. These operations restrict the model to go under the state of local minima.

We can formulate the class prediction for a given meme M_i using Softmax as the activation function in the final output layer as:

$$\hat{y} = P(Y_i | M_i, W, b) = \textit{softmax}(M_i W_i + b_i) \tag{21}$$

where, $\hat{y}$ is the prediction probability of selecting the i_{th} class (Y_i) given M_i, bias b_i, and weight matrix W_i ($i \in$(neg,neu,pos)). We use the categorical cross entropy as loss function with the following formula:

$$\mathcal{L} = -\sum [y \log \hat{y} + (1-y) \log(1-\hat{y})] \tag{22}$$

where, y is the original class and $\hat{y}$ is the predicted class of the meme.

3.8 Models

For our experiment, we develop the following models. The first is the official baseline model from SemEval, whereas the others are different variations of our proposed system.

Model 1 (Baseline Model) This model is the baseline model reported in SemEval2020 Task8 Subtask-A paper(Sharma et al., 2020). This model uses CNN + BiLSTM framework to extract the textual features and VGG-16 to extract the visual elements.

Model 2 (Only Text) The first model is the baseline model for the text part of memes. We use the CNN architecture to obtain the textual features. The framework of this model is discussed in Section 3.2. This textual feature is passed through the output layer with one softmax activation for the final prediction.

Model 3 (Text+ self-Attention) In this model, we use self-attention on the top of text features. The framework of this model is described in Section 3.3. The attended textual feature is passed through one dense layer, following a softmax layer with the final prediction.

Model 4 (Only Image) The architecture of this model is given in Section 3.4. This model uses a pre-trained VGG19 framework to obtain the region-specific features without attention to classify a meme into a specific category. Extracted visual features are fed to the output layer having softmax activation for the final prediction.

Model 5 (Image+ self-Attention) In this model, attention is used on the top of the visual features as mentioned in Section 3.5. After applying attention to the visual feature, it is passed through a softmax layer with three output neurons.

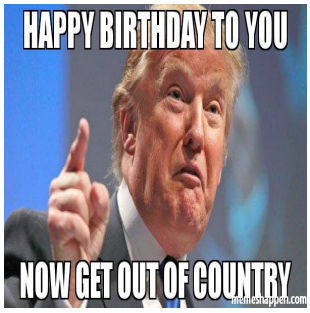
Input image			
	(a)	**(b)**	**(c)**
True label	negative	neutral	negative
Model2	positive	neutral	positive
Model3	neutral	neutral	neutral
Model4	negative	positive	positive
Model5	negative	negative	positive
Model6	neutral	positive	neutral
Model7	positive	positive	neutral
Model8	negative	neutral	negative

Figure 4: Outputs from different models

Model 6 (Text + Image with self-attention) In this model, we concatenate both textual T_i and visual feature V_i (c.f. Section 3.6). After obtaining a concatenated feature vector, we pass it through one dense layer. The dense layer follows the output layer with one softmax activation for the final prediction.

Model 7 (Proposed Model-1 (Text + Image with co-attention and bilinear pooling)) This is our first proposed model, described in Section 3.7.1.

Model 8 (Proposed Model-2 (Text + Image with co-attention and MFB)) This is our second proposed model which is described in Section 3.7.2.

Table 2: Data statistics

Data	Class	Statistics	Distribution
Train	Positive	4160	59.5%
	Neutral	2201	31.5%
	Negative	631	9.0%
	Total	6992	100%
Test	Positive	831	59.56%
	Neutral	439	31.44%
	Negative	126	9.02%
	Total	1396	100%

4 Datasets and Experiments

4.1 Datasets

To assess the significance of our proposed framework, we use a multimodal dataset given in SemEval2020 Task8 Sabtask A (Sharma et al., 2020). The dataset consists of 6,992 memes for training, where 10% is selected for validation, and testing is done on 1396 memes. Table 2 presents the summary of the dataset used.

4.2 Experimental Setup

For the experimental setup, we use keras with tensorflow at the backend. From the dataset distribution, it is visible that the dataset is skewed towards a positive class. To tackle this problem, we use the class weights for each class during implementation. We evaluate our system performance on the batch size of (16,32,64) and dropout rate as (0.2,0.3,0.4). We obtained the best performance using the batch size of 64 and the dropout rate of 0.4. During the training time for every model, we use the Adam optimizer with lr=3e-5, beta1=0.9, and beta2=0.999 for the loss optimization.

4.3 Result and Analysis

In this section, we discuss the performance of each model described in the above section. We report the results in the form of accuracy and F1-score. In Table 1, results of all the models are shown. The

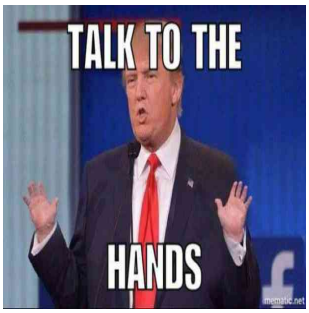
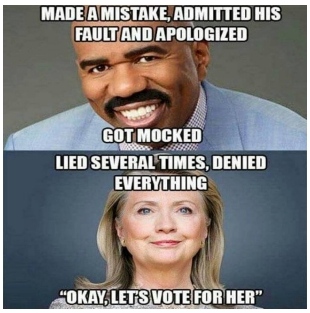
Input image			
True label	negative	negative	negative
Model 8	neutral	positive	neutral

Figure 5: Examples of miss-classification by the proposed framework

baseline model for text data obtained 49% accuracy with 32.78% F1-score. In contrast, when we applied attention to the top of the textual feature reported in Model 3, it gave us a decent baseline with 49% and 33.97% accuracy and F1-score, respectively. Similarly, Model 4 is a decent baseline model for only an image as an input with reported 29.36% accuracy and 34% F1-score. Furthermore, the visual attention model, i.e., Model 5, also demonstrates a comparatively good performance with a reported accuracy of 31.66%.

Table 3: Confusion matrix of the proposed model

	Negative	**Neutral**	**Positive**
Negative	26	18	82
Neutral	70	114	256
Positive	126	190	514

Model 6 is the framework where we merely concatenate the attended textual and visual feature vectors. We can observe in Table 1 that simple concatenation does not help the classifier to classify memes into its' right category effectively. The reported accuracy and F1-score for this model are 51% and 32.22%, respectively, which are -1.75% and $+0.56\%$ points increments (in terms of F1-score) when compared to Model 3 and Model 5, respectively.

To obtain a more robust multimodal classifier, we use our proposed deep learning framework mentioned in Section 3.7. Our model reported in Section3.7.1 i.e Model 7 performs better than all previously reported models. It shows $+3.1\%$ improvement in F1- score in comparison to Model 6. Furthermore, the significant growth in the accuracy as well as in the F1-score clearly shows the effectiveness of our proposed Model 8 mentioned in Section 3.7.2. We found the performance of Model 8 to have increased significantly in terms of F1-score by $+4.59\%$ and $+1.49\%$ when compared with Model 6 and Model 7, respectively. We find this improvement statistically significant as we performed the significance t-test conducted at a 5% significance level.

4.4 Detailed Analysis

We perform detailed quantitative and qualitative analysis of the output generated from our models to understand where our model succeeds and where our model fails. Results of all models are shown in Table 1. We take some example cases in Figure 4 where we evaluate the performance of all the models.

- For example (a), we can see that the visual model and our proposed model (text + visual) performs better than the other models.
- For example (b), it is shown that the textual model and our proposed model (text + visual) 3.7.2 performs well.
- In a similar way, for a given example (c), only the proposed model 3.7.2 shows the accurate output.

In Table 3, we report the confusion matrix of our proposed model. From the confusion matrix, we can identify the effectiveness of our proposed model. We can observe that using co-attention and an effective MFB based fused feature representation, the system can correctly capture the complex association between visual and textual representation.

We also perform qualitative analysis on the dataset to analyze the output from our proposed

model. We observe that due to the implicit nature of negative polarity memes, in few cases, our proposed multimodal system couldn't relate the textual and visual features properly, which results in miss-classification. We encountered a few complex examples where both textual and visual parts were neutral separately, but it became negative when we combined both the modalities. In such cases, our model was not able to produce a good result (c.f. Figure 5).

5 Conclusion

In this work, we have proposed an end-to-end CNN-based deep neural network that consists of co-attention that jointly reasons about textual and visual representation. Additionally, we incorporated one common portrayal of a meme by utilizing the multimodal factorized bilinear pooling of textual and visual features. By introducing these joint representations, we obtain more effective multimodal features to identify the sentiment of a given meme. From the quantitative and qualitative error analysis on the recently released *SemEval-2020 Task-8* (Sharma et al., 2020) dataset, we observed that our proposed method produces promising results with respect to the baseline models. In the future, we will investigate more fusion strategies to combine both the modalities effectively; and investigate methods to extract essential objects from the meme.

References

Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca J. Passonneau. 2011. Sentiment analysis of twitter data.

Farman Ali, Daehan Kwak, Pervez Khan, Shaker El-Sappagh, Amjad Ali, Sana Ullah, Kye Hyun Kim, and Kyung-Sup Kwak. 2019. Transportation sentiment analysis using word embedding and ontology-based topic modeling. *Knowledge-Based Systems*, 174:27–42.

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. *Proceedings of the 26th International Conference on World Wide Web Companion*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang. 2013. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *Proceedings of the 21st ACM International Conference on Multimedia*, MM '13, page 223–232, New York, NY, USA. Association for Computing Machinery.

Alexandra Cernian, Valentin Sgarciu, and Bogdan Martin. 2015. Sentiment analysis from product reviews using sentiwordnet as lexical resource. In *2015 7th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*, pages WE–15–WE–18.

Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *ICWSM*.

R. Dawkins. 2016. *The Selfish Gene*. Oxford Landmark Science. Oxford University Press.

Andrea Esuli and Fabrizio Sebastiani. 2006. SENTIWORDNET: A publicly available lexical resource for opinion mining. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Xing Fang and Justin Zhan. 2015. Sentiment analysis using product review data. *J Big Data*, 2.

Paula Fortuna, João Rocha da Silva, Juan Soler-Company, Leo Wanner, and Sérgio Nunes. 2019. A hierarchically-labeled portuguese hate speech dataset.

Vasileios Hatzivassiloglou and Kathleen McKeown. 1997. Predicting the semantic orientation of adjectives. In *ACL*.

Anthony Hu and Seth Flaxman. 2018. Multimodal sentiment analysis to explore the structure of emotions. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*.

Sardar Haider Waseem Ilyas, Zainab Tariq Soomro, Ahmed Anwar, Hamza Shahzad, and Ussama Yaqub. 2020. Analyzing brexit's impact using sentiment analysis and topic modeling on twitter discussion. In *The 21st Annual International Conference on Digital Government Research*, dg.o '20, page 1–6, New York, NY, USA. Association for Computing Machinery.

Walter Kasper and Mihaela Vela. 2011. Sentiment analysis for hotel reviews.

Vishal Keswani, Sakshi Singh, Suryansh Agarwal, and Ashutosh Modi. 2020. Iitk at semeval-2020 task 8: Unimodal and bimodal sentiment analysis of internet memes.

Efthymios Kouloumpis, Theresa Wilson, and Johanna D. Moore. 2011. Twitter sentiment analysis: The good the bad and the omg! In *ICWSM*.

Akshi Kumar, Kathiravan Srinivasan, Wen-Huang Cheng, and Albert Zomaya. 2020. Hybrid context enriched deep learning model for fine-grained sentiment analysis in textual and visual semiotic modality social data. *Information Processing Management*, 57.

Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2018. Benchmarking aggression identification in social media. In *TRAC@COLING 2018*.

Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Recurrent convolutional neural networks for text classification. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, page 2267–2273. AAAI Press.

Yann Lecun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature Cell Biology*, 521(7553):436–444. Funding Information: Acknowledgements The authors would like to thank the Natural Sciences and Engineering Research Council of Canada, the Canadian Institute For Advanced Research (CIFAR), the National Science Foundation and Office of Naval Research for support. Y.L. and Y.B. are CIFAR fellows. Publisher Copyright: © 2015 Macmillan Publishers Limited. All rights reserved.

Chenghua Lin and Yulan He. 2009. Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, page 375–384, New York, NY, USA. Association for Computing Machinery.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. In *NIPS*.

Navonil Majumder, Devamanyu Hazarika, Alexander F. Gelbukh, Erik Cambria, and Soujanya Poria. 2018. Multimodal sentiment analysis using hierarchical fusion with context modeling. *CoRR*, abs/1806.06228.

K. Mouthami, K. Nirmala Devi, and V. Murali Bhaskaran. 2013. Sentiment analysis and classification based on textual reviews. In *2013 International Conference on Information Communication and Embedded Systems (ICICES)*, pages 271–276.

Benet Oriol Sabat, Cristian Canton-Ferrer, and Xavier Giró i Nieto. 2019. Hate speech in pixels: Detection of offensive memes towards automatic moderation. *ArXiv*, abs/1910.02334.

Chhavi Sharma, Deepesh Bhageria, William Scott, Srinivas PYKL, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Björn Gambäck. 2020. Semeval-2020 task 8: Memotion analysis - the visuo-lingual metaphor! *CoRR*, abs/2008.03781.

Han-Xiao Shi and Xiao-Jun Li. 2011. A sentiment analysis model for hotel reviews based on supervised learning. In *2011 International Conference on Machine Learning and Cybernetics*, volume 3, pages 950–954.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition.

Ha-Nguyen Tran and Erik Cambria. 2018. Ensemble application of elm and gpu for real-time multimodal sentiment analysis. *Memetic Computing*, 10.

Quoc-Tuan Truong and Hady W. Lauw. 2019. Vistanet: Visual aspect attention network for multimodal sentiment analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):305–312.

Wei Wei and Jon Gulla. 2010. Sentiment learning on product reviews via sentiment ontology tree. pages 404–413.

Caiming Xiong, Victor Zhong, and Richard Socher. 2018. Dynamic coattention networks for question answering.

Jun-Ming Xu, Kwang-Sung Jun, Xiaojin Zhu, and Amy Bellmore. 2012. Learning from bullying traces in social media. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 656–666, Montréal, Canada. Association for Computational Linguistics.

Fan Yang, Xiaochang Peng, Gargi Ghosh, Reshef Shilon, Hao Ma, Eider Moore, and Goran Predovic. 2019. Exploring deep multimodal fusion of text and photo for hate speech classification. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 11–18, Florence, Italy. Association for Computational Linguistics.

Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. 2019. Deep modular co-attention networks for visual question answering. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6274–6283.