Exercise? I thought you said 'Extra Fries' ☺: Leveraging Sentence Demarcations and Multi-hop Attention for Meme Affect Analysis

Shraman Pramanick, Md Shad Akhtar, Tanmoy Chakraborty
Dept. of Computer Science & Engineering, IIIT-Delhi, India
{shramanp, shad.akhtar, tanmoy}@iiitd.ac.in

Abstract

Today's Internet is awash in memes as they are humorous, satirical, or ironic which make people laugh. According to a survey, 33% of social media users in age bracket [13 – 35] send memes every day, whereas more than 50% send every week. Some of these memes spread rapidly within a very short time-frame, and their virality depends on the novelty of their (textual and visual) content. A few of them convey positive messages, such as funny or motivational quotes; while others are meant to mock/hurt someone's feelings through sarcastic or offensive messages. Despite the appealing nature of memes and their rapid emergence on social media, effective analysis of memes has not been adequately attempted to the extent it deserves. Recently, in SemEval'20, a pioneering attempt has been made in this direction by organizing a shared task on 'Memotion Analysis' (meme emotion analysis). As expected, the competition attracted more than 500 participants with the final submission of [23 – 32] systems across three sub-tasks.

In this paper, we attempt to solve the same set of tasks suggested in the SemEval'20-Memotion Analysis competition. We propose a multi-hop attention-based deep neural network framework, called `MHA-Meme`, whose prime objective is to leverage the spatial-domain correspondence between the visual modality (an image) and various textual segments to extract fine-grained feature representations for classification. We evaluate `MHA-Meme` on the 'Memotion Analysis' dataset for all three sub-tasks - *sentiment classification*, *affect classification*, and *affect class quantification*. Our comparative study shows state-of-the-art performances of `MHA-Meme` for all three tasks compared to the top systems that participated in the competition. Unlike all the baselines which perform inconsistently across all three tasks, `MHA-Meme` outperforms baselines in all the tasks on average. Moreover, we validate the generalization of `MHA-Meme` on another set of manually annotated test samples and observe it to be consistent. Finally, we establish the interpretability of `MHA-Meme`.

Introduction

In recent years, Internet memes (or simply memes) have emerged as one of the most frequently circulated entities on social media platforms[1]. In general, memes describe a basic unit of cultural idea or symbol that can be transmitted from one mind to another and inherently, portray the opinion, resentment, fandom, along with the political, psychological, socio-cultural expression of a community. The behavior of memes is also distinctive as memes replicate and mutate, similar to genes in human evolution, during propagation on social media. Despite being so popular and entrenched on online media, it is extremely challenging to leverage automated methods to understand the inherent sentiment/emotion (affect) expressed by memes.

Motivation: Memes contain information of both textual and visual modalities; both the modalities often overlap, but are complementary sometimes. Therefore, to analyze the emotion expressed by a meme, both the modalities should be considered simultaneously; and at the same time, monotonous information must be eliminated. Two different memes may sometimes have the same image but can express completely different semantics based on just a few words in the text. For example, Figures 1a and 1b display exact same image, but vary in four different semantic classes. However, Figures 1c and 1d have rich visual information. Again, Figure 1e has high textual and very little visual content. For such variation of modality information, memes are often hard to classify.

State-of-the-art: Annotating memes into different sentiment and affect classes is another challenge, as emotion about memes highly depends upon an individual's perception of an array of aspects within society, and could thus vary from one person to another. This phenomenon is known as "*Subjective Perception Problem*" (Zhao et al. 2018), which leads to discrepancy in annotated data. Very recently, emotion analysis of memes has been portrayed as a separate task in SemEval-2020 (Sharma et al. 2020). A large dataset has been released as a part of the shared task. The winner of the shared task achieved macro-F1 scores of 0.35, 0.51, and 0.32, respectively for three tasks – sentiment analysis, affect classification and affect quantification (described later). The performance is not significantly high, which demands further work on these tasks.

Challenges: A meme M is an image consisting of two modalities – a background image I and some text T at the foreground, referring to a specific situation. The purpose of creating a meme is either to convey humorous or motivational messages, or to mock someone/something in a rhetorical

[1] https://tinyurl.com/526n8zpx

(a) [-1,1,1,1,0] (b) [0,0,0,0,0] (c) [1,1,1,1,0] (d) [0,1,1,1,1] (e) [1,1,1,1,0]

Figure 1: A few examples of memes from Memotion Analysis dataset (Sharma et al. 2020). Classification of the memes are in the following format: [Sentiment, Humor, Sarcasm, Offensive, Motivation]. For Sentiment, {-1, 0, 1} corresponds to negative, neutral, positive classes. For other cases, {0, 1} denotes absence and presence of corresponding affect.

way. The background image itself can have multiple sections usually representing a progressive story line; however, many memes convey their messages through a single image as well. Similarly, the foreground text can also be clustered into multiple sentences/phrases supporting the development or the depiction of the story line. A few example memes are shown in Figure 1.

Interpreting a meme is a challenging task often because of the implicit world knowledge and overlapping information. Two memes can have same image but can convey different semantics. Similarly, same text corresponds to two different events according to two background images. Moreover, the world knowledge plays a crucial part in establishing a relationship between a meme and its reference situation.

Proposed Method: The current study aims at analyzing the emotion of memes on three dimensions:

- **Sentiment classification –** whether a meme conveys *positive*, *negative*, or *neutral* sentiment;
- **Affect classification –** whether a meme conveys *humorous*, *sarcastic*, *offensive*, *motivational*, or a combination of the four affects;
- **Affect quantification –** what is the quantification of the expressed affect.

The foreground texts are critical in extracting the semantic-level information from meme. However, they require special attention depending upon their position in the meme and their reference to a specific region of the background image. Therefore, in the current work, we propose an attentive deep neural network architecture, called MHA−Meme (**M**ulti-**H**op **A**ttention for **Meme** Analysis), to carefully analyze the correspondence between the background image and each text segment at different spatial locations. To do so, at first, we perform OCR (optical character recognition) to extract texts from the meme, and segment them into l sequence of text depending upon their spatial positions. Next, we process each textual segment t_i separately by establishing their correspondence with the background image I. We employ two attention frameworks for exploiting the correspondence – a unimodal multi-hop attention framework (*aka* MHA attention), and an attention-based multi-modal fusion (*aka* ATMF). Subsequently, we combine all l sequence of the processed texts in a multi-hop attention framework for the classifications. The segment-level analysis enables us to leverage various fine-grained features for achieving the final goal.

Evaluation: We evaluate MHA−Meme on the recently published dataset by the SemEval-2020 shared task on 'Memotion Analysis' (Sharma et al. 2020). We perform extensive experiments for all three tasks and compare our results with various (in total 15) existing systems on the leader-board of the respective tasks. We observe improved performance against the state-of-the-art systems for all three tasks, in a range of $[+0.5\%, +2.2\%]$. Furthermore, we collected and annotated an additional set of 334 memes to validate the generalizability of MHA−Meme.

Analysis: We execute extensive ablation experiments with multiple existing baselines to demonstrate the importance of extracting fine-grained features through the segmented text and sophisticated attention mechanism for meme emotion classification. We also employ LIME (Locally Interpretable Model-Agnostic Explanations) (Ribeiro, Singh, and Guestrin 2016) to visualize the importance of relevant features considering it's prediction.

Our Contributions: The major contributions of the paper are as follows:

1. We leverage the correspondence between a meme and its constituent texts depending upon the spatial locations.
2. We propose an attentive framework that effectively selects and utilizes complementary features from textual and visual modalities to capture multiple aspects of emotions expressed by a meme.
3. We report benchmark results for all three tasks – sentiment classification, affect classification, and quantification.
4. In order to validate the generalizability of MHA−Meme, we collected and annotated an additional set of 334 memes and checked its performance. Furthermore, we establish the interpretability of MHA−Meme using LIME framework.

Reproducibility: To reproduce our results, we present detailed hyper-parameter configurations in Table 2 and the experiment section. Moreover, the full dataset and code for MHA−Meme is publicly available at https://github.com/LCS2-IIITD/MHA-MEME.

Related Work

Multimodal sentiment analysis (MSA) has gained increasing attention in recent years with the surge of multimedia data on the Internet and social media. Unlike text sentiment analysis (Akhtar et al. 2017) which is vastly studied, multimodal approaches make use of visual and acoustic modalities in addition to the textual modality, as a valuable source of information for accurately inferring emotion states. One of the key challenges in multi-modal framework is to utilize and fuse the relevant and complementary information for the prediction. Poria et al. (2017a) presented an overview of different categories of fusion techniques and potential performance improvements with multiple modalities.

Multimodal Fusion: There are mainly three types of fusion strategies for MSA – early, late and hybrid. Early-fusion directly integrates multiple sources of data into a single feature vector and uses a single classifier (Poria et al. 2016; Zadeh et al. 2017) for prediction from combined feature representation. Such early-fusion techniques can not exploit complementary nature of multiple modalities and often produces large feature vectors with redundancies. On the other hand, late-fusion refers to the aggregation of decisions from multiple sentiment classifiers, each trained on separate modalities (Cambria 2016; Cao et al. 2016). These schemes are based on the assumption that separate modalities are independent in the feature space, which is not always true in practice, and thus leading to limited performance when multiple modalities tend to be inter-connected. In contrast, hybrid-fusion employs an intermediate shared representation layer by merging multiple modality-specific paths and has been most successful in the literature. You et al. (2016) proposed a cross-modality consistent regression (CCR) scheme for joint textual-visual sentiment analysis. In the similar line, Aldeneh et al. (2017) compared pooling methods in the context of valence prediction. In another related work, Memory Fusion Network (MFN) (Zadeh et al. 2018) was proposed to associate a cross-view relevance score to each LSTM for MSA. Recently, Akhtar et al. (2019) introduced context-level inter-modal attention framework for simultaneously predicting sentiment and emotions of an utterance and suggested that multitask learning framework offers improvement over the single-task framework.

Meme Emotion Analysis: In the last few years, the growing ubiquity of memes on social media has been increasingly crucial to understand the opinion of a community. However, interpreting memes is challenging due to its inherent complex cognitive aspects of emotions. Although there has been a lot of work to understand the sentiment of other social media contents (Cambria 2016; Yue et al. 2019), such as textual or visual opinions, ratings, movie/product reviews, or recommendations, meme sentiment analysis has not been explored much. The pioneering attempt has recently been made by SemEval-2020 Task 8: Memotion Analysis-The Visuo-Lingual Metaphor! (Sharma et al. 2020). The top participants in this challenge have used a wide variety of methods for unimodal feature extraction, such as FFNN, Naive Bayes, ELMo (Peters et al. 2018), MMBT (Rahman et al. 2019), BERT (Devlin et al. 2019) for textual modality and Inception-ResNet (Szegedy et al. 2017), Polynet (Zhang et al. 2017), DenseNet (Huang et al. 2017) and PNASNet (Liu et al. 2018) for visual modality. Some of the participants have also employed a modality-specific deep neural ensembles to incorporate visual and textual information. However, none of these methods can guarantee non-redundant features from two highly correlated modalities to capture every aspect of emotions expressed by a meme.

To this end, we are the first to address this ill-posed problem by *introducing a multi-hop attention and image encoding filter based sentiment classifier network* which extracts complementary features from both modalities with overlapping information and achieves state-of-the-art performances on Memotion 1.0 dataset. We demonstrate the significance of spatial information for effective fusion of both modalities. Extensive experiments with our proposed `MHA-Meme` against various existing baselines reveal the importance of text segmentation and fine-grained feature extraction to capture multiple aspects of emotion for meme affect analysis.

Proposed Methodology

In this section, we describe our proposed system, `MHA-Meme`, for meme analysis. We take a meme as an input and extract the segmented text using Google OCR Vision API[2]. For each segmented text (t_i) and image[3] (I) pair, we encode them through Bidirectional LSTM (Hochreiter and Schmidhuber 1997) and VGG-19 (Simonyan and Zisserman 2014) networks, respectively. Further, we fine-tune the image encoding through a text-aware filter. The objective is to restrict the system to extract features from the spatial region of the text in the meme. Subsequent to the unimodal feature extraction step, we employ a novel unimodal multi-hop attention (MHA) framework to attend relevant features from the respective encoding.

For the affect classification and quantification, we learn all four affect dimensions together in a joint framework, and the intuition of multi-hop attention is to extract different set of features for different affect classes. We observe positive effect of the multi-hop attention for the sentiment classification as well, which further suggests that it is capable of extracting diverse features for the downstream tasks. In the next step, we employ an attention-based multi-modal fusion (ATMF) mechanism to leverage the correspondence between the textual and visual modalities.

We repeat the above procedure for each textual segment, $t_i, \forall i = [1...l]$, and compute its respective multimodal feature representations. Finally, we combine these representations through another multi-hop attention module to attend to relevant segments for the given meme. The context-aware attended representation is then fed to a sequence of fully-connected layers for the prediction. For the affect classification and quantification, our network includes separate fully-connected layer followed by softmax for each affect class. In subsequent subsections, we discuss these modules in details. Figure 2 presents the architecture of `MHA-Meme`, and the procedure is summarized in Algorithm 1.

[2]https://cloud.google.com/vision

[3]Image remains same for each segment.

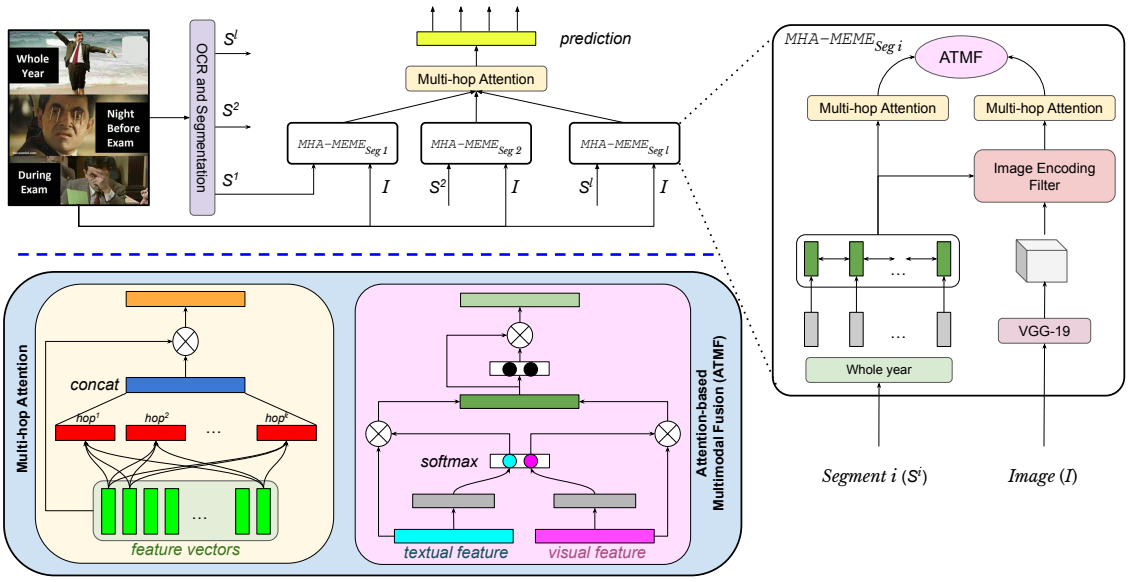


Figure 2: The architecture of the proposed model, MHA−Meme.

Unimodal Feature Extraction

This module extracts unimodal features for the textual and visual modalities of an Internet meme.

Textual Feature Extraction: We employ 200-dimensional GloVe word embedding (Pennington, Socher, and Manning 2014) for encoding each word of the textual segment. We attain 77.22% word coverage for the underlying dataset, whereas the remaining out-of-vocabulary words (OOVs) are initialized randomly. We train our network in dynamic mode to allow fine-tuning of the embedding layer via backpropagation.

We use a 2-layered BiLSTM for the textual feature extraction. It takes 200 dimensional word vectors as input and maps them to $2u$ dimensional hidden states, where u is the hidden dimension for each unidirectional LSTM. Mathematically, for the textual segment $t_i = (w^1, w^2, ..., w^n)$, we compute

$$\overrightarrow{h_t} = \overrightarrow{LSTM}(w_t, \overrightarrow{h_{t-1}}) \quad (1)$$

$$\overleftarrow{h_t} = \overleftarrow{LSTM}(w_t, \overleftarrow{h_{t+1}}) \quad (2)$$

$$h_t = [\overrightarrow{h_t}, \overleftarrow{h_t}] \quad (3)$$

where, $\overrightarrow{h_t} \in \mathbb{R}^u$, $\overleftarrow{h_t} \in \mathbb{R}^u$, and $h_t \in \mathbb{R}^{2u}$ are the forward, backward, and the combined hidden representations, respectively. Moreover, $H = (h_1, h_2, ..., h_n) \in \mathbb{R}^{n \times 2u}$. We use $u = 256$ in our experiments to match with visual feature dimension.

Image Encoding Filter: We employ pretrained VGG-19 (Simonyan and Zisserman 2014) to extract local feature maps for the images. Specifically, we extract $7 \times 7 \times 512$ feature maps from the last pooling layer (pool5) of VGG-19 and reshape it into a matrix $F = (f_1, .., f_m), m = 49$, where each $f_k \in \mathbb{R}^{512}$ corresponds to a local spatial region.

In general, visual modality contains complementary information from the textual modality. However, while dealing with meme images, we need to ensure that the OCR-extracted text and the text in the image do not establish a direct correspondence. Therefore, to restrict our model to focus aggressively on the textual parts of the image and extract redundant visual features, we aim to filter the redundant visual features in correspondence with the textual utterance (Lee et al. 2018; Liu, Zhang, and Gulla 2019).

Given $H \in \mathbb{R}^{n \times 512}$ and $F \in \mathbb{R}^{m \times 512}$, the image encoding filter mechanism begins with defining an affinity matrix $C \in \mathbb{R}^{n \times m}$, whose element c_{ij} denotes the similarity between the feature vector pair, $h_i \in \mathbb{R}^{512}$ and $f_i \in \mathbb{R}^{512}$: $C = tanh(H\mathbf{W}^b F^\top)$ where, $\mathbf{W}^b \in \mathbb{R}^{512 \times 512}$ is a correlation matrix to be learned during training.

Subsequently, we compute a normalized weight α_{ij}^h to denote the relevance of the i^{th} word to the spatial region j. Hence, the weighted summation of all word representation can be represented as,

$$a_j^h = \sum_{i=1}^{n} \alpha_{ij}^h h_i, \text{ where } \alpha_{ij}^h = exp(c_{ij}) / \sum_{i=1}^{n} exp(c_{ij}) \quad (4)$$

Since our goal is to emphasize the dissimilar features between the textual and visual modalities and to determine the importance of each region given the textual utterance, we define the relevance matrix R as the cosine distance between the attended sentence vector a_j^h and image region feature f_i –

$$R(f_i, a_j^h) = 1 - \frac{f_i^\top \cdot a_j^h}{\|f_i\| \|a_j^h\|} \quad (5)$$

Finally, the weighted summation of all regions gives the modified image representation U computed as,

$$U = \sum_{j=1}^{m} R(f_i, a_j^h) \cdot f_i \quad (6)$$

where, $R(f_i, a_j^h)$ acts as a filter for the image encoding f_i.

Multi-hop Attention for Unimodal Feature Representation

Every word in a sentence has a specific role in the semantic space; however, some words can have a higher relevance to a particular task. Hence, semantic attention mechanism has been proven to be extremely beneficial for many NLP tasks (Bahdanau, Cho, and Bengio 2015; Wang et al. 2016). Similarly, some visual regions are often more informative to represent certain emotions, and the literature (Lu et al. 2017; You, Jin, and Luo 2017) supports visual attention mechanism to be an effective solution.

Traditional self-attention mechanism typically focuses on a specific component of a sentence or image, reflecting only one aspect of the semantics. However, a meme can express multiple emotions together, e.g., a meme can be humorous as well as sarcastic. Therefore, we hypothesize that to cater to different tasks in a joint-framework, multiple hops of attention mechanism would attend to diverse and task-specific relevant features, i.e., one hop would attend to features relevant for the humor classification, while the other would focus on the sarcasm detection.

Following our intuition, we compute multi-hop attention as follows: Given a unimodal feature matrix (e.g., the textual features $H \in \mathbb{R}^{n \times 512}$ or the visual features $U \in \mathbb{R}^{m \times 512}$), we aim to extract k distinct set of relevant features corresponding to the input representation, where k is a hyper-parameter. The multi-hop attention mechanism takes H as an input, and outputs an attention weight matrix $A \in \mathbb{R}^{k \times n}$ respective to the textual representation. Similarly, we compute attention weight matrix $B \in \mathbb{R}^{k \times m}$ for the visual representation U.

$$A = softmax(\mathbf{W_{h2}}\ tanh(\mathbf{W_{h1}} H^\top)) \quad (7)$$

$$B = softmax(\mathbf{W_{u2}}\ tanh(\mathbf{W_{u1}} U^\top)) \quad (8)$$

where, $\mathbf{W_{h1}}, \mathbf{W_{u1}} \in \mathbb{R}^{d \times 512}$ and $\mathbf{W_{h2}}, \mathbf{W_{u2}} \in \mathbb{R}^{k \times d}$ are parameter matrices to be learned during training. The $softmax(\cdot)$ is performed along the second dimension of its input, and d is a hyper-parameter we can set arbitrarily.

The resulting embedding matrices M and N respectively for textual and visual modalities are computed using their respective attention weight matrices and unattended features.

$$M = A \otimes H \quad (9)$$

$$N = B \otimes U \quad (10)$$

where, $M, N \in \mathbb{R}^{k \times 512}$ are self-attended features for the textual and visual modalities.

Attention-based Multi-modal Fusion (ATMF)

ATMF utilizes an attention based mechanism to fuse the unimodal features from textual and visual modalities. Since same modality may have different contribution for different utterances and different meme samples, the attention-based multi-modal fusion network shows much better classification performance over concatenation and neural fusion techniques. Extending the attentive fusion network of (Gu et al. 2018), our ATMF module consists of two major parts – modality attention generation and weighted feature concatenation. In the first part, we use a sequence of dense layers followed by a softmax layer to generate the weighted scores for two given modalities. Each successive dense layer has smaller dimension than its previous one, thus forming a tower like architecture.

$$[s_t, s_v] = softmax(Dense(M, N)) \quad (11)$$

where, $[s_t, s_v]$ are the attention scores for the textual and visual modalities, respectively.

In the second part, the original unimodal features are weighted using their respective attention scores and concatenated together; and $(1 + s_t)M$ and $(1 + s_v)N$ denote residual+attended vectors for the textual and visual modalities, respectively.

$$P_F = tanh(\mathbf{W_F} \cdot [(1 + s_t)M, (1 + s_v)N] \quad (12)$$

where, $W_F \in \mathbb{R}^{512 \times 512}$ is a learnable parameter.

Moreover, we incorporate an additional attention layer specifically to reduce the effect of repetitive features captured in MHA for a short text segment (e.g., a single-word text segment). In the absence of sufficient words in a text segment, the multi-hop attention module would compute repetitive

Algorithm 1 **M**ulti-**H**op **A**ttention for **Meme**

Input: An Internet meme M
Output: Predicted class $\hat{y}_c\ \forall\ c \in \{\text{affect classes}\}$

procedure MHA-MEME(M)
  $\{t_1, t_2, ..., t_n\}, I \leftarrow$ `OCR` and `Segmentation` (M)
  for seg in $(1, n)$ **do** ▷ #segments
    $u \leftarrow 256$ ▷ LSTM dimension
    $H_{seg} \leftarrow BiLSTM(t_{seg}, u)$
    $F_{seg} \leftarrow VGG19_{pool5}(I)$
    $F^{filt}_{seg} \leftarrow Image\ Encoding\ Filter(H_{seg}, F_{seg})$
    $M_{seg} \leftarrow Multi\text{-}hop\ Attention(H_{seg})$
    $N_{seg} \leftarrow Multi\text{-}hop\ Attention(F^{filt}_{seg})$
    $X_{seg} \leftarrow ATMF(M_{seg}, N_{seg})$ ▷ feature fusion
  $Z \leftarrow Multi\text{-}hopAttention(Stack(X_1, ..., X_n))$
  $\hat{y}_c \leftarrow Classifier(Z)\ \forall\ c \in \{\text{affect classes}\}$
  return $\hat{y}_c$

procedure IMAGE ENCODING FILTER(H, F)
  $C \leftarrow H\mathbf{W^b}F^\top$ ▷ affinity matrix
  $a^h_j \leftarrow \sum_{i=1}^{n}((exp(c_{ij}) / \sum_{i=1}^{n} exp(c_{ij}))h_i)$
  $R(f_i, a^h_j) \leftarrow 1 - \dfrac{f_i^\top \cdot a^h_j}{\|f_i\| \|a^h_j\|}$
  $U \leftarrow \sum_{j=1}^{m} R(f_i, a^h_j) \cdot f_i$
  return U

procedure MULTI-HOP ATTENTION(H)
  $A \leftarrow softmax(\mathbf{W_{h2}}\ tanh(\mathbf{W_{h1}} H^\top))$
  $M \leftarrow A \otimes H$ ▷ attended features
  return M

procedure ATMF(M, N)
  $[s_t, s_v] \leftarrow \sigma(Dense(M, N))$ ▷ $\sigma - softmax$
  $\gamma_f \leftarrow \sigma(\mathbf{w_f}^\top \cdot tanh(\mathbf{W_F} \cdot [(1 + s_t)M, (1 + s_v)N]))$
  $X \leftarrow [(1 + s_t)M, (1 + s_v)N] \cdot \gamma_f^\top$
  return X

	#Memes	#Textual segments	Sentiment			Affects*			
			Pos	Neg	Neu	Humor	Sarcasm	Offense	Motivation
Train	6601	14032	3923	2092	586	5082	5162	4163	2406
Test_A	1879	4184	1111	594	173	1354	1376	1101	648
Test_B	334	748	184	88	62	265	212	244	142

Table 1: Dataset statistics of Memotion Analysis (Sharma et al. 2020). Test_A is the original test set (Sharma et al. 2020). Test_B is developed by us to validate the generalization of MHA−Meme. Affects*: A meme can belong to ≥ 1 affective classes.

features for k hops in both textual and visual modalities; thus we hypothesize that the inclusion of an additional attention layer would address this problem. Therefore, we compute the final segment-level multi-modal representation $x \in \mathbb{R}^{512}$ as follows:

$$\gamma_f = softmax(\mathbf{w_f}^\top \cdot P_F) \quad (13)$$

$$x = [(1+s_t)M, (1+s_v)N] \cdot \gamma_f^\top \quad (14)$$

where, $w_f \in \mathbb{R}^{512}$ is a learnable parameter.

Context-aware Classification

It is the final stage of our network where we leverage the contextual multimodal representation of each textual-segment for the classification. However, the sentiment or the affective dimension sometime relates to a few specific textual segments. Thus, we employ the multi-hop attention module at the meme-level. The objective is to highlight diverse features respective to the underlying tasks.

Let the matrix $X = [x_1, x_2, ..., x_l] \in \mathbb{R}^{512 \times l}$ contains the multimodal features for each textual segment, where l is the number of segments in a meme and $x_i \in \mathbb{R}^{512}$ is the multimodal feature representation for i^{th} segment. Finally, the multi-hop segment-level representation X^* is flattened and forwarded to the softmax layers the classification (i.e., one softmax layer for the sentiment classification and four softmax layers for affect classification and quantification each.

$$Z = softmax(X^* \cdot \mathbf{W_{soft}} + \mathbf{b_{soft}}) \quad (15)$$

$$\hat{y} = \arg\max_j (Z[j]) \;\; \forall j \in class \quad (16)$$

where, $\mathbf{W_{soft}}$ and $\mathbf{b_{soft}}$ are learnable weights.

Experiments

In this section, we present details of the dataset, data preprocessing steps and training methodologies.

Dataset: We conduct our evaluations on Memotion[4] 1.0 dataset (Sharma et al. 2020) (dubbed as **Memotion** dataset) which was recently released as part of the SemEval-2020 shared task on 'Memotion Analysis.' This dataset consists of $8,480$ manually annotated English memes from 52 unique and globally popular categories, e.g., *Hillary Clinton*, *Donald Trump*, *Minions*, *Baby godfather*, etc. The dataset was annotated through Amazon Mechanical Turk, where entities were annotated into three sentiment classes (i.e., *positive*, *neutral*, and *negative*) and four different affect classes (i.e., *Humorous*, *Sarcasm*, *Offensive*, and *Motivation*). The dataset also has quantification scores to which a particular affect is expressed. To address '*Subjective Perception Problem*' (Zhao et al. 2018), the annotation process was performed multiple times, and the final annotations were adjudicated based on majority voting. A brief statistics of the dataset is shown in Table 1. Each instance of the dataset consists of an image, representing a meme, and an unsegmented OCR-extracted text (i.e., an OCR text in a meme is represented as a sequence of words without any segment demarcations). The original dataset has $6,992$ and $1,879$ memes in the training and test (Test_A or T_A) sets, respectively.

Since we need segmented text (OCR_{Seg}) for building our model, we employed Google OCR Vision API to extract textual segments. However, during segmentation, we encountered some alignment issues with ~ 500 memes in the training set. We manually corrected most of the segmentation issues.

From the resultant $6,601$ memes in the training set, we extract $14,032$ textual segments; whereas, $4,184$ textual segments were extracted from $1,879$ meme in the test set. The distributions of sentiment and affect classes for both train and test sets are mentioned in Table 1. For the sentiment classification, we classify each meme into '*positive*', '*neutral*', or '*negative*' classes. The affect classification is a multi-label problem, where a meme can belong to more than one class, i.e., any combination of the '*humor*', '*sarcasm*', '*offensive*', and '*motivational*' affects is possible. The affect class quantification is a fine-grained classification task, which determines the extent of the expressed affects. The quantification of '*humor*' is '*not funny*', '*funny*', '*very funny*', and '*hilarious*'; whereas for sarcasm, it is '*not sarcastic*', '*general*', '*twisted meaning*', and '*very twisted*'. Similarly for offensive, the quantification labels are '*not offensive*', '*slightly offensive*', '*very offensive*', and '*hateful offensive* '. In contrast, the '*motivation*' affect has only two extents, i.e., '*not motivational*' and '*motivational*'.

We also collected and annotated an additional set of 334 memes, called Test_B (T_B), to validate the generalization of MHA−Meme. For annotation, we followed the guidelines as laid out by Sharma et al. (2020). Table 1 shows the statistics of Test_B.

Training: We train MHA−Meme using Pytorch framework on a NVIDIA Tesla T4 GPU, with 16 GB dedicated memory, with CUDA-10 and cuDNN-11 installed. We employ pre-trained GloVe (Pennington, Socher, and Manning 2014) Twitter embedding model and fine-tune them during training. The network is randomly initialized with a zero-mean

[4]http://alt.qcri.org/semeval2020/index.php?id=tasks

Hyper-parameter	Notation	Value
hidden units of BiLSTM	u	256
#dim for Dense layers	-	$[256, 64, 8]$
Multi-hop Attention		
#hops (unimodal)	k	30
#hops (multimodal)		10
#hidden-units (unimodal)	d	350
#hidden-units (multimodal)		100
Training		
Batch-size	-	8
Epochs	N	200
Optimizer	-	Adam
Loss	-	NLL
Learning-rate	α	0.005
Learning-rate-decay (/10K$iter$)	-	1e-4
Momentum	-	0.9
Class weights for imbalanced training data		
sentiment $[w_{pos}, w_{neu}, w_{neg}]$		$[1, 1.5, 2]$
affective - ***humor*** $[w_{nonhum}, w_{hum}]$		$[1.5, 1]$
affective - ***sarcasm*** $[w_{nonsar}, w_{sar}]$		$[1.5, 1]$
affective - ***offense*** $[w_{nonoff}, w_{off}]$		$[1.25, 1]$
affective - ***motivation*** $[w_{nonmot}, w_{mot}]$		$[1, 1.25]$

Table 2: Hyper-parameters of MHA−Meme.

Gaussian distribution with standard deviation 0.02.

From Table 1, we can observe label imbalance problem for both sentiment ([*positive* and *negative*] vs. *neutral*) and affect ([*humor*, *sarcasm*, and *offense*] vs. *motivational*) classification tasks. To minimize the effect of label imbalance in loss calculation, we assign larger weights for minority classes. We train our models using Adam (Kingma and Ba 2014) optimizer and negative log-likelihood (NLL) loss as the objective function. In Table 2, we furnish the details of hyper-parameters used for the training.

Baselines: We compare the performance of MHA−Meme with various existing models on the leader-board of three separate tasks. A brief description of each of these baselines is given below.

A. **Guoym:** Guo et al. (2020) trained five base classifiers to utilize five different types of data representation and combined their outputs through data-based and feature-based ensemble methods. Their system performed consistently well across three different sub-tasks and ranked 2^{nd}, 2^{nd}, and 1^{st} in sentiment classification, affect classification, and affect quantification, respectively.

B. **Vkeswani IITK:** Keswani et al. (2020) used a feed-forward neural network with Word2vec embedding for all the three subtasks. They utilized domain knowledge to improve the classification of context-specific memes.

C. **George.Vlad Eduardgzaharia UPB:** Vlad et al. (2020) developed a multimodal multi-task learning architecture that combines pre-trained ALBERT for text encoding with pre-trained VGG16 for image representation. They fused two modalities by simple feature concatenation.

D. **HonoMi Hitachi:** Morishita et al. (2020) fine-tuned four pre-trained visual (Inception-ResNet, PolyNet, SENet, and PNASNet) and textual (i.e., BERT, GPT-2, Transformer-XL, and XLNet) models, and combined them to capture the cross-modal correlations between the textual and visual modalities effectively.

E. **Souvik Mishra Kraken:** Gupta et al. (2020) proposed a hybrid neural Naïve-Bayes, Support Vector Machine and Logistic Regression to solve the multimodal classification problem.

F. **Mayukh Memebusters:** Sharma, Kandasamy, and Vasantha (2020) used transfer learning for image and text feature extraction. They employed attention-based recurrent (LSTM and GRU) architectures for the final prediction.

G. **Gundapu Sunil:** Walińska and Potoniec (2020) extracted textual and visual features using LSTM with GloVe word embeddings and pre-trained VGG16, respectively, and fused them for the classification.

H. **Nowshed CSECU KDE MA:** Chy, Siddiqua, and Aono (2020) introduced a convolution and BiLSTM based attentive framework to jointly learn visual and textual features from an input meme.

I. **Prhlt upv:** De la Peña Sarracén, Rosso, and Giachanou (2020) used pre-trained BERT to extract textual features and pre-trained VGG19 to extract visual features, and then combined two modalities using a simple concatenation-based fusion technique.

J. **Xiaoyu:** Guo, Ma, and Zubiaga (2020) proposed a system which is similar to Prhlt upv, except they found DenseNet outperforming ResNet when used for visual feature extraction.

K. **(1) Aihaihara**, **(2) Saradhix Fermi**, **(3) Hg**, **(4) Sourya Diptadas** and **(5) Jy930 Rippleai:** These teams participated in the SemEval'20-Memotion Analysis competition, but did not submit their system description paper in the workshop proceeding. We used their reported scores as outlined in Sharma et al. (2020).

All the other participants in the SemEval-2020 competition used similar multi-modal deep neural models; however, the difference in hyper-parameters distinguishes among their performance. Moreover, these systems did not regard the textual segmentation in their architectures. To this end, we are the first to analyze the interaction between visual and textual modalities at a fine-grained level, i.e., for each textual segment in a meme, we extract the fine-level feature representation in correspondence with the image. Table 4 shows a performance comparison of MHA−Meme with 15 different baselines taken from SemEval-2021.

Experimental Results

In this section, we report experimental results for our proposed model, MHA−Meme, and its variants for all three tasks. At the end, we also present comparative analysis against various baselines which include some of the best systems (current state-of-the-art) submitted as part of the 'Memotion Analysis' shared task in SemEval-2021 (Sharma et al. 2020).

	Models	Sentiment classification				Affect classication (Avg)				Affect quantification (Avg)			
		Macro F1		Micro F1		Macro F1		Micro F1		Macro F1		Micro F1	
		Test_A	Test_B	Test_A	Test_B	Test_A	Test_B	Test_A	Test_B	Test_A	Test_B	Test_A	Test_B
T	BiLSTM - OCR	0.338	0.373	0.509	0.572	0.421	0.455	0.542	0.570	0.302	0.310	0.420	0.438
	BERT - OCR	0.336	0.375	0.512	0.570	0.422	0.449	0.549	0.571	0.295	0.298	0.395	0.402
	BiLSTM - OCR_{Seg}	0.352	0.391	0.560	0.594	0.475	0.490	0.570	0.594	0.319	0.332	0.422	0.442
	BERT - OCR_{Seg}	0.351	0.384	0.538	0.580	0.471	0.482	0.563	0.581	0.311	0.316	0.418	0.425
I	InceptionV3	0.322	0.358	0.516	0.557	0.407	0.430	0.499	0.525	0.288	0.287	0.402	0.406
	V16	0.318	0.355	0.521	0.560	0.399	0.432	0.505	0.532	0.286	0.295	0.411	0.418
	V19	0.325	0.367	0.525	0.562	0.413	0.448	0.518	0.550	0.292	0.300	0.405	0.419
T+I	BERT - OCR_{Seg} + V19	0.356	0.410	0.585	0.624	0.508	0.529	0.620	0.645	0.325	**0.362**	0.424	0.435
	BiLSTM - OCR_{Seg} + V19	**0.376**	**0.426**	**0.608**	**0.635**	**0.523**	**0.545**	**0.682**	**0.698**	**0.333**	0.360	**0.430**	**0.444**

Table 3: Ablation results on multimodal inputs and various feature extraction mechanisms. For the affect classification and quantification tasks, we report average scores. T: Text, I: Image, V16: VGG16, V19: VGG19.

For evaluation, we adopt the official evaluation metric of macro-F1 score and perform all analyses and comparisons considering macro-F1 only. Moreover, we report results on both Test_A and Test_B sets. We additionally report micro-F1 during the ablation study as well.

Unimodal Evaluation

We present our obtained results in Table 3. At first, we experiment with unimodal inputs, i.e., we train separate models for the textual and visual representations. For the textual modality, we employ both BiLSTM and BERT (Devlin et al. 2019) for encoding the text. On the unsegmented OCR extracted text, we obtain macro-F1 scores of 0.338 and 0.336 in sentiment classification for the BiLSTM and BERT variants, respectively. Similarly for the affect classification and quantification, we yield 0.421 and 0.302 macro-F1 with BiLSTM and 0.422 and 0.295 macro-F1 with BERT.

Subsequently, we include OCR text segments, OCR_{Seg}, into our model and observe a performance gain in macro-F1 score for all tasks. The BiLSTM variant reports macro-F1 scores of 0.352 (+1.4%), 0.475 (+5.4%), and 0.319 (+1.7%) for the three tasks, respectively. Similarly, we obtain performance improvement of +1.3%, +5.0%, and +1.1% with BERT. The performance can be easily attributed to the textual segmentation; thus it supports our hypothesis that the different textual segments reflect different semantics and should be processed separately to extract the fine-grained features. We also observe that the performance of BERT, though comparable, is on the lower side as compared to BiLSTM with both segmented and unsegmented inputs. It could be because of shorter segmented text - for which an LSTM performs very well to capture the short-term dependencies. Since BERT is a contextual word embedding model, the lack of context could be a reason for its inferiority. Moreover, in case of memes, the text is often noisy and grammatically incorrect. Instead of using full sentences, memes repeatedly use sentence segments, which are uncommon in structured language resources.

For visual modality, we employ three widely-used pretrained image feature extractors – InceptionV3 (Szegedy et al. 2016), VGG-16 and VGG-19 (Simonyan and Zisserman 2014) networks, in the current work. With VGG-19, our model yields macro-F1 scores of 0.325, 0.413, and 0.292 for the sentiment classification, affect classification, and affect quantification, respectively. Compared to the other two tasks, we observe marginally better results with VGG-19, ranging in $0.2 - 0.5\%$ performance improvement. Therefore, we choose to use VGG-19 encoding as the visual feature extraction for all other experiments.

Bi-modal Evaluation

Finally, we combine the two available modalities in a single system and learn the multimodal interactions for the underlying tasks. Similar to the unimodal case, we experiment with both BiLSTM and BERT variants for the textual encoding. Subsequently, we fuse these encodings with VGG-19 visual encoding in MHA−Meme. The resultant BERT variant reports macro-F1 scores of 0.356, 0.508, and 0.325 for sentiment classification, affect classification, and affect quantification, respectively. For the same setup, BiLSTM-based MHA−Meme obtains 0.376, 0.523, and 0.332 macro-F1 scores, respectively. We observe macro-F1 improvements of +2.0%, +1.5%, and +0.8% against the BERT-based system for the three tasks, respectively. Therefore, we prefer BiLSTM encoding in our proposed MHA−Meme.

Furthermore, the bi-modal system obtains +2.4%, +4.8%, and +1.3% improvements against the best unimodal scores for the three tasks, respectively. As the complementary and diverse information are incorporated in a single network, improvements in bi-modal setup are not surprising. Moreover, we observe higher influence of the textual modality on the overall performance. We comprehend that this result is expected as meme text contains richer information than the meme image. For example, a single image is often used for multiple memes to describe the situation; however, the semantics of the meme is often established through text.

For each case, we show the results obtained on Test_B in Table 3. Similar to Test_A, we obtain better results with multimodal inputs compared to both text and image unimodal inputs. Moreover, the BiLSTM variant yields the best results on Test_B as compared to BERT variant on bi-modal input. Thus, we argue that MHA−Meme not only performs better on Test_A, but also generalizes well on unseen random samples

System	Sent.	Affect classification					Affect quantification				
		Hum	Sar	Off	Motiv	Avg	Hum	Sar	Off	Motiv	Avg
Bs*	0.218	0.512	0.506	0.491	0.491	0.500	0.248	0.241	0.230	0.484	0.301
A*	0.352^2	0.515	0.511^3	0.512	0.520^3	0.515^2	**0.271**^{1‡}	0.250	0.258	0.512	**0.322**^{1‡}
B*	**0.355**^{1‡}	0.473	0.508	0.499	0.474	0.489	0.262	**0.259**^{1‡}	0.264^2	0.474	0.314
C*	0.345	0.516^3	**0.516**^{1‡}	**0.522**^{2‡}	0.519	**0.518**^{1‡}	0.249	0.254	0.247	0.519	0.317^3
D*	0.341	0.521^2	0.441	0.491	0.512	0.491	0.264^3	0.254	0.241	0.517	0.319^2
E*	0.346	0.514	0.504	0.512	0.507	0.511^3	0.0	0.0	0.0	0.507	0.127
K.1*	0.350^3	-	-	-	-	-	-	-	-	-	-
F*	0.325	**0.529**^{1†}	0.485	**0.529**^{1†}	0.491	0.509	0.261	0.236	**0.265**^{1‡}	0.491	0.313
G*	0.339	0.502	0.499	0.479	0.498	0.494	0.236	0.230	0.262^3	0.521^3	0.312
H*	0.323	0.493	0.487	0.505	0.490	0.494	0.237	0.255^2	0.252	0.502	0.311
I*	0.335	0.510	0.513^2	0.506	0.509	0.509	0.256	0.244	0.248	0.509	0.314
J*	0.345	0.434	0.447	0.400	0.488	0.442	0.255	0.254^3	0.241	0.488	0.310
K.2*	0.248	0.502	0.494	0.496	**0.534**^{1†}	0.506	0.140	0.233	0.261	**0.534**^{1†}	0.292
K.3*	0.323	0.486	0.500	0.472	0.522^2	0.495	0.215	0.193	0.233	0.522^2	0.291
K.4*	0.349	0.514	0.495	0.486	0.494	0.497	0.265^2	0.245	0.246	0.494	0.312
K.5*	0.337	0.500	0.483	0.516^3	0.520	0.505	0.251	0.238	0.256	0.520	0.316
MM	**0.376**†	**0.527**‡	**0.520**†	0.517	**0.531**‡	**0.523**†	**0.271**†	**0.260**†	**0.268**†	**0.531**‡	**0.333**†

Table 4: Comparative study against baselines and various state-of-the-art systems. All scores are Macro-F1 as per the official evaluation metric of the 'Memotion Analysis' shared task (Sharma et al. 2020). Superscripts ^{1},^{2}, and ^{3} denote official rank of the system in the shared task. For each case, the best and the second ranked scores among all systems are denoted by dagger(†) and double-dagger(‡), respectively. The first batch of results (after baseline, *Bs*) denotes a set of top three ranked systems for the three tasks (on average). System*: Values taken from Sharma et al. (2020). MM: `MHA-Meme`.

in Test$_B$ and is consistent across the two test sets.

Comparative Study

For the comparative study, we pitch our proposed `MHA-Meme` against the best performing systems reported by the 'Memotion Analysis' shared task (Sharma et al. 2020). We take the official scores of these state-of-the-art and baseline systems from the task-description paper (Sharma et al. 2020) for comparison. We report comparative results in Table 4. We also highlight the official rank of the used baselines as per the shared task portal, the best and the second ranked methods. The first batch of results (after baseline) in Table 4 denotes a set of top three ranked systems for the three tasks.

Sentiment Classification: The baseline system of the shared task obtains macro-F1 score of 0.217 for sentiment classification. The top three submitted systems are 'Vkeswani IITK' (B), 'Guoym' (A), and 'Aihaihara' ($K.1$), and their reported macro-F1 scores on the test set are 0.354, 0.351, and 0.350, respectively. The narrow margins among these systems reveal that the meme sentiment analysis is a complex problem and significantly different from the tradition multimodal sentiment analysis.

In comparison, `MHA-Meme` yields macro-F1 score of 0.376 with the performance improvement of $+2.2\%$ compared to the top system, 'Vkeswani IITK' (0.354). We can relate the improvement to the better handling of the fine-grained features through the segmented text and sophisticated attention mechanism.

Affect Classification: For the affect classification, we present class-wise as well as average macro-F1 scores for all systems in Table 4. We observe that `MHA-Meme` reports new sota performance (0.519) for *sarcasm* classification with $+0.4\%$ improvement over the top system, 'George.Vlad Eduardgzaharia UPB' (C) (0.515). In the *humor* and *motivation* classification, `MHA-Meme` obtains second best macro-F1 scores of 0.526 and 0.531 - a difference of -0.3% with 'Mayukh Memebusters' (F) (0.529) in *humor* and 'Saradhix Fermi' (0.534) in *motivation*. For the offense classification task, we report third best with 0.517 macro-F1 score. An important point to observe here is that except 'George.Vlad Eduardgzaharia UPB', none of the class-wise best systems (i.e., 'Mayukh Memebusters' and 'Saradhix Fermi') ranked among top three on average. In comparison, `MHA-Meme` yields better average macro-F1 score against the top system, 'George.Vlad Eduardgzaharia UPB'. Moreover, fine-grained comparison further reveals that in four out of five comparisons (four affects and one average), `MHA-Meme` performs better than the top system. Also, `MHA-Meme` records best scores in two comparisons (sarcasm and average), second ranked in two comparisons (humor and motivation), and third ranked in one comparison (offense). In contrast, the top system achieves second rank in three comparisons (sarcasm, offense, and average), fourth rank in one comparison (humor), and sixth rank in one comparison (motivation). Thus, we argue that `MHA-Meme` is not only the best system on average macro-F1 score but also poses a higher degree of generalization at the task-level as well.

	Hops	Sentiment classification						Affect classification						Affect quantification					
		D-Fusion		AT-Fusion		ATMF		D-Fusion		AT-Fusion		ATMF		D-Fusion		AT-Fusion		ATMF	
		T_A	T_B	T_A	T_B	T_A	T_B	T_A	T_B	T_A	T_B	T_A	T_B	T_A	T_B	T_A	T_B	T_A	T_B
M1	S	33.6	37.2	34.2	38.1	34.5	38.5	50.1	52.2	50.4	52.7	50.5	52.9	30.7	32.4	31.4	33.0	31.8	33.5
	M	34.0	37.5	34.4	38.6	34.9	38.9	50.3	52.6	50.5	53.0	50.8	53.2	31.5	33.1	31.9	33.8	32.0	34.3
M2	S	35.5	38.6	358	39.1	37.0	40.9	51.0	52.8	51.2	53.3	51.7	54.0	32.2	34.4	32.6	34.8	32.9	35.2
	M	36.4	40.5	37.2	41.3	**37.6**	**42.6**	51.3	53.0	51.4	53.6	**52.3**	**54.5**	32.4	34.6	32.7	35.1	**33.3**	**36.0**

Table 5: Comparative study of different fusion mechanisms and effect of single-hop attention vs multi-hop attention (in %). **M1:** BiLSTM - OCR + VGG19. **M2:** BiLSTM - OCR_{Seg} + VGG19. **S:** Single hop; **M:** Multi hops

Affect Quantification: In this task, MHA−Meme performs even better than the affect classification task. We observe that four separate systems, namely 'Guoym' (humor), 'Vkeswani IITK' (sarcasm) (Keswani et al. 2020), 'Mayukh Memebusters' (offense), and 'Saradhix Fermi' (motivation), report best scores among the submitted systems respective to four affect dimensions (mentioned within parenthesis). Moreover, only one of these systems ('Guoym') ranks in top three as per the average macro-F1 score. In comparison, MHA−Meme yields four best scores (including the official ranking metric, i.e., the average case) and one second best score across all submitted systems.

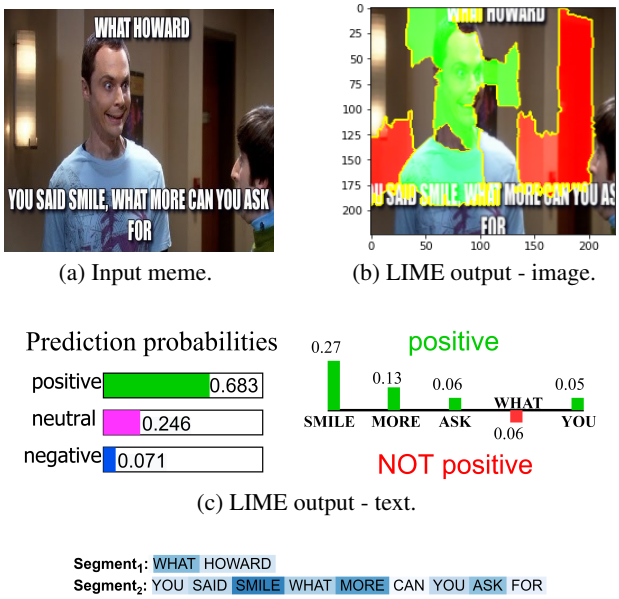
Across the three tasks, we observe three different systems ranked first in the competition, whereas, MHA−Meme reports state-of-the-art for all three cases. Moreover, for the class-wise affect classification and quantification, five separate systems rank top in eight setups (four each in affect classification and quantification). In comparison, MHA−Meme records state-of-the-art performances in four cases, second best in three cases and third best in one case.

Ablation Analysis

In this section, we present our analyses for two submodules of MHA−Meme – we report ablation studies of the multi-hop attention and attention-based multimodal fusion modules.

The notion of multi-hop attention was introduced by Lin et al. (2017) to represent the overall semantics of an utterance. This is specifically important for meme sentiment analysis as one single utterance (or textual segment) can often capture multiple emotions together. For example, a meme utterance can be sarcastic and offensive at the same time. To obtain complementary features for different objectives, we incorporate multiple hops of attention mechanism over the same input. For ablation, we show the performance of single-hop and multi-hop in Table 5. The incorporation of multi-hops yields ~ 2% improvement for different model variants on both Test_A and Test_B.

ATMF employs a hierarchical attention mechanism to amplify the contribution of important modality during fusion. To establish the efficacy of ATMF, we compare our fusion mechanism with two other variants: D-Fusion (Direct fusion) and AT-Fusion (Poria et al. 2017b). We first remove the textual segmentation operation at the input of the network and as a consequence, the whole text in each meme is treated as a single segment. Subsequently, we replace the ATMF layer with simple concatenation (Direct Fusion or D-Fusion)



(a) Input meme. (b) LIME output - image.

(c) LIME output - text.

(d) Attention weights as computed by MHA−Meme.

Figure 3: Example of explanation by LIME on both visual and textual modalities and visualization of attention weights over text tokens obtained from MHA−Meme.

and AT-Fusion layers for the experiments. AT-Fusion, which was originally introduced by Poria et al. (2017b), takes as an input audio, visual, and textual modalities and outputs an attention score for each modality. We modified the original AT-Fusion to exclude the audio modality. We report the obtained scores on Test_A and Test_B in Table 5. We can observe superiority of ATMF compared to D-Fusion and AT-Fusion in all experiments.

Interpretability of MHA−Meme

The problem of interpreting complex deep neural models is nontrivial, and at the same time, important for further exploration. Meme understanding is a complicated problem – it requires a high level of abstraction, background knowledge, and subjectivity, which are sometimes hard to expound even with human reasoning. Besides, Memotion 1.0 dataset complies memes of different topics, subjects, and genres, making the task even more challenging. Based on the prediction results,

we observe that the success of the model heavily depends on simultaneous image and text understanding. To comprehend the performance of MHA−Meme, we use LIME (Locally Interpretable Model-Agnostic Explanations) (Ribeiro, Singh, and Guestrin 2016) – a consistent model-agnostic explainer to explain the predictions in an explicated and faithful manner. It performs so by learning an interpretable model locally around the prediction.

We choose one sample meme from Test_A to visualize the explainability of our proposed model using the LIME framework. The example meme has two textual segments and is correctly classified by MHA−Meme ('positive' sentiment) as shown in Figure 3a. We apply LIME on both image and text individually, and analyze the importance of both modalities in the final classification. The prediction probabilities by MHA−Meme corresponding to positive, neutral, and negative classes are 0.683, 0.246, and 0.071, respectively. Figure 3b highlights the most contributing super-pixels to positive (green) and neutral (red) classes. As expected, the smiling face of the character, highlighted by green pixels, prominently contributes to the positive class, whereas the red pixels do not reveal any relevant information. Figure 3c demonstrates the contribution of different words from the meme text to positive and neutral/negative classes. The words 'SMILE' and 'MORE' have significant contributions to the positive class; removing these two words drastically reduces the prediction probability of the positive class. Moreover, we also plot the attention weights computed by MHA−Meme in Figure 3d for the same example. Evidently, the word 'SMILE' has the highest attention weight in the two segments, supporting the explanations by the LIME framework as well. Based on the above observations, we can conclude that the proposed MHA−Meme framework extracts and attends to the relevant features from the input representations.

Conclusion

In this paper, we addressed three tasks related to the affect analysis of a meme, namely, *sentiment classification*, *affect classification*, and *affect class quantification*. The sentiment classification has three labels [*positive*, *negative*, and *neutral*], while the affect classification and quantification have four affect dimensions [*humor*, *sarcasm*, *offense*, and *motivation*]. We proposed an attention-rich neural framework (called MHA−Meme) that analyzes the interaction between visual and textual modalities at fine-granular level, i.e., for each textual segment in a meme, it aims to extract the fine-level feature representation in correspondence with the image. We designed two attention mechanisms - a multi-hop attention module for the unimodal feature extraction and an attention-based multimodal fusion module for computing the interaction between the two modalities. Finally, we combined enriched multimodal representations of all segments via another multi-hop attention layer and forwarded it to the output layer for classification. We evaluated MHA−Meme on the recently released 'Memotion Analysis' dataset of SemEval-2020 shared task. We performed extensive experiments for each task and compared the obtained performances against 11 baseline systems (including the winners of the shared task). We observed performance improvements in the range $[+0.5\%, +2.2\%]$ for all three tasks. Furthermore, fine-grained result analysis revealed that MHA−Meme achieved consistently good performances (1st in four, 2nd in three, and 3rd in one) across eight affect dimensions, i.e., four each for affect classification and affect class quantification. In comparison, baseline systems did not report consistent performance for all the tasks or affect dimensions.

Acknowledgments

The work was partially supported by Wipro Pvt Ltd, India and CAI, IIIT-Delhi. T. Chakraborty would like to thank the support of the Ramanujan Fellowship (SERB).

References

Akhtar, M. S.; Chauhan, D. S.; Ghosal, D.; Poria, S.; Ekbal, A.; and Bhattacharyya, P. 2019. Multi-task Learning for Multi-modal Emotion Recognition and Sentiment Analysis. In *NAACL:HLT*, 370–379.

Akhtar, M. S.; Gupta, D.; Ekbal, A.; and Bhattacharyya, P. 2017. Feature Selection and Ensemble Construction: A Two-step Method for Aspect Based Sentiment Analysis. *Knowledge-Based Systems* 125: 116–35. ISSN 0950-7051.

Aldeneh, Z.; Khorram, S.; Dimitriadis, D.; and Provost, E. M. 2017. Pooling acoustic and lexical features for the prediction of valence. In *19th ACM ICMI*, 68–72.

Bahdanau, D.; Cho, K.; and Bengio, Y. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *ICLR*.

Cambria, E. 2016. Affective computing and sentiment analysis. *IEEE intelligent systems* 31(2): 102–107.

Cao, D.; Ji, R.; Lin, D.; and Li, S. 2016. A cross-media public sentiment analysis system for microblog. *Multimedia Systems* 22(4): 479–486.

Chy, A. N.; Siddiqua, U. A.; and Aono, M. 2020. CSECU_KDE_MA at SemEval-2020 Task 8: A Neural Attention Model for Memotion Analysis. In *SemEval-2020*, 1106–1111.

De la Peña Sarracén, G. L.; Rosso, P.; and Giachanou, A. 2020. PRHLT-UPV at SemEval-2020 Task 8: Study of Multimodal Techniques for Memes Analysis. In *SemEval-2020*, 908–915.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*, 4171–4186.

Gu, Y.; Yang, K.; Fu, S.; Chen, S.; Li, X.; and Marsic, I. 2018. Hybrid attention based multimodal network for spoken language classification. In *ACL*, volume 2018, 2379.

Guo, X.; Ma, J.; and Zubiaga, A. 2020. NUAA-QMUL at SemEval-2020 Task 8: Utilizing BERT and DenseNet for Internet Meme Emotion Analysis. In *SemEval-2020*, 901–907.

Guo, Y.; Huang, J.; Dong, Y.; and Xu, M. 2020. Guoym at SemEval-2020 Task 8: Ensemble-based Classification of Visuo-Lingual Metaphor in Memes. In *SemEval-2020*, 1120–1125.

Gupta, A.; Kataria, H.; Mishra, S.; Badal, T.; and Mishra, V. 2020. BennettNLP at SemEval-2020 Task 8: Multimodal sentiment classification Using Hybrid Hierarchical Classifier. In *SemEval-2020*, 1085–1093.

Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural computation* 9(8): 1735–1780.

Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely connected convolutional networks. In *CVPR*, 4700–4708.

Keswani, V.; Singh, S.; Agarwal, S.; and Modi, A. 2020. IITK at SemEval-2020 Task 8: Unimodal and Bimodal Sentiment Analysis of Internet Memes. In *SemEval-2020*, 1135–1140.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

Lee, K.-H.; Chen, X.; Hua, G.; Hu, H.; and He, X. 2018. Stacked cross attention for image-text matching. In *ECCV*, 201–216.

Lin, Z.; Feng, M.; Santos, C. N. d.; Yu, M.; Xiang, B.; Zhou, B.; and Bengio, Y. 2017. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130* .

Liu, C.; Zoph, B.; Neumann, M.; Shlens, J.; Hua, W.; Li, L.-J.; Fei-Fei, L.; Yuille, A.; Huang, J.; and Murphy, K. 2018. Progressive neural architecture search. In *ECCV*, 19–34.

Liu, P.; Zhang, L.; and Gulla, J. A. 2019. Dynamic attention-based explainable recommendation with textual and visual fusion. *Information Processing & Management* 102099.

Lu, J.; Xiong, C.; Parikh, D.; and Socher, R. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *CVPR*, 375–383.

Morishita, T.; Morio, G.; Horiguchi, S.; Ozaki, H.; and Miyoshi, T. 2020. Hitachi at SemEval-2020 Task 8: Simple but Effective Modality Ensemble for Meme Emotion Recognition. In *SemEval-2020*, 1126–1134.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *EMNLP*, 1532–1543.

Peters, M.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep Contextualized Word Representations. In *NAACL-HLT*, 2227–2237.

Poria, S.; Cambria, E.; Bajpai, R.; and Hussain, A. 2017a. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion* 37: 98–125.

Poria, S.; Cambria, E.; Hazarika, D.; Mazumder, N.; Zadeh, A.; and Morency, L.-P. 2017b. Multi-level multiple attentions for contextual multimodal sentiment analysis. In *ICDM*, 1033–1038. IEEE.

Poria, S.; Chaturvedi, I.; Cambria, E.; and Hussain, A. 2016. Convolutional MKL based multimodal emotion recognition and sentiment analysis. In *ICDM*, 439–448. IEEE.

Rahman, W.; Hasan, M. K.; Zadeh, A.; Morency, L.-P.; and Hoque, M. E. 2019. M-bert:Injecting multimodal information in the bert structure. *arXiv preprint arXiv:1908.05787* .

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. " Why should I trust you?" Explaining the predictions of any classifier. In *ACM KDD*, 1135–1144.

Sharma, C.; Bhageria, D.; Scott, W.; PYKL, S.; Das, A.; Chakraborty, T.; Pulabaigari, V.; and Gambäck, B. 2020. SemEval-2020 Task 8: Memotion Analysis- the Visuo-Lingual Metaphor! In *SemEval-2020*, 759–773.

Sharma, M.; Kandasamy, I.; and Vasantha, W. 2020. Memebusters at SemEval-2020 Task 8: Feature Fusion Model for Sentiment Analysis on Memes Using Transfer Learning. In *SemEval-2020*, 1163–1171.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* .

Szegedy, C.; Ioffe, S.; Vanhoucke, V.; and Alemi, A. A. 2017. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. In *AAAI*, 4278–4284.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *CVPR*, 2818–2826.

Vlad, G.-A.; Zaharia, G.-E.; Cercel, D.-C.; Chiru, C.; and Trausan-Matu, S. 2020. UPB at SemEval-2020 Task 8: Joint Textual and Visual Modeling in a Multi-Task Learning Architecture for Memotion Analysis. In *SemEval-2020*, 1208–1214.

Walińska, U.; and Potoniec, J. 2020. Urszula Walińska at SemEval-2020 Task 8: Fusion of Text and Image Features Using LSTM and VGG16 for Memotion Analysis. In *SemEval-2020*, 1215–1220.

Wang, Y.; Huang, M.; Zhu, X.; and Zhao, L. 2016. Attention-based LSTM for aspect-level sentiment classification. In *EMNLP*, 606–615.

You, Q.; Jin, H.; and Luo, J. 2017. Visual sentiment analysis by attending on local image regions. In *AAAI*, 231–237.

You, Q.; Luo, J.; Jin, H.; and Yang, J. 2016. Cross-modality consistent regression for joint visual-textual sentiment analysis of social multimedia. In *WSDM*, 13–22.

Yue, L.; Chen, W.; Li, X.; Zuo, W.; and Yin, M. 2019. A survey of sentiment analysis in social media. *Knowledge and Information Systems* 1–47.

Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; and Morency, L.-P. 2017. Tensor Fusion Network for Multimodal Sentiment Analysis. In *EMNLP*, 1103–1114. Copenhagen, Denmark.

Zadeh, A.; Liang, P. P.; Mazumder, N.; Poria, S.; Cambria, E.; and Morency, L. 2018. Memory Fusion Network for Multi-view Sequential Learning. In McIlraith, S. A.; and Weinberger, K. Q., eds., *AAAI*, 5634–5641.

Zhang, X.; Li, Z.; Change Loy, C.; and Lin, D. 2017. Polynet: A pursuit of structural diversity in very deep networks. In *CVPR*, 718–726.

Zhao, S.; Ding, G.; Huang, Q.; Chua, T.-S.; Schuller, B. W.; and Keutzer, K. 2018. Affective Image Content Analysis: A Comprehensive Survey. In *IJCAI*, 5534–5541.