IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 6 November 2023, accepted 24 December 2023, date of publication 28 December 2023,
date of current version 4 January 2024.

Digital Object Identifier 10.1109/ACCESS.2023.3347988

RESEARCH ARTICLE

Capturing the Concept Projection in Metaphorical Memes for Downstream Learning Tasks

SATHWIK ACHARYA[1], **BHASKARJYOTI DAS**[2], **AND T. S. B. SUDARSHAN**[1], **(Senior Member, IEEE)**
[1]Department of Computer Science and Engineering, PES University, Bengaluru, Karnataka 560085, India
[2]Department of Computer Science and Engineering in AI & ML, PES University, Bengaluru, Karnataka 560085, India

Corresponding author: Bhaskarjyoti Das (Bhaskarjyoti01@gmail.com)

ABSTRACT Metaphorical memes, where a source concept is projected into a target concept, are an essential construct in figurative language. In this article, we present a novel approach for downstream learning tasks on metaphorical multimodal memes. Our proposed framework replaces traditional methods using metaphor annotations with a metaphor-capturing mechanism. Besides using the significant zero-shot learning capability of state-of-the-art pretrained encoders, this work introduces an alternative external knowledge enhancement strategy based on ChatGPT (chatbot generative pretrained transformer), demonstrating its effectiveness in bridging the intermodal semantic gap. We propose a new concept projection process consisting of three distinct components to capture the intramodal knowledge and intermodal concept gap in the forms of text modality embedding, visual modality embedding, and concept projection embedding. This approach leverages the attention mechanism of the Graph Attention Network for fusing the common aspects of external knowledge related to the knowledge in the text and image modality to implement the concept projection process. Our experimental results demonstrate the superiority of our proposed approach compared to existing methods.

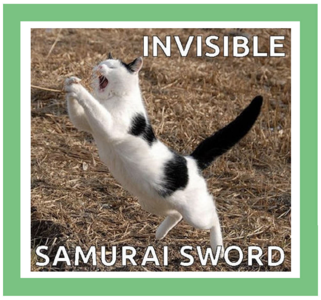
INDEX TERMS Memes, metaphor, concept projection, cognitive computing, multimodal machine learning, knowledge graph, large language models.

I. INTRODUCTION

Metaphors are at the intersection of computational linguistics, cognitive science, and psychology. They are a type of figurative language that is different from literal language and are rhetorical devices applied to pieces of literal language resulting in an 'emergent meaning'. Metaphors are different from other constructs of figurative language, i.e., sarcasm, irony, simile, satire, hyperbole, and humor. Three metaphor theories mostly inspire existing research about metaphor in Computational Linguistics. Metaphor Identification Procedure (MIP) [2] is built around the semantic contrast of the basic and contextual meaning aided by large annotated data. Selectional Preference Violation (SPV) Theory [3] is based on the abnormal word-pair association. According to Computational Metaphor Theory(CMT) [4], metaphors are creative cognitive constructs that map a source concept to a target concept using a common attribute. Typically, the source concept is more concrete, while the target concept is more abstract, requiring a cognitive mechanism. In the case of a multimodal meme, the source concept can be in text, whereas an image can represent the target concept. Such conceptual metaphors can be conventional or novel. Multimodality in such 'novel' cases adds additional semantic complexity. For example, in Figure 1, the concept transition has to occur between the source domain (in image) and the target domain (in text) with the help of external knowledge about samurai swords. The pair of a source and a target concept in a metaphor is also called the pair of a tenor and a vehicle [5] where the vehicle gives its property to the tenor. The concept projection between the tenor and the vehicle in a metaphor requires [6] substantial cognitive ability and contextual knowledge. While a function approximation of a cognitive mechanism can be attempted using a machine learning model, the knowledge aspect requires a specific knowledge input into the learning mechanism.

It is well established that images are more effective [7], [8] than pure text in communication. This explains the

The associate editor coordinating the review of this manuscript and approving it for publication was Laxmisha Rai.



- Metaphor occurence: True
- Category: Complementary
- Source domain: Air
- Target domain: Samurai sword
- Source modality: Image
- Target modality : Text
- Offensiveness: Non offensive
- Sentiment category : Anger
- Intention: Expressive

FIGURE 1. An example of a metaphorical meme found in the MET-meme data set [1] along with the annotations.

relative popularity of memes as a linguistic construct. Making meaning out of memes is a type of multimodal machine learning task [9] in which inputs are made of 'atomic units of information' of different modalities. Making a successful interpretation of a multimodal meme is inherently a multicontextual learning problem, as it also involves knowledge context in addition to the multimodality of the content itself. Hence, memes are frequently used in all formats of disinformation campaigns that are inherently multicontextual [10] in nature.

Metaphors have been well-researched in Natural Language Understanding (NLU) and Computer Vision. The two main tasks in linguistic metaphor research are metaphor identification and interpretation, where interpretation depends on successful identification. In existing linguistic metaphor research, metaphor interpretation must incorporate concept transition [11], [12] in a paraphrased rendering of the intended meaning. Even efforts towards metaphor identification are yet to be immensely successful [13] in the case of novel conceptual metaphors.

Compared to NLU and Computer Vision, more work must be done on metaphorical memes and consequent downstream learning tasks. Existing research has been primarily about downstream discriminative tasks that do not use metaphor detection as a building block, i.e., detection of hateful memes [14] and offensive memes [15], sarcasm detection [16], characterization of meme entities [17], and detection of sentiment [18], [19], [20], [21] as well as emotion in memes. Even for purely metaphor-centric research on memes, there needs to be more work in metaphor identification and interpretation.

For multimodal memes relying on conceptual metaphors, the literal meaning or local knowledge resident in the image and text of the meme is not enough for its semantic interpretation. A learning model tasked with interpreting the meme needs to implement the cognitive mechanism of concept projection with the help of global knowledge, i.e., a combination of the resident local knowledge and acquired global knowledge is necessary. The work described in this article addresses this requirement by acquiring this global knowledge and modeling the concept projection mechanisms. A three-pronged strategy of state-of-the-art visual-linguistic pretrained models with significant zero-shot learning capability for harnessing local knowledge, innovative use of ChatGPT for harnessing global knowledge, and a graph attention-based mechanism for modeling concept projection has been used in this work to achieve state-of-the-art performance in downstream learning tasks.

The remainder of this article is structured as follows. Section II reviews the state-of-the-art in the related areas. Section III discusses the proposed approach in detail. In Section IV, the data set used in this work and the implementation details are described. In Section V, the overall performance, analysis of the results, and ablation study are provided. Section VI concludes the article.

II. RELATED WORK

A. MULTIMODAL MACHINE LEARNING FOR MEMES

At a top level, the existing work in multimodal machine learning can be categorized as follows:

- **Fusion:** Several strategies have been proposed that rely on variants of concatenation methods such as [22], [23], [24], and [25], early fusion, late fusion, hybrid fusion, tensor fusion, hierarchical fusion etc. depending on how the different modalities are concatenated. For meme-related downstream tasks, it was found that early fusion is most effective [26] among fusion approaches. Zhao et al. [27], in the Facebook hateful meme challenge data set, reported the same using CLIP-based embedding of image and text modality, i.e., early fusion performed much better than other fusion strategies. However, the fusion approach does not utilize the intermodal semantic gap and is not expected to work well for metaphorical memes.

  Attention schemes between modalities are based on the assumption that the modalities are complementary and use the attention mechanism [28], [29], [30], [31], [32] between modalities to capture the intermodal information. It then concatenates this information with the information from individual modalities. However, the co-attention schemes work better for modalities that complement each other in visual-linguistic tasks such as Visual Question Answering. For contradicting information from modalities where one of the modalities has 'benign confounders' requiring concept transition, this approach is not the most effective. This issue was observed in the FaceBook Hateful Memes Challenge [33], [34] where the data set had 'benign confounders'. Therefore, this approach needs to be also revised for metaphorical memes.

  The latent feature vector approach projects the different modalities into a common latent feature space using methods such as the variational autoencoder(VAE) [35], [36]. This approach also needs to pay attention to the intermodal information and has not been used so far for meme-related tasks.
- **Visual Linguistic (VL) models:** The visual-linguistic models [37], [38], [39], [40] are an extension of the transformer-based architecture in NLP and follow

an overall strategy of pretraining and fine tuning. These models have reached state-of-the-art in many visual linguistic tasks. However, these models have challenges [41] such as cross-modal alignment, transferability issues such as cross-task gap, cross-lingual gap, overfitting, and distribution gap between training and deployment. Although the VL models work well for complementary modalities in applications such as Visual Question Answering(VQA), some of the above shortcomings were observed in the FaceBook Hateful Memes Challenge [33], [34] itself. The cross-modal alignment issue is expected to become prominent in the case of metaphorical memes. Similarly, as the text-based hate detection models were found to have bias [42], VL models for the hate meme detection task also exhibited bias [43] and consequent lack of generalizability. Without any specific effort to source global knowledge to address discordant modalities, researchers have used other alternatives with none of them achieving any outstanding results, i.e., sentiment [44], multitask learning [45], data augmentation [46], and an ensemble [47], [48] of models. Kiran et al. [49] have adopted an ensemble approach consisting of a Visual Linguistic(VL) model and cross-attention scheme with suitable domain augmentation to record a sizable performance in hate meme detection but still fail in cases of the semantic gap between modalities.

CLIP [50] is a recent significant effort to address the transferability challenge faced by VL models. CLIP learns a shared multimodal embedding by large-scale pretraining and provides zero-shot learning capability across training and deployment. Recently BLIP-2 [51] has implemented an efficient, cost-effective pretraining strategy while enhancing performance in zero-shot learning. Hence, in this work, we have used BLIP-2 [51] to harness the local knowledge available in each modality of the memes but devised other mechanisms to harness intermodal information.

- **Knowledge-based approach for discordant modalities:** This is an evolving area of research. Learning tasks [33], [52], [53], such as the detection of offensive memes, emotion, and propaganda techniques, have only had moderate success so far. The MediaEval 2022 NewsImages task [54] was also built around the complex semantic relationships [55] between the modalities.

  Few researchers have used the intermodality gap as a feature [44], [56], [57], [58], [59], [60] for the downstream machine learning task. Capturing inconsistency between modalities is effective only in specific scenarios like fake news detection, where this inconsistency itself is an attribute of fakeness. However, capturing inconsistency between modalities is insufficient to model concept transition in metaphors.

  Internal knowledge is the information in each modality and is stored as implicit knowledge as weights of the neural network of visual linguistic models. On the other hand, the main external knowledge sources are WordNet, Wikidata, DBPedia, ConceptNet [61], and Visual Genome [62]. The recent work on multimodal memes has mainly used ConceptNet [16], [17], [63], [64], [65] as an external knowledge source. The reason for choosing ConceptNet is that it encompasses several knowledge categories, such as part-whole relationships, utility relationships, factual knowledge, behavioral knowledge, common event knowledge, etc., which otherwise have to be sourced from individual knowledge repositories.

B. LARGE LANGUAGE MODELS AS A SOURCE OF WEB KNOWLEDGE

Web knowledge encompasses both external and internal knowledge [66] and offers a significant advantage. Therefore, the work described in this article relies on the web knowledge rather than common sense knowledge sourced from ConceptNet. The recent emergence of large language models such as ChatGPT provides an excellent opportunity [67] as a source of web knowledge in knowledge-intensive tasks. ChatGPT is powered by GPT-3.5, a language model trained on a wide range of text data from the internet, making it capable of answering a wide array of questions on numerous topics, from general knowledge to technical inquiries, offering a versatile and adaptable knowledge base. This, coupled with its ability to understand context and generate coherent responses, allows ChatGPT to provide more nuanced and detailed answers to questions, enhancing the quality and accuracy of the information.

ChatGPT has been successfully used in several knowledge-intensive tasks as a replacement for traditional knowledge sources such as Wikipedia, i.e., zero-shot information extraction [68], knowledge-based data augmentation [69], knowledge-based question-answering [70], text classification [71] and key phrase generation [72]. In key phrase generation, ChatGPT is seen to outperform [73] state-of-the-art models. The work described in this article has adopted ChatGPT as a source of web knowledge.

C. DOWNSTREAM LEARNING TASKS ON MEMES

Specifically, concerning the downstream learning tasks, such as the detection of intent, sentiment, and offensive nature of the content, the existing work is still a 'work in progress'. The challenge in these tasks is in capturing the authorial intent that depends on the literal meaning of the modalities and semiotic relations between modalities.

The harmful meme detection task has a significant overlap with the offensive meme detection task. Shivam Sharma et al. [14] in their survey on harmful meme detection, list complex abstraction, insufficient sample size, contextualization, and subjectivity in annotation as some of the critical challenges. Yehia Elkhatib et al. in 'Memes to an end' [74] point out that factors such as cultural aspect, demographics of users, and knowledge about the entities in the memes are essential to assess the offensive nature. Lanyu Shang et al. [15] proposed a framework capturing the analogy between modalities to detect the offensive nature of memes. It extended the existing deep learning-based approach using a fusion of global and local visual context captured by detected objects, text features, and user comments. When writing this article, Knowmeme [63] is the only existing work for offensive meme detection tasks that use knowledge context by building a knowledge graph based on ConceptNet.

More work needs to be done on the detection of intent in memes. Diaz and Ng [75] pointed out that capturing the hidden message and intention behind a meme is by itself a novel task that is challenged by the lack of a specific corpus with a suitable taxonomy of intent. Kruk et al. [76] proposed a multimodal data set based on Instagram posts, but so far, researchers have not been able to completely capture the 'complex meaning multiplication' mentioned in this work.

The sentiment and emotion detection tasks on memes have seen comparatively more work. The domain of sentiment analysis [77] and opinion mining of customer reviews [78] is not yet subjected to any advanced investigation for metaphorical memes. Most of the existing work in sentiment analysis [21] has been based on different fusion approaches except for a few [18], [19], [20] that employ a graph-based approach to capture complex intermodal relationships.

There is a single existing 'all-in-one' approach [79] for downstream learning tasks in multimodal memes, where a multitask deep learning framework has been used to learn all downstream tasks. However, that work does not use metaphor-capturing functionality as a building block.

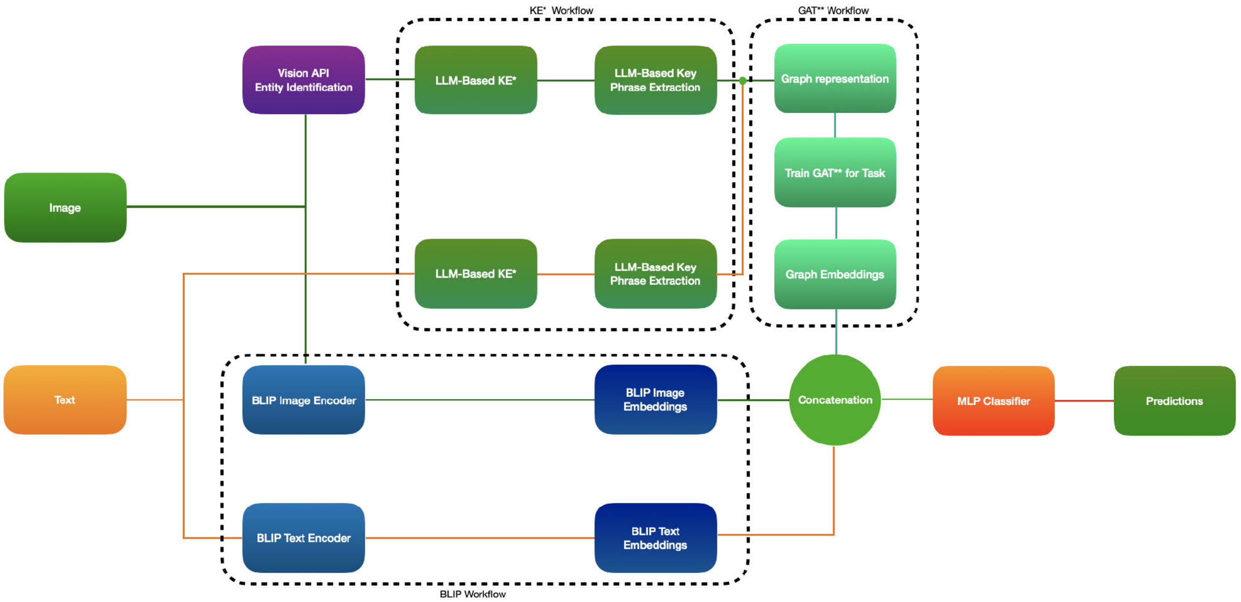
D. EXTERNAL KNOWLEDGE INTEGRATION STRATEGY

The predominant source of external knowledge in meme-related learning tasks has been ConceptNet. External knowledge integration strategies in the existing work with multimodal memes can be divided into two categories.

- **Strategies based on Attention and Fusion:** Shivam Sharma et al., in their AOMD framework [15], used comments on social media as the source of external knowledge along with the Cross-Attention mechanism in the model for detection of offensive memes. Shivam Sharma et al., in a recent work [17], have combined information from images, text, and knowledge from ConceptNet. Although a Graph neural network has been used to get a knowledge graph embedding of acquired knowledge from ConceptNet, the main concept transitioning mechanism is a state-of-the-art Optimal Transport Based Kernel Embedding Layer [80] replacing the traditional cross-attention mechanism. Vasiliki Kougia et al. in MEMEGRAPHS [64] acquire knowledge from Wikidata as text, use the scene graph to capture the local context in the image, and convert the entities in the scene graph to text. The above two pieces of information in the text are then concatenated with the meme text for the downstream learning task. These approaches mentioned here do not explicitly use the intermodal semantic gap.
- **Graph-based approaches:** Although graph neural network-based approaches have recently emerged as a natural mechanism for multicontextual learning [10], very few attempts so far have used this approach for downstream learning tasks in multimodal memes. Investigations in cognitive systems have also been undertaken using three-dimensional data points as point clouds that offer modeling opportunity [81]. Emotion, in particular, can be investigated [82] using a point-cloud dataset, and a graph neural network can be a possible approach [83]. However, multi-contextual cognitive constructs like sarcasm and metaphor have not been investigated yet in such a setup.
  In an interesting parallel, in multimodal sentiment analysis, researchers [18], [19], [20] have recently used a graph-based approach to include intermodal and intramodal information. Knowmeme [63] is a graph-based multicontextual learning approach to detect offensive memes. In this, entities are derived from both the content modalities, and the external knowledge is derived from ConceptNet. Both contexts are represented as nodes on a graph with edges between the nodes. A graph convolution network is then used to get the node embeddings and a self-attention mechanism is used. Tan Yue et al. proposed KnowleNet [16] for sarcasm detection in memes. In this work, meme text and annotation generated from meme images are taken as the first two modalities, and the meme image is taken as the third modality. The fourth modality is constructed from the Knowledge Graph extracted from ConceptNet using attribute words from the image and text modalities. The embeddings of image and text are concatenated with the semantic similarity derived from the first two modalities and the embedding of image and text entities from the knowledge graph.
  The works discussed above either use a graph convolution network to fuse the acquired knowledge together or use similarity as a proxy for the concept transition happening in a meme. The work described in this article has adopted a different strategy. For multicontextual learning, the Graph Attention Network(GAT) [84] offers a more robust mechanism to capture interactions between different contexts, and recently researchers have successfully used GAT in other domains such as aspect-level sentiment analysis [85], social trust [86], drug target interaction [87], sarcasm detection [88], and fake news detection [89]. Therefore, we have adopted GAT as a building block in the concept transitioning mechanism.

E. LIMITATIONS OF THE EXISTING WORK

1) In multimodal metaphorical memes, metaphors are a means to an end, such as offensive content, sentiment,



* KE - Knowledge Enhancement, ** GAT - Graph Attention Network

FIGURE 2. Workflow diagram consisting of the modality specific encoders, knowledge enhancement module and the task specific classifier.

and intent. Therefore, the success in downstream learning tasks depends on the successful detection of metaphors. However, existing research does not implement the metaphor detection task as a fundamental building block and instead relies on metaphor annotation.

2) The existing work on multimodal memes mainly relies on complementary and concordant modalities, with some very recent work that utilizes the semantic similarity between the modalities as an additional feature to capture the intermodal gap. However, capturing intermodal similarity may only partially model the concept projection across modalities.

F. CONTRIBUTIONS OF THIS WORK

1) The framework described in this article implements a metaphor-capturing mechanism instead of using the metaphor annotations available in the data set.
2) The work implements an alternative mechanism to integrate external knowledge. This work uses ChatGPT as a source of web knowledge instead of relying on ConceptNet as in the existing work. The results delivered prove the viability of this approach.
3) This work models the intermodal concept projection process in metaphorical memes through the attention mechanism of the Graph Attention Network to fuse the common aspects of external knowledge related to the entities present in individual modalities.

III. PROPOSED METHOD

In this section, we outline the methodology employed for the different tasks involving the multimodal memes from the MET-meme data set. As shown in Figure 2 and Algorithm 1, at a high level, our method consists of the following:

Algorithm 1 Multimodal Meme Subtask Label Estimation

```
Input: a set of $N$ meme posts, $M$ test meme posts, ChatGPT knowledge base $K$
Output: estimated labels for the subtask $\hat{y}$
Training Phase
for each $M_i$ in $N$ do
    Perform Knowledge enhancement using K and extract graph embeddings of $M_i$
    Extract visual and textual embeddings of $M_i$
    Concatenate the visual,textual and graph embeddings of $M_i$ to be fed into classifier
end for
Train classifier by minimizing $L$ (Eq. 3)
Classification Phase
Initialize: $\hat{y} = []$
for each $M_i$ in $M$ do
    Use the trained classifier to predict $\hat{y}_i$
    $\hat{y} \leftarrow \hat{y}_i$
end for
Output $\hat{y}$
```

- **Intramodal knowledge:** Modality-specific encoders encode the unimodal information(text and image) into embeddings.
- **Intermodal knowledge and concept projection:** Knowledge enhancement module, which consists of the pipeline used for external knowledge extraction followed by subsequent graph construction and graph representation to implement concept projection.

- **Classifier:** A classifier fuses the above-obtained embeddings(via concatenation) and feeds it to a downstream multilayer perceptron to obtain the predictions of the class labels for the subtask at hand.

A. MODALITY SPECIFIC ENCODER

In our research, the use of vector representations is essential to effectively capture the essential factors of variation inherent in both the text and the images associated with each meme. This comprehensive representation allows us to extract meaningful and semantically rich features for meme analysis. As our research primarily focuses on the downstream tasks of meme analysis, we place our emphasis on the innovative application of existing modality-specific encoders rather than developing new ones. We employ techniques of transfer learning, leveraging the power of pretrained models, to fulfill our encoding needs.

1) TEXT ENCODER

Given a meme caption x_t, we employ pretrained text encoders to acquire its vector representation. These embeddings encapsulate the semantic content of individual tokens and their contextual interrelationships within the sequence, offering a comprehensive textual representation.

These embeddings typically have the shape of $[1, t, 768]$. The first dimension (1) signifies the number of instances or samples processed, illustrating that the text encoder generates embeddings sequentially, one text sample at a time. The second dimension (t) denotes the number of tokens within the text sequence. Tokens, which are fundamental units of text such as words or subwords, are processed independently by the encoder. The third dimension (768) indicates the size or dimensionality of each token's embedding. These vectors encode both the inherent semantic meaning of the token and its contextual relevance within the input text.

To derive the final text embedding $e_t \in \mathbb{R}^{768}$ representing the entire caption, we compute the mean of the embeddings associated with the t tokens. While this averaging operation results in a loss of positional information, it effectively preserves the overarching context and semantics of the complete caption.

The calculation for the mean text embedding e_t can be expressed as follows:

$$e_t = \frac{1}{t}\sum_{i=1}^{t} x_i \tag{1}$$

where x_i represents the embedding of the i-th token in the sequence.

2) IMAGE ENCODER

Given a meme image x_v, we use pretrained image encoders to obtain its vector representation. These embeddings capture the semantic meaning of the different image patches, providing a rich image representation.

These embeddings are typically of shape [1,p,768]. The first dimension (1) represents the number of samples or instances. In this case, it indicates that the image encoder produces embeddings for a single image sample at a time. The second dimension (p) represents the regions of interest (RoI) or image patches extracted from the input image, which are then encoded separately. The third dimension(768) denotes the size of the embedding vector for each region of interest.

We finally obtain the image embedding $e_v \epsilon \mathbb{R}^{768}$ by taking the mean of the embeddings of p patches to obtain a global representation of the entire image, which can capture the overall context and the salient features. Use of individual region embeddings may lead to a sizeable final feature vector, which could be computationally expensive and lead to overfitting, especially in cases with limited data. Therefore, the final image embedding is the mean embedding value of the above patches. This becomes a concise yet informative representation of the input image so that it can be effectively used in various downstream tasks.

The calculation for the same is as follows:

$$e_v = \frac{1}{p}\sum_{i=1}^{p} x_i \tag{2}$$

where x_i represents the embedding of the i-th patch in the image.

B. KNOWLEDGE ENHANCEMENT MODULE

The work described in this article relies on the idea of augmenting the multimodal input with additional external knowledge about the image and textual inputs. To this end, we propose the usage of ChatGPT as a knowledge base and the subsequent representation of this knowledge in the form of a graph (illustrated in Figure 3). The knowledge enhancement module consists of the following:

1) EXTRACTION OF ENTITIES FROM THE MEME IMAGE

To obtain additional knowledge for a given meme image, it is imperative to recognize what the image is visually composed of. Identification of the different entities present in the meme is necessary because ChatGPT does not inherently accept image inputs. To circumvent this limitation, we use Google Cloud Vision API to identify the entities present and provide a brief caption for the meme image. Cloud Vision API allows one to easily integrate vision detection features into applications. This integration encompasses features such as image labeling, face and landmark detection, optical character recognition (OCR), and the ability to tag explicit content.

We have specifically chosen Google Cloud Vision API here over other SOTA image captioning and object detection models [90] because Google's Vision API leverages pretrained models that have been trained on a wide range of images and have learned to recognize a vast number of entities and objects. These models have already captured

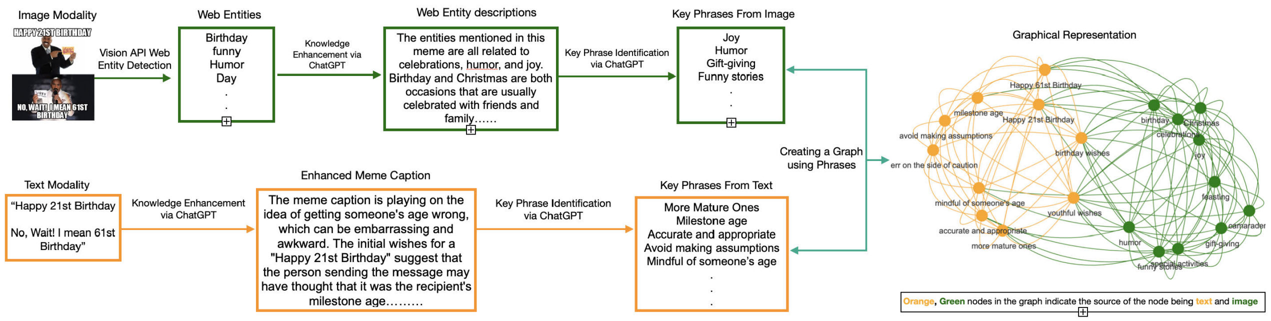


Image Modality
HAPPY 21ST BIRTHDAY
NO, WAIT! I MEAN 61ST BIRTHDAY
Vision API Web Entity Detection
Web Entities
Birthday
funny
Humor
Day
Knowledge Enhancement via ChatGPT
Web Entity descriptions
The entities mentioned in this meme are all related to celebrations, humor, and joy. Birthday and Christmas are both occasions that are usually celebrated with friends and family......
Key Phrase Identification via ChatGPT
Key Phrases From Image
Joy
Humor
Gift-giving
Funny stories
Creating a Graph using Phrases
Text Modality
"Happy 21st Birthday
No, Wait! I mean 61st Birthday"
Knowledge Enhancement via ChatGPT
Enhanced Meme Caption
The meme caption is playing on the idea of getting someone's age wrong, which can be embarrassing and awkward. The initial wishes for a "Happy 21st Birthday" suggest that the person sending the message may have thought that it was the recipient's milestone age.........
Key Phrase Identification via ChatGPT
Key Phrases From Text
More Mature Ones
Milestone age
Accurate and appropriate
Avoid making assumptions
Mindful of someone's age
Graphical Representation
Happy 61st Birthday
milestone age
Happy 21st Birthday
avoid making assumptions
err on the side of caution
birthday wishes
mindful of someone's age
youthful wishes
accurate and appropriate
more mature ones
birthday
Christmas
celebrations
joy
feasting
camarader
humor
gift-giving
funny stories
special activities
Orange, Green nodes in the graph indicate the source of the node being text and image

FIGURE 3. Knowledge enhancement module.

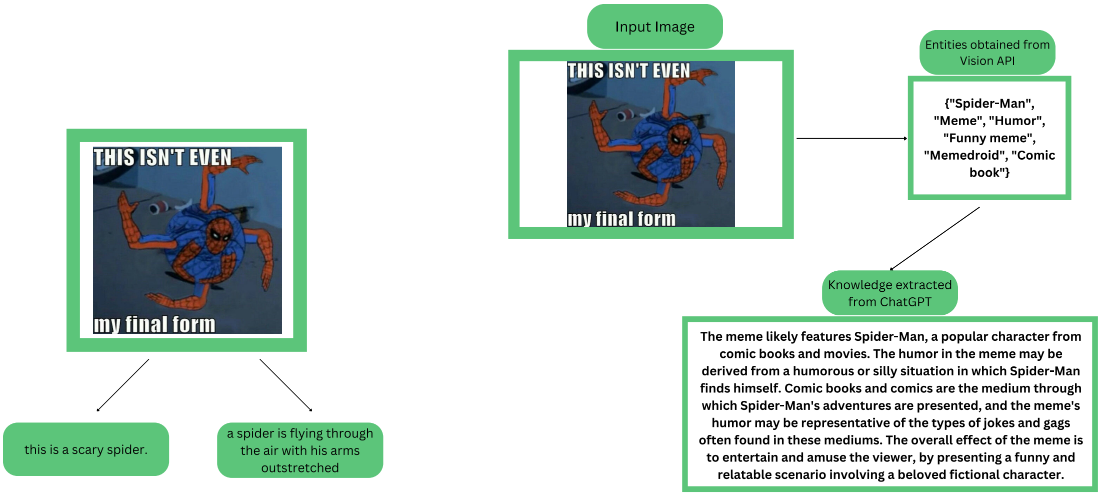

THIS ISN'T EVEN
my final form
this is a scary spider.
a spider is flying through the air with his arms outstretched
Input Image
THIS ISN'T EVEN
my final form
Entities obtained from Vision API
{"Spider-Man", "Meme", "Humor", "Funny meme", "Memedroid", "Comic book"}
Knowledge extracted from ChatGPT
The meme likely features Spider-Man, a popular character from comic books and movies. The humor in the meme may be derived from a humorous or silly situation in which Spider-Man finds himself. Comic books and comics are the medium through which Spider-Man's adventures are presented, and the meme's humor may be representative of the types of jokes and gags often found in these mediums. The overall effect of the meme is to entertain and amuse the viewer, by presenting a funny and relatable scenario involving a beloved fictional character.

FIGURE 4. Left Spiderman subfigure corresponds to the captions generated by GIT model(denoted by left arrow) and BLIP2(denoted by right arrow). The right Spiderman subfigure shows the knowledge enhancement via Vision API and ChatGPT.

a significant amount of visual knowledge, making them effective for various image analysis tasks. This is illustrated in Figure 4. The left subfigure in Figure 4 shows the captions generated for the given meme using GIT [91] and BLIP-2 [51] models. These models are chosen here for illustration due to their SOTA performance in image captioning tasks. However, these models, do not capture the necessary external knowledge required as compared to Vision API. This justifies the usage of Vision API over SOTA image captioning models to harness knowledge about the image modality.

2) QUERYING CHATGPT TO RETRIEVE RELEVANT KNOWLEDGE

As mentioned above, in this work, we propose the usage of ChatGPT as a source of external knowledge to provide additional information about both the image and text modality. The meme caption is fed to ChatGPT with the query as shown in Figure 3 with the prompt 'What does the following caption mean in the context of a meme?'. The set of entities obtained from Vision API is fed to ChatGPT with the prompt 'What does the set of these entities obtained from an image mean in the context of a meme?'. Let k_t and k_i be the knowledge extracted for the text and the image modality, respectively.

3) EXTRACTION OF KEY PHRASE FROM EXTRACTED KNOWLEDGE

The retrieved knowledge from ChatGPT, k_t and k_i are then subjected to key phrase extraction for subsequent graph construction. Key phrases capture the essence and main ideas of the text. To achieve the same, we again query and use ChatGPT to extract key phrases, as shown in Figure 3. The choice of ChatGPT over task-specific models such as Phraseformer [92], PromptRank [93], KeyBART [94] is mainly due to its better ability [72], [73] to capture the right key phrases.

4) GRAPH CONSTRUCTION AND SUBSEQUENT GENERATION OF GRAPH EMBEDDINGS

By converting the key phrases for k_t and k_i obtained from ChatGPT into nodes, the resulting graph embeddings can represent the semantic connections in terms of similarity, co-occurrence, or hierarchy between these important concepts, thus offering a more compact and meaningful representation of the extracted knowledge. More specifically, each node has an attribute called 'embeddings' that stores the embeddings of that particular phrase. These phrase embeddings are obtained by Phrase-BERT [95], which fine-tunes the BERT model toward the characteristics of phrases by using a data set of phrasal paraphrases.

The nodes mentioned above are then connected with an edge only if the cosine similarity of the embeddings of the phrases between the two nodes is higher than a given threshold, i.e., 0.2 in this case. The edges are also associated with a 'weight' attribute equal to the cosine similarity value. There are different ways to generate graph embeddings, i.e., random walk-based algorithms such as DeepWalk [96], node2vec [97], and Graph2Vec [98] and graph neural network approaches such as GraphSAGE [99], Graph Attention Network(GAT) [84] etc. For the reasons mentioned earlier, this work has adopted GAT due to its ability to capture intricate relationships within graph-structured data. GAT employs attention mechanisms to assign varying importance to neighboring nodes when aggregating information. This attention mechanism allows GAT to adaptively learn the significance of different connections, making it particularly well suited for capturing nuanced patterns and semantic information present in our graph data. This GAT model is first trained for that particular subtask (such as intention detection, etc) by formulating it as a graph classification problem. This trained model is then used to generate per-graph embeddings. The embeddings obtained is of shape $g_m \epsilon \mathbb{R}^{128}$.

IV. EXPERIMENTAL SETUP

A. DATA SET SELECTION

There are very few data sets available for downstream learning tasks on metaphorical memes. Existing data sets for the detection of the propaganda [52], offensive nature [100], and hate [101] are not specifically developed for multimodal metaphor research. Metaclue [102] is a recently published data set of multimodal metaphors where both the primary and secondary concepts are in the image. It means that all metaphors in this data set are 'image-dominating' type, and so this data set was not considered for our task. Both FigMemes [103] and recently published Irfl [104] are data sets for detecting different classes of figurative language and, therefore, were unsuitable for our work. Two data sets in particular were found to be suitable for investigating metaphorical memes, i.e., MET-meme [1] and Multimet [105]. MET-meme offered additional annotations towards metaphor understanding, i.e., source and target domain, along with a baseline accuracy for downstream machine learning tasks where the baseline model used the metaphor annotations. It also offered additional labels for offensiveness. Since the goal of this work was to capture concept transitions in conceptual metaphor without using metaphor annotations and evaluate the effectiveness of this approach for downstream learning tasks, MET-meme has been chosen for our work.

B. DATA SET

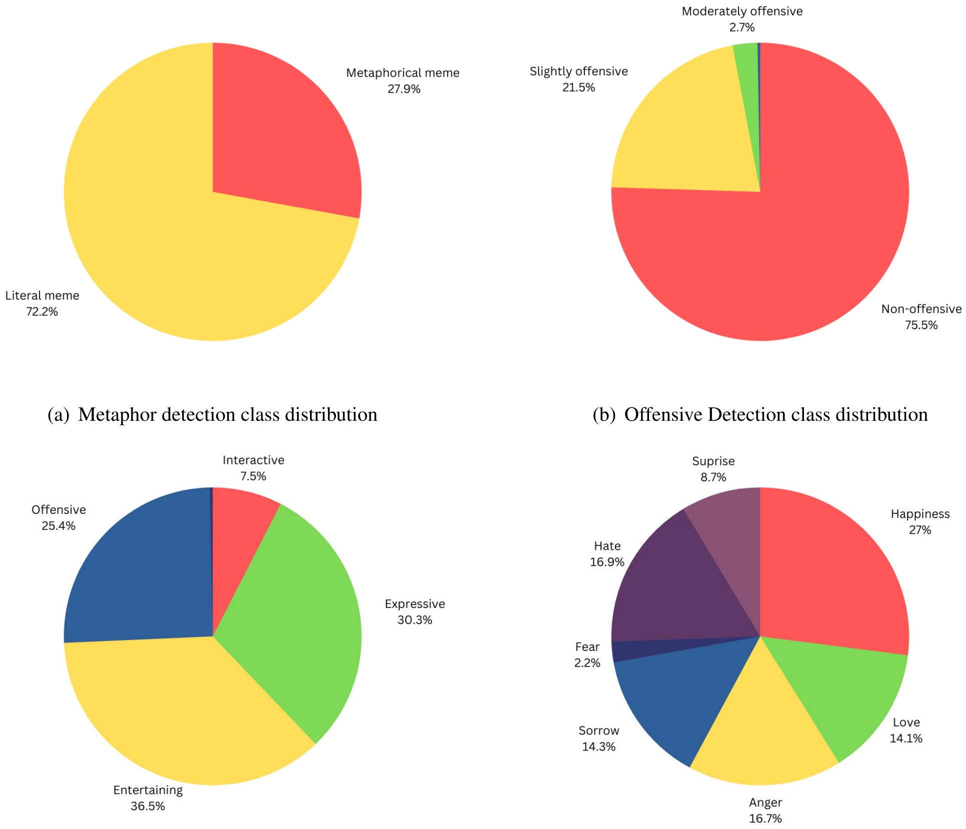
The MET-meme [1] data set is a large-scale multimodal (image+text) metaphor meme data set to support research on better understanding of memes. The data have been collected from public portals that include social media sites(Twitter and Weibo), meme sharing websites, Google images, and Baidu images. To ensure diversity and relevance, the data set spans different languages(Chinese and English in this case), genres, and themes. In total, there are 6045 Chinese and 4000 English memes, respectively. In this work, only the English meme data set has been used for experimentation.

To construct a comprehensive multimodal metaphor meme dataset, the authors of the MET-meme dataset sourced the data from various public platforms such as Twitter, Weibo, Google, and Baidu images. No personal data, such as user IDs or usernames, was retained to safeguard user privacy. Meme types were employed as search keywords for acquiring memes from Google, Baidu, and Weibo images. The English meme dataset, comprising 4000 text-image pairs, originated from the MEMOTION dataset [53] and Google search. Non-memes and duplicate text-image pairs were excluded, ensuring memes contained clear background images and textual content.

Two distinct annotation approaches were employed: one for conceptual metaphor understanding and the other for sentiment, intent, and offensiveness annotations. For metaphor annotation, the 'adjective-noun' and 'verb-noun' methods were used for source and target domains, respectively. Annotation tasks were conducted by 12 NLP postgraduate students and eight research assistants familiar with metaphor concepts. Semantic annotation tasks, encompassing sentiment, intention, and offensiveness, were outsourced to a professional crowdsourcing company. Annotators, presented with text and images randomly, utilized OCR APIs to extract meme text, which was manually proofread. Each meme underwent annotation by a minimum of 3 annotators. Stringent quality control measures, including statistical assessments like kappa score (k) in Fleiss kappa tests, were implemented to ensure inter-annotator agreement consistency.

The MET-meme data set proposes four tasks, namely, Metaphor detection, Sentiment Analysis, Intention Detection, and Offensiveness Detection. Figure 5 summarizes the class distribution across the four subtasks. In this article, we focus on all these subtasks.

1) **Metaphor detection**: This subtask focuses on whether a given meme has a metaphorical expression(metaphorical meme) or not(literal meme). These metaphorical memes are divided into three subcategories: text-dominant, image-dominant, and complementary. In the complementary setting, the image and text are required to understand the metaphorical information.
2) **Sentiment analysis**: Sentiment analysis, a crucial aspect of natural language processing, has received significant attention from numerous researchers [106], [107]. In MET-meme, the sentiment classification approach introduced in the 'Sentiment Vocabulary Ontology' [108] has been adopted with the definition of



(a) Metaphor detection class distribution

(b) Offensive Detection class distribution

(c) Intention Detection class distribution

(d) Sentiment Analysis class distribution

FIGURE 5. Data distribution of the classes across the respective subtasks.

seven distinct sentiment categories to suit the specific emotional style prevalent in memes, i.e., happiness, love, anger, sorrow, fear, hate, and surprise.

3) **Intention Detection**: The intentions behind memes can be categorized as expression, entertainment, and social integration, among others, as can be seen in some existing work [109], [110]. Taking inspiration from the eight intentions identified for tweets [76], the intention categories in this data set align with the unique characteristics of memes that can be interactive, expressive, purely entertaining, offensive, and others.

4) **Offensiveness Detection**: The offensive content within memes, encompass hate, racism, and misogyny, motivating the researchers [111], [112] into the task of offensive meme detection. The MET-meme data set has annotations for offensiveness, i.e., nonoffensive, slightly offensive, moderately offensive, and very offensive.

C. IMPLEMENTATION DETAILS

We have used ResNet50, BLIP2 image encoders to encode the image modality, and BERT, BLIP2 text encoders to encode the text modality. The reason for choosing the ResNet50 + BERT baseline is to compare with the MET-meme data set article's implementation with our knowledge enhancement module.

The BLIP-2 [51] model introduces an innovative vision-language pretraining approach that leverages pre-trained unimodal models, including a frozen image encoder and a frozen Large Language Model (LLM), while incorporating a novel component called the Querying Transformer (Q-Former). This lightweight transformer acts as a bridge between the fixed image encoder and the fixed LLM, using adaptable query vectors to extract relevant visual attributes aligned with textual guidance. The training process for the model comprises of two distinct phases: Vision-Language Representation Learning, aimed at aligning visual and textual representations by optimizing mutual information, and Vision-to-Language Generative Learning, where the Q-Former generates visual representations for the LLM to decode into precise text responses. This approach connects visual and textual data, facilitating the harmonization of representations and the creation of accurate feature vectors.

In this study, we use the BLIP2 feature extractor model that was pretrained on the COCO data set [113]. The embeddings produced by both the text and image encoders are of size $\mathbb{R}^{N*768}$ where N represents the batch size. In the case of the ResNet50 + BERT implementation, we use the output of the second last layer of ResNet50, which produces embeddings

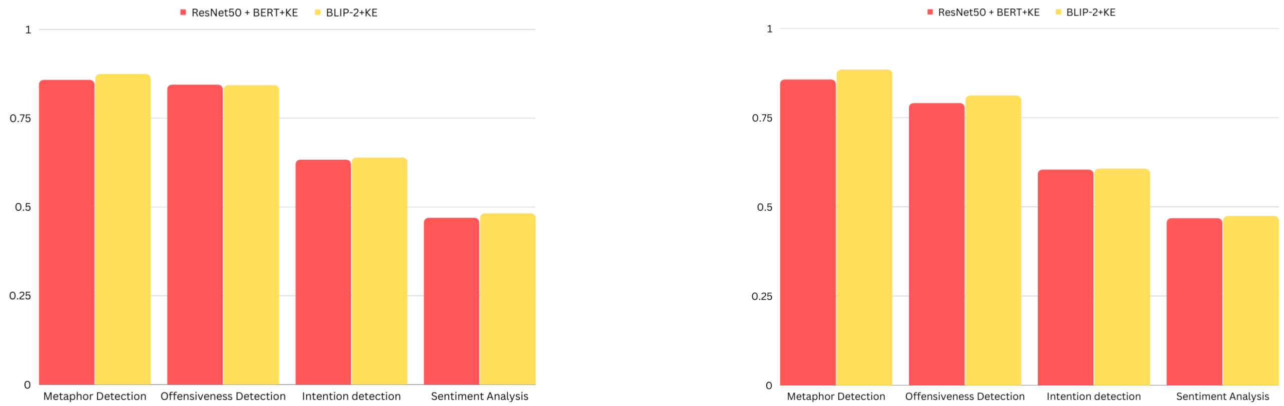


FIGURE 6. **The above figures correspond to the accuracy metrics of ResNet50 + BERT + KE and BLIP-2 + KE pipelines for the validation-set(left figure) and test-set(right figure).**

of size $\mathbb{R}^{N*2048}$. Likewise, the embeddings obtained from BERT are from the last hidden state and it is of size $\mathbb{R}^{N*768}$.

When implementing the Graph Attention Network, we use the Pytorch Geometric library [114]. The PhraseBERT model, which we use to vectorize the key phrases(extracted from the knowledge enhancement module), produces an embedding of size $p_k \epsilon \mathbb{R}^{768}$. Since this embedding is set as each node's attribute, the input feature for the Graph Attention Network is also 768. Our implementation consists of three layers of graph attention operators with a RELU activation function between them. In addition, there is also the multi-headed attention scheme(4 heads). This GAT model is trained for 100 epochs for each subtask, where each subtask is modeled as a graph classification task. In the case where no training is done, the learnable parameters and the attention weight matrix of the graph attention network are randomly initialized. This random initialization can lead to suboptimal convergence and low quality graph embeddings if the model starts with unfavorable weight values. To mitigate this, training of the GAT network is done before obtaining per graph embedding. The graph embeddings obtained is of size $g^k \epsilon \mathbb{R}^{128}$.

The above embeddings are then concatenated to obtain a final representation of size $\mathbb{R}^{N*1664}$ in the case of BLIP-2 [51] implementation and $\mathbb{R}^{N*2944}$ in the case of ResNet50 + BERT implementation. This is then passed onto a task-specific (depending on the number of class labels) classifier. The training of this classifier is done for 10 epochs using the Adam optimizer set with a learning rate of 0.005 for all the subtasks. The loss function used for the same is the Categorical Cross-entropy loss, as shown below:

$$\text{Categorical Cross-Entropy Loss} = -\sum_{i=1}^{N}\sum_{j=1}^{C} y_{ij}\log(p_{ij}) \tag{3}$$

We have used the Lavis [115] library to implement the BLIP2 image and text encoders. The ResNet50 + BERT baseline is implemented using the Timm [116] and transformers [117] library. The subsequent classifier has been implemented using the Pytorch library.

V. RESULTS AND DISCUSSIONS

A. OVERALL PERFORMANCE AND DISCUSSION

Table 1 presents an overview of the performance comparison in the four subtasks compared to the ResNet50 + BERT baseline [1]. Figure 6 compares the accuracy metrics of the ResNet50 + BERT + KE and the BLIP-2 + KE pipelines on validation and test set. Because this data set was very recently released, not much work has been carried out, so we cannot compare it with other benchmarks. The metric tracked here is the weighted F1 score as shown in Equation 6. The accuracy metric is also tracked across the four subtasks, which can be seen in Figure 6.

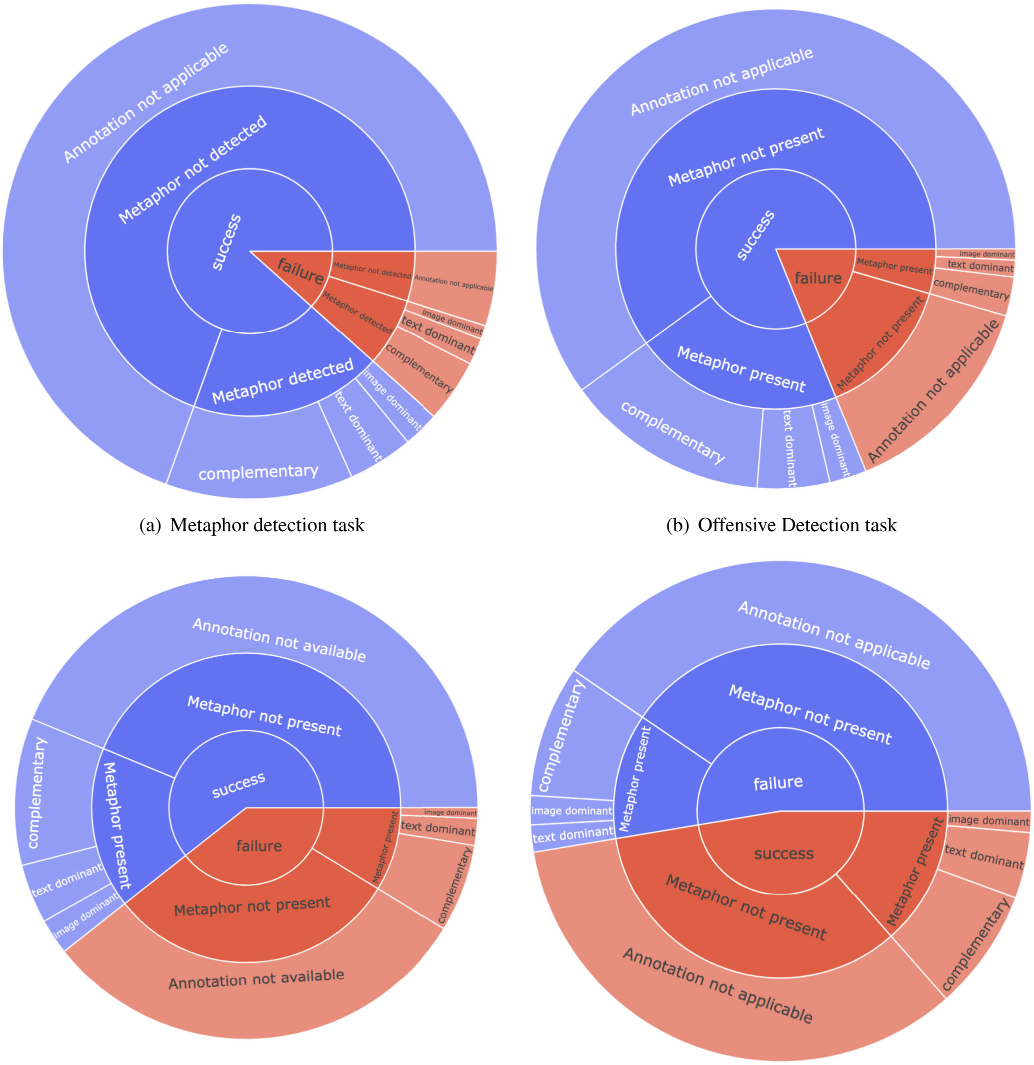
$$\text{Precision} = \frac{\sum_{i=1}^{N} w_i \cdot \text{TruePos}_i}{\sum_{i=1}^{N} w_i \cdot (\text{TruePos}_i + \text{FalsePos}_i)} \tag{4}$$

$$\text{Recall} = \frac{\sum_{i=1}^{N} w_i \cdot \text{TruePos}_i}{\sum_{i=1}^{N} w_i \cdot (\text{TruePos}_i + \text{FalseNeg}_i)} \tag{5}$$

$$\text{Weighted F1 Score} = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6}$$

As shown in Table 1, our approach using the BLIP2 models outperforms the baseline model in all subtasks in both the validation and the test set. The efficacy of our approach is further highlighted in the ResNet50 + BERT benchmark, which shows improvement in scores just with the inclusion of the knowledge enhancement module. This clearly shows that the scores have not been improved just because of more powerful SOTA image and text encoders but due to the inclusion of the knowledge enhancement module. This analysis is discussed further in the Ablation Study.

The sunburst charts in Figure 7 show a general trend, i.e., the knowledge enhancement module is able to capture the



(a) Metaphor detection task

(b) Offensive Detection task

(c) Intention Detection task

(d) Sentiment Analysis task

FIGURE 7. Sunburst chart showing the performance of BLIP-2 model with the knowledge enhancement module across the four tasks.

TABLE 1. Results across the four tasks using the baseline implementation, ResNet50 + BERT + KE module and BLIP-2 + KE module. Here KE stands for knowledge enhancement.

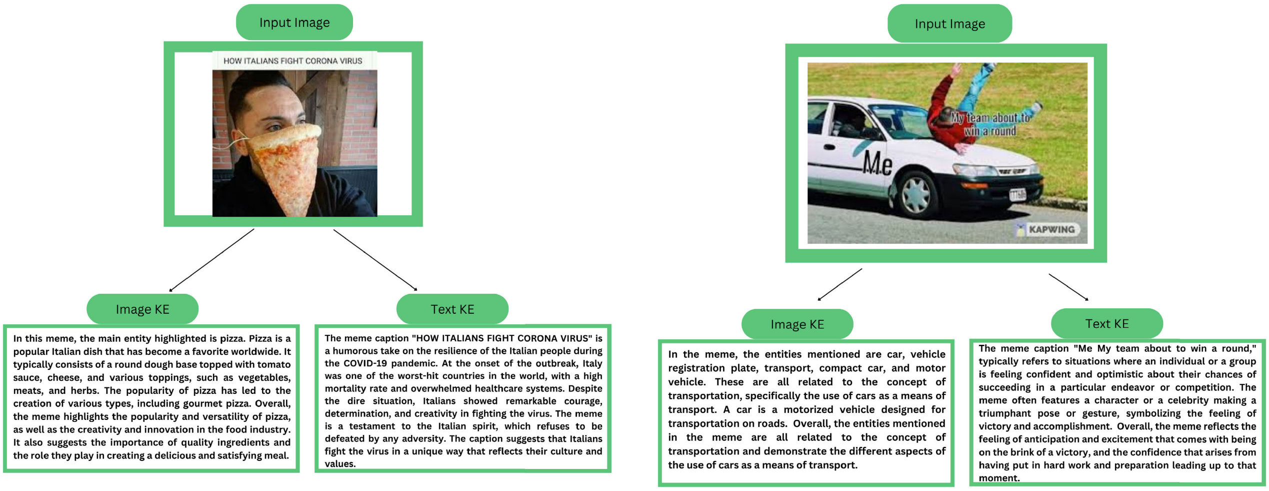
Model	Metaphor Understanding		Offensiveness detection		Intention detection		Sentiment analysis	
	Validation set	Test set	Validation set	Test set	Validation set	Test set	Validation set	Test set
ResNet50 + BERT (Baseline) [1]	0.7977	0.8239	0.6790	0.6839	0.3764	0.4165	0.2329	0.2768
ResNet50 + BERT + KE module	0.8579	0.8566	0.8167	0.7531	0.6192	0.5827	0.4406	**0.4386**
BLIP-2 + KE module	**0.871**	**0.881**	**0.8355**	**0.7938**	**0.631**	**0.5905**	**0.4439**	0.4356

metaphorical content across the subtasks without the explicit use of the metaphorical annotation.

The sunburst charts also show that memes having metaphors in the complementary setting are more in number with respect to the text-dominant setting and image-dominant setting. This shows the efficacy of our approach of using a Graph Attention Network to capture the common and modality-specific aspects of the concepts in the meme.

Recently, the Graph Transformers Network (GTN) [118], which is an extension of the transformer architecture in NLP into graphs, has been seen to beat the performance of GAT in terms of capturing the attention of a node compared to its neighbors, and has been recently used in a few GNN based architectures for diverse tasks, i.e., document classification [119], hateful discussion detection [120] and fake news detection [121]. One possibility is to use GTN to further improve the performance of the model proposed in this work.

It can also be seen that, across all the subtasks, the success cases with literal memes are more in number than metaphorical memes. This is a testament again to



In this meme, the main entity highlighted is pizza. Pizza is a popular Italian dish that has become a favorite worldwide. It typically consists of a round dough base topped with tomato sauce, cheese, and various toppings, such as vegetables, meats, and herbs. The popularity of pizza has led to the creation of various types, including gourmet pizza. Overall, the meme highlights the popularity and versatility of pizza, as well as the creativity and innovation in the food industry. It also suggests the importance of quality ingredients and the role they play in creating a delicious and satisfying meal.

The meme caption "HOW ITALIANS FIGHT CORONA VIRUS" is a humorous take on the resilience of the Italian people during the COVID-19 pandemic. At the onset of the outbreak, Italy was one of the worst-hit countries in the world, with a high mortality rate and overwhelmed healthcare systems. Despite the dire situation, Italians showed remarkable courage, determination, and creativity in fighting the virus. The meme is a testament to the Italian spirit, which refuses to be defeated by any adversity. The caption suggests that Italians fight the virus in a unique way that reflects their culture and values.

In the meme, the entities mentioned are car, vehicle registration plate, transport, compact car, and motor vehicle. These are all related to the concept of transportation, specifically the use of cars as a means of transport. A car is a motorized vehicle designed for transportation on roads. Overall, the entities mentioned in the meme are all related to the concept of transportation and demonstrate the different aspects of the use of cars as a means of transport.

The meme caption "Me My team about to win a round," typically refers to situations where an individual or a group is feeling confident and optimistic about their chances of succeeding in a particular endeavor or competition. The meme often features a character or a celebrity making a triumphant pose or gesture, symbolizing the feeling of victory and accomplishment. Overall, the meme reflects the feeling of anticipation and excitement that comes with being on the brink of a victory, and the confidence that arises from having put in hard work and preparation leading up to that moment.

FIGURE 8. The above two memes denote the failure cases where our model predicts incorrectly in the task of offensiveness detection for the meme to the left and in the task of intention detection for the meme to the right.

the knowledge enhancement module for having captured non-metaphorical information which subsequently helps in each of the subtasks.

However, it can be noted that the performance decreases with each subsequent task, i.e., from metaphor detection to sentiment analysis. We attribute this to the fact that, as we progress through these tasks, several factors come into play that challenge the overall performance of the model. First, a significant contributing factor is the increasing complexity of the tasks themselves. With each subsequent task, the number of distinct labels or categories expands. Second, these tasks often involve subtle differences in the definitions and interpretations of these labels. The task of discerning these nuanced variations can pose a formidable challenge. Third, our approach adopts a general querying strategy of ChatGPT for knowledge enhancement. While this strategy provides valuable information from a broad knowledge base, it does not tailor queries to the specific requirements of each downstream task. Task-specific querying could potentially yield more contextually relevant information that aligns with the subtleties of each task. Incorporating such task-specific querying mechanisms may enhance the ability of our model to excel in individual downstream tasks.

Although our research has made significant strides in understanding multimodal memes, it is important to acknowledge its limitations, particularly in cases involving metaphorical interpretations. This is highlighted in Figure 8. Our model failed in the task of offensiveness detection for the meme on the left side of this figure. As seen in the figure, the knowledge obtained from ChatGPT for the image and the text modality individually is relevant and factually correct. Although our model predicts it as non-offensive, it is evidently offensive due to its potential to perpetuate stereotypes. Despite our innovative approach to bridging the intermodal semantic gap, metaphors that heavily rely on visual cues and cultural nuances may present difficulties in model interpretation, as seen in this meme. One possible workaround for this can be to jointly query both the visual entities and the meme caption to ChatGPT instead of separately querying the visual entities and meme caption.

In the same Figure 8, our model fails in the task of intention detection for the meme on the right. It can be seen that the knowledge insights garnered here from the text modality are not relevant to the task at hand and are also factually incorrect. This discrepancy in prediction highlights the challenge of generating accurate modality-specific knowledge. While we leverage ChatGPT as a source of web knowledge to bridge the semantic gap in multimodal memes, it is clear that in this instance, ChatGPT did not provide the relevant information needed to understand the humor and satirical elements embedded in the meme. Modern memes often rely on cultural references, wordplay, and visual satire that may elude conventional knowledge sources.

B. ABLATION STUDY

To further assess the effectiveness of our approach, we conduct the following experiments:

1) EFFECTIVENESS OF CROSS ATTENTION BETWEEN THE MODAL SPECIFIC ENCODERS

In this analysis, as shown in Table 2, we check the performance of the model with the introduction of cross-attention between the modality-specific encoders. To do this, we experiment in the same pipeline as explained in Figure 2 with the addition of cross attention between the modality-specific encoders. The results in Table 2 indicate that the model performance remains mostly similar or marginally degrades with cross-attention. This observation can be rationalized by considering the concordant and discordant aspects of multimodal memes. In cases where the modalities exhibit concordance, cross-attention may not provide significant advantages and could introduce noise. On the contrary, in scenarios with discordant modalities, cross-attention might struggle to reconcile these disparities, potentially leading

TABLE 2. Results showing the effectiveness of cross attention between the model-specific encoders.

Task	Validation set without cross attention	Validation set with cross attention	Test set without cross attention	Test set with cross attention
Metaphor Understanding	**0.871**	0.847	**0.881**	0.845
Offensiveness detection	0.8355	**0.838**	**0.7938**	0.7811
Intention detection	**0.631**	0.6221	0.5905	**0.6002**
Sentiment category	**0.4439**	0.4434	**0.4356**	0.4267

TABLE 3. Results across the four tasks using only the BLIP-2 image and text encoders.

Task	Validation Set	Test Set
Metaphor Understanding	0.8664	0.8447
Offensiveness Detection	0.7443	0.6681
Intention Detection	0.4743	0.4472
Sentiment Analysis	0.3487	0.3534

TABLE 4. Results across the four tasks using only the knowledge enhancement module.

Task	Validation Set	Test Set
Metaphor Understanding	0.7341	0.7401
Offensiveness Detection	0.7256	0.6408
Intention Detection	0.4059	0.3695
Sentiment Analysis	0.2368	0.2383

to a decline in performance. This suggests that in certain cases, it is beneficial to allow the modalities to maintain their independence, as enforced by our results without cross-attention, to capture the nuanced characteristics of multimodal memes better. Since cross-attention does not bring in additional advantages when the knowledge-capturing mechanism is already in place, in the final configuration of the model, cross-attention has not been included.

2) EFFECTIVENESS OF KNOWLEDGE ENHANCEMENT MODULE

In this analysis, as shown in Table 3, we check the effectiveness of the knowledge enhancement module by experimenting with a pipeline where there are only modality-specific encoders. This configuration is not completely without knowledge, as BLIP2 pretrained models bring in significant zero-shot learning capability. We observe that the contribution of external web knowledge is much more significant in downstream tasks than in metaphor detection.

3) EFFECTIVENESS OF MODALITY SPECIFIC ENCODER

In this analysis, as shown in Table 4, we check the effectiveness of the modality-specific encoders, i.e., the text and image encoder, respectively, for the different tasks. To do so, we experiment in a pipeline with only the knowledge enhancement module and no modality-specific encoders. In essence, it is treated as a graph classification problem where we use the Graph Attention Network to classify each constructed graph.

The results indicate that a pipeline comprising the Knowledge Enhancement (KE) module surpasses the baseline results. This outcome underscores the importance of external knowledge integration, especially in enhancing the understanding of multimodal content by the model. Although the KE module proves beneficial, the results also reveal that a pipeline consisting solely of the KE module does not outperform configurations incorporating modality-specific encoders, individually or in combination. This is visible in the sentiment analysis downstream task, where content plays a relatively more important role. The key takeaway from this analysis is that the optimal approach involves a synergistic combination of modality-specific encoders and the knowledge enhancement module. This combination consistently delivers the best results across various tasks, highlighting its effectiveness and versatility.

VI. CONCLUSION

With multimodality becoming omnipresent on social media networks, multimodal memes have become an essential construct in figurative language with increasing usage of metaphors. Due to the multicontextual nature of such metaphorical memes, they present an interesting yet unsolved challenge in cognitive computing. The work described in this article has proposed and successfully demonstrates an innovative model for the concept projection process involved in understanding such multimodal metaphorical memes.

The correct interpretation of metaphorical memes may involve appreciating elements of humor and satire with joint consideration of textual and visual cues accentuated further by cultural nuances. Even though using ChatGPT as a source of external knowledge works for most cases, it may not always be accurate in scenarios involving cultural nuances with the complex multimodal interplay of humor and satire.

Future research can be taken up in two directions. First, to address the above shortcoming, a methodology capable of using external knowledge for resolving cultural nuances in a multimodal setup mixed with humor and satire can be investigated. Also, as discussed earlier, the presented methodology can be further improved upon by task-specific querying in ChatGPT and using a Graph Transformer Network in place of a Graph Attention Network. For example, a possible next step can be paraphrasing the intended meaning of metaphorical memes in an interpretable way, and that will be very impactful research in the domain of disinformation. Second, the proposed methodology can encourage many meaningful applications in the future. The same methodology can be suitably extended to other multimodal constructs of figurative language, such as sarcasm, irony, simile, and satire.

ACKNOWLEDGMENT

The authors would like to thank B. R. Om Amitesh for his contribution toward making of the flow chart and other diagrams and also would like to thank Dr. Sucheendra Palaniappan from The Systems Biology Institute, Japan, for providing them with research support.

REFERENCES

[1] B. Xu, T. Li, J. Zheng, M. Naseriparsa, Z. Zhao, H. Lin, and F. Xia, "MET-meme: A multimodal meme dataset rich in metaphors," in *Proc. 45th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2022, pp. 2887–2899.

[2] P. Group, "MIP: A method for identifying metaphorically used words in discourse," *Metaphor Symbol*, vol. 22, no. 1, pp. 1–39, Jan. 2007.

[3] Y. Wilks, "A preferential, pattern-seeking, semantics for natural language inference," *Artif. Intell.*, vol. 6, no. 1, pp. 53–74, 1975.

[4] G. Lakoff and M. Johnson, *Metaphors We Live By*. Chicago, IL, USA: Univ. Chicago, 1980.

[5] I. A. Richards, *The Philosophy of Rethoric*. Oxford Univ. Press, Dec. 1965.

[6] G. Philip, *Colouring Meaning: Collocation and Connotation in Figurative Language*. John Benjamins Publishing Company, Feb. 2011, pp. 1–248, doi: 10.1075/scl.43.

[7] M. Hameleers, T. E. Powell, T. G. L. A. Van Der Meer, and L. Bos, "A picture paints a thousand lies? The effects and mechanisms of multimodal disinformation and rebuttals disseminated via social media," *Political Commun.*, vol. 37, no. 2, pp. 281–301, Mar. 2020.

[8] Y. Li and Y. Xie, "Is a picture worth a thousand words? An empirical study of image content and social media engagement," *J. Marketing Res.*, vol. 57, no. 1, pp. 1–19, Feb. 2020.

[9] L. Parcalabescu, N. Trost, and A. Frank, "What is multimodality?" 2021, *arXiv:2103.06304*.

[10] B. Das, "Multi-contextual learning in disinformation research: A review of challenges, approaches, and opportunities," *Online Social Netw. Media*, vols. 34–35, May 2023, Art. no. 100247.

[11] S. Rai and S. Chakraverty, "A survey on computational metaphor processing," *ACM Comput. Surveys*, vol. 53, no. 2, pp. 1–37, Mar. 2021.

[12] M. Abulaish, A. Kamal, and M. J. Zaki, "A survey of figurative language and its computational detection in online social networks," *ACM Trans. Web*, vol. 14, no. 1, pp. 1–52, Feb. 2020.

[13] X. Tong, E. Shutova, and M. Lewis, "Recent advances in neural metaphor processing: A linguistic, cognitive and social perspective," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, 2021, pp. 4673–4686.

[14] S. Sharma, F. Alam, M. Shad Akhtar, D. Dimitrov, G. D. S. Martino, H. Firooz, A. Halevy, F. Silvestri, P. Nakov, and T. Chakraborty, "Detecting and understanding harmful memes: A survey," 2022, *arXiv:2205.04274*.

[15] L. Shang, Y. Zhang, Y. Zha, Y. Chen, C. Youn, and D. Wang, "AOMD: An analogy-aware approach to offensive meme detection on social media," *Inf. Process. Manag.*, vol. 58, no. 5, Sep. 2021, Art. no. 102664.

[16] T. Yue, R. Mao, H. Wang, Z. Hu, and E. Cambria, "KnowleNet: Knowledge fusion network for multimodal sarcasm detection," *Inf. Fusion*, vol. 100, Dec. 2023, Art. no. 101921.

[17] S. Sharma, A. Kulkarni, T. Suresh, H. Mathur, P. Nakov, M. S. Akhtar, and T. Chakraborty, "Characterizing the entities in harmful memes: Who is the hero, the villain, the victim?" 2023, *arXiv:2301.11219*.

[18] J. Chen and A. Zhang, "HGMF: Heterogeneous graph-based fusion for multimodal data with incompleteness," in *Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2020, pp. 1295–1305.

[19] Z. Li, J. Ma, X. Li, and X. Pan, "DFNM: Dynamic fusion network of intra- and inter-modalities for multimodal sentiment analysis," in *Proc. Int. Conf. Asian Lang. Process. (IALP)*, Dec. 2021, pp. 346–351.

[20] Z. Lin, B. Liang, Y. Long, Y. Dang, M. Yang, M. Zhang, and R. Xu, "Modeling intra- and inter-modal relations: Hierarchical graph contrastive learning for multimodal sentiment analysis," in *Proc. 29th Int. Conf. Comput. Linguistics*, 2022, pp. 7124–7135.

[21] P. Patwa, S. Ramamoorthy, N. Gunti, S. Mishra, S. Suryavardan, A. Reganti, A. Das, T. Chakraborty, A. Sheth, and A. Ekbal, "Findings of memotion 2: Sentiment and emotion analysis of memes," in *Proc. De-Factify, Workshop Multimodal Fact Checking Hate Speech Detection*. CEUR, 2022. [Online]. Available: https://ceur-ws.org/Vol-3199/paper21.pdf

[22] D. Lahat, T. Adali, and C. Jutten, "Multimodal data fusion: An overview of methods, challenges, and prospects," *Proc. IEEE*, vol. 103, no. 9, pp. 1449–1477, Sep. 2015.

[23] T. Baltrusaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019.

[24] R. R. Pranesh and A. Shekhar, "MemeSem:A Multi-modal framework for sentimental analysis of meme via transfer learning," in *Proc. 4th Lifelong Mach. Learn. Workshop ICML*. OpenReview.net, 2020. [Online]. Available: https://openreview.net/forum?id=Okmqu6xqXK

[25] A. Gandhi, K. Adhvaryu, S. Poria, E. Cambria, and A. Hussain, "Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions," *Inf. Fusion*, vol. 91, pp. 424–444, Mar. 2023.

[26] W. Zhang, G. Liu, Z. Li, and F. Zhu, "Hateful memes detection via complementary visual and linguistic networks," 2020, *arXiv:2012.04977*.

[27] B. Zhao, A. Zhang, B. Watson, G. Kearney, and I. Dale, "A review of vision-language models and their performance on the hateful memes challenge," 2023, *arXiv:2305.06159*.

[28] C. Song, N. Ning, Y. Zhang, and B. Wu, "A multimodal fake news detection model based on crossmodal attention residual and multichannel convolutional neural networks," *Inf. Process. Manag.*, vol. 58, no. 1, Jan. 2021, Art. no. 102437.

[29] T. Sachan, N. Pinnaparaju, M. Gupta, and V. Varma, "SCATE: Shared cross attention transformer encoders for multimodal fake news detection," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining*, Nov. 2021, pp. 399–406.

[30] Z. Jin, J. Cao, H. Guo, Y. Zhang, and J. Luo, "Multimodal fusion with recurrent neural networks for rumor detection on microblogs," in *Proc. 25th ACM Int. Conf. Multimedia*, Oct. 2017, pp. 795–816.

[31] R. Kumari and A. Ekbal, "AMFB: Attention based multimodal factorized bilinear pooling for multimodal fake news detection," *Exp. Syst. Appl.*, vol. 184, Dec. 2021, Art. no. 115412.

[32] S. Qian, J. Wang, J. Hu, Q. Fang, and C. Xu, "Hierarchical multi-modal contextual attention network for fake news detection," in *Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2021, pp. 153–162.

[33] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, and D. Testuggine, "The hateful memes challenge: Detecting hate speech in multimodal memes," 2020, *arXiv:2005.04790*.

[34] T. H. Afridi, A. Alam, M. N. Khan, J. Khan, and Y.-K. Lee, "A multimodal memes classification: A survey and open research issues," in *Proc. 3rd Int. Conf. Smart City Appl.* Cham, Switzerland: Springer, 2021, pp. 1451–1466.

[35] R. Jaiswal, U. P. Singh, and K. P. Singh, "Fake news detection using BERT-VGG19 multimodal variational autoencoder," in *Proc. IEEE 8th Uttar Pradesh Sect. Int. Conf. Electr., Electron. Comput. Eng. (UPCON)*, Nov. 2021, pp. 1–5.

[36] D. Khattar, J. S. Goud, M. Gupta, and V. Varma, "MVAE: Multimodal variational autoencoder for fake news detection," in *Proc. World Wide Web Conf.*, May 2019, pp. 2915–2921.

[37] L. Harold Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, "VisualBERT: A simple and performant baseline for vision and language," 2019, *arXiv:1908.03557*.

[38] J. Lu, D. Batra, D. Parikh, and S. Lee, "ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," 2019, *arXiv:1908.02265*.

[39] H. Tan and M. Bansal, "LXMERT: Learning cross-modality encoder representations from transformers," 2019, *arXiv:1908.07490*.

[40] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid, "VideoBERT: A joint model for video and language representation learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7463–7472.

[41] P. Xu, X. Zhu, and D. A. Clifton, "Multimodal learning with transformers: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 10, pp. 12113–12132, Oct. 2023.

[42] B. Kennedy, X. Jin, A. Mostafazadeh Davani, M. Dehghani, and X. Ren, "Contextualizing hate speech classifiers with post-hoc explanation," 2020, *arXiv:2005.02439*.

[43] M. S. Hee, R. K.-W. Lee, and W.-H. Chong, "On explaining multimodal hateful meme detection models," in *Proc. ACM Web Conf.*, Apr. 2022, pp. 3651–3655.

[44] A. Das, J. S. Wahi, and S. Li, "Detecting hate speech in multi-modal memes," 2020, *arXiv:2012.14891*.

[45] W. Dai, S. Cahyawijaya, Y. Bang, and P. Fung, "Weakly-supervised multi-task learning for multimodal affect recognition," 2021, *arXiv:2104.11560*.

[46] Y. Li, Z. Zhang, and H. Huang, "Enhance multimodal model performance with data augmentation: Facebook hateful meme challenge solution," 2021, *arXiv:2105.13132*.

[47] N. Muennighoff, "Vilio: State-of-the-art visio-linguistic models applied to hateful memes," 2020, *arXiv:2012.07788*.

[48] X. Zhong, "Classification of multimodal hate speech—The winning solution of hateful memes challenge," 2020, *arXiv:2012.01002*.

[49] A. Kiran, M. Shetty, S. Shukla, V. Kerenalli, and B. Das, "Getting around the semantics challenge in hateful memes," in *Computational Intelligence and Data Analytics*. Berlin, Germany: Springer, 2022, pp. 341–351.

[50] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, and J. Clark, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.

[51] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," 2023, *arXiv:2301.12597*.

[52] D. Dimitrov, B. Bin Ali, S. Shaar, F. Alam, F. Silvestri, H. Firooz, P. Nakov, and G. Da San Martino, "SemEval-2021 task 6: Detection of persuasion techniques in texts and images," 2021, *arXiv:2105.09284*.

[53] C. Sharma, D. Bhageria, W. Scott, S. Pykl, A. Das, T. Chakraborty, V. Pulabaigari, and B. Gamback, "SemEval-2020 task 8: Memotion analysis—The visuo-lingual metaphor!" 2020, *arXiv:2008.03781*.

[54] B. Kille, A. Lommatzsch, O. Ozgobek, M. Elahi, and D.-T. Dang-Nguyen, "News images in MediaEval 2021," in *Proc. MediaEval*, 2021, pp. 1–5.

[55] S. Rajesh, A. Krishnan, and B. Das, "An ensemble approach towards correlating articles and their corresponding images," in *Proc. Multimedia Eval. Workshop (MediaEval)*, in CEUR Workshop Proceedings. Bergen, Norway, Jan. 2023. [Online]. Available: https://ceur-ws.org/Vol-3583/paper43.pdf

[56] X. Zhou, J. Wu, and R. Zafarani, "Safe: Similarity-aware multi-modal fake news detection," in *Proc. Pacific–Asia Conf. Knowl. Discovery Data Mining*. Cham, Switzerland: Springer, 2020, pp. 354–367.

[57] J. Xue, Y. Wang, Y. Tian, Y. Li, L. Shi, and L. Wei, "Detecting fake news by exploring the consistency of multimodal data," *Inf. Process. Manag.*, vol. 58, no. 5, Sep. 2021, Art. no. 102610.

[58] S. Singhal, M. Dhawan, R. R. Shah, and P. Kumaraguru, "Inter-modality discordance for multimodal fake news detection," in *Proc. ACM Multimedia Asia*, Dec. 2021, pp. 1–7.

[59] A. Giachanou, G. Zhang, and P. Rosso, "Multimodal multi-image fake news detection," in *Proc. IEEE 7th Int. Conf. Data Sci. Adv. Analytics (DSAA)*, Oct. 2020, pp. 647–654.

[60] Y. Chen, D. Li, P. Zhang, J. Sui, Q. Lv, L. Tun, and L. Shang, "Cross-modal ambiguity learning for multimodal fake news detection," in *Proc. ACM Web Conf.*, Apr. 2022, pp. 2897–2905.

[61] R. Speer, J. Chin, and C. Havasi, "ConceptNet 5.5: An open multilingual graph of general knowledge," in *Proc. AAAI Conf. Artif. Intell.*, vol. 31, 2017, pp. 1–8.

[62] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *Int. J. Comput. Vis.*, vol. 123, no. 1, pp. 32–73, May 2017.

[63] L. Shang, C. Youn, Y. Zha, Y. Zhang, and D. Wang, "KnowMeme: A knowledge-enriched graph neural network solution to offensive meme detection," in *Proc. IEEE 17th Int. Conf. eScience (eScience)*, Sep. 2021, pp. 186–195.

[64] V. Kougia, S. Fetzel, T. Kirchmair, E. Cano, S. M. Baharlou, S. Sharifzadeh, and B. Roth, "Memegraphs: Linking memes to knowledge graphs," 2023, *arXiv:2305.18391*.

[65] D. Song, S. Ma, Z. Sun, S. Yang, and L. Liao, "KVL-BERT: Knowledge enhanced visual-and-linguistic BERT for visual commonsense reasoning," *Knowl.-Based Syst.*, vol. 230, Oct. 2021, Art. no. 107408.

[66] M. Lymperaiou and G. Stamou, "A survey on knowledge-enhanced multimodal learning," 2022, *arXiv:2211.12328*.

[67] J. Yang, H. Jin, R. Tang, X. Han, Q. Feng, H. Jiang, B. Yin, and X. Hu, "Harnessing the power of LLMs in practice: A survey on ChatGPT and beyond," 2023, *arXiv:2304.13712*.

[68] S. Ubani, S. Olcay Polat, and R. Nielsen, "ZeroShotDataAug: Generating and augmenting training data with ChatGPT," 2023, *arXiv:2304.14334*.

[69] H. Dai, Z. Liu, W. Liao, X. Huang, Y. Cao, Z. Wu, L. Zhao, S. Xu, W. Liu, N. Liu, S. Li, D. Zhu, H. Cai, L. Sun, Q. Li, D. Shen, T. Liu, and X. Li, "AugGPT: Leveraging ChatGPT for text data augmentation," 2023, *arXiv:2302.13007*.

[70] Y. Tan, D. Min, Y. Li, W. Li, N. Hu, Y. Chen, and G. Qi, "Can ChatGPT replace traditional KBQA models? An in-depth analysis of the question answering performance of the GPT LLM family," 2023, *arXiv:2303.07992*.

[71] Y. Shi, H. Ma, W. Zhong, Q. Tan, G. Mai, X. Li, T. Liu, and J. Huang, "ChatGraph: Interpretable text classification by converting ChatGPT knowledge to graphs," 2023, *arXiv:2305.03513*.

[72] M. Song, H. Jiang, S. Shi, S. Yao, S. Lu, Y. Feng, H. Liu, and L. Jing, "Is ChatGPT a good keyphrase generator? A preliminary study," 2023, *arXiv:2303.13001*.

[73] R. Martinez-Cruz, A. J. Lopez-Lopez, and J. Portela, "ChatGPT vs state-of-the-art models: A benchmarking study in keyphrase generation task," 2023, *arXiv:2304.14177*.

[74] Y. Elkhatib and K. Hill, "Memes to an end: A look into what makes a meme offensive," in *Proc. MISINFO Workshop Misinf. Integrity Social Netw.*, Apr. 2021. [Online]. Available: http://ceur-ws.org/Vol-2890/

[75] G. O. Diaz and V. Ng, "Unveiling hidden intentions," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, 2020, pp. 13550–13555.

[76] J. Kruk, J. Lubin, K. Sikka, X. Lin, D. Jurafsky, and A. Divakaran, "Integrating text and image: Determining multimodal document intent in Instagram posts," 2019, *arXiv:1904.09073*.

[77] S. Namitha, P. Sanjan, N. C. Reddy, Y. Srikar, H. Shanmugasundaram, and B. P. Andraju, "Sentiment analysis: Current state and future research perspectives," in *Proc. 7th Int. Conf. Intell. Comput. Control Syst. (ICICCS)*, May 2023, pp. 1115–1119.

[78] P. Sudhakaran, S. Hariharan, and J. Lu, "A framework investigating the online user reviews to measure the biasness for sentiment analysis," *Asian J. Inf. Technol.*, vol. 15, no. 12, pp. 1890–1898, 2016.

[79] D. S. Chauhan, S. Dhanush, A. Ekbal, and P. Bhattacharyya, "All-in-one: A deep attentive multi-task learning framework for humour, sarcasm, offensive, motivation, and sentiment on memes," in *Proc. 1st Conf. Asia–Pacific Chapter Assoc. Comput. Linguistics, 10th Int. Joint Conf. Natural Lang. Process.*, 2020, pp. 281–290.

[80] G. Mialon, D. Chen, A. d'Aspremont, and J. Mairal, "A trainable optimal transport embedding for feature aggregation and its relationship to attention," 2020, *arXiv:2006.12065*.

[81] Z. Xing, S. Zhao, W. Guo, F. Meng, X. Guo, S. Wang, and H. He, "Coal resources under carbon peak: Segmentation of massive laser point clouds for coal mining in underground dusty environments using integrated graph deep learning model," *Energy*, vol. 285, Dec. 2023, Art. no. 128771.

[82] Y. Yan, X. Wu, C. Li, Y. He, Z. Zhang, H. Li, A. Li, and L. Wang, "Topological EEG nonlinear dynamics analysis for emotion recognition," *IEEE Trans. Cognit. Develop. Syst.*, vol. 15, no. 2, pp. 625–638, Jun. 2023.

[83] Y. Zhou, F. Li, Y. Li, Y. Ji, G. Shi, W. Zheng, L. Zhang, Y. Chen, and R. Cheng, "Progressive graph convolution network for EEG emotion recognition," *Neurocomputing*, vol. 544, Aug. 2023, Art. no. 126262.

[84] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *Stat*, vol. 1050, no. 20, p. 48550, 2017.

[85] L. Huang, X. Sun, S. Li, L. Zhang, and H. Wang, "Syntax-aware graph attention network for aspect-level sentiment classification," in *Proc. 28th Int. Conf. Comput. Linguistics*, 2020, pp. 799–810.

[86] N. Jiang, J. Wen, J. Li, X. Liu, and D. Jin, "GATrust: A multi-aspect graph attention network model for trust assessment in OSNs," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 6, pp. 5865–5878, Jun. 2023.

[87] Z. Cheng, C. Yan, F.-X. Wu, and J. Wang, "Drug-target interaction prediction using multi-head self-attention and graph attention network," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 19, no. 4, pp. 2208–2218, Jul. 2022.

[88] J. Plepi and L. Flek, "Perceived and intended sarcasm detection with graph attention networks," 2021, *arXiv:2110.04001*.

[89] P. Kapadia, A. Saxena, B. Das, Y. Pei, and M. Pechenizkiy, "Co-attention based multi-contextual fake news detection," in *Complex Networks XIII*. Berlin, Germany: Springer, 2023, pp. 83–95.

[90] H. Sharma, M. Agrahari, S. K. Singh, M. Firoj, and R. K. Mishra, "Image captioning: A comprehensive survey," in *Proc. Int. Conf. Power Electron. IoT Appl. Renew. Energy Control (PARC)*, Feb. 2020, pp. 325–328.

[91] J. Wang, Z. Yang, X. Hu, L. Li, K. Lin, Z. Gan, Z. Liu, C. Liu, and L. Wang, "GIT: A generative image-to-text transformer for vision and language," 2022, *arXiv:2205.14100*.

[92] N. Nikzad-Khasmakhi, M.-R. Feizi-Derakhshi, M. Asgari-Chenaghlu, M.-A. Balafar, A.-R. Feizi-Derakhshi, T. Rahkar-Farshi, M. Ramezani, Z. Jahanbakhsh-Nagadeh, E. Zafarani-Moattar, and M. Ranjbar-Khadivi, "Phraseformer: Multimodal key-phrase extraction using transformer and graph embedding," 2021, *arXiv:2106.04939*.

[93] A. Kong, S. Zhao, H. Chen, Q. Li, Y. Qin, R. Sun, and X. Bai, "PromptRank: Unsupervised keyphrase extraction using prompt," 2023, *arXiv:2305.04490*.

[94] M. Kulkarni, D. Mahata, R. Arora, and R. Bhowmik, "Learning rich representation of keyphrases from text," 2021, *arXiv:2112.08547*.

[95] S. Wang, L. Thompson, and M. Iyyer, "Phrase-BERT: Improved phrase embeddings from BERT with an application to corpus exploration," 2021, *arXiv:2109.06304*.

[96] B. Perozzi, R. Al-Rfou, and S. Skiena, "DeepWalk: Online learning of social representations," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2014, pp. 701–710.

[97] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 855–864.

[98] A. Narayanan, M. Chandramohan, R. Venkatesan, L. Chen, Y. Liu, and S. Jaiswal, "Graph2vec: Learning distributed representations of graphs," 2017, *arXiv:1707.05005*.

[99] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[100] S. Suryawanshi, B. R. Chakravarthi, M. Arcan, and P. Buitelaar, "Multimodal meme dataset (MultiOff) for identifying offensive content in image and text," in *Proc. 2nd Workshop Trolling, Aggression Cyberbullying*, 2020, pp. 32–41.

[101] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, and D. Testuggine, "The hateful memes challenge: Detecting hate speech in multimodal memes," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 2611–2624.

[102] A. R. Akula, B. Driscoll, P. Narayana, S. Changpinyo, Z. Jia, S. Damle, G. Pruthi, S. Basu, L. Guibas, W. T. Freeman, Y. Li, and V. Jampani, "MetaCLUE: Towards comprehensive visual metaphors research," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 23201–23211.

[103] C. Liu, G. Geigle, R. Krebs, and I. Gurevych, "FigMemes: A dataset for figurative language identification in politically-opinionated memes," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2022, pp. 7069–7086.

[104] R. Yosef, Y. Bitton, and D. Shahaf, "IRFL: Image recognition of figurative language," 2023, *arXiv:2303.15445*.

[105] D. Zhang, M. Zhang, H. Zhang, L. Yang, and H. Lin, "MultiMET: A multimodal dataset for metaphor understanding," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics, 11th Int. Joint Conf. Natural Lang. Process.*, 2021, pp. 3214–3225.

[106] X. Zhou, Y. Yong, X. Fan, G. Ren, Y. Song, Y. Diao, L. Yang, and H. Lin, "Hate speech detection based on sentiment knowledge sharing," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics, 11th Int. Joint Conf. Natural Lang. Process.*, 2021, pp. 7158–7166.

[107] B. Geng, M. Yang, F. Yuan, S. Wang, X. Ao, and R. Xu, "Iterative network pruning with uncertainty regularization for lifelong sentiment classification," in *Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2021, pp. 1229–1238.

[108] L. Xu, H. Lin, Y. Pan, H. Ren, and J. Chen, "Constructing the affective lexicon ontology," *J. China Soc. Sci. Tech. Inf.*, vol. 27, no. 2, pp. 180–185, 2008.

[109] E. Dresner and S. C. Herring, "Functions of the nonverbal in CMC: Emoticons and illocutionary force," *Commun. Theory*, vol. 20, no. 3, pp. 249–268, Jul. 2010.

[110] C. Tauch and E. Kanjo, "The roles of emojis in mobile phone notifications," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput., Adjunct*, Sep. 2016, pp. 1560–1565.

[111] H. Rose Kirk, Y. Jun, P. Rauba, G. Wachtel, R. Li, X. Bai, N. Broestl, M. Doff-Sotta, A. Shtedritski, and Y. M. Asano, "Memes in the wild: Assessing the generalizability of the hateful memes challenge dataset," 2021, *arXiv:2107.04313*.

[112] C. Lou, B. Liang, L. Gui, Y. He, Y. Dang, and R. Xu, "Affective dependency graph for sarcasm detection," in *Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2021, pp. 1844–1849.

[113] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Computer Vision—ECCV*. Zurich, Switzerland: Springer, 2014, pp. 740–755.

[114] M. Fey and J. Eric Lenssen, "Fast graph representation learning with PyTorch geometric," 2019, *arXiv:1903.02428*.

[115] D. Li, J. Li, H. Le, G. Wang, S. Savarese, and S. C. H. Hoi, "LAVIS: A one-stop library for language-vision intelligence," in *Proc. 61st Annu. Meeting Assoc. Comput. Linguistics*, 2023, pp. 31–41.

[116] R. Wightman. (2019). *PyTorch Image Models*. [Online]. Available: https://github.com/rwightman/pytorch-image-models

[117] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and J. Davison, "Transformers: State-of-the-art natural language processing," in *Proc. Conf. Empirical Methods Natural Lang. Process., Syst. Demonstrations*, 2020, pp. 38–45.

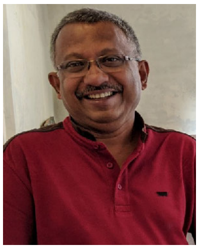
[118] E. Min, R. Chen, Y. Bian, T. Xu, K. Zhao, W. Huang, P. Zhao, J. Huang, S. Ananiadou, and Y. Rong, "Transformer for graphs: An overview from architecture perspective," 2022, *arXiv:2202.08455*.

[119] H. Zhang and J. Zhang, "Text graph transformer for document classification," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2020, pp. 1–6.

[120] L. Hebert, L. Golab, and R. Cohen, "Predicting hateful discussions on Reddit using graph transformer networks and communal context," in *Proc. IEEE/WIC/ACM Int. Joint Conf. Web Intell. Intell. Agent Technol. (WI-IAT)*, Nov. 2022, pp. 9–17.

[121] H. Matsumoto, S. Yoshida, and M. Muneyasu, "Propagation-based fake news detection using graph neural networks with transformer," in *Proc. IEEE 10th Global Conf. Consum. Electron. (GCCE)*, Oct. 2021, pp. 19–20.

SATHWIK ACHARYA received the bachelor's degree in computer science and engineering from PES University, Bengaluru, Karnataka, India, in 2023. He is currently a Systems Engineer with Hewlett Packard Enterprises. His leisure time is mainly dedicated to his research ventures in the frontiers of computational linguistics, multimodal deep learning, and systems biology.



BHASKARJYOTI DAS received the Graduate degree in electronics and telecommunication engineering from Jadavpur University, Kolkata, and the M.Tech. degree in computer science and engineering from Visvesvaraya Technological University (VTU), Belagavi. He is currently an Adjunct Professor with the Department of Computer Science and Engineering in AI & ML, PES University, Bengaluru, Karnataka, India. Besides being an Adjunct Professor, he is also a Research Scholar in CSE. He has more than 25 years of experience in the industry. He was an Engineer and in engineering management roles in diverse organizations, such as ISRO, NCR Corporation, Verifone Inc., Sun Microsystems, and Yahoo! His research interest includes machine learning with sparse labeled data for multicontextual problems.

T. S. B. SUDARSHAN (Senior Member, IEEE) received the Ph.D. degree in computer science and engineering from the Birla Institute of Technology and Science, Pilani, in 2007. He is currently the Dean of Research with PES University, Bengaluru, Karnataka, India, and a Professor with the Department of Computer Science and Engineering. He is also the Founder/Chairperson of IEEE RAS Bangalore Chapter.

• • •