Hierarchical Semantic Enhancement Network for Multimodal Fake News Detection

Qiang Zhang
University of Science and Technology of China
Hefei, Anhui, China
zq_126@mail.ustc.edu.cn

Jiawei Liu[*]
University of Science and Technology of China
Hefei, Anhui, China
jwliu6@ustc.edu.cn

Fanrui Zhang
University of Science and Technology of China
Hefei, Anhui, China
zfr888@mail.ustc.edu.cn

Jingyi Xie
University of Science and Technology of China
Hefei, Anhui, China
hsfzxjy@mail.ustc.edu.cn

Zheng-Jun Zha
University of Science and Technology of China
Hefei, Anhui, China
zhazj@ustc.edu.cn

ABSTRACT

The explosion of multimodal fake news content on social media has sparked widespread concern. Existing multimodal fake news detection methods have made significant contributions to the development of this field, but fail to adequately exploit the potential semantic information of images and ignore the noise embedded in news entities, which severely limits the performance of the models. In this paper, we propose a novel Hierarchical Semantic Enhancement Network (HSEN) for multimodal fake news detection by learning text-related image semantic and precise news high-order knowledge semantic information. Specifically, to complement the image semantic information, HSEN utilizes textual entities as the prompt subject vocabulary and applies reinforcement learning to discover the optimal prompt format for generating image captions specific to the corresponding textual entities, which contain multi-level cross-modal correlation information. Moreover, HSEN extracts visual and textual entities from image and text, and identifies additional visual entities from image captions to extend image semantic knowledge. Based on that, HSEN exploits an adaptive hard attention mechanism to automatically select strongly related news entities and remove irrelevant noise entities to obtain precise high-order knowledge semantic information, while generating attention mask for guiding cross-modal knowledge interaction. Extensive experiments show that our method outperforms state-of-the-art methods.

CCS CONCEPTS

• **Information systems** → **Social networks**; **Multimedia information systems**.

KEYWORDS

Fake news detection, Semantic information, Multimodal, Entity

ACM Reference Format:
Qiang Zhang, Jiawei Liu, Fanrui Zhang, Jingyi Xie, and Zheng-Jun Zha. 2023. Hierarchical Semantic Enhancement Network for Multimodal Fake News Detection. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29–November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3581783.3612423

1 INTRODUCTION

As social media platforms have become more and more embedded in people's lives, they have become the main source of access to information for public. Unfortunately, this has been accompanied by an 'explosion' of fake news [16, 25, 30]. Due to the confusing content of fake news, people are often misled by it, which in turn influences their judgement and decisions. It can also be used to distort and falsify facts to guide public opinion, which has a detrimental effect on social trust and stability [20, 26, 39]. Therefore, in order to stop the surge of fake news, there is an urgent need for automatic detection methods to identify fake news and enhance the trustworthiness of the social media ecosystem.

Fake news detection is a binary classification problem whose goal is to analyse news content so as to determine its authenticity. Traditional fake news detection focuses on textual content and relies on extracting semantic feature from the text, social media communication processes and user interactions to detect fake news [24, 45]. However, as multimedia technologies continue to evolve, rumour mongers are increasingly using multimodal content, such as attractive images, to get the attention of the public in order to facilitate faster dissemination. Therefore, the field of multimodal fake news detection is receiving more and more attention.

Some progress has been made in the field of multimodal fake news detection, yet existing methods [34, 41, 42] make insufficient use of image information. Some approaches simply extract image feature by VGG19 [31] or ResNet50 [13] pre-trained in ImageNet [6]. Some other methods combine frequency domain information as a complement to image features [46], or apply multimodal variational autoencoder to reconstruct textual and visual representations for reducing modality gap [18]. However, none of these methods make

[*] Corresponding author

MM '23, October 29–November 3, 2023, Ottawa, ON, Canada.

ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3612423

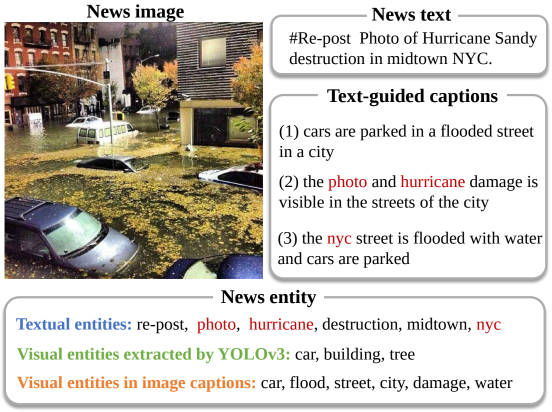
News image
News text
#Re-post Photo of Hurricane Sandy destruction in midtown NYC.
Text-guided captions
(1) cars are parked in a flooded street in a city
(2) the photo and hurricane damage is visible in the streets of the city
(3) the nyc street is flooded with water and cars are parked
News entity
Textual entities: re-post, photo, hurricane, destruction, midtown, nyc
Visual entities extracted by YOLOv3: car, building, tree
Visual entities in image captions: car, flood, street, city, damage, water

Figure 1: A multimodal news example, marked in red refers to textual entities serving as prompt subject vocabulary. This example shows that it is effective to apply textual information to guide image caption generation, while the extracted noisy entities should be filtered.

full use of the semantic information of the image, especially without combining text information for semantic extraction of image specific content. In addition, the simple extraction of image feature is not effective in reducing the modality gap between image and text features, which is not conducive to subsequent multimodal fusion. Apart from the basic features of news content, knowledge-level feature of news entities are also essential for predicting the reality of a sample. Knowledge graph (KG) consist of entities as graph nodes and relationships as edges with different types, which are rich in background knowledge information. Therefore, in order to improve the performance of fake news detection, a few of approaches use high-order knowledge semantic information of news entities as a source of objective evidence by incorporating it within the model [9, 38]. To acquire visual entities, these methods just utilize YOLOv3 [29] or Faster R-CNN [11] to detect the image [22, 36], which is not sufficient for the detection of visual entities. As shown in Figure 1, we can observe that the visual entities detected by YOLOv3 are scarce, while the entities contained in the image captions are often richer. Afterwards, these methods perform fake news detection by aligning and fusing visual and textual entities [22], or discovering the inconsistent semantic feature at the knowledge level [36]. But these methods tend to ignore the additional noise impact that comes with adding external knowledge. When extracting visual or textual entities, irrelevant entities are often identified, which tend to introduce varying degrees of noise information into the model. For instance, in Figure 1, the textual entities '*re-post*' and '*photo*' are not necessary for fake news detection, same for the visual entity '*tree*'. Therefore, removing the influence of the above mentioned irrelevant entities and obtaining more accurate high-order knowledge feature of the news is undoubtedly effective for fake news detection.

In this work, we propose a novel Hierarchical Semantic Enhancement Network (HSEN) for multimodal fake news detection by learning text-related image semantic and precise news high-order knowledge semantic information. It utilizes two modules to enhance news semantic features at different levels, namely the image semantic enhancement module and the knowledge semantic enhancement module. Specifically, for image semantic enhancement module, we identify textual entities as the prompt subject vocabulary to guide the BLIP [21] model for generating the image captions specific to the textual entities as the supplementary image semantic information. As shown in Figure 1, on the right are three types of image captions generated by our customised prompt format, where the first is the original caption without the addition of the prompt, which represents the global semantic information of the image. The second is generated by two textual entities as the prompt subject vocabulary, which represent the semantic interaction information between the two textual entities in the image. The third is generated by a single textual entity, which represents the local semantic information of the individual textual entity. We can observe that using textual entities to guide the generation of image captions could effectively combine the text and image semantic to extract cross-modal correlation information and eliminate the modality gap. Since the generation of image caption by the BLIP model is an autoregressive process and its gradient cannot be calculated, we apply reinforcement learning to discover the optimal prompt format to generate more valuable image captions.

For the knowledge semantic enhancement module, we identifies additional visual entities from image captions, which in turn effectively supplement the high-order knowledge information of the image. And we perform adaptive hard attention operation on visual and textual entity embedding to remove irrelevant noise entities and assign critical weights to relevant entities. Besides, we generate attention mask for the cross-modal knowledge interaction of visual and textual entity embedding based on the above hard attention operation. Finally, we perform multimodal fusion of image caption feature, text feature, image feature, and combine the filtered external knowledge feature to determine whether the news is real or fake. Extensive experiments on two datasets show that our multimodal fake news detection method HSEN outperforms state-of-the-art methods.

The main contributions of this paper are as follows:

- We propose a novel Hierarchical Semantic Enhancement Network for multimodal fake news detection by learning text-related image semantic and precise news high-order knowledge semantic information.
- We utilize textual entities as the prompt subject vocabulary and apply reinforcement learning to discover the optimal prompt format to guide the BLIP model in generating image captions specific to the corresponding textual entities.
- We extract additional visual entities from image captions to enrich the image knowledge information. Besides, we apply an adaptive hard attention mechanism to filter irrelevant knowledge and enhance relevant knowledge.

2 RELATED WORK

2.1 Text-based Fake News Detection

Text-based fake news detection models rely on extracting textual semantic features from the news textual content to determine the authenticity of fake news [10, 24, 45]. For example, Qian *et al.* [28] extracted multi-level text features from textual content and designed a generation module to produce user responses to aid fake news detection. Liu *et al.* [23] used a combination of recurrent networks and convolutional networks to capture global and local changes in user usage features along the news dissemination path. Yang *et al.* [48] constructed a heterogeneous information network to extract rich information between users, posts and user comments, and used an adversarial learning framework to improve model robustness. Khoo *et al.* [19] proposed a post-level attention model (PLAN), which used multi-head attention mechanism to model long-distance interactions between tweets for fake news detection.

2.2 Multimodal Fake News Detection

In recent years, multimodal fake news detection has received extensive attention as the content form of news often contains multimodal information such as image and text. For example, Zhou *et al.* [54] performed fake news detection by extracting text feature and image feature and comparing the cosine similarity between them. Wang *et al.* [41] used a multi-task learning framework to simultaneously perform fake news detection and event classification tasks, in which the event classification task learns event invariant features that contribute to fake news detection. Chen *et al.* [5] transformed heterogeneous unimodal features into a common feature space by contrast learning and quantified the ambiguity between text and image feature by KL divergence to adaptively aggregate unimodal features and multimodal features. Wang *et al.* [44] jointly modeled the text feature, image feature and knowledge concepts through the graph convolutional network to obtain global semantic representation. Sun *et al.* [36] performed fake news detection by capturing the inconsistent semantics at the cross-modal level and the content-knowledge level in a unified framework. However, these methods only focus on how to exploit the information at the feature level and knowledge level of multimodal content such as text and images, ignoring the extensive semantic information embedded in the images and the external noise impact introduced when adding external knowledge, which limits the performance of fake news detection systems.

2.3 Image Caption

Image caption is designed to generate textual descriptions of image and their methods are mainly based on the encoder-decoder architecture. The workflow is that the image content is input to a image encoder, after which the image feature is fed into a text decoder to generate image caption [1, 47]. Traditional text decoder typically applies LSTM network [14, 35], while more recent methods commonly utilize attention-based model to generate image captions [52, 43]. In controlled image caption generation, Chen *et al.* [4] proposed an abstract scene graphs(ASG) to fine-grainedly represent user intent and control image caption generation. Wang *et al.* [40] implemented prompt learning by embedding it into the image caption model to enable the transformation of multiple image caption styles. In terms of multimodal pre-training models, such as BLIP [21] and CoCa [50] have shown excellent performance on image caption benchmark, which can be used as backbone model to extract rich semantic information from images and generate object-specific image captions by the way of prompt, which is beneficial for fake news detection.

3 METHOD

3.1 Model Overview

In this work, we propose a novel Hierarchical Semantic Enhancement Network (HSEN) for multimodal fake news detection, whose architecture is shown in Figure 2. Specifically, HSEN consists of four modules for fusing multimodal information and performing fake news detection. (1) The image semantic enhancement module, which utilizes textual entities as the prompt subject vocabulary and applies reinforcement learning to find the optimal prompt format to generate multi-level image captions. (2) The multimodal fusion module, which encodes three types of news information: text, image and image captions, and performs fusion and enhancement of the three features. (3) The knowledge semantic enhancement module, which enlarges the image semantic knowledge and filters irrelevant knowledge and enhances relevant knowledge through the adaptive hard attention mechanism. (4) The model optimization module, which applies the multimodal representation information obtained from the above steps to perform binary classification. Next, we will describe the above modules in details.

3.2 Image Semantic Enhancement

Text-guided Image Caption. We apply the multimodal pre-training model BLIP as the base model for our image caption, which supports the use of image captions with prompt. We first utilize the entity linking tool TAGME to recognise the set of textual entities $\{\boldsymbol{E}_T^{\mathrm{i}}\}$ from news text $\boldsymbol{T}$. Then, we employ the textual entities as the prompt subject vocabulary and design the prompt format for guiding the BLIP model to generate three types image captions in total. The first is the original image caption generated by the BLIP model. The second is generated based on two textual entities as the subject vocabulary of the prompt. The prompt format is $P_c = \{P\}_1\{P\}_2...\{P\}_N\ \{textual\ entity\}\ and\{textual\ entity\}$, where $\{P\}_i$ represents the learnable prompt token and N represents the length of it; $\{textual\ entity\}$ represents the entity extracted from the news text, and the third is the local semantic information generated based on a single textual entity with the prompt format of $P_l = \{P\}_1\{P\}_2...\{P\}_N\ \{textual\ entity\}$. Because news text usually contain more entities and information than image, some textual entities may not have corresponding visual entities. Therefore, for the textual entities in the above prompts, we only utilize entities which are the same as the visual entities, and if there are not enough identical entities, they will be randomly selected from the remaining textual entities.

RLprompt. Due to the generation process of image caption is the autoregressive process, its gradient cannot be backward. Inspired by the method [7], we decided to apply the reinforcement learning for the training of learnable prompt tokens, the structure of which is illustrated in Figure 3. In which distilGPT-2 model is used as a continuous policy network to find the best learnable prompt tokens z^* from the vocabulary. Afterwards, we combine the learnable prompt tokens z with the textual entities as the complete prompt to guide the BLIP model for generating the image captions specific to the corresponding textual entities. At the same time, we calculate the reward based on the prediction results of the fake news detection model and apply soft Q-learning [12] to optimise the policy network. The details of this are described below:

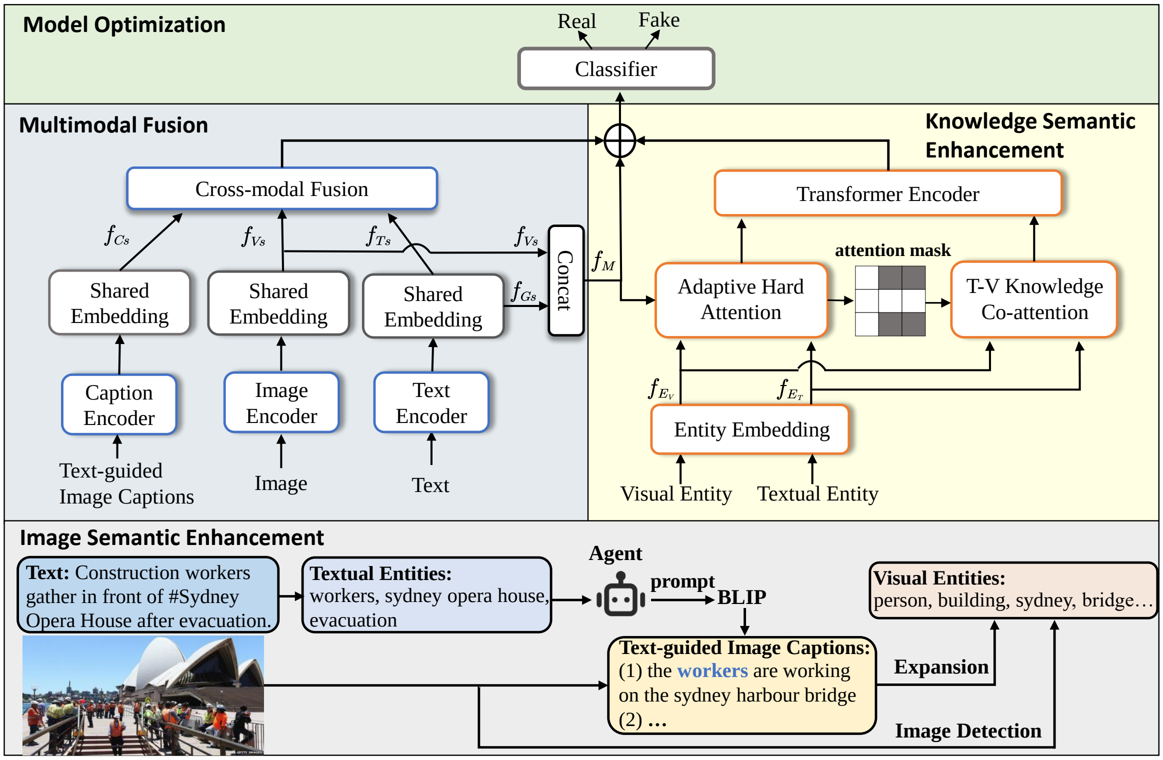
Model Optimization
Real
Fake
Classifier
Multimodal Fusion
Cross-modal Fusion
f_{Cs}
f_{Vs}
f_{Ts}
f_{Vs}
f_{Gs}
Shared Embedding
Shared Embedding
Shared Embedding
Concat
Caption Encoder
Image Encoder
Text Encoder
Text-guided Image Captions
Image
Text
Knowledge Semantic Enhancement
Transformer Encoder
f_M
attention mask
Adaptive Hard Attention
T-V Knowledge Co-attention
f_{E_V}
f_{E_T}
Entity Embedding
Visual Entity
Textual Entity
Image Semantic Enhancement
Text: Construction workers gather in front of #Sydney Opera House after evacuation.
Textual Entities: workers, sydney opera house, evacuation
Agent
prompt
BLIP
Visual Entities: person, building, sydney, bridge…
Text-guided Image Captions: (1) the workers are working on the sydney harbour bridge (2) …
Expansion
Image Detection

Figure 2: The overall architecture of the proposed HSEN. It contains four components: a image semantic enhancement module, a multimodal fusion module, a knowledge semantic enhancement module and a model optimization module.

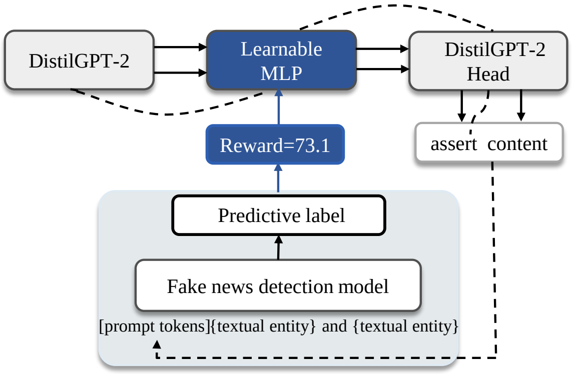

Figure 3: The architecture of the RLprompt.

We use the distilGPT-2 model as the policy network to explore the prompt space. The network trains only small MLP on the frozen distilGPT-2 model. Figure 3 illustrates the policy distilGPT-2's architecture. Specifically, we use distilGPT-2 to extract contextual embedding of the previous prompt tokens $z_{<t}$, and apply the added MLP layer to compute the adapted embedding, then pass the output to the original distilGPT-2 header to obtain the next prompt token probability. During training, we compute the MLP gradient by policy network backward. In the inference phase, we greedily select tokens to generate deterministic prompt tokens based on the trained policy.

We design a segmented reward function that encourages the generation of the optimal image caption by prompting, which in turn correctly classifies each fake news sample. And we calculate the reward by computing the gap between the probability of the correct label and the incorrect label. In this case, we first apply $P_z\left(y\right)$ to denote the probability of the correct label y predicted by the model, which is calculated as:

$$P_z(y) = \begin{cases} P_z(y \mid z, x) & \text{if } y = 1 \\ 1 - P_z(y \mid z, x) & \text{if } y = 0 \end{cases} \tag{1}$$

where z is the given prompt tokens, x is the training data and y is the true label. $P_z(y \mid z, x)$ represents the value of the final sigmoid output of the fake news detection model.

Then we denote the gap between the probability of the correct label and the incorrect label as $Gap_z\left(y\right) = P_z\left(y\right) - \left(1 - P_z\left(y\right)\right)$. The gap value is positive when the prediction is correct and negative otherwise. For correct predictions, we will multiply by a large number to indicate its desirability. The resulting reward function is thus as follows:

$$R\left(x, z, y\right) \;=\; \lambda_1^{1-y} \lambda_2^{y} Gap_z\left(y\right) \tag{2}$$

We also normalize the rewards by using input-specific means and standard deviations. Specifically, during prompt optimization, we sample multiple groups prompt tokens $z\left(x\right)$ for the input training data x, and compute the reward for each group of prompt tokens

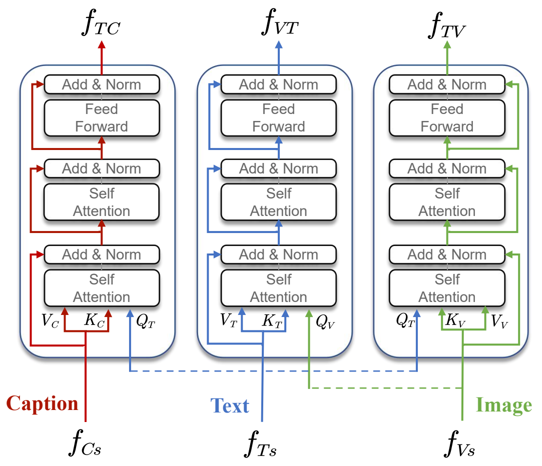

Figure 4: The architecture of the cross-modal fusion.

$z \in z(x)$. Afterwards, we compute the reward $z - score$ for the entire prompt tokens $z(x)$. The equation is as follows:

$$z - score(x, z, y) = \frac{R(x, z, y) - \text{mean}_{\mathbf{z} \in z(\mathbf{x})} R(x, z, y)}{\text{stdev}_{\mathbf{z} \in z(\mathbf{x})} R(x, z, y)} \tag{3}$$

3.3 Multimodal Fusion

Textual and Visual Encoding. This module consists of text encoder, image encoder and image caption encoder, which extracts basic features from the text $\boldsymbol{T}$, the image $\boldsymbol{I}$ and the generated image captions $\boldsymbol{C}$.

Textual Encoder. We utilize the pre-trained BERT [8] to map the sequence of text words of length L to an embedding sequence $\{\boldsymbol{x}_T^i\}_{i=1}^L$ of dimension 768. This embedding sequence is then processed by a bi-directional long and short-term memory network (Bi-LSTM) [15], which is further transformed by a linear layer to obtain the text features $\{f_T^i\}_{i=1}^L \in R^d$, with the following equation:

$$\{f_T^i\} = \boldsymbol{W}_T(Bi - LSTM(\boldsymbol{x}_T^i)) + \boldsymbol{b}_T \tag{4}$$

where $\boldsymbol{W}_T$ and $\boldsymbol{b}_T$ denote the learnable parameters of the fully connected (FC) layer.

We also take the output of the last layer of the Bi-LSTM forward and backward network as the global text feature f_G with the following equation:

$$f_G = \boldsymbol{W}_T(\overrightarrow{Bi - LSTM}(\boldsymbol{x}_t^i), \overleftarrow{Bi - LSTM}(\boldsymbol{x}_t^i)) + \boldsymbol{b}_T \tag{5}$$

where $\overrightarrow{Bi - LSTM}(\boldsymbol{x}_t^i)$ represents the output of the last layer of the forward network for Bi-LSTM and $\overleftarrow{Bi - LSTM}(\boldsymbol{x}_t^i)$ represents the output of the last layer of the backward network for Bi-LSTM.

Visual Encoder. For image $\boldsymbol{V}$, we apply the visual feature encoder of the BLIP model to generate image feature f_V. For the generated image captions $\boldsymbol{C}$, we utilize the textual feature encoder of the BLIP model to transform the image captions into a sequence of caption features $\{f_C^i\}$. The BLIP model will be completely frozen during the training process.

Cross-modal Fusion. To reduce the modality gap between image feature f_V, image caption feature f_C and text feature f_T and f_G, we first map them to the same feature space using the linear layer with shared weight, as described below:

$$\begin{aligned} f_{T_s} &= W_{\text{shared}} f_T \\ f_{V_s} &= W_{\text{shared}} f_V \\ f_{C_s} &= W_{\text{shared}} f_C \\ f_{G_s} &= W_{\text{shared}} f_G \end{aligned} \tag{6}$$

In multimodal news, the content of image and important news elements in the text are often related in some way. Therefore, we utilize the co-attention mechanism to capture the mutual information between text, image and image caption features and obtain enhanced cross-modal features, the architecture of which is shown in Figure 4. For instance, for the image enhanced text feature, we apply $\boldsymbol{Q}_V = f_{Vs}\boldsymbol{W}^Q$, $\boldsymbol{K}_T = f_{Ts}\boldsymbol{W}^K$ and $\boldsymbol{V}_T = f_{Ts}\boldsymbol{W}^V$ to compute their query matrix $\boldsymbol{Q}_V$, key matrix $\boldsymbol{K}_T$ and value matrix $\boldsymbol{V}_T$ respectively, where $\boldsymbol{W}^Q, \boldsymbol{W}^K, \boldsymbol{W}^V \in \mathbb{R}^{d \times \frac{d}{H}}$ is a linear transformation and H is the number of heads. Thus, we generate the image enhanced text feature by the following formula:

$$f_{VT} = \left(\Big\Vert_{h=1}^{H} softmax\left(\frac{\boldsymbol{Q}_V \boldsymbol{K}_T^T}{\sqrt{d}} \right) \boldsymbol{V}_T \right) \boldsymbol{W}_{VT} \tag{7}$$

where h denotes the h-th head and $W_{VT} \in R^{d \times d}$ represents the output linear transformation.

Thus, we can obtain the image enhanced text feature f_{VT}. Similarly, we can apply the text feature f_{Ts} to do separate feature enhancement operations on the image feature f_{Vs} and image caption feature f_{Cs}. So we obtain the text enhanced image feature f_{TV} and the text enhanced image caption feature f_{TC}. To further construct the internal connections, we model the above features separately using the self-attention sublayer and finally used the FC layer for the output of the cross-modal features. Eventually, we obtained the final outputs f_{VT}, f_{TV}, f_{TC} for the three features.

3.4 Knowledge Semantic Enhancement

Visual Entity Extraction. For the visual entity, we utilize the API from the Baidu OpenAI platform to recognise objects and celebrities from images. In order to increase the richness of the visual semantic knowledge, we extract additional visual entities from the image captions generated under the initial custom prompt, thus forming the set of visual entity $\{\boldsymbol{E}_V^{\text{i}}\}$.

Entity Embedding. To this point, we have obtained the set of textual entities $\{\boldsymbol{E}_T^{\text{i}}\}$ and visual entities $\{\boldsymbol{E}_V^{\text{i}}\}$ for the news. We link them to the Freebase [2] knowledge graph using the pre-trained entity representation TransE [3] to obtain the background knowledge feature f_{E_T} and f_{E_V} of the textual and visual entities.

Adaptive Hard Attention. The additional knowledge information extracted from the knowledge graph can provide rich external knowledge and evidence. However, not all entities contribute equally to the model, and some entities are completely irrelevant for fake news detection. To measure the weight of each entity and remove irrelevant entities, we propose an adaptive hard attention mechanism. Specifically, we perform hard attention operations on visual entity embedding f_{E_V} and textual entity embedding f_{E_T} through the global feature f_M for multimodal news, which is the concatenation of global text feature f_{G_s} and image feature f_{V_s}, as a way to assign different weight to each entity, and to eliminate irrelevant entities.

We first apply a linear layer to change the dimension of global multimodal news feature $\boldsymbol{f_M}$, so that its dimension is consistent with the entity embedding. To implement the news-textual entity embedding hard attention, we utilize the single-head attention and perform operations similar to the co-attention mechanism described above, but using $\boldsymbol{f_M}$ to obtain the query matrix $\boldsymbol{Q_{TE}}$ and textual entity embedding $\boldsymbol{f_{E_T}}$ to obtain the key matrix $\boldsymbol{K_{TE}}$ and the value matrix $\boldsymbol{V_{TE}}$. Then we can obtain the corresponding attention score $\boldsymbol{\alpha}$ and similarity $\boldsymbol{\beta}$ with the following equations:

$$\begin{aligned} \boldsymbol{\alpha} &= softmax\left(\frac{\boldsymbol{Q_{TE}K_{TE}^T}}{\sqrt{d}}\right) \\ \boldsymbol{\beta} &= \frac{\boldsymbol{Q_{TE}K_{TE}^T}}{\sqrt{d}} \end{aligned} \tag{8}$$

Here, we set a threshold for the attention score $\boldsymbol{\alpha}$, which is calculated as described below:

$$\delta = \frac{mean\,(\boldsymbol{\alpha})}{2} \tag{9}$$

When the attention score α_i is less than the threshold δ, We consider its corresponding entity to be irrelevant to the news and set its similarity β_i to $-\infty$. If the attention score is greater than the threshold δ, the original similarity β_i is kept unchanged with the following formula:

$$\beta_h = \begin{cases} \beta_i & \text{if } \alpha_i > \delta \\ -\infty & \text{if } \alpha_i < \delta \end{cases} \tag{10}$$

After that, we re-perform the *softmax* operation on the newly generated similarity β_h as the new attention score, where the attention score of irrelevant entities has been reset to 0 and the remaining entities has been enhanced. Then, we proceed to the subsequent attention operation steps to obtain the knowledge about the filtered textual entities f_{TE}. The formula is as follows:

$$\boldsymbol{f_{TE}} = (softmax\,(\boldsymbol{\beta_h})\,\boldsymbol{V_{TE}})\,\boldsymbol{W_{TE}} \tag{11}$$

Based on the same hard attention operation, we can obtain the knowledge about the filtered visual entity $\boldsymbol{f_{VE}}$.

Cross-modal Knowledge Interaction. We generate the attention mask of the cross-modal knowledge interaction of textual and visual entity embedding based on the above two hard attention operations, so as to obtain the complementary enhanced features of textual and visual entity embedding, while removing the influence of irrelevant entities. Firstly, we obtain the corresponding label sequence $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2$ based on the two similarities $\boldsymbol{\beta}_{h1}, \boldsymbol{\beta}_{h2}$ generated by the above hard attention operation, which is given by:

$$\eta_i = \begin{cases} 1 & \text{if } \beta_h^i \neq -\infty \\ 0 & \text{if } \beta_h^i = -\infty \end{cases} \tag{12}$$

where the relevant entities are set to 1 and the irrelevant entities are set to 0.

After that we obtain the $m \times n$ $mask_c$ by dotting $\boldsymbol{\eta_1}$ and $\boldsymbol{\eta_2}$ with the following equation:

$$mask_c = \begin{cases} 1 & \text{if } \eta_1^i = \eta_2^i = 1 \\ 0 & \text{if } \eta_1^i \neq \eta_2^i \end{cases} \tag{13}$$

We perform the co-attention operation on the visual entity embedding $\boldsymbol{f_{E_V}}$ and textual entity embedding $\boldsymbol{f_{E_T}}$ with the attention mask $mask_c$ respectively. The formula is as follows:

$$\begin{aligned} f_{E_T E_V} &= \left(softmax\left(\frac{\boldsymbol{Q_{E_T}K_{E_V}^T} \cdot mask_c}{\sqrt{d}}\right)\boldsymbol{V_{E_V}}\right)\boldsymbol{W_{E_T E_V}} \\ f_{E_V E_T} &= \left(softmax\left(\frac{\boldsymbol{Q_{E_V}K_{E_T}^T} \cdot mask_c^T}{\sqrt{d}}\right)\boldsymbol{V_{E_T}}\right)\boldsymbol{W_{E_V E_T}} \end{aligned} \tag{14}$$

After obtaining the augmented four entity knowledge features $f_{TE}, f_{VE}, f_{E_T E_V}, f_{E_V E_T}$, in order to further construct the internal connections, we first concatenate the above four entity knowledge features to turn them into a set of entity embedding. The output is then modelled using one-layer transformer encoder, which is the same as the final step of the multimodal fusion module. So far, we have obtained the feature of the news background knowledge $\boldsymbol{f_E}$.

3.5 Model Optimization

We have obtained the final representation of all features used for classification, which are image caption enhancement feature $\boldsymbol{f_{TC}}$, image original feature $\boldsymbol{f_{V_s}}$, image enhancement feature $\boldsymbol{f_{TV}}$, text original feature $\boldsymbol{f_{G_s}}$, text enhancement feature $\boldsymbol{f_{VT}}$ and knowledge feature $\boldsymbol{f_E}$. We concatenate the above features to form a more comprehensive feature representation, which is further transformed by a FC layer with a sigmoid activation function to predict the probability of fake news, as follows:

$$\hat{y} = \sigma\left(\boldsymbol{W_c}\left[\boldsymbol{f_{VT}};\boldsymbol{f_{TV}};\boldsymbol{f_{TC}};\boldsymbol{f_{G_s}};\boldsymbol{f_{V_s}};\boldsymbol{f_E}\right] + \boldsymbol{b_c}\right) \tag{15}$$

where $\boldsymbol{W_c}$ and $\boldsymbol{b_c}$ are the parameters of the classifier layer, and σ refers to the sigmoid function.

We then used the cross-entropy loss function as the loss for the whole model, which is formulated as described below:

$$\mathcal{L}_p = -y\log\left(\hat{y}\right) - (1-y)\log\left(1-\hat{y}\right) \tag{16}$$

4 EXPERIMENTS

4.1 Experimental Settings

Dataset. We evaluate the proposed method HSEN using two widely used datasets, Pheme [55] and Weibo [17]. The Pheme dataset consists of tweets from the Twitter platform and its content is based on five breaking news stories. The Weibo dataset was collected from the Xinhua News Agency and the Weibo platform, which contains content ranging from May 2012 to January 2016 and has been widely used in previous multimodal fake news detection work. Both of the above datasets have been pre-processed to ensure that each text has its corresponding image. Specifically, the Pheme dataset has 2225 tweets, of which 1577 are real and 648 are false, and the Weibo dataset has 7961 tweets, of which 3642 are true and 4319 are false.

Implementation Details. For the Pheme dataset, we set the length of the input text to 96 words and the number of embedding of both visual and textual entities to 5. For the Weibo dataset, we set the length of the input text to 128 words, the number of visual entity embedding to 5, and the number of textual entity embedding to 8. For the image caption module, we generated one original caption, two image captions guided by two textual entities, and two image captions guided by a single textual entity. For the parameter

Table 1: Performance comparison to the state-of-the-art methods on Pheme and Weibo datasets.

	Methods	Acc	Prec	Rec	F1
Pheme	EANN [41]	0.771	0.714	0.707	0.704
	MVAE [18]	0.776	0.735	0.723	0.728
	SAFE [54]	0.807	0.787	0.789	0.791
	SpotFake [33]	0.845	0.809	0.836	0.822
	KMGCN [44]	0.876	0. 876	0. 876	0. 876
	MM-MTL [51]	0.822	0.788	0.855	0.820
	DDGCN [37]	0.855	0.846	0.841	0.844
	MFAN [53]	0.887	0.871	0.856	0.862
	HESN	**0.908**	**0.886**	**0.890**	**0.888**
Weibo	EANN [41]	0.827	0.847	0.812	0.829
	MVAE [18]	0.824	0.828	0.822	0.823
	SAFE [54]	0.851	0.849	0.849	0.849
	SpotFake [33]	0.873	0.873	0.874	0.873
	CAFE [5]	0.840	0.840	0.841	0.840
	LIIMR [32]	0.900	0.882	0.823	0.847
	EM-FEND [27]	0.904	0.897	0.904	0.901
	BMR [49]	0.918	0.912	0.909	0.910
	HESN	**0.939**	**0.938**	**0.939**	**0.939**

settings, we set *Batch size = 16*, *Epoch = 80*, the learning rate to *5e-4* for the fake news detection model and fine-tune the BERT model with a learning rate of *1e-6*. The optimiser is the Adam. For the reinforcement learning model, we set the learning rate to *1e-4*, the reward function to $\lambda 1 = 180$ and $\lambda 2 = 200$. For the input training data, we generated 4 groups of prompt tokens for normative reward, while the number of prompt tokens in each group is 2. Besides, we employ the Adam optimizer to optimize the learnable MLP.

Evaluation Metrics. We apply the accuracy score as our evaluation metric, which is widely used for binary classification task. Considering the problem of category imbalance, we also utilize precision, recall and F1 score as complementary evaluation metrics.

4.2 Results and Discussion

To fully evaluate the proposed method, we compare it with several state-of-the-art methods on the Pheme and Weibo datasets and the results are shown in Table 1. Under the four evaluation metrics, HSEN achieves 90.8% accuracy and 93.9% accuracy respectively, both outperforming other comparative methods and demonstrating its superior performance.

The EANN and MVAE models apply both visual and textual information, but their performance is poor relative to other methods, which may be due to the use of Text-CNN or Bi-LSTM to extract text feature, whose text representation is weaker than pre-trained models such as BERT. The SAFE model is higher compared to EANN and MVAE, demonstrating that comparing the similarity of text and image is beneficial for fake news detection. The KMGCN and DDGCN models perform well on the Pheme dataset, suggesting that using graph network can extract news feature significantly, while both methods use external knowledge information, indicating that using additional knowledge is beneficial for fake news detection. The MFAN model achieved better results on Pheme dataset, demonstrating social graph information plays an important role in fake news detection. The EM-FEND model performs well on the Weibo dataset, indicating that using the high-order knowledge information of images and comparing the consistency of visual entities and textual entities is effective. The BMR model achieved the second best results on the Weibo dataset, suggesting the effectiveness of using multiple perspectives of news content to make crude predictions and reweighting each perspective feature through cross-modal consistency.

Table 2: Ablation study on the architecture of the image semantic enhancement module and the multimodal fusion module of HSEN on the two datasets.

	Methods	Acc	Prec	Rec	F1
Pheme	w/o Cap	0.879	0.848	0.870	0.859
	w/o RLprompt	0.893	0.865	0.888	0.876
	w/o Fuse	0.884	0.856	0.867	0.862
	HESN	**0.908**	**0.886**	**0.890**	**0.888**
Weibo	w/o Cap	0.902	0.902	0.903	0.903
	w/o RLprompt	0.910	0.909	0.909	0.909
	w/o Fuse	0.922	0.924	0.923	0.924
	HESN	**0.939**	**0.938**	**0.939**	**0.939**

Table 3: Ablation study on the architecture of the knowledge semantic enhancement module of HSEN on the two datasets.

	Methods	Acc	Prec	Rec	F1
Pheme	w/o Filter	0.886	0.861	0.863	0.862
	w/o Mask	0.902	0.885	0.876	0.880
	w/o Cap-E	0.891	0.868	0.870	0.869
	HESN	**0.908**	**0.886**	**0.890**	**0.888**
Weibo	w/o Filter	0.917	0.917	0.917	0.917
	w/o Mask	0.924	0.923	0.924	0.924
	w/o Cap-E	0.927	0.926	0.926	0.926
	HESN	**0.939**	**0.938**	**0.939**	**0.939**

Compared with other methods, our model HSEN outperforms other models, for which we can attribute the advantages of the HSEN model to several factors: (1) The use of BERT and BLIP as backbone, resulting in powerful text and image representation. (2) The image semantic enhancement module acquires valuable multi-level cross-modal correlation information, while using image caption to expand the visual entities, thus enriching the background knowledge information of the image. (3) The knowledge semantic enhancement module learns the precise news high-order knowledge semantic information.

4.3 Ablation Studies

The ablation experiments in Table 2 investigate the impact of the components in the image semantic enhancement module and multimodal fusion module of HSEN in terms of their performance on fake news detection. Specifically, variants of HSEN are described below: *w/o Cap* refers to not using image captions, *i.e.* no semantic information about the image is applied. *w/o RLprompt* refers to not using the reinforcement learning model to generate the optimal

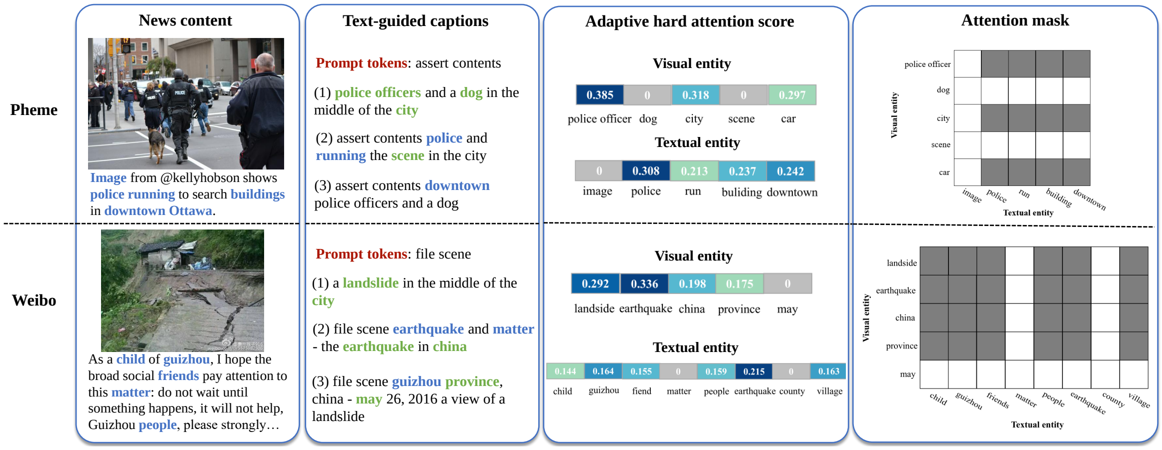
News content
Text-guided captions
Adaptive hard attention score
Attention mask
Pheme
Image from @kellyhobson shows police running to search buildings in downtown Ottawa.
Prompt tokens: assert contents
(1) police officers and a dog in the middle of the city
(2) assert contents police and running the scene in the city
(3) assert contents downtown police officers and a dog
Visual entity
0.385 0 0.318 0 0.297
police officer dog city scene car
Textual entity
0 0.308 0.213 0.237 0.242
image police run buliding downtown
police officer
dog
city
scene
car
Visual entity
image police run building downtown
Textual entity
Weibo
As a child of guizhou, I hope the broad social friends pay attention to this matter: do not wait until something happens, it will not help, Guizhou people, please strongly…
Prompt tokens: file scene
(1) a landslide in the middle of the city
(2) file scene earthquake and matter - the earthquake in china
(3) file scene guizhou province, china - may 26, 2016 a view of a landslide
Visual entity
0.292 0.336 0.198 0.175 0
landside earthquake china province may
Textual entity
0.144 0.164 0.155 0 0.159 0.215 0 0.163
child guizhou fiend matter people earthquake county village
landside
earthquake
china
province
may
Visual entity
child guizhou friends matter people earthquake county village
Textual entity

Figure 5: Two examples of correct recognition from the Pheme and Weibo datasets with extracted textual entities (marked in blue) and visual entities (marked in green), text-guided image captions, attention score and attention mask generated by the adaptive hard attention mechanism are shown.

prompt format, only custom prompt format is used to generate the image captions. *w/o Fuse* refers to not enhancing the three features using co-attention mechanism. These variants of HSEN perform significantly worse than the original HSEN, while *w/o Cap* has the worst performance among these variants. This indicates that: (1) Multi-level image captions can complement the semantic information of the image and improve model performance significantly. (2) Using reinforcement learning to discover the optimal prompt format can generate more valuable image captions. (3) The use of the co-attention mechanism is able to enhance the representation of the three features, further improving fake news detection performance.

The ablation experiments in Table 3 investigate the impact of the components in the knowledge semantic enhancement module. Specifically, *w/o Filter* refers to no knowledge semantic enhancement module is used and only the textual and visual entity embedding is fused by one layer of transformer encoder. *w/o Mask* refers to direct co-attention operation for textual and visual entity embedding without the usage of the generated attention mask. *w/o Cap-E* refers to the image caption is not used to supplement the visual entities. These variants of HSEN illustrate that: (1) filtering irrelevant entities and enhancing relevant entities is necessary, while generating the attention mask for guiding cross-modal knowledge interaction is effective, based on the above operations we can obtain more precise high-order knowledge semantic information. (2) Using image captions to supplement additional entities increases the richness of the visual entities and thus makes better use of the external knowledge information related to the image.

4.4 Visualization Results

Figure 5 gives two example of correct recognition from the Pheme and Weibo datasets, and provides the entities extracted from the text and image captions, the final selected prompt tokens and the generated image captions, as well as the attention score and the attention mask generated by the hard attention mechanism. Regarding the text-guided image captions, we can observe that although the final chosen prompt tokens is somewhat counter-intuitive, it is well placed to guide the multi-level text-related image captions, as well as adding the background knowledge embedded in the BLIP model to the image captions. For example, in the case of the Weibo dataset, the image caption can guide the generation of background knowledge such as '*earthquake*', '*china*', *etc*, which expanding the visual entities and facilitating fake news detection. Regarding the adaptive hard attention score, in the case of the Pheme dataset, the knowledge semantic enhancement module removes '*dog*', '*scene*' in the visual entities and '*image*' in the textual entities, which are obviously useless for fake news detection. This suggests that the module is able to reduce the influence of external knowledge noise. Meanwhile, the knowledge semantic enhancement module assigns the highest weight to '*police*', which is the most relevant to the news content. This indicates that the module can be effective in enhancing the representation of relevant knowledge. Regarding the attention mask, we discover that using the attention mask can significantly remove the influence of noise entities and guide cross-modal knowledge interaction better.

5 CONCLUSION

In this work, we propose a novel Hierarchical Semantic Enhancement Network (HSEN) for multimodal fake news detection. The proposed HSEN acquires cross-modal correlation information by discovering the optimal prompt format through reinforcement learning and guiding the BLIP model to generate image captions related to specific textual entities. Additionally HSEN extends the visual entities with the generated image captions and extracts precise news high-order knowledge semantic information through an adaptive hard attention mechanism. Extensive experiments on two datasets validate the effectiveness of the proposed method.

ACKNOWLEDGMENTS

This work was supported by National Key R&D Program of China under Grant 2020AAA0105702, National Natural Science Foundation of China (NSFC) under Grants 62225207, U19B2038, 62121002 and 62106245.

REFERENCES

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6077–6086.

[2] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, 1247–1250.

[3] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26.

[4] Shizhe Chen, Qin Jin, Peng Wang, and Qi Wu. 2020. Say as you wish: fine-grained control of image caption generation with abstract scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9962–9971.

[5] Yixuan Chen, Dongsheng Li, Peng Zhang, Jie Sui, Qin Lv, Lu Tun, and Li Shang. 2022. Cross-modal ambiguity learning for multimodal fake news detection. In *Proceedings of the ACM Web Conference 2022*, 2897–2905.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: a large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 248–255.

[7] Mingkai Deng, Jianyu Wang, Cheng-Ping Hsieh, Yihan Wang, Han Guo, Tianmin Shu, Meng Song, Eric P Xing, and Zhiting Hu. 2022. Rlprompt: optimizing discrete text prompts with reinforcement learning. *arXiv preprint arXiv:2205.12548*.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[9] Yaqian Dun, Kefei Tu, Chen Chen, Chunyan Hou, and Xiaojie Yuan. 2021. Kan: knowledge-aware attention network for fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence* number 1. Vol. 35, 81–89.

[10] Song Feng, Ritwik Banerjee, and Yejin Choi. 2012. Syntactic stylometry for deception detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 171–175.

[11] Ross Girshick. 2015. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, 1440–1448.

[12] Han Guo, Bowen Tan, Zhengzhong Liu, Eric P Xing, and Zhiting Hu. 2021. Text generation with efficient (soft) q-learning. *arXiv preprint arXiv:2106.07704*.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

[14] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. 2019. Attention on attention for image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4634–4643.

[15] Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

[16] Minyoung Huh, Andrew Liu, Andrew Owens, and Alexei A Efros. 2018. Fighting fake news: image splice detection via learned self-consistency. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 101–117.

[17] Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. 2017. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 25th ACM International Conference on Multimedia*, 795–816.

[18] Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. 2019. Mvae: multimodal variational autoencoder for fake news detection. In *The World Wide Web Conference*, 2915–2921.

[19] Ling Min Serena Khoo, Hai Leong Chieu, Zhong Qian, and Jing Jiang. 2020. Interpretable rumor detection in microblogs by attending to user interactions. In *Proceedings of the AAAI Conference on Artificial Intelligence* number 05. Vol. 34, 8783–8790.

[20] Dong Li, Jiaying Zhu, Menglu Wang, Jiawei Liu, Xueyang Fu, and Zheng-Jun Zha. 2023. Edge-aware regional message passing controller for image forgery localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8222–8232.

[21] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*. PMLR, 12888–12900.

[22] Peiguang Li, Xian Sun, Hongfeng Yu, Yu Tian, Fanglong Yao, and Guangluan Xu. 2022. Entity-oriented multi-modal alignment and fusion network for fake news detection. *IEEE Transactions on Multimedia*, 24, (Jan. 2022), 3455–3468.

[23] Yang Liu and Yi-Fang Wu. 2018. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In *Proceedings of the AAAI Conference on Artificial Intelligence* number 1. Vol. 32.

[24] Jing Ma, Wei Gao, Zhongyu Wei, Yueming Lu, and Kam-Fai Wong. 2015. Detect rumors using time series of social context information on microblogging websites. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, 1751–1754.

[25] Rahul Mishra. 2020. Fake news detection using higher-order user to user mutual-attention progression in propagation paths. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition workshops*, 652–653.

[26] Kartik Narayan, Harsh Agarwal, Surbhi Mittal, Kartik Thakral, Suman Kundu, Mayank Vatsa, and Richa Singh. 2022. Desi: deepfake source identifier for social media. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2858–2867.

[27] Peng Qi et al. 2021. Improving fake news detection by using an entity-enhanced framework to fuse diverse multimodal clues. In *Proceedings of the 29th ACM International Conference on Multimedia*, 1212–1220.

[28] Feng Qian, Chengyue Gong, Karishma Sharma, and Yan Liu. 2018. Neural user response generator: fake news detection with collective user intelligence. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*. Vol. 18, 3834–3840.

[29] Joseph Redmon and Ali Farhadi. 2018. Yolov3: an incremental improvement. *arXiv preprint arXiv:1804.02767*.

[30] Zhihua Shang, Hongtao Xie, Zhengjun Zha, Lingyun Yu, Yan Li, and Yongdong Zhang. 2021. Prrnet: pixel-region relation network for face forgery detection. *Pattern Recognition*, 116, 107950.

[31] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

[32] Shivangi Singhal, Tanisha Pandey, Saksham Mrig, Rajiv Ratn Shah, and Ponnurangam Kumaraguru. 2022. Leveraging intra and inter modality relationship for multimodal fake news detection. In *Companion Proceedings of the Web Conference 2022*, 726–734.

[33] Shivangi Singhal, Rajiv Ratn Shah, Tanmoy Chakraborty, Ponnurangam Kumaraguru, and Shin'ichi Satoh. 2019. Spotfake: a multi-modal framework for fake news detection. In *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*, 39–47. DOI: 10.1109/BigMM.2019.00-44.

[34] Shivangi Singhal, Rajiv Ratn Shah, Tanmoy Chakraborty, Ponnurangam Kumaraguru, and Shin'ichi Satoh. 2019. Spotfake: a multi-modal framework for fake news detection. In *2019 IEEE fifth International Conference on Multimedia Big Data (BigMM)*. IEEE, 39–47.

[35] Zeliang Song, Xiaofei Zhou, Zhendong Mao, and Jianlong Tan. 2021. Image captioning with context-aware auxiliary guidance. In *Proceedings of the AAAI Conference on Artificial Intelligence* number 3. Vol. 35, 2584–2592.

[36] Mengzhu Sun, Xi Zhang, Jianqiang Ma, and Yazheng Liu. 2021. Inconsistency matters: a knowledge-guided dual-inconsistency network for multi-modal rumor detection. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 1412–1423.

[37] Mengzhu Sun, Xi Zhang, Jiaqi Zheng, and Guixiang Ma. 2022. Ddgcn: dual dynamic graph convolutional networks for rumor detection on social media. In *Proceedings of the AAAI Conference on Artificial Intelligence* number 4. Vol. 36, 4611–4619.

[38] Yu-Wun Tseng, Hui-Kuo Yang, Wei-Yao Wang, and Wen-Chih Peng. 2022. Kahan: knowledge-aware hierarchical attention network for fake news detection on social media. In *Companion Proceedings of the Web Conference 2022*, 868–875.

[39] Menglu Wang, Xueyang Fu, Jiawei Liu, and Zheng-Jun Zha. 2022. Jpeg compression-aware image forgery localization. In *Proceedings of the 30th ACM International Conference on Multimedia*, 5871–5879.

[40] Ning Wang, Jiahao Xie, Jihao Wu, Mingbo Jia, and Linlin Li. 2022. Controllable image captioning via prompting. *arXiv preprint arXiv:2212.01803*.

[41] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. Eann: event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 849–857.

[42] Yaqing Wang, Fenglong Ma, Haoyu Wang, Kishlay Jha, and Jing Gao. 2021. Multimodal emergent fake news detection via meta neural process networks. In *Proceedings of the 27th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 3708–3716.

[43] Yiyu Wang, Jungang Xu, and Yingfei Sun. 2022. End-to-end transformer based model for image captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence* number 3. Vol. 36, 2585–2594.

[44] Youze Wang, Shengsheng Qian, Jun Hu, Quan Fang, and Changsheng Xu. 2020. Fake news detection via knowledge-driven multimodal graph convolutional networks. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*, 540–547.

[45] Ke Wu, Song Yang, and Kenny Q Zhu. 2015. False rumors detection on sina weibo by propagation structures. In *2015 IEEE 31st International Conference on Data Engineering*. IEEE, 651–662.

[46] Yang Wu, Pengwei Zhan, Yunjian Zhang, Liming Wang, and Zhen Xu. 2021. Multimodal fusion with co-attention networks for fake news detection. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2560–2569.

[47] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: neural image caption generation with visual attention. In *International Conference on Machine Learning*. PMLR, 2048–2057.

[48] Xiaoyu Yang, Yuefei Lyu, Tian Tian, Yifei Liu, Yudong Liu, and Xi Zhang. 2021. Rumor detection on social media with graph structured adversarial learning. In *Proceedings of the twenty-ninth International Conference on International Joint Conferences on Artificial Intelligence*, 1417–1423.

[49] Qichao Ying, Xiaoxiao Hu, Yangming Zhou, Zhenxing Qian, Dan Zeng, and Shiming Ge. 2022. Bootstrapping multi-view representations for fake news detection. (2022). arXiv: 2206.05741 [cs.CV].

[50] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*.

[51] Huaiwen Zhang, Shengsheng Qian, Quan Fang, and Changsheng Xu. 2021. Multi-modal meta multi-task learning for social media rumor detection. *IEEE Transactions on Multimedia*, 24, 1449–1459.

[52] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. Vinvl: revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5579–5588.

[53] Jiaqi Zheng, Xi Zhang, Sanchuan Guo, Quan Wang, Wenyu Zang, and Yongdong Zhang. 2022. Mfan: multi-modal feature-enhanced attention networks for rumor detection. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*. (July 2022), 2413–2419.

[54] Xinyi Zhou, Jindi Wu, and Reza Zafarani. 2020. : similarity-aware multi-modal fake news detection. In *Advances in Knowledge Discovery and Data Mining: 24th Pacific-Asia Conference, PAKDD 2020, Singapore, May 11–14, 2020, Proceedings, Part II*. Springer, 354–367.

[55] Arkaitz Zubiaga, Maria Liakata, and Rob Procter. 2017. Exploiting context for rumour detection in social media. In *Social Informatics: 9th International Conference, SocInfo 2017, Oxford, UK, September 13-15, 2017, Proceedings, Part I 9*. Springer, 109–123.