Boosting Entity-aware Image Captioning with Multi-modal Knowledge Graph

Wentian Zhao, Xinxiao Wu

***Abstract*—Entity-aware image captioning aims to describe named entities and events related to the image by utilizing the background knowledge in the associated article. This task remains challenging as it is difficult to learn the association between named entities and visual cues due to the long-tail distribution of named entities. Furthermore, the complexity of the article brings difficulty in extracting fine-grained relationships between entities to generate informative event descriptions about the image. To tackle these challenges, we propose a novel approach that constructs a multi-modal knowledge graph (MMKG) to associate the visual objects with named entities and capture the relationship between entities simultaneously with the help of external knowledge collected from the web. Specifically, we build a text sub-graph by extracting named entities and their relationships from the article, and build an image sub-graph by detecting the objects in the image. To connect these two sub-graphs, we propose a cross-modal entity matching module trained using a knowledge base that contains Wikipedia entries and the corresponding images. Finally, the MMKG is integrated into the captioning model via a graph attention mechanism. Extensive experiments on both GoodNews and NYTimes800k datasets demonstrate the effectiveness of our method.**

***Index Terms*—Image Captioning, Named Entity, Knowledge Graph**

I. Introduction

Different from the conventional image captioning [1], [2], [3], [4], [5] [6], [7] that describes common objects and their relationships, entity-aware image captioning focuses on generating informative descriptions of named entities and specific events presented in the images by utilizing the background knowledge in the associated articles. For instance, the image in Figure 1 shows the scene where a famous racer is celebrating his victory. A conventional captioning model may describe the general semantic in the image such as "Four men are holding trophies on a podium.", while the entity aware captioning model can leverage the relevant background knowledge to generate more expressive description such as "Alonso is celebrating victory with his Toyota teammates". Entity-aware image captioning is closer to the human cognition process that integrates prior knowledge for understanding and interpreting scenes [8], and has attracted increasing attention in the fields of computer vision and natural language processing [9], [10], [11], [12], [13]. However, this task still presents two key challenges. First, since the distribution of named entities is heavily imbalanced and long-tailed, it is difficult to select the named entities relevant to the image from the articles. Second, the news articles typically contain complex sentences where related entities might be far apart, making it non-trivial to exploit the fine-grained relationships between the named entities for generating descriptions about the events in the image.

Article Text: …, Alonso's victory completes the second part of motorsport's 'triple crown' for the two-time Formula 1 world champion. Twice a Monaco Grand Prix winner, he now wants to win the Indianapolis 500. Only Englishman Graham Hill has won all three classic races in the entire history of motorsport.

General Caption: Four men holding trophies are standing on a podium.

Entity-aware Caption: Alonso's victory, alongside his Toyota team-mates, completes the second part of the Spaniard's motorsport 'triple crown'

Fig. 1: An example of entity-aware image captioning. The named entities in type "PERSON" are marked in red.

There have been some attempts to select accurate named entities for the caption. The template-based methods [9], [11] first generate template captions with placeholders, and then fill the placeholders using the named entities in the relevant sentences. Several end-to-end methods have also been proposed, including using word-level contextual information to draw named entities [13] and applying byte-pair encoding to generate rare words in named entities [12]. The aforementioned named entity selection strategies heavily depend on the contextual information of the named entities in the article but neglect the association between named entities and visual cues in the image.

Recently there have been a few endeavours to utilize the background knowledge in the articles for concrete event description. Many methods extract the knowledge by encoding the article text at article level [9], sentence level [11] or word level [12]. In [13], a more fine-grained attention mechanism is

Wentian Zhao is with the Beijing Laboratory of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology, Beijing 100081, China (e-mail: wentian_zhao@bit.edu.cn).

Xinxiao Wu is with the Beijing Laboratory of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology, Beijing 100081, China, and also with the Guangdong Provincial Laboratory of Machine Perception and Intelligent Computing, Shenzhen MSU-BIT University, Shenzhen 518172, China (e-mail: wuxinxiao@bit.edu.cn).

proposed to progressively concentrate on the text information from the sentence level to the word level. Most of these methods employ sequence encoding models for captioning and lack ability to capture the entity relationships. This limits their applications to the difficult cases where the article presents more complex events with multiple named entities in different sentences.

To overcome these limitations, we propose a multi-modal knowledge graph (MMKG) to explicitly model the association between visual objects and named entities, and simultaneously capture the fine-grained relationships between named entities for entity-aware image captioning. Since there are a wide range of named entities involved in the articles and the distribution of real-world named entities is often long-tailed, it is extremely difficult to learn the association between visual objects and named entities from the training data. Therefore, we attempt to explore the external knowledge from the web which provides rich and comprehensive multi-modal information about named entities. To be more specific, starting with collecting an external multi-modal knowledge base from Wikipedia, we then train a cross-modal entity matching module that connects a textual sub-graph extracted from the article and an image sub-graph extracted from the image, in order to construct MMKG. Finally, MMKG along with the input image and the news article is encoded via a graph attention mechanism into a captioning model to generate the entity-aware descriptions.

In summary, our contributions are as follows:

- We propose a novel entity-aware captioning method that constructs multi-modal knowledge graphs to choose accurate named entities and refine relevant events for generating informative descriptions.
- We design a novel cross-modal entity matching module that is effectively trained using a multi-modal knowledge base collected from Wikipedia to facilitate modeling the association between visual objects and named entities.
- Extensive experiments on two large-scale news image captioning datasets, namely GoodNews [11] and NYTimes800k [12], have verified the superiority of our method compared with the state-of-the-art methods.

II. Related Work

A. Image Captioning

Image captioning has received increasing attention due to the advances in computer vision and natural language processing. Attention mechanism is used to selectively focus on certain objects or regions in the image at each time step [14], [15], [16], [3] or model the visual context that evolves over time [17]. Graph neural networks [18], [19], [20] are used to model the spatial and semantic relationships between objects in the image. Transformer [21] is also employed by recent methods to encode the pair-wise relationship between visual features [22], [5], [23], [7] or syntactic structure of the sentence [6].

B. Entity-aware Image Captioning

Entity-aware image captioning is attracting increasing attention in recent years. Some existing methods generate entity-aware captions by utilizing hashtags or named entities retrieved from the web [10], [24]. Several other methods extract the background knowledge from the associated news articles for generating the named entities [9], [10], [11], [13], [12], [25], which are closer to our method. Among these methods, the early works [9], [10], [11] generate entity-aware captions by two separate steps: a template caption is first generated with placeholders indicating the entity types, and then the placeholders are filled by selecting the named entities from the news article. For instance, Biten *et al*. [11] first generate a template caption by attending to both the image and the sentences in the news article, and then draw the named entities from the sentences with the highest attention weights.

More recently, a few end-to-end methods [13], [12], [25], [26] are proposed to generate entity-aware captions in one pass. Tran *et al*. [12] encode the news article at word level using a pre-trained language model, and handle the rare words in the named entities using byte-pair encoding [27]. Hu *et al*. [13] first retrieve the sentences that are most relevant to the image, and then attend to the retrieved sentences. Yang *et al*. [26] integrate templates that describe the key elements of the caption, i.e. who, what, when, where, why and how, to the entity-aware captioning model. Zhang *et al*. [28] uses CLIP [29] to transfer the visual concepts to linguistic space, and fine-tunes BART [30] with multi-modal prompts to generate the entity-aware caption.

Although remarkable progress has been achieved through leveraging textual information from the news article at the sentence or word level in the aforementioned methods, the explicit association between the named entities and visual objects is still under-explored, as well as the fine-grained relationship between named entities. This paper focuses on constructing MMKG to associate the visual objects with named entities and capture the entity relationships, simultaneously.

C. Multi-modal Knowledge Graph

Multi-modal knowledge graphs introduce information in multiple modalities, such as images, videos or text, to represent the entities and relations. Several recent studies have validated the effectiveness of MMKGs in different fields. For instance, some methods [31], [32] incorporate images as additional features of the entities in the knowledge graph to learn better entity representations for knowledge graph completion and triple classification. Kannan *et al*. [33] construct multi-modal knowledge base to excavate the facts in deep learning literatures. A MMKG based recommendation system is proposed in [34], where images and text entities are introduced to model user behaviour.

To the best of our knowledge, our method is the first to employ MMKGs in entity-aware image captioning. Compared with the methods mentioned above, it is more challenging to construct MMKGs for the images and the associated news articles, since the relationship between the entities in the news article and the visual objects in the image is unknown. Therefore, we leverage the external knowledge from the web to train a cross-modal entity matching module that establishes the connections between entities in different modalities.

D. Relationship Modeling in Visual Recognition

Relationship modeling is a powerful technique that has been applied to various visual recognition tasks, such as classification and video action recognition. For instance, in the field of image classification, Shu *et al*. [35] model the relationship between web text and images by transferring the semantic knowledge in web text domain to image domain using deep transfer networks (DTN). Tang *et al*. [36] improve DTN by adding representation-shared and parameter-shared constraints. Du *et al*. [37] perform fine-grained visual recognition by learning how the object parts interact with each other in multi-granularity fashion. Guo *et al*. [38] propose a few-shot fine-grained visual classification method that calibrates the class centers by modeling the correlation between the query sample and the support samples. In action recognition, Shu *et al*. [39] model the correlation between RGB features and skeleton features using modal-wise attention and channel-wise attention.

The above methods mainly model the relationship between samples in different modalities or features at multiple granularities. In comparison, our method explore modeling the relationship between the named entities in the news article and the visual objects in the image using multi-modal knowledge graph.

III. Our Method

A. Overview

In this paper, we propose a MMKG for entity-aware image captioning. It explicitly models the association between named entities and visual objects in the image and simultaneously captures the fine-grained relationships between named entities in the article. MMKG consists of a text sub-graph and an image sub-graph. The text sub-graph models the interaction between the named entities in the article text, where the nodes represent the named entities and relationships between named entities in the text. The directed edges in the text sub-graph represent the connections between named entities and relationships. The image sub-graph represents the visual objects detected in the image. To connect these two sub-graphs for generating the complete knowledge graph, we introduce a cross-modal entity matching module that measures the similarity between the named entities in the text sub-graph and the visual objects in the image sub-graph. To facilitate the training of the cross-modal entity matching module, we collect an external multi-modal knowledge base from Wikipedia and Google Images. Finally, the captioning model encodes MMKG as well as the image and article to generate the entity-aware captions. Figure 2 illustrates the framework of our method.

B. External Multi-modal Knowledge Base

The external multi-modal knowledge base contains pairs of named entities and images, formulated as $D^M = \{(e_i, v_i)|_i\}$, where e_i and v_i denote the named entity and the corresponding image, respectively. The named entities in D^M are collected from the news articles in the training splits of the GoodNews and NYTimes800k datasets. We perform named entity recognition on all the news articles and keep the named entities that can be represented by images, including persons, organizations, artifacts and facilities. The abstract named entities, e.g. numbers and dates, are not included. To reduce the ambiguity in the original text, we perform entity linking [40] to connect the recognized named entities in the news article to Wikipedia entries.

Ideally, the image corresponding to a named entity should be representative, i.e. the most salient object in the images reflects the entity. For instance, the image that best reflects a person is the person's portrait, and the image corresponding to a building is the close-up of the building. Note that the purpose of the images in each Wikipedia page is to demonstrate the named entity itself, and we believe that the images in D^M collected from Wikipedia and the top image search results from Google Images are with superior representativeness. For each named entity that is linked to a Wikipedia entry and can be represented by images, we collect the first image in its corresponding Wikipedia page and the top three images searched from Google Images. We detect the objects or faces in these images, and the objects or faces with the largest bounding box in each image are added to the multi-modal knowledge base, resulting in a total of about 490,000 pairs of named entities and images.

C. Cross-modal Entity Matching Module

The cross-modal entity matching module measures the similarity between a named entity e_i and an image v_j. A pre-trained language model and a pre-trained CNN are used to encode the vector representations of e_i and v_j, respectively, denoted as $\boldsymbol{u}_{e_i}$ and $\boldsymbol{u}_{v_j}$. The cross-modal entity matching module is trained to map the vectors $\boldsymbol{u}_{e_i}$ and $\boldsymbol{u}_{v_j}$ into a common embedding space, where the similarity between positive entity-image pairs is larger than any negative pairs by a margin δ.

We use the pairs of named entities and images in the multi-modal knowledge base to train the cross-modal entity matching module. The loss function used to train the cross-modal entity matching module is formulated as

$$\begin{aligned} L_r = &\max_{e'}(\delta + \mathrm{sim}(e', v) - \mathrm{sim}(e, v))_+ \\ &+ \max_{v'}(\delta + \mathrm{sim}(e, v') - \mathrm{sim}(e, v))_+, \end{aligned} \tag{1}$$

where the pairs (e', v) and (e, v') denote negative samples, $\mathrm{sim}(e, v)$ denotes the similarity between e and v, and $(x)_+ = \max(x, 0)$.

D. Multi-modal Knowledge Graph

Compared with conventional knowledge graphs that only contains entities extracted from text, MMKGs introduce additional modality of entities and relationships, such as images. Formally, MMKG is denoted as $G^M = \langle V, E \rangle$, where V denotes the entity set and E denotes the edge set. The entity set V contains both the entities from the article and the visual objects from the image. We construct the MMKG G^M by first building a text sub-graph G^T and an image sub-graph G^I, and then connecting these two sub-graphs using the cross-modal

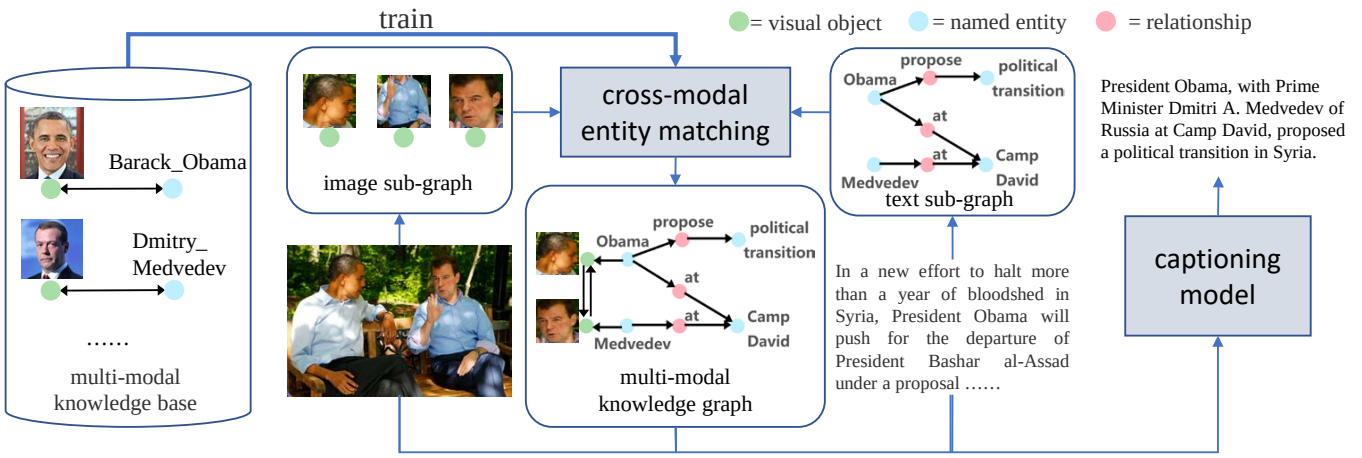


Fig. 2: The framework of our proposed method. The left part shows the external multi-modal knowledge base containing named entities and their corresponding images, which are used to train the cross-modal entity matching module. The middle part shows the generation process of the multi-modal knowledge graphs. An image sub-graph and a text sub-graph are extracted from the input image and the article text, respectively. The multi-modal entity matching module connects the related entities in the two sub-graphs to construct the multi-modal knowledge graph. The right part shows the captioning model, which encodes the image, the article and the multi-modal knowledge graph to generate an entity-aware caption.

entity matching module. An example of MMKG is shown in Figure 3.

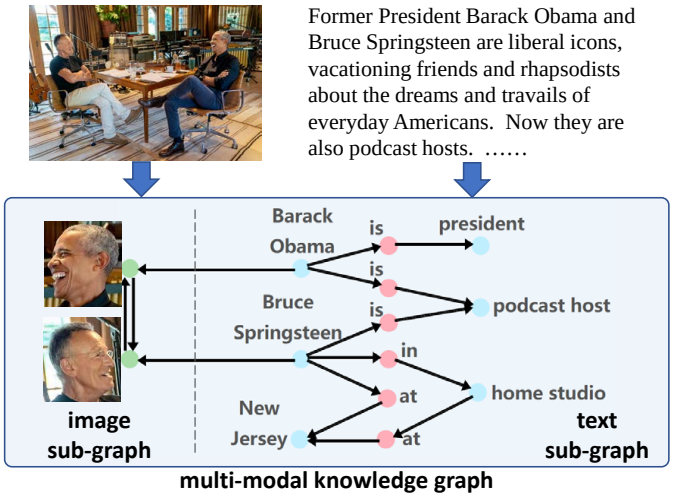


Fig. 3: An example of the constructed multi-modal knowledge graph, which consists of an image sub-graph (the left part of the box) and a text sub-graph (the right part of the box).

To build the text sub-graph, denoted as $G^T = \langle V^T, E^T \rangle$, where V^T and E^T denote the entities in the news article and the edges that connects them, we perform information extraction and coreference resolution by utilizing the Stanford CoreNLP toolkit [41]. Specifically, for an article T, the output of information extraction is a set of triples $R^T = \{\langle e_i^h, e_i^r, e_i^t \rangle|_i\}$, where e_i^h, e_i^r and e_i^t denote the head entity, the relation and the tail entity, respectively. For each triple $\langle e_i^h, e_i^r, e_i^t \rangle$, we create two directed edges $e_i^h \rightarrow e_i^r$ and $e_i^r \rightarrow e_i^t$. To simplify the graph structure, for the same entity that appears more than once in the article, we only keep the entity that appears first. We use the pre-trained Wikipedia2Vec vector [42] as the embedding $\boldsymbol{u}_{e_i^h}$ if a named entity e_i^h is linked to a Wikipedia entry. Otherwise, the embedding $\boldsymbol{u}_{e_i^h}$ is obtained by encoding the corresponding text using RoBERTa.

To construct the image sub-graph, denoted as $G^I = \langle V^I, E^I \rangle$, we use the YOLOv3 object detector [43] and the MTCNN [44] network to detect the objects and faces in the image, and the detected objects and faces are denoted by $V^o = \{v_i^o|_i\}$ and $V^f = \{v_i^f|_i\}$, respectively. The entity set V^I of the image sub-graph contains both the set of objects and the set of faces, i.e. $V^I = V^o \cup V^f$. The embedding $\boldsymbol{u}_{v_i^o} \in \mathbb{R}^{2048}$ of the object v_i^o is extracted using pre-trained ResNet-152 model [45]. The embedding $\boldsymbol{u}_{v_i^f} \in \mathbb{R}^{2048}$ of face v_i^f is obtained using $\boldsymbol{u}_{v_i^f} = \boldsymbol{W}_f \boldsymbol{u}'_{v_i^f}$, where $\boldsymbol{u}'_{v_i^f} \in \mathbb{R}^{512}$ is the face embedding extracted by pre-trained FaceNet model [46], and $\boldsymbol{W}_f \in \mathbb{R}^{2048 \times 512}$ is a learnable parameter. For each pair of nodes $\langle v_i^\sharp, v_j^\sharp \rangle$ in the image sub-graph, we connect two directed edges $v_i^\sharp \rightarrow v_j^\sharp$ and $v_j^\sharp \rightarrow v_i^\sharp$, where $\sharp \in \{o, f\}$.

The similarities between the entities in the image sub-graph and the text sub-graph are measured by the cross-modal entity matching module. For each entity pair $\langle e_i^*, v_j^\sharp \rangle$ that satisfies $\text{sim}(e_i^*, v_j^\sharp) > 0.4$, we create a directed edge $e_i^* \rightarrow v_j^\sharp$ that connects them, where $\text{sim}(e_i^*, v_j^\sharp)$ denotes the similarity between e_i^* and $v_j^\sharp$. $* \in \{h, r, t\}$ and $\sharp \in \{o, f\}$.

E. Entity-aware Captioning Model

The generated multi-modal knowledge graph G^M, the input image I and the associated news article T are encoded by the encoder of the entity-aware captioning model. The image I is encoded using the pre-trained ResNet-152, and the output before the last pooling layer is taken and flattened into a matrix $\boldsymbol{X}^I = \{\boldsymbol{x}_j^I|_j\}$, $\boldsymbol{x}_j^I \in \mathbb{R}^{2048}$.

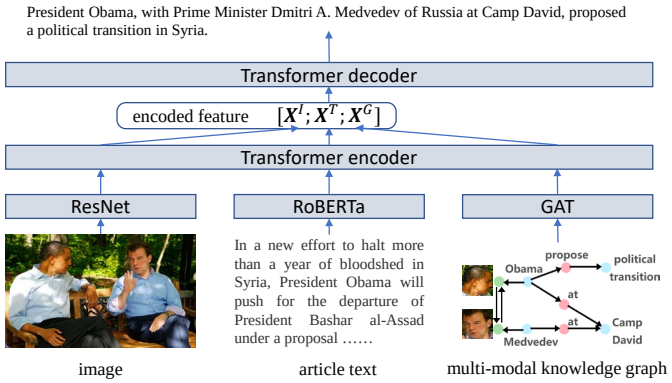


Fig. 4: The architecture of the captioning model.

The news article T is encoded by the pre-trained RoBERTa [47] into a sequence of subword units $\{w_1, w_2, ..., w_{L_T}\}$, where L_T denotes the length of the subword unit sequence. The output of the last layer is used as the representation, denoted as $\boldsymbol{X}^T = \{\boldsymbol{x}_j^T|_j\}$, $\boldsymbol{x}_j^T \in \mathbb{R}^{1024}$.

To obtain rich representations of the entities in MMKG, we encode the nodes in the MMKG using a two-layer graph attention network (GAT) [48]. The node representations in MMKG, denoted as $\boldsymbol{U}^G = \boldsymbol{U}^T \cup \boldsymbol{U}^I$, is used as the input of the GAT. $\boldsymbol{U}^T = \{\boldsymbol{u}_{e_i^*}|e_i^* \in V^T\}$ and $\boldsymbol{U}^I = \{\boldsymbol{u}_{v_j^\sharp}|v_j^\sharp \in V^I\}$ denote the vector representations of the nodes in the text sub-graph G^T and the image sub-graph G^I, respectively. $\boldsymbol{u}_{e_i^*}$ and $\boldsymbol{u}_{v_j^\sharp}$ denote the initial vector representations of named entity e_i^* and visual object $v_j^\sharp$, respectively.

The output of the last GAT layer is used as the input to the decoder, denoted as $\boldsymbol{X}^G = \{\boldsymbol{x}_{e_i^*}|e_i^* \in V^T\} \cup \{\boldsymbol{x}_{v_j^\sharp}|v_j^\sharp \in V^I\}$ where $\boldsymbol{x}_{e_i^*}$ and $\boldsymbol{x}_{v_j^\sharp}$ denote the vector representations of named entity e_i^* and visual object $v_j^\sharp$ that are encoded by the graph attention network, respectively.

The decoder of the captioning model generates the tokens in the entity-aware captions sequentially and consists of N identical Transformer layers. The initial input to the decoder is denoted as $\boldsymbol{X}_0 = [\boldsymbol{X}^T; \boldsymbol{X}^I; \boldsymbol{X}^G]$, where the operator $[;]$ denotes matrix concatenation. At the t-th time step, the decoder predicts the probability of the current token $\boldsymbol{p}_t \in \mathbb{R}^D$ using the initial input and the embeddings of the previously generated subword units $\boldsymbol{M}_{t-1} = \{\boldsymbol{m}_0, \boldsymbol{m}_1, ..., \boldsymbol{m}_{t-1}\}$, where D denotes the vocabulary size and $\boldsymbol{m}_i$ denotes the embedding of the i-th subword unit.

When training the captioning model, we fix the parameters of the pre-trained RoBERTa and the pre-trained CNN in the encoder. The parameters of GAT and the decoder are optimized using the following cross-entropy loss function:

$$L_p = -\sum_{t=1}^{|Y|} \log p(w_t|w_1, w_2, ..., w_{t-1}), \tag{2}$$

where Y denotes the length of the ground-truth caption, and w_i denotes the i-th token in Y. The overall structure of the captioning model is shown in Figure 4.

IV. Experiments

A. Datasets and Evaluation Metrics

The images, captions and news articles in the GoodNews dataset and the NYTimes800k dataset are collected from the New York Times, and each image is annotated with one ground-truth caption. We employ the dataset split of the GoodNews dataset in [12], where the training, validation and testing splits contain 445,259 images, 19,448 images and 24,461 images, respectively. The average lengths of the news articles and the ground-truth captions are 451 words and 18 words, respectively. The NYTimes800k dataset contains 763,217 images, 7,777 images and 21,977 images for training, validation and testing, respectively. Compared to the GoodNews dataset, NYTimes800k is more complex since the average article length and the average caption length of NYTimes800k are 974 words and 18 words, respectively.

We use standard image captioning metrics, including Bleu-4 [49], METEOR [50], ROUGE-L [51] and CIDEr [52] to evaluate the similarity of the generated captions to the ground-truth captions. Since the goal of the model is to generate entity-aware captions, we also evaluate the named entities in the genrated sentences. Specifically, the Spacy toolkit [53] is used to recognize the named entities in both ground-truth sentences and generated sentences. Following [12], we count the exact string matches between all the named entities in the ground-truth captions and the generated captions, to calculate the F1 score for the generated named entities. We also evaluate the performance of generating long-tail named entities by reporting F1 scores of the named entities that appear less than 100 times in the training split. The F1 score of all the named entities and the long-tail named entities are denoted as F1 and L-F1, respectively.

B. Implementation Details

To detect the objects in the image I, we use the YOLOv3 detector [43] and filter out the object bounding boxes with a confidence score of less than 0.3, while a maximum of 64 objects are kept for each image. To detect the faces, we use the MTCNN [44] network. Following the practice in [12], we keep no more than 4 face bounding boxes with the highest confidence scores for each image. To extract the features of the input images as well as the detected objects and faces, we use the pre-trained ResNet-152 [45] model.

The hidden dimension of the decoder in the captioning model is set to 1024. The sub-word vocabulary of the decoder is identical to the sub-word vocabulary of the pre-trained RoBERTa [47] model, which contains about 50,000 sub-word units. We set the maximum length of the article text T and the caption Y to 512 tokens and 50 tokens, respectively. We optimize the parameters of the model using the Adam optimizer [54] and the weight decay is set to 1×10^{-5}. We apply L_2 regularization to the parameters and clip the gradient by the norm of 0.1. The batch size is set to 16, and the model is trained for 409,600 steps in total. In the first 20,000 training steps, the learning rate increases from 1×10^{-7} to 1×10^{-4}, and linearly decreases in the remaining training steps.

Dataset	Decoder	Method	B-4	M	R	C	F1	L-F1
GoodNews	LSTM	ICECAP [13]	1.96	6.01	15.70	26.08	12.03	-
	Transformer	Transform and Tell [12]	6.05	-	21.40	53.80	20.30	7.13
	Transformer	Ours	6.82	11.42	22.83	60.84	25.81	8.34
	JoGANIC	JoGANIC [26]	6.34	10.78	21.65	59.19	22.60	-
	JoGANIC	JoGANIC+MSTR+NEE(auto)	6.83	11.25	23.05	61.22	24.22	-
	JoGANIC	Ours+JoGANIC	6.83	11.52	23.14	62.03	25.37	8.25
	BART	NewsMEP[28]	8.30	12.23	23.17	63.99	23.33	-
	BART	Ours+BART	**8.31**	**12.32**	**23.22**	**64.15**	**23.39**	**8.43**
NYTimes800k	Transformer	Transform and Tell [12]	6.30	-	21.70	54.40	23.34	5.57
	Transformer	Ours	6.71	9.60	22.13	57.97	25.14	6.35
	JoGANIC	JoGANIC [26]	6.39	10.75	22.38	56.54	25.41	-
	JoGANIC	JoGANIC+MSTR+NEE(auto)	6.79	10.93	22.80	59.42	26.39	-
	JoGANIC	Ours+JoGANIC	6.77	11.22	22.96	59.94	26.23	6.81
	BART	NewsMEP[28]	**9.57**	13.02	23.62	65.85	27.55	-
	BART	Ours+BART	9.53	**13.30**	**23.89**	**66.43**	**27.71**	**7.72**

TABLE I: Evaluation results of entity-aware captioning on the GoodNews dataset and the NYTimes800k dataset. B-4, M, R, C, F1 and L-F1 are the abbreviations of Bleu-4, METEOR, ROUGE-L, CIDEr, the F1 score of all named entities and the F1 score of long-tail named entities, respectively. The best results are marked in Bold. The "JoGANIC" in the "Decoder" column denotes using the decoder proposed by [26], and "BART" denotes using the decoder in [28].

		GoodNews					NYTimes800k				
Method	Decoder	B-4	R	C	F1	L-F1	B-4	R	C	F1	L-F1
w/o graph	Transformer	5.34	19.32	47.70	18.34	7.02	5.31	19.73	48.51	19.87	5.79
w/o text graph		5.68	19.57	47.92	18.53	7.41	5.72	19.89	48.90	19.93	5.93
w/o image graph		6.01	20.02	49.75	19.37	7.13	6.10	20.08	52.72	21.07	6.10
w/o matching		6.02	20.13	50.24	19.01	7.07	6.15	21.03	53.31	22.58	6.13
w/o matching+complete		5.99	20.03	52.89	20.79	7.34	6.45	21.27	56.19	23.02	6.10
Ours		**6.82**	**22.83**	**60.84**	**25.81**	**8.34**	**6.71**	**22.13**	**57.97**	**25.14**	**6.35**
w/o graph	JoGANIC	6.34	21.65	59.18	22.60	-	6.39	22.38	56.54	25.41	-
w/o text graph		6.57	22.02	59.23	22.59	7.95	6.59	22.42	56.69	25.43	5.90
w/o image graph		6.60	22.15	59.22	22.04	8.06	6.62	22.49	57.39	25.57	5.97
w/o matching		6.65	22.13	60.02	22.76	8.09	6.60	22.63	58.40	25.67	6.23
w/o matching+complete		6.70	22.45	61.30	23.75	8.13	6.68	22.67	57.93	26.03	6.35
Ours		**6.83**	**23.14**	**62.03**	**25.37**	**8.25**	**6.77**	**22.96**	**59.94**	**26.23**	**6.81**

TABLE II: The results of ablation studies on the GoodNews dataset and the NYTimes800k dataset.

C. Comparison with State-of-the-Art Methods

We compare our method with end-to-end entity aware captioning methods, including "ICECAP"[13], "Transform and Tell" [12], "JoGANIC"[26] and "NewsMEP" [28]. "ICECAP" [13] uses LSTM to generate the entity-aware caption, while "Transform and Tell" [12] and "JoGANIC" uses Transformer [21] as the decoder. In particular, "JoGANIC" introduces additional component-specific Transformer blocks, together with multi-span text reading mechanism (MSTR) that encodes more than 512 tokens in the article and entity embedder (NEE) that uses pre-trained Wikipedia2vec [42] representations of the named entities. "NewsMEP" designs a contextual prompts module to select named entities, and utilizes the decoder of BART [30] to generate entity-aware captions. For fair comparison, we re-implement the component-specific Transformer blocks in "JoGANIC" and use these blocks to replace the last layer of our model, denoted as "Ours+JoGANIC". We also use the decoder of NewsMEP [28] that is initialized by the parameters of BART, denoted as "Ours+BART".

Table I shows the results of the aforementioned state-of-the-art methods and our method. We observe that our method achieves the best results in terms of most evaluation metrics on both datasets, especially Bleu-4 and CIDEr, which validates the superority of the proposed MMKG on entity-aware image captioning. The entity F1 score of our method outperforms all existing methods on both datasets, which indicates that by modeling the association between the named entities and the visual cues in the image, our method selects the named entities from the news articles more accurately. Our method also outperforms "Transform and Tell" in terms of the F1 score of long-tail named entities, which indicates that our method selects the long-tail named entities more appropriately. When using the decoder of "JoGANIC" and BART, our method ("Ours+JoGANIC" and "Ours+BART") outperforms "JoGANIC" and "NewsMEP" on most evaluation metrics, which implies that our method works effectively with different

Subset	Method	B-4	R	C
GoodNews w/ people	w/o graph	5.39	21.55	53.26
	w/o graph+JoGANIC	6.70	21.46	59.03
	Ours	6.31	21.60	58.41
	Ours+JoGANIC	**6.79**	**22.34**	**61.03**
GoodNews w/o people	w/o graph	4.11	19.78	37.58
	w/o graph+JoGANIC	6.13	22.79	59.30
	Ours	5.22	19.92	45.10
	Ours+JoGANIC	**6.85**	**24.37**	**60.07**
NYTimes800k w/ people	w/o graph	5.41	21.03	53.47
	w/o graph+JoGANIC	6.24	22.18	56.03
	Ours	6.39	22.04	56.73
	Ours+JoGANIC	**6.73**	**23.02**	**57.87**
NYTimes800k w/o people	w/o graph	4.12	19.75	39.50
	w/o graph+JoGANIC	6.52	22.46	57.19
	Ours	5.70	20.06	47.12
	Ours+JoGANIC	**6.72**	**22.34**	**58.03**

TABLE III: Evaluation results on captions that involve people (w/ people) and captions that do not contain people (w/o people).

kinds of decoders.

D. Ablation Studies

To verify the effect of each component in our method, we conduct ablation studies on the GoodNews dataset and the NYTimes800k dataset by using the Transformer as the decoder. We evaluate the following variants of our model: **(1) w/o graph**: To evaluate the effect of the MMKG, the entire MMKG is removed, and the input to the captioning model only includes the region features of the image $\boldsymbol{X}^I$ and the token-level features of the news article $\boldsymbol{X}^T$; **(2) w/o text graph**: To evaluate the contribution of the text sub-graph G^T, the text sub-graph is removed and the input to the decoder includes $\boldsymbol{X}^T$, $\boldsymbol{X}^I$ and the features of the objects and faces in the image; **(3) w/o image graph**: To evaluate the contribution of the image sub-graph, we remove the objects and faces in G^I. The input to the decoder includes $\boldsymbol{X}^T$, $\boldsymbol{X}^I$ and the representations of the nodes in the text sub-graph G^T; **(4) w/o matching, w/o matching+complete**: To validate the effectiveness of the cross-modal entity matching module, we remove the cross-modal entity matching module and use two different strategies of constructing the multi-modal knowledge graph: removing all the edges between the image sub-graph and the text sub-graph (**w/o matching**), and connecting all the possible edges between the image sub-graph and the text sub-graph (**w/o matching+complete**).

The results of ablation studies are shown in Table II. From these results, we have the following observations: First, the performance of "w/o graph" significantly degrades on all the metrics, which indicate that the MMKG is beneficial to describing the events in the image as well as selecting named entities. Second, when using only the image sub-graph or the text sub-graph, the values of most metrics slightly increase, demonstrating that either sub-graphs contribute to improving the performance. Third, compared to the image sub-graph, using the text sub-graph achieves better results, indicating that the background knowledge in the news article is of greater importance than the objects and faces in the image. Fourth, the captioning performance degrades when the edges between the image sub-graph and the text sub-graph are all removed or all connected, which validates that associating the named entities to the corresponding objects helps to generate more accurate captions. Finally, our full model performs best when the text sub-graph and the image sub-graph are both encoded, validating the effectiveness of learning the association between the named entities and visual objects.

Dataset	Method	B-4	R	C
GoodNews	w/o graph	5.27	17.32	40.45
	w/o text graph	5.29	17.93	41.26
	w/o image graph	5.72	18.21	43.37
	Ours	5.89	19.19	45.22
	w/o graph+JoGANIC	5.82	19.10	42.03
	w/o text graph+JoGANIC	5.97	19.30	42.34
	w/o image graph+JoGANIC	5.99	19.21	42.45
	Ours+JoGANIC	**6.03**	**19.85**	**45.97**
NYTimes800k	w/o graph	5.37	17.23	42.83
	w/o text graph	5.41	18.10	42.82
	w/o image graph	5.74	18.21	46.59
	Ours	5.97	19.20	47.27
	w/o graph+JoGANIC	5.45	18.04	43.27
	w/o text graph+JoGANIC	5.50	18.23	46.93
	w/o image graph+JoGANIC	5.56	18.97	47.98
	Ours+JoGANIC	**5.99**	**19.75**	**49.53**

TABLE IV: Evaluation results of event description on the GoodNews dataset and the NYTimes800k dataset.

Dataset	Named entity recall
GoodNews	78.4
NYTimes800k	76.7

TABLE V: The recall of the named entities extracted from the article text on GoodNews and NYTimes800k.

Dataset	Object F1 score
GoodNews	77.4
NYTimes800k	80.1

TABLE VI: The F1 scores of the objects and faces detected in the images of GoodNews and NYTimes800k.

E. Performance of Generating Different Types of Captions

Since multiple types of named entities are relevant to the news images, an ideal entity-aware captioning model should generate different types of named entities accurately. To evaluate the effect of the proposed MMKG on generating different types of named entities, we divide the test split of GoodNews and NYTimes800k into a subset where the ground-truth captions involve people's names ("w/ people") and and another subset where the ground-truth captions do not contain people's names ("w/o people"). The test split of GoodNews contains 15,883 images with people and 7,330 images without people,

Dataset	Method	Object Detection	Face Detection	B-4	M	R	C	F1	L-F1
GoodNews	Ours	YOLOv3	MTCNN	6.82	11.42	22.83	60.84	25.81	8.34
	Ours	VinVL	TinaNet	**6.82**	**11.44**	**22.89**	**60.93**	**25.89**	**8.48**
NYTimes800k	Ours	YOLOv3	MTCNN	**6.71**	9.60	22.13	57.97	25.14	6.35
	Ours	VinVL	TinaNet	6.70	**9.62**	**22.17**	**58.20**	**25.23**	**6.44**

TABLE VII: The results of using different object detector and face detector.

respectively. The test split of NYTimes800k consists of 15,099 images that involve people and 6,877 images without people, respectively. From the results in Table III, we observe that compared with "w/o graph", our method makes improvements on both subsets, which indicate that the proposed MMKG is capable of generating captions with people's names as well as describing images without people.

F. Performance of Describing Events

In addition to selecting named entities correctly, an entity aware captioning model is also required to describe the events depicted in the images properly. To evaluate the performance of event description, the influence of named entities should be eliminated. We replace the named entities in the ground-truth captions and the generated captions with corresponding class labels and report the Bleu-4, METEOR, ROUGE-L and CIDEr metrics. From the results in Table IV, we observe that both the image sub-graph and the text sub-graph contribute to event description. The model using the text sub-graph performs better than the model using image sub-graph, which indicates that the named entity relationships in the text sub-graph play a more important role in describing the events in the image.

G. Evaluation of Information Extraction and Object Detection

To better understand the construction of multi-modal knowledge graphs, we evaluate the quality of the intermediate results, namely the output of the information extraction model as well as the object detection and face detection model. We evaluate the quality of information extraction by comparing the triplets extracted from the article text and the ground-truth captions of the images in the article, respectively. The triplets extracted from the ground-truth captions are regarded as the references, and we calculate the recall of the triplets extracted from the article text. The recall on GoodNews and NYTimes800k are shown in Table V. From these results, we observe that the information extraction model captures a large portion of the triplets that are related with the images.

In terms of object and face detection, we randomly select 500 images from both datasets, and ask the human annotators to annotate the bounding boxes of the objects and faces that are relevant to the named entities in the ground-truth captions. The detected objects and faces that have an IoU (intersection over union) larger than 0.8 with the annotated bounding boxes are considered positive samples, and we calculate the F1 score of all the detected objects. The results in Table VI indicates that the object detection model and face detection model effectively detects the objects and faces in the images.

To evaluate the effect of object detection model and face detection model on entity-aware captioning, we conduct additional experiments by using the object detector from VinVL [55] and face detector TinaNet [56] to construct the image sub-graph. The experiments are conducted using Transformer as the decoder, and the results on GoodNews and NYTimes800k are shown in Table VII. We observe that by using VinVL and TinaNet, the performance of generating entity-aware captions is improved in terms of the standard evaluation metrics and the F1 scores of named entities, indicating that using more advanced object detector and face detector helps to generate more accurate entity-aware captions.

H. Qualitative Results

We show some examples of the generated entity-aware captions in Figure 5. As illustrated in the figure, our model correctly selects multiple types of named entities inculding persons, places and times. For example, in Figure 5(b), our method correctly describes the person name in the image, while the model without MMKG uses another person name in the article. The cases where multiple named entities are related to the image is also well handled by our model, thanks to the modeling of entity relationships using the MMKGs. For instance, in Figure 5(e) where the image shows multiple persons, our model successfully identifies all the persons. It is also interesting to observe that even if the visual objects are not clear, our method still works well, which further validates the advantage of the entity relationship captured by the multi-modal graph using the external knowledge. Taking Figure 5(d) for example, though the Renault's chairman Jean-Dominique Senard (the person on the left) is shadowed and is difficult to recognize, our model correctly describes both person names using the relationships involving him and the Nissan's executive, Hitoro Saikawa (the person on the right). Apart from the images that mainly depicts people, our method also describes the images that only involves the common objects (Figure 5 (c) and (f)) accurately. For instance, our method associates the football in the image to "The Brazuca" in the article text and utilizes the facts about the football in the article to generate accurate description.

The examples of multi-modal knowledge graphs are shown in Figure 7. From these figures, we observe that the multi-modal knowledge graphs are capable of modeling the fine-grained relationship between named entities in the news article. For instance, in Figure 7(a), though the image only shows one person (Novak Djokovic), the relationship between the person and another person (Alexander Zverev) is captured by the text sub-graph. In Figure 7(b), the multi-modal knowledge

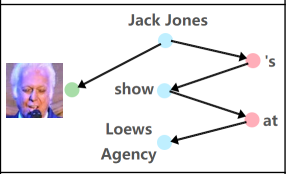
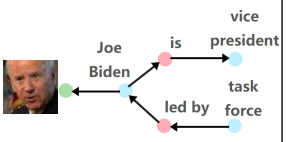
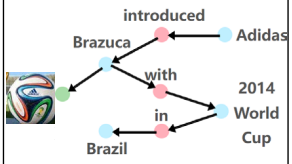
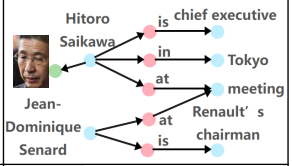
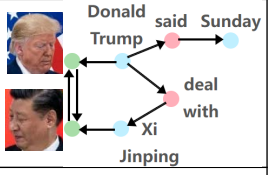
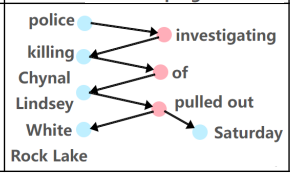
	Image	Captions	Generated MMKG
(a)		**Ground-truth:** Jack Jones performing at Feinstein 's at Loews Regency. **w/o graph:** Jack Jones (√)performing at Lolette Hotel (×) on Wednesday night. **Ours:** Jack Jones (√) played his show at Feinstein's (√) at Loews Regency (√) on Wednesday. **Ours+JoGANIC:** Jack Jones (√) performing Feinstein's (√) at Loews Regency (√) on Wednesday night.	
(b)		**Ground-truth:** Vice President Joseph R. Biden Jr. on Thursday at one of the meetings he conducted to gather information on gun violence. **w/o graph:** President Obama (×) spoke about the ban on gun control at the White House on Thursday (√). **Ours:** Vice President Joseph R. Biden Jr. (√) on Thursday (√) at the White House. **Ours+JoGANIC:** Vice President Joseph R. Biden Jr. (√) at a meeting on Thursday (√).	
(c)		**Ground-truth:** Adidas on Tuesday unveiled the Brazuca, the official match ball of the 2014 World Cup in Brazil. **w/o graph:** The Brazuca (√), designed for the 2014 World Cup (√) , will be used in September 2012 (×). **Ours:** The Brazuca (√) will be used for the 2014 World Cup (√) in Brazil (√). **Ours+JoGANIC:** The Brazuca (√) is the official match ball for the 2014 FIFA World Cup (√)	
(d)		**Ground-truth:** Renault 's chairman , Jean - Dominique Senard , left , and Nissan 's chief executive , Hiroto Saikawa , in Tokyo on Thursday . **w/o graph:** Nosuki Ghosn (×), the chief executive of Nissan, in Tokyo last month (×). **Ours:** The Renault chief, Jean - Dominique Senard (√), and Hitoro Saikawa (√), the chief executive of Nissan, at the company 's headquarters in Tokyo (√) on Thursday (√). **Ours+JoGANIC**: Hiroto Saikawa (√), Nissan's former chairman, at a news conference in Tokyo(√).	
(e)		**Ground-truth:** President Trump and President Xi Jinping of China at a bilateral meeting at the Group of 20 in Osaka , Japan, in June . **w/o graph:** President Xi Jinping (√) of China in Beijing on Monday. **Ours:** President Trump (√) and President Xi Jinping (√) of China at the Japan (√) on Sunday. **Ours+JoGANIC:** President Trump with President Xi Jinping of China at the Group of 20 summit meeting in Paris on Saturday.	
(f)		**Ground-truth:** White Rock Lake in Dallas , where the body of a transgender woman was recovered by the police on Saturday. **w/o graph:** A view of White Rock Lake (√) in Dallas (√) on Monday (×). **Ours:** A view of the White Rock Lake (√), where Ms. Lindsey (√) was found dead on Saturday (√). **Ours+JoGANIC:** White Rock Lake in Dallas, where Ms. Lindsey was found dead on Saturday.	

Fig. 5: Qualitative results on the GoodNews dataset (a, b, c) and the NYTimes800k dataset (d, e, f). "Ground-truth", "w/o graph", "Ours" and "Ours+JoGANIC" denote the ground-truth caption, the caption generated without the MMKG, the caption generated using MMKG, and the caption generated using the MMKG together with the decoder of "JoGANIC", respectively. The named entities in the captions are colored and underlined. Due to space limitation, the rightmost column only shows part of the constructed MMKGs. √ and × denote the named entities that are described correctly and the named entities that are irrelevant to the image, respectively. The named entities, relationships and the visual objects in the multi-modal knowledge graph are represented by blue dots, red dots and green dots, respectively.

graph correctly captures the relationship between the person (Sebastian Vettel) and his racing team (Red Bull Infiniti).

I. Discussion on Limitations

In this section, we analyze the failure cases in the generated captions and discuss the limitations of the proposed method. Some failure cases are shown in Figure 6, where the captioning model generates incorrect human names for different reasons. In Figure 6(a), the face detection model fails to detect a human face (corresponding to "Prime Minister Wen Jiabao"), and the named entity is not connected to the human face in the multi-modal knowledge graph. This issue can be alleviated by using more advanced face detection model. In Figure 6(b), though the human face in the image is detected, the cross modal entity matching module does not connect the named entity ("Ronald Westbrook") to the correct human face since the image of the named entity is not available in the external knowledge base. We are going to address this issue by collecting a larger knowledge base that covers a wider range of named entities, or introducing knowledge graph completion methods to reason about the relationship between the named entities and the visual objects that are unknown to the cross-modal entity matching module. In Figure 6(c) and (d), the small human faces are not detected by the face detection model. In such cases, the captioning model either misses some human names or generate wrong human names, which can also be addressed by using better face detection models.

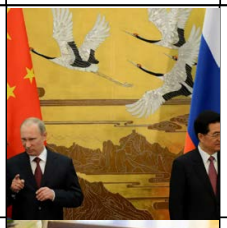
	Image	Captions
(a)		**Ground-truth:** Chancellor Angela Merkel of Germany with Prime Minister Wen Jiabao during a welcoming ceremony in Beijing on Thursday. **w/o graph:** President Xi Jinping (×) of China, center, and other European Union leaders in Beijing (√) on Thursday (√). **Ours:** Chancellor Angela Merkel (√) of Germany and President Emmanuel Macron (×) of France at a news conference in Beijing (√) on Thursday.
(b)		**Ground-truth:** Ronald Westbrook, 72, was a retired Air Force officer. **w/o graph:** Joe Hendrix (×), a retired Air Force officer, at a news conference in Atlanta on Friday. **Ours:** Joe Hendrix (×), a retired Air Force officer, was shot to death on Nov. 27.
(c)		**Ground-truth:** Ford's development of the Mustang under Mr. Iacocca first put him in the public eye. He rode in one in 1965 between Donald N. Frey, left, who led the Mustang design and engineering work, and Henry Ford II, the automaker's powerful chairman. **w/o graph:** Mr. Iacocca (√)'s car, shown in an undated photo. **Ours:** Mr. Iacocca (√) on the assembly line at the Ford (√) plant in Detroit in 1979. Mr. Iacocca helped transform big business's relationship with Washington.
(d)		Ground-truth: Chris Urmson, left, who previously led a self-driving car project at Google, and Sterling Anderson, formerly of Tesla, at the Aurora office in Palo Alto, Calif., last year. w/o graph: Mike Manley (×), chief executive of Fiat Chrysler, at the company's headquarters in San Francisco (×). Ours: Mike Manley (×), chief executive of Fiat Chrysler, at Palo Alto, Calif. (√)

Fig. 6: Failure cases on the GoodNews dataset (a and b) and the NYTimes800k dataset (c and d), where the errors are marked in cross (×). In the first case, the face detection model fails to detect an occluded face and the captioning model is uncertain about the human name. In the second case, the visual information related to the named entity "Ronald Westbrook" is not available in the external knowledge base and the cross-modal entity matching module fails to connect the named entity in the text sub-graph to the human face in the image sub-graph. In the last two cases, the human faces are too small to detect, and the captioning model misses some human names or generates wrong human names.

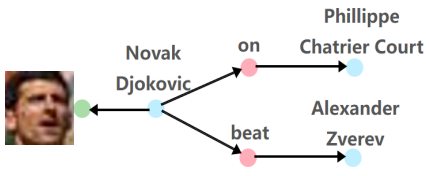
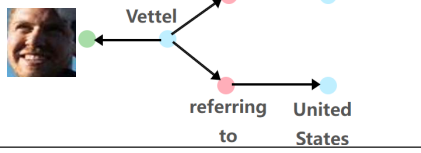
	Image	News Article	Multi-modal Knowledge Graph	Captions
(a)		…… In a match that lasted 22 minutes longer, Djokovic beat fifth-seeded Alexander Zverev on Philippe Chatrier Court, 7-5, 6-2, 6-2. Djokovic and Thiem will play at Roland Garros for the third time in four years. ……		**GT:** Novak Djokovic defeated Alexander Zverev at the French Open on Thursday. **Ours:** Novak Djokovic (√) during his 7-5, 6-2, 6-2 victory over Alexander Zverev (√) on Thursday (√) at the French Open.
(b)		Many drivers in the glamour-rich, cash-poor sport attribute Vettel's success to his Red Bull Infiniti partnership, as if he were behind the wheel of a Brink's truck. …… "You have low expectations when you come here," Vettel said, referring to the United States, "because you don't know how much people know about Formula One." ……		**GT:** Sebastian Vettel, seeking a record eighth straight Formula One victory, will race in the United States Grand Prix on Sunday. **Ours:** Sebastian Vettel (√) of Red Bull Infiniti at the United States (√) Grand Prix.

Fig. 7: Qualitative results of generated multi-modal knowledge graphs on the GoodNews dataset (a) and the NYTimes800k dataset (b). In the multi-modal knowledge graphs, the named entities, relationships and the visual objects are represented by blue dots, red dots and green dots, respectively.

Dataset	Method	Event score
GoodNews	Transform and Tell [12]	2.14
	Ours	2.46
	Ours+JoGANIC	2.58
NYTimes800k	Transform and Tell [12]	2.20
	Ours	2.49
	Ours+JoGANIC	2.73

TABLE VIII: The average scores given by the human annotators that reflect the quality of the events described by the entity-aware captions on GoodNews and NYTimes800k.

J. Human Evaluation

To evaluate the quality of the entity aware captions more thoroughly, we recruit 20 volunteers in total and perform human evalaution on 200 images in the GoodNews dataset. We first evaluate the quality of the generated captions by showing the news article, the image and the generated caption and asking the human annotators to rate the event described in the caption by giving a "event score" from 0 to 3 (a higher score is better). Each caption is scored by at least two annotators. From the average event scores in Table VIII, we observe that "Ours" performs better than "Transform and Tell" in describing the events depicted by the image, and "Ours+JoGANIC" further improves the performance.

In addition, we also evaluate the quality of the multi-modal knowledge graphs using a separate human evalaution interface. The named entities and the corresponding visual objects in the multi-modal knowledge graphs are shown, and the annotators are asked to rate whether the visual objects are matched to the correct named entities. For the captions with different event scores, we calculate the portion of the correctly matched visual objects in the corresponding multi-modal knowledge graphs. The distribution of the correctly matched visual objects with respect to the event score of the corresponding entity-aware captions is shown in Figure 8. We observe that for the captions that have higher event score, the portion of the correctly matched visual objects in the corresponding multi-modal knowledge graphs is larger, which indicates that the multi-modal knowledge graphs of high quality helps the captioning model to generate better entity-aware captions.

V. Conclusion

We present a novel entity-aware image captioning method that constructs a MMKG by exploring external knowledge from the web. Our method can simultaneously associate visual cues with named entities and capture the fine-grained relationships between named entities, thus succeeding in extracting accurate entities and refining concrete events. The proposed method first constructs a text sub-graph that consists of the named entities and their relationships, and an image sub-graph containing the visual objects in the image, and then connects the similar named entities and visual objects using a cross-modal entity matching module. Extensive experiments on two large-scale entity-aware image captioning datasets, GoodNews and NYTimes800k, demonstrate the effectiveness of the proposed method. In the future, we are going to investigate more advanced caption decoders and linguistic features to further improve the captioning performance.

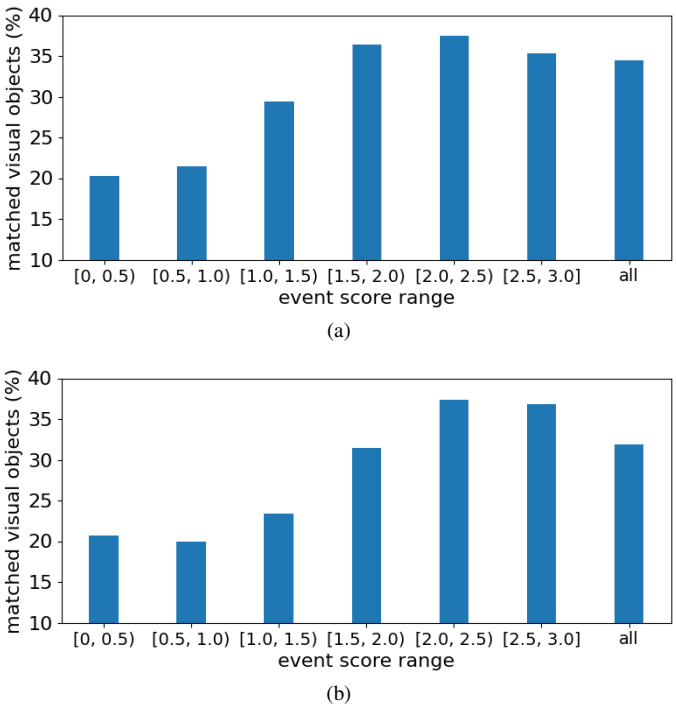


Fig. 8: The distribution of the correctly matched visual objects with respect to the event score on the GoodNews dataset (a) and the NYTimes800k dataset (b). The label under each bar denotes a range of event score, and the height of each bar reflects the portion of the correctly matched visual objects in the MMKGs corresponding to the captions whose event scores lie in the range.

VI. Acknowledgment

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 62072041.

References

[1] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *CVPR*, 2015.
[2] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *CVPR*, 2015.
[3] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *CVPR*, 2018.
[4] L. Huang, W. Wang, J. Chen, and X.-Y. Wei, "Attention on attention for image captioning," in *ICCV*, 2019.
[5] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, "Meshed-memory transformer for image captioning," in *CVPR*, 2020.
[6] C. Yan, Y. Hao, L. Li, J. Yin, A. Liu, Z. Mao, Z. Chen, and X. Gao, "Task-adaptive attention for image captioning," *IEEE Transactions on Circuits and Systems for Video technology*, vol. 32, no. 1, pp. 43–51, 2021.
[7] Y. Wang, J. Xu, and Y. Sun, "End-to-end transformer based model for image captioning," *arXiv e-prints*, pp. arXiv–2203, 2022.
[8] J. S. Wynn, J. D. Ryan, and M. Moscovitch, "Effects of prior knowledge on active vision and memory in younger and older adults." *Journal of Experimental Psychology: General*, 2020.

[9] A. Ramisa, F. Yan, F. Moreno-Noguer, and K. Mikolajczyk, "Breakingnews: Article annotation by image and text processing," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2017.

[10] D. Lu, S. Whitehead, L. Huang, H. Ji, and S.-F. Chang, "Entity-aware image caption generation," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 4013–4023.

[11] A. F. Biten, L. Gomez, M. Rusinol, and D. Karatzas, "Good news, everyone! context driven entity-aware captioning for news images," in *CVPR*, 2019.

[12] A. Tran, A. Mathews, and L. Xie, "Transform and tell: Entity-aware news image captioning," in *CVPR*, 2020.

[13] A. Hu, S. Chen, and Q. Jin, "Icecap: Information concentrated entity-aware image captioning," in *ACM Multimedia*, 2020.

[14] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *ICML*, 2015.

[15] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[16] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua, "Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5659–5667.

[17] Z.-J. Zha, D. Liu, H. Zhang, Y. Zhang, and F. Wu, "Context-aware visual policy network for fine-grained image captioning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 2, pp. 710–722, 2019.

[18] T. Yao, Y. Pan, Y. Li, and T. Mei, "Exploring visual relationship for image captioning," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 684–699.

[19] X. Yang, K. Tang, H. Zhang, and J. Cai, "Auto-encoding scene graphs for image captioning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 10 685–10 694.

[20] J. Wang, W. Wang, L. Wang, Z. Wang, D. D. Feng, and T. Tan, "Learning visual relationship and context-aware attention for image captioning," *Pattern Recognition*, vol. 98, p. 107075, 2020.

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[22] G. Li, L. Zhu, P. Liu, and Y. Yang, "Entangled transformer for image captioning," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 8928–8937.

[23] Y. Luo, J. Ji, X. Sun, L. Cao, Y. Wu, F. Huang, C.-W. Lin, and R. Ji, "Dual-level collaborative transformer for image captioning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 3, 2021, pp. 2286–2293.

[24] S. Zhao, P. Sharma, T. Levinboim, and R. Soricut, "Informative image captioning with external sources of information," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 6485–6494.

[25] Z. Yang and N. Okazaki, "Image caption generation for news articles," in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 1941–1951.

[26] X. Yang, S. Karaman, J. Tetreault, and A. Jaimes, "Journalistic guidelines aware news image captioning," in *EMNLP*, 2021.

[27] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," *arXiv preprint arXiv:1508.07909*, 2015.

[28] J. Zhang, S. Fang, Z. Mao, Z. Zhang, and Y. Zhang, "Fine-tuning with multi-modal entity prompts for news image captioning," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 4365–4373.

[29] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[30] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 7871–7880.

[31] R. Xie, Z. Liu, H. Luan, and M. Sun, "Image-embodied knowledge representation learning," in *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 2017, pp. 3140–3146.

[32] H. Mousselly-Sergieh, T. Botschen, I. Gurevych, and S. Roth, "A multimodal translation-based approach for knowledge graph representation learning," in *SEM@NAACL-HLT*, 2018.

[33] A. V. Kannan, D. Fradkin, I. Akrotirianakis, T. Kulahcioglu, A. Canedo, A. Roy, S.-Y. Yu, M. Arnav, and M. A. Al Faruque, "Multimodal knowledge graph for deep learning papers and code," in *CIKM*, 2020.

[34] R. Sun, X. Cao, Y. Zhao, J. Wan, K. Zhou, F. Zhang, Z. Wang, and K. Zheng, "Multi-modal knowledge graphs for recommender systems," in *CIKM*, 2020.

[35] X. Shu, G.-J. Qi, J. Tang, and J. Wang, "Weakly-shared deep transfer networks for heterogeneous-domain knowledge propagation," in *Proceedings of the 23rd ACM international conference on Multimedia*, 2015, pp. 35–44.

[36] J. Tang, X. Shu, Z. Li, G. Qi, and J. Wang, "Generalized deep transfer networks for knowledge propagation in heterogeneous domains," *ACM Trans. Multim. Comput. Commun. Appl.*, vol. 12, no. 4s, pp. 68:1–68:22, 2016. [Online]. Available: https://doi.org/10.1145/2998574

[37] R. Du, J. Xie, Z. Ma, D. Chang, Y.-Z. Song, and J. Guo, "Progressive learning of category-consistent multi-granularity features for fine-grained visual classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 12, pp. 9521–9535, 2021.

[38] Y. Guo, R. Du, X. Li, J. Xie, Z. Ma, and Y. Dong, "Learning calibrated class centers for few-shot classification by pair-wise similarity," *IEEE Transactions on Image Processing*, vol. 31, pp. 4543–4555, 2022.

[39] X. Shu, J. Yang, R. Yan, and Y. Song, "Expansion-squeeze-excitation fusion network for elderly activity recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 8, pp. 5281–5292, 2022.

[40] V. I. Spitkovsky and A. X. Chang, "A cross-lingual dictionary for english wikipedia concepts," 2012.

[41] C. D. Manning, M. Surdeanu, J. Bauer, J. R. Finkel, S. Bethard, and D. McClosky, "The stanford corenlp natural language processing toolkit," in *ACL*, 2014.

[42] I. Yamada, A. Asai, J. Sakuma, H. Shindo, H. Takeda, Y. Takefuji, and Y. Matsumoto, "Wikipedia2vec: An efficient toolkit for learning and visualizing the embeddings of words and entities from wikipedia," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2020, pp. 23–30.

[43] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[44] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, 2016.

[45] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[46] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *CVPR*, 2015.

[47] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[48] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *International Conference on Learning Representations*, 2018.

[49] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *ACL*, 2002.

[50] M. Denkowski and A. Lavie, "Meteor universal: Language specific translation evaluation for any target language," in *Proceedings of the ninth workshop on statistical machine translation*, 2014.

[51] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004.

[52] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *CVPR*, 2015.

[53] M. Honnibal and I. Montani, "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing," 2017, to appear.

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

[55] P. Zhang, X. Li, X. Hu, J. Yang, L. Zhang, L. Wang, Y. Choi, and J. Gao, "Vinvl: Revisiting visual representations in vision-language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5579–5588.

[56] Y. Zhu, H. Cai, S. Zhang, C. Wang, and Y. Xiong, "Tinaface: Strong but simple baseline for face detection," *arXiv preprint arXiv:2011.13183*, 2020.