MAGIC: Multimodal relational Graph adversarial inference for Diverse and Unpaired Text-Based Image Captioning

Wenqiao Zhang 1*, Haochen Shi 2*, Jiannan Guo 1*, Shengyu Zhang 1 Qingpeng Cai 3, Juncheng Li 1†, Sihui Luo 4, Yueting Zhuang 1†

1 Zhejiang University, 2 Université de Montréal, 3 National University of Singapore, 4 Ningbo University
{wenqiaozhang, jiannan, sy_zhang junchengli, yzhuang}@zju.edu.cn, haochen.shi@umontreal.ca, qingpeng@comp.nus.edu.sg, luosihui@nbu.edu.cn

Abstract

Text-based image captioning (TextCap) requires simultaneous comprehension of visual content and reading the text of images to generate a natural language description. Although a task can teach machines to understand the complex human environment further given that text is omnipresent in our daily surroundings, it poses additional challenges in normal captioning. A text-based image intuitively contains abundant and complex multimodal relational content, that is, image details can be described diversely from multiview rather than a single caption. Certainly, we can introduce additional paired training data to show the diversity of images’ descriptions, this process is labor-intensive and time-consuming for TextCap pair annotations with extra texts. Based on the insight mentioned above, we investigate how to generate diverse captions that focus on different image parts using an unpaired training paradigm. We propose the Multimodal relational Graph adversarial inference (MAGIC) framework for diverse and unpaired TextCap. This framework can adaptively construct multiple multimodal relational graphs of images and model complex relationships among graphs to represent descriptive diversity. Moreover, a cascaded generative adversarial network is developed from modeled graphs to infer the unpaired caption generation in image–sentence feature alignment and linguistic coherence levels. We validate the effectiveness of MAGIC in generating diverse captions from different relational information items of an image. Experimental results show that MAGIC can generate very promising outcomes without using any image–caption training pairs.

Introduction

Visual captioning aims to help visually impaired people to quickly understand the image content. Existing methods focus on describing visual objects that have achieved impressive or even super-human performance (Yu et al. 2019; Cornia et al. 2020; Zhang et al. 2021a; Ding et al. 2020; Zhang et al. 2021a,b). However, most prominent captioning methods fail to “read” the associated texts in the visual scene. Therefore, text-based image captioning (TextCap) (Sidorov et al. 2020) is proposed in this study to caption models, recognize visual content, read text, understand both modalities jointly, and write a succinct caption. Such task is very practical that encourages machines to generate fine-grained captions (e.g., read the price of commodities in Fig.1), which is helpful to visually disabled people to understand surroundings more accurately, but put forward new difficulties to existing captioning models meanwhile.

Intuitively, human-made images with text contain more semantic entities (e.g., objects, attributes, and texts) and complex relations (e.g., correlation and constraint) than natural images (Hou et al. 2019; Zhang et al. 2020b,a; Li et al. 2022). Common single captioning models experience difficulty in representing a holistic image understanding. These models also hinder the diversity of image descriptions, which is crucial for blind people concerning different aspects of an image (e.g., price comparison, discount, brand and position of commodities in Fig.1). In fact, we humans tend to focus on different aspects of image content by our preferences, personality and sentiment, this applies to the annotators with no exception (Yang, Zhang, and Cai 2020b). Definitely, we could ask annotators for more paired data (e.g., multiview of an image - diverse descriptions), thereby force captioning model to generate varied captions from multiple views of an image. However, in the Textcap task as well as many vision-language applications, such large-scale annotations are not readily available, and are both time-consuming and labor-intensive to acquire. In these scenarios, unsupervised/unpaired methods (Lample et al. 2017; Gu et al. 2019; Caron et al. 2020) that can
learn vision to language or conversely from unpaired training data are highly desirable. Based on this insight, our goal is to break away from single and paired captioning conventions and conduct the unpaired captioning paradigm, which encourages TextCap models to generate diverse textual descriptions to comprehend images better.

However, the diverse and unpaired TextCap is challenging and can be summarized as three aspects: 1) **Descriptive Diversity**, as shown in Fig.1, three sentences focus on different and interrelated multimodal information. It is hard to determine which fine-grained multimodal details are semantically related and ignore the unconcerned content to mine the diversity of image content from multiple views; 2) **Relational Modeling**, how to model complex intra- and cross-relationships among the above multimodal contents, some of which even seem irrelevant intuitively. For instance, there are various relational semantics in Fig.1: (a) *intra-relationship*, a price of “17.88” denotes the “rollback” while “19.88” indicates the original price, a lady is located next to the commodity; (b) *cross-relationship*, the word “Hydr oxycut” is the brand of a commodity and its price of “17.88” is located above a lady. Thus, appropriately model the relational semantic meanings is the premise to generate a plausible caption; 3) **Unpaired Learning Paradigm**, there have been few attempts for unpaired captioning (Gu et al. 2018, 2019), mainly relies on aligning/sharing latent semantic space across the image-sentence domain. Although promising, these methods may not mature enough to apply to unpaired Textcap. The multimodal relationships among images are complex and abundant. The characteristics distribution among images and descriptions are apparently larger than general captioning.

Overall, we propose a novel method for diverse and unpaired TextCap called Multimodal relAtional Graph adversarial inferenCe (MAGIC) to address these challenges. As shown in Fig.1, our framework comprises the following parts: 1) **Multimodal Relational Graph Encoder**, this module aims to encode an image to multiple semantic embeddings based on content diversity adaptively. The intuition behind this is: when human beings composing different image descriptions, they prefer to focus on a few salient regions, then query the contextually relevant content part and clarify their relationships to describe. To mimic this process, we first introduce a central object-aware pool that adaptively identifies the key objects which are supposed to reflect humans’ intentions. Then construct their multimodal relational graphs via query relation-aware multimodal semantic meaning. The relational contexts among graphs are finally modeled by GCN (Berg, Kipf, and Welling 2017) into embeddings to represent varying levels of details of an image; 2) **Sentence Auto-Encoder**, this module aims to reconstruct the sentences by encoding the pre-extracted scene graph from the sentence corpus. Encoded embedding sets can learn the language inductive bias (Yang, Zhang, and Cai 2020a) from the sentence corpus to provide explicit signals for subsequent unpaired mapping and decode lexical words appropriately; 3) **Unpaired Adversarial Caption Inference**, last but not least, semantic consistency throughout image and sentence domain is crucial to achieving unpaired learning. For this purpose, we develop a cascaded generative adversarial network(CGAN). CGAN first performs feature alignment from image to sentence to capture the cross-modal semantic consistency, then learns linguistic coherence by a language discriminator to generate more accurate and fluent captions. To summarize, the major contributions of our paper are as follows:

- A novel perspective is presented for the TextCap task to simultaneously liberate from the single and paired captioning training mode.
- The MAGIC framework can appropriately construct and encode diverse multimodal relational graphs to observe an image from multiview and learn to generate unpaired captions via cascaded adversarial inference.
- The superiority of MAGIC is demonstrated in terms of both accuracy and diversity using automatic metrics and human judgments.

**Related Work**

**Diverse Captioning.** The ability of understanding and reasoning different modalities is a longstanding and challenging goal of artificial intelligence (Rennie et al. 2017; Zhang et al. 2020a, 2019; Li et al. 2021, 2020, 2019; Chen et al. 2021). Image captioning is one of the most important tasks and recent works are focus on generating multiple captions to describe an image. (Johnson, Karpathy, and Fei-Fei 2016) is the first to propose the dense captioning task, which detects and describes diverse regions in the image. (Guo et al. 2019) design a multi-style image captioning model that learns to map images into attractive captions of multiple styles. (Xu et al. 2021) develops anchor-centered graphs that select different textual contents to generate that multiple captions. However, the works mentioned above are more relevant to holistic visual understanding by paying more attention to unimodal information (e.g., objects, regions or texts). They require much paired data to learn the captioning diversity and may hardly model an image’s complex multimodal relational details.

**Unpaired Captioning.** First attempt (Gu et al. 2018) of unpaired image captioning utilize pivot language corpus and align it to the target language, connect the pivot language sentences in different domains by shared word embeddings. (Gu et al. 2019) uses scene graph as representations of the image and the sentence, and maps the scene graphs in their feature space through cycle-consistent adversarial training. The most recent work (Liu et al. 2021) employs the semantic relationships to bridge the gap between vision and language domains for unpaired caption generation. However, due to the complex multimodal relationships in TextCap task, the discrepancy of feature distribution and semantic gap across image-sentence domain is larger than image captioning. The above methods may fail to generate accurate captions appropriately in an unpaired setting.

**Method**

This section describes our diverse and unpaired TextCap framework, we will describe each module of it and introduce the objectives and strategy for training.
Problem Formulation

Before presenting our method, we first introduce some basic notions and terminologies. Given a text-based image \( I \), the goal is to generate diverse natural language sentences \( S = \{S_i\}_{i=1}^{N_s} \) with unpaired learning, where \( N_s \) is the predefined number of generated sentences to satisfy the user’s intention. We use the following training pipeline \( C(\cdot, \theta) \) to show how the diverse and unpaired captions are generated:

\[
C((S|I); (\theta_T, \theta_S)) = E_T((G^T|I); \theta_T) \quad U_A((G^S|G^T); \theta_T) \quad D_S((S|G^S); \theta_S)
\]

(1)

Where \( E_T(\cdot; \theta) \) is image encoder aims to model \( N_b \) multimodal relational graphs (MRGs) \( G^T = \{G^T_i\}_{i=1}^{N_b} \) learned from image \( I \), then translate \( G^T \) to appropriate sentence graph \( G^S \) through unpaired cross-domain alignment \( U_A(\cdot; \theta) \), the aligned results can be decoded to diverse captions \( S = \{S_i\}_{i=1}^{N_s} \) by sentence decoder \( D_S(\cdot; \theta) \). Notice that \( E_T(\cdot; \theta) \) and \( U_A(\cdot; \theta) \) share parameters \( \theta_T \) to learn feature alignment across domain, \( D_S(\cdot; \theta) \) is a pretrained sentence decoder from textual corpus, \( \theta_S \) are fixed as explicit supervised signals to learn \( \theta_T \) to generate unpaired captions.

Fig.2 shows the proposed Multimodal relational Graph adversarial interference (MAGIC) framework \( C(\cdot, \theta) \) consists of three main modules: Multimodal Relational Graph Encoder \( E_T(\cdot; \theta) \), Sentence Auto-Encoder \( D_S(\cdot; \theta) \) and Unpaired Caption Adversarial Inference \( U_A(\cdot; \theta) \). We use the TextCap task as an example to illustrate our method.

Multimodal Relational Graph Encoder

This module aims to select the central objects dynamically as the central object-aware pool, then constructs their MRGs adaptively, which can represent the different fine-grained contents. Last, GCN model intra- and cross-relationships to learn MRG embeddings.

Multimodal Feature Extractor. Given a text-based image \( I \), we firstly initialize it as its corresponding \( N_v \) visual object features and \( N_t \) text token features by Faster RCNN (Ren et al. 2016) and Rosetta OCR (Borisuyk, Gordo, and Sivakumar 2018), respectively. In order to represent visual and text features at a unified level, follow (Sidorov et al. 2020), we adopt a single-layer MLP to map the above features as \( d \)-dimensional vectors, the unified visual and text embedding sets of image \( I \) given by \( \{o_i\}_{i=1}^{N_v} \) and \( \{t_j\}_{j=1}^{N_t} \), respectively.

Central Object-Aware Pool. Inspired by the human-being attention mechanism to describe an image, which usually focuses on a salient object or region, then from “one-to-many”, queries its relevant information and ignores the unconcerned contents to compose fine-grained sentences. We imitate this process by introducing a central object-aware pool (COP) that is supposed to reflect the human preference.
and select objects that may need additional attention adaptively. In detail, given the $i$th object embedding $\mathbf{o}_i$, we employ soft attention to each text token embedding in $\mathbf{T}$. Then $\mathbf{o}_i$ and attended textual latent vector $\mathbf{t}_i^a$ are concatenated as the augmented object embedding $\mathbf{o}_i^a = [\mathbf{o}_i, \mathbf{t}_i^a]$, the embedding set is given by $\mathbf{O}_a = \{\mathbf{o}_i^a|_{i=1}^{N_k}\}$. To decide the central objects from $\mathbf{O}_a$, we adopt a MLP to predict a score for each object:

$$S^c_{object} = \text{softmax}(\mathbf{f}_m(\mathbf{O}_a))$$

(2)

$S^c_{object}$ contains $N_k$ scores for each element that indicates the importance weight of each object. Then we select $N_k$ objects $\mathbf{O}^c = \{\mathbf{o}_i^a|_{i=1}^{N_k}\}$ with the highest scores as COP.

**Multimodal Relational Graph Embedding.** In analogy to the human inference of contextual information with a given focused object, we first query its intra- and cross-relations for each central object, then construct their corresponding MRGs $\mathbf{G}_k^T = \{\mathbf{G}_k^T|_{i=1}^{N_k}\}$ to represent the diversity of image $I$. In detail, for the $k$th central object $\mathbf{o}_k$, its MRG defined as $\mathbf{G}_k^T = (\mathbf{N}_T^T, \mathbf{E}_T)$. We devote $\mathbf{N}_T^T$ as three relational types according to the contextual information: relationship nodes $\mathbf{n}^r$ attribute nodes $\mathbf{n}^a$ and text nodes $\mathbf{n}^t$. On the one hand, intra-relational nodes $\mathbf{n}^r$ and $\mathbf{n}^a$ corresponds to the visually neighbour objects has relationships, attributes for object $\mathbf{o}_k^a$, respectively.

$$\mathbf{N}_T^T\{\mathbf{n}^r, \mathbf{n}^a\} = \{\{\mathbf{W}_o^r\mathbf{a}_{i=1}^{N_a}\}\}, \text{relationship nodes};$$

$$\mathbf{N}_T^T\{\mathbf{n}^a\} = \{\{\mathbf{W}_o^a\mathbf{a}_{i=1}^{N_a}\}\}, \text{attribute nodes}; (3)$$

where $\mathbf{W}_o^r \in \mathbb{R}^{d \times d}$ and $\mathbf{W}_o^a \in \mathbb{R}^{d \times (d+4)}$ are the relational embedding matrix, we omit the bias term for simplicity. $\mathbf{s}_i$ the positional embedding to distinguish different attribute nodes connected with the object $\mathbf{o}_k^a$.

On the other hand, to capture the cross-relationship among central object $\mathbf{o}_k^a$ and text tokens, we perform soft attention weight on texts $\mathbf{T}$ to query the semantically textual embeddings as text nodes $\mathbf{n}^t$.

$$\mathbf{N}_T^T\{\mathbf{n}^t\} = \{\mathbf{a}^t_{i=1}^{N_t}\}, \text{text nodes}. (4)$$

where $\mathbf{a}^t_{i}$ is the attention score, $\epsilon$ is a pre-defined threshold. $\text{sum}(\cdot)$ is counting function aims to connect more related contextual texts for $\mathbf{o}_k^a$ if attention score $\mathbf{a}^t_{i}$ greater than $\epsilon$.

Thus far, the MRG $\mathbf{G}_k^T = (\mathbf{N}_T^T, \mathbf{E}_T)$ with three relational nodes $\{\mathbf{n}^r, \mathbf{n}^a, \mathbf{n}^t\}$ is obtained. To model the semantic relations among the above MRG, we introduce a multi-layer GCN (Marcheggiani and Titov 2017; Guo et al. 2021), which can capture mutual relations between neighbour nodes and embed the graph structure into latent vectors.

$$\mathbf{g}_k^T = \delta_r(W_{m_0}\mathbf{o}_k^a + \sum_{\mathbf{n}_i^r \in \mathbf{N}_T^T} \mathbf{W}_{m_1}\mathbf{n}_i)$$

(5)

where $\delta_r$ is the ReLU activation function, $W_{m_0}$ and $W_{m_1}$ $\in \mathbb{R}^{d \times d}$ are the transformation matrices. In this multi-layer GCN, first layer brings contexts for each node from its direct neighbor nodes. Meanwhile, stacking multiple layers enables the encoding of broad contexts in the graph. After such graph pooling, the MRG embedding $\mathbf{g}_k^T$ for object $\mathbf{o}_k^a$ is obtained. As a consequence, the content diversity of image $I$ can be regarded as the collection $\mathbf{g}_k^T = \{\mathbf{g}_k^T|_{k=1}^{N_k}\}$ of $N_k$ central objects.

**Sentence Auto-Encoder.** We design a sentence auto-encoder that fully capture the language properties by self-reconstructing sentence. In particular, we extract the scene graph $\mathbf{G}^S$ from the sentence corpus, and then decode the sentence from latent vector that is encoded from the $\mathbf{G}^S$ (sentence $\rightarrow \mathbf{G}^S$ $\rightarrow$ latent vector $\mathbf{g}^S$ $\rightarrow$ reconstructed sentence).

**Scene Graph Encoder.** We use the scene graph parser provided by (Anderson et al. 2016) for sentences, where a syntactic dependency tree is built by (Klein and Manning 2003) and then a rule-based method (Schuster et al. 2015) is applied for transforming the tree to a scene graph.

Given a sentence $\hat{S}$ from TextCap corpus, its corresponding scene graph involved the high-level semantic concepts (e.g., objects, relationships between two adjacent objects and attributes of objects) and textual tokens, to align subsequently image-sentence domain better, we encode sentences in a similar way in MRG encoding. In detail, for each conceptual object $\mathbf{c}_i^k$ in $\hat{S}$, we define its relational graph as $\mathbf{G}_k^S = (\mathbf{N}_S^S, \mathbf{E}_S)$, where $\mathbf{N}_S^S$ denote three relational type, relationship nodes $\mathbf{c}_i^k$, attribute nodes $\mathbf{a}_i^k$ and associated textual nodes $\mathbf{e}_i^k$. The global semantic meaning of sentence $\hat{S}$ can be regarded as the collection of object relational graphs $\mathbf{G}_k^S = \{\mathbf{G}_k^S|_{i=1}^{M_k}\}$, where $M_k$ is the object number in sentence $\hat{S}$. Similarly, the latent embedding of sentence is learned by the multi-layer GCN which using same structure in image graph encoding but independent parameters.

$$g^S = \frac{1}{M_k} \delta_r(\sum_{i=1}^{M_k}(\mathbf{W}_{e0}\mathbf{e}_{i} + \sum_{\mathbf{c}_i^k \in \mathbf{G}_k^S} \mathbf{W}_{c1}\mathbf{c}_{i}))$$

(6)

where $\mathbf{W}_{e0}$ and $\mathbf{W}_{c1} \in \mathbb{R}^{d \times d}$. Thus, the scene graph embedding of sentence $\hat{S}$ is given by $\mathbf{g}^S$.

**Sentence Decoder.** The decoder part is essential in the way that it translates the semantic latent vector $\mathbf{g}^S$ into the textual sentence. However, most text tokens in sentence corpus only appear few times. This long-tailed word distribution specifically demonstrates the large variance in text occurring, which is challenging to model using a fixed word vocabulary. For this challenge, following M4C (Sidorov et al. 2020), we adopt different classifiers for common vocabulary and candidate texts to predict words. In detail, to capture long-range dependency in sequence modeling, a general LSTM is employed to predict the word $\hat{y}_t$ in common vocabulary based on the scene graph embedding $\mathbf{g}^S$. Then a dynamic pointer network that makes final word prediction on the basis of text candidates $\mathbf{T}^f$ and $\hat{y}_t$.

$$\hat{y}_t = l^m_{lm}(\text{LSTMDEC}(\mathbf{h}_{t-1}, \mathbf{g}^S)),$$

$$\mathbf{y}^t = \text{argmax}([\hat{y}_t, \text{DPN}(\hat{y}_t, \mathbf{T}^f)]);$$

(7)

where $l^m_{lm}()$ is a linear classifier for common vocabulary, $\text{DPN}()$ denotes the dynamic pointer network. We denote notation $\mathbf{y}_{0:T}$ refers to a sequence of predicted words $\{\mathbf{y}_0, \cdots, \mathbf{y}_T\}$, the captioning loss can be computed by:

$$\mathcal{L}(S; \theta_S) = -\log \frac{1}{T} \sum_{i=0}^{T} \mathbb{P}(\mathbf{y}_i|_{\mathbf{y}_{0:T-1}, \mathbf{g}^S, \mathbf{T}^c}))$$

(8)
Unpaired Adversarial Caption Inference

To achieve this unpaired captioning, we develop a CGAN that contains two modules, namely, cross-domain alignment and language discriminator. We first perform the feature alignment across domains and then distinguish whether the semantic structure of decoded caption is real or fake using the language discriminator. Such a cascaded mechanism can train a captioning model in an unsupervised manner and improve coherence among generated sentences.

Cross-Domain Alignment. Given two unpaired graph embeddings sets $G^I = \{g_i^{Ii}\}_{i=1}^{N_I}$ and $G^S = \{g_i^{Si}\}_{i=1}^{N_S}$ from image and sentence domain, respectively, where $N_I$ and $N_S$ are the capacity of corresponding corpus. The goal of feature alignment is to enforce the $G^I$ to be close to the $G^S$ distribution. We follow (Zhu et al. 2017; Gu et al. 2019) that align feature space through cycle-consistent adversarial training. Differently, that the huge domain discrepancy, we leverage recent advances in Optimal Transport (Peyré, Cuturi et al. 2019) to encourage the cross-modal semantic coherence. In detail, the wasserstein distance (Shen et al. 2018) is employed as the loss value to learn the discrimination.

Implementation Details. We add a unknown token and the sentences are “cut” at a maximum length of 20 words. We also report RE-4 that captures a degree of N-gram repetition of unique N-grams (N=1,2) to the total number of words. We add a Unknown token to replace the words out of the vocabulary list. The vocabulary has 6,736 words, and each word is represented by a 300-dimensional vector.

Training Algorithm

We first benchmark our MAGIC for unpaired and diverse captioning on the TextCaps dataset (Sidorov et al. 2020) to verify its effectiveness, and then discuss MAGIC’s property with controlled studies.

Dataset and Setting

Dataset. The TextCaps dataset collected 28,408 images with 142,040 captions from Open Image V3 dataset. Each image has five captions with verified texts through the Rosetta OCR system (Borisuyk, Gordo, and Sivakumar 2018) and human annotators. Follows TextVQA (Singh et al. 2019), this dataset splits for training (21,953), validation (3,166), and test (3,289) sets.

Implementation Details. The number of visual objects is $N_v = 100$ and the number of OCR tokens is $N_t = 50$. For TextCap generation, we tokenized the texts on whitespace, and the sentences are “cut” at a maximum length of 20 words. We add a Unknown token to replace the words out of the vocabulary list. The vocabulary has 6,736 words, and each word is represented by a 300-dimensional vector.

Metrics. We use the microsoft coco caption evaluation 1, which includes BLEU (Papineni et al. 2002) , METEOR (Denkowski and Lavie 2014), ROUG-L (ROUGE 2004), CIDEr (Vedantam, Lawrence Zitnick, and Parikh 2015) and SPICE (Anderson et al. 2016) to evaluate the captioning results. To evaluate the captioning diversity, we use Div-1 and Div-2 scores (Li et al. 2015) that measure a ratio of unique N-grams (N=1,2) to the total number of words. We also report RE-4 that captures a degree of N-gram repetition (N=4) in a description and SelfCIDEr (Wang and Chan 2019) to study the semantic level diversity.

1https://github.com/tylin/coco-caption/
Comparison of Methods. Diverse and unpaired Textcap is a fire-new task that existing captioning algorithms are not applied to compare. We simplified MAGIC as the baseline and conducted an ablation study to ensure a fair comparison and investigate contributions of individual components in MAGIC using the following versions: ISM-LSTM is a simplified version of MAGIC that maps image features from the image corpus to the sentence corpus using cycle GAN and then decodes the caption with LSTM. MAGIC (w/o CO) is similar to (Yao et al. 2018) and encodes the global spatial graph as the image embedding to decode a single caption. MAGIC (w/o TA) considers all text tokens in the MRGs without query attention to examine the impact on the diversity of captions. MAGIC (w/o CG) utilizes the original GAN to align the image-sentence domain to investigate the impact on the accuracy of captions. MAGIC (w/o LD) generates captions without the language discriminator to assess the importance of the cascaded adversarial mechanism.

Table 1: Captioning and diversity results on TextCap validation set. DIC. indicates the results from diverse image captions, † indicates whether the model is a variant of MAGIC. Larger values indicate better performance, except for small value is better for RE-4. The top two scores on each metric are in bold. Acronym notations of each method see in comparison of methods.

<table>
<thead>
<tr>
<th>Method</th>
<th>DIC.</th>
<th>BLEU</th>
<th>METEOR</th>
<th>ROUG</th>
<th>SPICE</th>
<th>CIDEr</th>
<th>Div-1</th>
<th>Div-2</th>
<th>RE-4</th>
<th>SelfCIDEr</th>
</tr>
</thead>
<tbody>
<tr>
<td>ISM-LSTM</td>
<td>15.2</td>
<td>17.6</td>
<td>38.4</td>
<td>11.8</td>
<td>31.1</td>
<td></td>
<td>25.2</td>
<td>34.2</td>
<td>5.3</td>
<td>41.9</td>
</tr>
<tr>
<td>MAGIC (w/o CO)†</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAGIC (w/o TA)†</td>
<td>✓</td>
<td>21.7</td>
<td>18.3</td>
<td>39.5</td>
<td>12.3</td>
<td>71.3</td>
<td>28.8</td>
<td>37.4</td>
<td>4.8</td>
<td>45.4</td>
</tr>
<tr>
<td>MAGIC (w/o CG)†</td>
<td>✓</td>
<td>20.1</td>
<td>18.4</td>
<td>37.1</td>
<td>12.2</td>
<td>65.3</td>
<td>26.8</td>
<td>36.2</td>
<td>5.3</td>
<td>46.1</td>
</tr>
<tr>
<td>MAGIC (w/o LD)†</td>
<td>✓</td>
<td>20.8</td>
<td>20.1</td>
<td>41.1</td>
<td>12.8</td>
<td>73.9</td>
<td>28.2</td>
<td>38.3</td>
<td>4.5</td>
<td>48.8</td>
</tr>
<tr>
<td>MAGIC</td>
<td>✓</td>
<td>22.2</td>
<td>20.5</td>
<td>42.3</td>
<td>13.8</td>
<td>76.6</td>
<td>29.7</td>
<td>40.9</td>
<td>3.8</td>
<td>49.6</td>
</tr>
</tbody>
</table>

Table 2: The captioning and diversity results of different $N_k$ of central object-aware pool.

<table>
<thead>
<tr>
<th>COP Capacity</th>
<th>Results of different COP Size</th>
<th>Diversity Evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N_k=1$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$N_k=2$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$N_k=4$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$N_k=5$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$N_k=3$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3: The captioning and diversity results of different COP construction. Center, large and random means that the central object is selected by different rules.

Experimental Results

Captioning Results. The overall quantitative results of our model and baseline on the validation set of the Textcap dataset are listed in Tab. 1. The trained baseline model (ISM-LSTM) achieves the minimum captioning score, thereby indicating that it fails to describe the text-based image due to missing text awareness. The excellent performance of MAGIC and its variants in comprehending multimodal content compared with the baseline verifies the effectiveness of our proposed model. Notably, the results of other variants are better than those of MAGIC (w/o CO) and MAGIC (w/o CG). Our analysis showed that MAGIC (w/o CO) encodes an image with global content with abundant multimodal details, but its ground truth captions describe the image locally.

Diversity Results. Tab. 1 also presents the comparison of diversity results. The top three objects are selected as the central object-aware pool in MAGIC variants to ensure a fair comparison. Notably, diverse captioning models of MAGIC outperform the single captioning baseline of ISM-LSTM by a large margin. Generated various captions effectively respect MRGs, especially in the SelfCider score, which focuses on semantic similarity. We conduct additional diversity analyses in the following section.

In-Depth Analysis

We further validate the several key issues of the proposed method by answering three questions as follows.

**Q1: How does the structure of central object-aware pool (COP) affect the captioning diversity?** On the one hand, we conduct an ablation study to illustrate the impact of different sizes of COP on diverse and unpaired TextCap. The results are presented in Tab. 2. MAGIC with a pool size of $N_k=1,2,3$ presents the solid and approximate TextCap results, and MAGIC with a pool size of $N_k=4,5$ obtains competitive results. For this situation, we found that there are objective semantic overlaps among annotated image descriptions, i.e., for an image, some ground truths of an image from the sentence corpus are similarly semantic in the sentence domain. This similarity may cause confusion in the alignment of different multimodal relational graphs to the sentence domain in CGAN and lead to slight perfor-
Multimodal Relational Graphs

ISM-LSTM: A bottle of beer sits next to a bottle of beer.
MAGIC 1: A can of energy drink sits next to a bottle of mountain dew.
MAGIC 2: A can of energy drink on a desktop with a computer.
MAGIC 3: A bottle of mountain dew and a can of energy drink on a desk.

MAGIC 1: A book cover for Shadow of the watching star.
MAGIC 2: A page of a book that says shadow of the watching star.
MAGIC 3: A book called the shadow of the watching star by William Sarabande.

Figure 3: Two qualitative cases of MAGIC from TextCaps validation set. For better visualisation, we contruct the multimodal relational graphs based on generated diverse captions.

Figure 4: Visualization of features (1000 samples) 2D space by t-SNE (Van der Maaten and Hinton 2008).

Figure 5: The pie charts each comparing the two methods in human evaluation.

Diversity
Accuracy
Fluency

Comparative
MAGIC
MAGIC (w/o CG)

12% 35% 53%
21% 43% 36%
28% 31% 41%

Q3: How about the diversity of the visual concepts and texts for an image with multi-captions? We additionally conduct a qualitative analysis of the descriptive diversity of MAGIC in Fig. 3. We present the multiple MRGs ($N_k=3$) for each image on the basis of generated captions. The excellent results showed that MAGIC produces more accurate and fine-grained captions than baseline. In this case(a), MAGIC models complex multimodal relationships correctly and describes an image with different fine-grained contents adaptively. As presented in third caption, for intra-relationship, MAGIC can capture the semantic relationship “can-sit nexts to-bottle”, “bottle-on-desk”, and “can-on-desk” correctly; for cross-relationship, MAGIC can model the semantic relationship “bottle-of-energy drink” and “can-of-mountain dew” appropriately. However, the baseline ISM-LSTM describes the image with error semantic information and semantic structure.

Conclusion

We focus on diverse and unpaired text-based image caption generation. The MAGIC framework is proposed to achieve these objectives. MAGIC can adaptively build multiple multimodal relational graphs on the basis of selected central objects. A cascaded generative adversarial network is then developed to infer the unpaired caption generation in the image–sentence feature alignment level and linguistic coherence. The experimental results verified that MAGIC can effectively generate very promising and diverse captions without using any paired training data. The results of our study can provide new insights into visual captioning and guidance for future investigations on vision and language.

Acknowledgments

This work has been supported in Apart by National Key Research and Development Program of China (2018AAA0101900).


