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, hcshi, jiannan, 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

---

*These authors contributed equally to this work.

†Corresponding author.

1. The price of $ 17.88 that is **above** a lady.
2. A **Walmart sign** that **says** Rollback $17.88 is **above** a **shelf** of weight loss products.
3. The **hydroxycut** **has** 33% bonus on sale for $ 17.88 **instead of** $ 19.88.
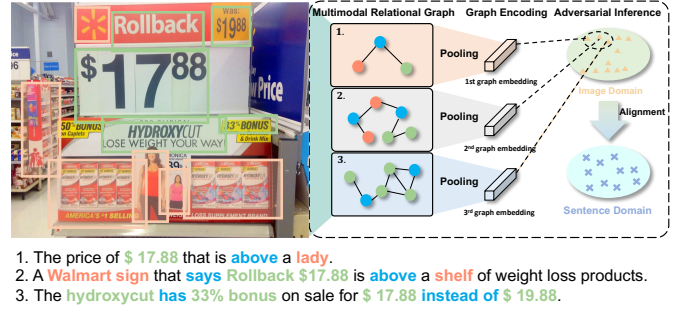
Figure 1: An example of how the diverse and unpaired captions generated. Orange, green and blue represent the objects, texts and multimodal relationships, respectively.

jointly, and write a succinct caption. Such task is very practical that encourage machine 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). 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 "17.8" denotes the "rollback" while "19.88" indicates the original price, a lady is located next to the commodity; (b) *cross-relationship*, the word "Hydroxycut" 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 task 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.
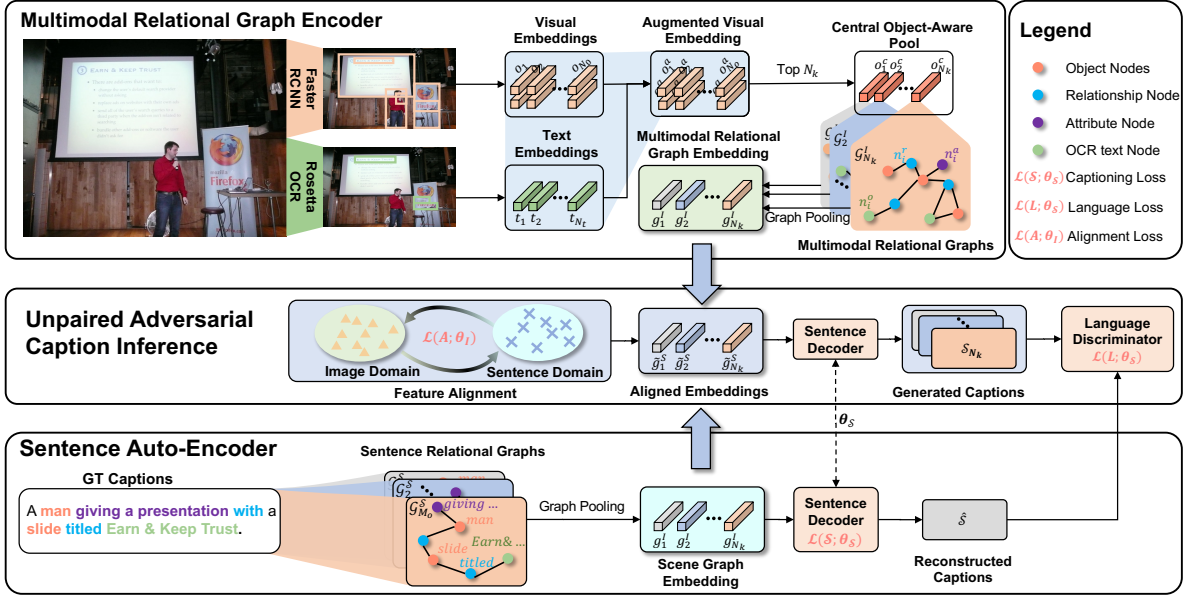
Figure 2: Overview of our framework. (a) *Multimodal Relational Graph Encoder* encodes multiple multimodal relational graphs based on central object-aware pool adaptively to represent the diversity of image content; (b) *Sentence Auto-Encoder* pre-extract the scene graph from the sentence corpus, then reconstructing the sentences from the latent vectors that are encoded from the scene graph; (c) *Unpaired Adversarial Caption Inference*, a cascaded generative adversarial network developed to reason the accurate and fluent captions with unpaired learning by aligning cross-domain features and distinguishing linguistic structure.

Problem formulation

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

$$
\underbrace{C((\mathcal{S}|\mathcal{I});(\theta_{\mathcal{I}}, \theta_{\mathcal{S}}))}_{\text{TextCap Model}} = \underbrace{E_{\mathcal{I}}((\mathcal{G}^{\mathcal{I}}|\mathcal{I});\theta_{\mathcal{I}})}_{\text{Image Encoder}}
$$
$$
\underbrace{U_{\mathcal{A}}((\mathcal{G}^{\mathcal{S}}|\mathcal{G}^{\mathcal{I}});\theta_{\mathcal{I}})}_{\text{Unpaired Alignment}} \underbrace{D_{\mathcal{S}}((\mathcal{S}|\mathcal{G}^{\mathcal{S}});\theta_{\mathcal{S}})}_{\text{Sentence Decoder}} \quad (1)
$$

$E_{\mathcal{I}}(\cdot;\theta)$ is image encoder aims to model N_k multimodal relational graphs (MRGs) $\mathcal{G}^{\mathcal{I}} = \{\mathcal{G}_i^{\mathcal{I}}|_{i=1}^{N_k}\}$ learned from image $\mathcal{I}$, then translate $\mathcal{G}^{\mathcal{I}}$ to appropriate sentence graph $\mathcal{G}^{\mathcal{S}}$ through unpaired cross-domain alignment $U_{\mathcal{A}}(\cdot;\theta)$, the aligned results can be decoded to diverse captions $\mathcal{S} = \{\mathcal{S}_i|_{i=1}^{N_k}\}$ by sentence decoder $D_{\mathcal{S}}(\cdot;\theta)$. Notice that $E_{\mathcal{I}}(\cdot;\theta)$ and $U_A(\cdot;\theta)$ share parameters $\theta_{\mathcal{I}}$ to learn feature alignment across domain, $D_{\mathcal{S}}(\cdot;\theta)$ is a pretrained sentence docoder from textual corpus, $\theta_{\mathcal{S}}$ are fixed as explicit supervised signals to learn $\theta_{\mathcal{I}}$ to generate unpaired captions.

Fig.2 shows the proposed Multimodal relAtional Graph adversarIal InferenCe (MAGIC) framework $C(.;\theta)$ consists of three main modules: Multimodal Relational Graph Encoder $E_{\mathcal{I}}(\cdot;\theta)$, Sentence Auto-Encoder $D_{\mathcal{S}}(\cdot;\theta)$ and Unpaired Caption Adversarial Inference $U_A(\cdot;\theta)$. We use the TextCap task as an example to illus- trate 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 $\mathcal{I}$, we firstly initialize it as its corresponding N_o visual object features and N_t text token features by Faster RCNN (Ren et al. 2016) and Rosetta OCR (Borisyuk, 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 $\mathcal{I}$ given by $\mathbf{O} = \{\mathbf{o}_i|_{i=1}^{N_o}\}$ and $\mathbf{T} = \{\mathbf{t}_j|_{j=1}^{N_t}\}$, 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 i^{th} object embedding $\mathbf{o}_i$, we employ soft attention to each text token embedding in $\mathbf{T}$. Then $\mathbf{o}_i$ and attened 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 $\mathbf{O}^a = \{\mathbf{o}_i^a|_{i=1}^{N_o}\}$. To decide the central objects

from $\mathbf{O}^a$, we adopt a MLP to predict a score for each object:

$$S_{object}^c = \text{softmax}(l_m^c(\mathbf{O}^a)) \qquad (2)$$

S_{object}^c contains N_o scores for each element that indicates the importance weight of each object. Then we select N_k objects $\mathbf{O}^c = \{\mathbf{o}_i^c|_{i=1}^{N_k}\}$ 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- relationships for each central object, then construct their corresponding MRGs $\mathcal{G}^{\mathcal{I}} = \{\mathcal{G}_i^{\mathcal{I}}|_{i=1}^{N_k}\}$ to represent the diversity of image $\mathcal{I}$. In detail, for k^{th} central object $\mathbf{o}_k^c$, its MRG defined as $\mathcal{G}_k^{\mathcal{I}} = (\mathcal{N}^{\mathcal{I}}, \mathcal{E}^{\mathcal{I}})$. We devote $\mathcal{N}^{\mathcal{I}}$ 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 $\mathbf{o}_k^c$, respectively.

$$\mathcal{N}^{\mathcal{I}}\{\mathbf{n}^r, \mathbf{n}^a\} = \begin{cases} \{W_r \mathbf{o}_i\}_{i=1}^{N_o-1}, & \text{relationship nodes;} \\ \{W_a[\mathbf{o}_i, s_i]\}_{i=1}^{N_o-1}, & \text{attribute nodes.} \end{cases} \qquad (3)$$

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

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

$$\mathcal{N}^{\mathcal{I}}\{\mathbf{n}^o\} = \{a_i^t \mathbf{t}_i\}_{i=1}^{N_t - sum(a_i^t > \epsilon)}, \quad \text{text nodes.} \qquad (4)$$

where a_i^t is the attention score, ϵ is a pre-defined threshold. $sum()$ is counting function aims to connect more related contextual texts for o_k^c if attention score a_j^t greater than ϵ.

Thus far, the MRG $\mathcal{G}_k^{\mathcal{I}} = (\mathcal{N}^{\mathcal{I}}, \mathcal{E}^{\mathcal{I}})$ 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 neighboured nodes and embed the graph structure into latent vectors.

$$\mathbf{g}_k^{\mathcal{I}} = \delta_r(W_{m_0}\mathbf{o}_k^c + \sum_{\hat{\mathbf{n}}_i \in \mathcal{N}^{\mathcal{I}}} W_{m_1}\hat{\mathbf{n}}_i) \qquad (5)$$

where δ_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 $\mathbf{g}_k^{\mathcal{I}}$ for object $\mathbf{o}_k^c$ is obtained. As a consequence, the content diversity of image $\mathcal{I}$ can be regarded as the collection $\mathbf{g}^{\mathcal{I}} = \{\mathbf{g}_k^{\mathcal{I}}|_{k=1}^{N_k}\}$ 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 $\mathcal{G}^{\mathcal{S}}$ from the sentence corpus,

and then decode the sentence from latent vector that is encoded from the $\mathcal{G}^{\mathcal{S}}$ (sentence $\rightarrow \mathcal{G}^{\mathcal{S}} \rightarrow$ latent vector $g^{\mathcal{S}} \rightarrow$ reconstructed sentence).

Sence 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 $\mathbf{c}_i^o$ in $\hat{S}$, we define its relational graph as $\mathcal{G}_i^{\mathcal{S}} = (\mathcal{N}^{\mathcal{S}}, \mathcal{E}^{\mathcal{S}})$, where $\mathcal{N}^s$ devote three relational type, relationship nodes $\mathbf{c}^r$, attribute nodes $\mathbf{c}^a$ and associated textual nodes $\mathbf{c}^t$. The global semantic meaning of sentence $\hat{S}$ can be regarded as the collection of object relational graphs $\mathcal{G}^{\mathcal{S}} = \{\mathcal{G}_i^{\mathcal{S}}|_{i=1}^{M_o}\}$, where M_o is the object number in sentence $\mathcal{S}$. Similarly, the latent embedding of sentence is learned by the multi-layer GCN which using same structure in image graph encoding but independent parameters.

$$g^{\mathcal{S}} = \frac{1}{M_o}\delta_r(\sum_{i=1}^{M_o}(W_{c0}\mathbf{c}_i^o + \sum_{\hat{\mathbf{c}}_i \in \mathcal{N}^{\mathcal{S}}} W_{c1}\hat{\mathbf{c}}_i)) \qquad (6)$$

where W_{c0} and $W_{c1} \in \mathbb{R}^{d \times e}$, Thus, the scene graph embedding of sentence $\hat{\mathcal{S}}$ is given by $\mathbf{g}^{\mathcal{S}}$.

Sentence Decoder. The decoder part is essential in the way that it translates the semantic latent vector $\mathbf{g}^{\mathcal{S}}$ 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 $\hat{\mathbf{y}}_t$ in common vocabulary based on the scene graph embedding $\mathbf{g}^{\mathcal{S}}$. Then a dynamic pointer network that makes final word prediction on the basis of text candidates $\mathbf{T}^c$ and $\hat{\mathbf{y}}_t$.

$$\begin{aligned} \hat{\mathbf{y}}_t &= l_m^w(\text{LSTM}_{\text{DEC}}(\mathbf{h}_{t-1}, \mathbf{g}^{\mathcal{S}})), \\ \mathbf{y}^t &= \text{argmax}([\hat{\mathbf{y}}_t, DPN(\hat{\mathbf{y}}_t, \mathbf{T}^c)]); \end{aligned} \qquad (7)$$

where $l_m^w(.)$ is a linear classifier for common vocabulary, $DPN(.)$ denotes the dynamic pointer network. We devote 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:

$$\mathcal{L}(\mathcal{S}; \theta_{\mathcal{S}}) = -\log \cdot \sum_{i=0}^{T}(P(\mathbf{y}_t|(\mathbf{y}_{0:T-1}, \mathbf{g}^{\mathcal{S}}, \mathbf{T}^c))) \qquad (8)$$

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 $\mathbf{G}^{\mathcal{I}} = \{\mathbf{g}_i^{\mathcal{I}}|_{i=1}^{N_{\mathcal{I}}}\}$ and $\mathbf{G}^{\mathcal{S}} = \{\mathbf{g}_i^{\mathcal{S}}|_{i=1}^{N_{\mathcal{S}}}\}$ from image and sentence domain, respectively, where $N_{\mathcal{I}}$ and $N_{\mathcal{S}}$ are the capacity of corresponding corpus. The goal of feature alignment is enforce the $\mathbf{G}^{\mathcal{I}}$ to be close to the $\mathbf{G}^{\mathcal{S}}$ distribution. We follow (Zhu et al. 2017; Gu et al. 2019) that align feature space through cycle-consistent adversarial training. Differently, considering 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 discriminator, to align domain X to domain Y, the distribution distance $W(X, Y)$ can be defined and corresponding loss can be minimized as follows:

$$
\begin{aligned}
W(X, Y) &= \inf_{\gamma \sim \Pi(X,Y)} \mathbb{E}_{(\mathbf{x},\mathbf{y}) \sim \gamma}[||\mathbf{x} - \mathbf{y}||] \\
\mathcal{L}(X \to Y; \theta) &= \mathbb{E}_{\mathbf{y} \sim Y}[D_Y(\mathbf{y})] - \mathbb{E}_{\mathbf{x} \sim X}[D_Y(\mathbf{x})]
\end{aligned}
\tag{9}
$$

where $\gamma \sim \Pi(X, Y)$ is the set of possible joint distributions for the combination of $\mathbf{x}$ and $\mathbf{y}$. D_Y is a discriminator to distinguish the origin source of the latent vector if from Y. Such loss is more effectively maintain the semantic coherence between two domain with huge discrepant characteristics. Thus, the cycle loss of feature alignment from $\mathbf{G}^{\mathcal{I}}$ to $\mathbf{G}^{\mathcal{S}}$ can be defined as follows:

$$
\begin{aligned}
\mathcal{L}(A; \theta_{\mathcal{I}}) =&\lambda_A(\mathcal{L}(M_{\mathcal{S}}(\mathbf{G}^{\mathcal{I}}) \to \mathbf{G}^{\mathcal{S}}) + \mathcal{L}(M_{\mathcal{I}}(\mathbf{G}^{\mathcal{S}}) \\
&\to \mathbf{G}^{\mathcal{I}}) + \lambda_C \mathcal{L}(M_{\mathcal{I}}(\mathbf{G}^{\mathcal{I}}) \leftrightarrow M_{\mathcal{I}}(\mathbf{G}^{\mathcal{S}})))
\end{aligned}
\tag{10}
$$

where λ_A and λ_C are weight hyper-parameters to control the feature alignment, $M_{\mathcal{I}}$ and $M_{\mathcal{S}}$ are mapping function to image domain and sentence domain, respectively. $\mathcal{L}(M_{\mathcal{I}}(\mathbf{G}^{\mathcal{I}}) \leftrightarrow M_{\mathcal{I}}(\mathbf{G}^{\mathcal{S}}))$ is the cycle consistency loss to regularize the training. Subsequently, the aligned image-sentence embeddings $\mathbf{G}^{\tilde{\mathcal{S}}} = \{\tilde{\mathbf{g}}_i^{\mathcal{S}}|_{i=1}^{N_{\mathcal{I}}}\}$ and text candidates T^c are feed into sentence decoder to generate captions.

Language Discriminator. Notably, the cycle loss can minimize the distribution distance among image and sentence domains. However, the characteristic distribution of aligned image embeddings $\mathbf{G}^{\tilde{\mathcal{S}}} = \{\tilde{\mathbf{g}}_i^{\mathcal{S}}|_{i=1}^{N_{\mathcal{I}}}\}$ is between image domain and sentence domain. To capture the language characteristics further, we introduce a language discriminator D_L that focuses on the language structure of an individual sentence. Here, we want to ensure fluency and language accuracy that may be lacking in the cross-domain alignment. In addition to generated captions from the sentence decoder, D_L is given negative inputs with a mixture of randomly shuffled words or repeated phrases within a sentence.

$$
\begin{aligned}
\mathcal{L}(L; \theta_{\mathcal{S}}) =&\lambda_L(\mathbb{E}_{\mathbf{s} \sim S}[\log(D_L(\mathbf{s}))] - \mathbb{E}_{\mathbf{s}^r \sim S^r} \\
&[1 - \log(D_L(\mathbf{s}^r))])
\end{aligned}
\tag{11}
$$

where λ_L is weight hyper-parameter and D_L is a Bidirectional LSTM. S^r is a set in which sentences with random N-grams. This language discriminator verifies not only the semantic fluency among sentences but also improves the captioning accuracy.

Training Algorithm

Algorithm 1 (in the appendix) presents the pseudocode of our MAGIC algorithm for diverse and unpaired TextCap. First, we pretrain the sentence auto-encoder by reconstructing sentences and the language discriminator as well as disrupting the order of words from the sentence corpus. The MAGIC framework then learns diverse multimodal relational graph embeddings by querying the semantic meaning of central objects. Finally, the developed CGAN aligns cross-modal embeddings from the image domain to the sentence domain and then checks the fluency and accuracy of generated captions using the language discriminator to generate unpaired and diverse captions magically.

Experiments

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 (Borisyuk, 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 $N_o = 100$ 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.

---

[1] https://github.com/tylin/coco-caption/

	DIC.	Captioning Evaluation					Diversity Evaluation			
Method		BLEU	METEOR	ROUG	SPICE	CIDEr	Div-1	Div-2	RE-4 ↓	SelfCIDEr
ISM-LSTM		15.2	17.6	38.4	11.8	31.1	25.2	34.2	5.3	41.9
MAGIC (w/o CO)[†]		20.2	18.1	39.5	12.3	71.3	28.8	37.4	4.8	45.4
MAGIC (w/o TA)[†]	✓	**21.7**	**20.3**	**42.0**	**13.6**	**75.9**	**28.1**	**40.4**	**3.6**	**49.1**
MAGIC (w/o CG)[†]	✓	20.1	18.4	37.1	12.2	65.3	26.8	36.2	5.3	46.1
MAGIC (w/o LD)[†]	✓	20.8	20.1	41.1	12.8	73.9	28.2	38.3	4.5	48.8
MAGIC	✓	**22.2**	**20.5**	**42.3**	**13.8**	**76.6**	**29.7**	**40.9**	**3.8**	**49.6**

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.

	Results of differnet COP Size			
COP Capacity	BLEU	CIDEr	Div-2	SelfCIDEr
$N_k=1$	**22.3**	**77.3**	**41.9**	**50.3**
$N_k=2$	21.8	75.8	**41.1**	49.3
$N_k=4$	21.6	75.3	39.3	47.8
$N_k=5$	20.8	73.4	38.1	47.5
$N_k=3$	**22.2**	**76.6**	40.9	**49.6**

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

	Results of different COP construction			
Central Object	BLEU	CIDEr	Div-2	SelfCIDEr
Center	18.7	72.3	38.3	44.2
Large	**20.1**	**75.9**	**40.2**	**49.3**
Random	17.1	67.8	36.9	46.6
Top Socre	**22.2**	**76.6**	**40.9**	**49.6**

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.

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.

Experimental Results

Captioning Results. The overall qualitative 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.

The CGAN may fail to align the cross-domain appropriately. Compared with MAGIC, MAGIC (w/o CO) with original GAN may fail to capture cross-domain semantic consistency effectively. In contrast, MAGIC with CGAN can improve the quality of unpaired captions in feature alignment and semantic structure levels.

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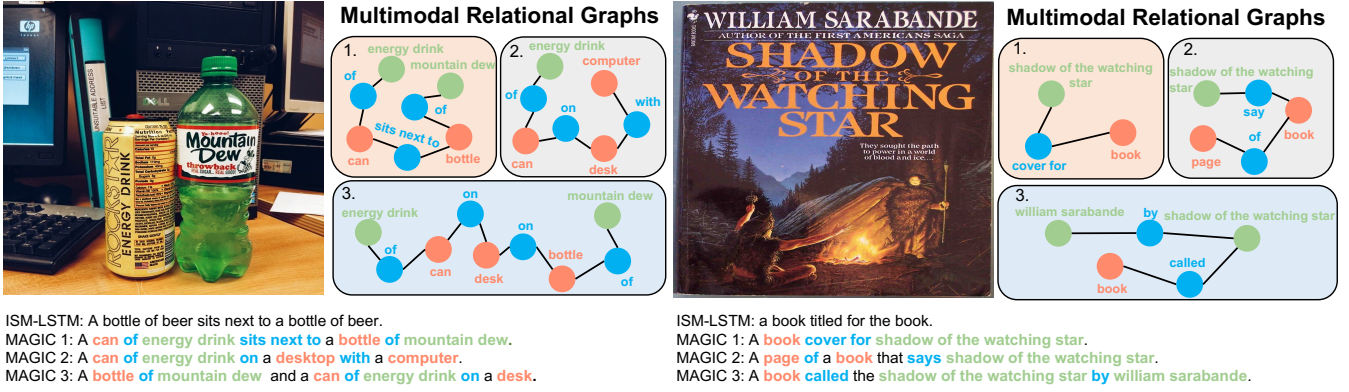
Figure 3: Two qualitative cases of MAGIC from TextCaps validation set. For better visualisation, we contruct the multimodal relational graphs based on generated diverse captions.
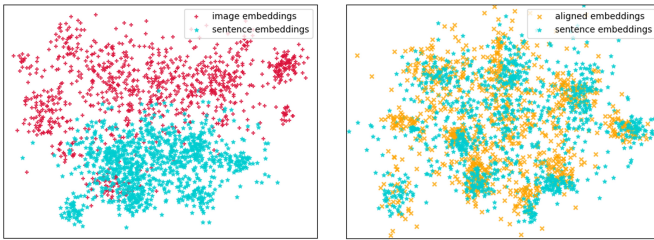

Figure 4: Visualization of features (1000 samples) 2D space by t-SNE (Van der Maaten and Hinton 2008).

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

mance degradation. On the other hand, the quantitative results according to rule- and learning-based central object selections for OAP are reported. Tab.3 demonstrated that all rule-based methods suffer from low metric performance while the learning-based central object can accurately and diversely describe an image.

Q2: How much improvement of unpaired TextCap quality has the MAGIC brought? The MAGIC can improve the unpaired TextCap effectively by feature alignment and semantic coherence distinguishing. First, we visualize the feature alignment to indicate that our MAGIC can capture the semantic consistency across the image–sentence domain Fig. 4. To better understand how satisfactory are the sentences generated by CGAN, we conducted a human study with five experienced workers to evaluate the generated descriptions from MAGIC and MAGIC (w/o CG). Workers evaluated 100 images which are randomly sampled from the validation set for each pairwise comparison. Fig. 5 shows that employing MAGIC can describe images more diversely, accurately and fluently.

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
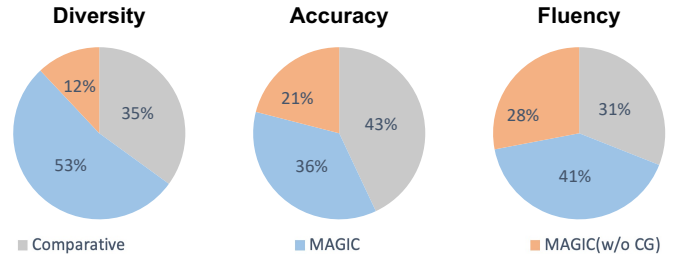
correctly and describes an image with different fine-grained contents adaptively. As presented in third caption, for intra-relationship, MAGIC can capture the semantic relationship "*can-sit nexts to-bottle*", "*bottle-on-desk*", and "*can-on-desk*" correctly; for cross-relationship, MAGIC can model 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

References

Anderson, P.; Fernando, B.; Johnson, M.; and Gould, S. 2016. Spice: Semantic propositional image caption evaluation. In *European conference on computer vision*, 382–398. Springer.

Berg, R. v. d.; Kipf, T. N.; and Welling, M. 2017. Graph convolutional matrix completion. *arXiv preprint arXiv:1706.02263*.

Borisyuk, F.; Gordo, A.; and Sivakumar, V. 2018. Rosetta: Large scale system for text detection and recognition in images. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 71–79.

Caron, M.; Misra, I.; Mairal, J.; Goyal, P.; Bojanowski, P.; and Joulin, A. 2020. Unsupervised learning of visual features by contrasting cluster assignments. *arXiv preprint arXiv:2006.09882*.

Chen, T.; Pu, T.; Wu, H.; Xie, Y.; Liu, L.; and Lin, L. 2021. Cross-domain facial expression recognition: A unified evaluation benchmark and adversarial graph learning. *IEEE transactions on pattern analysis and machine intelligence*.

Cornia, M.; Stefanini, M.; Baraldi, L.; and Cucchiara, R. 2020. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10578–10587.

Denkowski, M.; and Lavie, A. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, 376–380.

Ding, S.; Qu, S.; Xi, Y.; and Wan, S. 2020. Stimulus-driven and concept-driven analysis for image caption generation. *Neurocomputing*, 398: 520–530.

Gu, J.; Joty, S.; Cai, J.; and Wang, G. 2018. Unpaired image captioning by language pivoting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 503–519.

Gu, J.; Joty, S.; Cai, J.; Zhao, H.; Yang, X.; and Wang, G. 2019. Unpaired image captioning via scene graph alignments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10323–10332.

Guo, J.; Shi, H.; Kang, Y.; Kuang, K.; Tang, S.; Jiang, Z.; Sun, C.; Wu, F.; and Zhuang, Y. 2021. Semi-Supervised Active Learning for Semi-Supervised Models: Exploit Adversarial Examples With Graph-Based Virtual Labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2896–2905.

Guo, L.; Liu, J.; Yao, P.; Li, J.; and Lu, H. 2019. Mscap: Multi-style image captioning with unpaired stylized text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4204–4213.

Hou, J.; Wu, X.; Qi, Y.; Zhao, W.; Luo, J.; and Jia, Y. 2019. Relational reasoning using prior knowledge for visual captioning. *arXiv preprint arXiv:1906.01290*.

Johnson, J.; Karpathy, A.; and Fei-Fei, L. 2016. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4565–4574.

Klein, D.; and Manning, C. D. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st annual meeting of the association for computational linguistics*, 423–430.

Lample, G.; Conneau, A.; Denoyer, L.; and Ranzato, M. 2017. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*.

Li, J.; Galley, M.; Brockett, C.; Gao, J.; and Dolan, B. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.

Li, J.; Tang, S.; Wu, F.; and Zhuang, Y. 2019. Walking with mind: Mental imagery enhanced embodied qa. In *Proceedings of the 27th ACM International Conference on Multimedia*, 1211–1219.

Li, J.; Tang, S.; Zhu, L.; Shi, H.; Huang, X.; Wu, F.; Yang, Y.; and Zhuang, Y. 2021. Adaptive hierarchical graph reasoning with semantic coherence for video-and-language inference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1867–1877.

Li, J.; Wang, X.; Tang, S.; Shi, H.; Wu, F.; Zhuang, Y.; and Wang, W. Y. 2020. Unsupervised reinforcement learning of transferable meta-skills for embodied navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12123–12132.

Liu, F.; Gao, M.; Zhang, T.; and Zou, Y. 2021. Exploring Semantic Relationships for Unpaired Image Captioning. *arXiv preprint arXiv:2106.10658*.

Marcheggiani, D.; and Titov, I. 2017. Encoding sentences with graph convolutional networks for semantic role labeling. *arXiv preprint arXiv:1703.04826*.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Peyré, G.; Cuturi, M.; et al. 2019. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6): 355–607.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2016. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6): 1137–1149.

Rennie, S. J.; Marcheret, E.; Mroueh, Y.; Ross, J.; and Goel, V. 2017. Self-critical sequence training for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7008–7024.

ROUGE, L. C. 2004. A package for automatic evaluation of summaries. In *Proceedings of Workshop on Text Summarization of ACL, Spain*.

Schuster, S.; Krishna, R.; Chang, A.; Fei-Fei, L.; and Manning, C. D. 2015. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Proceedings of the fourth workshop on vision and language*, 70–80.

Shen, J.; Qu, Y.; Zhang, W.; and Yu, Y. 2018. Wasserstein distance guided representation learning for domain adaptation. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Sidorov, O.; Hu, R.; Rohrbach, M.; and Singh, A. 2020. Textcaps: a dataset for image captioning with reading comprehension. In *European Conference on Computer Vision*, 742–758. Springer.

Singh, A.; Natarajan, V.; Shah, M.; Jiang, Y.; Chen, X.; Batra, D.; Parikh, D.; and Rohrbach, M. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8317–8326.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4566–4575.

Wang, Q.; and Chan, A. B. 2019. Describing like humans: on diversity in image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4195–4203.

Xu, G.; Niu, S.; Tan, M.; Luo, Y.; Du, Q.; and Wu, Q. 2021. Towards Accurate Text-based Image Captioning with Content Diversity Exploration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12637–12646.

Yang, X.; Zhang, H.; and Cai, J. 2020a. Auto-encoding and distilling scene graphs for image captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Yang, X.; Zhang, H.; and Cai, J. 2020b. Deconfounded image captioning: A causal retrospect. *arXiv preprint arXiv:2003.03923*.

Yao, T.; Pan, Y.; Li, Y.; and Mei, T. 2018. Exploring visual relationship for image captioning. In *Proceedings of the European conference on computer vision (ECCV)*, 684–699.

Yu, J.; Li, J.; Yu, Z.; and Huang, Q. 2019. Multimodal transformer with multi-view visual representation for image captioning. *IEEE transactions on circuits and systems for video technology*, 30(12): 4467–4480.

Zhang, W.; Shi, H.; Tang, S.; Xiao, J.; Yu, Q.; and Zhuang, Y. 2021a. Consensus graph representation learning for better grounded image captioning. In *Proc 35 AAAI Conf on Artificial Intelligence*.

Zhang, W.; Tang, S.; Cao, Y.; Pu, S.; Wu, F.; and Zhuang, Y. 2019. Frame augmented alternating attention network for video question answering. *IEEE Transactions on Multimedia*, 22(4): 1032–1041.

Zhang, W.; Tang, S.; Cao, Y.; Xiao, J.; Pu, S.; Wu, F.; and Zhuang, Y. 2020a. Photo Stream Question Answer. In *Proceedings of the 28th ACM International Conference on Multimedia*, 3966–3975.

Zhang, W.; Tang, S.; Su, J.; Xiao, J.; and Zhuang, Y. 2021b. Tell and guess: cooperative learning for natural image caption generation with hierarchical refined attention. *Multimedia Tools and Applications*, 80(11): 16267–16282.

Zhang, W.; Wang, X. E.; Tang, S.; Shi, H.; Shi, H.; Xiao, J.; Zhuang, Y.; and Wang, W. Y. 2020b. Relational graph

learning for grounded video description generation. In *Proceedings of the 28th ACM International Conference on Multimedia*, 3807–3828.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 2223–2232.