Beyond a Pre-Trained Object Detector:
Cross-Modal Textual and Visual Context for Image Captioning

Chia-Wen Kuo
Georgia Tech
albert.cwkuo@gatech.edu

Zsolt Kira
Georgia Tech
zkira@gatech.edu

Abstract

Significant progress has been made on visual captioning, largely relying on pre-trained features and later fixed object detectors that serve as rich inputs to auto-regressive models. A key limitation of such methods, however, is that the output of the model is conditioned only on the object detector's outputs. The assumption that such outputs can represent all necessary information is unrealistic, especially when the detector is transferred across datasets. In this work, we reason about the graphical model induced by this assumption, and propose to add an auxiliary input to represent missing information such as object relationships. We specifically propose to mine attributes and relationships from the Visual Genome dataset and condition the captioning model on them. Crucially, we propose (and show to be important) the use of a multi-modal pre-trained model (CLIP) to retrieve such contextual descriptions. Further, object detector models are frozen and do not have sufficient richness to allow the captioning model to properly ground them. As a result, we propose to condition both the detector and description outputs on the image, and show qualitatively and quantitatively that this can improve grounding. We validate our method on image captioning, perform thorough analyses of each component and importance of the pre-trained multi-modal model, and demonstrate significant improvements over the current state of the art, specifically +7.5% in CIDEr and +1.3% in BLEU-4 metrics.

1. Introduction

For vision-and-language (VL) tasks such as generating textual descriptions of an image (image captioning) [1, 9, 24], it is crucial to encode the input image into a representation that contains relevant information for the downstream language task. Earlier works use an ImageNet pretrained model to encode the input image [24, 51], while recent works achieve much better performance by using objects detected by an object detector (e.g. Faster R-CNN [41]
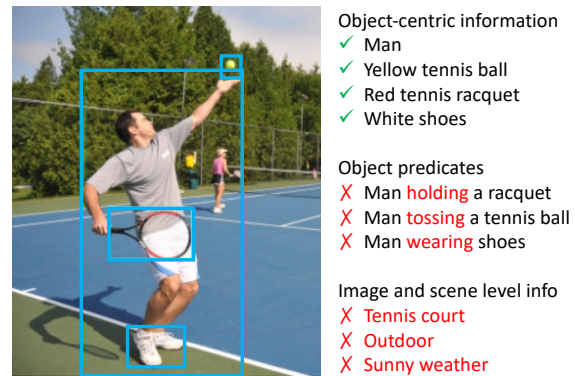

Figure 1. Most existing VL methods encode the input image by a set of objects detected by a frozen pre-trained object detector. This set of detected objects may be able to provide object-centric information such as object classes, locations, and attributes, but may fail to encode other information also crucial for the target VL tasks such as object predicates and image/scene level information.

pre-trained on Visual Genome [27]) [1,25,39]. The detected objects encode more fine-grained information from the input image such as object classes, locations, and attributes, hence achieving substantially better performance.

Despite the success of encoding the input image with detected objects, the object detector is *pre-trained* on datasets such as Visual Genome and kept *frozen* during the training of the target VL task (on a different dataset). This leads to two major issues as illustrated in Figure 1: (1) the detector may be good at encoding object-centric information but not at many other kinds of information necessary for the target VL task such as the relationship between objects and image/scene level information; and (2) the conditional relationship between the detected objects and the input image is not jointly optimized for the target VL task so that features computed by the object detector cannot be refined before sending into the VL model, potentially resulting in poor features that are difficult to ground, for example.

For (1), most existing works follow prior works [1] to pre-train the object detector on Visual Genome for object

detection and attribute classification. This implies that the object features may be good at encoding object-centric information such as classes, locations, and attributes, but not at encoding other crucial information. Take image captioning as an example; as shown in Figure 1, such crucial information includes relationships between objects (object predicates), image/scene level information, etc. Therefore, the first objective of this paper is to provide information complementary to the detected objects.

Inspired by the way the Visual Genome dataset is constructed, we propose to provide complementary but necessary information in the form of contextual text descriptions for image sub-regions. However, generating the descriptions for image sub-regions requires training another image captioning model, which by itself may not be an easy task. Therefore, we propose to turn the text generation problem into a *cross-modal retrieval problem*: given an image sub-region, retrieve the top-k most relevant text descriptions from a description database. One way of doing cross-modal retrieval is to search for visually similar images and return the paired text of that image [14,18,35,46]. However, we posit that we can effectively leverage recent advances in cross-modal pre-training on large-scale image and text pairs, CLIP [40], to *directly* retrieve relevant text given an image. CLIP has two branches, CLIP-I and CLIP-T which encode image and text, respectively, into a global feature representation, and is trained to pull paired image and text together and push unpaired ones apart. We show in Section 4.3 that the text descriptions retrieved by CLIP are more relevant to the image query compared to those retrieved indirectly by visual similarity. The retrieved text descriptions by CLIP provide rich and complementary information, thus leading to substantial performance improvement.

For (2), in most existing works the pre-trained object detector is kept frozen when training the target VL task. This implies that the conditioning relationship between the detected objects and the input image is not jointly optimized with the target VL task. Consequently, the information from a transferred object detector may not result in high-quality features that can be effectively used by the captioning model, for example in grounding them to words. Therefore, the second objective of this paper is to strengthen the conditioning relationship between the detected objects and the input image by optimizing this relationship jointly with the target VL task.

To strengthen the conditional relationship for (2), we should first encode the input image into a global feature representation in a way that preserves as much information relevant to the target VL task as possible. In this paper, we choose CLIP-I, the image branch of CLIP model, as the image encoder. Since CLIP is also pre-trained on a cross-modal VL task, we show in Section 4.3 that it can better encode information relevant to the target VL tasks compared

to models pre-trained on image-only datasets. We then use a fully connected (FC) layer, which is jointly optimized with the target VL task, to model the conditional relationship.

In this paper, we validate our proposed method on the VL task of image captioning. By addressing the two issues from using a frozen pre-trained object detector described above, our method improves one of the SoTA image captioning models $\mathcal{M}^2$ by +7.2% in CIDEr and +1.3% in BLEU-4. In summary, we make the following contributions:

- Identify the potential issues of using detected objects from a frozen pre-trained object detector to encode the input image for image captioning.
- Propose a cross-modal retrieval module by leveraging the cross-modal joint embedding space by CLIP to retrieve a set of contextual text descriptions that provide information complementary to detected objects.
- Propose an image conditioning module to strengthen and jointly optimize the conditional relationship between the detected objects and the input image such that the features are more effective and support tasks such as grounding.
- Improve SoTA object-only baseline model by a substantial margin and provide thorough quantitative and qualitative analyses for the proposed two modules and the design choices made within each module.

2. Related Works

In the early stage of image captioning, researchers used an image encoder such as ResNet [16] to encode the input image into a global pooled representation [2, 3, 6, 11, 13, 15, 22, 24, 34, 51, 54, 61]. Captions are then generated conditioned on the encoded global feature. The major issue of using a global pooled representation is that information from the input image is heavily compressed and lost during the encoding process. For example, all salient objects are fused and spatial information is discarded. Therefore, follow-on works discard the pooling layer and use grid features from the ConvNet to encode more fine-grained details of the input image [4, 23, 33, 42, 56].

In order to further encode more fine-grained details of the input image, Anderson et al. [1] propose to encode the input image with a set of objects detected by a frozen object detector. The object detector is pre-trained on Visual Genome for object detection and attribute classification. The detected objects are represented by the RoI-pooled features from the object detector. With the set of detected objects, fine-grained and richer information from the input image such as salient objects, object classes, locations, attributes, etc, are encoded for the downstream VL task, leading to substantial performance improvements. Because of its great success, encoding the input image with a set of detected objects has become a standard approach in recent VL works [1, 17, 19, 20, 25, 39, 52], as well as in VL pre-training [7, 28, 32, 45, 49].
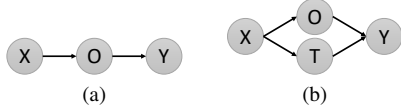
Figure 2. Graphical models of **(a)** most existing image captioning models, where X is the input image, O is a set of objects detected by a frozen pre-trained object detector; and **(b)** our proposed model with a newly introduced node T, which represents a set of text descriptions of image sub-regions.

Despite the wide adoption of encoding the input image with a set of detected objects, as we argued in Section 1, the pre-trained object detector may not properly encode other crucial information such as object predicates and image/scene level information as illustrated in Figure 1. In order to encode more complete information of the input image by a pre-trained object detector, Zhang et al. [63] propose VinVL where the object detector is pre-trained on much larger training corpora that combine multiple public annotated object detection datasets. Therefore, richer visual objects, attributes, and concepts in the input image are encoded. Nevertheless, the object detector is still pre-trained to encode object-centric information, and other information, e.g. interaction between objects, is not optimized for in the pre-training of the object detector.

To encode information in addition to the detected objects, Li et al. [29] propose OSCAR that includes object tags detected in the image in the form of text for VL pre-training. Different from our work, the motivation of including object tags in OSCAR is to use them as anchor points to bridge vision and language modalities for VL pre-training. Also, object predicates are not provided by the object tags.

The image captioning methods discussed above typically generate general descriptions of an input image. Furthermore, we cannot control the captioning model to generate diverse or more specific captions that focus on particular objects in the input image. Therefore, another line of image/video captioning works focus on generating customized captions with additional control signals [8, 36, 37, 43, 53]. This line of works also includes information external of the input image/video such as hashtags [8], or news documents [53] for caption generation. Different from standard image captioning methods and our method that train and evaluate the trained model on MS-COCO [31], this line of works requires additional annotation or labeling during the construction of such custom datasets. Our method does not require additional annotations of the input image and focuses on the standard MS-COCO benchmark.

3. Method

3.1. Graphical Model

Most existing works model the image captioning problem with a graphical model shown in Figure 2a, where

given an input image X, a set of objects O are detected by a frozen pre-trained object detector, and the caption Y is generated conditioned on O. The graphical model with the chain structure shown in Figure 2a can be derived as:

$$p(\boldsymbol{y}|\boldsymbol{x}) = \prod_i p(y_i|\boldsymbol{x}, y_{1:i-1})$$

$$= \prod_i \sum_O p(\boldsymbol{o}|\boldsymbol{x}, y_{1:i-1})p(y_i|\boldsymbol{x}, \boldsymbol{o}, y_{1:i-1})$$

$$= \prod_i \sum_O p(\boldsymbol{o}|\boldsymbol{x})p(y_i|\boldsymbol{x}, \boldsymbol{o}, y_{1:i-1}) \quad (1)$$

$$= \prod_i \sum_O p(\boldsymbol{o}|\boldsymbol{x})p(y_i|\boldsymbol{o}, y_{1:i-1}) \quad (2)$$

$$\simeq \prod_i p(y_i|\boldsymbol{o}, y_{1:i-1}) \quad (3)$$

where $p(y_i|\boldsymbol{o}, y_{1:i-1})$ is modeled as an auto-regressive caption generation model. Between Equation 1 and Equation 2, it is assumed that $\boldsymbol{o}$ completely encode all necessary information of $\boldsymbol{x}$ so that y_i is conditionally independent of $\boldsymbol{x}$. Between Equation 2 and Equation 3, researchers typically take the argmax and threshold to select a fixed set of detected objects from the object detector. From the graphical model derived above, we can clearly identify the two major issues arising from a frozen pre-trained object detector. First, the assumption that $\boldsymbol{o}$ completely encodes all necessary information of $\boldsymbol{x}$. In practice, the object detector pre-trained on Visual Genome for object detection and attribute prediction may fail to encode crucial information of $\boldsymbol{x}$ such as the relationship between objects and image/scene level information. Second, the conditional relationship between the detected objects $\boldsymbol{o}$ and the input image $\boldsymbol{x}$ is computed by a frozen pre-trained object detector and is not jointly optimized with the target image captioning task. Therefore, the features computed by the frozen pre-trained object detector cannot be refined before sending into the auto-regressive caption generation model, leading to potentially poor features especially given that they are trained on a different dataset.

To mitigate issue (1), one naïve solution is to pre-train the object detector to predict other information such as the predicates between objects so that more complete information can be encoded by $\boldsymbol{o}$. However, it is still an open research question to effectively train a network to model the interaction between objects, especially across datasets [30,50,55,57,62]. Therefore, in this paper, we propose to insert another node T into the model, as shown in Figure 2b, to encode information complementary to O without re-training the object detector. By including both O and T, more complete information of X is encoded, and thus the conditional independence assumption between Equation 1 and Equation 2 are better supported. We can similarly de-
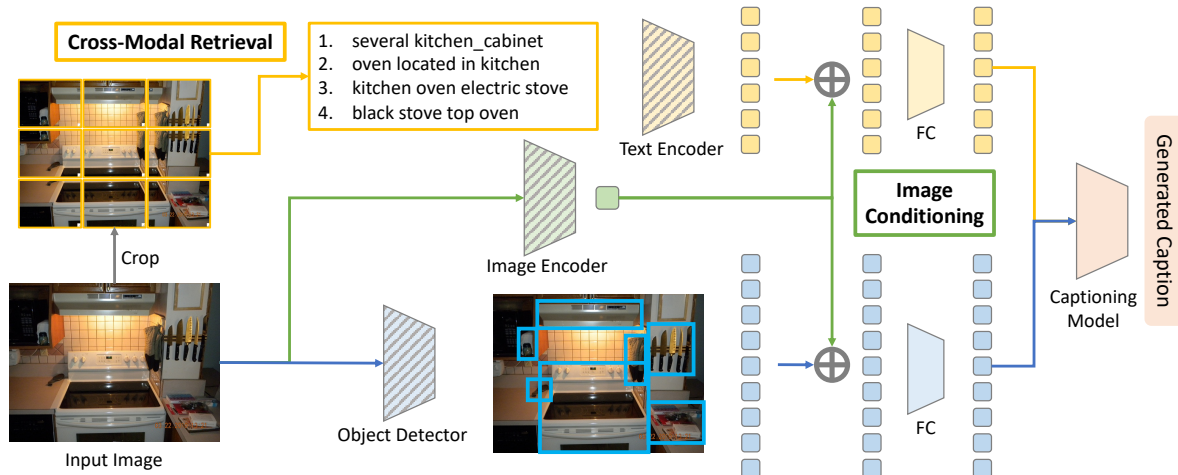
Figure 3. Model architecture. We propose (1) a cross-modal retrieval module to retrieve a set of contextual text descriptions that provide information complementary to the detected objects as shown in the yellow box. We also propose (2) an image conditioning module to strengthen the conditional relationship between the detected objects and the input image as shown in the green box. The models with shaded patterns (text encoder, image encoder, and object detector) are pre-trained and kept frozen. Only the FCs and the captioning model are trained for the target VL task. The $\bigoplus$ symbol represents concatenation operation along the feature dimension. Each token ($\square$ symbol) represents a d-dimensional feature vector. The image features (green token) are broadcast before the concatenation operation.

rive the graphical model with the newly introduced node T:

$$p(\boldsymbol{y}|\boldsymbol{x}) \simeq \prod_i p(y_i|\boldsymbol{o}, \boldsymbol{t}, y_{1:i-1}) \qquad (4)$$

Please see supplementary for the full derivation.

To mitigate issue (2), we propose to use a fully connected (FC) layer to refine the features of each detected object conditioned on the features of input image X. The FC layer is jointly optimized with the training objective of the image captioning task in order to strengthen the conditional relationship between O and X, and we show such feature refinement can lead to features supporting qualitatively and quantitatively improved grounding and results.

We illustrate the overall model in Figure 3. To address issue (1), we introduce a cross-modal retrieval module (yellow box) to retrieve a set of text descriptions T that encode information complementary to detected objects O from the input image. To address issue (2), we introduce an image conditioning module (green box) to strengthen the conditional relationship between the detected object and the input image. Key to our method is that both of these approaches will allow us to leverage large-scale cross-modal models that have recently been introduced. In the rest of this section, we describe how to obtain T in Section 3.2 and how to model the conditional relationship in Section 3.3.

3.2. Text Descriptions

In the last section, we introduce T, a set of text descriptions providing information complementary to the detected objects O. Imagine when one is asked to describe an image,

he/she may first focus on local regions of the image and then gradually merge the local information to generate the final description of the whole image. Similarly, we propose to generate text descriptions for image sub-regions as shown in Figure 4 so that those descriptions contain more details and provide more complete information of the input image that can be merged in the later stage. Instead of training another captioning model for generating descriptions of image sub-regions, which by itself may not be an easy task, we propose to retrieve top-k most relevant descriptions from a description database for each image sub-region, thus turning this into a cross-modal retrieval problem. We describe the three steps for cross-modal retrieval as follows.

Description database The descriptions database is the source of relevant text descriptions for an image sub-region, and we choose the top-k most relevant ones. In this paper, we propose to parse the annotations from the Visual Genome dataset (which is already commonly used to train object detectors on) to construct the description database. Instead of taking the *region descriptions* from Visual Genome, which contain many similar sentences (and we show to be inferior in experiments), we parse the annotations of *attributes* and *relationships*. Specifically, the attribute annotations take the form of "attribute-object" pairs. We first convert the object name to its synset canonical form and then collect all the "attribute-object" pairs. On the other hand, the relationship annotations take the form of "subject-predicate-object" triplets. We similarly convert the subject and object names to their synset canonical forms and then collect all the "subject-predicate-object" triplets. Fi-

(a) Original image and the retrieved top-5 most relevant text descriptions.	(b) Image five crops and the retrieved top-5 text descriptions for the crop in the blue box.	(c) Image nine crops and the retrieved top-5 text descriptions for the crop in the blue box.

1. several kitchen_cabinet
2. with stainless steel surfaces kitchen
3. oven located in kitchen
4. kitchen oven electric stove
5. black stove top oven

1. knife arranged beside rack
2. knife are beside oven
3. knife hanging of rack
4. knife on rack
5. knife organizer

1. toaster_oven sitting on counter
2. control_panel on front of stove
3. control_panel on stove
4. kitchen_appliance on stove
5. toaster_oven on countertop

Figure 4. Retrieved top-5 most relevant text descriptions for **(a)** the original image, **(b)** the image five crops, and **(c)** the image nine crops. For five crops and nine crops, we show the retrieved text descriptions for the crop in the blue box.

nally, we merge all the collected "attribute-object" pairs and "subject-predicate-object" triplets and remove duplicates to construct the descriptions database.

Text description retrieval. Our goal is to retrieve the top-k most relevant text descriptions from the description database given a query of an image sub-region. This involves two sub-problems: (1) how to generate image sub-regions and (2) how to perform cross-modal retrieval between image and text. For (1), we propose to generate the five crops (Figure 4b) and nine crops (Figure 4c) of the original image. These crops may contain multiple objects rather than just a single salient object, which is beneficial for capturing the interaction between objects if we are able to retrieve good text descriptions for the crop. For (2), we propose to leverage the cross-modal joint embedding from CLIP [40] for this cross-modal retrieval problem. The CLIP model has two branches: the image branch CLIP-I and text branch CLIP-T that encode image and text into a global feature representation, respectively. CLIP is trained on large-scale image and text pairs such that paired image and text are pulled together in the embedding space while unpaired ones are pushed apart. With the pre-trained CLIP model, the cross-modal retrieval problem becomes a nearest neighbor search in CLIP's cross-modal embedding space. Specifically, we use CLIP-T to encode all text descriptions in the description database as the searching keys. The image sub-region from five crops and nine crops as well as the original image is encoded by CLIP-I into a query. We then search in the description database for the text descriptions with the top-k highest cosine similarity scores. Therefore, we will have a set of retrieved text descriptions $T = \{t_{i,j,k} | i \in \{\text{original, five, nine}\}, j \in \{1, 2, ..., \#\text{crops}_i\}, k \in \{1, 2, ..., \text{top-}k\}\}$, where subscript i represents whether it is from the original image, five

crop, or nine crop sets; subscript j represents the j-th crop for t_i (e.g. top-left, bottom right, etc. for five crop); $\#\text{crops}_{\{\text{original, five, nine}\}}$ equals to $\{1, 5, 9\}$, respectively; and subscript k represents the top-k-th retrieval. Some examples of the top-5 results are shown in Figure 4.

Text encoding. After retrieving the set of text descriptions $T = \{t_{i,j,k}\}$, we use a pre-trained text encoder to encode each of them into a global representation. In this paper, we use a frozen pre-trained CLIP-T as the text encoder as CLIP is similarly pre-trained on a VL task so that it could better encode relevant information for the target VL task from the retrieved text descriptions. The three steps described above, from constructing the description database to searching it and finally encoding the retrieved text descriptions for T, can be done offline for each image in the benchmark dataset in the same way as the detected objects O. To further distinguish between different i (original, five crop, or nine crop) and j (j-th crop for t_i), we add a learnable embedding to $t_{i,j,k}$ for different i and j.

3.3. Image Conditioning

In Section 3.1, we propose to model and strengthen the conditional relationship between the detected object O and the input image X so that the features computed by the object detector can be refined before sending into the captioning model. Since the text descriptions are also retrieved offline by the pre-trained CLIP model, we similarly want to strengthen the conditional relationship between the retrieved text descriptions T and the input image X. As shown in the green box of Figure 3, we propose to condition each detected object and retrieved text description on the input image, and model this conditional relationship by fully connected (FC) layers.

To condition the detected objects $\boldsymbol{o}$ and the retrieved

text descriptions t on the input image x, we first encode x into a global representation $f_x \in \mathbb{R}^{d_x}$. We require that the encoded representation for x preserves as much information relevant to the target VL task as possible. In this paper, we use a frozen pre-trained CLIP-I as the image encoder as it is similarly pre-trained on a VL task such that it could better encode information relevant to the target VL task from the input image. We use the following notations: $o = \{o_1, o_2, ..., o_n\}$, where each $o \in \mathbb{R}^{d_o}$, is a set of objects detected by a frozen pre-trained object detector; and $t = \{t_{i,j,k} | \forall i, j, k\}$, where each $t_{i,j,k} \in \mathbb{R}^{d_t}$, is the set of retrieved text descriptions encoded by CLIP-T (described in Section 3.2). We then model the conditional relationship as:

$$\hat{o}_m = \text{drop}(\text{fc}_o(\text{norm}_o([o_m, f_x])))$$
$$\hat{t}_{i,j,k} = \text{drop}(\text{fc}_{t_i}(\text{norm}_{t_i}([t_{i,j,k}, f_x]))),$$

(5)

where $[\cdot, \cdot]$ is concatenation along the feature dimension, *norm* is a layer normalization layer, and *drop* is a dropout layer. Note that we encode each t_i separately with different FC layers and norm layers as those are text descriptions retrieved for different granularity (original, five crops, or nine crops). Finally, we collect the image conditioned sequences: $\hat{o} = \{\hat{o}_1, \hat{o}_2, ..., \hat{o}_n\}$, $\hat{t}_i = \{\hat{t}_{i,j,k} | \forall j, k\}$.

3.4. Image Captioning

Incorporating the image conditioned objects $\hat{o}$ and text descriptions $\hat{t}_i$ into an image captioning model is simple. As shown in Equation 1, an image captioning model is typically an auto-regressive model $p(y_i | o, y_{1:i-1})$, which takes a sequence of detected objects o as input. Therefore, without modifying the image captioning model, we simply need to concatenate the image conditioned objects $\hat{o}$ and text descriptions $\hat{t}_i$ along the sequence dimension as $z = [\hat{o}, \hat{t}_{\text{original}}, \hat{t}_{\text{five}}, \hat{t}_{\text{nine}}]$, and feed z into it in place of o as $p(y_i | z, y_{1:i-1})$. The model can then be trained with the commonly used maximum log-likelihood loss for word prediction and fine-tuned with the RL loss using CIDEr score as reward [9, 42] in the same way as before.

4. Experiments

4.1. Implementation Details

In this paper, we incorporate our method into one of the state-of-the-art image captioning models, $\mathcal{M}^2$ [9], and train and evaluate our method on the Karpathy split [24] of the MS-COCO dataset [31]. We tune the top-k parameter on the validation set and find that the performance saturates at $k = 12$. Therefore, we set $k = 12$ across all the experiments. In our proposed model shown in Figure 3, the image encoder CLIP-I and text encoder CLIP-T are both frozen. Only the FC layers are trainable, which contains one order fewer parameters compared to the image captioning model.

Method	B-1	B-4	M	R	C	S
SCST [42]	-	34.2	26.7	55.7	114.0	-
Up-Down [1]	79.8	36.3	27.7	56.9	120.1	21.4
RFNet [23]	79.1	36.5	27.7	57.3	121.9	21.2
HIP [60]	-	38.2	28.4	58.3	127.2	21.9
GCN-LSTM [59]	80.5	38.2	28.5	58.3	127.6	22.0
SGAE [58]	80.8	38.4	28.4	58.6	127.8	22.1
ORT [17]	80.5	38.6	28.7	58.4	128.3	22.6
AoANet [19]	80.2	38.9	29.2	58.8	129.8	22.4
$\mathcal{M}^2$ [9]	80.8	39.1	29.1	58.4	131.2	22.6
$\mathcal{M}^{2\dagger}$	80.2	38.4	29.1	58.4	128.7	22.9
$\mathcal{M}^{2\dagger}$ + Ours	**81.5**	**39.7**	**30.0**	**59.5**	**135.9**	**23.7**

Table 1. Image captioning results on the test set of MS-COCO Karpathy split [24]. We incorporate our proposed method into the baseline image captioning model $\mathcal{M}^2$ using their released code. For fair comparison, we also show the performance of the released code (denoted as $\mathcal{M}^{2\dagger}$), which is slightly lower than those reported in the paper.

Model	Detector Pre-training	Transformer Pre-training	B-4	C	S
$\mathcal{M}^{2\dagger}$	VinVL [63]	None	40.5	135.9	23.5
$\mathcal{M}^{2\dagger}$ + Ours	VinVL	None	**41.4**	**139.9**	**24.0**
OSCAR [29]	VG [27]	6.5M	40.5	137.6	22.8
VinVL [63]	VinVL	8.85M	40.9	140.4	**25.1**
OSCAR + Ours	VG	6.5M	**41.3**	**142.2**	24.9

Table 2. Detector pre-training *v.s.* transformer pre-training. When combined with other advanced pre-training techniques, our proposed method achieves competitive performance against VinVL [63], state-of-the-art method with *both* large-scale detector and transformer pre-training.

4.2. Main Results

We first compare with the trained-from-scratch methods in Table 1. We show the results on the test set with cross-entropy training and then SCST [42] RL fine-tuning. With complementary information provided by the retrieved text descriptions and image conditioning, our method improves the baseline model $\mathcal{M}^{2\dagger}$ by $+7.2\%$ in CIDEr and $+1.3\%$ in BLEU-4, and compares favorably with all previous trained-from-scratch methods across all metrics.

We then compare with the methods with more advanced pre-training techniques in Table 2, and show that when combined together, our method performs competitively against state-of-the-art VinVL [63]. Specifically, VinVL improves the object features by pre-training a larger object detector model on large training corpora that combine multiple object detection datasets, as opposed to the conventional approaches that pre-train the object detector on the Visual Genome (VG) dataset. On the top half of Ta-

Text	Image	B-1	B-4	C	S
		75.74	35.47	112.39	20.41
	✓	77.33	36.96	116.84	21.41
✓		77.07	37.12	116.99	21.30
✓	✓	**77.45**	**37.74**	**118.87**	**21.45**

Table 3. Ablation study for the proposed text descriptions (denoted as *Text*) and image conditioning (denoted as *Image*). The first row without *Text* and *Image* corresponds to the baseline model using only detected objects for encoding the input image.

ble 2, we can see that our method is able to provide information complementary to VinVL pre-trained detector and achieves better performance compared with $\mathcal{M}^{2\dagger}$ + VinVL detector. On the other hand, recent methods (*e.g.* OSCAR [29] and VinVL [63]) propose to pre-train the cross-modal transformer on large corpora of image-caption pairs and achieve SoTA performance. On the lower half of Table 2, by combining our method with OSCAR [29], our method achieves competitive performance compared to VinVL, which requires *both* large-scale detector pre-training and transformer pre-training. Lastly, by comparing OSCAR and OSCAR+Ours, we verify our claim that our proposed method indeed provides information in addition to objects, as OSCAR explicitly includes object tags of the image as part of the inputs to the model.

4.3. Analysis

To verify the effectiveness of the proposed text description module and the image conditioning module, as well as the design choices made within each module, we provide detailed analyses in this section.

Ablation study. We ablate the two major components proposed in this paper: (1) text descriptions and (2) image conditioning. We use the $\mathcal{M}^2$ model trained with cross-entropy loss as the baseline model, which uses only the detected objects to encode the input image. Results are shown in Table 3. We can see that adding either one of text descriptions or image conditioning brings substantial performance improvement compared to the baseline model. Overall, with these two modules combined, our proposed method achieves $+6.5\%$ performance improvement in CIDEr and $+2.3\%$ in BLEU-4. This means the proposed two components are indeed capable of providing information complementary to the detected objects, and the complementary information is beneficial for image captioning. Qualitatively, we show some top-5 retrieved text descriptions in Figure 4, where we can see that the retrieved text descriptions well describe the image sub-regions and provide complementary information such as object predicates.

Text descriptions retrieval by visual similarity. In Section 3.2, we propose to leverage the cross-modal joint
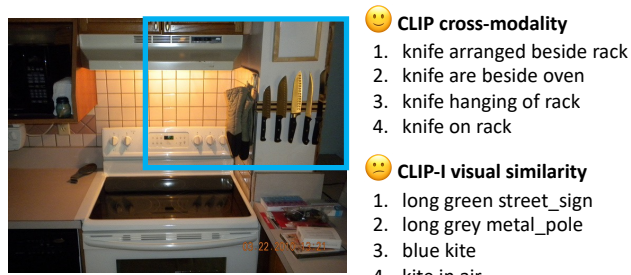

Figure 5. Retrieved text descriptions for the image sub-region in the blue box from *(top)* cross-modal joint embedding by CLIP *v.s.* *(bottom)* visual similarity by CLIP-I.

Encoder	B-1	B-4	C	S
R-101 [16]	75.82	36.27	112.77	20.54
ViT [12]	76.00	35.13	112.53	20.52
MoCo-v2 [5]	76.03	35.35	112.06	20.51
CLIP-I [40]	75.87	36.52	113.66	20.66
Object only baseline	75.74	35.47	112.39	20.41
Ours (CLIP)	**77.07**	**37.12**	**116.99**	**21.30**

Table 4. Retrieving text descriptions by visual similarity with different image encoders.

embedding from CLIP to retrieve text descriptions for each image crop. Another common way of performing cross-modal retrieval is to search for visually similar images and return the paired text of the image [14, 18, 35, 46]. In the analysis, we first encode the image crop as query and encode the bounding boxes (bboxes) annotations from Visual Genome as key with an image encoder (either R-101 [16], MoCo-v2 [5], ViT [12] or CLIP-I [40]). Then the top-k bboxes with the highest cosine similarity scores are selected and the text descriptions associated with the selected bboxes are returned as the retrieved text descriptions. Finally, we train the $\mathcal{M}^2$ model with detected objects and the retrieved text descriptions. We can see in Table 4 that incorporating text descriptions retrieved by visual similarity does not bring much performance improvement compared to the object-only baseline. Qualitatively, we can see in Figure 5 that many of the retrieved text descriptions are irrelevant to the query image. On the other hand, by leveraging the cross-modal joint embedding by CLIP, the retrieved text descriptions are highly relevant to the query image. Therefore, the retrieved text descriptions can provide complementary and relevant information, which is beneficial.

How to retrieve text descriptions? In Section 3.2, we describe how to construct the description database where the text descriptions are retrieved from, and how to retrieve text descriptions by image crops (see Figure 4). In this subsection, we answer the following questions: (1) Is

Caption dataset	B-4	C	S
-	35.47	112.39	20.41
COCO	**37.37**	**117.17**	21.19
CC	37.30	117.13	21.14
VG (Ours)	37.12	116.99	**21.30**

Table 5. Text descriptions retrieved from different image captioning datasets.

Query	B-4	C	S
-	35.47	112.39	20.41
Whole	36.39	115.98	20.94
Five	37.04	116.73	21.21
Nine	36.95	116.03	21.10
All (Ours)	**37.12**	**116.99**	**21.30**

Table 6. Text descriptions retrieved by using different image crops as query.

Conditioning Method	#Tokens	B-4	C	S
-	50	35.47	112.39	20.41
TF-V	51	36.66	116.01	21.16
TF-G	100	36.75	116.22	21.36
FC (Ours)	50	**36.96**	**116.84**	**21.41**

Table 7. Different image conditioning methods. See text for detailed descriptions of TF-V and TF-G. We use CLIP-I to encode the input image.

Encoder	Pre-trained Dataset	B-4	C	S
R-101 [16]	IN-1K [10]	35.64	113.20	21.00
BiT [26]	JFT-300M [47]	36.08	114.00	20.95
ViT [12]	IN-21K [10]	35.97	113.19	20.82
CLIP-I [40]	400M [40]	**36.96**	**116.84**	**21.41**

Table 8. Image conditioning with image encoded by different pre-trained models. We use the FC method to incorporate the encoded image features.

the final performance sensitive to the description database? (2) How effective is the proposed text retrieval method by using different image crops as query? We train the $\mathcal{M}^2$ model with detected objects and the text descriptions retrieved with different methods. For (1), we construct the description databse on different image captioning datasets including Visual Genome [27], MS-COCO [31], and Conceptual Captions [44]. In Table 5, we show that the performance is *not* sensitive to the description databases the text descriptions are retrieved from. For (2), we ablate the text retrieval strategies of using the whole image, five crops, nine crops or the combination of all the above in Table 6. We can see that using image crops to retrieve more fine-grained text descriptions is beneficial. Qualitative results of the retrieved fine-grained text descriptions using different image crops are shown in Figure 4.

Modeling of image conditioning. In Section 3.3, we model the conditional relationship between the detected objects and the input image by an FC layer. In this subsection, we answer the following questions: (1) What is the better way to model image conditioning? (2) What is the better pre-trained image encoder to encode the input image for image conditioning? We train the $\mathcal{M}^2$ model with detected objects with different image conditioning methods. For (1), an alternative is to treat the image features as an additional token, send them into the captioning model together with the set of detected objects, and let the transformer module inside the captioning model learn the conditional relationship. The image features could be a single token of its global vector representation [21, 45, 64] (denoted as TF-V) or a sequence of tokens of the grid features (denoted as TF-G). In Table 7, we see that our proposed simple FC method is the most effective. Since the computational complexity grows

quadratically with respect to the input sequence length for a transformer model, substantially increasing the length of the input sequence (#Tokens) such as TF-G is undesirable. By using an FC layer to fuse the features of the detected objects and the input image, we do not increase the length of input tokens. On the other hand, even though TF-V only increase #Tokens by one, it underperforms our proposed FC method.

For (2), we also propose to use CLIP-I to encode the input image for image conditioning in Section 3.3. We claim that CLIP-I pre-trained on a similar VL task is capable of preserving as much relevant information from the input image as possible compared to the models pre-trained on image-only datasets such as ImageNet (IN) 1K/21K [10] or JFT-300M [47]. To verify this claim, we compare image encoded by different image encoders, R-101 [16], BiT [26], ViT [12], and CLIP-I [40]. In Table 8, we can see that using CLIP-I, which is pre-trained on V+L tasks, as the image encoder performs substantially better than using R-101, BiT, and ViT, which are all pre-trained on image-only datasets.

How does image conditioning help. In Section 3.3, we claim that jointly optimizing the conditional relationship between detected objects and input image helps refine the object features to aid with grounding. To verify this, we train the captioning model on the Flickr30k [38] dataset, which provides the grounding annotations between image and caption pairs. Following the standard approach as other image captioning papers, we find the most attended object using Integrated Gradient [48] for each word. Out of 1,014 validation images, our method that refines the object features by image conditioning correctly localizes ***421 objects***, while the baseline $\mathcal{M}^2$ only localizes ***287 objects***. More qualitative results can be found in the supplementary.

5. Conclusion

In this paper, we address limitations of using pre-trained frozen object detectors as the sole input to auto-regressive models in image captioning. We specifically propose to add an auxiliary branch in the graphical model, leveraging advances in large pre-scaled multi-modal models to retrieve (from the same dataset that the object detector is pre-trained on) contextual attribute and relationship descriptions. Further, we refine both the detector outputs and retrieved context descriptors in an image-conditioned manner, through a simplified architectural design that avoids significant computational overhead, and show that such conditioning improves quantitative results. We perform thorough analysis demonstrating that both retrieved text and image conditioning improve results (and jointly even more so), that the multi-modal CLIP model is uniquely suited to our approach, and grounding improvements. We also demonstrate significant performance improvements, up to +7.5% in CIDEr and +1.3% over already strong state of art.

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018. 1, 2, 6

[2] Fuhai Chen, Rongrong Ji, Jinsong Su, Yongjian Wu, and Yunsheng Wu. Structcap: Structured semantic embedding for image captioning. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 46–54, 2017. 2

[3] Fuhai Chen, Rongrong Ji, Xiaoshuai Sun, Yongjian Wu, and Jinsong Su. Groupcap: Group-based image captioning with structured relevance and diversity constraints. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1345–1353, 2018. 2

[4] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng 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*, pages 5659–5667, 2017. 2

[5] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 7

[6] Xinlei Chen and C Lawrence Zitnick. Mind's eye: A recurrent visual representation for image caption generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2422–2431, 2015. 2

[7] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020. 2

[8] Cesc Chunseong Park, Byeongchang Kim, and Gunhee Kim. Attend to you: Personalized image captioning with context sequence memory networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 895–903, 2017. 3

[9] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10578–10587, 2020. 1, 6

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009. 8

[11] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015. 2

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 7, 8

[13] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. From captions to visual concepts and back. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1473–1482, 2015. 2

[14] Yunchao Gong, Liwei Wang, Micah Hodosh, Julia Hockenmaier, and Svetlana Lazebnik. Improving image-sentence embeddings using large weakly annotated photo collections. In *European conference on computer vision*, pages 529–545. Springer, 2014. 2, 7

[15] Jiuxiang Gu, Gang Wang, Jianfei Cai, and Tsuhan Chen. An empirical study of language cnn for image captioning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1222–1231, 2017. 2

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2, 7, 8

[17] Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. Image captioning: Transforming objects into words. *arXiv preprint arXiv:1906.05963*, 2019. 2, 6

[18] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899, 2013. 2, 7

[19] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on attention for image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4634–4643, 2019. 2, 6

[20] Lun Huang, Wenmin Wang, Yaxian Xia, and Jie Chen. Adaptively aligned image captioning via adaptive attention time. *Advances in Neural Information Processing Systems*, 32:8942–8951, 2019. 2

[21] Jiayi Ji, Yunpeng Luo, Xiaoshuai Sun, Fuhai Chen, Gen Luo, Yongjian Wu, Yue Gao, and Rongrong Ji. Improving image captioning by leveraging intra-and inter-layer global representation in transformer network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1655–1663, 2021. 8

[22] Xu Jia, Efstratios Gavves, Basura Fernando, and Tinne Tuytelaars. Guiding the long-short term memory model for image caption generation. In *Proceedings of the IEEE international conference on computer vision*, pages 2407–2415, 2015. 2

[23] Wenhao Jiang, Lin Ma, Yu-Gang Jiang, Wei Liu, and Tong Zhang. Recurrent fusion network for image captioning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 499–515, 2018. 2, 6

[24] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015. 1, 2, 6

[25] Lei Ke, Wenjie Pei, Ruiyu Li, Xiaoyong Shen, and Yu-Wing Tai. Reflective decoding network for image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8888–8897, 2019. 1, 2

[26] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 491–507. Springer, 2020. 8

[27] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. 2016. 1, 6, 8

[28] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 2

[29] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020. 3, 6, 7

[30] Yikang Li, Wanli Ouyang, Bolei Zhou, Kun Wang, and Xiaogang Wang. Scene graph generation from objects, phrases and region captions. In *Proceedings of the IEEE international conference on computer vision*, pages 1261–1270, 2017. 3

[31] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 3, 6, 8

[32] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*, 2019. 2

[33] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 375–383, 2017. 2

[34] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). *arXiv preprint arXiv:1412.6632*, 2014. 2

[35] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24:1143–1151, 2011. 2, 7

[36] Cesc Chunseong Park, Byeongchang Kim, and Gunhee Kim. Attend to You: Personalized Image Captioning with Context Sequence Memory Networks. In *CVPR*, 2017. 3

[37] Cesc Chunseong Park, Byeongchang Kim, and Gunhee Kim. Towards personalized image captioning via multimodal memory networks. *IEEE transactions on pattern analysis and machine intelligence*, 41(4):999–1012, 2018. 3

[38] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015. 8

[39] Yu Qin, Jiajun Du, Yonghua Zhang, and Hongtao Lu. Look back and predict forward in image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8367–8375, 2019. 1, 2

[40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2, 5, 7, 8

[41] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99, 2015. 1

[42] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7008–7024, 2017. 2, 6

[43] Philipp Rimle, Pelin Dogan-Schönberger, and Markus Gross. Enriching video captions with contextual text. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 5474–5481. IEEE, 2021. 3

[44] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 8

[45] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*, 2020. 2, 8

[46] Chen Sun, Chuang Gan, and Ram Nevatia. Automatic concept discovery from parallel text and visual corpora. In *Proceedings of the IEEE international conference on computer vision*, pages 2596–2604, 2015. 2, 7

[47] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 843–852, 2017. 8

[48] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pages 3319–3328. PMLR, 2017. 8

[49] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019. 2

[50] Junjiao Tian and Jean Oh. Image captioning with compositional neural module networks. *IJCAI*, 2019. 3

[51] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption gen-

erator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015. 1, 2

[52] Li Wang, Zechen Bai, Yonghua Zhang, and Hongtao Lu. Show, recall, and tell: Image captioning with recall mechanism. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12176–12183, 2020. 2

[53] Spencer Whitehead, Heng Ji, Mohit Bansal, Shih-Fu Chang, and Clare Voss. Incorporating background knowledge into video description generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3992–4001, 2018. 3

[54] Qi Wu, Chunhua Shen, Lingqiao Liu, Anthony Dick, and Anton Van Den Hengel. What value do explicit high level concepts have in vision to language problems? In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 203–212, 2016. 2

[55] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5410–5419, 2017. 3

[56] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015. 2

[57] Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. Graph r-cnn for scene graph generation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 670–685, 2018. 3

[58] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-encoding scene graphs for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10685–10694, 2019. 6

[59] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In *Proceedings of the European conference on computer vision (ECCV)*, pages 684–699, 2018. 6

[60] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Hierarchy parsing for image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2621–2629, 2019. 6

[61] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4651–4659, 2016. 2

[62] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5831–5840, 2018. 3

[63] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2021. 3, 6, 7

[64] Xuying Zhang, Xiaoshuai Sun, Yunpeng Luo, Jiayi Ji, Yiyi Zhou, Yongjian Wu, Feiyue Huang, and Rongrong Ji. Rst-

net: Captioning with adaptive attention on visual and nonvisual words. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15465–15474, 2021. 8