Bridging the Gap between Vision and Language Domains for Improved Image Captioning

Fenglin Liu
ADSPLAB, School of ECE
Peking University
fenglinliu98@pku.edu.cn

Xian Wu
Medical AI Lab
Tencent
kevinxwu@tencent.com

Shen Ge
Medical AI Lab
Tencent
shenge@tencent.com

Xiaoyu Zhang
School of ECE
Peking University
zxy2019@pku.edu.cn

Wei Fan
Medical AI Lab
Tencent
Davidwfan@tencent.com

Yuexian Zou*
ADSPLAB, School of ECE
Peking University
zouyx@pku.edu.cn

ABSTRACT

Image captioning has attracted extensive research interests in recent years. Due to the great disparities between vision and language, an important goal of image captioning is to link the information in visual domain to textual domain. However, many approaches conduct this process only in the *decoder*, making it hard to understand the images and generate captions effectively. In this paper, we propose to bridge the gap between the vision and language domains in the *encoder*, by enriching visual information with textual concepts, to achieve deep image understandings. To this end, we propose to explore the textual-enriched image features. Specifically, we introduce two modules, namely Textual Distilling Module and Textual Association Module. The former distills relevant textual concepts from image features, while the latter further associates extracted concepts according to their semantics. In this manner, we acquire textual-enriched image features, which provide clear textual representations of image under no explicit supervision. The proposed approach can be used as a plugin and easily embedded into a wide range of existing image captioning systems. We conduct the extensive experiments on two benchmark image captioning datasets, i.e., MSCOCO and Flickr30k. The experimental results and analysis show that, by incorporating the proposed approach, all baseline models receive consistent improvements over all metrics, with the most significant improvement up to 10% and 9%, in terms of the task-specific metrics CIDEr and SPICE, respectively. The results demonstrate that our approach is effective and generalizes well to a wide range of models for image captioning.

CCS CONCEPTS

• **Computing methodologies** → **Scene understanding**; *Natural language generation.*

*Corresponding author. Also with Peng Cheng Laboratory.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
MM '20, October 12–16, 2020, Seattle, WA, USA
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3414004

KEYWORDS

Image Captioning, Image Representations, Textual Concepts, Attention Mechanism

ACM Reference Format:
Fenglin Liu, Xian Wu, Shen Ge, Xiaoyu Zhang, Wei Fan, and Yuexian Zou. 2020. Bridging the Gap between Vision and Language Domains for Improved Image Captioning. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3394171.3414004

1 INTRODUCTION

Different from the low-level vision tasks, e.g., image denoising [19] and image super-resolution [52], vision-and-language tasks, such as image captioning [8] and visual question answering [4], belong to the category of high-level vision tasks. In particular, for the task of image captioning, there are great disparities between source (vision) and target (language) domains. For example, under the unsupervised learning setting in image captioning, Feng et al. [10] shows that the image captioning system needs to take extra supervised information (i.e., textual concepts [9, 33, 49]) as input to ensure the relevance of the generated descriptive sentence to the input image. The reason is that the textual concepts can be object words (e.g., *car*), attribute words (e.g., *wooden*) or relationship words (e.g., *sitting*), which carry important visual information in the language domain. To the contrary, in image denoising, Krull et al. [19] shows that the systems only need to take the original image as input. This phenomena indicates that there are indeed great disparities between visual and textual domains. As a result, in image captioning, where the image representations are used for text-oriented purposes, it is often desirable to translate the implicit low-level visual information to explicit high-level textual information [43]. It also suggests that the systems should **1)** first link the image modality to the text modality, then **2)** generate a semantically and grammatically correct descriptive sentence based on such textual guided visual information.

However, existing systems [3, 27, 47] entangle these two steps in one single model, i.e., the *decoder*. And due to the large disparities between vision and language domains, the *decoders* in such models have to devote most of their capabilities on conducting the step **1)**, and thus be distracted from finishing step **2)** with high quality. Especially, the *decoders*, e.g., LSTMs [12] and Transformers [39], are

effective at propagating information across the decoding process and generating a semantically and grammatically correct descriptive sentence, and is not good at conducting step **1)**. As a result, ignoring the possibility to construct two separate models to disentangle the two steps makes it hard for these systems to understand the images and generate captions efficiently. In this work, we argue that exploring another separate model to perform the step **1)** is helpful to bridge the gap between vision and language domains and improve the performance.

To achieve that, we propose to conduct the step **1)** in the *encoder* by enriching image features with textual concepts. Specifically, we provide complete semantics information for the image features. In implementations, our approach consists of a Textual Distilling Module (TDM) and a Textual Association Module (TAM). Take Figure 2 for example, the TDM focuses on distilling textual concepts to the corresponding visual objects, like "woman", "tennis", "field", "standing" and "playing" in the red box. Since the textual concepts are single words and only represent a fraction of a semantics, we further introduce the TAM to group related concepts to form a complete semantics. For instance, we group three separated concepts "woman", "tennis" and "playing" into a phrase "woman playing tennis". In this manner, we can connect visual objects to well-organized textual concepts and in turn generate textual-enriched image features. Such a process is essential in the *encoder* to bridge the gap between vision and language domains, and is able to obtain textual-enriched image features generally not explicitly learned by existing systems. It is worth noticing that since we enrich the image features with textual concepts in the *encoder*, the textual-enriched image features can be easily integrated into existing image captioning models. An example for equipping with our approach in baseline models is illustrated in Figure 1.

Overall, the contributions of this paper are as follows:

- For achieving a deep image understanding efficiently, we introduce a novel approach to bridge the gap between vision and language domains in the encoding process, which provides a solid bias for image captioning.
- We implement the proposed approach by enriching the image features with textual concepts. First, we introduce the Textual Distilling Module to distill relevant textual concepts for each visual object; Then, we introduce the Textual Association Module to group these concepts according to their semantics, and enrich the image features with grouped textual concepts.
- The textual-enriched image features are able to bring consistent performance gains to existing models. According to our experimental results on the Flickr30k and MSCOCO datasets, after equipping textual-enriched image features, the baseline models receive up to 10% gain in CIDEr and 9% gain in SPICE. Besides, we further validate the importance of modeling the relationships between vision and language domains, rather than simple incorporating them.

2 RELATED WORK

In this section, we will introduce the related work from three aspects: 1) image captioning; 2) image representations and 3) dense captioning.

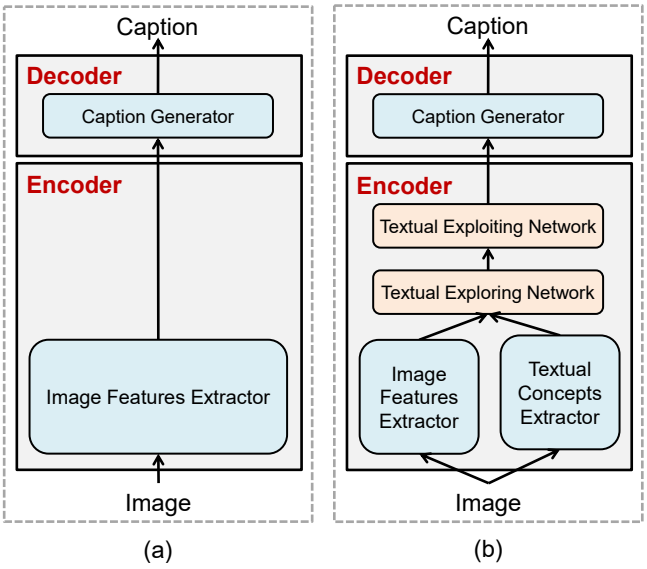


Figure 1: Illustration of how to use our approach on the baseline image captioning systems. (a) The data flow in the baseline model and (b) the data flow in the baseline w/ proposal model. It is worth noticing that we focus on bridging the gap between the vision and language domains in the encoding process, which means when training the baseline w/ proposal model, the proposal is trained together with baseline. Especially we preserve the original experimental settings and training strategy.

2.1 Image Captioning

In recent years, a large number of neural systems have been proposed for image captioning [3, 41, 44, 46, 47]. The state-of-the-art approaches [3, 11] incorporate the attention mechanism [6, 44] to translate low-level image representations (Region-CNN image features) to high-level textual information (image captions) [43]. However, these state-of-the-art systems do not use image concepts (textual concepts) to guide caption generation, making it difficult for the decoder to generate textual captions directly from low-level image features.

In order to provide high-level image representations to the decoder, Wu et al. [43], You et al. [50] and Yao et al. [49] proposed to use a set of textual concepts as textual features. However, 1) these textual concepts are not associated with image features; 2) the relationships among the individual textual concepts are not well explored. In brief, textual concepts are not fully explored in above mentioned methods. In our approach, we consider both the association of textual concepts with image features and the relationship between the concepts, which achieves a better textual understanding of images.

2.2 Image Representations

In vision-and-language tasks, such as image captioning [8], visual question answering [4], visual storytelling [13] and referring expressions [17], it is vital to understand the input images effectively and generates fine-grained image representations. In the literature,

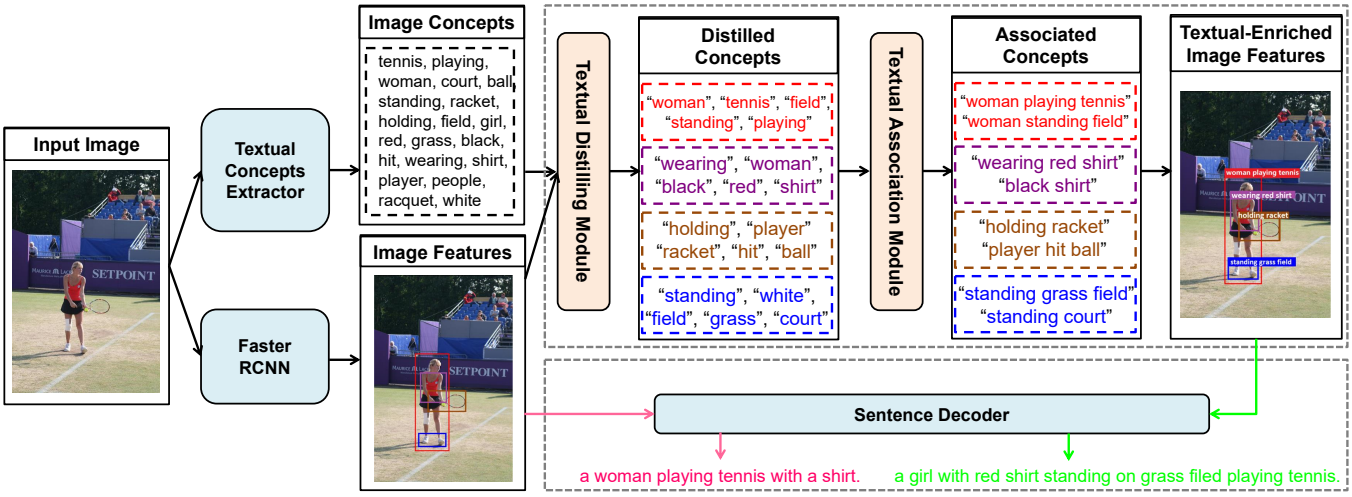


Figure 2: Illustration of the proposed approach. The image features is captured by Faster-RCNN [35] and the textual concepts are extracted by a textual concepts extractor [9]. The Textual Distilling Module then adaptively distill the related textual concepts for each region feature. The Textual Association Module focuses on associating the distilled textual concepts of the image features to come up with some phrases, and form a unified textual-enriched image representation. The color Pink and Green denote the caption generated by the vanilla image features and the textual-enriched image features, respectively, based on the Up-Down model [3]. As we can see, the textual-enriched image features helps the baseline model to generate more complete and coherent captions.

a number of neural approaches have been proposed to obtain image representations in various forms. The most-widely used methods are to extract image features using CNNs or Region-CNNs, where the former split an image into a uniform grid of visual regions and the latter produce object-level image features based on predicted bounding boxes. For image captioning, Fang et al. [9], Liu et al. [28], Wu et al. [43], You et al. [50] augmented the information source with textual concepts that are given by a concept extractor, which is trained to find the most frequent words in the captions. Regardless of the type of image representations, relationships among the individual parts of representations (regions or concepts) are not defined, which should be essential to a deep understanding of images. Recently, Yao et al. [47] and Liu et al. [26, 27] explored the visual relationships among the individual parts of representations, which provides a solid basis for downstream tasks. Specifically, Yao et al. [47] and Liu et al. [26, 27] attempted to use graph networks and attention mechanism to explore visual relationships, respectively. The graph-based approaches explicitly model the spatial and semantic relationships of visual information, while the attention-based methods accomplish that in implicit ways. Besides, Liu et al. [25] focused on exploring the textual relationships among the individual textual concepts to perform the unpaired image captioning task. In our approach, we not only explored both the relationship between the concepts, but also considered the association of textual concepts with image features.

It is worth noticing that several pre-trained vision-and-language models [1, 21, 30, 36–38, 56] have been proposed to learn task-agnostic joint representations of vision and language (a.k.a. vision-language representations) for various tasks. However, most existing systems only use the region-of-interests (RoIs) / video frames as the image / video features and do not consider to learn such joint representations by incorporating the textual concepts. As a result, there are still huge gaps between the vision and language domains.

2.3 Dense Captioning

Our approach also relates to the effort of dense captioning [15, 45]. However, compared with our approach, the major difference lies in the training labels. These dense captioning methods [15, 45] explicitly provide a series of region captions for training, while our model does not require such additional ground truth labels. Instead, we learn the regional textual information implicitly from the complete descriptions of images without explicit region alignment. Therefore our approach can be directly applied to existing image captioning models.

3 APPROACH

As shown in Figure 2, there are three main steps to generate text-enriched image representations: (1) Image encoder and textual concepts extractor: this step extracts image features and textual concepts from images; (2) Textual distilling module: this module distills related textual concepts for visual objects in images; (3) Textual association module: this module is necessary since the distilled textual concepts are independent and do not associate with each other, e.g., the three words *riding*, *boy* and *bike* should be associated together as a phrase *boy riding bike* to represent a complete semantics. With the above steps, we are able to generate textual-enriched image features, which can improve the performance of image captioning task and increase the model interpretability at the same time. In the following sections, we will describe these three steps in detail.

3.1 Image Encoder and Concept Extractor

Given an image, we extract two types of information: image features and textual concepts. For the image features, we apply the Region-CNN model, which is pre-trained on Visual Genome [18] and is proposed by Anderson et al. [3], and acquire a $\boldsymbol{v} \in \mathbb{R}^d$ vector; For the textual concepts, which contain rich visual semantics, and have been used to provide explicit high-level semantic information of an image [43]. Following the previous works [20, 27, 28, 49], we apply the concept extractor proposed by Fang et al. [9], which is built upon a weakly-supervised approach of Multiple Instance Learning [53]. In particular, the concept extractor is trained on the MSCOCO caption dataset for 1,000 visual concepts. Given the input image, this extractor will output a set of words as textual concepts. The extracted textual concepts can be either objects (e.g. *bike*, *tree*), attributes (e.g. *green*, *young*), or relationships (e.g. *riding*, *wearing*), representing explicit high-level information of the image [43, 50]. For each image, only the top $m = 20$ textual concepts are selected. We represent the extracted textual concepts with a set of vectors: $\boldsymbol{T} = \{\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_m\} \in \mathbb{R}^{m \times e}$ where $\boldsymbol{w}_i \in \mathbb{R}^e$ refers to the embedding of the i^{th} concept.

3.2 Textual Distilling Module

In previous step, we manage to detect objects in the image and represent each object with a image feature $\boldsymbol{v}$, we also acquire textual concepts from this image. Textual Distilling Module targets to find the most relevant textual concepts from the set $\boldsymbol{T}$ for each image feature $\boldsymbol{v}$. To achieve this, we apply the attention mechanism [31, 44]. According to the attention mechanism, we feed the image feature and the concept embeddings into a single layer neural network to generate the attention scores $\boldsymbol{S} \in \mathbb{R}^m$:

$$\boldsymbol{S} = \boldsymbol{w}_S \tanh(\boldsymbol{W}_{S,v}\boldsymbol{v} \oplus \boldsymbol{W}_{S,T}\boldsymbol{T}^{\mathrm{T}}) \tag{1}$$

where $\boldsymbol{W}_{S,v} \in \mathbb{R}^{m \times d}$, $\boldsymbol{W}_{S,T} \in \mathbb{R}^{m \times e}$ and $\boldsymbol{w}_S \in \mathbb{R}^m$ are the learnable parameters.[1] According to the computed attention scores $\boldsymbol{S}$, we select top-k related textual concepts as the textual enrichments of the current input region feature $\boldsymbol{v}$. This is done by selecting k largest scores and recording their index positions in the $\boldsymbol{S}$ to form a set $\boldsymbol{P}$ [54], where k is a hyper-parameter. Then, we update the value of each position S_i in the attention scores $\boldsymbol{S}$ by the formula below:

$$S_k(i) = \begin{cases} S(i), & \text{if } i \in \boldsymbol{P} \\ -\infty, & \text{if } i \notin \boldsymbol{P} \end{cases} \tag{2}$$

After that, a softmax function is applied on $\boldsymbol{S}_k \in \mathbb{R}^m$ to generate the attention distribution $\boldsymbol{\alpha}_k \in \mathbb{R}^m$ over the m textual concepts:

$$\boldsymbol{\alpha}_k = \mathtt{softmax}(\boldsymbol{S}_k) \tag{3}$$

Based on the attention distribution, the top-k most relevant textual concepts can be computed as below:

$$\boldsymbol{T}_k = \boldsymbol{\alpha}_k * \boldsymbol{T} \tag{4}$$

Since we assign $-\infty$ to the textual concepts attention scores that are unrelated to the current region feature $\boldsymbol{v}$, the attention weights would be 0 after the application of the softmax function, ensuring our intention of keeping top-k textual concepts. By using this method, we can distill the most relevant k textual concepts, and $\boldsymbol{T}_k \in \mathbb{R}^{m \times e}$ is taken to be the resulting textual enrichment of the region feature.

[1] For conciseness, all the bias terms of linear transformations in this paper are omitted.

3.3 Textual Association Module

As shown in Figure 2, through the Textual Distilling Module, we can relate the textual concepts with image features. However, these extracted textual concepts are single words and represent only a fraction of semantics. Therefore, when we finished the identification of associated single-word textual concepts, we combine them together to come up with some phrases, and further form a complete semantics. Such a complete semantics can be used to enhance the textual representations of image features, which is essential for a deep understanding of images.

To form a complete semantics, we proceed as follows: 1) combine independent semantics representation vectors to a unified semantics representation vector; 2) associate multiple aspects of the extracted top-k textual concepts with multiple vector representations (as shown in Figure 2). We accomplish the above two steps simultaneously by applying the self-attention mechanism provided by Lin et al. [23]. The self-attention mechanism is defined as:

$$\boldsymbol{A} = \mathtt{softmax}(\boldsymbol{W}_A \tanh(\boldsymbol{W}_{A,T}\boldsymbol{T}_k^{\mathrm{T}})) \tag{5}$$

$$\boldsymbol{U} = \boldsymbol{A}\boldsymbol{T}_k \tag{6}$$

where $\boldsymbol{W}_{A,T} \in \mathbb{R}^{m \times e}$ and $\boldsymbol{W}_A \in \mathbb{R}^{n \times m}$ are learnable parameters. n is the number of different association aspects. It means that in order to extract multiple aspects of the extracted top-k textual concepts to form n phrases, we repeat the above steps n times with different learnable parameters. In particular, similar to the Textual Distilling Module, for each $\boldsymbol{a}_i \in \boldsymbol{A}$, we also apply the top-$r$ attention mechanism to extract each phrase. Through the Textual Association Module, we can establish the connections of each individual textual concept to enhance the power of overall textual information.

3.4 Textual-Enriched Image Features

Next, we combine the original region features $\boldsymbol{v} \in \mathbb{R}^d$ along with the corresponding textual information $\boldsymbol{U} \in \mathbb{R}^{n \times e}$ of $\boldsymbol{v}$ to get the textual-enriched image features:

$$\boldsymbol{v}_T = \mathtt{LayerNorm}(\boldsymbol{W}_v\boldsymbol{v} + \boldsymbol{W}_U\mathtt{MeanPooling}(\boldsymbol{U})) \tag{7}$$

where $\boldsymbol{W}_v \in \mathbb{R}^{d \times d}$ and $\boldsymbol{W}_U \in \mathbb{R}^{d \times e}$ are learnable parameters; LayerNorm stands for Layer Normalization [5].

To sum it up, we first apply Textual Distilling Module to extract the most relevant textual concepts, which are further associated with each other by Textual Association Module. In this way, we can extract the textual representations of the image features, and further generate textual concepts enriched image representations, providing a good prior for the following image-based text generation task.

4 EXPERIMENTS

In this section, we first describe two benchmark datasets and the settings, as well as some widely-used metrics. Then we evaluate the proposed approach from three perspectives: (1) the effectiveness of applying our proposed textual-enriched image features to existing works; (2) the effect of incorporating extra information. (3) the online leaderboard performance.

Table 1: Performance on the MSCOCO Karpathy test split [16]. All values in this paper are reported in percentage (%). Higher is better in all columns. All the baselines enjoy a comfortable improvement with the proposed approach. Additionally, we report the performance of the current published state-of-the-art models ORT [11] and HIP [48], as we can see, the *Transformer w/ proposal* outperforms the ORT and HIP substantially in major metrics, which proves our arguments and demonstrates the effectiveness of our approach.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
ORT [11]	80.5	-	-	38.6	28.7	58.4	128.3	22.6
HIP [48]	-	-	-	39.1	28.9	59.2	130.6	22.3
NIC [41]								
Baseline	77.9	60.1	46.3	34.2	26.2	55.7	112.3	19.7
w/ proposal	**79.5**	**63.9**	**50.1**	**37.1**	**27.8**	**58.1**	**120.9**	**21.0**
Adaptive [31]								
Baseline	79.5	62.0	48.7	35.9	27.3	56.1	117.6	20.3
w/ proposal	**80.1**	**64.2**	**49.9**	**37.5**	**27.9**	**57.9**	**121.4**	**21.3**
Up-Down [3]								
Baseline	79.8	63.7	49.5	36.3	27.7	56.9	120.1	21.4
w/ proposal	**80.7**	**65.1**	**50.4**	**37.9**	**28.4**	**58.5**	**124.8**	**22.5**
Transformer [39]								
Baseline	80.7	65.6	51.2	39.2	28.9	58.9	129.4	22.4
w/ proposal	**81.0**	**66.1**	**51.5**	**39.6**	**29.2**	**59.1**	**131.9**	**22.7**

Table 2: Performance on the Flickr30k Karpathy test split. B-n, M, R, C and S are short for BLEU-n, METEOR, ROUGE-L, CIDEr and SPICE, respectively. The GVD [55] is the published state-of-the-art model on Flickr30k dataset. As we can see, we outperforms the state-of-the-art model substantially in terms of CIDEr, which further demonstrates the effectiveness of our approach.

Methods	Flickr30k					
	B-1	B-4	M	R	C	S
GVD [55]	69.9	27.3	22.5	-	62.3	16.5
Adaptive [31]						
Baseline	70.5	26.7	21.0	48.1	57.1	14.6
w/ proposal	**71.8**	**27.9**	**21.6**	**49.3**	**62.7**	**15.9**
NBT [32]						
Baseline	71.4	27.8	21.7	48.8	60.2	15.6
w/ proposal	**73.3**	**29.5**	**22.0**	**49.7**	**65.6**	**16.3**

4.1 Datasets, Metrics and Settings

In this section, we will give a detailed introduction of our used datasets, metrics and settings for evaluation.

4.1.1 Datasets. Our reported results are evaluated on the popular Flickr30k [51] and MSCOCO [8] datasets, which contain 31k images and 12k images, respectively, and each image in the datasets is annotated with 5 sentences. The results are reported using the widely-used publicly-available splits in the work of Karpathy and Li [16]. The MSCOCO validation and test set contain 5,000 images each, and the number is 1,000 images for Flickr30k. Following common practice [3, 29, 31, 41], we replace caption words that occur less than 5 times in the training set with the generic unknown word token <UNK>, resulting in a vocabulary of 9k words for MSCOCO and 7k words for Flickr30k.

4.1.2 Metrics. The metrics SPICE [2], CIDEr [40], BLEU [34], METEOR [7] and ROUGE [22] are used in our tests for performance evaluation. They are widely used and could be reported by the MSCOCO captioning evaluation toolkit [8]. In particular, SPICE [2] is based on scene graph matching, and CIDEr [40] is built upon on n-gram matching. They both incorporate the consensus of a set of references for an example. BLEU [34] and METEOR [7] are originally proposed for machine translation evaluation. ROUGE [22] is designed for measuring the quality of summaries. Among them, SPICE and CIDEr are specifically designed to evaluate image captioning systems and will be the main considering metrics.

4.1.3 Experimental Settings. For fair comparisons, we use the RCNN-based image features provided by Anderson et al. [3]. Specifically, Anderson et al. [3] uses the Faster R-CNN [35], which is pre-trained on Visual Genome [18] dataset, to detect objects. For textual concepts, we use the textual concepts prediction model pre-trained by Fang et al. [9] for 1,000 words. The caption words and the textual concept words share the same embeddings. For our proposal, d and e stands for the hidden size and the word embedding size of the baseline decoder, respectively. The number of textual concepts in the entire image is set to 20, which means $m = 20$. Based on the average performance of all the baselines on the validation set, for the value of k, r and n, we select 10, 4, 10 for them, respectively.

Table 3: Evaluation of representative systems which further take the textual concepts as input. The proposed approach can further improve the already strong baselines in all metrics. The significant improvements come from the textual-enriched image features rather than the simple incorporation of extra features. Besides, it also shows the importance of enriching image features with textual concepts.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
ATT-FCN [50]								
Baseline	78.3	61.7	48.3	35.7	27.4	57.3	117.8	20.7
w/ proposal	**79.4**	**64.0**	**49.9**	**36.6**	**28.1**	**58.0**	**120.1**	**21.1**
LSTM-A3 [49]								
Baseline	78.6	62.5	48.6	35.7	27.1	56.3	116.2	20.1
w/ proposal	**79.7**	**63.7**	**50.1**	**36.7**	**28.0**	**57.5**	**119.1**	**20.6**
LSTM-A4 [49]								
Baseline	78.2	61.9	48.8	35.9	27.3	56.5	113.9	20.3
w/ proposal	**79.4**	**64.4**	**50.1**	**36.8**	**28.0**	**57.5**	**121.3**	**22.2**
LSTM-A5 [49]								
Baseline	78.4	62.8	48.9	36.5	27.5	57.1	116.8	20.5
w/ proposal	**79.9**	**64.4**	**49.7**	**36.5**	**27.7**	**57.7**	**119.6**	**21.8**

Since our focus is to provide textual-enriched image features, we preserve the original parameter settings, and training strategy for all the baselines.

4.2 Baselines

We experiment with two types of baseline models that only use image features and further incorporate textual concepts:

4.2.1 Models based only on image features. For the baselines, we replace the original image features with our textual-enriched image features to evaluate the performance gain. We consider NIC [41], Adaptive [31], Up-Down [3], Transformer [39] and NBT [32] models, which depend entirely on image features to generate image captions.

4.2.2 Models further incorporate textual concepts. We further validate our approach on models that incorporate both visual and textual concepts information. This is to verify whether the improvements are due to our modeling towards image features and textual concepts, rather than the introduction of additional textual concepts.

LSTM-A [49] uses a series of models (LSTM-A3,4,5) to combine image features and textual concepts. We will study the effect of our method on them. LSTM-A3 feeds the textual concepts in the first decoding step and the image feature in the second step. LSTM-A4 provides the textual concepts for the decoder at the beginning and leave image features for the subsequent steps. In contrast to LSTM-A4, LSTM-A5 reverse the order by first providing image features to the decoder. We will also study ATT-FCN [50], which uses semantic attention mechanism to further incorporates textual concepts.

4.3 Experimental Results

In this section, we first compare the proposal with representative models. Next, we study the effects of incorporating textual features to demonstrate that the significant improvements come from the unified textual representations rather than the simple incorporation of extra features, i.e., textual concepts.

4.3.1 Comparisons with Representative Models. We compare our proposal with some representative models on Flickr30k dataset and MSCOCO dataset. The results on MSCOCO Karpathy test split and Flickr30k Karpathy test split are reported in Table 1 and 2, respectively. By using our proposed textual-enriched image features, improvements of up to 10% and 9% in terms of CIDEr and SPICE can be achieved, respectively, demonstrating the effectiveness and generalization capabilities of our method to a wide range of models. The improvements on SPICE scores, which correlate the best with human judgment [2], suggest that staring from bridging the gap between vision and language domains point of view helps a lot to generate coherent and human-like captions. Specifically, our proposed approach of using textual-enriched image features is able to obtain significant improvements, when applied to the Transformer captioning model. The applied captioning model outperforms current published state-of-the-art models ORT [11] and HIP [48], which further demonstrates the effectiveness of our approach.

4.3.2 Effect of Incorporating Extra Information. To verify whether the significant improvements that our proposal brings to the representative baseline models are due to the introduction of additional textual concepts, we validate our approach on some models that further incorporate textual concepts. As shown in Table 3, improvements of 6% and 9% with respect to CIDEr and SPICE are achieved respectively when applying the proposal to the models that employ both image features and textual concepts. The experimental results demonstrate that the improvement comes from the textual-enriched image features rather than simple incorporation of textual concepts. It is worth noticing that LSTM-A4, which feeds the image features to the decoder at every time step, performs poorly compared to

Table 4: Leaderboard performance on the online MSCOCO evaluation server. c5 means comparing to 5 references and c40 means comparing to 40 references. We outperform all the published works in major metrics.

MSCOCO	BLEU-1		BLEU-2		BLEU-3		BLEU-4		METEOR		ROUGE-L		CIDEr	
	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40
LSTM-A [49]	78.7	93.7	62.7	86.7	47.6	76.5	35.6	65.2	27.0	35.4	56.4	70.5	116.0	118.0
Up-Down [3]	80.2	95.2	64.1	88.8	49.1	79.4	36.9	68.5	27.6	36.7	57.1	72.4	117.9	120.5
CAVP [24]	80.1	94.9	64.7	88.8	50.0	79.7	37.9	69.0	28.1	37.0	58.2	73.1	121.6	123.8
RFNet [14]	80.4	95.0	64.9	89.3	50.1	80.1	38.0	69.2	28.2	37.2	58.2	73.1	122.9	125.1
GLIED [27]	80.1	94.6	64.7	88.9	50.2	80.4	38.5	70.3	28.6	37.9	58.3	73.8	123.3	125.6
GCN-LSTM [47]	-	-	65.5	89.3	50.8	80.3	38.7	69.7	28.5	37.6	58.5	73.4	125.3	126.5
SGAE[47]	81.0	95.3	65.6	89.5	50.7	80.4	38.5	69.7	28.2	37.2	58.6	73.6	123.8	126.5
HIP [48]	**81.6**	**95.9**	**66.2**	**90.4**	**51.5**	81.6	**39.3**	71.0	28.8	38.1	59.0	74.1	127.9	130.2
Ours	80.9	95.7	65.7	**90.4**	51.2	**82.2**	**39.3**	**72.2**	**29.5**	**39.0**	**59.2**	**74.6**	**129.0**	**131.6**

Table 5: Results of quantitative analysis of our approach based on two baselines, i.e., LSTM-A4 [49] and Up-Down [3]. For a better understanding of the differences, we further list the breakdown of SPICE F-scores. TDM and TAM stands for Textual Distilling Module and Textual Association Module, respectively. We can see that the w/ TDM has a higher Attributes and Color scores than the baselines, and the TAM brings significant improvements in Relations and Count. As we can see, incorporating the proposal (i.e., w/ TDM + TAM) directly on the baselines, leads to overall improvements.

Methods	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE					
								All	Objects	Attributes	Relations	Color	Count
LSTM-A4 [49]													
Baseline	78.2	61.9	48.8	35.9	27.3	56.5	113.9	20.3	37.7	8.8	6.0	6.9	7.9
w/ TDM	78.8	63.5	49.6	36.2	27.7	57.1	118.0	21.3	38.9	10.2	5.9	11.1	9.8
w/ TDM + TAM	**79.4**	**64.4**	**50.1**	**36.8**	**28.0**	**57.5**	**121.3**	**22.2**	**39.9**	**11.2**	**7.0**	**11.8**	**15.4**
Up-Down [3]													
Baseline	79.8	63.7	49.5	36.3	27.7	56.9	120.1	21.4	39.1	10.0	6.5	11.4	18.4
w/ TDM	80.8	64.8	49.9	36.9	28.0	57.6	122.4	22.1	39.7	10.6	6.2	12.7	18.5
w/ TDM + TAM	**80.7**	**65.1**	**50.4**	**37.9**	**28.4**	**58.5**	**124.8**	**22.5**	**40.8**	**10.8**	**8.6**	**13.8**	**21.0**

LSTM-A3, which only conditions the decoding on the image features at the first two steps. This can be attributed to the weakness of the image features without textual enriching, as the error brought by the vanilla image features may accumulate with each step of the RNN-based decoder [42, 49].

In Table 5, which shows sub-category scores of SPICE, the proposed approach does especially well in attributes and relations, which requires semantic and deep understanding of images. From the results we can see that employing the textual-enriched image features provides a solid basis for describing images. In all, the baseline scores are promoted by up to 10% and 9% in terms of CIDEr and SPICE, respectively, verifying the effectiveness of the proposed method. It also indicates that our approach are less prone to the variations of model structures, hyper-parameters (e.g., learning rate and batch-size), and learning paradigms.

4.3.3 Online MSCOCO Evaluation. For online evaluation[2], we submit an ensemble of seven "Transformer w/ proposal" models to the leaderboard and compare with the published state-of-the-art methods. As shown in Table 4, compared with the state-of-the-art models, our approach achieves the best results in major metrics, which further demonstrates the effectiveness of our approach.

[2]https://competitions.codalab.org/competitions/3221

5 ANALYSIS

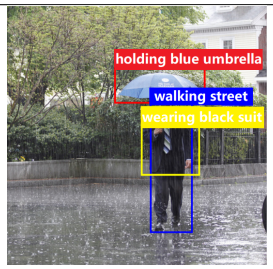
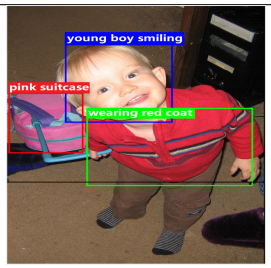
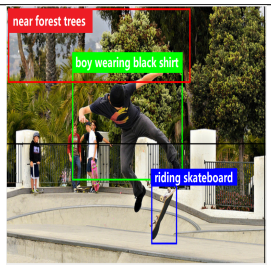
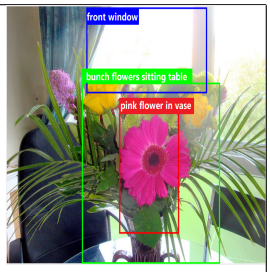
In this section, incremental studies are conducted to verify the effectiveness of each component in the proposal. Furthermore, some examples are given to show the effect of our approach.

5.1 Quantitative Analysis

We select the LSTM-A4 [49] and Up-Down [3] to conduct a series of studies to investigate the contribution of each component in the proposed approach and the results are shown in Table 5.

5.1.1 Effect of Textual Distilling Module. As shown in Table 5, we can find that the Textual Distilling Module promotes all baselines over almost all sub-categories. As expected, the Textual Distilling Module is good at distilling related textual concepts in the current region features, especially in *Attributes* and *Color*, which is verified by the significant improvements of *Attributes* scores and

Table 6: Examples of the captions generated by different methods. The first and second lines are the ground truth and the captions generated by Up-Down. The third and fourth lines are caption results by adding TDM and the full proposal with the both TDM and TAM modules, respectively. From the generated captions, we can find that the w/ TDM helps the baseline to generate more detailed captions in attributes and colors for each object. The Full Model helps the baseline to generate captions that are more comprehensive in relations and objects.

Comparison of Models				
Reference	a man with suit holding an umbrella walking down street in the rain.	a toddler looking up and smiling while pulling a pink bag.	a man flying through the air while riding a skateboard.	a bunch of flowers are on a clear glass table.
Baseline	a man holding an umbrella in the rain.	a boy standing in front of a suitcase.	a man doing a trick on a skateboard in the air.	a vase of flowers sitting on top of a table.
w/ TDM	a man in a black suit holding a blue umbrella.	a young boy standing in front of a pink suitcase.	a boy doing a trick on a skateboard in the air.	a bunch of flowers in a vase sitting on table.
Full Model	a man in a black suit walking on street in the rain with an umbrella.	a young boy in a red coat standing in front of a pink suitcase.	a boy doing a trick on a skateboard in the air with a forest nearby.	a bunch of flowers in a vase sitting on table in front of a window.

Color scores. The reason is that the image feature usually contains a specific object, and the Textual Distilling Module tends to distill the most relevant attributes and colors to depict the specific object. However the Textual Distilling Module is less accurate in associating the textual concepts, resulting in the impaired performance in *Relations*.

5.1.2 Effect of Textual Association Module. As expected, Table 5 shows that the Textual Association Module is good at associating the distilled textual concepts as an unified representation, which is demonstrated by the increased scores in *Relations* and *Count*. With the abundant and enriched textual information introduced by our approach, a variety of baseline models turn the image features into the deep and textual-enriched image understandings, resulting in significant performance improvements, which validate the effectiveness of our approach.

5.2 Qualitative Analysis

We show the captions generated by the Up-Down baseline model, the baseline w/ TDM and the baseline w/ TDM + TAM (Full Model) to analyze the strength of our proposal intuitively. Table 6 shows the baseline model could already generate fluent and descriptive sentences of the input images. However, the conveyed information is rather limited. The w/ TDM is good at describing objects, bringing more details in *attributes* and *colors*, but is less specific in *relations*. The TAM portrays the *relations* and brings more objects by associating textual concepts. As a result, the generated captions of the full model is more complete and coherent, which further proves our arguments and demonstrates the effectiveness of our approach.

6 CONCLUSIONS

We focus on bridging the gap between vision and language domains by enriching image features with textual concepts, which provides a solid basis for describing images. We explore the textual representations of image features to describe salient image regions on the textual level. We propose the Textual Distilling Module and Textual Association Module to explore the abundant and enriched textual information for achieving a deep image understanding. Extensive experiments on the widely-used Flickr30k and MSCOCO image captioning datasets validate the effectiveness of our method. Our proposed solution successfully promotes the performance of all the strong baselines across all metrics over the board, with the most significant improvement up to 10% and 9%, in terms of CIDEr and SPICE, respectively. The results demonstrate the generalization ability of our approach to a wide range of existing systems. The qualitative analysis and the SPICE sub-categories scores show that the generated captions are complete and coherent in comparison with existing methods. Besides, we further validate the importance of modeling the relationships between vision and language domains, rather than simple incorporating them.

ACKNOWLEDGMENTS

This paper was partially supported by National Engineering Laboratory for Video Technology - Shenzhen Division, Shenzhen Municipal Development and Reform Commission (Disciplinary Development Program for Data Science and Intelligent Computing). Special acknowledgements are given to Aoto-PKUSZ Joint Lab for its support. We thank all the anonymous reviewers for their constructive comments and suggestions.

REFERENCES

[1] Chris Alberti, Jeffrey Ling, Michael Collins, and David Reitter. 2019. Fusion of Detected Objects in Text for Visual Question Answering. In *EMNLP*.

[2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. SPICE: Semantic Propositional Image Caption Evaluation. In *ECCV*.

[3] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-Up and Top-Down Attention for Image Captioning and VQA. In *CVPR*.

[4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *ICCV*.

[5] Lei Jimmy Ba, Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer Normalization. *arXiv: 1607.06450* (2016).

[6] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv: 1409.0473* (2014).

[7] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *ACL Workshop*.

[8] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. 2015. Microsoft COCO Captions: Data Collection and Evaluation Server. *arXiv: 1504.00325* (2015).

[9] Hao Fang, Saurabh Gupta, Forrest N. Iandola, Rupesh Kumar Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C. Platt, C. Lawrence Zitnick, and Geoffrey Zweig. 2015. From Captions to Visual Concepts and Back. In *CVPR*.

[10] Yang Feng, Lin Ma, Wei Liu, and Jiebo Luo. 2019. Unsupervised Image Captioning. In *CVPR*.

[11] Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. 2019. Image Captioning: Transforming Objects into Words. In *NeurIPS*.

[12] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9, 8 (1997), 1735–1780.

[13] Ting-Hao (Kenneth) Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross B. Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, and Margaret Mitchell. 2016. Visual Storytelling. In *HLT-NAACL*.

[14] Wenhao Jiang, Lin Ma, Yu-Gang Jiang, Wei Liu, and Tong Zhang. 2018. Recurrent Fusion Network for Image Captioning. In *ECCV*.

[15] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. 2016. DenseCap: Fully Convolutional Localization Networks for Dense Captioning. In *CVPR*.

[16] Andrej Karpathy and Fei-Fei Li. 2015. Deep visual-semantic alignments for generating image descriptions. In *CVPR*.

[17] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L. Berg. 2014. ReferItGame: Referring to Objects in Photographs of Natural Scenes. In *EMNLP*.

[18] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. 2017. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision* 123, 1 (2017), 32–73.

[19] Alexander Krull, Tim-Oliver Buchholz, and Florian Jug. 2019. Noise2Void - Learning Denoising From Single Noisy Images. In *CVPR*.

[20] Guang Li, Linchao Zhu, Ping Liu, and Yi Yang. 2019. Entangled Transformer for Image Captioning. In *ICCV*. IEEE, 8927–8936.

[21] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. *arXiv: 2004.06165* (2020).

[22] Chin-Yew Lin and Eduard H. Hovy. 2003. Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. In *HLT-NAACL*.

[23] Zhouhan Lin, Minwei Feng, Cícero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A Structured Self-Attentive Sentence Embedding. In *ICLR*.

[24] Daqing Liu, Zheng-Jun Zha, Hanwang Zhang, Yongdong Zhang, and Feng Wu. 2018. Context-Aware Visual Policy Network for Sequence-Level Image Captioning. In *ACMMM*.

[25] Fenglin Liu, Meng Gao, Tianhao Zhang, and Yuexian Zou. 2019. Exploring Semantic Relationships for Image Captioning without Parallel Data. In *ICDM*.

[26] Fenglin Liu, Yuanxin Liu, Xuancheng Ren, Xiaodong He, and Xu Sun. 2019. Aligning Visual Regions and Textual Concepts for Semantic-Grounded Image Representations. In *NeurIPS*.

[27] Fenglin Liu, Xuancheng Ren, Yuanxin Liu, Kai Lei, and Xu Sun. 2019. Exploring and Distilling Cross-Modal Information for Image Captioning. In *IJCAI*.

[28] Fenglin Liu, Xuancheng Ren, Yuanxin Liu, Houfeng Wang, and Xu Sun. 2018. simNet: Stepwise Image-Topic Merging Network for Generating Detailed and Comprehensive Image Captions. In *EMNLP*.

[29] Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. 2020. Federated Learning for Vision-and-Language Grounding Problems. In *AAAI*.

[30] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *NeurIPS*.

[31] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing When to Look: Adaptive Attention via a Visual Sentinel for Image Captioning. In *CVPR*.

[32] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2018. Neural Baby Talk. In *CVPR*.

[33] Yingwei Pan, Ting Yao, Houqiang Li, and Tao Mei. 2017. Video Captioning with Transferred Semantic Attributes. In *CVPR*.

[34] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *ACL*.

[35] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *NIPS*.

[36] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. VL-BERT: Pre-training of Generic Visual-Linguistic Representations. In *ICLR*.

[37] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. VideoBERT: A Joint Model for Video and Language Representation Learning. In *ICCV*.

[38] Hao Tan and Mohit Bansal. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In *EMNLP*.

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *NIPS*.

[40] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based image description evaluation. In *CVPR*.

[41] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *CVPR*.

[42] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2017. Show and Tell: Lessons Learned from the 2015 MSCOCO Image Captioning Challenge. *PAMI* 39, 4 (2017), 652–663.

[43] Qi Wu, Chunhua Shen, Lingqiao Liu, Anthony R. Dick, and Anton van den Hengel. 2016. What Value Do Explicit High Level Concepts Have in Vision to Language Problems?. In *CVPR*.

[44] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *ICML*.

[45] Linjie Yang, Kevin D. Tang, Jianchao Yang, and Li-Jia Li. 2017. Dense Captioning with Joint Inference and Visual Context. In *CVPR*.

[46] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. 2019. Auto-Encoding Scene Graphs for Image Captioning. In *CVPR*.

[47] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. 2018. Exploring Visual Relationship for Image Captioning. In *ECCV*.

[48] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. 2019. Hierarchy Parsing for Image Captioning. In *ICCV*.

[49] Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. 2017. Boosting Image Captioning with Attributes. In *ICCV*.

[50] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image Captioning with Semantic Attention. In *CVPR*.

[51] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL* (2014).

[52] Yuan Yuan, Siyuan Liu, Jiawei Zhang, Yongbing Zhang, Chao Dong, and Liang Lin. 2018. Unsupervised Image Super-Resolution Using Cycle-in-Cycle Generative Adversarial Networks. In *CVPR Workshops*.

[53] Cha Zhang, John C. Platt, and Paul A. Viola. 2006. Multiple Instance Boosting for Object Detection. In *NIPS*.

[54] Guangxiang Zhao, Junyang Lin, Zhiyuan Zhang, Xuancheng Ren, Qi Su, and Xu Sun. 2019. Explicit Sparse Transformer: Concentrated Attention Through Explicit Selection. *arXiv: 1912.11637* (2019).

[55] Luowei Zhou, Yannis Kalantidis, Xinlei Chen, Jason J. Corso, and Marcus Rohrbach. 2019. Grounded Video Description. In *CVPR*.

[56] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J. Corso, and Jianfeng Gao. 2020. Unified Vision-Language Pre-Training for Image Captioning and VQA. In *AAAI*.