AEGAN: Attribute-Enhanced Graph Attention Network for Image Captioning

Maofu Liu, Wei Wang, Yinwei Wei, *Member, IEEE,*

***Abstract*—In this work, we focus on incorporating textual attribute information into visual object representation to improve the performance of image captioning. Therefore, we construct an attribute-enhanced scene graph to explicitly model the correlation between visual object and its corresponding textual attribute and accordingly develop an attribute-enhanced graph attention network (AEGAN) for image caption generation. In particular, we devise a multi-instance extractor to learn the textual attributes for the visual objects from the ground truth captions. Based on the attribute-enhanced scene graph established by the object, relation, and textual attribute nodes, a dual graph attention network model conducts to embed the textual attributes and relations into object nodes to enrich their representations. Finally, we adopt a multi-task strategy to train AEGAN, where a node-level classification task avoids over-smooth problem during the graph convolutional operation. The experimental results on the MSCOCO dataset show that our proposed method has significantly outperformed the baseline, and achieved at the state-of-the-art performance under the evaluation metrics.**

***Index Terms*—image captioning, attribute-enhanced scene graph, dual graph attention network, multi-instance extractor.**

I. Introduction

IMAGE captioning, a challenging and promising vision and language task [1], [2], attracts much attention from computer vision and natural language processing communities. It focuses on discovering the semantic signals from visual information and generating the natural language text to describe the content of the given image.

Towards this end, the Neural Image Caption (NIC) [3] generator adopted Convolution Neural Network (CNN) [4] as an encoder to encode the given image into a global vector, and then the vector as input to the Long Short Term Memory (LSTM) decoder [5] to generate a description. Based on NIC, Xu *et al.* [6] proposed an attention mechanism to find the visual clues for the generated text. Afterward, Anderson *et al.* [7] introduced an up-down attention mechanism into the image captioning models, concentrating on capturing the information corresponding to the objects in the given image.

Beyond representing the isolated objects, recently, some efforts [8], [9] have been dedicated to explicitly modeling the relation between visual object and the candidate attributes by constructing the scene graph for the image. With the obtained scene graphs, these models conduct graph convolutional operations to enrich the representations of object nodes by injecting their local structure information, which optimizes the caption generation.

Besides the remarkable performance, we have observed that most of the attributes in scene graph implied from the pre-trained model on Visual Genome (VG) dataset do not appear in the ground truth caption and bring noise to the caption generation due to the gap between the VG dataset and the image captioning target dataset, *i.e.*, MSCOCO. In addition, we also find that the preceding models have omitted the attributes in the visual object representation learning, which hinders the models from generating discriminative captions for similar images. Taking us in the context of real-world applications, like the information retrieval [10] and personalized recommendation [11], [12], the discriminative and fine-grained caption can help the users to locate the targets (*e.g.,* videos, images, news) from similar candidates. Therefore, in this work, we aim to explicitly model the textual attributes of visual objects in the image and enhance the object representations for the high-quality image caption.

However, achieving the goal is nontrivial due to the following challenges:

- Although the pre-trained models can capture the attributes from the given image, they also can hurt the performance due to the gap between pre-trained VG and MSCOCO datasets. To justify our argument, we do the statistics over these datasets and exhibit their top-10 high-frequency attributes in captions, as shown in Figure 1(a). Considering the difference between two distributions, how to extract the textual attribute features from the region-level visual information according to the description is the first challenge in this work.
- By analyzing the images and captions, we can find that: 1) the textual attributes in the captions cannot be independent with their corresponding objects, and 2) the textual attributes of visual objects in the images are diverse, as shown in Figure 1(b). Hence, how to explicitly model the correlation between textual attributes and visual objects, and encode the diverse textual attributes associated with the same object is the second challenge.
- It is necessary to incorporate the textual attribute features into the correlated visual object in object representation learning. However, how to fuse these heterogeneous signals is another challenge we are facing in the task.

To solve outlined challenges, we develop an Attribute-Enhanced Graph Attention Network (AEGAN), which consists

M. Liu and W. Wang are with the School of Computer Science and Technology, Wuhan University of Science and Technology, China. (e-mails: liumaofu@wust.edu.cn; wangwei8024@gmail.com).

Y. Wei is with the School of Computing, National University of Singapore, Singapore. (e-mail: weiyinwei@hotmail.com).

Manuscript received April 19, 2005; revised August 26, 2015.

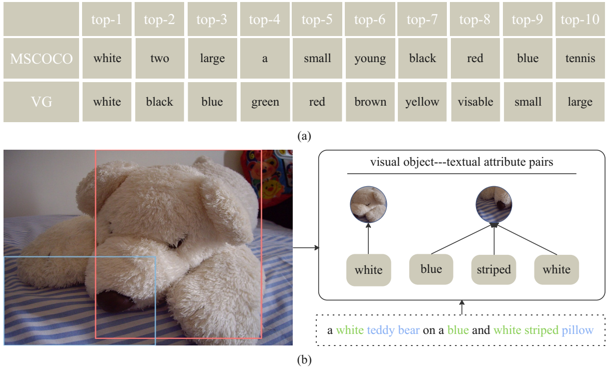


Fig. 1. (a) Top-10 attributes in the target MSCOCO dataset and the pre-trained VG dataset. (b) Pairs of visual objects and their textual attributes extracted from the given image and the caption.

of the attribute-enhanced scene graph construction, Graph Attention Network (GAT) based image encoder, and caption decoder equipped with the attention-language LSTMs. In particular, we propose a multi-instance based attribute extractor to extract textual attributes of visual objects from the description and train it on the MSCOCO dataset. With the extracted attributes, we introduce them into the scene graph, termed the attribute-enhanced scene graph, to explicitly construct the correlation between textual attributes and visual objects. To enrich the object representation in the encoder, we devise a Dual GAT Model (DGM) to embed the textual attribute and relation signals into the object, respectively. To be more specific, for each object, the DGM aggregates the information in terms of their affinities to their attributes or relations and fuses them to enhance the object representation. Afterward, we adopt the attention-language LSTMs based decoder to map the enriched representations to the natural language captions.

To optimize the proposed model, we adopt the multi-task optimization strategy. In addition to the caption generation, we classify each node supervised by their types *i.e.,* object, attribute, and relation, which is motivated by alleviating the over-smoothing problem in Graph Neural Network (GNN) [13] models. To demonstrate the effectiveness of our proposed method, we conduct extensive experiments on the MSCOCO dataset. With both supervised and reinforcement learning, our proposed model outperforms several state-of-the-art models, like Up-Down [7], SGAE [8], MT-I [14] and MT-II [14], by a margin, and achieved the state-of-the-art performance. In a nutshell, the contributions of our work can be summarized as follows:

- We explore the effectiveness of textual attribute and explicitly construct the correlation between object and its textual attribute in image captioning.
- We develop an Attribute-Enhanced Graph Attention Network (AEGAN) to extract the textual attributes by our devised multi-instance based attribute extractor and generate the image caption upon the constructed attribute-enhanced scene graph with the DGM.
- By conducting extensive experiments on the benchmark dataset, we demonstrate our proposed model outperforms the state-of-the-art performance.

II. Related Work

A. *Image Captioning*

In recent years, the image captioning community has made progress mainly in the following three aspects: 1) Vinyals *et al.* [3] designed the NIC image captioning model with the encoder-decoder framework and first adopted the CNN as the encoder to encode the input image and the LSTM as a decoder to generate the description in an end-to-end manner. Nowadays, the encoder-decoder framework becomes a popular paradigm for image captioning. 2) Attention mechanisms [6], [7], [15] were proposed to find the visual clue of each word in the image, increase the model interpretability, and improve the performance of caption generation. Recently, some studies [8], [14], [16], [17] committed to constructing scene graph to mine semantic information in the image. Under these circumstances, a new structural attention mechanism [16] was proposed to align linguistic words and visual semantic units for image captioning. 3) Instead of teacher forcing, reinforcement learning [18]–[20] which optimized the metrics (*i.e.*, BLEU and CIDEr) could solve the problem of exposure bias and the inconsistency of evaluation metrics between training and testing. In this paper, we followed the encoder-decoder paradigm. We firstly construct an attribute-augmented scene graph to mine semantic information and then propose the DGM as an encoder for fine-grained graph representation learning. And then, we adopt the attention-language LSTMs [7] as a decoder for word prediction by focusing on the nodes in the scene graph. Lastly, we optimized our model with supervised and reinforcement learning.

B. *Multi-instance Learning*

In multi-instance learning (MIL) [21], each bag consists of multiple instances. Whereinto, the bag equips with the label but the instance does not. The MIL can be coarsely divided into two types, *i.e.*, instance-level and bag-level, according to the learning manner. The instance-level method treats each instance in the bag equally and the category of the bag is jointly determined by all the instances. The bag-level method learns a representation for each bag and predicts the category of the bag based on the representation. In image captioning, some efforts [14], [22], [23] tried to use the MIL framework to improve the performance. Fang *et al.* [22] and Yao *et al.* [23] selected the most common 1000 words in the training set as the attributes of the image, and then extracted the attributes of the image using MIL to improve the performance of image captioning. Shi *et al.* [14] proposed a MIL method to recognize the predicate between visual objects in terms of the ground truth caption and improved the performance significantly. However, the above MIL methods are instance-level, which tend to ignore the association among instances in the same bag. Meanwhile, the bag-level method always loses the detail of each instance. In this paper, we proposed a MIL method to extract the textual attribute of the visual object from the ground truth caption. In our method, the representation of each instance contains not only its local detailed information but also the global information of the bag in which it is located.

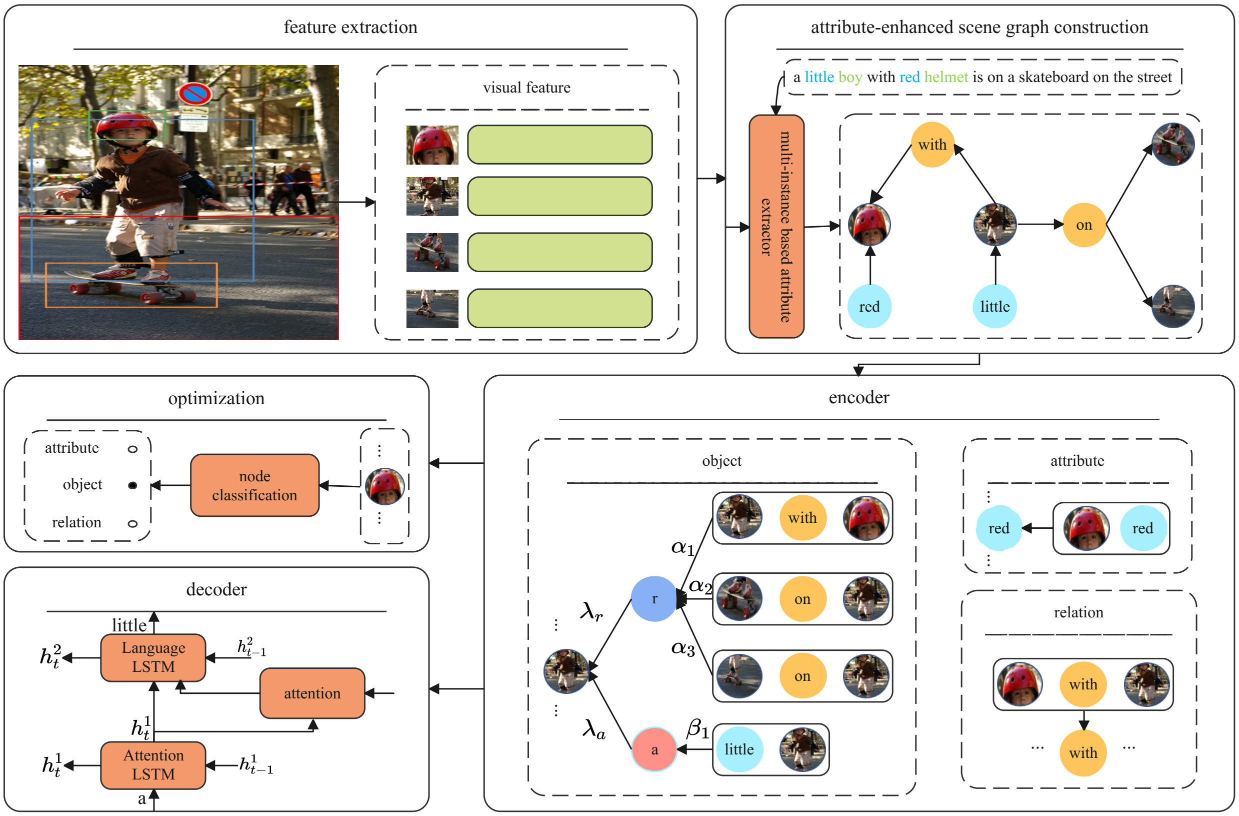


Fig. 2. An overview of our method. Given an image and a ground truth caption, we firstly generate a attribute-enhanced scene graph. Subsequently, we learn node representation for the object, relation, and attribute in the scene graph. Finally, we use the node representation for node classification and caption generation.

C. Graph Representation Learning based on Textual Attribute

Benefit from the rapid development of graph representation learning, in recent years, a lot of work committed to constructing scene graph to mine the semantic information in the image, and the graph neural networks are used to encode the scene graph for image captioning. In existing studies, Yao *et al.* [24] explored the visual relation between objects and used Graph Convolutional Network (GCN) [25] to integrate the visual relation into the representation of the object. Some research [8], [14], [16] constructed the scene graph to mine the semantic information in the image and used GCN to aggregate the relation information into the object representation. However, attribute information is crucial for a fine-grained description of the object. The preceding methods extracted attributes using the pre-trained model on the VG dataset. Due to the gap between the VG and target datasets, it brings noise to the caption generation. Meanwhile, they ignored the integration of attribute information into object representation and did not distinguish the importance of various attributes. Therefore, in this paper, we devised the MIL model to extract textual attributes of visual objects from the description, which fills the gap between the pre-trained and target datasets, and designed the DGM to aggregate relation and textual attribute information into the object representation according to their affinities, separately.

III. Method

Following prior works [3], [6], [26], we adopt the popular encoder-decoder paradigm to represent the visual content of the image and generate the caption in natural language form, which is widely used in image captioning. To this end, by feeding an image I, the encoder is able to detect its salient objects [27] and represent them as $\mathcal{V} = \{\mathbf{v}_{o_1}, \mathbf{v}_{o_2}, \cdots, \mathbf{v}_{o_n}\}$ with the pre-trained models, where $\mathbf{v}_{o_j} \in \mathbb{R}^{2048\times 1}$ denotes the visual feature of j-th salient object and n means the number of salient objects in the image. Using the obtained representations of the image, the decoder can generate the caption for the given image, which can be denoted as $\mathcal{S} = \{w_1, w_2, \cdots, w_T\}$. Whereinto, w_j and T represent the j-th word in the caption and length of generated sentence, respectively.

In our work, we propose to harness the textual attributes of visual objects to enrich the representation of images and accordingly develop an attribute-enhanced graph attention network for image captioning, as illustrated in Figure 2. In particular, we first extract textual attributes of visual objects detected in the image from the ground truth caption using the multi-instance based attribute extractor and incorporate them with the scene graph. Based on the attribute-enhanced scene graph, we perform the graph convolutional operations to inject the informative signal from the textual attributes and relations into object nodes. With the enhanced representations of nodes, we could generate the captions through the decoder module. In addition, we introduce an auxiliary task, *i.e.,* the nodes

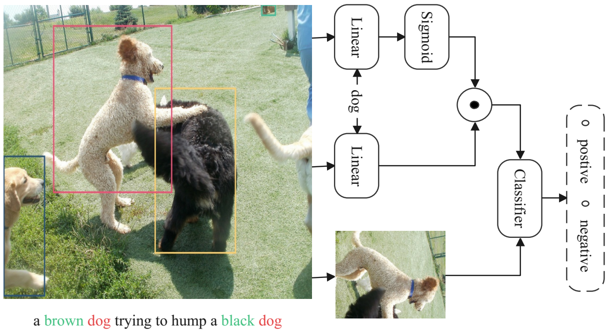


Fig. 3. An overview of the multi-instance based attribute extractor. The regions depicted by colorful bounding boxes in the image construct a positive bag for attributes "brown" and "black" according to the description. An instance in the bag passing the module on the right will output the probability of owning one attribute.

classification, to alleviate the over-smoothing representations of nodes.

A. Attribute-enhanced Scene Graph Construction

In this section, we detail our designed multi-instance based attribute extractor and attribute-enhanced scene graph generation. To extract the attributes without noise for visual objects, we commit to training an attribute extractor on the ground truth datasets, which can bridge the gap of the attribute distributions over pre-trained and ground truth datasets. We adopt the weakly-supervised learning strategy to optimize the attribute extractor due to the difficulty of visual object and textual attribute alignment. Therefore, we design a multi-instance based attribute extractor to deploy the weakly supervised learning.

The multi-instance based attribute extractor learns the affinities between bags (*i.e.*, clusters of instances) and labels to approximate the alignment of object-attribute pairs instead of modeling the correlation between instances (*i.e.*, visual objects) and labels (*i.e.*, textual attributes). We extract the visual objects from each image and cluster them according to their categories. As for the labels, we use the object category in each bag to collect the corresponding words describing their attributes in the captions.

As shown in Figure 3, the objects, which can be detected from the given image and depicted by the colorful bounding boxes, hold the same category, *i.e.,* dog. Moreover, we can also extract two object-attribute pairs, *i.e.*, brown-dog and black-dog, from the description. Therefore, the visual regions construct a positive bag for attributes "brown" and "black". We use $\mathcal{B} = \{\mathbf{i}_1, \mathbf{i}_2, \cdots, \mathbf{i}_K\}$ to represent the bag, where K is the number of instances in the bag and may be different in different bags, and $\mathbf{i}_k$ denotes the k-th instance in $\mathcal{B}$. The representation of instance $\mathbf{i}$ should include its local detailed information and the global information of the bag. The local detailed information can distinguish different instances in the same bag, while the global information can discriminate between different bags. We use $\mathbf{b}$ denotes the global information of $\mathcal{B}$. Since the instances in the same bag must be from the same image and have the same category, we can obtain $\mathbf{b}$ from the image feature and the category information. To this end, we design a gating mechanism denoted by $gate$ to select the global information from the candidate $\tilde{\mathbf{b}}$, derived from the image feature, and word vector of the category, to represent $\mathbf{b}$.

$$\tilde{\mathbf{b}} = W^b[I; \mathbf{e}_c] + \mathbf{b}_1, \tag{1}$$

$$gate = \sigma(W^g[I; \mathbf{e}_c] + \mathbf{b}_2), \tag{2}$$

$$\mathbf{b} = gate \odot \tilde{\mathbf{b}}, \tag{3}$$

where $W^b, W^g \in \mathbb{R}^{512\times 2348}$ are trainable matrices, $I \in \mathbb{R}^{2048\times 1}$ is the image feature, $\mathbf{b}_1, \mathbf{b}_2 \in \mathbb{R}^{512\times 1}$ are biases, and $\mathbf{e}_c \in \mathbb{R}^{300\times 1}$, [;], and $\odot$ denote the word vector of the category, concatenation operation, and hadamard product, respectively.

With the global information $\mathbf{b}$, we concatenate the visual feature of instance $\mathbf{i}_k$ and $\mathbf{b}$ to represent instance $\mathbf{i}_k$.

$$\mathbf{i}_k = [\mathbf{v}_{\mathbf{i_k}}; \mathbf{b}], \tag{4}$$

where $\mathbf{v}_{\mathbf{i}_k}$ represents the visual feature of instance $\mathbf{i}_k$.

We input instance $\mathbf{i}_k$ into the attribute classifier to get the probability $p^{a_j}_{\mathbf{i}_k}$ that instance $\mathbf{i}_k$ owns candidate attribute a_j. The probability that $\mathcal{B}$ owns a_j is calculated using "noisy-OR" [28]:

$$\mathbf{p}^{a_j}_{\mathcal{B}} = 1 - \prod_{\mathbf{i}_k \in \mathcal{B}} (1 - p^{a_j}_{\mathbf{i}_k}). \tag{5}$$

Since the confidence of alignment between object and attribute decreases with the number increment of instances in the bag. Therefore, we assign different weights to the loss of the bag according to the number of instances in the bag. For attribute a_j, we treat the bag without any instance owning a_j as the negative bag $\tilde{\mathcal{N}}_{a_j}$, and otherwise, the bag is regarded as the positive one and denoted by $\mathcal{N}_{a_j}$. Based on the cross-entropy loss, we reformulate the loss function as,

$$\begin{aligned} L(I) \quad &= -\textstyle\sum_{j=1}^{M}[(1 - \frac{1}{1+\sqrt{C_{\mathcal{N}_{a_j}}}}) * \log p^{a_j}_{\mathcal{N}_{a_j}} \\ &+(1 - \frac{1}{1+\sqrt{C_{\tilde{\mathcal{N}}_{a_j}}}}) * \log{(1 - p^{a_j}_{\tilde{\mathcal{N}}_{a_j}})}], \end{aligned} \tag{6}$$

where M is the number of candidate attributes, and $C_{\mathcal{N}_{a_j}}$ and $C_{\tilde{\mathcal{N}}_{a_j}}$ are the numbers of instances in positive bag and negative bag, respectively.

The object attributes in our model are extracted from the ground truth captions. Based on textual attributes, we establish the attribute-enhanced scene graph with three kinds of nodes, *i.e.,* object, attribute, and relation, indicated by o, a, and r, respectively. Towards this end, we adopt Faster-RCNN to identify and localize the objects contained in the image. And then, relation detector [14], [29] is designed to detect the relations among objects. To explicitly construct the correlation between attributes and objects, we put forward a multi-instance based attribute extractor to extract the textual attributes of objects from the description and build the link of them, as illustrated in Figure 2. Whereinto, we use o_i, $a_{i,k}$, and r_{ij} to represent i-th object in the scene graph, k-th attribute of object o_i, and the relation between object o_i and o_j, respectively. To specify the existence of an edge between nodes, we obey the following policies:

- If object o_i holds attribute $a_{i,k}$, a directed edge is used to connect from $a_{i,k}$ to o_i.
- If there is a relation r_{ij} between objects o_i and o_j, two directed edges are used to connect from o_i to r_{ij} and from r_{ij} to o_j, respectively.

B. Encoder

1) Graph Embedding: There are three types of nodes, *i.e.*, object, attribute, and relation, in the attribute-enhanced scene graph. We compute an initial representation for each node in the scene graph.

We use $\mathbf{e}$ to represent the textual feature of the node in the scene graph, respectively. For example, $\mathbf{e}_{o_i}$, $\mathbf{e}_{r_{ij}}$, and $\mathbf{e}_{a_{i,k}} \in \mathbb{R}^{1000\times 1}$ denote the textual features of o_i, r_{ij}, and $a_{i,k}$, respectively.

Object Embedding g_{o_i}: We use the visual and textual features of the object o_i to initialize the embedding representation. To alleviate the gap between the source and target domains, we firstly map the visual feature to the textual vector space.

$$\mathbf{e}^v_{o_i} = W^v \mathbf{v}_{o_i} \tag{7}$$

$$\begin{cases} \mathbf{g}_{o_i} = ReLU(W^g_o[\mathbf{e}^v_{o_i}; \mathbf{e}_{o_i}]), \\ \mathbf{g}_{o^a_i} = ReLU(W^g_{o^a}[\mathbf{e}^v_{o_i}; \mathbf{e}_{o_i}]), \\ \mathbf{g}_{o^r_i} = ReLU(W^g_{o^r}[\mathbf{e}^v_{o_i}; \mathbf{e}_{o_i}]), \end{cases} \tag{8}$$

where $W^v \in \mathbb{R}^{1000\times 2048}$ and W^g_o, $W^g_{o^a}$, $W^g_{o^r} \in \mathbb{R}^{1000\times 2000}$ are trainable matrices, $ReLU$ is an activation function, and $\mathbf{g}_{o_i}$, $\mathbf{g}_{o^a_i}$, $\mathbf{g}_{o^r_i} \in \mathbb{R}^{1000\times 1}$ denote the initial representations of object o_i. Whereinto, $\mathbf{g}_{o_i}$, $\mathbf{g}_{o^a_i}$, and $\mathbf{g}_{o^r_i}$ are used to fuse heterogeneous information, aggregate attribute information, and incorporate relation information, respectively.

Relation Embedding $\mathbf{r_{ij}}$: Since the $relation$ does not have an obvious corresponding region in the image, we just use the textual feature to initilize the representation of relation r_{ij}.

$$\mathbf{g}_{r_{ij}} = ReLU(W^g_r \mathbf{e}_{r_{ij}}), \tag{9}$$

where $\mathbf{g}_{r_{ij}} \in \mathbb{R}^{1000\times 1}$ denotes the initial representation of relation r_{ij}, and $W^g_r \in \mathbb{R}^{1000\times 1000}$ is a trainable matrix.

Attribute Embedding $a_{i,k}$: Similar to the $relation$, we use the textual feature to initialize the representation of attribute $a_{i,k}$.

$$\mathbf{g}_{a_{i,k}} = ReLU(W^g_a \mathbf{e}_{a_{i,k}}), \tag{10}$$

where $\mathbf{g}_{a_{i,k}} \in \mathbb{R}^{1000\times 1}$ denotes the initial representation of attribute $a_{i,k}$, and $W^g_a \in \mathbb{R}^{1000\times 1000}$ is a trainable matrix.

2) DGM: We design the DGM to enhance the object, attribute, and relation features. The DGM consists of two graph attention networks, Attribute Graph Attention Network (AGAT) and Relation Graph Attention Network (RGAT).

Object Enhancement $\mathbf{x}_{o^a_i}$ and $\mathbf{x}_{o^r_i}$: For one object, it has different attributes and the importance of attribute is not consistent. At this point, AGAT is designed to assign different weights to the various attributes. The weights α_i for attributes are calculated as follows,

$$s^a_{i,k} = \phi_a(W^a_1 \mathbf{g}_{o^a_i} + W^a_2[\mathbf{g}_{o^a_i}; \mathbf{g}_{a_{i,k}}]), \tag{11}$$

$$\alpha_i = softmax(\mathbf{s}^a_i), \tag{12}$$

where $s^a_{i,k}$ and $\mathbf{s}^a_i$ are the attention weight of attribute $a_{i,k}$ and attention weight vector of attributes of object o_i, respectively, and ϕ_a is a fully-connected layer.

We aggregate the attribute information into the representation of object o_i based on the weights.

$$\mathbf{x}_{o^a_i} = \mathbf{g}_{o^a_i} + \sum_{k=1}^{A(o_i)} \alpha_{i,k} \cdot ReLU(W^x_{o^a}([\mathbf{g}_{o^a_i}; \mathbf{g}_{a_{i,k}}])), \tag{13}$$

where $W^x_{o^a} \in \mathbb{R}^{1000\times 2000}$ is a learnable matrix, $A(o_i)$ is the number of attributes of object o_i, and $\mathbf{x}_{o^a_i} \in \mathbb{R}^{1000\times 1}$ represents the enhanced object representation using attributes.

RGAT is designed to assign weights to the relations. The weights β_i for the relations are calculated as follows,

$$s^r_{i,j} = \phi_r(W^r_1 \mathbf{g}_{o^r_i} + W^r_2[\mathbf{g}_{o^r_i}; \mathbf{g}_{r_{i,j}}; \mathbf{g}_{o^r_j}]), \tag{14}$$

$$\beta_i = softmax(\mathbf{s}^r_i), \tag{15}$$

where W^r_1, $W^r_2 \in \mathbb{R}^{512\times 1000}$ are trainable matrices, $s^r_{i,j}$ and $\mathbf{s}^r_i$ are the attention weight of relation r_{ij} and attention vector of relations with object o_i, respectively, and ϕ_r is a fully-connected layer.

We aggregate the relation information into the representation of object o_i based on the weights.

$$\mathbf{x}_{o^r_i} = \mathbf{g}_{o^r_i} + \sum_{j=1}^{R(o_i)} \beta_{i,j} \cdot ReLU(W^x_{o^r}([\mathbf{g}_{o^r_i}; \mathbf{g}_{r_{ij}}; \mathbf{g}_{o^r_j}])), \tag{16}$$

where $W^x_{o^r} \in \mathbb{R}^{1000\times 3000}$ is a learnable matrix, $R(o_i)$ and $\mathbf{x}_{o^r_i} \in \mathbb{R}^{1000\times 1}$ denote the number of relation triplets with object o_i and the enhanced object representation using relations, separately.

In order to fully evaluate the importance of attributes and relations from different perspectives, we adopt the multi-head attention mechanism to aggregate information.

$$\begin{cases} \mathbf{x}_{o^a_i} = \mathbf{g}_{o^a_i} + \frac{1}{\mathcal{D}} \sum_{d=1}^{D} \sum_{k=1}^{A(o_i)} \mathbf{x}^d_{o^a_i}, \\ \mathbf{x}_{o^r_i} = \mathbf{g}_{o^r_i} + \frac{1}{\mathcal{D}} \sum_{d=1}^{D} \sum_{j=1}^{R(o_i)} \mathbf{x}^d_{o^r_i}, \end{cases} \tag{17}$$

where $\mathcal{D}$ is the number of attention mechanisms, $\mathbf{x}^d_{o^a_i}$ and $\mathbf{x}^d_{o^r_i}$ are enhanced object representations calculated by the d-th attention mechanism as aforementioned.

Relation Enhancement $\mathbf{x}_{r_{ij}}$: The relation r_{ij} enhances its representation by incorporating the object.

$$\mathbf{x}_{r_{ij}} = \mathbf{g}_{r_{ij}} + ReLU(W^x_r([\mathbf{g}_{o^r_i}; \mathbf{g}_{r_{ij}}; \mathbf{g}_{o^r_j}])), \tag{18}$$

where $W^x_r \in \mathbb{R}^{1000\times 3000}$ is a trainable matrix, and $\mathbf{x}_{r_{ij}} \in \mathbb{R}^{1000\times 1}$ denotes the enhanced relation representation of $\mathbf{g}_{r_{ij}}$.

Attribute Enhancement $\mathbf{x_{a_{i,k}}}$: The attribute $a_{i,k}$ incorporates the object information to enhance the attribute representation.

$$\mathbf{x}_{a_{i,k}} = \mathbf{g}_{a_{i,k}} + ReLU(W^x_a([\mathbf{g}_{o^a_i}; \mathbf{g}_{a_{i,k}}])), \tag{19}$$

where $W^x_a \in \mathbb{R}^{1000\times 2000}$ is a trainable matrix, and $\mathbf{x}_{a_{i,k}} \in \mathbb{R}^{1000\times 1}$ represents the enhanced attribute representation of $\mathbf{g}_{a_{i,k}}$.

3) Heterogeneous Information Fusion: $\mathbf{x}_{o_i^a}$ and $\mathbf{x}_{o_i^r}$ are enhanced representations including attributes and relations, separately. The enhanced representation x_{o_i} of object o_i should contain both attribute and relation information. To this end, we aggregate $\mathbf{x}_{o_i^a}$ and $\mathbf{x}_{o_i^r}$ based on the similarity scores with the initial object representation $\mathbf{g}_{o_i}$. The whole calculation process is as follows,

$$\lambda_a = \frac{exp(\mathbf{g}_{o_i} \cdot \mathbf{x}_{o_i^a})}{exp(\mathbf{g}_{o_i} \cdot \mathbf{x}_{o_i^a}) + exp(\mathbf{g}_{o_i} \cdot \mathbf{x}_{o_i^r})}, \tag{20}$$

$$\lambda_r = \frac{exp(\mathbf{g}_{o_i} \cdot \mathbf{x}_{o_i^r})}{exp(\mathbf{g}_{o_i} \cdot \mathbf{x}_{o_i^a}) + exp(\mathbf{g}_{o_i} \cdot \mathbf{x}_{o_i^r})}, \tag{21}$$

$$\mathbf{x}_{o_i} = \lambda_a \mathbf{x}_{o_i^a} + \lambda_r \mathbf{x}_{o_i^r}. \tag{22}$$

where $(\cdot)$ denotes the interior product, and λ_a and λ_r are similarity scores.

C. Decoder

Inspired by the prior work [7], we adopt the attention-language LSTMs model as the decoder to generate the image caption. In particular, we use the bottom LSTM to calculate the weights of the nodes in the scene graph, which measures the importance of the nodes for the caption generation. As for the top LSTM, it can generate the words by incorporating the node representations with the corresponding weights. The decoder is formulated as,

$$\mathbf{h}_t^2 = LSTM(\mathbf{h}_{t-1}^2, [\mathbf{h}_t^1; \mathbf{c}_t]), \tag{23}$$

$$p(w_t | w_{<t}, I) = softmax(f_w(\mathbf{h}_t^2)), \tag{24}$$

where $\mathbf{h}_{t-1}^2$ and $\mathbf{h}_t^2$ are the hidden states of top LSTM at time $t-1$ and t, $\mathbf{h}_t^1$ is the hidden state of bottom LSTM at time t, $\mathbf{c}_t$ is an attention value, f_w is a fully-connected layer, and $w_{<t}$ represents words that have been generated before time t.

L_{word} represents the loss of generated words, which is calculated based on the ground truth caption $\mathcal{S}^* = \{w_1^*, w_2^*, \cdots, w_T^*\}$, where T denotes the length of the ground truth caption, as follows,

$$L_{word} = -\sum_{t=1}^{T} \log p(w_t^* | w_{<t}^*, I). \tag{25}$$

D. Optimization

Beyond the optimal objective generation, we introduce a classification task to supervise the optimization of our proposed model, which is able to alleviate the over-smooth problem through distinguishing the node types. Specifically, with the obtained representations $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$ of scene graph nodes associated with their types $\mathcal{C}^* = \{c_1^*, c_2^*, \cdots, c_N^*\}$, where N denotes the number of nodes, we calculate the probability distribution $p(c_i|\mathbf{x}_i)$ for each node. Formally,

$$p(c_i|\mathbf{x}_i) = softmax(f_c(\mathbf{x}_i)), \tag{26}$$

where f_c is a fully-connected layer.

The L_{node} denotes the loss of nodes classification, which is calculated based on their types. Formally,

$$L_{node} = -\sum_{i=1}^{N} \log p(c_i^* | \mathbf{x}_i). \tag{27}$$

Therefore, we combine the generation and classification loss functions to formulate the new objective function in our task, formally,

$$L = L_{word} + \gamma * L_{node}, \tag{28}$$

where γ is the hyper-parameter to combine the classification and generation tasks.

Afterward, we also employ reinforcement learning [20] to emphasize the sentence-level generation in the caption $\mathcal{S}$ and take CIDEr as the reward to fine-tune the learned parameters,

$$R_{RL} = E_{\mathcal{S} \sim p(\mathcal{S}|I,\mathcal{V})}(r(\mathcal{S}^*, \mathcal{S})), \tag{29}$$

where r denotes the function of reward.

IV. Experiments

A. Dataset and Metrics

1) MSCOCO: It is widely used in image captioning. We conduct the experiments on a benchmark version presented by Karpathy, whose image-caption pairs are split into training, validation, and testing sets in ratio 113287:5000:5000. We perform an abundance of ablation experiments on the MSCOCO benchmark [30] and compare it to other state-of-the-art models using the Karpathy split.

2) Metrics: In this paper, we adopt six standard metrics to evaluate the generated captions, including BLEU [31], METEOR [32], ROUGE-L [33], CIDEr [34] and SPICE [35].

B. Implementation Details

1) Attribute Extraction: We select the 200 most frequent attributes as the bag labels appearing in the captions. Attributes can be adjectives, quantifiers, nouns, or verbs that modify the object. Since the object categories in the pre-trained and ground truth datasets are inconsistent, we make an object mapping from the ground truth captions to the pre-trained data [14]. In the training phase, we filter out the bags holding the negative tag for all candidate attributes to maintain the stable ratio of positive and negative bags. After the filtering, the bags in the training, validation, and testing sets are 344712, 15133, and 15092, respectively. We select the attribute extractor model with the highest F1 score.

2) Encoder and Decoder: We adopt the dual-GAT model as an encoder to encode the semantic information contained in the image. In each GAT layer, the number of attention heads is set to 2. And we implement the decoder following the work [7].

3) Training and Testing Strategy: In the training phase, the batch size is 64. In cross-entropy training, the initial learning rate and decay are 3e-4 and 0.8 every three epochs, respectively. The model has trained for 30 epochs in the period. In reinforcement learning, the initial learning rate is 3e-5 and remains unchanged. The model has trained for another 30 epochs. In the testing phase, the beam size is 2 to generate a description for the input image.

TABLE I

THE PERFORMACES OF AGAN AND OTHER STATE-OF-THE-ART METHODS ON MSCOCO DATASET (KARPATHY SPLIT). B@1, B@4, ME, RG, CD AND SP DENOTE BLEU-1, BLEU-4, METEOR, ROUGE-L, CIDER-D AND SPICE, RESPECTIVELY.

	Cross Entropy						CIDEr-D Optimizing					
	B@1	B@4	ME	RG	CD	SP	B@1	B@4	ME	RG	CD	SP
SCST (2017)	-	31.3	26.0	54.3	101.3	-	-	33.3	26.3	55.3	111.4	-
LSTM-A (2017)	75.4	35.2	26.9	55.8	108.8	20.0	78.6	35.5	27.3	56.8	118.3	20.8
StackCap (2018)	76.2	35.2	26.5	-	109.1	-	78.6	36.1	27.4	-	120.4	-
Up-Down (2018)	77.2	36.2	27.0	56.4	113.5	20.3	79.8	36.3	27.7	56.9	120.1	21.4
CAVP (2018)	-	-	-	-	-	-	-	38.6	28.3	58.5	126.3	21.6
GCN-LSTM (2018)	77.3	36.8	27.9	57.0	116.0	20.9	80.5	38.2	28.5	58.3	127.6	22.0
SGAE (2019)	77.6	36.9	27.7	57.2	116.7	20.9	80.8	37.4	28.4	58.6	127.8	22.1
VSUA (2019)	-	-	-	-	-	-	-	38.4	28.5	58.4	128.6	22.0
MT-I (2020)	78.1	38.4	28.2	58.0	119.0	21.1	80.8	38.9	28.8	58.7	**129.6**	22.3
MT-II (2020)	77.9	38.0	28.1	57.6	117.8	21.3	80.5	38.6	28.7	58.4	128.7	22.4
AEGAN (Ours)	**78.4**	**38.9**	**28.7**	**58.1**	**120.1**	**22.0**	**83.1**	**40.9**	**29.4**	**59.9**	128.6	**23.1**

C. Quantitive Analysis

To prove the effectiveness of our method, we compare it with the state-of-the-art baselines. The introduction of these baselines is as follows:

- SCST [20]: Reinforcement learning is used to optimize the sequence-level reward CIDEr score for solving the problem of exposure bias and inconsistent evaluation of training and testing.
- LSTM-A [23]: It integrates attributes extracted from the image into the image captioning framework.
- StackCap [36]: It proposes a coarse-to-fine multistage prediction framework for image captioning.
- Up-Down [7]: It proposes a combined bottom-up and top-down attention mechanism to pay attention to the salient image regions.
- CAVP [37]: It sets up a sentinel to determine whether the visual information helps generate the current word.
- GCN-LSTM [24]: It explores the relationship between visual objects and uses GCN to encode the object and relationship.
- SGAE [8]: It incorporates the language inductive bias into the encoder-decoder image captioning framework for more human-like captions.
- VSUA [16]: It exploits the alignment between linguistic words and visual semantic units for image captioning.
- MT-I and MT-II [14]: It makes full use of the information of ground truth captions to generate a description for the input image.

From Table I, we can find that AEGAN outperforms the state-of-the-art baselines in most cases, as illustrated in Table I. It demonstrates the positive influence of the textual attribute information in image captioning and the effectiveness of our proposed model.

Diving into the comparison in terms of two optimization strategies (*i.e.,* cross-entropy loss minimization *v.s.* reinforcement learning), we also have two key findings.

- The improvement of our proposed model with reinforcement learning is more significant than that of other baselines. We own the significant improvement to the fine-grained information modeling. Different from the baselines, AEGAN models much more fine-grained cues, which provides the information source to capture their corresponding words. On the contrary, the information is ignored by the encoders of baselines. These results again verify the contributions of our proposed model.
- Although the reward in the reinforcement learning is calculated by CIDEr, the achieved result *w.r.t.* CIDEr is suboptimal. By comparing the generated captions of various models, we suggest that the phenomenon may be caused by the *tf-idf* based metric (*i.e.,* CIDEr). Due to more attention on the attributes generation, it is inevitable to affect the generation of verbs whose *tf-idf* value tends to be higher than the attributes'.

TABLE II

THE PERFORMANCES (CROSS-ENTROPY OPTIMIZATION) OF ABLATIVE BASELINES.

	Cross Entropy					
	B@1	B@4	ME	RG	CD	SP
Base	78.1	38.2	28.5	57.8	118.0	21.7
Visual Attribute	77.9	38.1	28.6	57.4	118.1	21.4
Textual Attribute	78.3	38.5	28.5	58.0	119.0	21.8
DGM /w 1 layer	78.5	38.4	28.6	58.1	119.3	21.9
DGM /w 2 layers	78.4	38.8	28.7	58.0	119.0	21.8
DGM /w 3 layers	78.4	38.5	28.6	58.0	118.4	21.8
$\gamma = 0.1$	78.4	38.9	28.7	58.1	119.9	21.9
$\gamma = 0.2$	78.4	38.9	28.7	58.1	120.1	22.0
$\gamma = 0.3$	78.4	38.8	28.6	58.1	119.7	21.8
$\gamma = 0.4$	78.2	38.5	28.5	58.0	118.6	21.8
AEGAN	78.4	38.9	28.7	58.1	120.1	22.0

D. Ablative Studies

As shown in Table II, to justify our proposed model, we do ablative studies based on four groups.

- **Base model:** It forgoes the attribute nodes in the scene graph.
- **Visual and Textual Attribute models:** They integrate the attribute into the scene graph and then use GCN for graph representation learning. Whereinto, visual attributes

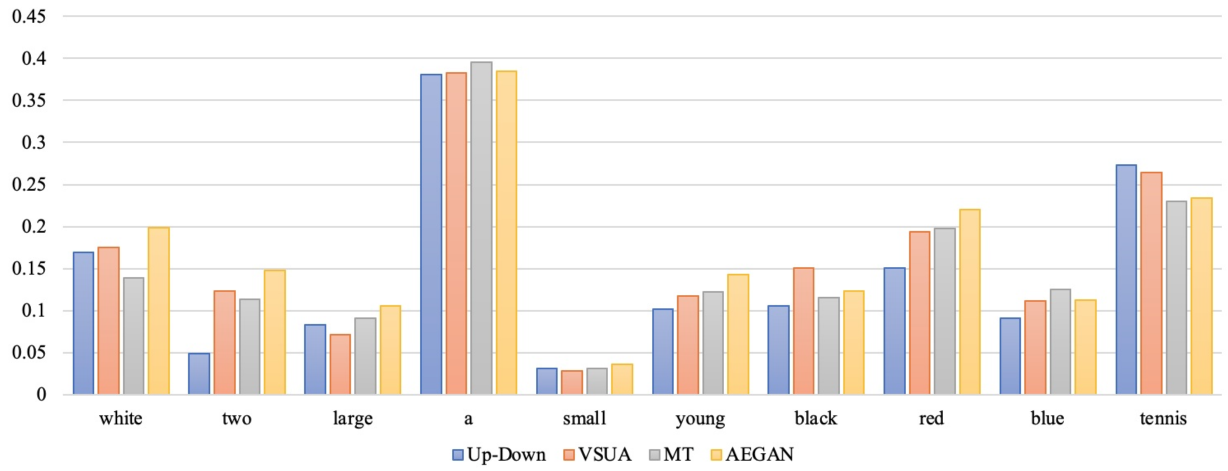


Fig. 4. The recall of captions generated by four state-of-the-art models on the top-10 common attributes. The Up-Down, VSUA and MT are the three state-of-the-art models, and the AEGAN is our proposed model in this paper.

are derived from the pre-trained model on the VG dataset and textual attributes are extracted from the ground truth captions by the multi-instance based attribute extractor.

- **Dual GAT Models (DGMs):** They adopt the DGM over the attribute-enhanced scene graph for optimization without the node classification. In this group, three variants hold different numbers of graph attention layers.
- **Multi-task models:** They introduce a node classification task to avoid the over-smoothing problem. There are four variants with different values for parameter γ in this group.

From Table II, we have the following findings.

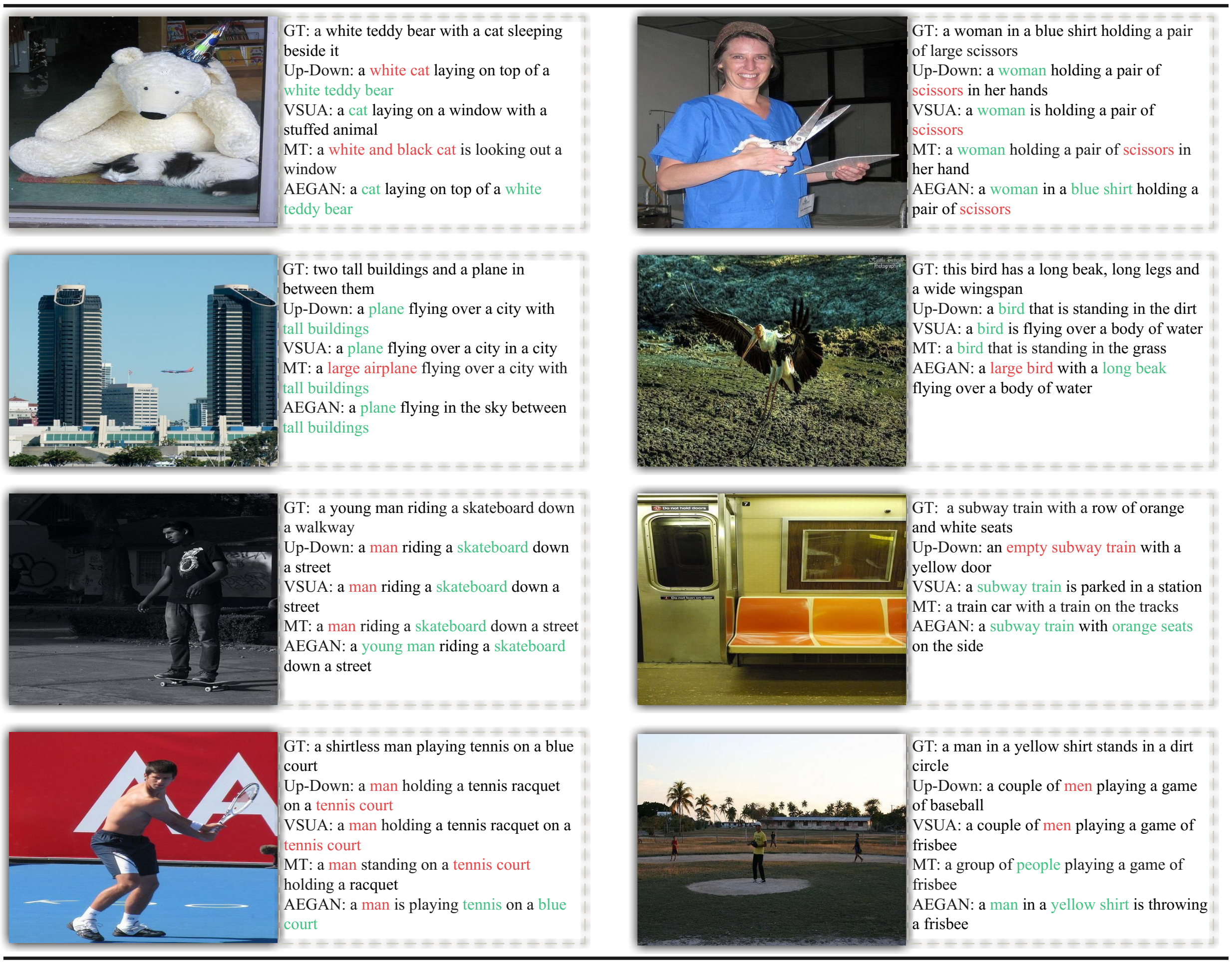
- The performance of **Textual Attribute model** is better than those of **Base model** and **Visual Attribute model** in most cases. We believe that the extracted attributes not only enrich the object representation but capture the fine-grained information to optimize the generated captions as well. The textual attributes extracted by the multi-instance attribute extractor can fill the gap between the pre-trained and the target datasets. Therefore, we indicate that the textual attribute hidden in the ground truth caption can help object representation learning and fine-grained caption generation.
- DGM with one layer is superior to the one without GAT, *i.e.,* **Textual Attribute model**. Based on the observation, we consider that the various attributes or relations probably exhibit the different effects on the object representation learning and the DGM could model them to optimize the image caption generation. By comparing the model performances with various layers, we find that the more layers, the worse the performance. We hold that due to the small number of nodes and node types in the scene graph, even if the number of layers is relatively small, there may still be an over-smoothing problem.
- Compared with **Multi-task models**, **DGMs** achieve the suboptimal results *w.r.t.* B@4, ME, CD, and SP. We believe that it is mainly caused by the over-smoothing representation during the graph convolutional operations, especially when the scene graph has fewer nodes and node types. We select the model whose parameter γ is set to 0.2 as the final model, *i.e.*, AEGAN. Equipped with the node classifier, AEGAN is able to alleviate the over-smoothing problem and generate a more suitable caption for the given image.

E. Qualitative Analysis

1) Attribute Statistics: To justify that the description generated by AEGAN can describe objects in a fine-grained manner by utilizing the attributes in the ground truth captions, we calculate the recall of the top-10 common attributes in descriptions generated by several state-of-the-art models. As shown in Figure 4, the recall score of AEGAN with six attributes, *i.e.*, white, two, large, small, young, and red, is the highest in the four models. We can find that the description generated by AEGAN contains more attributes and can describe the object in a more fine-grained way.

2) Case Study: To visiually demonstrate our proposed model [38], [39], we randomly select eight image-caption pairs to illustrate that the caption generated by AEGAN holds the discriminative and fine-grained attributes. As shown in Figure 5, each image accompanies five different captions, where GT means the ground truth caption, and Up-Down, VSUA, MT, and AEGAN represent the captions generated by corresponding models. Moreover, we mark the attribute-object pairs consistent with the ground truth in green and the inconsistent ones in red.

From the images associated with the captions, we find that AEGAN can collect the fine-grained attributes of objects involved in the ground truth captions, such as “white” for “teddy bear” and “tall” for “building”. However, the other models ignore attribute information during the caption generation. For instance, the caption generated by Up-Down discards “young” for “man”, and the captions from VSUA and MT neglect “in yellow shirt” for “man”. Our proposed model can capture the attributes from visual information and generate a fine-grained description for the given image.



GT: a white teddy bear with a cat sleeping beside it
Up-Down: a white cat laying on top of a white teddy bear
VSUA: a cat laying on a window with a stuffed animal
MT: a white and black cat is looking out a window
AEGAN: a cat laying on top of a white teddy bear

GT: a woman in a blue shirt holding a pair of large scissors
Up-Down: a woman holding a pair of scissors in her hands
VSUA: a woman is holding a pair of scissors
MT: a woman holding a pair of scissors in her hand
AEGAN: a woman in a blue shirt holding a pair of scissors

GT: two tall buildings and a plane in between them
Up-Down: a plane flying over a city with tall buildings
VSUA: a plane flying over a city in a city
MT: a large airplane flying over a city with tall buildings
AEGAN: a plane flying in the sky between tall buildings

GT: this bird has a long beak, long legs and a wide wingspan
Up-Down: a bird that is standing in the dirt
VSUA: a bird is flying over a body of water
MT: a bird that is standing in the grass
AEGAN: a large bird with a long beak flying over a body of water

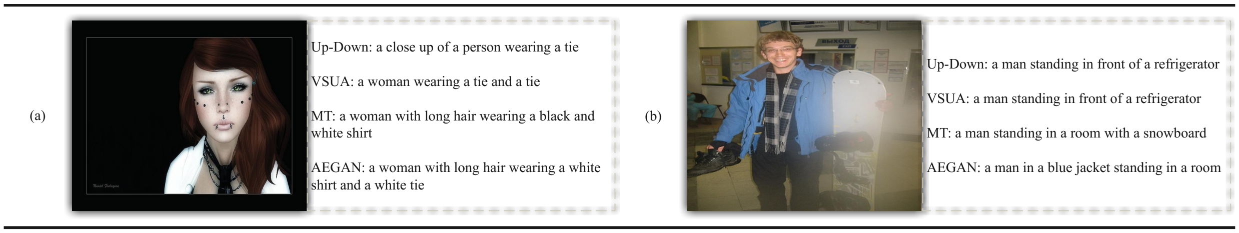
GT: a young man riding a skateboard down a walkway
Up-Down: a man riding a skateboard down a street
VSUA: a man riding a skateboard down a street
MT: a man riding a skateboard down a street
AEGAN: a young man riding a skateboard down a street

GT: a subway train with a row of orange and white seats
Up-Down: an empty subway train with a yellow door
VSUA: a subway train is parked in a station
MT: a train car with a train on the tracks
AEGAN: a subway train with orange seats on the side

GT: a shirtless man playing tennis on a blue court
Up-Down: a man holding a tennis racquet on a tennis court
VSUA: a man holding a tennis racquet on a tennis court
MT: a man standing on a tennis court holding a racquet
AEGAN: a man is playing tennis on a blue court

GT: a man in a yellow shirt stands in a dirt circle
Up-Down: a couple of men playing a game of baseball
VSUA: a couple of men playing a game of frisbee
MT: a group of people playing a game of frisbee
AEGAN: a man in a yellow shirt is throwing a frisbee

Fig. 5. The captions generated by different models. The green and red fonts indicate the consistence and inconsistence with the ground truth caption, respectively.



Up-Down: a close up of a person wearing a tie
VSUA: a woman wearing a tie and a tie
MT: a woman with long hair wearing a black and white shirt
AEGAN: a woman with long hair wearing a white shirt and a white tie

Up-Down: a man standing in front of a refrigerator
VSUA: a man standing in front of a refrigerator
MT: a man standing in a room with a snowboard
AEGAN: a man in a blue jacket standing in a room

Fig. 6. Two examples are selected to show the shortcomings of AEGAN.

Beyond explicitly modeling the attributes, we can also find that the information facilitates the corresponding object representation. In particular, AEGAN can capture several objects, like “shirt”, “beak”, and “seat”, from the image, whereas, these objects are neglected by the other models. According to this finding, we conclude that the attribute-enhanced scene graph not only helps the fine-grained caption generation but enriches the object representations, which benefits to generating and describing the objects.

We select two sample images from the dataset to discuss the AEGAN shortcomings, as shown in Figure 6. In the caption generated by AEGAN for the image in Figure 6(a), the textual attribute of object “tie” is “white”, whereas the visual attribute is “black” in the sample image. The reason is that the multi-instance based attribute extractor extracts the wrong attribute. As we all know, MIL optimizes the model on the dataset that does not rely on a large amount of artificially annotation information. However, it is not as accurate as supervised learning. Therefore, extracting the wrong attributes of objects is inevitable. For the sample image in Figure 6(b), in the caption generated by AEGAN, the object “man” has more fine-grained description information, *i.e.*, “in a blue jacket”. However, the fine-grained description of some content may cause AEGAN to ignore the other contents in the image. For example, the objects, “refrigerator” and “snowboard”, contained in the image, are described by other models but not by AEGAN.

V. Conclusion and Future Work

In this paper, we focus on explicitly constructing the correlation between visual object and textual attribute to enhance the scene graph, so as to optimize the image captioning models. To this end, we develop an attribute-enhanced scene graph attention network to generate the caption for the given image. In particular, we design a multi-instance based attribute extractor to extract textual attributes from the ground truth captions. For the object representation and scene graph construction, we also refine the object structure with attribute information and enhance the scene graph to handle attribute-object and object-object correlations. Moreover, the DGM is proposed to aggregate attributes and relations according to their similarities to the objects. By conducting extensive experiments over the MSCOCO dataset, we demonstrate that our proposed model, *i.e.*, AEGAN, outperforms the state-of-the-art performance. Moreover, we visually exhibit that the generated captions contain more fine-grained information.

In the future, we will still focus on filling the gap between the pre-trained and target datasets. For example, most objects in the image recognized by the object detector pre-trained on the VG dataset never show in the ground truth caption. Therefore, recognizing objects from the given image according to the ground truth caption is a new problem. However, the MIL is seemingly not suitable for solving the problem. How to effectively deal with this problem is the focus of our future work.

References

[1] M. Yang, W. Zhao, W. Xu, Y. Feng, Z. Zhao, X. Chen, and K. Lei, "Multitask learning for cross-domain image captioning," *IEEE Transactions on Multimedia*, vol. 21, no. 4, pp. 1047–1061, 2018.

[2] L. Wu, M. Xu, J. Wang, and S. Perry, "Recall what you see continually using gridlstm in image captioning," *IEEE Transactions on Multimedia*, vol. 22, no. 3, pp. 808–818, 2019.

[3] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3156–3164.

[4] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[5] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[6] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*. PMLR, 2015, pp. 2048–2057.

[7] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6077–6086.

[8] X. Yang, K. Tang, H. Zhang, and J. Cai, "Auto-encoding scene graphs for image captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 685–10 694.

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

[10] F. Liu, Z. Cheng, C. Sun, Y. Wang, L. Nie, and M. Kankanhalli, "User diverse preference modeling by multimodal attentive metric learning," in *Proceedings of the 27th ACM International Conference on Multimedia*. ACM, 2019, p. 1526–1534.

[11] Y. Wei, X. Wang, L. Nie, X. He, R. Hong, and T.-S. Chua, "Mmgcn: Multi-modal graph convolution network for personalized recommendation of micro-video," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 1437–1445.

[12] Y. Wei, X. Wang, L. Nie, X. He, and T.-S. Chua, "Graph-refined convolutional network for multimedia recommendation with implicit feedback," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 3541–3549.

[13] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE transactions on neural networks*, vol. 20, no. 1, pp. 61–80, 2008.

[14] Z. Shi, X. Zhou, X. Qiu, and X. Zhu, "Improving image captioning with better use of caption," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 7454–7464.

[15] J. Lu, C. Xiong, D. Parikh, and R. 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*, 2017, pp. 375–383.

[16] L. Guo, J. Liu, J. Tang, J. Li, W. Luo, and H. Lu, "Aligning linguistic words and visual semantic units for image captioning," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 765–773.

[17] S. Chen, Q. Jin, P. Wang, and Q. Wu, "Say as you wish: Fine-grained control of image caption generation with abstract scene graphs," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9962–9971.

[18] S. Liu, Z. Zhu, N. Ye, S. Guadarrama, and K. Murphy, "Improved image captioning via policy gradient optimization of spider," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 873–881.

[19] M. Ranzato, S. Chopra, M. Auli, and W. Zaremba, "Sequence level training with recurrent neural networks," 2015.

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

[21] J. Foulds and E. Frank, "A review of multi-instance learning assumptions," *The knowledge engineering review*, vol. 25, no. 1, pp. 1–25, 2010.

[22] H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt *et al.*, "From captions to visual concepts and back," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1473–1482.

[23] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei, "Boosting image captioning with attributes," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4894–4902.

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

[25] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[26] J. H. Tan, C. S. Chan, and J. H. Chuah, "Comic: Toward a compact image captioning model with attention," *IEEE Transactions on Multimedia*, vol. 21, no. 10, pp. 2686–2696, 2019.

[27] W. Ji, X. Li, L. Wei, F. Wu, and Y. Zhuang, "Context-aware graph label propagation network for saliency detection," *IEEE Transactions on Image Processing*, vol. 29, pp. 8177–8186, 2020.

[28] C. Zhang, J. Platt, and P. Viola, "Multiple instance boosting for object detection," *Advances in neural information processing systems*, vol. 18, pp. 1417–1424, 2005.

[29] R. Zellers, M. Yatskar, S. Thomson, and Y. Choi, "Neural motifs: Scene graph parsing with global context," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5831–5840.

[30] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick, "Microsoft COCO captions: Data collection and evaluation server," *arXiv preprint arXiv:1504.00325*, 2015.

[31] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[32] M. Denkowski and A. Lavie, "METEOR universal: Language specific translation evaluation for any target language," in *Proceedings of the ninth workshop on statistical machine translation*, 2014, pp. 376–380.

[33] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.

[34] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4566–4575.

[35] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "SPICE: Semantic propositional image caption evaluation," in *European conference on computer vision*. Springer, 2016, pp. 382–398.

[36] J. Gu, J. Cai, G. Wang, and T. Chen, "Stack-captioning: Coarse-to-fine learning for image captioning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[37] D. Liu, Z.-J. Zha, H. Zhang, Y. Zhang, and F. Wu, "Context-aware visual policy network for sequence-level image captioning," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 1416–1424.

[38] Y. Wei, Z. Cheng, X. Yu, Z. Zhao, L. Zhu, and L. Nie, "Personalized hashtag recommendation for micro-videos," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 1446–1454.

[39] Y. Wei, X. Wang, W. Guan, L. Nie, Z. Lin, and B. Chen, "Neural multimodal cooperative learning toward micro-video understanding," *IEEE Transactions on Image Processing*, vol. 29, pp. 1–14, 2019.