Article

Middle-Level Attribute-Based Language Retouching for Image Caption Generation

Zhibin Guan ¹, Kang Liu ^{1,*} , Yan Ma ¹, Xu Qian ¹ and Tongkai Ji ^{1,2}

¹ School of Mechanical Electronic & Information Engineering, China University of Mining & Technology (Beijing), Beijing 100083, China; gzbcumtb@163.com (Z.G.); macumtb@163.com (Y.M.); xuqian@cumtb.edu.cn (X.Q.); jtk@gdei.com.cn (T.J.)

² G-Cloud Technology Corporation, Cloud Computing Center, Chinese Academy of Sciences, Dongguan 523808, China

* Correspondence: Kangliu@cumtb.edu.cn

Received: 27 August 2018; Accepted: 2 October 2018; Published: 9 October 2018

Abstract: Image caption generation is attractive research which focuses on generating natural language sentences to describe the visual content of a given image. It is an interdisciplinary subject combining computer vision (CV) and natural language processing (NLP). The existing image captioning methods are mainly focused on generating the final image caption directly, which may lose significant identification information of objects contained in the raw image. Therefore, we propose a new middle-level attribute-based language retouching (MLALR) method to solve this problem. Our proposed MLALR method uses the middle-level attributes predicted from the object regions to retouch the intermediate image description, which is generated by our language generation model. The advantage of our MLALR method is that it can correct descriptive errors in the intermediate image description and make the final image caption more accurate. Moreover, evaluation using benchmark datasets—MSCOCO, Flickr8K, and Flickr30K—validated the impressive performance of our MLALR method with evaluation metrics—BLEU, METEOR, ROUGE-L, CIDEr, and SPICE.

Keywords: image captioning; middle-level attributes; language retouching; MLALR

1. Introduction

Image caption generation is attractive research in the field of artificial intelligence and has emerged in recent years. It is an interdisciplinary subject combining computer vision (CV) and natural language processing (NLP). It focuses on generating a readable sentence to depict the visual content of a given image. The representative applications of image caption generation are assisting visually impaired people to perceive the visual content of the surrounding scenes, providing visual intelligence to chatting robots, image retrieval, and semantic visual research. Therefore, image caption generation is a significant part of scene understanding.

Benefiting from the rapid development of deep learning, many novel methods have been proposed for image caption generation in recent years. Researchers have used convolution neural networks (CNNs) [1] to extract the visual information of a given image, and to generate the corresponding language description by a decoder, such as a recurrent neural network (RNN) [2]. Briefly, most methods of image caption generation are based on an encoder-decoder framework. CNN-based models are the most common encoders, such as VGG16 [3] and ResNet-101 [4]. Meanwhile, RNN-based models are commonly used encoders, such as Gated Recurrent Unit (GRU) [5], Bidirectional RNNs (Bi-RNNs) [6], Long Short-Term Memory cell (LSTM) [7], etc.

However, most methods for image caption generation based on the encoder-decoder framework have an obvious limitation. The image feature extracted by CNNs is a fixed-length feature vector,

which may lose significant identification information of objects. For example, the attention-based methods of image captioning, which use an attention mechanism to decide which part of the input information is important, cannot fully exploit the role of the attention mechanism. The contribution of an attention mechanism is that it can fuse several feature vectors by assigning a weight for each input feature vector, and then generating one fused feature vector. Nonetheless, the attention mechanism may weaken the role of some significant information, such as the shape of objects. The attribute-based methods of image caption generation extract high-level semantic attributes from the raw image, such as “running”, “standing”, “birds”, etc., and then directly inject them into the language generation model. The obvious problem with these methods is that they ignore the role of the middle-level attributes of objects contained in the raw image, such as color, texture, etc.

Therefore, we propose a new middle-level attribute-based language retouching (MLALR) method for image caption generation to solve the problem discussed earlier. Our MLALR method is based on an encoder-decoder framework. As opposed to other methods, MLALR extracts the middle-level attributes from the object regions and then uses them as supplementary information to retouch the intermediate image caption generated by our language generation model. It is worth noting that the middle-level attributes are not included in the input information of our attention mechanism. We tried to separate the middle-level attributes and use sequential means to generate the final image caption. One advantage of the sequential method is that the role of middle-level attributes will not be weakened by the attention mechanism. After that, the well-described natural language sentence of the raw image can be generated by our method.

The method we propose consists of four key parts: (1) image feature extraction; (2) middle-level attribute prediction; (3) language generation model; and (4) language retouching of the intermediate image caption. The first two parts of our method are used to extract valuable information from the raw image. The last two parts of our method mainly focus on generating the intermediate image caption and retouching it, respectively.

In the first part of the study, we used the ResNet-101 model to extract the global image feature, since it can extract accurate high-level semantic features compared with VGG16. Moreover, given that the accuracy and robustness of faster-RCNN model [8] are much greater than those of other models, it can simultaneously generate object features and object regions. Therefore, we used faster-RCNN to extract the local image features and regions of objects. In the second part, we combined a series of VGG16 classifiers into an ensemble attribute predictor to predict the middle-level attributes from the object regions. In the third part, since the LSTM cell can maintain long-term dependence of information to a certain extent, we use a stacked two-layer RNN with LSTM cell to generate intermediate image captions. In the last part, the extracted middle-level attributes were applied to retouch the intermediate image caption and to generate the final well-described image caption.

Finally, the main contributions of our work are as follows:

- We propose a new MLALR method for image captioning.
- We extracted different information from the raw image, including the global image feature, local image features, and the middle-level attributes, all of which can complement each other.
- We used the predicted middle-level attributes to retouch the intermediate image caption and generate the final well-described image caption.
- Evaluation results indicate that our method can achieve impressive performance with several popular evaluation metrics, including BLEU [9], METEOR [10], ROUGE-L [11], CIDEr [12], and SPICE [13].

Remainder. The remainder of this paper is organized as follows. We review the related work of image caption generation in Section 2. In Section 3, we introduce our proposed method in detail. Then, the results and discussion are given in Section 4. Finally, we summarize our work with future research in Section 5.

2. Related Work

In this section, we review the related work of image caption generation. The commonly used framework for image captioning is the encoder-decoder framework, which has been extensively studied recently [14–22]. The related work can be broadly divided into two categories: attention-based methods and attribute-based methods.

2.1. Attention-Based Methods

An attention mechanism is a powerful algorithm which is inspired by human's visual attention. It is widely used in CV and NLP. Therefore, various attention-based methods have been proposed for image caption generation.

The “hard” and “soft” attention methods, proposed by [23], are used to better understand the visual content of the raw image. In [23], the “hard” and “soft” attention methods were respectively applied to two different image captioning frameworks. “Hard” attention is not dependent on all hidden states of the language model, and the gradient in “hard” attention needs to be estimated by Monte Carlo-based sampling. On the other hand, “soft” attention is used to calculate the weights of all input feature vectors and to generate an encoded feature vector. Therefore, “soft” attention is a parameterized method, and it can be embedded in the language model for training directly. Most existing methods use “soft” attention as the basic attention mechanism.

To solve the problem of the visual attention mechanism, i.e., that it is active all the time, Ref. [24] proposed an adaptive encoder-decoder framework to automatically decide whether to use the attention mechanism or not. In [24], some words could be generated by previous words, such as “sign” after “behind a red stop”. In this case, the attention mechanism is unnecessary. On the other hand, the attention mechanism is active when generating other words which have little relationship with the previously generated words, such as “of”, “the”, etc. The shortcoming of the method in [24] is that the attention mechanism might be suppressed all the time when generating image captions.

Furthermore, to address problems related to object missing and object misjudgment, Ref. [25] proposed a global-local attention (GLA) method to solve these issues during the image captioning process. The advantage of the method in [25] is that the image features are split into two parts, one of which is used as the local image features and is integrated by “soft” attention. The drawback of GLA is that the model is not fine-tuned in different datasets. In [26], the authors proposed a novel attention-based model for automatic image captioning named “Areas of attention”. It can be trained without bounding-box supervision. The contribution of the method in [26] is that the corresponding image areas can be marked when generating words at each time step.

The authors in [27] proposed a sequence-to-sequence RNN for image caption generation. Different from previous methods, [27] made the input image a sequence of detected objects to generate the corresponding captions by using an attention mechanism. The advantage of the method in [27] is that a sequential attention layer is introduced when generating each word. However, the middle-level attributes of the object were not taken into consideration in [27].

Furthermore, Ref. [28] proposed a bottom-up and top-down method for image caption and visual question answering (VQA), which also uses an attention mechanism to integrate the objects and other salient image regions. The advantage of the method in [28] is that “soft” and “hard” attention are combined to achieve image captioning.

As described above, the attention mechanism is usually used to integrate different pieces of image information during the word generation process. The power of the attention mechanism relies on extracting sufficient image information.

2.2. Attribute-Based Methods

There are various attribute-based methods that have been proposed for image caption generation. In [29], the high-level semantic information of the raw image was extracted for image captioning

and VQA, and this achieved impressive performance improvement. The advantage of the method from [29] is that non-parametric attribute prediction and parametric attribute prediction are used as two different methods to extract the visual attribute labels.

Moreover, in [30], an LSTM-A method was proposed for image captioning. The authors in [30] explored the impact of high-level visual attribute labels by using five different frameworks with different insertion locations of attributes, including “Leveraging only Attributes”, “Inserting Image First”, “Feeding Attribute First”, “Inputting Image each time step”, and “Inputting Attributes each time step”. In [30], the role of high-level visual attributes, such as “outdoor”, “riding”, and “market”, can be observed.

Meanwhile, the middle-level attributes of images are integrated by low-level image features in the method proposed in [31], which enhanced the precision of the attribute labels. The low-level image features are tiny information of objects, including edges, corners, pixels, and so on.

Furthermore, in terms of some specific application scenarios, such as supermarkets, researchers took user-contributed tags as the image information to recognize the specific objects contained in the raw image. The user-contributed tags are attributes which can reflect the user’s attention, such as “camera” or “surfboard” held by users, etc. For example, Ref. [32] combined the visual attention and user attention simultaneously for social image captioning. In addition, due to the good performance and robustness of some existing image generation methods, they can also be extended and applied to other fields, such as VQA and video captioning [29,33,34]. The main difference between image captioning and VQA is whether a machine can respond well to the input question information.

From all of the above methods, we can observe that the process of image caption generation can be broadly divided into several parts, including visual feature extraction from the raw images, information fusion of visual features and attribute labels, and descriptive sentence generation.

The problem with the existing image captioning methods is that they use an encoder-decoder framework to generate the final image caption directly, which ignores the role of the middle-level attributes of objects. Therefore, our research focuses on using the middle-level attributes of objects to retouch the intermediate image caption generated by our language model and to generate a final well-described image caption.

3. Materials and Methods

In this section, we introduce the contributions of this paper, including global and local image feature extraction, the prediction of the middle-level attributes of objects, the language generation model, and language retouching of the intermediate image caption. The framework of our proposed method can be seen in Figure 1. The main purpose of our work is to correct descriptive errors in the intermediate image caption and to make the final image caption more accurate.

Figure 1. The framework of our proposed middle-level attribute-based language retouching method.

The main idea of our research is to use the middle-level attributes to retouch the intermediate image caption generated by our language generation model, which can solve the problem discussed before.

3.1. Image Feature Extraction

The image features we used are the global image feature and local image features. The global image feature is used to make the language generation model have a general understanding of the raw image. Meanwhile, the local image features are used as fine-grained information for image captioning.

In our method, the global image feature is extracted by ResNet-101 [4], which was pre-trained on the ImageNet classification dataset [35]. Since the number of the last fully connected layer of the ResNet-101 model is 2048, the extracted global image feature of each image is a 2048-dimension vector, denoted as $\mathcal{G}$ (see Equation (1)):

$$\mathcal{G} \in \mathbb{R}^{2048}. \quad (1)$$

The local image features we use are the feature vectors of object regions. Hence, we used faster-RCNN [8] as the local image feature extractor, as it can generate object features and object regions separately. The object features are used as local image features, and the object regions are used to predict the middle-level attribute labels, as described in next subsection (see Table 1).

Table 1. Possible objects and object regions contained in raw images.

Categories	Optional Values
Objects	child, man, woman, desk, polar bear, rock, cat...
Object regions	the corresponding bounding-box: (P_x, P_y, P_w, P_h)

The column named “optional values” in Table 1 denotes the possible objects and the corresponding object regions in the raw images. The bounding-box (P_x, P_y, P_w, P_h) can be used to denote an object region, where P_x and P_y represent the coordinates of the center point of the bounding-box, and P_w and P_h denote the width and height of the bounding-box, respectively.

Since the basic deep neural network of faster-RCNN is ResNet-152, the generated object feature vectors are all 2048-dimension, denoted as set $\mathcal{O}$ (see Equation (2)).

$$\mathcal{O} = \{o_1, o_2, \dots, o_K\}, o_j \in \mathbb{R}^{2048}, j \in (1, K). \quad (2)$$

3.2. Middle-Level Attributes Prediction

In this part, we mainly focus on predicting the middle-level attribute labels from object regions, which can be used to retouch the intermediate image description generated by our language generation model. The process can be broadly divided into three steps, including extracting the object regions, predefining the middle-level attributes, and training and applying the middle-level attribute predictors.

From the above subsection, we know that the object features and object regions of the raw image can be generated simultaneously by the faster-RCNN model. Hence, we used the object regions as the raw data to predict the valuable middle-level attribute labels.

Since the categories of the objects contained in the raw image can be recognized by faster-RCNN, we only need to predict the middle-level attribute labels of these objects. We roughly used some valuable middle-level attributes of human and non-human objects predefined in previous works [36,37].

As described in [36], the images in the PubFig dataset consist of positive and negative examples, all of which have been labelled with “0” or “1” for each attribute to indicate whether an attribute exists or not. The attributes are in [36], including expression, lighting, scene, etc. Thus, the PubFig dataset can also be used for training binary classifiers to recognize the presence or absence of describable aspects of visual appearance, such as gender, age, etc.

Therefore, the attribute labels used to retouch a human's appearance in our research are a subset of those used in [36]. The details of the human attributes are shown in Table 2, including gender, age, and hair color.

Table 2. Attributes of humans.

Categories	Optional Values
Gender	Male, female
Age	Child, youth, middle aged, senior
Hair color	Black, white, blonde, brown, gray

The ImageNet dataset [35] contains a subset [37] which has 9600 images collected from 384 synsets, and each image is paired with 25 object attributes. The 25 semantic appearance attributes in [35] can be divided into four categories, including “color”, “pattern”, “shape”, and “texture”. Hence, the middle-level attributes of non-human objects used in this research are the semantic appearance attributes, which are similar to those in [37] and are illustrated in Table 3.

Table 3. Attributes of non-human objects.

Categories	Optional Values
Shape	Long, round, rectangular, square
Color	Black, white, gray, blue, green, red, pink, yellow, orange, brown, violet
Texture	Furry, smooth, rough, shiny, metallic, vegetation, wooden, wet

Finally, we combined multiple attribute predictors into an ensemble attribute predictor to predict the valuable middle-level attributes from the object regions. We used the VGG16 [3] model as the basic classifier, which was trained six times separately for different aims, as shown in Table 4.

We used the PubFig dataset to train our attribute predictors VGG16(GENDER), VGG16(AGE), and VGG16(HAIR COLOR) to predict human attributes. The attribute predictors VGG16(SHAPE), VGG16(COLOR), and VGG16(TEXTURE) were pre-trained on the subset of the ImageNet dataset to recognize the middle-level attributes of non-human objects.

Table 4. Attribute predictors with different aims.

Classifiers	Classification Aims
VGG16(GENDER)	Male, female
VGG16(AGE)	Child, youth, middle aged, senior
VGG16(HAIR COLOR)	Black, white, blonde, brown
VGG16(SHAPE)	Long, round, rectangular, square
VGG16(COLOR)	Black, white, gray, blue, green, red, pink, yellow, orange, brown, violet
VGG16(TEXTURE)	Furry, smooth, rough, shiny, metallic, vegetation, wooden, wet

After that, these pre-trained attribute predictors were combined into an ensemble attribute predictor to predict the middle-level attributes from the object regions generated by faster-RCNN. It is worth noting that only when the probability of a predicted attribute is greater than a predefined threshold is the predicted attribute applied to retouch the intermediate image description. Here, the threshold was set to 70%.

3.3. Language Generation Model

The language generation model we used in this work is a stacked two-layer RNN with LSTM cells. The advantage of the LSTM cell is that it can maintain long-term dependence of information to

a certain extent. Additionally, the two-layer RNNs are used to reserve the information of the global image feature and local image features, respectively.

The detailed calculations of the LSTM cell are denoted as Equations (3)–(5). The input, forget, output, and memory gate are respectively denoted as $i_t^{(l)}$, $f_t^{(l)}$, $o_t^{(l)}$, and $g_t^{(l)}$, where $l \in \{1, 2\}$ indicates the first layer or the second layer of our language generation model.

$$\begin{aligned} i_t^{(l)} &= \sigma \left(W_{ix} x_t^{(l)} + W_{ih} h_{t-1}^{(l)} + b_i^{(l)} \right) \\ f_t^{(l)} &= \sigma \left(W_{fx} x_t^{(l)} + W_{fh} h_{t-1}^{(l)} + b_f^{(l)} \right) \\ o_t^{(l)} &= \sigma \left(W_{ox} x_t^{(l)} + W_{oh} h_{t-1}^{(l)} + b_o^{(l)} \right) \\ g_t^{(l)} &= \varphi \left(W_{gx} x_t^{(l)} + W_{gh} h_{t-1}^{(l)} + b_g^{(l)} \right), \end{aligned} \quad (3)$$

$$c_t^{(l)} = o_t^{(l)} \odot c_{t-1}^{(l)} + i_t^{(l)} \odot g_t^{(l)}, \quad (4)$$

$$h_t^{(l)} = o_t^{(l)} \odot \varphi \left(c_t^{(l)} \right), \quad (5)$$

where $x_t^{(l)}$ denotes the input information of the LSTM cell at time step t , $l \in \{1, 2\}$. Similarly, $h_t^{(l)}$ means the reserved information of a different layer of our language model at time step t . The parameters W_* and b_* are shared and need to be learned in all time steps. $\odot$ denotes element-wise multiplication. σ and φ are the activation functions, shown as Equations (6) and (7).

$$\sigma(z) = \frac{1}{1 + e^{-z}}. \quad (6)$$

$$\varphi(z) = \tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}. \quad (7)$$

The main difference between the first layer and the second layer of our language generation model is the input information. As shown in Figure 2, at time step t , the input information of the first layer, $x_t^{(1)}$, consists of the global image feature $\mathcal{G}$ and the previous generated word s_{t-1} . Hence, the calculation of $x_t^{(1)}$ can be defined as Equation (8).

$$x_t^{(1)} = \mathcal{G} + W_s s_{t-1}, \quad (8)$$

where s_{t-1} belongs to the set S , which is a sentence in a form of a word sequence generated by our language generation model. It is referred to as Equation (9).

$$S = \{s_0, s_1, \dots, s_L\}, \quad (9)$$

where L is the length of the generated descriptive sentence.

What is more, the input information of the second layer contains the fused feature vector v_t and the hidden state $h_t^{(1)}$. In our model, the fused feature vector v_t is calculated from the local image features $\mathcal{O}$ and the hidden state $h_{t-1}^{(2)}$ by using the attention mechanism. It is worth noting that the middle-level attributes are not included in the input information of our attention mechanism. In this way, the role of the middle-level attributes will not be weakened by the attention mechanism.

The cosine function is used to measure the similarity between object feature vector o_j and the hidden state $h_{t-1}^{(2)}$, as shown in Equation (10).

$$\mathcal{M}_t(o_j) = \cos(o_j, h_{t-1}^{(2)}) = \frac{o_j h_{t-1}^{(2)}}{|o_j| |h_{t-1}^{(2)}|}, j \in [1, K]. \quad (10)$$

Then, the weight value $\alpha(o_j)$ of each object feature vector o_j can be calculated, as shown in Equation (11).

$$\alpha_t(o_j) = \frac{\mathcal{M}_t(o_j)}{\sum_j \mathcal{M}_t(o_j)}, j \in [1, K]. \quad (11)$$

After that, the fused feature vector v_t at time step t can be calculated according to the weight value $\alpha(o_j)$ and the object feature vector o_j , which is referred to as Equation (12).

$$v_t = \sum_{j=1}^k \alpha_t(o_j) o_j. \quad (12)$$

Hence, the input information of the second layer, $x_t^{(2)}$, is calculated as Equation (13).

$$x_t^{(2)} = W_\phi v_t + W_h h_t^{(1)}. \quad (13)$$

Furthermore, the generated probability of each word s_t is calculated based on the previously generated words $\{s_1, s_2, \dots, s_{t-1}\}$ and the global and local image features; the detailed operation is shown as Equation (14).

$$p(s_t | s_0, \dots, s_{t-1}, \mathcal{G}, \mathcal{O}) = \text{softmax}(W_p h_t^{(2)} + b_p). \quad (14)$$

Therefore, the generated probability of the entire sentence can be calculated by the product of the probability of each word, as shown in Equation (15).

$$p(s_0, s_1, \dots, s_L) = \prod_{t=1}^L p(s_t | s_0, \dots, s_{t-1}, \mathcal{G}, \mathcal{O}). \quad (15)$$

The loss function of our language model is the negative log-likelihood loss, which is defined as Equation (16).

$$L(S) = - \sum_{t=1}^L \log p(s_t | s_0, \dots, s_{t-1}, \mathcal{G}, \mathcal{O}). \quad (16)$$

Finally, we used Self-Critical Sequence Training (SCST) [38] to achieve CIDEr optimization. The negative expected reward is defined as Equation (17), where r indicates the score function.

$$L(\theta) = -\mathbb{E}_{S \sim p}[r(S)]. \quad (17)$$

The gradient of the negative expected reward can be calculated as Equation (18):

$$\nabla_\theta L(\theta) = -(r(S) - r(\hat{S})) \nabla_\theta \log p(S). \quad (18)$$

where $r(\hat{S})$ is the reward obtained by the current model.

The image description generated by our language generation model is used as an intermediate image caption, which will be retouched by the predicted middle-level attribute labels.

Figure 2. Language Generation model.

3.4. Language Retouching of Intermediate Image Captioning

As we mentioned before, the generated intermediate image caption loses the information of the descriptive middle-level attributes. Therefore, the intermediate image caption is retouched by using the predicted middle-level attributes. It consists of two steps: (1) traversing the intermediate image caption to search for the same fragment as the key index, and (2) replacing the searched fragments with the corresponding short phrases.

Before retouching the intermediate image caption, the middle-level attribute labels generated by our ensemble attribute predictor are combined with the object label to form a short phrase (examples in Table 5). The object labels and the corresponding object regions are generated by the faster-RCNN model. Here, the short phrase is generated without grammar rules. Since the structure of the short phrase is relatively simple, we only need to arrange the middle-level attributes in order, just like a static template. For human objects, the order of middle-level attributes is “age”, “gender”, “hair color”. For non-objects, the order of words is “shape”, “color”, “texture”, and object label.

Table 5. The generated short phrases.

Index	Middle-Level Attributes	Short Phrases
polar bear	white, furry	white furry polar bear
rock	gray, rough	gray rough rock
woman	youth, female, brown hair	youth female with brown hair

In the search step, each object label is used as a key index to search for the same fragment from the intermediate image caption. After that, the searched fragment is replaced by a short phrase according to the key index. For example, in the intermediate image caption, “a polar bear is standing on a rock with its mouth open”, the word “rock” is replaced by “gray rough rock” after one instance of the searching and replacing steps. After all searching and replacing steps are completed, the language retouching of the intermediate image caption is complete.

4. Results and Discussion

4.1. Datasets

The datasets we used in our work are the MS COCO dataset [39], Flickr8K dataset [40], Flickr30K dataset [41], Pubfig dataset [36], and a subset [37] of the ImageNet dataset. The first

three datasets are well-known datasets and can be used for object detection, image segmentation, and captioning. The PubFig dataset and the subset of the ImageNet dataset were used for training our middle-level attribute predictors (see Table 4) to predict the valuable attributes of human and non-human objects, respectively.

The official MS COCO dataset consists of 82,783 training images, 40,504 validation images, and 40,775 test images. However, in our work, we used the 'Karpathy' split [24] for reporting results, as in previous works. Therefore, the MS COCO dataset is split into 113,287 training images, 5000 validation images, and 5000 test images. Additionally, the Flickr8K dataset contains 6000 training images, 1000 validation images, 1000 testing images. The Flickr8K dataset is an extension of the Flickr8K dataset, which consists of 31,783 images. We used 28,000 images for training, 1000 images for validation, and 1000 images for testing. The above three datasets all contain five reference captions for each image.

The PubFig dataset consists of $\sim 10,000$ images, and we used ~ 7000 images for training, 1500 images for validation, and 1500 images for testing. The subset of ImageNet contains 9600 images, which are split into 6900 training images, 1350 validation images, and 1350 testing images.

4.2. Evaluation Metrics

We adopted multiple metrics to evaluate the performance of our method for image captioning, including BLEU [9], METEOR [10], ROUGE-L [11], CIDEr [12], and SPICE [13].

BLEU can be used to evaluate the co-occurrences of n-grams between the reference sentences and the generated captions. METEOR is based on the harmonic mean of uni-gram precision and recall. Its evaluation is at the corpus level. ROUGE-L is based on the Longest Common Subsequence (LCS) of reference sentences and generated captions and is used to capture sentence-level structure.

Different from the above evaluation metrics, CIDEr and SPICE are human consensus metrics. CIDEr can measure the similarity of the generated captions against a set of human-written sentences, and SPICE is used to measure how effectively the image captions recover attributes, objects, and the relationships between them.

4.3. Experiment Setting and Result

The experiment described in our paper was separated into two parts, including the training part and test part (see Figure 3). In the training part, we aimed to obtain the model parameters by training our language generation model and middle-level attribute predictors. In the test part, we separately used the language generation model and the ensemble attribute predictor to generate the intermediate image caption and the middle-level attribute labels. Then, the attribute labels were used to retouch the intermediate image caption to generate a final well-described sentence.

In our language model, the Adam algorithm is used to optimize the cost function (see Equation (16)). The value of alpha and beta used for the Adam algorithm are equal to 0.9 and 0.999, respectively. The embedding size, which is used to translate the generated word to the feature vector, is set to 1000. The number of hidden cells (LSTM) is equal to 1000, and the attention hidden size is set to 512. Furthermore, the sentence generation strategy we used in this research is the beam search method, which can ensure the quality of the generated intermediate sentence.

Finally, based on the old top-n sentences at the previous moment of each time step, the new top-n best sentences with higher probabilities can be selected. In our research, the value of n was set to 3. The GPU (graphic processing unit) we used is Nvidia TITAN X (PASCAL).

Figure 3. Experiment sketch. The experiments are separated into two parts, including the training phase and test phase.

4.3.1. Results of Middle-Level Attribute Prediction

In the phase for predicting the middle-level attributes, the pre-trained VGG16 predictors are combined into an ensemble attribute predictor to predict the attribute labels of human and non-human objects. When the probability of the predicted attribute of each VGG16 predictor is not more than the predefined threshold, the generated attribute label is not used to retouch the intermediate image caption.

Figure 4 illustrates the results of the predicted middle-level attributes of humans. We show four representative results, including the girl with brown hair (i.e., Figure 4a), young female with blonde hair (i.e., Figure 4b), middle-aged man with black hair (i.e., Figure 4c), and senior man with brown hair (i.e., Figure 4d). From the results of the predicted mid-level attributes, we can observe that our ensemble attribute predictor can accurately predict the identification information of humans, which covers people of different ages, genders, and hair colors.

Furthermore, we also display the results of the predicted middle-level attributes of non-human objects (see Figure 5). The displayed images cover four different categories, including animal (i.e., Figure 5a), artifact (i.e., Figure 5b), natural object (i.e., Figure 5c), and plant (i.e., Figure 5d). As observed, our ensemble attribute predictor can accurately predict the color information and texture information of non-human objects. For instance, the color 'green' and the texture 'vegetation' of cabbage are predicted precisely (see Figure 5d). Nevertheless, the shape information of animals may be lost (see Figure 5a). The reason is that the predicted shape labels of animals, including long, round, rectangular, and square, all have low probabilities. Therefore, the predicted shape labels of animals would not be applied to retouch the intermediate image caption.

Figure 4. Predicted results of middle-level attributes of humans.

Figure 5. Predicted results of middle-level attributes of non-human objects.

4.3.2. Results of Retouched Image Captioning

We compared our proposed method with recent state-of-the-art methods for image captioning in the literature, such as soft attention [23], hard attention [23], Log-bilinear [21], ATT [18], F-G attention [42], GLA [25], and Topdown [28]. ‘OUR’ indicates our proposed method. Since the evaluation results of these methods on the MS COCO, Flickr8K, and Flickr30K datasets have been published publicly, we can compare with them directly. The evaluation results were generated by the coco-caption code.

Referring to Tables 6–8, the detailed comparison results of our method with previous works are shown, and we observe that our proposed method achieves impressive performance when using the predicted middle-level attribute labels to retouch the intermediate image caption generated by the language model. Our main assumption is that the global image feature provides an overview of the raw image, which is significant for the machine to understand the visual content in a rough manner. However, in order to accurately identify the details of an image, the local image features and the middle-level attributes need to be taken into account.

Table 6. Evaluation results of our method on the MS COCO dataset.

Model	BLEU-1	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
soft attention [23]	70.7	24.3	23.9	—	—	—
hard attention [23]	71.8	25.0	23.0	—	—	—
Log Bilinear [21]	70.8	24.3	20.0	—	—	—
ATT [18]	70.9	30.4	24.3	—	—	—
F-G Attention [42]	72.5	25.9	24.5	—	—	—
GLA [25]	72.5	31.2	24.9	53.3	96.4	—
Topdown [28]	74.5	33.4	26.1	54.4	105.4	19.2
OUR (before retouch)	77.6	36.5	27.8	56.9	114.9	20.9
OUR (after retouch)	79.5	37.2	27.9	57.1	118.2	21.6

Table 7. Evaluation results of our method on the Flickr8K dataset.

Model	BLEU-1	BLEU-4	METEOR	ROUGE-L	CIDEr
soft attention [23]	67.0	19.5	18.9	—	—
hard attention [23]	67.0	21.3	20.3	—	—
Log-Bilinear [21]	65.6	17.7	17.3	—	—
F-G Attention [42]	69.4	23.8	22.6	—	—
GLA [25]	57.2	14.8	16.9	36.2	41.9
OUR (before retouch)	70.7	24.6	23.4	39.1	51.2
OUR (after retouch)	72.5	25.1	24.2	40.7	53.4

Table 8. Evaluation results of our method on the Flickr30K dataset.

Model	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
soft attention [23]	66.7	19.9	18.5	—	—
hard attention [23]	66.9	19.9	18.5	—	—
Log-Bilinear [21]	60.0	17.1	16.9	—	—
F-G Attention [42]	68.4	21.4	20.0	—	—
GLA [25]	56.8	14.6	16.6	36.2	41.9
OUR (before retouch)	70.5	23.8	20.9	38.6	50.7
OUR (after retouch)	71.4	24.3	21.4	39.9	52.8

Moreover, we observe that the performance of our method for image caption generation achieves obvious improvement when using middle-level attribute labels to retouch the intermediate image description. Our conjecture is that the middle-level attributes can provide fine-grained identification information of objects, which avoids the loss of visual information. The main difference between middle-level attributes and object features is that the middle-level attribute information is more detailed and more accurate, which can highlight deep identification information of objects compared with object features.

A shortcoming of the attention mechanism is that it will weaken partial input information and generate one fused feature vector. If we try to inject the middle-level attributes and object features into the attention mechanism simultaneously, the role of the middle-level attributes will be weakened, since the middle-level attributes contain fine-grained identification information of objects.

Furthermore, we display some results of image captioning on the MS COCO dataset, Flickr8K dataset, and Flickr30K dataset, referring to Figures 6 and 7. Figure 6 illustrates the sampled MS COCO images and their corresponding descriptions of whether or not to use attributes. In Figure 6a–c, we can observe that the final image captions generated by our method accurately depict the details of the objects when using the middle-level attributes to retouch the intermediate image caption, including the age and gender of humans, and color and texture of non-human objects. Meanwhile, from Figure 7, which illustrates the sampled Flickr images and their descriptions, we can observe results similar to Figure 6.

However, in Figure 6d, the generated final image caption has a descriptive error when using the middle-level attribute to retouch the intermediate image caption. The word “boat” is retouched by “yellow” and “white”, simultaneously. The reason that the middle-level attribute negatively affected the generated caption is that there are two of the same object—“boat”—with different colors, respectively. Therefore, we need to be cautious about such issues in future work. The optional approach is to generate multiple image captions which can depict different aspects of a given image. Similarly, the youth male is mistakenly described as a middle-aged male in Figure 7d.

Finally, the results above show that our proposed method can solve the problem of missing middle-level attributes, to some degree, and can make the final image caption more accurate.

Figure 6. Results of image captioning on MS COCO. Red color indicates the positive example, and blue color indicates the negative example.

Figure 7. Results of image captioning on Flickr8K and Flickr30k. (a,b) show the results of some Flickr8K images, (c,d) show the results of some Flickr30K images. Red color indicates the positive example, and blue color indicates the negative example.

5. Conclusions

In this research, we tried to solve a problem in existing methods, i.e., that the encoder-decoder framework is used to generate the final image caption directly, which may ignore significant identification information of the middle-level attributes of the raw image, resulting in a generated image description of the object that is not accurate enough.

We propose an MLALR method for image caption generation, which can solve the aforementioned problem and make the final generated image caption more accurate. Our proposed MLALR method first uses the global and the local image features to generate the intermediate image caption. Then, it uses the middle-level descriptive attributes, which are predicted from the object regions of the raw image, to retouch the intermediate image caption according to the object index.

We validated our proposed method with several well-known evaluation metrics—BLEU, METEOR, ROUGE-L, CIDEr, and SPICE. The evaluation results and the final generated image descriptions show that our proposed method can correct descriptive errors in the intermediate image description to some degree and make the final generated image caption more accurate.

However, the middle-level attributes used in this research are describable aspects of visual appearance. We did not consider that whether people wear ornaments or not, such as glasses,

necklaces, watches, etc. Therefore, we will try to extract the information of ornaments from the original image to make the final image captioning more detailed in future work.

Author Contributions: Funding acquisition, T.J.; Investigation, Z.G., K.L. and Y.M.; Methodology, Z.G. and K.L.; Validation, K.L.; Writing—original draft, Y.M.; Writing—review & editing, X.Q. and T.J.

Funding: This work was partially funded by the National Key R&G Program of China under Grant [2018YFB1004600], partially funded by the National Key Research and Development Program of China under Grant [2016YFB10005000].

Acknowledgments: The authors are grateful for the constructive advice on the revision of the manuscript from the anonymous reviewers.

Conflicts of Interest: The authors declare that they have no competing interests.

References

1. Wei, Y.; Xia, W.; Huang, J.; Ni, B.; Dong, J.; Zhao, Y.; Yan, S. CNN: Single-label to Multi-label. *arXiv* **2014**, arXiv:1406.5726.
2. Mikolov, T.; Karafiát, M.; Burget, L.; Černocký, J.; Khudanpur, S. Recurrent neural network based language model. In Proceedings of the Eleventh Annual Conference of the International Speech Communication Association, Chiba, Japan, 26–30 September 2010; pp. 1045–1048.
3. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arxiv:1409.1556.
4. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. *arXiv* **2015**, arXiv:1512.03385.
5. Cho, K.; van Merriënboer, B.; Gülçehre, Ç.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *arXiv* **2014**, arXiv:1406.1078.
6. Schuster, M.; Paliwal, K.K. Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.* **1997**, *45*, 2673–2681. [[CrossRef](#)]
7. Hochreiter, S.; Schmidhuber, J. Long Short-term Memory. In *Neural Computation*; MIT Press: MA, USA, 1997; pp. 1735–1780.
8. Ren, S.; He, K.; Girshick, R.B.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal. *arXiv* **2015**, arXiv:1506.01497.
9. Papineni, K.; Roukos, S.; Ward, T. BLEU: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Philadelphia, PA, USA, 7–12 July 2002; pp. 311–318.
10. Denkowski, M.; Lavie, A. Meteor Universal: Language Specific Translation Evaluation for Any Target Language. In Proceedings of the Ninth Workshop on Statistical Machine Translation, Association for Computational Linguistics, Baltimore, MD, USA, 26–27 June 2014; pp. 376–380.
11. Lin, C.Y. ROUGE: A Package for Automatic Evaluation of Summaries. In Proceedings of the ACL-04 Workshop: Text Summarization Branches Out, Barcelona, Spain, 25–26 July 2004; pp. 74–81.
12. Vedantam, R.; Zitnick, C.L.; Parikh, D. CIDEr: Consensus-based Image Description Evaluation. *arXiv* **2014**, arXiv:1411.5726.
13. Anderson, P.; Fernando, B.; Johnson, M.; Gould, S. SPICE: Semantic Propositional Image Caption Evaluation. In *ECCV*; Springer: Cham, Switzerland, 2016; pp. 382–398.
14. Mao, J.; Xu, W.; Yang, Y.; Wang, J.; Yuille, A.L. Explain Images with Multimodal Recurrent Neural Networks. *arXiv* **2014**, arXiv:1410.1090.
15. Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and tell: A neural image caption generator. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3156–3164, 1063–6919.
16. Karpathy, A.; Li, F.-F. Deep Visual-Semantic Alignments for Generating Image Descriptions. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 664–676. [[CrossRef](#)] [[PubMed](#)]
17. Donahue, J.; Hendricks, L.A.; Rohrbach, M.; Venugopalan, S.; Guadarrama, S.; Saenko, K.; Darrell, T. Long-Term Recurrent Convolutional Networks for Visual Recognition and Description. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 677–691. [[CrossRef](#)] [[PubMed](#)]

18. You, Q.; Jin, H.; Wang, Z.; Fang, C.; Luo, J. Image Captioning with Semantic Attention. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 4651–4659.
19. Fu, K.; Jin, J.; Cui, R.; Sha, F.; Zhang, C. Aligning Where to See and What to Tell: Image Captioning with Region-Based Attention and Scene-Specific Contexts. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2321–2334. [[CrossRef](#)] [[PubMed](#)]
20. Kinghorn, P.; Zhang, L.; Shao, L. A region-based image caption generator with refined descriptions. *Neurocomputing* **2018**, *272*, 416–424. [[CrossRef](#)]
21. Kiros, R.; Salakhutdinov, R.; Zemel, R. Multimodal neural language models. In Proceedings of the International Conference on Machine Learning, Beijing, China, 21–26 June 2014; pp. 595–603.
22. Chen, M.; Ding, G.; Zhao, S.; Chen, H.; Liu, Q.; Han, J. Reference Based LSTM for Image Captioning. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; pp. 3981–3987.
23. Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.C.; Salakhutdinov, R.; Zemel, R.S.; Bengio, Y. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. *arXiv* **2015**, arXiv:1502.03044.
24. Lu, J.; Xiong, C.; Parikh, D.; Socher, R. Knowing When to Look: Adaptive Attention via a Visual Sentinel for Image Captioning. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 3242–3250.
25. Li, L.; Tang, S.; Zhang, Y.; Deng, L.; Tian, Q. GLA: Global-Local Attention for Image Description. *IEEE Trans. Multimed.* **2018**, *20*, 726–737. [[CrossRef](#)]
26. Pedersoli, M.; Lucas, T.; Schmid, C.; Verbeek, J. Areas of Attention for Image Captioning. *arXiv* **2016**, arXiv:1612.01033.
27. Liu, C.; Sun, F.; Wang, C.; Wang, F.; Yuille, A.L. MAT: A Multimodal Attentive Translator for Image Captioning. *arXiv* **2017**, arXiv:1702.05658.
28. Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; Zhang, L. Bottom-Up and Top-Down Attention for Image Captioning and VQA. *arXiv* **2017**, arXiv:1707.07998.
29. Wu, Q.; Shen, C.; Liu, L.; Dick, A.; van den Hengel, A. What Value Do Explicit High Level Concepts Have in Vision to Language Problems? In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition CVPR, Las Vegas, NV, USA, 27–30 June 2016; pp. 203–212.
30. Yao, T.; Pan, Y.; Li, Y.; Qiu, Z.; Mei, T. Boosting Image Captioning with Attributes. In Proceedings of the IEEE International Conference on Computer Vision ICCV, Venice, Italy, 22–29 October 2017; pp. 4904–4912.
31. Hu, C.; Miao, J.; Su, Z.; Shi, X.; Chen, Q.; Luo, X. Precision-Enhanced Image Attribute Prediction Model. In Proceedings of the 2017 IEEE Trustcom/BigDataSE/ICSS, Sydney, Australia, 1–4 August 2017; pp. 866–872.
32. Wang, L.; Chu, X.; Zhang, W.; Wei, Y.; Sun, W.; Wu, C. Social Image Captioning: Exploring Visual Attention and User Attention. *Sensors* **2018**, *18*, 646. [[CrossRef](#)] [[PubMed](#)]
33. Pan, Y.; Yao, T.; Li, H.; Mei, T. Video Captioning with Transferred Semantic Attributes. *arXiv* **2016**, arXiv:1611.07675.
34. Gao, L.; Guo, Z.; Zhang, H.; Xu, X.; Shen, H.T. Video Captioning with Attention-Based LSTM and Semantic Consistency. *IEEE Trans. Multimed.* **2017**, *19*, 2045–2055. [[CrossRef](#)]
35. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. In Proceedings of the 25th International Conference on Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–6 December 2012; Curran Associates, Inc.: Red Hook, NY, USA, 2012; pp. 1097–1105.
36. Kumar, N.; Berg, A.C.; Belhumeur, P.N.; Nayar, S.K. Attribute and simile classifiers for face verification. In Proceedings of the 12th International Conference on Computer Vision ICCV, Kyoto, Japan, 29 September–2 October 2009; pp. 365–372.
37. Russakovsky, O.; Li, F.F. Attribute Learning in Large-Scale Datasets. In Proceedings of the Trends and Topics in Computer Vision, Crete, Greece, 10–11 September 2012; pp. 1–14.
38. Rennie, S.J.; Marcheret, E.; Mroueh, Y.; Ross, J.; Goel, V. Self-critical Sequence Training for Image Captioning. *arXiv* **2016**, arXiv:1612.00563.
39. Lin, T.; Maire, M.; Belongie, S.J.; Bourdev, L.D.; Girshick, R.B.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. *arXiv* **2014**, arXiv:1405.0312.
40. Hodosh, M.; Young, P.; Hockenmaier, J. Framing Image Description As a Ranking Task: Data, Models and Evaluation Metrics. *J. Artif. Intell. Res.* **2013**, *47*, 853–899. [[CrossRef](#)]

41. Young, P.; Lai, A.; Hodosh, M.; Hockenmaier, J. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Trans. Assoc. Comput. Linguist.* **2014**, *2*, 67–78.
42. Chang, Y.S. Fine-grained attention for image caption generation. *Multimed. Tools Appl.* **2018**, *77*, 2959–2971. [[CrossRef](#)]

© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).