P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2023.29.03.034

INTERNATIONAL FOCUS ON IMAGE DESCRIPTION

Ms T Usha Durga, Assistant Professor, ushadurga.kotipally@gmail.com, Rishi UBR Women's College, Kukatpally, Hyderabad 500085

Abstract

In recent years, the task of automatically generating image description has attracted a lot of attention in the field of artificial intelligence. Benefitting from the development of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), many approaches based on the CNN-RNN framework have been proposed to solve this task and achieved remarkable process. However, there remain two problems to be tackled in that most of existing methods only use imagelevel representation. One problem is object missing that there may miss some important objects when generating the image description and the other is misprediction that it may recognize one object to a wrong category. In this paper, to address the two problems, we propose a new method called global-local attention (GLA) for generating image description. The proposed GLA model utilizes attention mechanism to integrate objectlevel features with image-level feature. Through this manner, our model can selectively pay attention to objects and context information concurrently. Therefore, our proposed GLA method can generate more relevant image description sentences, and achieves the state-of-the-art performance on the well-known Microsoft COCO caption dataset with several popular evaluation metrics — CIDEr, METEOR, ROUGE-L and BLEU-1,2,3,4.

I. INTRODUCTION

RECENTLY, image description has become more and more important and received much attention in the field of computer vision. It is a higher-level and more complicated cognitive task than fundamental perceptual tasks, such as image classification task [1], [2], [3], [4], object localization [5] and detection task [6], [7], [8], [9], image retrieval task [10], [11], [12] and so on. The goal of image description is to make computer understand what is happening in a given image and establish the bridge of image and natural language. So it not only involves computer vision technologies, but also requires natural language processing technologies. Automatically generating image caption is very meaningful for image understanding and can be used to many applications, such as, helping to improve the performance of multi-modal image retrieval, helping the blind navigation and so on.

For humans, it is very easy to vividly and accurately describe the content of an image. However, for computer, it remains a challenging task to automatically generate image caption in that the model need address the following key points in one model: (1) Object: the model should be able to accurately recognize objects as much as possible. (2) Relationship: the model needs to correctly identify what relationships among objects have. (3) Scene: the model can determine what scenes those objects are in. (4) Description: the model can express the above components by using

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2023.29.03.034

grammatically correct natural language sentences.

So far, many pioneering approaches have been proposed for solving the task of image description and achieved significant progress. These methods can be broadly divided into three categories according to the way of sentence generation [14]: (1) template-based method; (2) transfer-based method; (3) neural network-based method.

Baseline: A group of people standing on top of a snow covered slope.
Ours: A group of people on **skis** standing on a snow covered slope.

Baseline: A group of young men playing a game of soccer.
Ours: A boy and a **child** fly **kites** in a field.

Baseline: A person is holding a hot dog in a box.
Ours: A **picnic table** with a plate of food on it.

Baseline: A woman is eating a doughnut with a fork.
Ours: A woman holding a **cell phone** and smiling.

Fig. 1. Illustration of problems of the object missing (top left: missing "skis", bottom left: missing "picnic table") and misprediction (top right: mispredicting "kite" as "soccer", bottom right: mispredicting "cell phone" as "doughnut"). The baseline caption is generated by using the LRCN [13] method. Our caption is generated by utilizing the proposed GLA method.

Template-based method [15], [16], [17], [18] firstly recognizes that what kind of objects there are in an image by using object category classifiers, then recognizes attributes of these objects by using attribute classifiers and relationships among these objects by using relationship classifiers, and finally uses rigid sentence templates to form these components to complete sentences. The kind of method is intuitive, but needs complex data processing and is not enough flexible to generate meaningful sentence due to the limitation of sentence template.

Transfer-based method [19], [20], [21] firstly retrieves a similar image from annotated image caption dataset, and subsequently directly transfers the description of the retrieved image to the query image. This kind of method can generate more grammatically correct and natural sentences. Nevertheless, the generated sentences may not correctly express the visual content of query image due to the differences between query image and retrieved image.

Inspired by the recent advance of neural network in image recognition [1], [3], [22], [4], [6] and natural language processing [23], [24], [25], [26], [27], [28], [29], neural network-based method has been rapidly applied to automatically generate image/video caption [30], [31], [14], [32], [33], [13], [34], [35], [36], [37], [38], [39] and has been made great success. This kind of method is primarily based on the encoder-decoder framework which is proposed for sequenceto-sequence learning via utilizing Recurrent Neural Network (RNN) [28], [29]. These methods based on this framework firstly utilize an encoder to transform the input to a fixed length vector representation, and then use a decoder to decode the vector to the corresponding output.

Although the encoder-decoder framework is very efficient, there is a big limitation of it. Since the encoder needs to encode all the input sequence information to a fix length vector, the vector can not represent the

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2023.29.03.034

whole sequence information and the former information may be diluted by the later information along the sequence becoming longer. Therefore, the decoder can not obtain enough input information and lead to inaccurate decoding. Furthermore, an attention-based encoder-decoder framework [27], [40] is proposed to solve the problem by introducing attention weights. The attention weights make the encoder dynamically focus on more important parts of input sequence. Compared with the previous encoder-decoder framework, the attention-based encoder-decoder framework just encodes a subset of input sequence to a fixed length vector instead of encoding all input sequence. Thus, the decoder can selectively make the best of input information.

For image caption methods, the encoder is always based on Convolutional Neural Network (CNN) and the decoder is always based on RNN. These methods extract the image feature using a CNN encoder, and then utilize a language model, RNN or its variants, such as Long Short-Term Memory (LSTM) [41], Gated Recurrent Unit (GRU) [28], Bi-directional Recurrent Neural Network (BRNN) [42] et al., to decode the image feature to meaningful sentence. Compared with the previous two kind of methods, neural network-based method can generate more coherent and relevant sentences thanks to the ability of capturing dynamic temporal information of RNN and the good representation ability of CNN.

However, there are some limitations of most existing neural network-based methods due to the mere use of global representation in image-level. As shown in Fig. 1, these limitations can result in the problem of object missing and the problem of object misprediction. Some objects may not be recognized with only utilizing global features. As shown in the top left and bottom left pictures of Fig. 1, the “skis” and “picnic table” are missed. Besides, global features are extracted at a coarse level which may result in incorrect recognition and cause the problem of object misprediction during the process of description generation. As shown in the top right and bottom right pictures of Fig. 1, the “kite” is mispredicted as “soccer” and the “cell phone” is mispredicted as “doughnut”. In order to tackle these issues and obtain more accurate description, we take advantage of local features at object-level to address the problem of object missing. Moreover, inspired by attentionbased encoder-decoder framework, we integrate local features with global features via attention mechanism to reserve context information to address the problem of misprediction.

Therefore, we propose an global-local attention (GLA) method for image caption, which is our major contribution. The proposed GLA method can selectively focus on semantically more important regions at different time while keeping global context information through integrating local features with global features via attention mechanism. We compare our GLA with several baselines over the well-known MS COCO image caption dataset. Our results show that the proposed GLA method achieves the state-of-the-art performance

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2023.29.03.034

with different evaluation metrics. A preliminary version of this paper has appeared as a full paper in the Thirty-First AAAI Conference on Artificial Intelligence 2017 [43]. Compared to our previous conference paper, in this paper, we provide technical details about our previous image caption framework, present extended results with more comparisons and datasets, and offer an indepth analysis of the effect of different number of objects and the property of attention mechanism. Besides, we also apply our method to multi-model retrieval applications, image-tocaption retrieval and caption-to-image retrieval and the experiments show that our model achieves significant improvement compared with exiting methods.

The remainder of this paper is organized as follows. We first review the existing related works in Section II. Secondly, we present the overview of our framework in Section III. Thirdly, we elaborate on the details of the proposed GLA method in Section IV. Then, we evaluate our GLA on different datasets and analyze the performance in Section V. Finally, we draw our conclusive remarks with discussions for future work in Section VI.

II. RELATED WORK

In this section, we mainly review the related work from the following three aspects. First, we review recent image caption methods which are based on deep neural networks and the limitations of these methods. Second, we briefly introduce the attention-based image caption approaches. Finally, we introduce some works on object detection which are related with our proposed methods.

Deep neural network-based image caption. With the successful application of deep neural network in the task of image recognition and machine translation, the task of automatically generating image description also makes significant progress. There exist several effective methods [30], [31], [14], [32], [33], [13] based on deep neural networks.

As mentioned above, these approaches consider generating image description as a translation process. They directly translate an image to a sentence via utilizing the encoderdecoder framework [28] which is originally introduced in machine translation task. In general, this paradigm firstly uses a deep CNN as the encoder which encodes an image to a static representation, and then uses a RNN as the decoder which decodes this static representation to a meaningful sentence. The generated sentence should be grammatically correct and well describe the content of the image as much as possible.

To address the task of image description, in [32], Mao et al. propose a multimodal RNN (m-RNN) model which can also be used for image and sentence retrieval. The proposed mRNN additionally utilizes a multimodal layer to connect the language model and the CNN together. Similarly, Karpathy et al. [33] propose an alignment model via a multimodal embedding layer. This alignment model can align segments of sentence with the regions of the

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2023.29.03.034

corresponding image that they describe. Replacing the basic RNN by LSTM, a more powerful RNN model, Vinyals et al. [31] propose an endto-end model named NIC by combining deep CNN with LSTM for the problem. Furthermore, to address the problem of “drift away” or “lose track” of the image content, Jia et al. [14] propose gLSTM model, an alternative extension of LSTM. This model utilizes semantic information extracted from image as input along with the whole image to generate image descriptions. Donahue et al. [13] propose Long-term Recurrent Convolutional Network (LRCN) which combines convolutional layers and long-range temporal recursion for visual recognition and description.

However, as shown in Fig. 1, we notice that the above mentioned approaches may suffer from the problems of object missing and misprediction in that those methods encode the whole image to a static global feature vector. To overcome these problems, in this paper, we propose to integrate objectlevel features with image-level features for generating image caption via the widely used attention mechanism. In the next section, we brief some related works based on attention mechanism.

Attention mechanism in image caption and machine translation. Recently, attention mechanism has been widely used and proved to be important and effective in the field of natural language processing [27] and computer vision [30], [44], [37], [45]. In fact, the essence of attention mechanism is to assign positive weights to different parts to indicate the importance of these parts.

Attention mechanism is originally introduced in machine translation task [27]. In [27], Bahdanau et al. exploit BRNN with attention mechanism for machine translation. This approach is able to automatically search the part of the source sentence which is most relevant to a target word. Then, attention mechanism is introduced into image/video understanding task. Xu et al. [30] explore two kinds of attention mechanism for image caption, i.e, soft-attention and hardattention, and analyze how the attention mechanism works in the process of generating image caption via visualization manner. In [37], Yao et al. address video caption task through capturing global temporal structure among video frames with a temporal attention mechanism which is based on softalignment method. This temporal attention mechanism makes the model dynamically focus on key frames which are more relevant with the predicted word. ATT [44] proposes to utilize semantic concept to improve the performance. This method firstly obtains semantic concept proposals by utilizing different approaches, such as, k-NN, multi-label ranking and so on, and then integrates these concept proposals into one vector via the attention mechanism. The integrated vector is finally used to guide language model to generate description.

Different from soft/hard attention method [30] and ATT method [37], our proposed GLA method integrates local representation at object-level with global representation at

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2023.29.03.034

imagelevel through attention mechanism, whose aim is to address aforementioned problems of object missing and misprediction. Due to these methods which use only global frame-level features which cannot avoid problems of object missing and misprediction. Instead of considering semantic concepts or attributes used in ATT [44], we directly apply image visual feature with attention mechanism to image caption. RA [45] proposes a complicated pipeline to obtain important regions from selective search region proposals [46] and combines them with scene-specific contexts to generate image caption. Compared with ATT and RA methods, our GLA method is simpler and the performance is much better than RA method.

Object Detection. With the great success achieved by deep learning technology, object detection has also made significant progress. R-CNN [47] stands out as one of the notable landmarks in the process of object detection tasks. It takes advantage of high quality region proposals (selective search method [46]) and CNN features. This pipeline mainly contains four procedures: (1) Extracting region proposals which likely contain objects via region proposal methods; (2) Extracting CNN features of these region proposals via CNNs; (3) Classifying these proposals through classifier trained with CNN features; (4) Localizing these objects via bounding box regression methods.

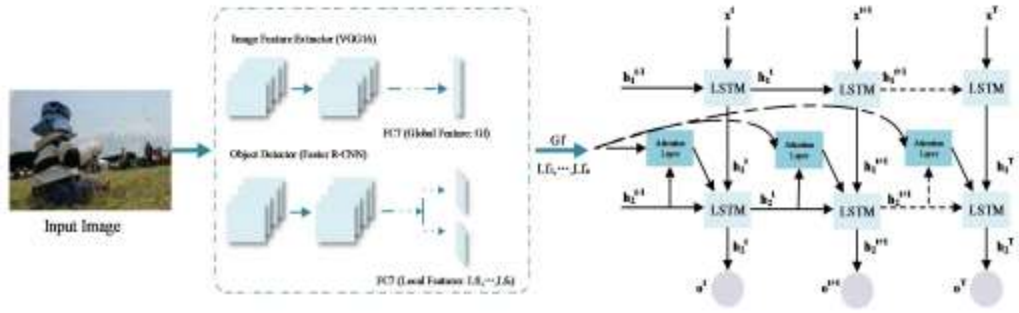


Fig. 2. Illustration of our proposed GLA image caption framework based on objects attention mechanism.

III. FRAMEWORK OVERVIEW

As illustrated in Fig. 2, we propose a novel image caption framework for automatically describing the content of an image based on object attention. The proposed framework mainly consists of three important processes.

First of all, we propose to make full use of global feature and local features for automatically generating image description. The global feature which contains the context information of an image is extracted from deep CNN. The local features which keep more precise object information are extracted through Faster R-CNN method.

Secondly, we integrate global feature with local features via attention mechanism. In this step, the attention mechanism can choose more important objects and assign them greater weights. This makes the model selectively focus on some important entities as well as considering the image's context information.

Finally, we utilize a language model to generate a description sentence with the dynamic integrated feature. In this paper, we use a stacked two-layer LSTM as the language model which can effectively capture the long-term sequence information.

IV. GLOBAL-LOCAL ATTENTION MODEL

In this section, we introduce our proposed GLA method for automatically generating image description in detail.

A. Global and Local Features Extraction

To make the computer understand an image, we firstly need to extract image features. There are mainly two kinds of features, global feature and local feature. Global feature reserves the comprehensive information of an image which is a kind of coarse-grained feature. Local feature usually contains the fine-grained information of objects. In our model, we explore the effect of the two types of feature in the task of image caption. Benefitting from the good performance of CNNs in the task of image classification and object detection, we consider extracting global feature with VGG16 model [3] and local features with Faster R-CNN [6].

For global feature denoted as Gf, we represent it as a 4096- dimension vector, the fc7 layer feature extracted from VGG16 net. This VGG16 net is trained on ImageNet classification dataset. For local features, we select k objects in an image and their features are denoted as {Lf1, ..., Lfk} which is a set of 4096-dimension vectors extracted from fc7 layer for each object bounding box. These k objects are the top-k objects which are chosen according to the classification confidence scores obtained from Faster R-CNN. The Faster R-CNN model is pre-trained on ImageNet classification dataset and then finetuned on the MS COCO detection dataset.

Therefore, each image is finally represented by a set of 4096-dimension vectors I={Gf,Lf1,...,Lfk}. In our experiments, we set k to 10 since the number of object contained in an image is usually below 10 and the experiments verify the effectiveness of this conjecture.

B. Global-local Attention Mechanism

Obviously, for a generated sentence, each word usually corresponds to different entity of the image. Therefore, when generating an image description, the model should be able to focus on different entities in each time step. For this purpose, we adopt attention mechanism to integrate the local features with the global features according to the following Eq. 1:

$$\Psi^{(t)}(I) = \alpha_0^{(t)} Gf + \sum_{i=1}^{k} \alpha_i^{(t)} Lf_i, \qquad (1)$$

where Ψt (I) is the final integrated image representation at time t. α (t) i denotes the attention weight of each feature i at time t and satisfies the constraint ∑k i=0 α (t)=1.

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2023.29.03.034

During sentence generation procedure, this attention mechanism dynamically weights the importance of each entity by assigning it with one positive weight α (t) i . Through this manner, the model can selectively focus on the salient objects and keep the scene information at the same time.

The weight α (t) i measures the importance of each feature i and relevancy to the history information at each time step t. Thus, it can be determined by the previous output h t−1 and the feature fi ∈ {Gf, Lf0, ..., Lfn} via the following equations:

$$\beta_i^{(t)} = w^T \varphi(W_h h^{(t-1)} + W_o f_i + b), \qquad (2)$$

$$\alpha_i^{(t)} = \frac{\beta_i^{(t)}}{\sum_{j=0}^{n} \beta_j^{(t)}}, \qquad (3)$$

where β (t) i denotes the relevance score of each feature fi with the generated word.

The weight α (t) i is computed through normalizing the relevance score β (t) i with softmax regression. h (t−1) contains the history information which is output by the previous hidden state of RNN. W, Wh, Wo and b are the parameters shared by all features at all time steps. φ is the element-wise Hyperbolic Tangent activation function which is defined as the Eq. 4:

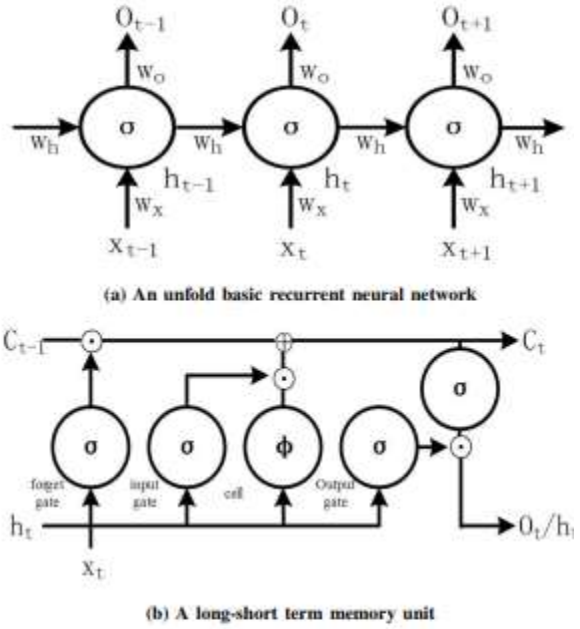
$$tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}. \qquad (4)$$

C. Image Description Generation

For language model, we choose to use a stacked twolayer LSTM to generate sentence due to the outstanding performance of RNN in the task of neural machine translation [27]. LSTM is one kind of advanced RNN which can effectively capture dynamic long-term temporal information with a distinctive unit.

We first introduce the basic RNN structure as shown in the top row of Fig. 3. A basic RNN essentially maintains a hidden state ht in each time step t and overwrites this hidden state with input xt and the previous hidden state ht−1. Usually, we use Back-Propagation Through Time (BPTT) algorithm to train RNNs. However, the gradients of RNNs tend to vanish with this BPTT algorithm due to the chain rule of derivative. Thus, the basic RNNs suffer from the problem of long-range dependency caused by vanshing and exploding gradients in the training process. Then, to solve this issue, the LSTM is designed to combat them via a gating mechanism. The bottom figure of Fig. 3 shows the structure of a LSTM unit. A LSTM unit mainly consists of four part, a memory cell, input gate, output gate and forget gate. The forget gate determines how much the history information will be retained. The input gate decides what new information should be reserved and the output gate determines how much of the hidden state should be exposed to next process. All the three gates receive the input information xt (word or input image) and the previous hidden state ht−1 and input these information into activation functions. The memory cell also has the same input with these gates, but with different activation functions.

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2023.29.03.034



(a) An unfold basic recurrent neural network

(b) A long-short term memory unit

Fig. 3. Illustration of an un-fold basic RNN and a LSTM ui represents logic sigmoid function. ϕ represents hyperbolic tangent fur ⊙ represents the multiplication operation and ⊕ represents sum opera

In our model, the LSTM is trained for predicting each word st with which the model composes a complete description sentence. To predict each word st, the model should know the image visual information I as well as the predicted words {s0, s1, ..., st−1}. This process can be defined by predicting the word probability on the condition that I and {s0, s1, ..., st−1}, i.e. p(st|I, s0, s1, ..., st1), are known. Specifically, we only input the image information into the second LSTM layer. The detailed operations of the first LSTM layer is listed as following Eq. 5:

$$
\begin{aligned}
&x^t = w_x s_t\\
&i_1^t = \sigma(w_{is}^1 x^t + w_{ih}^1 h_1^{(t-1)} + b_i^1)\\
&f_1^t = \sigma(w_{fs}^1 x^t + w_{fh}^1 h_1^{(t-1)} + b_f^1)\\
&o_1^t = \sigma(w_{os}^1 x^t + + w_{oh}^1 h_1^{(t-1)} + b_o^1)\\
&c_1^t = f_1^t \otimes c_1^{t-1} \oplus i_1^t \otimes \phi(w_{cs}^1 x_1^t + w_{ch}^1 h_1^{t-1})\\
&h_1^t = o_1^t \otimes c_1^t
\end{aligned}
\tag{5}
$$

where each word is represented as an one-hot vector st whose dimension is equal to the vocabulary size. The second LSTM layer is computed as the following Eq. 6:

$$
\begin{aligned}
&I^t = \Psi^t(I)\\
&i_2^t = \sigma(w_{iI}^2 I^t + w_{ih}^{12} h_1^{(t-1)} + w_{ih}^2 h_2^{(t-1)} + + b_i^2)\\
&f_2^t = \sigma(w_{fI}^2 I^t + w_{fh}^{12} h_1^{(t-1)} + w_{fh}^2 h_2^{(t-1)} + b_f^2)\\
&o_2^t = \sigma(w_{oI}^2 I^t + w_{oh}^{12} h_1^{(t-1)} + w_{oh}^2 h_2^{(t-1)} + b_o^2)\\
&c_2^t = f_2^t \otimes c_2^{t-1} \oplus i_2^t \otimes \phi(w_{cs}^2 x_2^t + w_{ch}^2 h_2^{t-1})\\
&h_2^t = o_2^t \otimes c_2^t\\
&P(s_t|I, s_0, s_1, ..., s_{t-1}) = Softmax(w_p h_2^t)
\end{aligned}
\tag{6}
$$

where w∗ and b∗ are parameters that the model should learn and shared by all time steps in our model.

Our final goal is to obtain an optimal description sentence of the input image. This can be generated by computing the probability of all the observing words. As the Eq. 7 shows, the probability of a sentence is the product of probability of each word given an image and all the words before current time step.

$$
p(s_0, s_1, ..., s_m) = \prod_{i=0}^{m} p(s_i|I, s_0, ..., s_{i-1}) \tag{7}
$$

Thus, the loss function final can be defined as Eq. 8. It is the sum of the probability log likelihood of each word.

$$
L(I, S) = \sum_{i=0}^{m} log(p(s_t|I, s_0, ..., s_i)) \tag{8}
$$

We use stochastic gradient descent to optimize the above objective function over the whole training set. The configuration of hyperparameters is introduced in the next section. In test procedure, there are two

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2023.29.03.034

strategies for generating description sentence of a given image. The first strategy is essentially a greedy method. At each time step, we sample the next word which has the maximum probability from the probability distribution until we sample the end sign word or the sentence reaches to the maximum length. The second strategy is beam search method. At each time step, we select the top-m best sentences and then sample new best top-m sentences based on the previous top-m sentences. In this paper, we sample sentences by using these two strategies to evaluate our method. Particularly, in beam search strategy, we can obtain the best sentence when the value of m is set to 3.

V. EXPERIMENTS

In this section, we first introduce our implement details. Second, we brief several popular metrics used in our experiments. Third, we introduce the datasets used for validating our proposed method. Finally, we introduce our experiments configuration and analyze the results in detail.

A. Experiment Setup

LSTM Hyperparameter Configuration. We implement our global-local attention model based on LRCN framework [13], an open-source implementation of RNN. For hyperparameters of training the LSTM, we set the momentum of stochastic gradient descent to 0.9. The learning rate is initially set to 0.01 and then is decreased every 20000 iterations. The final training iteration is set to 120, 000. The clip gradient used in LSTM is set to 10. The node number of LSTM's hidden layer is set to 1000.

Selection of Top-k Object. In our method, one of the important part is how to select the top-k objects. In order to obtain better objects, in our experiments, we firstly use Faster R-CNN to obtain some region proposals and then use NonMaximum Suppress (NMS) algorithm to pick out better region proposals. After executing NMS, we sort the remained region proposals according to these regions confidences in descending order. Finally, we choose the top-k region proposals as the final object bounding boxes. If there are no more than k region proposals after NMS, we randomly sample some region proposals from the remainders to make enough k objects.

Finetuning of Faster R-CNN. In order to extract better object features, we firstly finetune Faster R-CNN model over MS COCO object detection dataset. The base model is pretrained on ImageNet object detection dataset. MS COCO object detection task shares the same images with image caption task. Therefore, we keep the same splits with the image caption dateset for training which will be introduced in the following introduction of dataset. The finetuning process is almost the same with the pre-training process [6]. The initial learning rate for finetuning is set to 0.001. The momentum of stochastic gradient descent is set to 0.9 and the weight decay is set to 0.0005.

B. Evaluation Metrics

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2023.29.03.034

There have been proposed various criteria for evaluating the generated sentences. However, how to evaluate the quality of the descriptions remains to be a challenging problem. Therefore, in order to accurately verify the performance of our model, we use multiple metrics to evaluate the GLA method, i.e. METEOR [51], ROUGE-L [52], CIDEr [53] and BLEU1,2,3,4 [54].

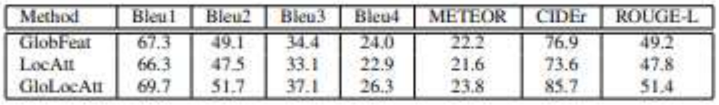
BLEU is the most popular metric for evaluating the generated sentence in machine translation task. This metric is only based on the n-gram precision. In our experiments, we validate the performance with 1,2,3,4-gram individually. To fix some of the problems of BLEU metric, METEOR is designed which is based on the harmonic mean of unigram precision and recall. Different with the BLEU metric, the METEOR seeks correlation at the corpus level. ROUGE-L is designed for measuring the common subsequence with maximum length between target sentence and source sentence. CIDEr is used to evaluate the generated descriptions using human consensus.

C. Datasets

In order to prove the effectiveness of our proposed method, we conduct several experiments on the following three popular datasets.

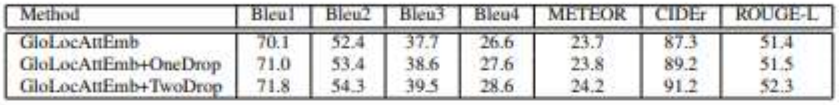
Flickr 8k Dataset. The first dataset is the popular Flickr 8k dataset [55] which consists of 8,000 images in total and each image is annotated with 5 English sentences. In order to fairly compare with existing methods, we keep 6,000 images for training, 1,000 images for validation and 1,000 images for testing which is the same splits with the existing methods.

TABLE I
MPARISON EXPERIMENTS FOR EXPLORING THE EFFECT OF IMAGE DESCRIPTION WITH GLOBAL FEATURE, LOCAL FEATURES AND FUSION OF THE TWO FEATURES OVER MS COCO DATASET RESPECTIVELY .



Method	Bleu1	Bleu2	Bleu3	Bleu4	METEOR	CIDEr	ROUGE-L
GlobFeat	67.3	49.1	34.4	24.0	22.2	76.9	49.2
LocAtt	66.3	47.5	33.1	22.9	21.6	73.6	47.8
GloLocAtt	69.7	51.7	37.1	26.3	23.8	85.7	51.4

TABLE II
COMPARISON EXPERIMENTS FOR EXPLORING THE EFFECT OF OUR PROPOSED IMAGE DESCRIPTION WITH DROPOUT MECHANISM OVER MS COCO DATASET.



Method	Bleu1	Bleu2	Bleu3	Bleu4	METEOR	CIDEr	ROUGE-L
GloLocAttEmb	70.1	52.4	37.7	26.6	23.7	87.3	51.4
GloLocAttEmb+OneDrop	71.0	53.4	38.6	27.6	23.8	89.2	51.5
GloLocAttEmb+TwoDrop	71.8	54.3	39.5	28.6	24.2	91.2	52.3

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2023.29.03.034

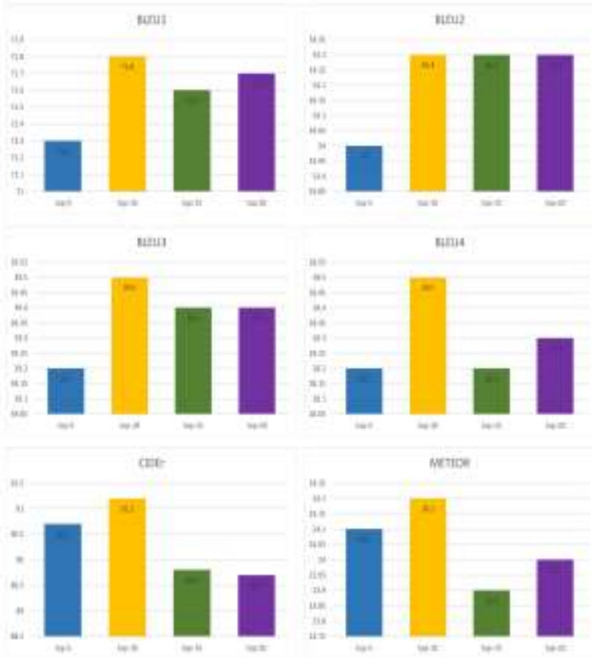


Fig. 4. Illustration of the performance with different top-*k* objects on six well-known metrics — BLEU1, BLEU2, BLEU3, BLEU4, CIDEr and METEOR The abscissa represents the top-*k* type. The ordinate denotes the accuracy of image caption. Each subfigure represents the performance of different *k* on one of these metrics. Each color denotes the performance of different metric on one kind of *k*. We explore the performance by using four different *k* — 5,10,15,20.

D. Image Caption Experiments and Results

The determination of the number of top-k objects. How many objects to choose is also important for generating description. In order to obtain proper number of objects, in the first section, we analyze the effect of different number of objects for image caption. Here, we choose 5, 10, 15 and 20 objects of each image to verify the performance respectively.

Fig. 4 illustrates the results of MS COCO image caption dataset with different top-k objects via several metrics. Each subfigure shows the accuracy of different k with the same metric and each color shows the accuracy of the same k with different metrics. These experiments use greedy strategy to generate description sentences.

From this figure, we observe that it achieves best performance when we choose top-10 objects in each image. This may because that the average number of objects in the dataset is blew 10. When the number of objects is smaller, there may miss some important object which causes incomplete description. On the contrary, when the number of objects is larger, there may have some non-objects to influence the prediction. According to this experiment, we finally choose that k is equal to 10 to conduct our following experiments.

Evaluation on Image-level and Object-level Information for Image Description on MS COCO Dataset. In the task of image recognition, global features and local features have been proved to be important. Therefore, in this section, we first conduct three experiments to explore the effect of local features, global feature and the fusion form of these two features for automatically describing image. The detail configuration of the three experiments are listed as follows:

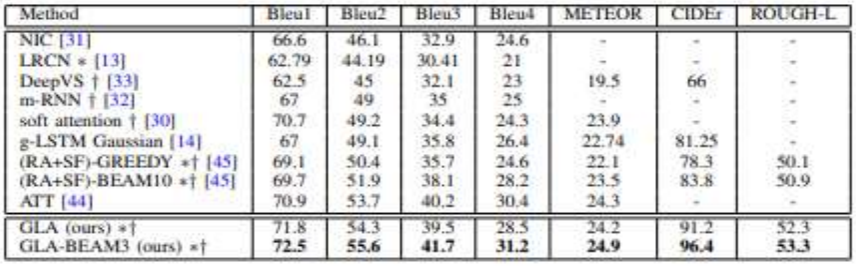
- GloFeat: Only employing the image-level feature Gf which is extracted from VGG16.

- LocAtt: Only employing object-level features {Lf1, ..., Lfk} which are extracted from Faster R-CNN, and then integrating them with attention mechanism.

- GloLocAtt: Employing these two kinds of features, and then integrating them with attention mechanism to generate description.

TABLE III

COMPARISON WITH SEVERAL STATE-OF-THE-ART MODELS IN TERMS OF BLEU-1,2,3,4, METEOR, CIDER, ROUGE-L AND METEOR OVER MS COCO DATASET. - INDICATES UNKNOWN SCORES. † INDICATES THAT THE MODEL HAS THE SAME DECODER WITH OURS, THAT IS, THE SAME CNN MODEL FOR IMAGE REPRESENTATION. * INDICATE THAT THE MODEL HAS THE SAME ENCODER - THE LANGUAGE MODEL FOR GENERATING SENTENCE DESCRIPTION WITH OURS.



Method	Bleu1	Bleu2	Bleu3	Bleu4	METEOR	CIDEr	ROUGH-L
NIC [31]	66.6	46.1	32.9	24.6	-	-	-
LRCN * [13]	62.79	44.19	30.41	21	-	-	-
DeepVS † [33]	62.5	45	32.1	23	19.5	66	-
m-RNN † [32]	67	49	35	25	-	-	-
soft attention † [30]	70.7	49.2	34.4	24.3	23.9	-	-
g-LSTM Gaussian [14]	67	49.1	35.8	26.4	22.74	81.25	-
(RA+SF)-GREEDY *† [45]	69.1	50.4	35.7	24.6	22.1	78.3	50.1
(RA+SF)-BEAM10 *† [45]	69.7	51.9	38.1	28.2	23.5	83.8	50.9
ATT [44]	70.9	53.7	40.2	30.4	24.3	-	-
GLA (ours) *†	71.8	54.3	39.5	28.5	24.2	91.2	52.3
GLA-BEAM3 (ours) *†	**72.5**	**55.6**	**41.7**	**31.2**	**24.9**	**96.4**	**53.3**

Baseline: A group of birds are standing in the water.
Ours: A group of **ducks swimming** in a lake.

Baseline: A group of people playing a game of frisbee.
Ours: A group of people playing **soccer** on a field.

Baseline: A man is riding a elephant with a trunk.
Ours: A man is **standing** next to a large elephant.

Baseline: A bear is walking through a tree in the woods.
Ours: A bear is **sitting** on a tree branch.

Baseline: A sheep standing in a field with a green and white sheep.
Ours: A sheep standing in a field with a **sheep**.

Baseline: A brown cow standing in a field of grass.
Ours: A group of cows **standing next to each other**.

Baseline: A man riding a wave on top of a surfboard.
Ours: A man is standing on a **mountain** overlooking a mountain.

Baseline: A table with a plate of food and a cup of coffee.
Ours: A table with a bunch of **fruit** and a glass of wine.

Fig. 5. Illustrating the comparison of some sampled image caption results with existing method results. The baseline caption is generated by LRCN method which only exploits image-level features. The second caption is generated with our proposed GLA method.

- GloLocAttEmb: Employing integrated feature and adding one linear transform layer to reduce the feature dimension.

- GloLocAttEmb+OneDrop: On the basis of the first experiment, adding one dropout layer after the second LSTM layer.

- GloLocAttEmb+TwoDrop: On the basis of the second experiment, adding one dropout layer after the first LSTM layer.

Tab. II shows the comparison results of the three experiments. Compared with GloLocAtt experiment, the performance of GloLocAttEmb experiment has some improvement in that the linear transform layer makes the feature more distinctive. Furthermore, the dropout experiments improve the performance since dropout can reduce overfitting in some degree. The performance is better by adding two dropout layers.

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2023.29.03.034

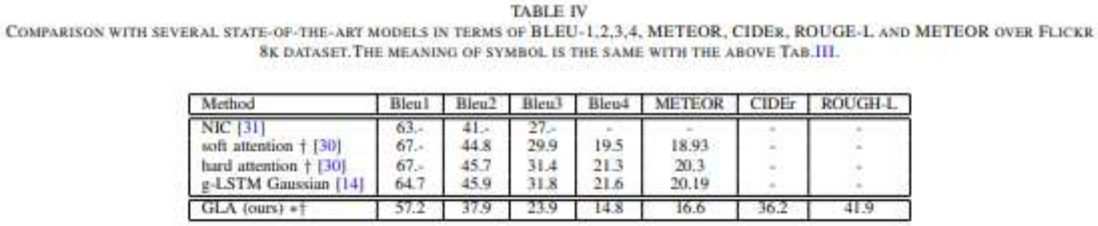


TABLE IV

COMPARISON WITH SEVERAL STATE-OF-THE-ART MODELS IN TERMS OF BLEU-1,2,3,4, METEOR, CIDER, ROUGE-L AND METEOR OVER FLICKR 8K DATASET.THE MEANING OF SYMBOL IS THE SAME WITH THE ABOVE TAB.III.

Method	Bleu1	Bleu2	Bleu3	Bleu4	METEOR	CIDEr	ROUGH-L
NIC [31]	63.-	41.-	27.-	-	-	-	-
soft attention † [30]	67.-	44.8	29.9	19.5	18.93	-	-
hard attention † [30]	67.-	45.7	31.4	21.3	20.3	-	-
g-LSTM Gaussian [14]	64.7	45.9	31.8	21.6	20.19	-	-
GLA (ours) *†	57.2	37.9	23.9	14.8	16.6	36.2	41.9

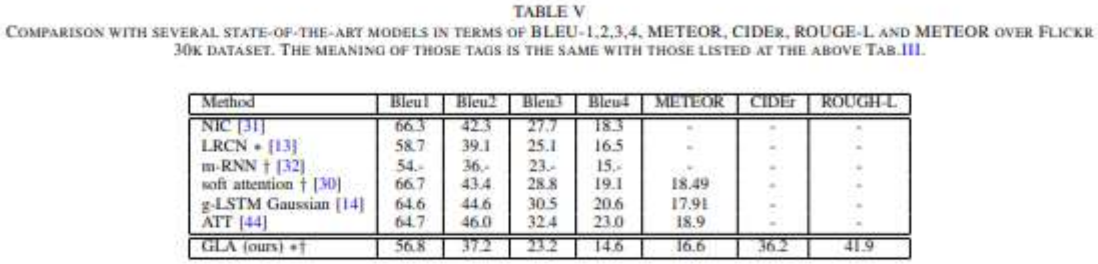


TABLE V

COMPARISON WITH SEVERAL STATE-OF-THE-ART MODELS IN TERMS OF BLEU-1,2,3,4, METEOR, CIDER, ROUGE-L AND METEOR OVER FLICKR 30K DATASET. THE MEANING OF THOSE TAGS IS THE SAME WITH THOSE LISTED AT THE ABOVE TAB.III.

Method	Bleu1	Bleu2	Bleu3	Bleu4	METEOR	CIDEr	ROUGH-L
NIC [31]	66.3	42.3	27.7	18.3	-	-	-
LRCN * [13]	58.7	39.1	25.1	16.5	-	-	-
m-RNN † [32]	54.-	36.-	23.-	15.-	-	-	-
soft attention † [30]	66.7	43.4	28.8	19.1	18.49	-	-
g-LSTM Gaussian [14]	64.6	44.6	30.5	20.6	17.91	-	-
ATT [44]	64.7	46.0	32.4	23.0	18.9	-	-
GLA (ours) *†	56.8	37.2	23.2	14.6	16.6	36.2	41.9

Comparison with the State-of-the-art Methods. In the final, we compare our GLA method with several state-ofthe-art methods, such as, g-LSTM [14], NIC [31], m-RNN [32], LRCN [13], DeepVS [33], soft/hard attention [30] and ATT [44], on all the above three datasets. In the following tables, we use "GLA" to represent our proposed method whose configuration is the same with "GloLocAttEmb+TwoDrop" experiment. We sample sentences with greedy strategy in all the above experiments. Since beam search can approximately obtain sentence with the maximum probability, we try beam search in our experiment. When the m is set to 3, we can get the best performance. The best result is denoted as "GLA+BEAM3".

Among the above state-of-the-art methods, there are some differences with each other. The first difference is the encoder used for representing images. LRCN exploits AlexNet to extract image-level features. ATT, g-LSTM, and NIC use GoogLeNet to extract image-level features. DeepVS, m-RNN and soft/hard attention use the same encoder with our to obtain image-level representation. To make fair comparison, we first compare our approach with methods which use the same encoder. The results show that the performance is improved significantly on different metrics.

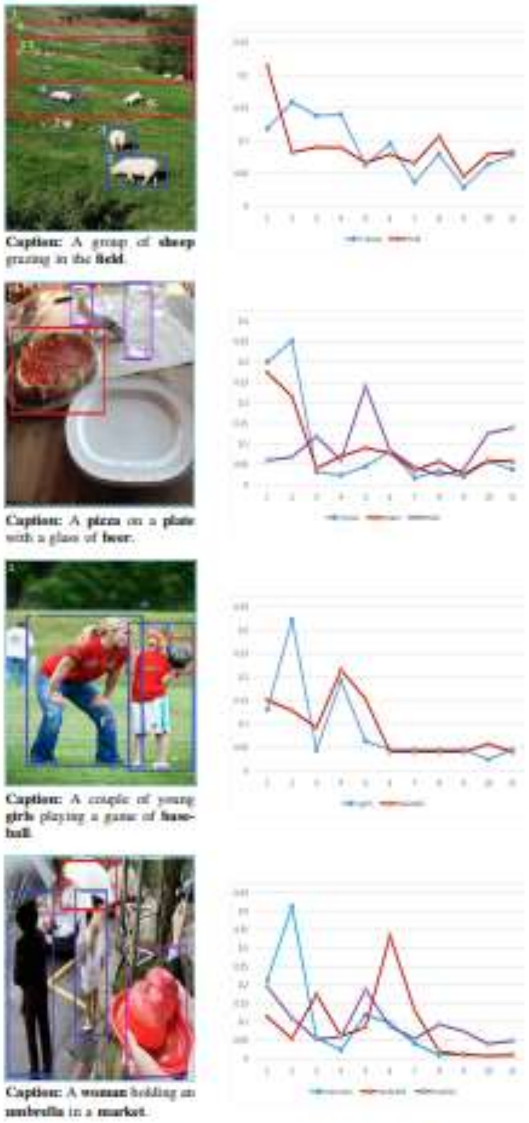
The second difference is the decoder used for generating sentence. Same as our method, LRCN employs a stacked twolayer LSTM to generate image description. NIC, g-LSTM, and soft/hard attention use one layer LSTM network as language model. ATT and m-RNN use the basic RNN as decoder. DeepVS employs the BRNN to generate image caption. Here, compared with the same decoder, the performance of our GLA method also has better performance.

We show our results over MS COCO dataset in Tab. III. By comparing our GLA model with the existing methods on MS COCO dataset, we note that our approach achieves best performance in that our method can capture more detailed object information. From Fig. 5 which illustrates the sampled

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2023.29.03.034

images and their descriptions generated by LRCN [13] model and GLA model, we can see that GLA model can generate more relevant descriptions. The results show our method can solve the problems of objects missing and misprediction in some degree.

Then we show our results over Flickr 8k dataset in Tab. IV and over Flickr 30k dataset in Tab. V. However, from those two tables, we can find that the performance is decreased compared with the existing method. There are two reasons to account for this phenomenon. On the one hand, our model can not jointly train the feature extractor and the language model. On the other hand, the flickr datasets lack object information which causes that we can not finetune the Faster R-CNN on flickr dataset. This two reasons cause that our model can not be well generalized for flickr datasets.



Caption: A group of **sheep** grazing in the **field**.

Caption: A **pizza** on a **plate** with a glass of **beer**.

Caption: A couple of young **girls** playing a game of **baseball**.

Caption: A **woman** holding an **umbrella** in a **market**.

Fig. 6. Illustration of the attention weights of objects. In each row, th left subfigure shows the sampled image and its description generated by ou method. Each box corresponds to one object and the soft blue box denote the whole image. These boxes in each image are a subset of the top-k object which are generated by Faster R-CNN. The right subfigure is a line cha which shows the weight of each entity when generating a certain word.

Through these subfigures, we can clearly observe that the attention mechanism assigns the objects with greater weights which are relevant to the predicted word. For example, in the first row, when generating the word “sheep”, the entities with number two, three and four have greater weights. This proves that our model can dynamically focus on more relevant objects when describing the context of an image.

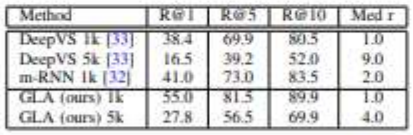
E. Retrieval Experiments and Results over MS COCO Dataset

The improvement of image caption can promote the multimodal retrieval task. In addition to our previous work [43], we also conduct two kinds of retrieval experiments, image-tocaption retrieval task and caption-to-image retrieval task, for further verifying the performance of proposed method over MS COCO dataset. Here, we use two configurations of test set. One test is the same with caption task which consists of 5,000 images. The other consists of 1,000 images sampled from the original 5,000 images. We also use R@K (k=1, 5, 10) and Med r for evaluation of the two tasks.

R@K is the recall rate with top K return results. The higher the R@K is, the better the performance of retrieval is. Med r is the median rank of the first retrieved groundtruth results. In contrast with R@K, the lower the Med r is, the better the performance of retrieval is.

Tab. VI shows the comparison results of our method with previous methods. Image-to-caption retrieval is to find the most relevant caption given an image. The results show that our method outperforms these methods over two test datasets. For 1k validation dataset, our method improves over 15% of R@1. For 5k validation dataset, our method improves about 10% of R@1.

TABLE VI
COMPARISON WITH EXISTING MODELS IN TERMS OF R@1, R@5, R@10 AND MED R OVER MS COCO DATASET. 1K DENOTES THAT WE SAMPLE 1K IMAGES FROM VALIDATION DATASET FOR TESTING THE PERFORMANCE. 5K DENOTES THAT WE SAMPLE 5K IMAGES FROM VALIDATION DATASET.



Method	R@1	R@5	R@10	Med r
DeepVS 1k [33]	38.4	69.9	80.5	1.0
DeepVS 5k [33]	16.5	39.2	52.0	9.0
m-RNN 1k [32]	41.0	73.0	83.5	2.0
GLA (ours) 1k	55.0	81.5	89.9	1.0
GLA (ours) 5k	27.8	56.5	69.9	4.0

Tab. VII shows the results of caption-to-image retrieval. Caption-to-image retrieval is to find the most relevant image for a given image description. Our model also performs well in this task. Compared with the existing methods, our method improves over 10% of R@1 in 1k test dataset and improves about 8% of R@1 in 5k test dataset.

VI. CONCLUSION

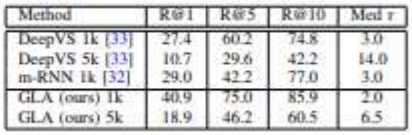
We propose a novel method for generating image description which achieves better performance on the MS COCO benchmark compared with the previous approaches. Our proposed method combines image-level and object-level information via attention mechanism. Compared with state-of-theart approaches, our method not only can capture the global information, but also obtain local object information. We prove this by doing quantitative analysis of the attention weights along with the sentence generation procedure. Consequently, our method generates more relevant and coherent natural language sentences which can describe the context of images.

However, our current GLA method is not end-to-end which can not jointly train the CNN part and the language model. Thus, we will try how to integrate the object detector

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2023.29.03.034

with image feature extractor so as to train and test our model endto-end.

TABLE VII
COMPARISON WITH EXISTING MODELS IN TERMS OF R@1, R@5, R@10 AND MED R OVER MS COCO DATASET. 1K DENOTES THAT WE SAMPLE 1K IMAGES FROM VALIDATION DATASET FOR TESTING THE PERFORMANCE. 5K DENOTES THAT WE SAMPLE 5K IMAGES FROM VALIDATION DATASET.



Method	R@1	R@5	R@10	Med r
DeepVS 1k [33]	27.4	60.2	74.8	3.0
DeepVS 5k [33]	10.7	29.6	42.2	14.0
m-RNN 1k [32]	29.0	42.2	77.0	3.0
GLA (ours) 1k	40.9	75.0	85.9	2.0
GLA (ours) 5k	18.9	46.2	60.5	6.5

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in Advances in neural information processing systems, 2012, pp. 1097–1105.

[2] S. Tang, Y.-T. Zheng, Y. Wang, and T.-S. Chua, “Sparse ensemble learning for concept detection,” IEEE Transactions on Multimedia, vol. 14, no. 1, pp. 43–54, 2012.

[3] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” CoRR, vol. abs/1409.1556, 2014.

[4] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proceedings of the IEEE International Conference on Computer Vision, 2016.

[5] S. Tang, Y. Li, L. Deng, and Y.-D. Zhang, “Object localization based on proposal fusion,” IEEE Transactions on Multimedia, vol. 19, no. 9, pp. 2015–2116, 2017.

[6] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” in Neural Information Processing Systems, 2015.

[7] C. Szegedy, D. Erhan, and A. T. Toshev, “Object detection using deep neural networks,” Mar. 1 2016, uS Patent 9,275,308.

[8] W. Ouyang, X. Wang, X. Zeng, S. Qiu, P. Luo, Y. Tian, H. Li, S. Yang, Z. Wang, C.-C. Loy et al., “Deepid-net: Deformable deep convolutional neural networks for object detection,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2403–2412.

[9] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 779–788.

[10] C. Kang, S. Xiang, S. Liao, C. Xu, and C. Pan, “Learning consistent feature representation for cross-modal multimedia retrieval,” IEEE Transactions on Multimedia, vol. 17, no. 3, pp. 370–381, 2015.

[11] S. Bu, Z. Liu, J. Han, J. Wu, and R. Ji, “Learning high-level feature by deep belief networks for 3-d model retrieval and recognition,” IEEE Transactions on Multimedia, vol. 16, no. 8, pp. 2154–2167, 2014.

[12] F. Radenovic, G. Tolias, and O. Chum, “Cnn image retrieval learns ´ from bow:

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2023.29.03.034

Unsupervised fine-tuning with hard examples," in European Conference on Computer Vision. Springer, 2016, pp. 3–20.