BiTMulV: Bidirectional-Decoding Based Transformer with Multi-view Visual Representation

Qiankun Yu[1,2], XueKui Wang[3], Dong Wang[4], Xu Chu[1,2], Bing Liu[4(✉)], and Peng Liu[1(✉)]

[1] National Joint Engineering Laboratory of Internet Applied Technology of Mines, China University of Mining and Technology, Xuzhou 221008, Jiangsu, China
yqk166@cumt.edu.cn , liupeng@cumt.edu.cn
[2] School of Information and Control Engineering, China University of Mining and Technology, Xuzhou 221116, Jiangsu, China
[3] Alibaba Group, Hangzhou 311121, Zhejiang, China
[4] School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, Jiangsu, China
liubing@cumt.edu.cn

Abstract. The Transformer-based image captioning models have made significant progress on the generalization performance. However, most methods still have two kinds of limitations in practice: 1) Heavily rely on the single region-based visual feature representation. 2) Not effectively utilize the future semantic information during inference. To solve these issues, we introduce a novel bidirectional-decoding based Transformer with multi-view visual representation **(BiTMulV)** for image captioning. In the encoding stage, we adopt a modular cross-attention block to fuse both grid features and region features by virtue of multi-view visual feature representation, which realizes full exploitation of image context information and fine-grained information. In the decoding stage, we design the bidirectional decoding structure, which consists of two parallel and consistent forward and backward decoders, to promote the model to effectively combine the history with future semantics for inference. Experimental results on the MSCOCO dataset demonstrate that our proposal significantly outperforms the competitive models, improving by **1.5 points** on the CIDEr metric.

Keywords: Image captioning · Transformer · Multi-view visual representation

1 Introduction

In the past years, multi-modal learning has achieved remarkable progress with the success of deep learning. Image captioning [1], as one of the most challenging multi-modal learning tasks, integrates technologies in the fields of both

S. Yu et al. (Eds.): PRCV 2022, LNCS 13534, pp. 735–748, 2022.
https://doi.org/10.1007/978-3-031-18907-4_57

natural language generation (NLG) and computer vision (CV). It aims to semantically describe an image using a fluent natural language sentence and has widely applied in many fields, such as image retrieval [2], robot interaction [3], and digital library [4]. Inspired by the Sequence-to-Sequence (S2S) model for neural machine translation (NMT) [5], mainstream image captioning approaches follow a popular encoder-decoder framework [6–8], which consists of a convolutional neural network (CNN) that encodes visual features from the input image, and a recurrent neural network (RNN) that serves as the decoder to generate the caption based on the visual features.

Recently, more and more researchers have incorporated the Transformer model into image captioning, considering that it enables to learn the semantic-rich encoding vector, capture the better long-distance dependence and support higher parallel operations. Therefore, image captioning models with the Transformer inherently have better performance compared with traditional methods. In terms of input visual features, existing image captioning approaches with the Transformer can be roughly divided into two classes: grid features based models and region features based models. In terms of caption decoder structure, most works focus on the left-to-right (L2R) paradigm. Despite the success of the current image captioning approaches, there still exist two major limitations as follows: 1) The region visual features cannot fully cover important details of the entire image, thus weakening the representation ability of the image used to guide the generation of captions. 2) Due to the masked self-attention mechanism, the transformer decoder can only learn the sequence information based on the partially generated words, and cannot effectively utilize the future predicted information to learn complete semantics.

To address the first limitation, we explore a novel method to coordinate region visual features and grid visual features. The grid visual features have the advantage of fine granularity and contain enough image context information, while the region visual features have adequate information of core image objects. To combine the merits of the two methods, we adopt a multi-view learning approach to integrate the two visual features, taking region visual features as the query input and grid visual features as the key-value pair input, respectively. As a result, the combination of grid visual features and region visual features can sensibly enhance the feature representation of the image. **To address the second limitation**, we design a bidirectional caption decoder, in which the forward and backward decoding processes have the same methodology but different learning directions of the sentence.

To sum up, a novel Transformer-based image captioning approach is proposed in this article, and the main contributions are given below:

1) An multi-view (MV) visual representation model is proposed, which can cooperatively encode region visual features and grid visual features, to extract the rich contextual information of an input image.
2) An end-to-end bidirectional multi-modal Transformer is presented, which can learn both history and future sentence semantics to generate more accurate image captions.

3) Extensive experiments are conducted on the MSCOCO dataset to validate the effectiveness of the proposed model. The evaluation results demonstrate that our model outperforms the mainstream models on most metrics.

2 Related Work

2.1 Image Captioning

There have been extensive studies and improvements on image caption. Researchers initially focused on template-based approaches [9,10] and search-based approaches [11]. Due to the development of deep learning in CV and NLP, recent works generally adopt an encoder-decoder framework. For instance, Vinyals et al. [7] proposed Convolutional Neural Networks (CNN) as an image encoder and then adopted LSTM as an encoder to generate corresponding descriptive sentences. In order to accurately understand the objects in the image, Anderson et al. [12] utilized up-down attention based on a ResNet within pre-trained object detector instead of a traditional convolution neural network (CNN) to extract region-based image features. Since the graph convolution neural network (GCN) can effectively extract spatial features, Yao et al. [13] applied a new GCN with Long Short-Term Memory (dubbed as GCN-LSTM) architecture that can enrich region-level representations and eventually enhances image captioning.

Despite the success of the above methods, RNN-based models are limited by their representation power. The new Transformer-based models with self-attention mechanism further improved the ability of representation and achieved SOTA (state-of-the-art) results in multi-modal tasks. For instance, to address the internal covariate shift problem inside self-attention, Guo et al. [14] employed Normalized Self-Attention that fixes the distribution of hidden activations in the Self-Attention mechanism. Pan et al. [15] incorporated Bilinear pooling which performs well in fine-grained visual recognition into image captions, and proposed X-Linear Attention Networks to achieve multi-modal input interaction. Huang et al. [16] extended the traditional attention mechanism and introduced the Attention on Attention (AOA) model, which can filter out attention vectors with low correlation with query vectors. Different from the semantic features of global multi-view features, Yu et al. [17] proposed Multi-modal Transformer that uses different object detectors to extract region-based local multi-view features, which can maintain the fine-grained semantic features of the image. Motivated by the prior research, we present a novel model based on an encoding-decoding framework, which improves visual representation by fusing two complementary visual features.

2.2 NMT

Recently, the research of captioning tasks is inspired by work related to NMT. Generally, most NMT decoders generate translations in a left-to-right (L2R) paradigm. However, the right-to-left (R2L) contexts are also crucial for translation predictions, since they can provide complementary signals to the models. In this paper, we focus on work with bidirectional decoding structures. For instance,

Liu et al. [18] introduced to find the best candidate from the combined N-best list via the joint probability of L2R and R2L models. To explore bidirectional decoding for NMT, Zhou et al. [19] construct a decoder that integrates forward attention and backward attention. Zhang et al. [20] constructed two Kullback-Leibler divergence regularization methods, which improve the coordination of translation sequences generated by L2R and R2L decoders. Bidirectional decoding structures for image captioning have also been successfully attempted, for example, Wang et al. [21] proposed a multimodal bi-directional LSTM method to realize end-to-end training for image captioning. Different from these RNN-based bidirectional decoders models, our proposed **BiTMulV** is a novel bidirectional decoding Transformer, which further incorporates the L2R and R2L structure.

3 Methodology

3.1 Model Architecture

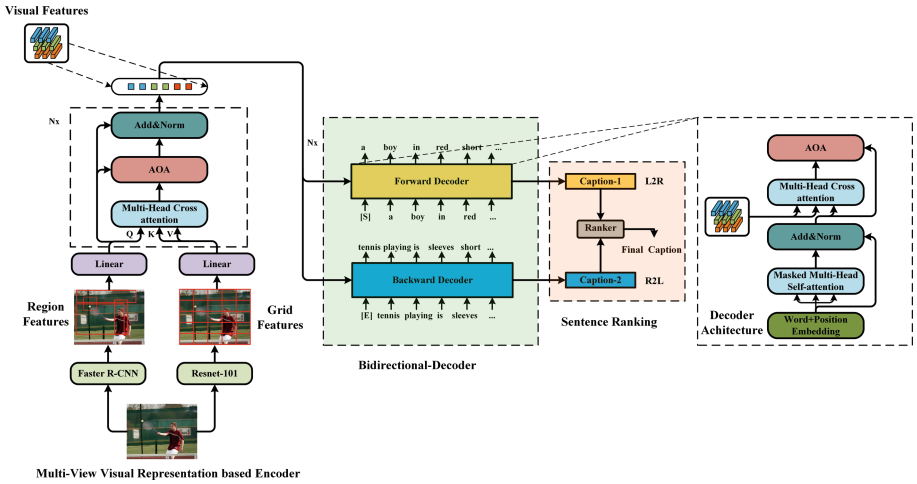


Fig. 1. Overview of the framework of our **BiTMulV** model architecture. Our **BiTMulV** for image captioning consists of three components. The Multi-View based encoder fuses region features and grid features, which enables reasoning of visual object relationships via a modular cross-attention block. In the decoder of **BiTMulV**, we propose the bidirectional decoder that learns the contextual semantics from the history and future. The Sentence Ranking method relies on the L2R sentence and the R2L sentence generated by the bidirectional decoder to select the final descriptive caption.

The visual representation of images indirectly affects the accuracy of generating descriptive captions in image captioning research. In this paper, we would like to make full exploitation of the visual sequence by fusing two different visual features. Moreover, we also would like to generate more accurate sentences by

learning global semantics. Specifically, we describe one model for image captioning, which is an end-to-end framework that is composed of an image encoder and a bidirectional decoder as shown in Fig. 1. The image encoder takes region visual features and grid visual features as its input. The visual features are then fed into the bidirectional encoder to obtain the attended multi-view visual representation with cross-attention learning. The bidirectional decoder predicts the next word recursively by exploiting the attended multi-view visual representation and the previous word. Finally, the final caption is selected from the L2R text and the R2L text generated by the bidirectional encoder via our proposed ranking mechanism.

3.2 Multi-view Visual Representation Based Encoder

Multi-visual (MV) Feature Extraction. Given the input images, the region features and grid features are extracted from the pre-trained object detector. The region features are extracted by the off-the-shelf Faster-RCNN pre-trained on Visual Genome [18]. For the grid features, we leverage the 49 grid visual features from the last convolutional layer of ResNet-101 [15].

AOA. The AOA [16] is proposed to eliminate the interference of irrelevant query vectors to obtain refined attention results. The vector N is concatenated by the attention result and the attention query, which is transformed into the information vector T via linear transformation and transformed into K via linear transformation and nonlinear activation function. (as shown in Fig. 1):

$$N = Concat(F_{Attention}, Q) \tag{1}$$

$$T = Linear(N), K = Sigmoid(Linear(N)) \tag{2}$$

To obtain the attended information H, the AOA incorporates the attention gate into the information vector by using element-wise multiplication:

$$H = T \odot k \tag{3}$$

Encoder Architecture. Considering the grid features contain rich visual context information, we combine it with region based features to facilitate the visual representation of our model. However, the region and grid features are unaligned, we adopt an unaligned multi-view image encoder, to fuse them (as shown in Fig. 1.). For the convenience of expression, the extracted grid features and region features from an image can be denoted as $G = \{g_1, g_2, g_3, ..., g_n\}$ and $R = \{r_1, r_2, r_3, ..., r_n\}$, respectively, where $r_n \in R^{d_1}, g_m \in R^{d_2}$ and $n \neq m$. Notably, we choose the region features R as the primary view and the grid visual features G as the secondary view and adopt the multi-head cross-attention (MHCA) block exploiting them:

$$Q = W_Q R, K = W_K G, V = W_V G \tag{4}$$

$$F_{MHCA} = MHCA(Q, K, V) = Concat(head_1, head_2, ..., head_m)w^o \tag{5}$$

$$head_m = att(Q, K, V) \tag{6}$$

where $W_Q, W_K, W_V \in R^{d \times d_h}$ are three linear transformation matrices, and $W^O \in R^{m*d \times d_h}$. The multi-head attention mechanism consists of m parallel heads with an independent scaled dot-product attention function. Afterwards, the AOA module is applied to filter out information vectors that are not relevant to the attention query and keeps only the useful ones, and then is followed by residual concatenations and layer normalization:

$$A = AOA(F_{MHCA}, R) \tag{7}$$

$$E = LayerNorm(R + A) \tag{8}$$

To generate more abstract and distinctive visual features for the bidirectional encoder, the image encoder adopts deep stacking and the L-th encoder block ${A^L}_{encoder}$ takes the results from the L-1-th encoder block E^{N-1} as follows:

$$E^N = {A^L}_{encoder}(E^{N-1}) \tag{9}$$

3.3 Bidirectional Decoder

Decoder Architecture. Generally, the standard Transformer decoder follows the L2R sentence generation paradigm, which utilizes the mask attention mechanism to implement the unidirectional modeling of the sequence. To alleviate the inability of the mask self-attentive mechanism in standard Transformers to explore the past and future context information of a sequence, we propose a novel bidirectional decoder architecture, which consists of a forward decoder and a backward decoder.

Based on the multi-view visual representations learned by the image encoder, the bidirectional decoder produces L2R and R2L sentence descriptions simultaneously for an image. Given a sequence of captions for an image as $S = \{s_1, s_2, s_3, ..., s_n\}$, the sentence is first tokenized into words and trimmed to a maximum length of 18 words. Note that each word in the sentence is represented as $s_n \in R^{300}$by using the Glove word embedding. The bi-directional decoder is implemented with two parallel decoders with the same structure. The forward decoder is fed with the L2R sequence $\overrightarrow{S} = (s_1, s_2, s_3, s_4, s_5, s_6, ...s_n)$, while the R2L sequence $\overleftarrow{S} = (s_n, ...s_6, s_5, s_4, s_3, s_2, s_1)$is fed to the backward decoder. Since two decoders have similar architectures, we mainly introduce the structure of the forward encoder.

Specifically, we first employ the masked multi-head self-attention (MMHSA) mechanism, which can characterize word to word relationships in sentences:

$$Q = W_Q \overrightarrow{S}, K = W_K \overrightarrow{S}, V = W_V \overrightarrow{S} \tag{10}$$

$$F_{MMHSA} = Masked - Mulihead(Q, K, V) \tag{11}$$

where $W_Q, W_K, W_V \in R^{d \times d}$ are three linear transformation matrices and the masked multi-head attention results are also normalized by residual concatenations and layer normalization:

$$\overrightarrow{A} = LayerNorm(\overrightarrow{S} + F_{MMHA}) \tag{12}$$

Afterwards, the second MHCA module is used to impose the multi-view visual representations-guided attention on the caption words:

$$F_{MHCA} = MHCA(W_Q \overrightarrow{A}, W_K V, W_V V) \tag{13}$$

where V denotes the multi-view visual features vector. Once again, we apply the AOA to measure how well visual features and text sequence features are related:

$$\overrightarrow{W} = AOA(F_{MHCA}, \overrightarrow{A}) \tag{14}$$

Similar to the image encoder, the bi-directional decoder consists of N identical decoder layers stacked in sequence. The text vector $\overrightarrow{w}$ is projected to M-dimensional space by a linear embedding layer, where M is the size of the vocabulary. Finally, the softmax layers are leveraged for the prediction of the probability of the next word.

Sentence Ranking. In the inference stage, we adopt the beam search strategy [17] with a beam size of 3 to improve the diversity of generated captions. Since the forward decoder and backward decoder generate two sentences via beam search, we design a sentence-level ranking mechanism, which compares the probabilities of the two sentences and selects the one with the largest probability as the final caption.

3.4 Training and Objectives

The proposed model is trained in two stages as same as a standard practice in image captioning [16,20,22,27]. we first pre-train our model by optimizing cross-entropy (XE) loss. For training both the forward and backward decoders, the joint loss is formulated as follows by averaging the cumulative forward and backward losses:

$$L_{XE}\theta = \text{AVE}(\overrightarrow{L}_{XE}\theta + \overleftarrow{L}_{XE}\theta) \tag{15}$$

$$\overrightarrow{L}_{XE}(\theta) = -\sum_{t=1}^{T} \log(p_\theta(y_t^* | y_{1:t-1}^*)) \tag{16}$$

$$\overleftarrow{L}_{XE}(\theta) = -\sum_{t=1}^{T} \log(p_\theta(y_t^* | y_{t+1:1}^*)) \tag{17}$$

where $y_{t+1:1}^*$ and $y_{1:t-1}^*$ are the ground truth of text sequences, θ is the parameters of our **BiTMulV**.

Subsequently, reinforcement learning is used to optimize non-differentiable metrics. Specifically, we employ a variant of the self-critical sequence training [22] on sequences sampled via beam search:

$$L_{RL}(\theta) = -E_{y_{1:T} \sim p_\theta}(R(y_{1:T})) \tag{18}$$

where the $R(\bullet)$ is employing the CIDEr-D as reward.

4 Experiments

4.1 Experimental Setup

Datasets. Our proposed method is trained, validated, and tested on the Microsoft benchmark dataset MSCOCO [23]. It contains 123,287 images and is each equipped with five different human-annotated captions. In particular, in order to ensure the effectiveness of offline experiments, we adopt the widely used Karpathy split [15], of which 113287 images are used for training and 5000 images are used for testing and verification.

Evaluation Metrics. The standard evaluation metrics include BLEU [23], ROUGE-L [24], SPICE [25], METEOR [26], CIDEr [27], which are used to evaluate the quality of the model.

Implementation Details. Our model is carried out on two NVIDIA RTX 2080 GPU, utilizing AOANET [16] as the baseline model. We pre-processed the annotated sentences, discarded words less than 6 times or did not appear in the Glove pre-training vocabulary, and finally formed a vocabulary with 9568 words. The dimension of the extracted visual feature vectors is 2048. Both the number of encoder layers and the number of decoder layers are set to 6.

We first train the proposed model by utilizing the word-level cross-entropy loss for 35 epochs with a mini-batch size of 15. The Adam optimizer is adopted with an initial learning rate of 1.5e-4, annealed by 0.65 every 3 epochs. To alleviate the exposure bias of cross-entropy optimization, our **BiTMulV** for all experiments is further trained using 10 epochs of self-critical [22] (SCST) loss with an initial learning rate of 1e-5 and annealed by 0.5.

4.2 Ablation Study

In this section, we analyze the effectiveness of the proposed method. The ablation study focused on: 1) The effect of Multi-View visual representation, 2) The effect of bidirectional decoder.

Table 1. Performance comparisons with different methods. The result is reported after the self-critical training stage

Method	B@1	B@4	M	R	C	S
Enc(Grid)+Dec(U)	78.5	36.6	27.4	57.2	124.5	20.6
Enc(Region)+Dec(U)	79.9	38.2	28.8	58.2	128.6	21.4
Enc(Region+Grid)+Dec(U)	80.5	38.4	29.4	59.0	130.6	22.2
Enc(Grid)+Dec(B)	78.8	36.8	27.9	57.5	124.9	20.8
Enc(Region)+Dec(B)	80.1	38.2	29.1	58.3	129.5	21.5
Full:BiTMulV	**80.8**	**38.8**	**29.6**	**59.5**	**131.3**	**22.7**

Enc (Grid/Region/Region+Grid)+Dec (U/B): In the comparison methods, Enc (Grid/Region) indicates that the image encoder adopts the grid

features or the region features, and Dec (U/B) means that the encoder adopts the unidirectional structure or the bidirectional structure.

1) The effect of Multi-view (MV) visual representation. To analyze the effect of our proposed **BiTMulV** visual representation, we implement validation experiments, as shown in Table 1. We can observe that the model Enc(Region)+Dec(U) performs better than the model Enc(Grid)+Dec(U) on all evaluation metrics. Compared with the Enc(Region)+Dec(U), the performance of the Enc(Region+Grid)+Dec(U) is significantly improved. Specifically, the Enc(Region+Grid)+Dec(U) achieves 2 and 0.6 increments over the metrics of CIDEr and BLEU-1, as shown in Table 1.

2) The effect of the bidirectional decoder. To further demonstrate the effectiveness of the proposed bidirectional decoder, we also compare the unidirectional decoder with bidirectional decoder for performance analysis. As illustrated in Table 1, the performance of both Enc(Grid)+Dec(B) and Enc(Region)+Dec(B) can be boosted (from **124.5 CIDEr to 124.9 CIDEr** and from **128.6 CIDEr to 129.5 CIDEr,** respectively). The same growing trend can be observed (from **129.5 CIDEr to 131.3 CIDEr**) after combining the MV visual representation method with bidirectional decoding(**Full: BiTMulV**), which further confirms the effectiveness of the bidirectional decoder.

4.3 Quantitative Analysis

Results on the Karpathy Test Split: Table 2 report the performance of our model and the competitive models on the Karpathy test split. Including SCST [18], a gradient-guided optimization method based on reinforcement learning that can effectively train non-differentiable metrics. We compare our method with the state-of-the-art methods, including SCST [22], LSTM-A [24], Up-Down [12], RFNet [25], GCN-LSTM [13], LBPF [26], SGAE [27], AoANet [16]. Up-Down introduces an attention mechanism that integrates bottom-up and top-down for fine-grained image understanding. GCN-LSTM uses GCN as the image semantic encoder and LSTM as the decoder. LSTM-A incorporates high-level

Table 2. Performance comparisons on Offline COCO Karpathy test split.

	Cross-entropy Loss						SCST Loss					
Model	B@1	B@4	M	R	C	S	B@1	B@4	M	R	C	S
SCST	-	30.0	25.9	53.4	94.0	0.0		34.2	26.7	55.7	114.0	0.0
LSTM-A	75.4	35.2	26.9	55.8	108.8	20.0	78.6	35.5	27.3	56.8	118.3	20.8
Up-Down	77.2	36.2	27.0	56.4	113.5	20.3	79.8	36.3	27.7	56.9	120.1	21.2
RFNet	76.4	35.8	27.4	56.8	112.5	20.5	79.1	36.5	27.7	57.3	121.9	21.2
GCN-LSTM	77.3	36.8	27.9	57.0	116.3	20.9	80.5	38.2	28.5	58.3	127.6	22.0
LBPF	77.8	37.4	28.1	57.5	116.4	21.2	80.5	38.3	28.5	58.4	127.6	22.0
SGAE	-	-	-	-	-	-	80.8	38.4	28.4	58.6	127.8	22.1
AoANet	77.4	37.2	28.4	57.5	**119.8**	21.3	80.2	**38.9**	29.2	58.8	129.8	22.4
Ours	**77.9**	**38.1**	**28.8**	**58.8**	118.1	**21.6**	**80.8**	38.8	**29.6**	**59.5**	**131.3**	**22.7**

image attributes into the CNN-RNN framework to facilitate sentence generation. RFNet adopts multiple CNNs to encode fusion features for images and insert a recurrent fusion procedure. LBPF fuses visual information from the past as well as semantic information in the future to improve the performance of caption generation. SGAE constructs a shared dictionary with induction bias to guide language generation. AOANet filters out irrelevant attention vectors by constructing the interaction of "information vector" and "attention gate".

For a fair comparison, we respectively utilize the cross-entropy loss and SCST to train all the models in the single model setting, as shown in Table 2. It can be observed that our model surpasses the previous state-of-the-art models in terms of BLEU-1, BLEU-4, METEOR, ROUGE-L, and SPICE and is slightly worse than the baseline model AOANet in terms of cider when optimized with cross-entropy. After optimizing the CIDEr-D score, our proposed method improves by 0.6 points on BLEU-1, 0.4 points on METEOR, 0.7 points on ROUGE-L and achieves a significant improvement of **1.5 points** in comparison with AOANet.

4.4 Qualitative Analysis and Visualization

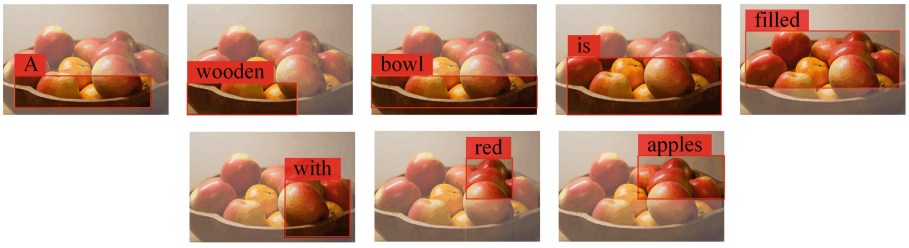
Qualitative Analysis. As shown in Fig. 2, we generate four captions for the sampled images, where "GT" indicates ground truth Sentences.

Image	Caption
	Baseline: Two cats laying on top of a bed.
	Ours: A black and white cat laying on a big bed.
	GT1: A couple of cats laying on top of a bed.
	GT2: Two cats laying on a big bed and looking at the camera.
	GT3: A couple of cats on a mattress laying down
	Baseline: A short train on a train track near trees.
	Ours: A yellow train on a train track near trees.
	GT1:A train traveling down tracks through rural countryside
	GT2:A yellow is traveling down the track in the country.
	GT3:A short train is coming down the train tracks.
	Baseline: A double-decker bus on a road
	Ours: A yellow double-decker bus driving down the street next to a blue bus
	GT1:A bus stops at an intersection outside on the street.
	GT2:A yellow double-decker bus on a road next to a blue bus
	GT3:A yellow double-decker bus driving down a street
	Baseline: A boy is running the bases.
	Ours: A child in orange is running bases on the playground.
	GT1:A boy in an orange shirt running the bases.
	GT2:A young boy is running the bases at a game.
	GT3:A child running for a base during a baseball game.

Fig. 2. Examples of captions generated by our method and the baseline model, as well as the corresponding ground truths.

Generally, the sentences generated by our proposed method are more accurate and descriptive than the baseline model. In detail, our proposed model is superior following two aspects:

1) Our proposed **BiTMulV**model could help understand visual contextual information and focus on fine-grained information to realize the alignment of the image and the captions. For example, the baseline model in the first example does not realize the color of the two cats, while our **BiTMulV** model effectively captures the "black" and the "white". In the third example, the blue car and the yellow car in the image are related to each other, but the baseline model does not recognize the blue car. On the contrary, our model effectively captures the context information of the image and recognizes the connection between the blue car and the yellow car.
2) Secondly, our **BiTMulV** model is more accurate and effective in counting objects of the Multi-object images. In the third example, the baseline model describes only one bus, while our method describes two buses. In the fourth example, our model recognizes images that describe the child, short, playground, and bases, while the baseline model does not capture short and playground.



(a) Baseline: A wooden bowl is filled with red apples

(b) BiTMulV: A wooden bowl full of apples and oranges

Fig. 3. Visualization.

Visualization. To better qualitatively evaluate the generated results with our proposed **BiTMulV** method, we visualize the attended image regions during the caption generation processes for Baseline and **BiTMulV**in Fig 3. It can be seen that our **BiTMulV** model correctly aligns image regions to the words, while the baseline model ignores some significant regions and then generates inaccurate captions. For example, the baseline model attends to the applies, while the oranges are not recognized. In contrast, by exploiting Multi-View Visual Representation for multi-modal reasoning, our **BiTMulV** model accurately localizes the oranges region to generate the "oranges".

5 Conclusions

In this paper, we propose a novel image captioning model (**BiTMulV**) based on the multi-view visual representation and bidirectional decoding structure. On one hand, we make use of image contextual information and fine-grained information by combining region visual features with grid visual features. On the other hand, the **BiTMulV** adopts the explicit bidirectional decoding structure, which can exploit both historical semantics and future semantics to guide model learning. The quantitative and qualitative experiments demonstrate the superiority of our proposed method over the existing deep image captioners.

Acknowledgement. This work was supported by the National Natural Science Foundation of China 61971421, the open fund for research and development of key technologies of smart mines (H7AC200057) and the Postgraduate Research & Practice Innovation Program of Jiangsu Province (KYCX21_2248).

References

1. Lu, J., Xiong, C., Parikh, D., Socher, R.: Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 375–383 (2017)
2. Revaud, J., Almazán, J., Rezende, R.S., Souza, C.R.D.: Learning with average precision: Training image retrieval with a listwise loss. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5107–5116 (2019)
3. Chen, C., Liu, Y., Kreiss, S., Alahi, A.: Crowd-robot interaction: Crowd-aware robot navigation with attention-based deep reinforcement learning. In: 2019 International Conference on Robotics and Automation (ICRA), pp. 6015–6022 (2019)
4. Fox, E.A., Ingram, W.A.: Introduction to digital libraries. In: Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020, pp. 567–568 (2020)
5. Sen, S., Gupta, K.K., Ekbal, A., Bhattacharyya, P.: Multilingual unsupervised NMT using shared encoder and language-specific decoders. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 3083–3089 (2019)
6. Chen, C., Mu, S., Xiao, V., Ye, Z., Wu, V., Ju, Q.: Improving image captioning with conditional generative adversarial nets. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 8142–8150 (2019)

7. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3156–3164 (2015)
8. Xu, K., et al.: Show, attend and tell: Neural image caption generation with visual attention. In: International Conference on Machine Learning, pp. 2048–2057 (2015)
9. Yang, Y., Teo, C.L., Daum´e III, H., Aloimonos, Y.: Corpus-guided sentence generation of natural images. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 444–454 (2011)
10. Kilkarni, G., et al.: Babytalk: understanding and generating simple image descriptions. IEEE Trans. Pattern Anal. Mach. Intell. **35**(12), 2891–2903 (2013)
11. Farhadi, A., et al.: Every picture tells a story: Generating sentences from images. In: European Conference on Computer Vision, pp. 15–29 (2010)
12. Anderson, P., et al.: Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6077–6086 (2018)
13. Yao, T., Pan, Y., Li, Y., Mei, T.: Exploring visual relationship for image captioning. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 684–699 (2018)
14. Guo, L., Liu, J., Zhu, X., Yao, P., Lu, S., Lu, H.: Normalized and geometry-aware self-attention network for image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10327–10336 (2020)
15. Pan, Y., Yao, T., Li, Y., Mei, T.: X-linear attention networks for image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10971–10980 (2020)
16. Huang, L., Wang, W., Chen, J., Wei, X.Y.: Attention on attention for image captioning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4634–4643 (2019)
17. Yu, J., Li, J., Yu, Z., Huang, Q.: Multimodal transformer with multi-view visual representation for image captioning. IEEE Trans. Circuits Syst. Video Technol. **30**(12), 4467–4480 (2019)
18. Liu, L., Utiyama, M., Finch, A., Sumita, E.: Agreement on target-bidirectional neural machine translation. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 411–416 (2016)
19. Zhou, L., Zhang, J., Zong, C.: Synchronous bidirectional neural machine translation. Trans. Assoc. Comput. Linguist. **7**, 91–105 (2019)
20. Zhang, Z., Wu, S., Liu, S., Li, M., Zhou, M., & Xu, T.: Regularizing neural machine translation by target-bidirectional agreement. In Proceedings of the AAAI Conference on Artificial Intelligence, pp. 443–450 (2019)
21. Wang, C., Yang, H., Meinel, C.: Image captioning with deep bidirectional LSTMs and multi-task learning. In: ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 14(2s), 1–20 (2018)
22. Rennie, S.J., Marcheret, E., Mroueh, Y., Ross, J., Goel, V.: Self-critical sequence training for image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7008–7024 (2017)
23. Lin, T.Y., et al.: Microsoft coco: Common objects in context. In European Conference on Computer Vision, pp. 740–755 (2014)
24. 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, pp. 4894–4902 (2017)

25. Jiang, W., Ma, L., Jiang, Y. G., Liu, W., Zhang, T.: Recurrent fusion network for image captioning. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 499–515 (2018)
26. Qin, Y., Du, J., Zhang, Y., Lu, H.: Look back and predict forward in image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8367–8375 (2019)
27. Yang, X., Tang, K., Zhang, H., Cai, J.: Auto-encoding scene graphs for image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10685–10694 (2019)