Image Caption Method Combining Multi-angle with Multi-modality

Yang Zhenyu*, Zhang Jiao
Qilu University of Technology (Shandong Academy of Science)
No.3501 University Road of Changqing District, 250353, Jinan Shandon, China
e-mail: 15853190997@163.com, 1841736565@qq.com

Abstract—**Image caption generation technology has attracted the interest of researchers due to its wide application in practical applications. It involves two vital areas of artificial intelligence: image processing and natural language processing. The existing method is to predict the generation of the next word based on the image features and the words generated in the previous state. However, it ignores the important role of text information. In this paper, we propose an image caption generation method that combines multi-angle with multimodality. The model firstly uses the fusion features of the global and local images as input. Picture description of the first sentence is generated using the baseline encoding-decoding model. The image caption which is generated firstly is then input into the sentence encoding network to generate a semantic feature vector of the first sentence. Then, the local visual feature vector of the image and the semantic eigenvector of the first sentence which are two different modal features, are combined and input into the attention-based language generation model to generate the next sentence. This allows our model to generate multi-angle descriptions in a targeted manner.**

Keywords-image captio; multi-angl; multi-modality; feature extraction; text generation

I. INTRODUCTION

Image caption generation technology and image semantic analysis, image annotation and semantic extraction have all been greatly developed and improved. In recent years, deep convolutional networks have achieved a series of breakthroughs in image classification and image recognition [1]. This has had a great positive impact on the development of image caption generation technology. Research on multimodal learning is also emerging. In fact, multimodal research is mostly used for the semantic information of text to assist the task of computer vision (CV). Such as image recognition [2], [3], image relationship detection [4], zero sample migration learning [5], visual question and answer (VQA)[6] and so on. In these tasks, the embedded semantic information contained in these words enhances the recognition of CV. In our paper, the semantic information contained in the word embedding is used to improve the text description effect of the image. The traditional description method has a single angle and a lack of content for the image content. It cannot fully describe the content displayed in the image. Simply describing the content of the image in one sentence is no longer sufficient for the public. Therefore, domain learning combing multi-angle with multi-modality can more fully describe the image content.

Our main contributions in this article are as follows: (1) The global and local features of the image are extracted by the network model based on the deep residual network[7], and the image features are fused by the visual attention mechanism. (2) The model embeds the fused image features into the language generation model, Long Short-Term Memory Networks (LSTM), to generate the first sentence of image. (3) The model inputs the first description generated into the encoder network of sentence to extract important vectors of word information. Then, the encoded sentence vector and the local feature vector of the image are jointly input into language generation model based on the Attention-based Long Short Term Memory Network to generate the next sentence. By analogy, until we reach the upper limit of our default sentence.

II. RELATED WORK

As a challenging and meaningful research task, image caption generation has quickly attracted the attention of researchers from all fields. In particular, in recent years, deep learning has been developed continuously. Deep learning-based image caption methods can generate more flexible and updated text descriptions without the constraints of templates and rules. It can also describe images that have never been seen before. Mao et al. [8] merged image information and text information obtained by DCNN into the same cyclic neural network (m-RNN) and integrated the image information into a sequence of text generation and achieved good results. The literature [9] used CNN to extract image features and replaced RNN with LSTM. In [10], the attention mechanism [11] was added to the image caption generation network model. Thus, the model could determine the focus of the image features autonomously when generating the text sequence. This method simulates the "attention" transfer process of human vision and can promote the generation process of word sequences. The generated sentences are more similar to human expression habits. In [12], an adaptive attention structure was designed. This model was an extension of the LSTM model. The addition of a
visual space vector in the hidden layer greatly improved the performance of the model. Liu et al. [13] proposed an image caption framework based on a self-retrieving module, which generated distinctive subtitles.

978-1-7281-4778-9/19/$31.00 ©2019 IEEE

Many tasks such as image classification, image detection and image recognition have benefited from the development of multimodal research. In the literature[14], the research direction and application of multimodal learning in deep learning are introduced in detail. The author proposed that multimodal learning is the real direction of artificial intelligence. Multimodality representations learn to eliminate redundancy by Complementary between multiple modes. This modality is represented by a joint representation to learn a better representation of the feature.

III. MODEL

Our paper proposes a new image caption generation method. The flow chart of structure is shown in Fig.1.

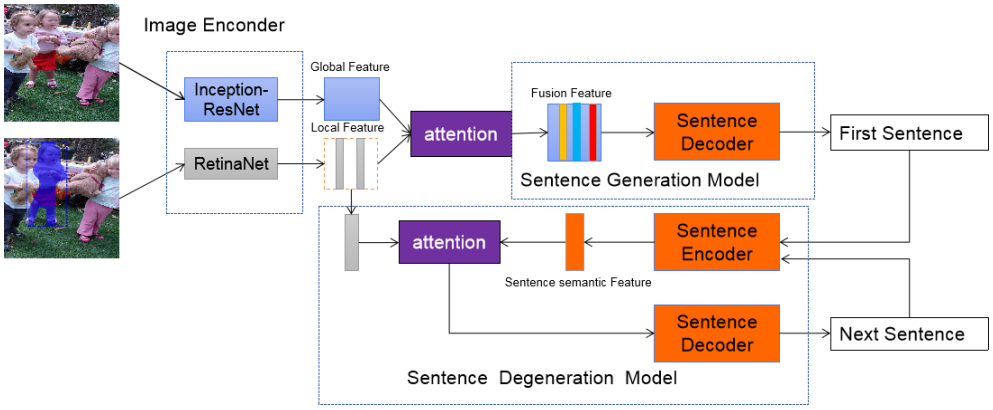

Figure 1. Image structure diagram of image caption method combining multi-angle multi-modality.

A. Image Encoding

The fusion of image features is very important for image representation. Global features typically contain contextual information around objects, while local features always contain fine-grained information about objects. Benefiting from the deep Inception-ResNet in image classification and deep RetinaNet's powerful capabilities in image target recognition, we use the Inception-ResNet to extract the global features of the image. The model transforms them into uniform-sized feature vectors through feature transformation. We combine the two by visual attention and input them into the baseline language generation model to generate the first sentence image caption. The process of image coding is as follows:

1) Global Image Feature: For the global feature of the image $H_{i,c}(x)$, we use the Inception-ResNet network. The Inception-ResNet is a deep model that is stacked by multiple convolutional modules. We extract the feature vector of the $1 \times 1 \times 1792$ dimension of the last pooling layer and re-adjust to the uniform feature size of 1×1×1024 by feature transformation: $H_{i,c}(x) = \{x_1, x_2, \dots x_L\}, x_i \in R^D$. Where L represents the number of feature vectors. D represents the dimension of the feature vector. This output will be fused with the output of the RetinaNet network in the visual attention module.

2) Local Image Feature: For the local image featureI_{box}, we use the RetinaNet network to extract local features of candidate regions. We mainly use the features of the ROI Pooling area. RetinaNet is a combination of residual network ResNet and pyramid network FPN. It can improve the accuracy and detect small targets. It can generate feature maps of higher quality. We select the $1 \times 1 \times 1024$ dimensional image feature of the layer before the pooling layer as the embedding vector： $\mathrm{I}_{box} = \{\mathrm{I}_{box1}, \mathrm{I}_{box2}, \dots \mathrm{I}_{boxn}\}$. Where box$n$ presents the number of feature vectors.

3) Fusion of Local and Global Image Features: We use a visual attention mechanism to fuse these two features:

$$V^t = \alpha_0^t H_{i,c}(x) + \sum_{i=1}^{n} \alpha_i^t \mathrm{I}_{box} \qquad (1)$$

where α_i^t represents the attention weight of each image feature at time t, and$\sum_{i=0}^{n} \alpha^t = 1$.

The sentence generation process dynamically weights each feature by assigning a positive weight α_i^t. Thus, our method can selectively focus on different objects at different time s and consider their context information at the same time. Note that the weight α_i^t measures the importance of each feature at time t and the correlation of each feature with previous information. Therefore, it can be calculated based on the previous information and each feature $I_i \in \{H_{i,c}(x), \mathrm{I}_{box1}, \mathrm{I}_{box2}, \dots \mathrm{I}_{boxn}\}$ by using the following formula:

$$\beta_i^t = W^t Tanh(W_0 I_i + W_h h_{t-1}^n + b) \qquad (2)$$

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

where β_i^t represents the association score of the feature I_i with the previously generated word. The weight α_i^t is obtained by normalizing β_i^t with softmax regression. h_{t-1}^n is the previous hidden state output, which is introduced in the next section. $W^t, W_h,$ and b are parameters that are learned by our model and shared by all features in all time steps. $Tanh$ is our activation function.

B. *Sentence Generation Model*

In general, the first sentence in multi- sentences contains summary information and advanced information in the image. Therefore, we have adopted a relatively popular generation model based on LSTM. We improve the input vectors of previous models through visual attention rather than simply use image features extracted from deep convolutional neural networks as input. Before entering the LSTM, the fused visual feature vectors are transformed with a fully connected layer to have the same dimension as the word embedding. We use a single layer of LSTM [15] for sentence decoding. The hidden state of the LSTM and the initial value of the cell are set to zero. The model mainly uses the fusion visual feature of the image encoder to predict the initial input of the first word in the first sentence and predicts the next word. The first description of the image is then generated. The model is trained as a training set with sentence description of the image, producing a summary of the image. The model is jointly trained with the sentence regeneration model, and the first sentence generated is used as a partial initialization input for the sentence regeneration model. In all the LSTM modules used in this article, the dimensions of the word embedding and the hidden state are 512 and 1024, respectively.

C. *Sentence Regeneration Model*

We propose a sentence regeneration model. The model takes combination of the first sentence vector generated and the partial image feature of the specific candidate region as input. It sequentially generates a plurality of statements. The sentence regenerating model consists of two main parts: a sentence encoder and a sentence decoder.

1) Sentence Encoder: A sentence encoder is used to extract semantic vectors from text descriptions. Our paper presents two well-known text encoders. The first is the Bidirectional Long Short Term Memory network (Bi-LSTM) [16]. It can encode context information and be better than traditional LSTM. The second is the use of one-dimensional convolutional neural networks for sentence coding. In Bi-LSTM, each word corresponds to two hidden states, and one direction corresponds to one. For the first sentence $\mathrm{r} = \{w_t\}_{t=1}^{T}$, we first embed each word w_t into the vector set e_t with one hot vector, and use the bidirectional Bi-LSTM to encode the entire expression. The last hidden representation of each word is the concatenation of hidden vectors in both directions. This hidden vector contains information about the entire sentence centered on w_t. It can be expressed by the following formula:

$$e_t = embedding(w_t) \quad (4)$$

$$\vec{h}_t = \overrightarrow{LSTM}(e_t, \vec{h}_{t-1}) \quad (5)$$

$$\overleftarrow{h}_t = \overleftarrow{LSTM}(e_t, \overleftarrow{h}_{t-1}) \quad (6)$$

$$h_t = [\vec{h}_t, \overleftarrow{h}_t] \quad (7)$$

But in a sentence, not every word can express the meaning of the sentence equally. Therefore, we introduce attention mechanisms to extract words that are important to the meaning of the sentence. The representations of these information words are then grouped together to form a sentence feature vector. In order to enable neural networks to automatically place "attention" on these words, we designed a word-based attention model. Its calculation formula is as follows:

$$u_t = \tanh(W_w h_t + b_w) \quad (8)$$

$$\alpha_t = \frac{\exp(u_{it}^T u_w)}{\sum_t \exp(u_{it}^T u_w)} \quad (9)$$

$$s_i = \sum_t \alpha_{it} h_t \quad (10)$$

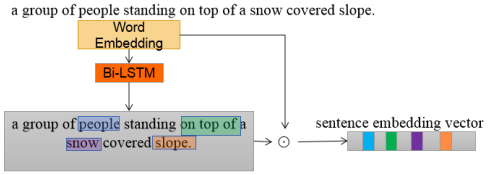

Figure 2." Structure diagram of sentence encoder.

In other words, we first input h_t through a single layer of MLP to obtain u_t, which is a hidden representation of h_t. We measure the importance of words using the similarity of the word-level context vectors u_w and u_t. The importance of each word is calculated by a normalized weight α_t obtained by a softmax function. Then, we obtain a representation of each sentence by weighting the output of the network Bi-LSTM. The structure diagram is shown in Fig 2:

Inspired by the literature[17], the one-dimensional convolutional neural network is applied to sentence encoding. Our CNN model embeds 512-dimensional words as input and has three convolutional layers to learn hierarchical features. Each convolutional layer has a kernel size of 3, a step size of 1, and 1024 feature channels. The max pooling operation is applied to the feature map u_i extracted from every convolutional layer, producing a 1024-dimensional feature vector. The last sentence is characterized by linking different level of feature vectors together: $\mathrm{u} = [u_1, u_2, \ldots u_n]$. In the fourth part, we compare the experimental results of the proposed two encoder.

2) Sentence Decoder: We use the attention mechanism to fuse the local visual features of a particular candidate region with the first sentence vector generated previously. It acts as a multimodal input to the sentence decoder, generating the next description. The sentence decoder is a superimposed two-layer LSTM model. The local image feature vector is converted to the initial input of the 2-layer LSTM. The learning encoding process of the previous sentence leads our model to generate the next sentence. We repeat this process until reach the upper limit for the number of generated sentences.

In order to focus different sentences on different regions of the image and capture the dependencies between sentences, we adopt a attention mechanism. The semantic features of the preceding sentence and regional visual representations are provided through a fully connected layer and a softmax layer to obtain the attention distribution of the image region. Firstly, we calculate the attention weights in the region as:

$$\mathrm{a} = W_{att} Tanh(W_v v + W_s s) \quad (11)$$

Where $v \in R^{d_v \times k}$ is the local visual feature of the region learned by the image encoder. $s \in R^{d_s}$ represents the vector encoding described in the first sentence. $W_{att} \in R^{1\times k}, W_v \in R^{k\times d_v}, W_v \in R^{k\times d_s}$ are the parameters of the attention network. $d_v = 1024$ is the dimension of the local visual feature. $d_s = 2048$ is the dimension of the sentence feature.($d_s = 2048$ is used for Bi-LSTM encoders. $d_s = 3072$ is used for CNN sentence encoder.) Next, we standardize all areas to get the attention distribution:

$$\alpha_i = \frac{\exp(a_i)}{\sum_i^n a_i} \qquad (12)$$

Where a_i is the ith vector in vectors a. Finally, we calculate the weighted visual representation:

$$v_{att} = \sum_{i=1}^{k} \alpha_i v_i \qquad (13)$$

The initial input of the sentence decoder is a joint weighted representation of local image feature for particular candidate region and a semantic feature vector. When generating different sentences, the attention model focuses on different areas of the image based on the context of the previous sentence. Then it filters out features or regions that are not related to the current sentence. The model cannot directly see the sentence encoding, so it is unlikely that the semantic input will be over-fitting. For model expressions with and without attention to the module, comparisons of performances can be seen in the experimental section.

D. Parameter Learning

The models that we propose are all trained by the Adam optimizer. The initial learning rate is set to le-4. The learning rate is attenuated by 0.9 every 5 steps. The batch size is set to 460 for training. In the training process, we adopted a mandatory strategy. In the next time period, we provide the generator with basic real words and sentences. In the test process, in order to improve the efficiency of the test, a greedy search method is used to generate words and sentences at each time stage. The previously generated words and sentences will be input to the decoder as part of the initial input of the next word and sentence. Our sentence regeneration model will generate multiple descriptive sentences until it reaches the upper limit of number of sentences.

All modules are jointly trained in an end-to-end manner. Each training sample is a set (I, S). Where I represents an image and S represents a sentence description. Given the training samples (I, S), our first sentence generation model calculates the probability distribution $\mathrm{p_{first}}$. By normalization, we can get the real sentence distribution $\mathrm{p_{real}} = \mathrm{S}/\|\mathrm{S}\|$. This training step has a cross entropy loss $\mathrm{L_{first}}$ between $\mathrm{p_{first}}$ and $\mathrm{p_{real}}$. We combine the first sentence generated by the model and the local image features into the sentence regeneration model. There is a cross entropy loss $\mathrm{L_w}$ in the word generation training process. Therefore, the overall sentence description generates training loss L as:

$$\mathrm{L} = \gamma_l L_l + \gamma_w \sum_{s=1}^{S} L_w(p_s^t, w_s^t) \qquad (14)$$

IV. EXPERIMENT

To compare with existing models, we conducted a large number of experiments to verify the validity of our model. We used a variety of evaluation criteria, multiple data sources and model architectures. Based on the speed of the previous model, the model is more unique, more targeted, and more detailed in image captioning.

A. Dataset introduction and processing

The datasets Flickr and MSCOCO are popular training datasets. This article mainly uses MSCOCO datasets and Flickr datasets for training and testing. The network model used in our paper has been proven to surpass the original network model[18][19] in terms of parameter efficiency and final performance through experiments. In order to evaluate our proposed model, we use several popular evaluation indicators: BLEU@N, METEOR[20] and CIDER[21]. All metrics are calculated using the code published by the COCO Evaluation Server[22].

B. Experimental analysis

In the image encoding part, the following experiments are mainly carried out. In order to verify the validity of the fusion of the global and local image features proposed in our paper, three comparative experiments were carried out for the first sentence generated by the model: Global-based image caption, Local-based image caption and image caption based on the fusion of global and local features.

1) Contrast experiments of global image features, local image features and fusion of image feature: In order to verify the important role of the global features of the image and the local fine-grained features in image caption generation, we conducted multiple sets of comparative experiments by evaluating the first description generated by the sentence generation model. For fair comparison, we all use network LSTM to generate the first sentence description. This process mainly includes the following parts:

a) The first sentence of the image caption content is generated using only the global image feature $H_{i,c}(x)$ extracted by the deep network. We conducted comparative experiments using the networks VGG16, VGG19 and Inception-ResNet, respectively. As shown in table 2. We extract the 4096-dimensional image features of the fc7 layer of VGG16, the image features of the fc7 layer of VGG19 and the 1×1×1792-dimensional image features of the final pooled layer of Inception-ResNet. They are tuned to a uniform input size through feature conversion. They adjust to a uniform input-size through feature conversion and are used as the initial input to the LSTM. We also show the visualization results of image feature extraction from three network models. Examples of a resulting for an image based on the global image feature of the modal, as the first sentence shown in Fig.5.

As can be seen from Table 1, using the image encoders of the networks VGG16 and VGG19, the values on the popular score indicators of the entire model differed little. However, the Inception-ResNet network get a good score for image caption model. Therefore, the final image encoder uses the

network based on Inception-ResNet to extract the global image features.

TABLE I. " FOR THE FIRST SENTENCE OF THE MODEL GENERATION, VGG16, VGG19 AND INCEPTION-RESNET WHICH ARE THREE DIFFERENT IMAGE ENCODERS BASED ON GLOBAL FEATURE, ARE USED TO COMPARE THE SCORES OF THE ENTIRE MODEL

Model	B@1	B@2	B@3	B@4	Meteor
VGG16-LSTM	66.6	45.1	30.4	20.3	--
VGG19-LSTM	66.8	44.6	29.8	21.8	--
ours	67.8	49.7	34.6	24.9	22.8

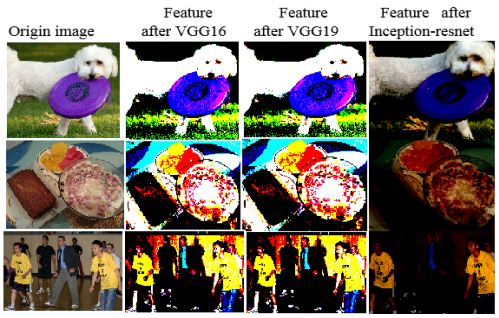

Figure 3." Examples of visualization results of image feature extraction for network models VGG16, VGG19 and Inception-ResNet.

b) The first fine-grained image feature $I_{box} = \{I_{box1}, I_{box2}, \ldots I_{boxn}\}$ extracted by the deep region detection network is used to generate the first sentence of the image content. We use the extraction network based on Faster-RCNN for the local feature of image and RetinaNet network to conduct comparative experiments. We extract the local features of the candidate regions, mainly using the image features of the ROI Pooling region. As is shown in Table 3. An example of image description based on a local image feature model is shown in the second sentence of Fig.4.

From Table 2, we can see that the evaluation score based on the local image feature is not much higher than the model based on the global image feature. Sometimes the score will be higher. The effect of using the network based on RestinaNet is generally higher than the score using the Faster-RCNN network. Therefore, the final model of our paper uses the network RestinaNet for local image feature extraction.

TABLE II. " FOR THE FIRST SENTENCE OF MODEL GENERATION, TWO DIFFERENT IMAGE ENCODER OF LOCAL IMAGE FEATURE, FASTER-RCNN AND RETINANET, ARE USED TO COMPARE PERFORMANCES ACROSS THE MODEL SCORES

Model	B@1	B@2	B@3	B@4	Meteor
Faster-RCNN-LSTM	44.1	45.1	33.4	22.3	20.9
RetinaNet-LSTM	66.8	47.5	33.7	22.8	21.1
ours	67.8	49.7	34.6	24.9	22.8

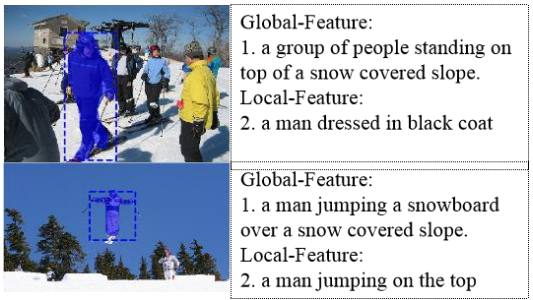

Figure 4." Descriptions of the test image based on the model of global image and the local image features.

c) The first sentence of the image content is generated using the fusion features V^t of the global and local image features. We use the fusion features of the global image features extracted by the network Inception-resnet and the local image features extracted by RetinaNet to carry out model training. The contrast of the above models is shown in Table 3.

From Table 3, we find that the scoring effect based on fused image features is much higher than the scoring results based on global or partial image features. Usually we focus on important objects in the image, but small objects may not be overlooked. The best performance can be obtained by the coding method based on the fusion feature.

TABLE III. " FOR THE FIRST SENTENCE OF THE MODEL GENERATION, COMPARISON TEST RESULTS OF DESCRIPTION BASED ON GLOBAL FEATURES, LOCAL FEATURES AND FUSED IMAGE FEATURES

Model	B@1	B@2	B@3	B@4	Meteor	CIDER
Global-only	67.8	49.7	34.6	24.9	22.8	77.1
Local-only	66.8	47.5	33.7	22.8	21.1	73.6
Global-local	70.1	51.7	37.6	26.8	23.8	86.2

2) In the sentence regeneration model experiments: For comparison, we reproduce the baseline model of the two image captions[23][24]. For all models, we use the same pre-trained RetiaNet encoder based on local image feature. Since Bi-LSTM encoding achieved better performance than convolutional encoding in experiments, our final model of sentence encoder use Bi-LSTM. We also implement a baseline model with no attention. In the inadvertent statement regeneration model, the text semantic encoding learned by the statement encoder is used as the initial hidden state and unit state of the sentence decoder.

From Table 4 we can see that our last attention-grabbing model has a significant improvement over the baseline model in all of the evaluation indicators. In addition, although the hierarchical model[25] achieved a fairly high evaluation score, the generated description content was single, lacking in relevance and broad. In contrast, the description generated by our proposed model contains multiple objects, which are rich in content and clear. As is shown in Fig.5.

TABLE IV. " WE COMPARE OUR TO TWO BASELINE MODELS. WE USE THE BLEU, METEOR, AND CIDER INDICATORS TO EVALUATE DESCRIPTIONS GENERATED BY THE TEST SET

Methods	B@1	B@2	B@3	B@4	Meteor	CIDER
DenseCao	33.18	16.92	8.54	4.54	12.66	12.51
Hierarchical generation	41.90	24.11	14.23	8.69	15.95	13.52
Ours-re-conv	41.60	29.80	21.70	16.30	22.70	30.90
Ours-re-BiLSTM	42.30	30.70	22.30	16.50	23.60	32.20
Ours-re-attention	46.40	35.80	27.00	19.50	27.40	36.60

The focus of our paper is to generate multiple description based on different angles for an image. This model combines the input of two different modalities of image and text for end-to-end training to achieve our expected experimental results. As we can see from Fig. 5, for the single sentence generated, although the image content can be summarized, it is not targeted. The main reason is that the image features and text information are not fully utilized. Although our model does not produce very good new sentences that have never appeared in the training set. This may be due to the difficulty of learning the correct grammar from a small corpus. Because the objective function of training does not consider syntactic correctness. We anticipate that, to address these limitations, we need a larger, better data set, a new training strategy and a new evaluation metric that considers both keyword accuracy and grammatical accuracy.

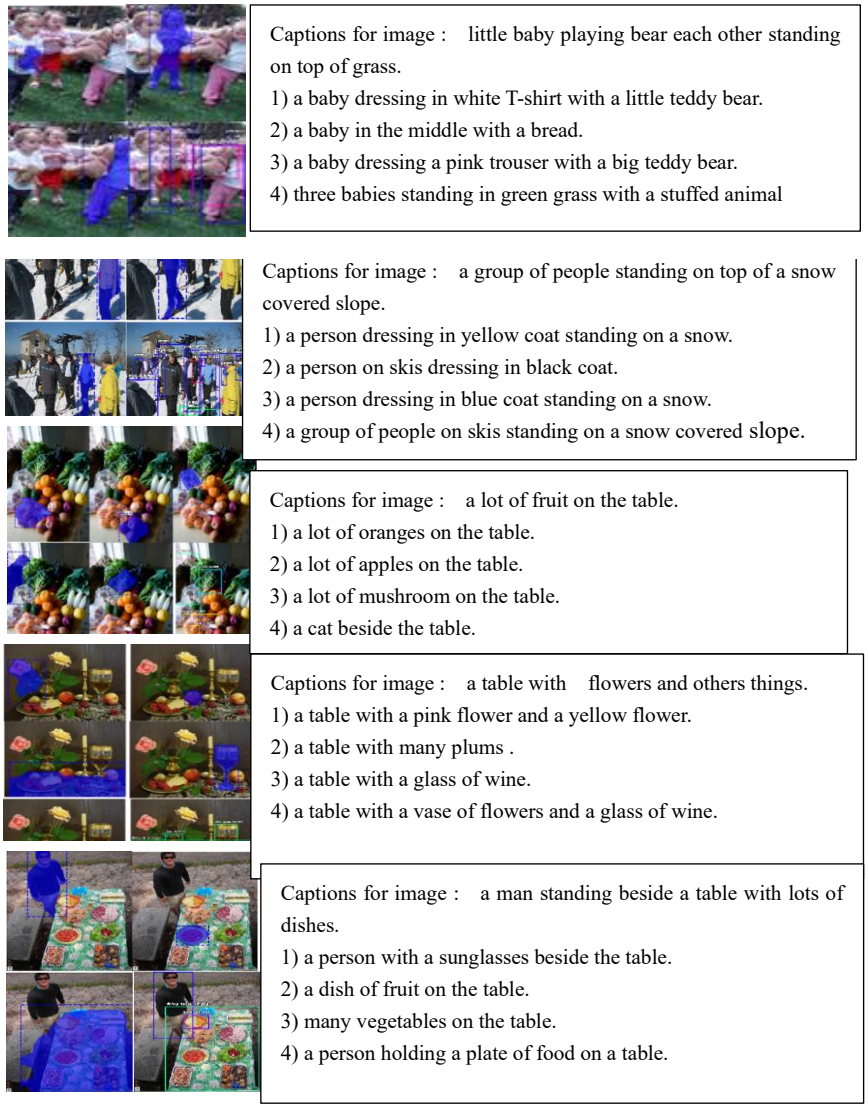

Figure 5. Examples of the image caption about the effect of our model.

V. CONCLUSION

We have designed an image caption generation method that combines multi-angle and multimodality. Comparing with the previous image caption techniques, we clearly observe the unique properties of the model. The score of the description statement has been greatly improved.

Although the descriptions produce exciting results, we believe this is just the beginning. In the future, it is possible to consider the description of image content through autonomous learning. We can also consider optimizing our model with a more efficient feedback architecture. The future direction will consider a system that can more specifically describe the identification descriptions of traffic signs and medical images. We will describe the anchoring to a given image property and location, responding to user-specified questions or tasks in a timely manner, and evaluating the higher-level goals (such as robots) through the application.

ACKNOWLEDGEMENT

This work was supported by and Shandong Natural Science Fund Project(No.ZR2017LF021) and Key Research and Development Program of Shandong Province (No.2017XCGC0605).We would like to thank the developers of TensorFlow, the authors for their open source code by Oriol Vinyals, Alexander Toshev, and the support of the Provincial Natural Science Foundation.

REFERENCE

[1] M.D.Zeiler and R.Fergus. Visualizing and understanding convolutional neural networks. In ECCV, 2014.

[2] Srivastava N, Salakhutdinov R. Multimodal Learning with Deep Boltzmann Machines. JMLR.org, 2012.

[3] Frome A, Corrado G S, Shlens J, et al. Devise: A deep visual-semantic embedding mode. International Conference on Neural Information Processing Systems. Curran Associates Inc. 2013.

[4] Radenović, Filip, Tolias G, Chum, Ondřej. CNN Image Retrieval Learns from BoW: Unsupervised Fine-Tuning with Hard Examples. 2016.

[5] Socher R, Ganjoo M, Sridhar H, et al. Zero-Shot Learning Through Cross-Modal Transfer. 2013.

[6] Lu J , Yang J , Batra D , et al. Hierarchical Question-Image Co-Attention for Visual Question Answering. 2016.

[7] Xuelei Li. FPGA Accelerates Deep Residual Learning for Image Recognition. Proceedings of 2017 IEEE 2nd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC 2017), 2017, pp.4.

[8] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang and Alan L. Yuille, Deep Captioning with Multimodal Recurrent Neural Networks (m-RNN). arXiv:1412.6632 / ICLR 2015.

[9] Vinyals O, Toshev A, Bengio S, et al. Show and tell: A neural image caption generator.//CVPR2015: Proceedings of the 2015 International Conference on Computer Vision and Pattern Recognition. Washington, DC: IEEE Computer Society, 2015.pp. 3156-3164.

[10] Liang R, Zhu Q X, Liao S J, et al. Deep natural language description method for video based on multi-feature fusion[J]. Journal of Computer Applications, 2017, 37(4),p.1179-1184.

[11] Xu K, Ba J, Kiros R, et al. Show, attend and tell: Neural image caption generation with visual attention.Computer Science, 2015,pp. 2048-2057.

[12] Lu J, Xiong C, Parikh D, et al. Knowing when to look: Adaptive attention via a visual sentinel for image captioning.//CVPR2017: Proceedings of the 2017 International Conference on Computer Vision and Pattern Recognition. Washington, DC: IEEE Computer Society, 2017:,pp.3242–3250.

[13] Xihui Liu, Hongsheng Li, Jing Shao, Dapeng Chen, Xiaogang Wang. Show Tell and Discriminate: Image Captioning by Self-retrieval with Partially Labeled Data. European Conference on Computer Vision (ECCV), 2018. In arXiv: [2018-08-30] https://arxiv.org/abs/1803.08314.

[14] Baltrušaitis T, Ahuja C, Morency L P. Multimodal Machine Learning: A Survey and Taxonomy. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, (99),pp.1-1.

[15] Hochreiter, S, Schmidhuber, J.: Long short-term memory. Neural Comput. 1997,9(8),pp.1735–1780.

[16] Graves, A., Schmidhuber, J.: Framewise phoneme classification with bidirectional lstm and other neural network architectures. Neural Netw. 2005, 18(5–6), pp.602–610.

[17] Conneau, A., Kiela, D, Schwenk, H., Barrault, L., Bordes, A: Supervised learning of universal sentence representations from natural language inference data. arXiv preprint arXiv:1705.02364,2017.

[18] Karpathy A, Fei-Fei L. Deep visual-semantic alignments for generating image descriptions. Proceedings of the IEEE conference on computer vision and pattern recognition. 2015,pp.3128-3137.

[19] Vinyals O, Toshev A, Bengio S, and Erhan D. Show and tell: A neural image caption generator. Computer Vision and Pattern Recognition (CVPR), IEEE Conference on. IEEE, (2015),pp.3156-3164.

[20] Banerjee S, Lavie A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. 2005,pp.65-72.

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

[22] Chen X, Fang H, Lin T Y, et al. Microsoft COCO captions: Data collection and evaluation server. arXiv preprint,arXiv:1504.00325, 2015.

[23] Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: CVPR, 2015,pp.3156–3164.

[24] Krause, J., Johnson, J., Krishna, R., Fei-Fei, L.: A hierarchical approach for generating descriptive image paragraphs. In: CVPR, 2017,pp.3337–3345.

[25] Z. Yang, D. Yang, C. Dyer, X. He, A. J. Smola, and E. H. Hovy. Hierarchical attention networks for document classification. In HLT-NAACL, 2016.