 CMC, vol. 59, no.2, pp.575-589, 2019

Feedback LSTM Network Based on Attention for Image Description Generator

Zhaowei Qu[1, *], Bingyu Cao[1], Xiaoru Wang[1], Fu Li[2], Peirong Xu[1] and Luhan Zhang[1]

Abstract: Images are complex multimedia data which contain rich semantic information. Most of current image description generator algorithms only generate plain description, with the lack of distinction between primary and secondary object, leading to insufficient high-level semantic and accuracy under public evaluation criteria. The major issue is the lack of effective network on high-level semantic sentences generation, which contains detailed description for motion and state of the principal object. To address the issue, this paper proposes the Attention-based Feedback Long Short-Term Memory Network (AFLN). Based on existing codec framework, there are two independent sub tasks in our method: attention-based feedback LSTM network during decoding and the Convolutional Block Attention Module (CBAM) in the coding phase. First, we propose an attention-based network to feedback the features corresponding to the generated word from the previous LSTM decoding unit. We implement feedback guidance through the related field mapping algorithm, which quantifies the correlation between previous word and latter word, so that the main object can be tracked with highlighted detailed description. Second, we exploit the attention idea and apply a lightweight and general module called CBAM after the last layer of VGG 16 pretraining network, which can enhance the expression of image coding features by combining channel and spatial dimension attention maps with negligible overheads. Extensive experiments on COCO dataset validate the superiority of our network over the state-of-the-art algorithms. Both scores and actual effects are proved. The BLEU 4 score increases from 0.291 to 0.301 while the CIDEr score rising from 0.912 to 0.952.

Keywords: Image description generator, feedback LSTM network, attention, CBAM.

1 Introduction

The image description generator is an analytical study which can generate a natural language description expressing the meaning of an image [Wu, Shen, Hengel et al. (2016)]. It is a frontier, widely used, and significant research, which can generate accurate image description for blind people, children enlightenment learning, visual understanding on search engine and so on.

[1] Beijing Key Laboratory of Network System and Network Culture, Beijing University of Posts and Telecommunications, Beijing, 100876, China.

[2] Department of Electrical and Computer Engineering, Portland States University, Portland, OR 97207-0751, USA.

[*] Corresponding Author: Zhaowei Qu. Email: zwqu@bupt.edu.cn.

The first problem with existing image description generator algorithms is the lack of accuracy rate for detailed description. Based on the codec technology, the generation process of image description is to integrate text information through RNN after encoding the image with CNN, and then generate a smooth image description with the target language, which utilize the relevance at the time level of RNN. The mainstream research methods range from the simple NIC [Vinyals, Toshev, Bengio et al. (2015)] to the various network structures based on codec technology, such as bidirectional-LSTM [Wang, Yang, Bartz et al. (2016)], Guiding-LSTM [Jia, Gavves, Fernando et al. (2016)], and visual semantic alignment model [Karpathy and Fei-Fei (2014)]. These methods continually update the accuracy rate under the public evaluation criteria.

The second problem with existing descriptor approaches is the lack of high-level semantic sentences which can distinguish between primary and secondary objects in image, leading plain description and unsmooth overall result. It becomes hard to succeed on effect and evaluation scores only by making minor improvement to the codec framework. With the fact that each image has a main object, instead of straightforward simplicity, the description should emphasize the action and state of details of the object, which can be smoother comprehensively and more prominent in particulars. Also, the network relies on a rational codec algorithm in an excellent end-to-end network structure [Girshick, Donahue, Darrell et al. (2014)].

To address the issues above, we propose Attention-based Feedback LSTM Network (AFLN) for image description generator. Based on codec framework, we innovatively proposed attention-based feedback LSTM network in the decoding stage and the CBAM for enhancing the feature expression. Our network performance has been proved of significant effect through experiments. There are three main contributions in this paper. First, we present attention-based feedback LSTM network. The LSTM network provides feedback guidance for the latter generated word. In the decoding stage, the previous generated word guides the generation process of the next word through this feedback network, while effectively performing feedback tracking on the object state and action, and generate detailed descriptions finally. Second, we generate a name-feature dictionary in the encoding phase. This cache structure integrates the image-detected name-bbox vocabulary and CNN extracted coding features, and generates the name-feature dictionary through the related field mapping algorithm for feedback attention. Third, we apply CBAM during the coding phase. The latest CBAM enhances the expression of image extraction, and effectively add attention guidance while less increasing the burden of model training.

The paper is organized as follows. Section 2 discusses previous work on image description, which provides readers with a general idea of related work. Section 3 explains the basic design concept and presents the approach in detail. Section 4 contains our experiments and evaluations and Section 5 presents our conclusion.

2 Related work

Related work based on codec architecture solutions.

The successful practice of RNN in machine translation field breaks the traditional template-based technical bottleneck of img2text generation, such as m-RNN proposed by

Mao et al. [Mao, Xu, Yang et al. (2014)], GNMT proposed by Google [Wu, Schuster, Chen et al. (2016)], and model for learning paragraph vector [Zeyu, Qiangqiang, Yijie et al. (2018)]. These breakthroughs lead to refocus on the advantages of LSTM network in long-term dependency information [Hochreiter and Schmidhuber (1997)]. The LSTM unit allows time-associated features to be well abstractly expressed, which help the picture better semantic [Graves (2012)], just like the NIC model proposed by Vinyals, which creatively uses LSTM as a decoder [Vinyals, Toshev, Bengio et al. (2016)]. But in order to avoid image noise and overfitting, the NIC only inputs image features at the beginning of the decoding phase, resulting in the reduction of image information in later LSTM units and the accuracy of image description [Donahue, Hendricks, Guadarrama et al. (2015)].

Aiming to address the problem of inaccuracy, research jumps into the study of visual semantic alignment of image main body. For instance, the visual semantic alignment model leverage large image sentence datasets by treating the sentences as weak labels, in which contiguous segments of words correspond to some particular, but unknown location in the image [Huang, Wu and Wang (2018)]. Also, the semantic matching model of image text proposed by Carrara et al. [Carrara, Esuli, Fagni et al. (2017)], organizes text features by correct semantic order to guide visual representation, which can translate textual information into visual representations through learning. Compared with NIC, the idea of using depth graphic features to embed embedding has a great advantage in image local description. However, due to the large differences in visual semantics, pixel-level image representations often lack high-level semantic information, forcing the image body features to correspond to text semantics, resulting in a description of an object that tends to be the same single description, lacking high-level semantic information in pixel-level features, which force the image body features to correspond to text semantics [Mingxing, Yang, Hanwang et al. (2019)].

Aiming to address the problem of insufficient detail, there are two research routes. First, Kinghorn proposed a codec architecture based on region object detector [Kinghorn, Zhang and Shao (2017)]. And Chen proposed a bidirectional mapping algorithm between image and text description, which construct dynamically a visual representation of the scene image [Chen and Zitnick (2015)]. Second is about attention. For instance, the SCA-CNN model combines the attention mechanism of space and channel direction in CNN to dynamically modulate the sentence generation context in multilayer feature maps. Anderson proposed a bottom-up and top-down combined attention mechanism that calculates attention at the level of objects and other significant image areas [Anderson, He, Buehler et al. (2018)]. Ding assigns different weights to the embedding process according to the correlation between the image and texts, and maximizes the consensus reference features of the target image and the consensus scores corresponding to the generated description [Guiguang, Minghai, Sicheng et al. (2018)].

Above mentioned methods first use CNN to encode the image to obtain visual features, and then decode to generate an image description combining text features. However, features of images and texts are unidirectional in these one-encode-one-decode models, leading no feedback attention both in encoding and decoding phase. When there are multiple objects in the image, each object is always bluntly expressed regardless of the primary and secondary, resulting in poor coupling of the generated words, insufficient

description of the object details, and lack of overall fluency. In addition, the image detection stage uses RCNN, which contains huge parameters and repeated calculations. Fortunately, this has been better optimized by Faster R-CNN [Wang, Shrivastava and Gupta (2017)], which is our choice in the phase of image detection.

Based on the above, this paper proposes Attention-based Feedback LSTM Network (AFLN) which combines the retrieval feedback and attention mechanism, resulting in overall fluency and full guidance on tracking and feedback on the details of the object. Compared with the current mainstream methods, our model is more accurate, and the results are verified experimentally.

3 Attention-based feedback LSTM network

3.1 Network architecture

AFLN effectively stitches the feedback LSTM network and CBAM together with the existing codec framework to make the feature encoding expression better in the coding stage and the effective feedback tracking in the decoding stage. Excellent performance of AFLN displays fluent description which contains detailed action and state of main object in the image well.

Based on the advantages of the codec framework without limiting the input and output modalities, many previous research methods are added basic spatial or channel attention within CNN [Chen, Zhang, Xiao et al. (2017)]. However, our design is more direct and effective, using the weight mapping algorithm and the feedback LSTM module to achieve feedback guidance of the previous word to the next word. Additionally, CBAM is applied for enhancing feature representation to AFLN, bringing rising accuracy of the image description.

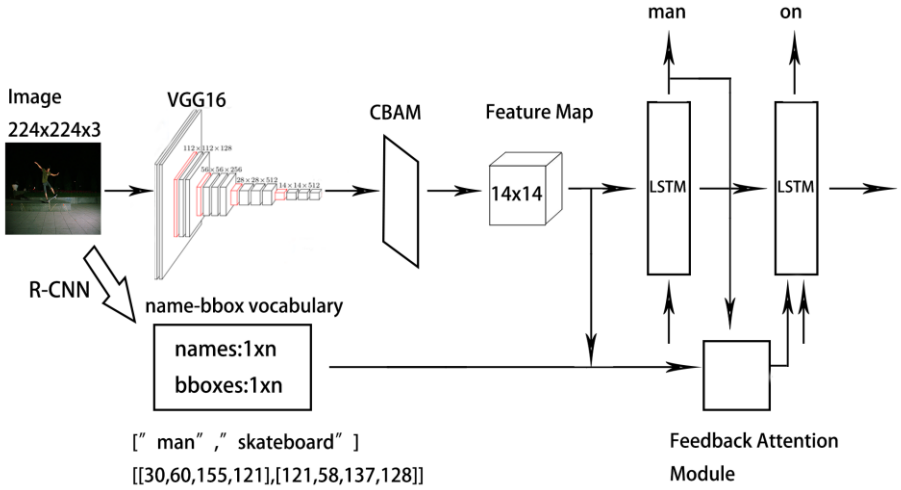


Figure 1: Attention-based feedback LSTM network

The feature map of the image is obtained by VGG 16 and CBAM, and the dictionary is obtained by Faster R-CNN. The generated word of LSTM finds the match in the

dictionary by each word, strengthens the characteristics of a region to provide attention-based feedback guidance on the next generated word.

3.2 Feedback attention module on LSTM

To address the problem of insufficient correlation between the generated LSTM unit weights and the gradient explosion, we design the Feedback Attention Module. Firstly, the weights of multiple LSTM units in the codec framework are the same, resulting in insufficient correlation between the generated preceding and following descriptors, so that the state and action of the image objects stay in a shallow description. Feedback attention module describes an object in tracking manner, which uses the related field mapping algorithm to personalize the correlation between each feedforward and the next generated word. This can accurately highlight the details of the image. Secondly, we use multiple identical LSTM unit superpositions to complete each descriptor generation, which results in gradient explosion or gradient disappearing due to the large number of layers when the description is too long [Wen, Xu, Yan et al. (2017)]. In the case of a limited number of LSTM layers, we additionally exploit gradient truncation to train a feedback attention module to achieve more accurate tracking of details of the object features while avoiding gradient explosion. Through the related field mapping algorithm to build quantitative correlation between each of the feedforward words and the next generated word, we can trace the object to describe details accurately.

3.2.1 Name-feature dictionary

We use Faster R-CNN for image preprocessing [Ren, He, Girshick et al. (2017)], so that AFLN can notice surrounding features of main object. The output of the Faster R-CNN is made into a dictionary consisting of the object name and the corresponding bounding box, called name-bbox dictionary. In order to balance accuracy and performance, we chose the top five bounding boxes to enter the dictionary.

When VGG 16 and CBAM get the feature map and combine the name-bbox dictionary to generate the name-feature dictionary, feature in the dictionary completes the correlation between features through the feedback mechanism in LSTM. Special correlation between the object and surrounding details can be better captured through the feedback attention module to generate a natural language description of the image.

3.2.2 Locate and feedback

AFLN adds new Locate-Feedback mechanism to the CNN-LSTM architecture. For the t^{th} moment of the i^{th} image, the word w_{it} generated by the LSTM is searched in the dictionary dic_i of the i^{th} image. If there is a s^{th} bounding box $bbox_{is}$ corresponding to w_{it}, then our model uses $bbox_{is}$ to process the feature map output by cnn. This process is called locate.

$$f_{locate}(w_{it}, dic_i) = \begin{cases} bbox_{is}, \exists s, s.t\ w_{it} = dic_{is} \\ O, otherwise \end{cases} \tag{1}$$

When object features corresponding to previous generated word are located, AFLN processes the feature map according to the obtained $bbox_{is}$. Then the model inputs

continuously the operation to the next time of the LSTM after the connection operation with the state h_t at the previous time until the LSTM network outputs the end character. The concat layer implements the splicing of input data. This layer has two identical parameters.

$$h_{t+1} = LSTM(concat([h_t, f_{attention}(bbox_{is}, feature\_map)])) \quad (2)$$

For example, the network structure described in Fig. 1, when the LSTM outputs the word "man", the dictionary obtained by Faster R-CNN matches the word "man". Based on the attention-based feedback module, the attention enhanced feature map can be obtained by the bbox of the man's area, which bringing rising description accuracy of the action or state of main object.

3.2.3 Related field mapping algorithm

Weights of the features of the bbox of different objects are different in the feedback. It is natural to design the weighting algorithm according to the degree of association. In order to realize the attention-based feedback module, if the feedback of each word directly recalculates the corresponding coding feature of CNN through the $bbox_{is}$ area of the object, the calculation amount will be larger. Attention mechanism and weighting algorithm are very instructive for the design of our mapping algorithm in the attention-based feedback module.

We propose a related field mapping algorithm for generating a name-feature dictionary. The following shows how to use $bbox_{is}$ to process the feature map, which is the process of generating a name-feature dictionary using the name-bbox dictionary. Our model selects VGG16 as the encoder and selects the $conv_{5\_3}$ layer as the output feature map with a spatial resolution of 14×14. We need to find a mapping between the $bbox_{is}$ area of the original image and an area of the feature map to make a partial selection of the feature map.

$$\text{f: bbox} \rightarrow \text{related field} \quad (3)$$

Naturally, we should choose the output area of the $bbox_{is}$ area after cnn. However, as the convolution depth increases, the resolution gradually becomes smaller, and the information of the entire image is gradually overlapped. Therefore, it can only be said that a certain area of the feature map is associated with the original $bbox_{is}$ area rather than a complete convolution relationship.

$$\text{related field}_{is} \neq \text{cnn}(bbox_{is}) \quad (4)$$

In order to solve the above problem, we choose to calculate the related field of the corresponding region from the overall feature map that has been calculated in the image. Therefore, we have designed a weighting algorithm to represent the relationship between the feature map on spatial and $bbox_{is}$ region of the original image. Specifically, if a point has no information about the point of the region other than the $bbox_{is}$, then its weight remains unchanged at 1, otherwise it should be reduced. Let the convolution kernel size be k×k. After one convolution, the number of points outside the related field is n_o, and the number of points inside is n_i, then the weight W of the result of the convolution operation centered on this position is

$$W = \frac{n_i}{k^2}, n_o + n_i = k^2 \tag{5}$$

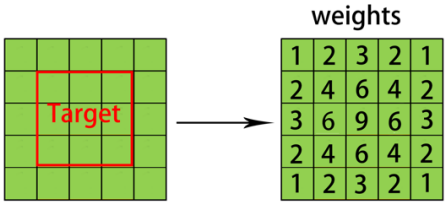


Figure 2: Weight algorithm. Taking a 3×3 bbox in a 5×5 image as an example, assuming that the convolution kernel is 3×3, the number in the right figure indicates the number of points where the area covered by the convolution kernel overlaps with the target area. Normally, it should be normalized so that the center (the number of overlapping points is 9) has a weight of 1, and the other positions have a weight of (0, 1)

AFLN utilizes the first 5 blocks of VGG 16. For the x^{th} convolutional layer, the above algorithm to calculate the weight matrix W_{conv_x} is used. For the y^{th} pooling layer, we calculate the weight matrix of the pooling layer according to the weight value as input.

$$W_{pool\_y} = pool_y(W_{conv}) \tag{6}$$

Finally, the process of simulating the VGG network using the following formula

$$W_{related\_field} = \prod_{y=1}^{block\_num} \prod_{x=1}^{conv\_num} pool_y(W_{block_y, conv\_x}) \tag{7}$$

Since getting the related field weights, $f_{attention}$ can be calculated

$$f_{attention}(bbox_{is}, feature\_map) = W_{related\_field} \cdot feature\_map \tag{8}$$

Based on the related field mapping algorithm for calculating the $bbox \rightarrow feature$, AFLN uses the obtained name-feature dictionary to calculate the feedback data.

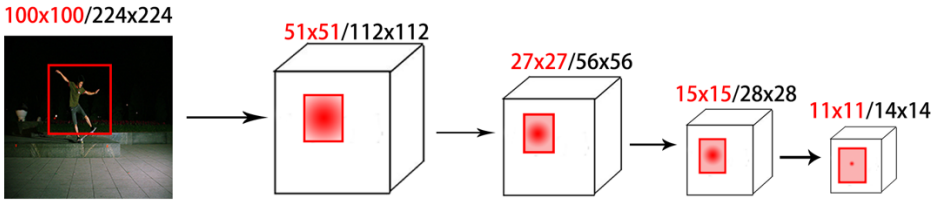


Figure 3: $Bbox \rightarrow feature$ related field mapping process. Mapping from bounding box to related field, proportion of related field increases while the weights is gradually decreasing

3.3 Convolutional block attention module

The advantage of the newly proposed Convolutional Block Attention Module (CBAM) in 2018 is that it accurately and expresses the image features while almost not increasing the

burden of model training. The module is precise and brisk, mainly used for attention guidance [Woo, Park, Lee et al. (2018)]. We innovatively add the CBAM after VGG 16 image convolutional extraction to enhance expression of the coding features, which achieves effective tracking match between the generated words and the corresponding object features.

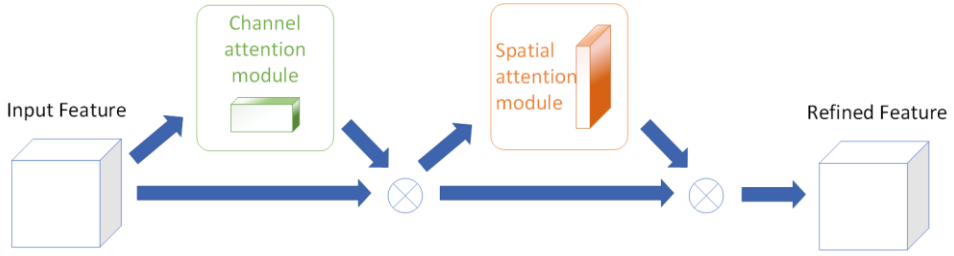


Figure 4: Convolutional Block Attention Module. The module has two sequential submodules: channel and spatial. The intermediate feature map is adaptively refined through CBAM at every convolutional block of deep network

4 Experiment

Extensive experiments on the COCO dataset show that image description generated from AFLN is natural and smooth, with excellent performance and improved scores. we will briefly introduce the experimental dataset and evaluation indicators in the followings, and then report our experimental results and comparative experimental verification.

4.1 Dataset

AFLN is experimenting with the Microsoft COCO dataset. There are 91 categories in the COCO dataset, which are less than the ImageNet and SUN categories, but there are many images in each class, which is beneficial to obtain more features of a particular scene. Compared to the PASCAL VOC, it has more classes and image. The COCO dataset provides five artificially annotated subtitles for each image, which contains 82,783 training images and 40,504 verification images [Lin, Maire, Belongie et al. (2014)]. Since most images contain multiple objects and important contextual information, creating a challenging test platform for image generation descriptions.

4.2 Evaluation indicators

The evaluation process is to compare the image generation description with the reference description in the corresponding dataset, and calculate the score according to the algorithm of the evaluation indicators. The higher the score, the better the machine translates. In this paper, we use the mainstream automatic evaluation indicators BLEU 4, CIDEr for double verification, and measure the accuracy of the generated description. The following is a brief introduction to the algorithmic ideas and features of the two evaluation indicators.

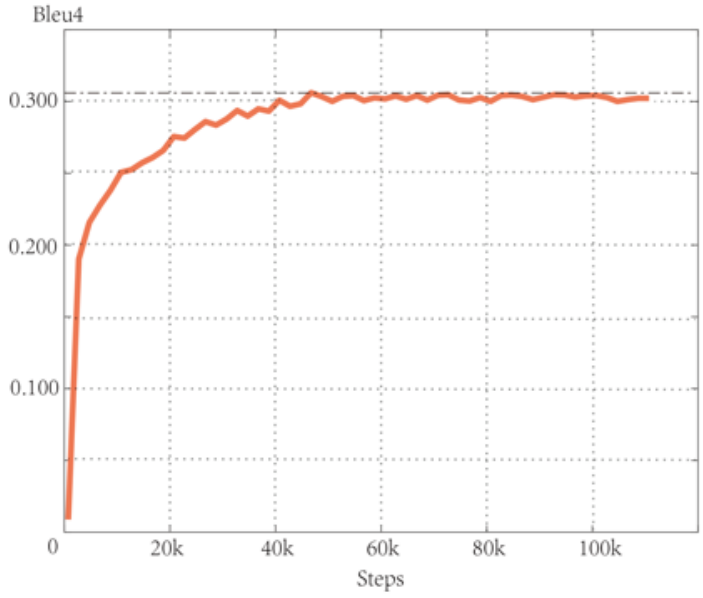
BLEU (Bilingual Evaluation Understudy) was first proposed by IBM, focusing on the accuracy-based similarity measure of machine translation and reference translation. The

core idea is to compare the degree of coincidence between words and phrases and reference translations in machine translations, and to introduce BP (Brevity Peanlty) to solve the tendency of the algorithm itself for short texts. BLEU is essentially the calculation of the cooccurrence frequency of two sentences. The value of calculation can measure the degree of agreement between the two sentences. The advantage of BLEU is that the granularity it considers is N-gram (N=1, 2, 3, 4) rather than words, considering longer matching information. The disadvantage of BLEU is that no matter what kind of n-gram granularity, it will be treated equally if it is matched (for example, the importance of verb matching should be intuitively greater than the article). BLEU is easy to fall into the trap of common words and short sentences, giving a higher score. Although this indicator has some obvious shortcomings, it has been shown to have a good correlation with human assessment [Agarwal and Lavie (2008)].

CIDEr makes up for the unfairness of BLEU's assessment of common words and essays [Vedantam, Zitnick and Parikh (2015)]. Compared to BLEU for machine translation only, CIDEr is widely used in image/video description, automatic summary evaluation and other fields. Based on the vector space model, CIDEr treats each sentence as a document, expresses it as a form of tf-idf vector, and then calculates the cosine similarity of the reference caption and the caption generated by the model as a score.

4.3 Result

Fig. 5 shows faster convergence speed and greater maximum of accuracy. The scores of Bleu 4 and CIDEr increase as the number of training increases Bleu 4 eventually converges around 0.301 (the limit is about 0.32), and CIDEr finally converges around 0.95 (the limit is about 0.96). In addition, our model also demonstrates excellent performance when describing images in detail, and Fig. 6 shows the specific description of the model.



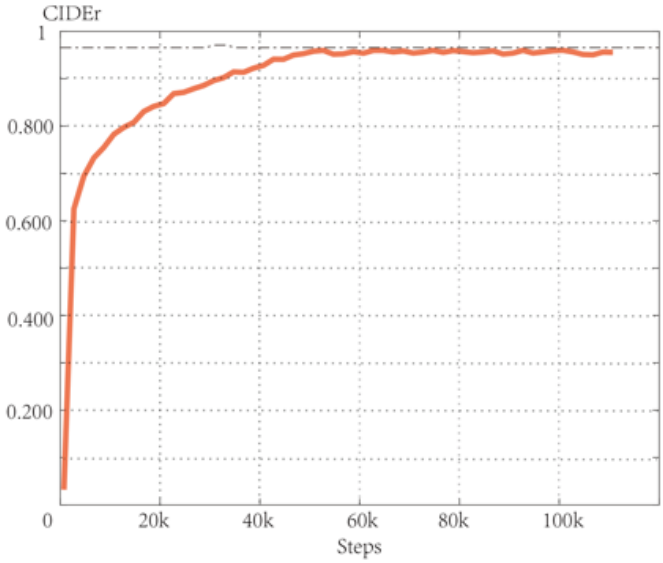


Figure 5: Growth process of BLEU 4 and CIDEr during training

Figure 6: Performance of experiment with on our network. As showed, our model pays attention on the remotes beside a cat(row1) and the swinging racket playing by a tennis player(row2)

4.4 Contrast experimental results

In order to confirm the role of the modules, we conducted four sets of comparative experiments including baseline, only the CBAM, only the attention-based feedback module and the combination of CBAM and attention-based feedback module. Fig. 7 shows the performance of the four models on Bleu x (x=1,2,3,4) and CIDEr. Both the attention-based feedback module and the CBAM strengthen the ability to describe images,

while the attention-based feedback module is slightly better than the CBAM.

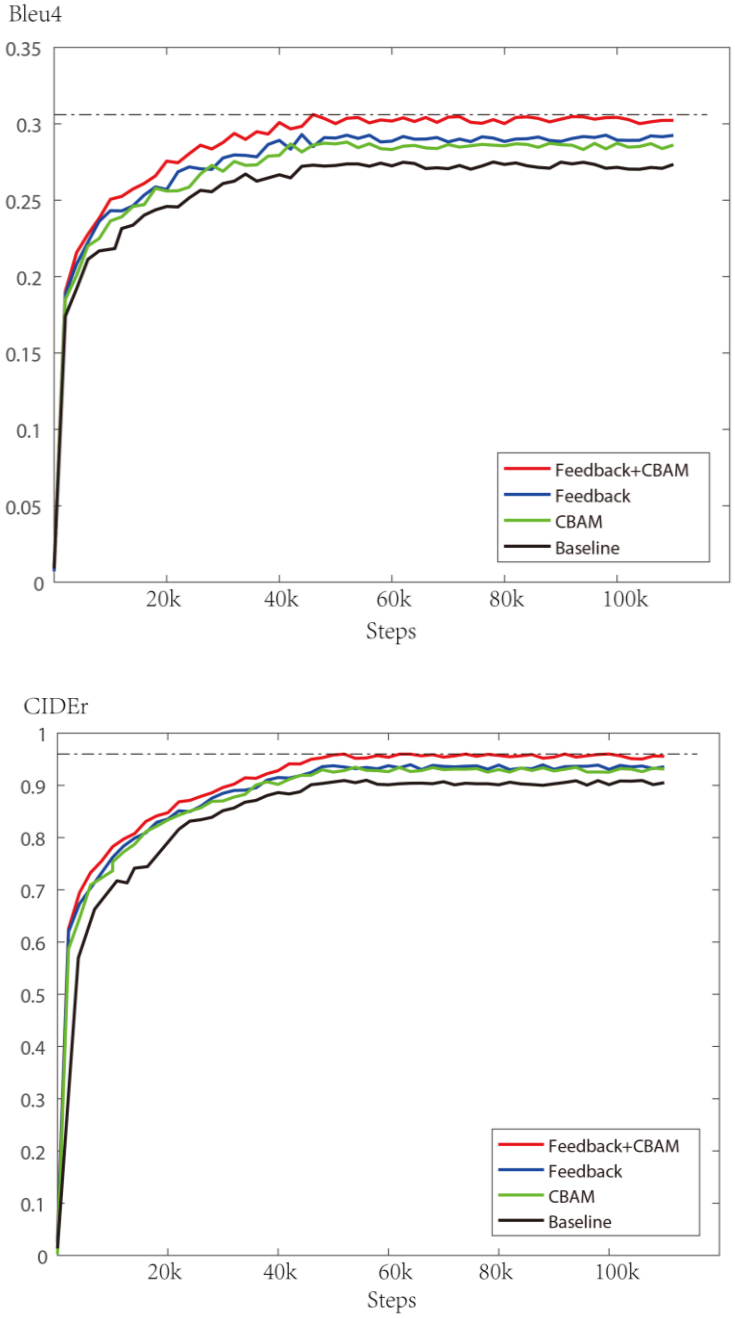


Figure 7: Comparison of BLEU 4 and CIDEr growth curves for the model. The BLEU 4 score increased from 0.291 to 0.301, and the CIDEr score increased from 0.912 to 0.952

Table 1: Comparison of Bleu x (x=1,2,3,4) and CIDEr evaluation scores of four models

Model	Bleu 1	Bleu 2	Bleu 3	Bleu 4	CIDEr
Neuraltalk2	0.716	0.543	0.393	0.291	0.912
CBAM only	0.718	0.545	0.4	0.283	0.925
Feedback only	0.721	0.551	0.41	0.288	0.93
CBAM+Feedback	**0.724**	**0.554**	**0.42**	**0.301**	**0.952**

See Tab. 1, the results show that each score grows with the network changed from that contains single module such as Neuraltalk2, CBAM, Feedback LSTM module to that contains both CBAM and Feedback LSTM module. The BLEU 4 score increased from 0.291 to 0.301, and the CIDEr score increased from 0.912 to 0.952.

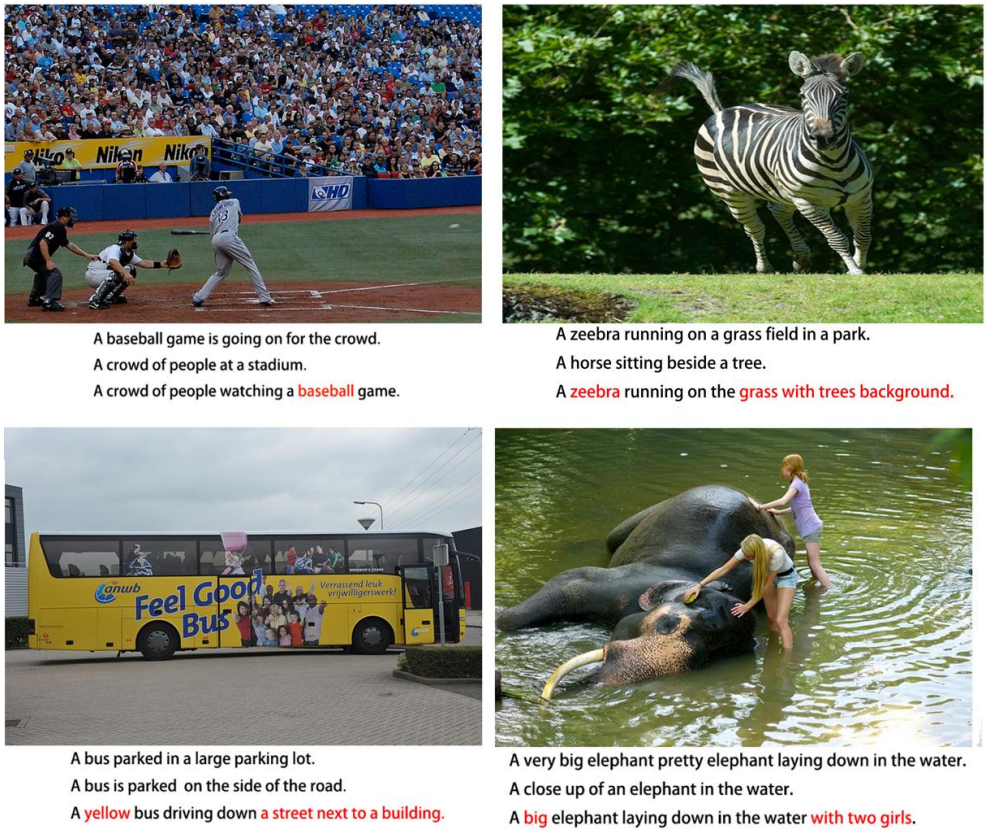


Figure 8: Comparison of descriptions of several models. From top to bottom, the captions are given by ground truth, neuraltalk2, and our model, which contains more details of main object. Our model gives more detailed and explicit description, especially the details around the key object in the picture. For example, sentence of the second image notices the court which is the play venue, while the third not only gives a description of bus, but also notes the street and building surrounding the bus

4.5 Result analysis

As experiment results show, the feedback LSTM network and CBAM play significant roles. The scores of Bleu 4 and CIDEr show that the use of both can speed up the training speed and enhance the value of the final convergence. Shown as image de-scriptions, AFLN plays excellent performance of high-level semantics and is more sensitive to the spatial position, noticing diverse surround details based on previous key description of the object.

5 Conclusion

In this paper, we present Feedback LSTM Network Based on Attention, an end-to-end network for image description generator. The network combines the retrieval feedback and attention mechanism, resulting in overall fluency and full guidance on tracking and feedback details of the object. We stitch the CBAM and the feedback attention module together with the existing codec solution to make the feature expression better in the coding stage and effective feedback tracking in the decoding stage. Our excel-lent network describes the details and actions of the main object well without losing the fluent description of the image.

AFLN is more accurate compared to current mainstream methods, with both quali-ty and quantity verified. Our experiments on the coco dataset show the accuracy of the model and the fluency of the description. The BLEU 4 score increased from 0.291 to 0.301, and the CIDEr score increased from 0.912 to 0.952.

Acknowledgement: This research study is supported by the National Natural Science Foundation of China (No. 61672108).

References

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M. et al. (2018): Bottom-up and top-down attention for image captioning and vqa. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6077-6086.

Agarwal, A.; Lavie, A. (2008): Meteor, m-bleu and m-ter: evaluation metrics for high-correlation with human rankings of machine translation output. *Proceedings of the Third Workshop on Statistical Machine Translation*, pp. 115-118.

Carrara, F.; Esuli, A.; Fagni, T.; Falchi, F.; Moreo Fernández, A. (2018): Picture it in your mind: generating high level visual representations from textual descriptions. *Information Retrieval Journal*, vol. 21, no. 2-3, pp. 208-229.

Chen, X.; Zitnick, C. L. (2015): Mind's eye: learning a recurrent visual representation for image caption generation. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2422-2431.

Chen, L.; Zhang, H.; Xiao, J.; Nie, L.; Shao, J. et al. (2017): Sca-cnn: spatial and channel-wise attention in convolutional networks for image captioning. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6298-6306.

Donahue, J.; Hendricks, L. A.; Guadarrama, S.; Rohrbach, M.; Venugopalan, S. et

al. (2017): Long-term recurrent convolutional networks for visual recognition and description. *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 39, no. 4, pp. 677-691.

Graves, A. (2008): Supervised sequence labelling with recurrent neural networks. *Studies in Computational Intelligence*, pp. 385.

Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. (2014): Rich feature hierarchies for accurate object detection and semantic segmentation. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580-587.

Guiguang, D.; Minghai, C.; Sicheng, Z.; Hui, C.; Jungong, H. et al. (2018): Neural image caption generation with weighted training and reference. *Cognitive Computation*. pp. 1-15.

Hochreiter, S.; Schmidhuber, J. (1997): Long short-term memory. *Neural Computation*, vol. 9, no. 8, pp. 1735-1780.

Huang, Y.; Wu, Q.; Wang, L. (2018): Learning semantic concepts and order for image and sentence matching. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6163-6171.

Kinghorn, P.; Zhang, L.; Shao, L. (2018): A region-based image caption generator with refined descriptions. *Neurocomputing*, vol. 272, pp. 416-424.

Jia, X.; Gavves, E.; Fernando, B.; Tuytelaars, T. (2015): Guiding the long-short term memory model for image caption generation. *IEEE International Conference on Computer Vision*, pp. 2407-2415.

Karpathy, A.; Li, F. F. (2017): Deep visual-semantic alignments for generating image descriptions. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 39, no. 4, pp. 664-676.

Lin, T. Y.; Maire, M.; Belongie, S.; Bourdev, L.; Girshick, R. et al. (2014): Microsoft coco: common objects in context. *European Conference on Computer Vision*, pp. 740-755.

Mao, J.; Xu, W.; Yang, Y.; Wang, J.; Huang, Z. et al. (2014): Deep captioning with multimodal recurrent neural networks (M-RNN). *arXiv.* cs.cv. 1412.6632.

Ren, S.; He, K.; Girshick, R.; Jian, S. (2017): Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137-1149.

Mingxing, Z.; Yang, Y.; Hanwang, Z.; Yanli, J.; Tao, S. H. et al. (2019): More is better: precise and detailed image captioning using online positive recall and missing concepts mining. *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 32-44.

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

Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. (2017): Show and tell: lessons learned from the 2015 mscoco image captioning challenge. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 39, no. 4, pp. 652-663.

Vedantam, R.; Zitnick, C. L.; Parikh, D. (2015): CIDEr: consensus-based image

description evaluation. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4566-4575.

Wu, Y.; Schuster, M.; Chen, Z.; Le, Q. V.; Norouzi, M. et al. (2016): Google's neural machine translation system: bridging the gap between human and machine translation. *arXiv*. cs.cv. 1609.08144.

Wang, X.; Shrivastava, A.; Gupta, A. (2017): A-fast-RCNN: hard positive generation via adversary for object detection. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3039-3048.

Wu, Q.; Shen, C.; Hengel, A. V. D.; Wang, P.; Dick, A. (2018): Image captioning and visual question answering based on attributes and external knowledge. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 40, no. 6, pp. 1367-1381.

Wang, C.; Yang, H.; Bartz, C.; Meinel, C. (2016): Image captioning with deep bidirectional lstms. *Proceedings of the 24th ACM International Conference on Multimedia*, pp. 988-997.

Woo, S.; Park, J.; Lee, J. Y.; Kweon, I. S. (2018): CBAM: convolutional block attention module. *The European Conference on Computer Vision*, pp. 3-19.

Wen, W.; Xu, C.; Yan, F.; Wu, C.; Wang, Y. et al. (2017): Terngrad: ternary gradients to reduce communication in distributed deep learning. *Neural Information Processing Systems*.

Zeyu, X.; Qiangqiang, S.; Yijie, W; Chenyang, Z. (2018): Paragraph vector representation based on word to vector and CNN learning. *Computers, Materials & Continua*, vol. 55, no. 2, pp. 213-227.