ISSN: 2688-8645

Application of Prior Mask Attentional Mechanism in Image Question Answering

Ziwei Liu

Electronic information, Southwest Minzu University, Chengdu, 61000, China

Abstract

Video Q&A is one of the research hotspots in the field of deep learning, which is widely used in security and advertising systems. In the framework of attentional mechanism, a model of preempirical MASK attentional mechanism was established. The Faster R-CNN model was used to extract key frames and object labels in videos, and three kinds of attention al weights were applied to them and the text features of questions. The MASK was used to MASK irrelevant answers, so as to enhance the interpretability of the model. Experimental results show that the accuracy of this model in video question answering task reaches 62%, and compared with VQA+, SA+ and other video question answering models, this model has faster prediction speed and better prediction effect.

Keywords

VQA; Computer vision; Natural language processing; Attention mechanism; MASK model.

1. Introduction

With the improvement of communication technology, video has become the biggest One of the information carriers. "A picture is worth a thousand words" vividly illustrates images As a medium of information, and video carries more information. So, how do you make a computer understand what's in a video and become a scholar Our research hotspot. Visual Question Answering Visual QA) [1] task is the parent task of video comprehension, which is simply described Given an image and a problem related to the image content, The computer answers the questions based on the understanding of the image content and the questions Case. The sub-tasks of image question answering include image pattern recognition and natural language Word processing.

Video Question Answering, Video QA) is more challenging. At present, video question and answer research development is relatively slow, one of the important The reason is the cost of collating data sets, including collection and annotation High, and the video related processing technology is not mature enough. Video mining The consequence of not intercepting core pieces of content very well during the set process is that On the one hand, if the video time is too long, it will lead to the increase of irrelevant information, which is difficult Attract enough people to answer and annotate the questions, on the other hand .Too short will lead to insufficient information, resulting in the misunderstanding of respondents.

Based on the existing attention mechanism framework, this paper proposes the prior MASK attention mechanism model. The key frames of the video were extracted and the upward attention mechanism Faster R-CNN model was used to obtain the features of the key frames and the object labels in the key frames, and the features and objects were labeled In order to improve the accuracy of video q&A, three kinds of attention weights were applied to the tabs and the question text, and the prior MASK was used to shield irrelevant answers.

2. Related Work

Based on the existing attention mechanism framework, this paper proposes the prior MASK attention mechanism model. The key frames of the video were extracted and the upward attention mechanism Faster R-CNN model was used to obtain the features of the key frames and the object labels in the key frames, and the features and objects were labeled In order to improve the accuracy of video q&A, three kinds of attention weights were applied to the tabs and the question text, and the prior MASK was used to shield irrelevant answers. At present, there are few researches on video question answering in academia and industry, but there are many researches on its parent task, image question answering [1], and great progress has been made. From the perspective of model, image q&A task mainly focuses on the fusion of image features and text features, so as to achieve end-to-end training. In the field of image processing, with the continuous development of network structure, the use of convolutional neural network (CNN) for image feature expression has become the mainstream. In addition, natural language processing has also developed rapidly. From the early word bag model [2] and Word2vec [3] to the recent natural language processing pre-training models Bert [4] and XLNet [5], computers can extract grammatical and semantic features to abstractly extract text features. In recent years, how to effectively fuse text features with image features has become a hot research topic in the image question answering task, which combines image processing with natural language processing. In 2015, ZHOU et al. [6] proposed the baseline of image question and answer, introduced the iBOWIMG model, as shown in FIG. 1. VGGNet [7] network was used to extract image features, and word bag coding was implemented for questions and answers [2], and then image features were combined with question features. The probability of each answer is output through the classification layer, and the error calculation with the real answer is carried out to realize the gradient return, to achieve the purpose of training.

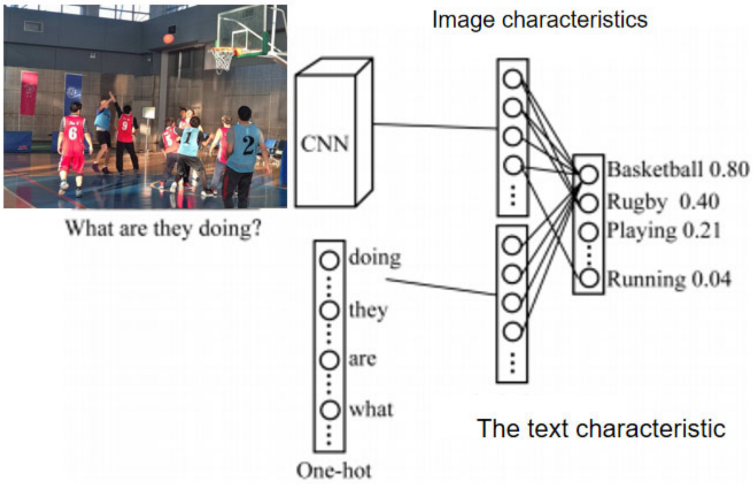


Figure 1. IBOWIMG model structure

Literature [1] using neural network for image feature extraction, using cycle LSTM neural networks [8] implementation issues text feature extraction, the characteristics of two phase joining together so as to achieve the purpose of the training, in addition, the use of language model from COCO [9] automatically generated problems in image annotation, also stipulates that the answer must be a word, Contains four themes: object, quantity, color, and location, but it supports only one question and the answer can only be one word, which has no practical significance. Literature [10] proposed the attention mechanism and applied it to the field of image question answering. In literature [11], a convolution kernel is formed after feature expression of the problem and convolution operation is carried out with the image to obtain a regional concern map in the image space, so as to extract features more accurately. Literature [12] proposed the attention model method combining top-down and bottom-up, and applied it to visual scene understanding and visual question answering system. Among them, the bottom-up attention model (generally Faster R-CNN [13]) is used to extract regions of interest in images and obtain object features. The top-down attention model is used to learn the corresponding weights of features to realize the deep understanding of visual images. The method proved its effectiveness by winning first place in the 2017 VQA Challenge. The above are all related research of image question answering. It can be seen that image question answering system has made remarkable progress. Literature [14] proposed the C3D model Type [15] was combined with features extracted from ResNet model [16] and Word2vec extracted from Glove [17] again to conduct end-to-end training of attention mechanism. Since then, video q&A has gradually become the focus of researchers. As shown in Figure 2, the video contains time series, so the q&A task is more difficult than the image.

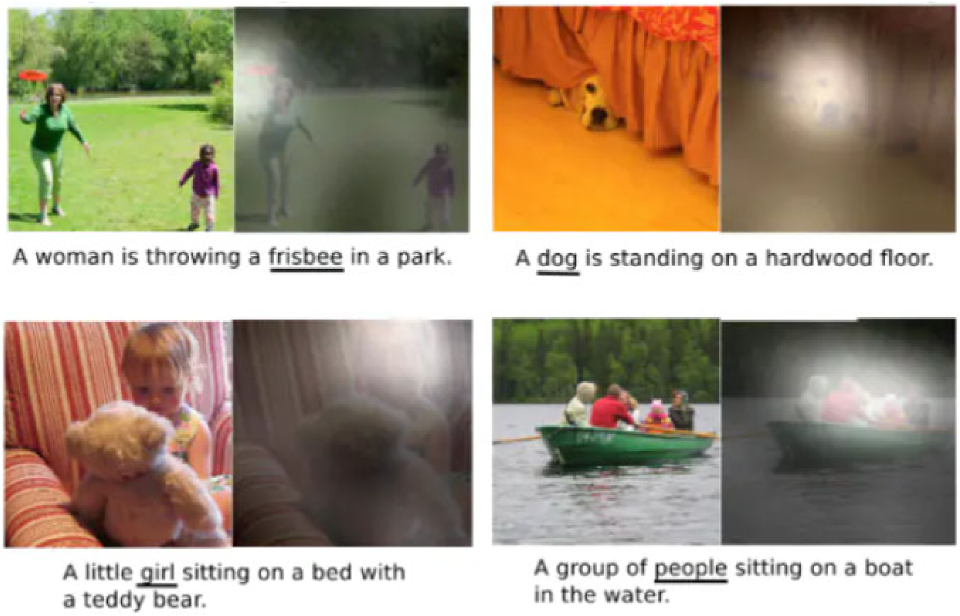

A woman is throwing a frisbee in a park.
A dog is standing on a hardwood floor.
A little girl sitting on a bed with a teddy bear.
A group of people sitting on a boat in the water.

Figure 2. Picture q&A and video Q&A

3. Methodology

At present, most video q&A schemes do not consider the correlation between text features and video features, but such correlation features in the answer.

How to make the model capture both the question and the interest point of the video is particularly critical. Therefore, this paper proposes a VQA model based on the preempirical MASK attention mechanism for general video question answering tasks, and the model structure is shown in Figure 3. The model input is N question Q, a video, question prior information prior and video label attr. After Word2vec, the question text is weighted with video features and video label attr for attention mechanism. Finally, the network output is multiplied with prior information between the network output period, and the result is called the prior MASK. The final output of the network is the predicted answers to N questions.

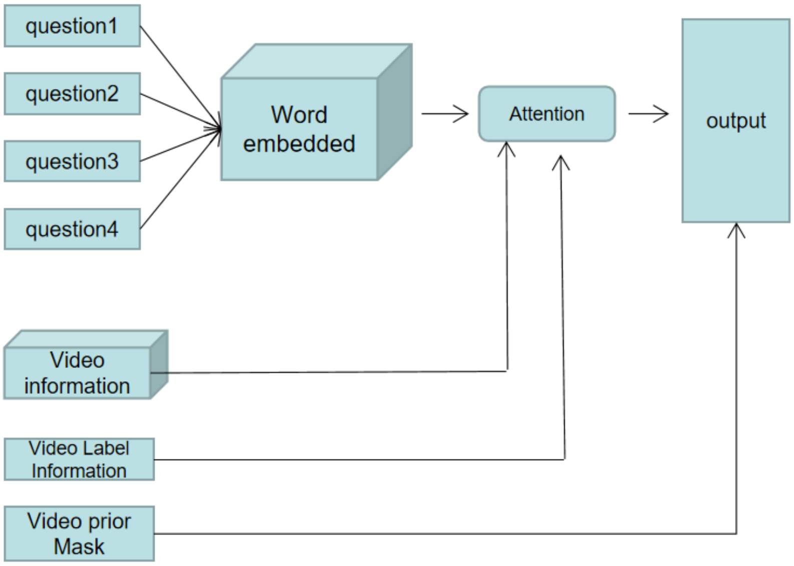


Figure 3. Picture Prior Mask model

As can be seen from Figure 3, the model input is N text questions. Since the public data set is that 1 video corresponds to multiple questions (for example, 1 video corresponds to 5 related questions in the ZJB-VQA data set), better generalization features can be achieved through this multi-input learning method.

In problem processing, the text is unified into the same length. According to the results in Table 1, the average length of the text is about 8 words, and the longest is 18 words. The model performs better when the text is 14 words as the input length of the question. In video processing, there are many useless frames in video, so how to deal with redundant frames becomes the focus of the task. If all frames are used for training, the training time will be greatly increased and the machine is required to be high. Therefore, FFMPEG, an open source audio and video processing tool, is used for video key frame extraction in this paper. In this paper, the number of key frames is set as Lv, and the extracted frames are supplemented when the extracted frames are less than Lv, and the extra Lv frames are compressed into one frame. Finally, each video is processed into an IMAGE set of L-frame, and Faster R-CNN is used as a feature extraction tool. As shown in Figure 6, Faster R-CNN is a target detection model, which can not only detect the target, but also mark the category to which the target belongs and the coordinate position in the image with the border.

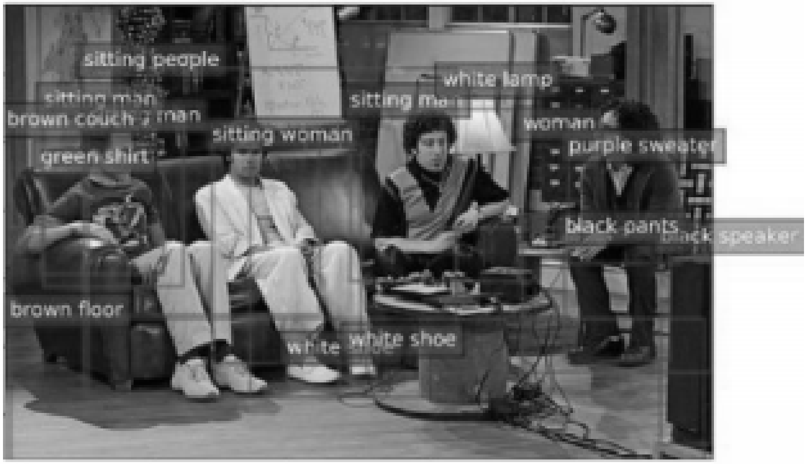


Figure 4. Fast R-CNN model detection example

In this paper, the Faster R-CNN model is adopted, which takes the output of the last layer of the network as the feature and uses an IoU threshold for screening. For each region I, Vi represents the feature of this region. In the video question and answer task, the dimension of feature is M (M is 2 048 in the pre-training model). The first P objects with high confidence in this region are given. Therefore, for a frame in the video, the dimension of the output of Faster R-CNN is (PM), and the features of dimension (Lv PM) and P object labels with maximum confidence in each frame are extracted by feature extraction for each video. For Lv frames, there are a total of Lv´P labels, and w labels with the highest frequency are selected as the final input label of the model. The pre-training ResNET-101 CNN model in the Faster R-CNN used in the scheme is trained based on ImageNet. The combination of top-down and bottom-up attention model based on Faster R-CNN was used for feature extraction, and the threshold of confidence was set as 0.2, which could make the Faster R-CNN get more tags with higher reliability. At this time, the method obtained (LPM) dimension size characteristics and W video labels. For the attention module and network output module in the network structure, this paper proposes 3 attention models and 3 prior maskattention models, namely, grain-attention, Attr-block-attention and time-spatial-attention. The three capture the relationship between the video and the question text from different angles.

3.1. Temporal-attention Model

As shown in Figure 5, the grain-attention model weighted the attention of the problem features and the video features so that the model could capture the key points in the video according to the problems. The hadamard dot product was performed between the global average sampling results of N problems and the video features as the input of the next network. For the treatment of the problem features, this paper uses double - layer dual - directional LSTM network representation.

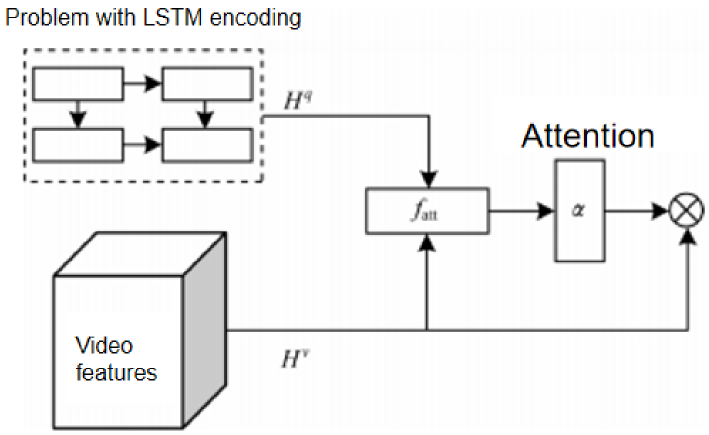


Figure 5. Picture Prior Mask model

3.2. Attr-block-attention Model

The grain-attention model focuses on the relationship between video features and problems, while the Attr-block-attention model focuses on problems And the attention between tags extracted through Faster R-CNN in the video. Consider the example "What's on the table?" , Faster R-CNN extracted the "table" label and easily weighted the label with the problem. The structure of the atr-block-attention model is shown in Figure 6. Using atr-block-attention allows the model to find interesting points among the visual frequency tags and N questions. Learn more important information from the model.

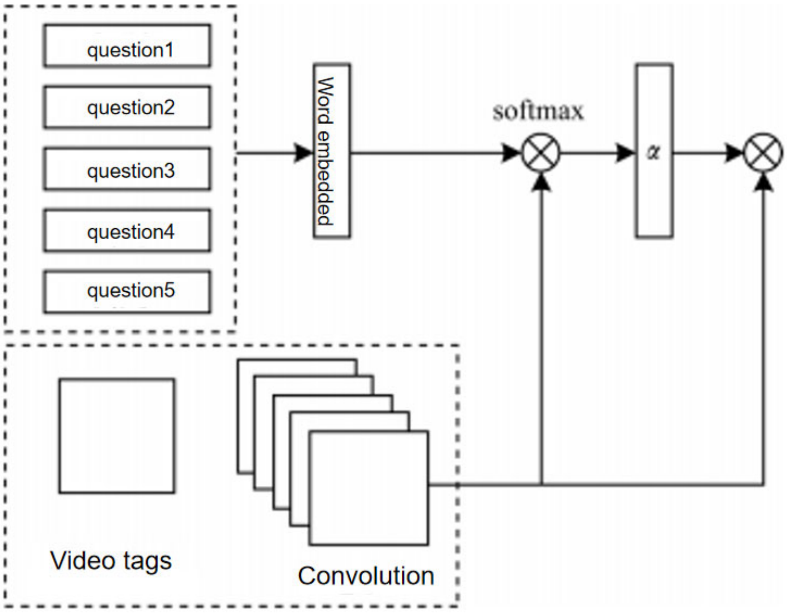


Figure 6. attr-block-attention model

3.3. A Priori Mask

There are various output answers of the network, but the answer space for a certain kind of question is limited, as shown in Figure 3, zJB-VQA number

The problem types of data sets mainly focus on how many, color, doing, yes/not and WHERE. For example, for the type of yes/ NOT problem, the answer is only yes or not, and no other answer is possible. Therefore, this paper uses prior MASK to control the answer within a fixed output space, so as to improve the prediction performance of the network.

3.4. Domestic and Foreign Methods

At present, the research methods of video question answering by domestic and foreign scholars are mainly divided into joint embedding, video description and attention mechanismAs follows:

1.Joint embedding is the most common method in video q&A task. It uses convolutional neural network to extract video features, and at the same time uses recursive neural network to extract feature expression of question text. Then, video features and question features are splicing together and directly input into the model to generate the probability of each answer. For video features, currently most of the pre-training training models in ImageNet [18] (such as VGGNet [19], ResNet [16] and GoogleNet [20]) are used to extract image features. For problem text, LSTM is mostly usedAnd GRU [21] to extract text features. Re-watching and re-watcher are put forward in literature [22] to imitate the human behavior of constantly watching videos when reading questions, and then the two mechanisms are combined into unforgettable -watche model.

2.Video description method, which converts a video into a sentence described in natural language. The method converts the video into text and uses natural language processing to get the answer to the question. Literature [23] proposed a Layered Memory Network,LMN) model, which extracts words and sentences from movie or TV subtitles, uses LMN to generate video expression, and then converts questions and videos into text through semantic matching to generate answers. In literature [24], pre-trained Faster R-CNN [13] model was used to first obtain the target and position attribute information in each frame of image, learn the subtitle information in the video to obtain relevant visual labels, and then input the acquired regional features (target and position attributes), video features and question text features into the model. To get the answer to the question.

3.The attention mechanism model [25], which was first proposed in machine translation [26] task, recognizes the weight of different parts of sentences in the recurrent neural network, thus making the neural network pay attention to different words. Attention mechanism has achieved good results in machine translation tasks and has gradually become a research hotspot in the field of video question answering. Literature [27] proposed a Joint Sequence Fusion (JSFusion) model. Joint Semantic Tensor (JST) uses dense Hadamard products between multi-module sequences to generate 3D tensors, and then uses the self-attention mechanism of learning to highlight 3D matching vectors. Convolutional Hierarchical Decoder (CHD) discovers the local alignment fraction of 3D tensors generated by JST module through convolution and convolution gate module. As a universal method, the model can be applied to a variety of multimodal sequence data pairs. It is also used for video retrieval, video q&A, multiple choice and blank filling tasks. Literature [28] uses a dual attention mechanism that integrates video features and problem features to solve visionAsk questions frequently. In literature [29], two Attention mechanisms, Appearance and Motion, were used to strengthen the relationship between problems and videos, and a variant of RNN, AMU (Attention Memory Unit), was used to further deal with problems, so as to improve the predictive performance of the model.

4. Experimental Results and Analysis

1) VQA+ model, which is an extension of image question answering method [1], adopts ResNet [16] network for feature extraction, uses LSTM to complete feature extraction, and then inputs into the classification network to get the question answer. The VQA+ model is the basic scheme adopted by the 2nd and 3rd place winners of the Jikang Cup.

2) SA+ model [31], which extracts word features from question text through LSTM, and then combines them with features of video frames and inputs them intoClassification network, get the answer to the question.

3) R-ANL model [28], which is an attribute enhanced attentional network learning representation method, adopts multi-step reasoning and attribute enhanced attentionA combination of forces to get the answer.

4) DLAN model [32], which adopts a hierarchical approach to solve video q&A problems and obtains problems according to the importance of the problem Video expression to answer questions. Since THE LSTM network cannot be parallel, Transformer model [33] changes the LSTM network of traditional attention mechanism into a parallelized matrix operation, and then carries out attention weighting to achieve the goal of training timing data. Due to the rapid development of Gpus, the network parameters of Transformer can be large without extending the training iteration time of the network. In addition, all experiments in this paper have multiple questions for one video. During the experiment, it was found that multi-input training can speed up the training and improve the accuracy. The comparison results of multi-input training and individual training are shown in Table 2. It can be seen that compared with individual training mode, multi-input training mode is used in ZJB-VQA data set to accelerate training iteration and improve accuracy.

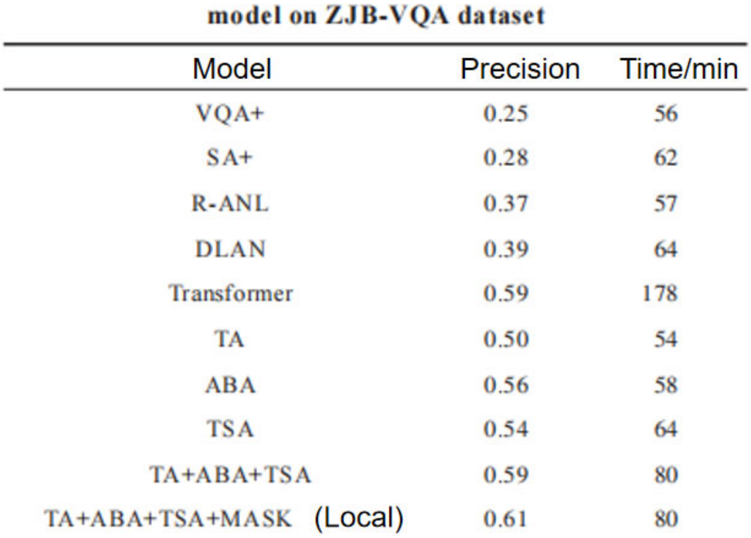
4.1. Parameter Setting and Experimental Environment

For ZJB-VQA data set, the number of text questions is set to 5 in this paper, which is consistent with the training set of ZJB-VQA official data set. At the same time, the number of Lv key frames of visual frequency is set to 40. The image feature size obtained by the Faster R-CNN is 2 048, parameter P is set to 36, and the labeling number w obtained through the Faster R-CNN is96. Lq of sentence length is 14, O of hidden layer of network is 256, and L 'of frame used after downsampling is 16. The experimental configuration in this paper adopts 64 GB memory, one GTX 1080Ti display card and I7 CPU.

4.2. Model Comparison Experiment

VQA+ model and SA+ model are the benchmark models of video question answering, while R-ANL model, DLAN model and Transformer model adopt attention mechanism. According to the above three models, grain-attention (TA), Atr-block-attention (ABA), time spatial-attention (TSA) and prior MASK, this paper conducted experiments respectively. The results are shown in Table 1.

Table 1. Performance comparison results of each model on ZJB-VQA dataset



model on ZJB-VQA dataset

Model	Precision	Time/min
VQA+	0.25	56
SA+	0.28	62
R-ANL	0.37	57
DLAN	0.39	64
Transformer	0.59	178
TA	0.50	54
ABA	0.56	58
TSA	0.54	64
TA+ABA+TSA	0.59	80
TA+ABA+TSA+MASK (Local)	0.61	80

As can be seen from Table 3:

1) The generalization ability of models using attentional mechanism is better than that without attentional mechanism Using models of attentional mechanisms.

2) The three attention mechanism components (TA, ABA, TSA) proposed in this paper greatly improved the model accuracy, and the model accuracy of VQA+ inattentional mechanism was improved to more than 0.50. When the three attention mechanisms were combined, the accuracy of 0.59 could be obtained. Although the accuracy of Transformer model is also 0.59, it greatly increases the training cost and the training time is about twice that of the model in this paper.

3) The prior MASK proposed in this paper can further improve the model accuracy. Since the prior MASK only processes the output of the last layer of the network, it does not increase the training cost.

Figure 7 shows the training accuracy and loss value results of the model in this paper on the ZJB-VQA data set. Among them, the training set and verification set are divided in a ratio of 8：2. As can be seen from Figure 10, when the number of training iterations reaches 25, the convergence of the training set begins to fluctuate and the loss value drops gently, indicating that the model has reached saturation. The accuracy of validation set increases slowly, and the model sets an early stop when it reaches 30 iterations, so as to ensure the generalization performance of the model and avoid the occurrence of fitting problems.

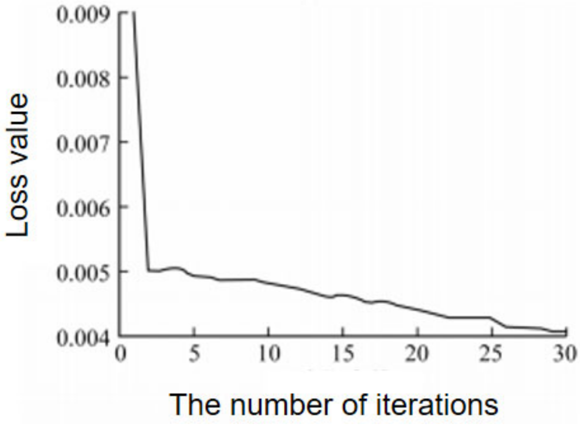
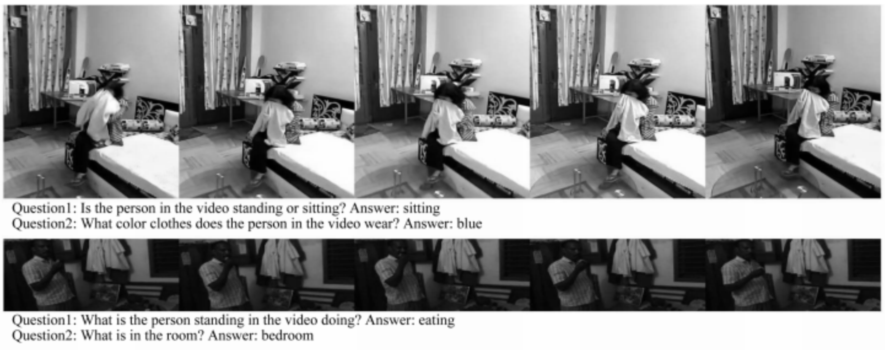


Figure 7. Experimental results of the proposed model on ZJB-VQA dataset

4.3. Model Comparison Experiment

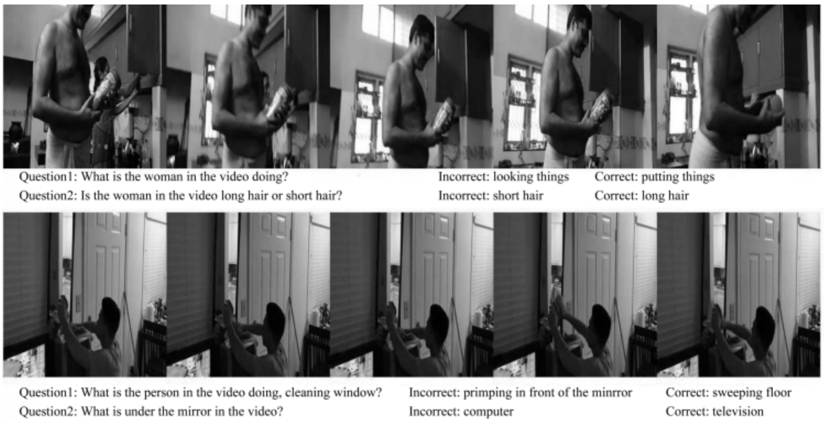
According to the video content of the ZJB-VQA data set shown in an example, the questions are proposed and the answers are predicted by using the model in this paper, as shown in Figure 8 and 9. Figure 8 Is a video of a woman sitting by the bed. The question Is "Is the person in the video standing or sitting?" What color clothes do the person in the video wear? , the predicted answers of the model in this paper are "sitting" and "blue" respectively, indicating that the predicted results of the model are consistent with the video scene. However, the model in this paper still has shortcomings. In Figure 9, in the first video, a man holds a bottle and a woman puts things into the cabinet. The question is "What is the woman in the video doing? The model predicted "looking things" and the answer was "Putting things", so it can be seen that the model has limitations in predicting some scenes involving complex actions.Along with the movement of the target, the sink node timely notifies the sensor nodes in the relevant detection area to join in the process of target tracking. Figure 1 is the flow chart of the moving target tracking process.



Question1: Is the person in the video standing or sitting? Answer: sitting
Question2: What color clothes does the person in the video wear? Answer: blue

Question1: What is the person standing in the video doing? Answer: eating
Question2: What is in the room? Answer: bedroom

Figure. 8 Example of correct model prediction



Question1: What is the woman in the video doing? Incorrect: looking things Correct: putting things
Question2: Is the woman in the video long hair or short hair? Incorrect: short hair Correct: long hair

Question1: What is the person in the video doing, cleaning window? Incorrect: primping in front of the minrror Correct: sweeping floor
Question2: What is under the mirror in the video? Incorrect: computer Correct: television

Figure 9. Example of incorrect model prediction

5. Conclusion

In this paper, a model of prior MASK attentional power machine is constructed for video question answering task. Three kinds of attentional power are used to focus on the visual frequency and the interesting points of the question from different angles. Step by step to improve the mold performance. Experimental results show that compared with VQA+ and SA+ models, this model has higher accuracy and faster speed. The model in this paper won the champion of video question and answer group in the "Zhijiang Cup" ARTIFICIAL intelligence Competition in 2018, which verified its effectiveness. Later, a deeper network (such as the ResNET-152 model) will be used to extract the features of the video keyframes, or natural language pre-training models such as BERT and XLNet are used to extract problem characteristics, so as to improve the pre-measurement speed and accuracy of the vQ model.

References

[1] ANTOL S,AGRAWAL A,LU J,et al. VQA:visual question answering[J]. International Journal of Computer Vision,2017,123(1):4-31.

[2] TURNEY P,PANTEL P. From frequency to meaning: vector space models of semantics[J].Journal of Artificial Intelligence Research,2010,(1):141-188.

[3] MIKOLOV T,CHEN K,CORRADO G,et al. Efficient estimation of word representations in vector space [EB/OL].[2019-11-10]. http://export..arxiv. org/pdf/ 1301.3781.

[4] DEVLIN J,CHANG M W,LEE K,et al. BERT:pretraining of deep bidirectional transformers for language understanding[EB/OL.[2019-11-10. https://tooob. com/api/objs/read/noteid/28717995/.

[5] YANG Zhilin,DAI Zihang,YANGYiming,etal.XLNet:generalized autoregressive pretraining for language understanding[EB/OL].[2019-11-10]. https://arxiv. org/ abs/1906. 08237.

[6] ZHOU B L,TIAN Y D,SUKHBAATAR S,et al. Simple baseline for visual question answering[EB/OL].[2019-11-10]. http://de. arxiv. org/pdf/1512. 02167.

[7] SIMONYAN K,ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL].[2019- 11-10]. https://arxiv. org/abs/1409. 1556.

[8] HOCHREITER S,SCHMIDHUBER J.Long short-term memory[J]. Neural Computation,1997,9(8):1735-1780.

[9] LIN T Y,MAIRE M,BELONGIE S,et al. Microsoft coco: common objects in context[C]//Proceedings of European Conference on Computer Vision. Berlin,Germany: [Springer,2014:740-755.

[10] VASWANI A,SHAZEER N,PARMAR N,et al. Attention is all you need[EB/OL]. [2019-11-10]. https://arxiv. org/abs/1706. 03762.

[11] XU H,SAENKO K. Ask,attend and answer:exploring question-guided spatial attention for visual question answering[C]//Proceedings of European Conference on Computer Vision. Berlin,Germany:Springer,2016:156-163.

[12] ANDERSON P,HE X,BUEHLER C,et al. Bottom-up and top-down attention for image captioning and visual question answering[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C. ,USA:IEEE Press,2018:6077-6086.

[13] REN S,HE K,GIRSHICK R,et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2017,39(6):1137-1149.

[14] JANG Y,SONG Y L,YU Y,et al. TGIF-QA:toward spatiotemporal reasoning in visual question answering[EB/OL]. [2019-11-10]. https://arxiv. org/pdf/1704. 04497. pdf.

[15] TRAN D,BOURDEV L,FERGUS R,et al. Learning spatiotemporal features with3Dconvolutional networks[EB/OL][.2019-11-10].https://arxiv. org/abs/1412. 0767.

[16] HE Kaiming,ZHANG Xiangyu,REN Shaoqing,et al. Deep residual learning for image recognition[C]// Proceedings of IEEE Conference on Computer Visionand Pattern Recognition. Washington D. C. ,USA:IEEE Press,2016:770-778.

[17] MU J Q,BHAT S M,VISWANATH P. All-but-the-top: simple and effective postprocessing for word representa tions[EB/OL].[2019-11-10]. https://arxiv. org/abs/1702. 01417.

[18] DENG J,DONG W,SOCHER R,et al. ImageNet:a largescale hierarchical image database[C]//Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C. ,USA:IEEE Press,2009:45-69.

[19] SIMONYAN K,ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL][.2019- 11-10]. https://arxiv. org/abs/1409. 1556.

[20] SZEGEDY C,LIU N W,JIA N Y,et al. Going deeper with convolutions[C]//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C. ,USA:IEEE Press,2015:12-26.

[21] CHUNG J,GULCEHRE C,CHO K,et al. Empirical evaluation of gated recurrent neural networks on sequence modeling[EB/OL]. [2019-11-10]. https://arxiv. org/abs/1412. 3555.

[22] CHU W,XUE H,ZHAO Z,et al. The forgettablewatcher model for video question answering[J]. Neurocomputing,2018,314:386-393.

[23] WANG Bo,XU Youjiang, HAN Yahong,et al. Movie question answering:remembering the textual cues for layered visual contents[EB/OL].[2019-11-10]. https:// arxiv. org/pdf/1804. 09412. pdf.

[24] LEI J,YU L,BANSAL M,et al. Tvqa:localized, compositional video question answering [EB/OL]. [2019-11-10]. https://www. aclweb. org/anthology/D18- 1167. pdf.

[25] ZHANG Jing,CHEN Qingkui. Analysis of crowd congestion degree in narrow space based on attention mechanism[J]. Computer Engineering,2020,46(9):254-260,267.(in Chinese)

[26] LI Yachao,XIONG Deyi,ZHANG Min. A survey of neural machine translation[J]. Chinese Journal of Computers,2018,41(12):2734-2755.(in Chinese)

[27] YU Y,KIM J,KIM G. A joint sequence fusion model for video question answering and retrieval [C]// Proceedings of European Conference on Computer Vision. Berlin,Germany:Springer,2018:471-487.

[28] YE Yunan,ZHAO Zhou,LI Yimeng,et al.Video question answering via attribute-augmented attention network learning[C]//Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York,USA:ACM Press,2017: 829-832.

[29] XU Dejing,ZHAO Zhou,XIAO Jun,et al. Video question answering via gradually refined attention over appearance and motion[C]//Proceedings of the 25th ACM International Conference on Multimedia. New York,USA:ACM Press, 2017:1645-1653.

[30] LIANG Lili. Research on video question answering based on deep learning method[D]. Harbin:HarbinUniversity of Science and Technology,2019.(in Chinese)

[31] YAO L,TORABI A,CHO K,et al. Describing videos by exploiting temporal structure[C]//Proceedings of IEEE International Conference on Computer Vision. Washington D. C. ,USA:IEEE Press,2015:4507-4515.

[32] DONAHUE J,HENDRICKS L A,ROHRBACH M,et al. Long-term recurrent convolutional networks for visual recognition and description[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2014,39(4): 677-691.

[33] SUN C,MYERS A,VONDRICK C,et al. Videobert:a joint model for video and language representation learning[EB/OL].[2019-11-10]. https://arxiv.org/pdf/ 1904. 01766. pdf.