Multi-modal Feature Fusion Based on Variational Autoencoder for Visual Question Answering

Liqing Chen, Yifan Zhuo, Yingjie Wu, Yilei Wang(✉), and Xianghan Zheng

College of Mathematics and Computer Science, Fuzhou University,
Fuzhou, Fujian Province, China
yilei@fzu.edu.cn

Abstract. Visual Question Answering (VQA) tasks must provide correct answers to the questions posed by given images. Such requirement has been a wide concern since this task was presented. VQA consists of four steps: image feature extraction, question text feature extraction, multi-modal feature fusion and answer reasoning. During multi-modal feature fusion, outer product calculation is used in existing models, which leads to excessive model parameters, high training overhead, and slow convergence. To avoid these problems, we applied the Variational Autoencoder (VAE) method to calculate the probability distribution of the hidden variables of image and question text. Furthermore, we designed a question feature hierarchy method based on the traditional attention mechanism model and VAE. The objective is to investigate deep questions and image correlation features to improve the accuracy of VQA tasks.

Keywords: Visual Question Answering · Multi-modal feature fusion · Variational Auroencoder · Attention mechanism

1 Introduction

Visual Question Answering (VQA) [1] tasks must provide correct answers to the questions posed by given images. In comparison with the traditional Question Answering system, the search and reasoning parts must be based on the image content. This system contains the knowledge of target location detection, scene classification and knowledge reasoning. VQA tasks can be easily expanded to other tasks and play a significant role in various practical scenarios such as automobile navigation, medical system and education system.

In this paper, we propose a multi-modal feature fusion method for combining image and question features. The central idea is to use Variational Autoencoder

Student Paper. This work is supported by the Natural Science Foundation of Fujian Province of China (2017J01754). This work is supported by the Natural Science Foundation of Fujian Province of China (2018J01799).

Z. Lin et al. (Eds.): PRCV 2019, LNCS 11858, pp. 657–669, 2019.
https://doi.org/10.1007/978-3-030-31723-2_56

(VAE) [2] to calculate the hidden coding of image and question features and then fuse them in the hidden layer to obtain the associated image and question representations for improved answer reasoning. We use the fusion method to the basic model and verify its validity on the VQA 2.0 dataset. Subsequently, in decoding the attention weight, the sampling method is added to the attention mechanism of VQA tasks to increase randomness and the hierarchical attention mechanism model is designed by using hidden variables, and a further generalized attention mechanism weighting matrix, which can weight image and question features, is generated. Experimental results show that our model further improves the accuracy of VQA tasks.

The remainder of this paper is presented as follows: In Sect. 2, we introduce the relevant work in recent years. In Sect. 3, we present the implementation details of the model. In Sect. 4, we compare basic models and our model on VQA 2.0 dataset. In Sect. 5, we conclude this paper.

2 Related Work

2.1 Visual Question Answering

VQA tasks have been proposed in 2015. In recent studies, majority of the methods in VQA are based on neural networks. Convolution Neural Networks (CNNs) [3] are generally used to extract image features, whereas Recurrent Neural Networks (RNNs) [4] are utilized to extract question features. Then, the two features are fused to form a new feature, which is used for answer reasoning (see Fig. 1).

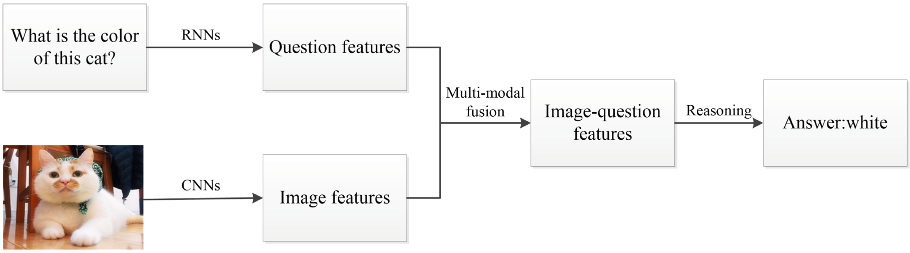


Fig. 1. Simple model of the VQA task.

Recently, most VQA tasks use VGGNet [5] and ResNet [6] to extract image features. Girshick et al. [7] proposed the use of Fast Region-based Convolutional Network (Fast-R-CNN)to extract image features consisting of multiple objects and obtained new state-of-the-art results. By contrast, almost all VQA models use GRU [8] and LSTM [9] to extract question features. These two models can efficiently obtain the question contextual information. Multi-model feature fusion is a method for associate images with question textual information.

Fukui et al. [10] first introduced the bilinear model into multi-modal feature fusion in VQA. They proposed the Multi-modal Compact Bilinear (MCB) pooling method and achieved good results. Then Yu et al. [11] designed a Multi-modal Factorized High-order (MFH) pooling method to improve the result further.

VQA reasoning method is simple. We must develop a limited quantity answer set in accordance with the frequency of the answers and perform classification tasks on it.

2.2 Attention Mechanism

Recently, attention mechanism, which finds the most deserving word or phrase in the text, has been successfully applied in the field of natural language processing. In the VQA task, researchers use the attention mechanism to discover picture areas that are most related to the semantic information of the question (see Fig. 2).

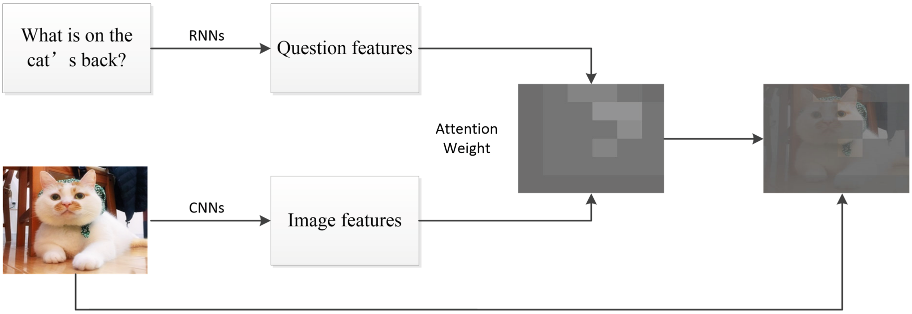


Fig. 2. Simple attention mechanism model for VQA.

[12] proposed a question-oriented image attention mechanism. This method assigns attention weights to image features based on the question features. [13] introduced a collaborative attention mechanism to associated images and questions that became the baseline method at that time. [14] firstly introduced the multi-objective feature extraction method in the field of target detection into the VQA model and named it as bottom-up attention. Compared with the method of weighting the attention of the whole image, this model can directly focus on the image target itself by weighting the entire object's attention, which has been significantly improved and has become one of the best models in the field of VQA.

2.3 Variational Autoencoder

VAE is a probabilistic approximation model based on variational inference and autoencoder structure. Suppose two variables x and z, the variational inference uses simple distribution $q(z)$ to approximate complex posterior distribution

$p(z|x)$ and Kullback-Leibler (KL) distance for measuring the distance between probability distributions:

$$KL(q(z)||\mathrm{p}(z|x)) = \int q(z) \ln \frac{q(z)}{p(z|x)} dz \tag{1}$$

The smaller the KL distance, the closer the two probability distributions are. The goal is to minimize the KL distance. Further derivation form is as follows:

$$\ln p(x) - KL(q(z)||p(z|x)) = \int q(z) \ln p(x|z) dz - KL(q(z)||p(z)) \tag{2}$$

VAE assumes that $q(z)$ obeys a normal distribution $N\left(\mu, \sigma^2\right)$, and $p(z)$ obeys a normal distribution $N(0, I)$. The optimization objectives of the model can be expressed as:

$$L = E_{x \sim p(x)}[-\ln q(x|z) + KL(p(z|x)||q(z))], z \sim p(z|x) \tag{3}$$

Here,$-\ln q(x|z)$ indicates the distance of the generating and real values. The KL distance can be calculated by:

$$KL\left(N\left(\mu, \sigma^2\right)||N(0, I)\right) = \frac{-\sum \log\left(\sigma^2\right) - d + \sum\left(\sigma^3\right) + \mu^T \mu}{2} \tag{4}$$

The structure of VAE (see Fig. 3) makes it a generating model. By sampling and decoding the probability distribution on the trained model, new data are generated with the same distribution as the training data. Therefore, this model is widely used in the field of image generation with Generative Adversarial Networks (GAN) [15].

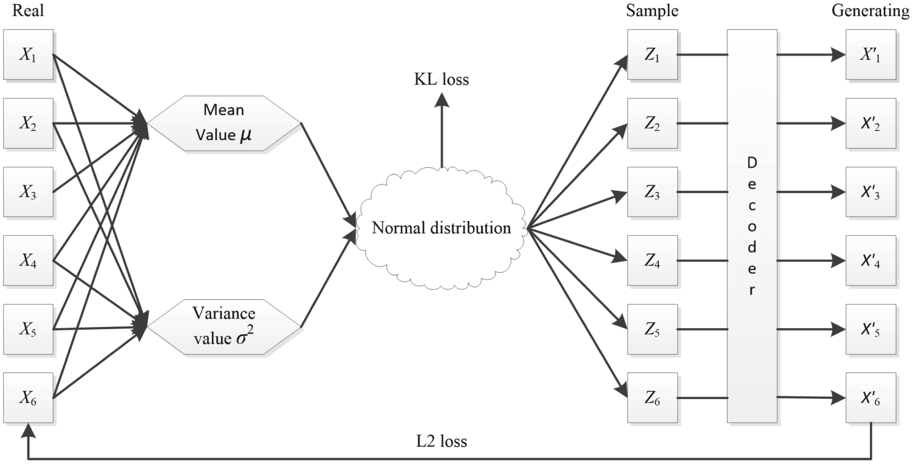


Fig. 3. Model of the Variational Autoencoder.

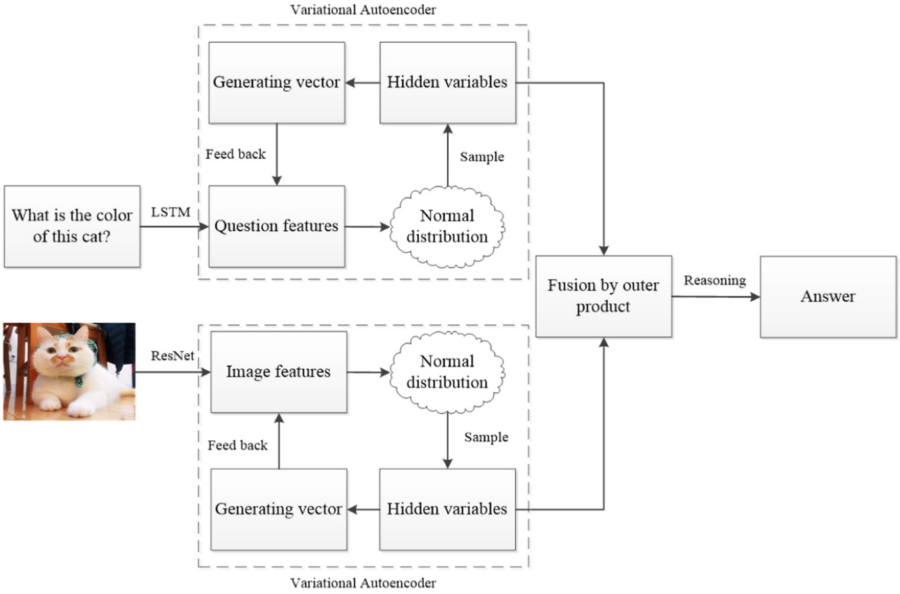


Fig. 4. Structure of multi-modal feature fusion model.

3 Proposed Method

3.1 Multi-modal Feature Fusion

Traditional VQA fusion methods only consider the external representation of features instead of the important hidden links between images and questions, thereby losing information during fusion.

Currently, multi-modal feature fusion methods are based on the calculation of the outer product or approximate outer product of the features. This case limits the scope of application because numerous parameters and computational loads are required during calculation, and the dimension reduction methods of the optimization calculation process are sensitive to the super-parameters and slow convergence speed of the model.

Our first work is to attempt to use VAE to solve the above-mentioned problem (see Fig. 4). In this model, we use ResNet to extract image features and LSTM to extract question features. Then we apply the VAE model mentioned in Sect. 2.3 to calculate the hidden vector probability distribution of the features. Finally, the hidden variables of features are sampled and fused.

The algorithm for calculating the probability distribution of hidden variables is shown as Algorithm 1. The extracted image hidden variables are multiplied by the question hidden variables and the results are input into the full connection layer. By locally adjusting the model structure, several different models are obtained, which mainly include calculating the distribution of hidden variables for image features only, for question features only, and simultaneously for image features and question features. In order to ensure the association between image and question, we fuse the image feature hidden variable with the question feature hidden variable, then use the merged feature to decode. After that, we attempts to fuse the image feature hidden variable and the question feature hidden variable into the multi-modal decomposition bilinear pooling method.

Algorithm 1. Probability distribution calculation of hidden variables.

```
Input:
    image features or question features,F
Output:
    Distribution parameters of latent variables,$(\mu, \sigma)$;
    loss value, $loss$;
$f \leftarrow \mathrm{Relu}\left(W_I F + b_I\right)$;
$\mu \leftarrow W_\mu f + b_\mu$;
$\sigma \leftarrow W_\sigma f + b_\sigma$;
$kld\_loss \leftarrow \frac{1}{2}\left(1 - \|\mu\|^2 - \|\sigma\|^2 - \log\left(\sigma^2\right)\right)$;
$z \sim N(0, I)$;
$z' \leftarrow \mu + \sigma z$;
$F' \leftarrow W_{F2}\left(W_{FI} z' + b_{FI}\right) + b_{F2}$;
$l2\_loss \leftarrow \|F - F'\|^2$
$loss \leftarrow kld\_loss + l2\_loss$
return $(\mu, \sigma)$,$loss$;
```

3.2 Variational Attention Mechanism

Our second work is to introduce a variational attention mechanism in the process of multi-modal feature fusion to reduce the complexity of the parameters.

We use Faster-R-CNN to extract multi-target features from images. Then, an implicit variable model of attention is established on the basis of bottom-up attention mechanism model and variational inference. Finally, a method for the multi-sample fusion of attention weighted features is designed (see Fig. 5).

For a further a generalized expression of attention weight of the feature of local question, the feature of the question text need to be generated hierarchically. However, the above model still directly calculates the attention weight for the image many times, and the number of parameters required for the untreated image will lead to inefficiency, then we sampled the hidden variables of the image several times. Therefore, we sampled the hidden variables of the image several times, and calculated the attention weight of the image features by combining with the previous features of each layer. The number of parameters can be greatly reduced and the training speed of the model can be improved due to the low dimension of the problem text features after hidden variable coding and layering (see Fig. 6). Algorithm 2 shows the algorithm for question feature hierarchy.

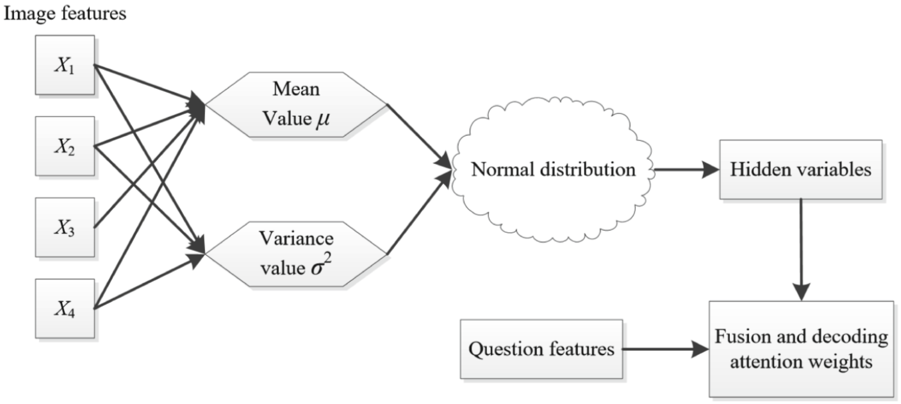
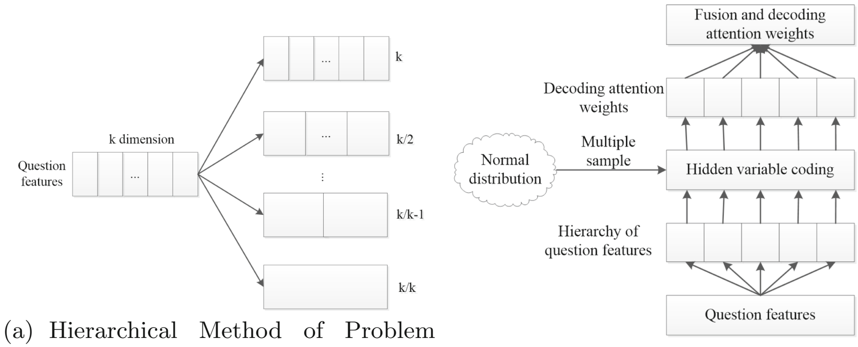


Fig. 5. Calculation of attention weight with VAE.



(a) Hierarchical Method of Problem Characteristics.

(b) Variational Attention method.

Fig. 6. Variational attention mechanism model.

4 Experiment

4.1 Datasets

The dataset used in the experiment is test-dev of VQA 2.0. The images are obtained from MS-COCO dataset, including 123,287 images, of which 72,738 are used for training and 38,948 for testing. Each image has a corresponding question and answer. The evaluation of answers can be divided into three types: yes/no, number and others. The three types correspond to judgment, counting and open questions respectively. On the VQA2.0 dataset, the calculation of accuracy rate does not directly measure the proportion of the correct answered samples. The calculation formula is as follows:

$$Acc = \frac{1}{M}\sum_{i=1}^{M}\min\left\{\frac{\text{human that provided that answer}}{3}, 1\right\} \tag{5}$$

Algorithm 2. Variational attention mechanism for question feature hierarchy.

```
Input:
    The image features, $I$;
    The question features, $F$;
    The question features length, $k$;
    The threshold of $KL$, $r$;
Output:
    The attention weight matrix, $Iatt$ ;
    loss value, $loss$;
$f \leftarrow \text{Relu}\,(W_1 I + b_1)$;
$\mu \leftarrow W_\mu f + b_\mu$;
$\sigma \leftarrow W_\sigma f + b_\sigma$;
$loss \leftarrow \max\left(r, \frac{1}{2}\left(1 - \|\mu\|^2 - \|\sigma\|^2 - \log\left(\sigma^2\right)\right)\right)$;
for $i = 1$ to $k$ do
    Initialize $Q_-$ Level $_i$;
    $j = 0$;
    while $j < |Q_i|$ do
        $Q_-\text{Level}_i.\text{add}\,(Q_{i,j})$;
        $j = j + i$;
    end while
end for
Initialize $Iatt_-Level$
for $i = 1$ to $k$ do
    $z \sim N(0, I)$;
    $z' \leftarrow \mu + \sigma z$;
    $Q_-z \leftarrow \text{merge}\,(Q_-Level_i, z')$
    $Iatt_-Level_i \leftarrow \text{Softmax}\,(\text{Conv}\,(Q_-z))$
end for
$Iatt \leftarrow \text{Sumpooing}(\frac{Iatt_-Level}{k})$
return $Iatt$,$loss$;
```

where M represents the total number of tested samples, and "humans that provided that answer" indicates the number of answers predicted by the model consistent with those manually collected by VQA 2.0.

4.2 Configurations

In the training process of this study, the neural networks built by different models use uniform hyperparameters. Table 1 shows the key parameters, in which Weight_VQAVae denotes the weights used in the multi-modal feature fusion method. Weight_VQAVaeAtt indicates the weights used in variational attention mechanism.

4.3 Results

Multi-modal Feature Fusion Results. The different structures of the local model are as follows, and Table 2 shows the experimental results.

Table 1. Key parameters of our methods.

Parameters	Weight_VQAVae	Weight_VQAVae Att
Batch Size	128	64
Loss	KLDivloss	KLDivloss
Learning rate	0.001	0.007
Learning rate decay	0.5	0.5
Decay step	20000	20000
Training interval	60000	100000
Drop	0.5	0.5
Hidden code size	128	128
Variational weight	-	0.000005
Image number	-	36
Optimizer	Momentum	Momentum

Table 2. Experimental results of multi-modality feature fusion method.

I_h	Q_h	$Concat$	OP	$Merge$	Yes/No	Number	Others	All
		✓			74.14	36.18	45.45	55.02
✓		✓			75.32	35.84	45.12	55.25
	✓	✓			75.45	35.85	44.21	54.74
✓	✓	✓			74.85	36.13	45.09	55.09
✓			✓		76.55	37.24	47.06	56.85
	✓		✓		76.39	36.56	45.28	56.02
✓	✓		✓		77.09	36.66	45.33	56.24
✓		✓		✓	77.73	36.74	48.36	57.87
✓			✓	✓	**77.68**	**37.56**	**48.37**	**58.01**

* I_h: Image feature hidden variable
* Q_h: Problem feature hidden variable
* $Concat$: Fusion feature using stitching method
* $OuterProduct(OP)$: Fusion feature using outer product method
* $Merge$: Binding of hidden variables based on bilinear pooling

Experimental results show that implicit vector coding using image features, encoding without question features, and multi-modal feature fusion using outer products are significant improvements in the accuracy of the VQA model. We rename $I\_h + Q + OP + Merge$ with the best experimental results on the model as VQAVae, jointly use the training and verification sets to train the model, and evaluate the model accuracy on the test set. Table 3 shows the results compared with the existing basic VQA model.

Table 3. VQAVae compared with existing base models.

Method	Yes/No	Number	Others	All
IBOWING [16]	76.5	35.0	42.6	55.7
DPPnet [17]	80.7	37.2	41.7	57.2
Norm LSTM I+Q [1]	80.5	36.8	43.1	57.8
AYN [18]	78.4	36.4	46.3	58.4
AMA [19]	81.0	38.4	45.2	59.2
MCB [10]	81.2	35.1	49.3	60.8
MFB [11]	79.02	39.21	50.57	61.0
VQAVae	**80.92**	**39.69**	**50.92**	**61.48**

Results show that the proposed multi-modal feature fusion method based on variational inference outperforms most of the existing basic VQA models. The possible original meaning is that the question text is a discrete word sequence, and the image features are further continuous. In decoding the new features and calculating the error with the original features, the image features can be efficiently restored to the original features. During training, the difference value can be easily optimized as part of the loss value, so that the coding probability of the hidden variable coding can be calculated further accurately.

Variational Attention Mechanism Results. We name the used methods as follow:

* ResNet (Res): Extracting image features using a residual network
* Fast-R-CNN (FRC): Extracting multi-target image features using Fast-R-CNN
* Qatt: Problem-oriented self-attention mechanism
* Iatt: Problem-oriented attention mechanism for images
* Concat: Splicing combines multiple sampling features
* Average (Ave): Weighted average blends multiple sampling features.

Table 4 presents the compared experimental results. Then we rename VQAVaeAtt as the model with the best experimental results on the verification, jointly use the training and verification sets to train the model, and calculate the model accuracy on the test set. Table 5 shows the results compared with the existing basic VQA model.

The proposed attention mechanism method based on VAE outperforms most existing VQA models. Because (1) the attention mechanism is modeled as an implicit variable model and the probability distribution of attention weight is calculated through VAE; (2) multiple attention weight sampling is added to the model and (3) the attention weight of image is calculated combined with the feature information of subsection questions, which is helpful for obtaining additional information. To sum up, modeling the effective method to model the attention mechanism as an implicit variable model based on the complete image attention weighted feature is a novel and effective method.

Table 4. Experimental results of the variational attention method

Res	FRC	Iatt	Qatt	Concat	Ave	Yes/No	Number	Others	All
✓						77.68	37.56	48.37	58.01
✓		✓				78.64	38.37	49.9	60.43
	✓	✓				80.11	40.28	52.7	61.35
	✓	✓	✓			80.05	41.77	52.64	61.53
	✓	✓	✓	✓		80.42	41.04	53.68	52.15
	✓	✓	✓		✓	**80.4**	**40.67**	**54.13**	**62.23**

Table 5. VQAVaeAtt compared with existing base models.

Method	Yes/No	Number	Others	All
DPPnet [18]	80.7	37.2	41.7	57.2
SMem [12]	80.9	37.3	43.1	58.0
NMN [20]	81.2	38.0	44.0	58.6
SAN [13]	81.1	36.6	46.1	58.7
HieCoAtt [21]	79.7	38.7	51.7	61.8
MRN [22]	81.9	39.0	53.0	63.18
MCB [10]	82.2	37.7	54.8	64.2
VQAVaeAtt	**81.79**	**42.76**	**55.5**	**64.89**

5 Conclusion

This study investigates the feature fusion of VQA, including multi-modal feature fusion and attention mechanism. VAE is introduced to overcome the limitations of existing methods, and greatly improves the accuracy of the VQA model. The main contribution of this work includes two parts:

The VAE is introduced to calculate the probability distribution of hidden variables of image and question text features, and a multi-modal feature fusion method based on hidden variables is designed to reduce the computational complexity of the model effectively. Furthermore, the random sampling increases the anti-over-fitting of the model. Comparative experiments show that the model effectively improves the accuracy of VQA tasks.

We attempted to introduce a variational attention mechanism in the process of multi-modal feature fusion. Based on the VAE model, the question text information is used to guide the autoencoding of image features in accordance with their attention weights. As such, a hierarchical attention mechanism, which effectively reduces the number of parameters required by the model, is established. The comparative experiments show that the variational attention mechanism can further improve the model accuracy in VQA tasks.

References

1. Agrawal, A., et al.: AQA: visual question answering. Int. J. Comput. Vis. **123**, 4–31 (2017)
2. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
3. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
4. Elman, J.L.: Finding structure in time. Cogn. Sci. **14**, 179–211 (1990)
5. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
7. Girshick, R.: Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448 (2015)
8. Cho, K., et al.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 (2014)
9. Donahue, J., et al.: Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2625–2634 (2015)
10. Fukui, A., Park, D.H., Yang, D., Rohrbach, A., Darrell, T., Rohrbach, M.: Multimodal compact bilinear pooling for visual question answering and visual grounding. arXiv preprint arXiv:1606.01847 (2016)
11. Yu, Z., Yu, J., Xiang, C., Fan, J., Tao, D.: Beyond bilinear: generalized multimodal factorized high-order pooling for visual question answering. IEEE Trans. Neural Netw. Learn. Syst. **14**, 1–13 (2018)
12. Xu, H., Saenko, K.: Ask, attend and answer: exploring question-guided spatial attention for visual question answering. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9911, pp. 451–466. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46478-7_28
13. Yang, Z., He, X., Gao, J., Deng, L., Smola, A.: Stacked attention networks for image question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 21–29 (2016)
14. Anderson, P., et al.: Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6077–6086 (2018)
15. Goodfellow, I., et al.: Generative adversarial nets. In: Advances in Neural Information Processing Systems, pp. 2672–2680 (2014)

16. Zhou, B., Tian, Y., Sukhbaatar, S., Szlam, A., Fergus, R.: Simple baseline for visual question answering. arXiv preprint arXiv:1512.02167 (2015)
17. Noh, H., Hongsuck Seo, P., Han, B.: Image question answering using convolutional neural network with dynamic parameter prediction. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 30–38 (2016)
18. Malinowski, M., Rohrbach, M., Fritz, M.: Ask your neurons: a neural-based approach to answering questions about images. In: Proceedings of the IEEE international conference on computer vision, pp. 1–9 (2015)
19. Wu, Q., Wang, P., Shen, C., Dick, A., van den Hengel, A.: Ask me anything: free-form visual question answering based on knowledge from external sources. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4622–4630 (2016)
20. Andreas, J., Rohrbach, M., Darrell, T., Klein, D.: Neural module networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 39–48 (2015)
21. Lu, J., Yang, J., Batra, D., Parikh, D.: Hierarchical question-image co-attention for visual question answering. In: Advances In Neural Information Processing Systems, pp. 289–297 (2016)
22. Noh, H., Han, B.: Training recurrent answering units with joint loss minimization for VQA. arXiv preprint arXiv:1606.03647 (2016)