ISSN: 1813-4890

Research on bilinear pooling feature fusion optimization driven by recurrent structure

Boxiang Liu [a], Xianglian Chen [b]

School of Information Technology and Engineering, Tianjin University of Technology and Education, Tianjin 300222, China;

[a]box0707@163.com, [b]xiangliang_chen@163.com

Abstract

In multimodal visual question answering tasks, feature fusion technology plays a crucial role in enhancing the model's understanding and reasoning capabilities. This paper proposes a feature fusion method that integrates recurrent structures and bilinear pooling technology, aiming to enhance the model's ability to explore complex interactions between image and text features. Specifically, recurrent structures enable the model to explore feature interactions from multiple perspectives through applying different linear transformations in multiple iterations, effectively alleviating the information redundancy and burstiness issues present in traditional bilinear pooling technology. Furthermore, this method can be organically combined with existing improved bilinear pooling methods to further enhance their performance. Experimental results show that, compared to bilinear pooling and its improved methods alone, the proposed method achieves significant performance improvements in visual question answering tasks, validating the effective enhancement of recurrent structures on feature fusion methods and bringing new research ideas and solutions to the field of multimodal feature fusion.

Keywords

Multimodal, Visual Question Answering, Bilinear Pooling, Recurrent Structure.

1. Introduction

In the field of artificial intelligence, multimodal tasks integrate multi-source information such as visual and textual data, and Visual Question Answering (VQA) is a typical example. VQA requires models to understand visual content in images and accurately answer questions by combining textual information. Effective feature fusion methods are key to improving the performance of VQA models, and bilinear pooling technology has garnered significant attention. Bilinear pooling effectively captures complex interactions between image and text features by mapping the two modalities to different spaces and performing bilinear mapping to obtain a fused feature representation. Compared to simple feature fusion methods, it can more fully exploit inter-modal information and significantly enhance model performance.

However, bilinear pooling suffers from issues of information redundancy and suddenness, which affect stability and accuracy. To address these limitations, researchers have proposed various

improvement methods, which have somewhat enhanced robustness and efficiency, but there is still room for further improvement. In light of this, this paper proposes a feature fusion method that combines bilinear pooling technology with a recurrent structure. By leveraging different linear transformations in different iterations through the recurrent structure, the method explores feature interactions from multiple perspectives, enhancing the model's capacity and flexibility to further improve the performance of VQA models.

2. Research background

Visual Question Answering (VQA), lying at the intersection of computer vision and natural language processing, has recently gained widespread attention. Researchers both domestically and internationally have proposed numerous VQA models, which typically follow these four steps: visual feature extraction (image feature extraction), textual feature extraction (question feature extraction), feature fusion, and answer generation.

Joint embedding methods, which map visual and textual features to a shared feature space to facilitate information interaction and reasoning, form the foundation of VQA models. Early research, such as the "Neural-Image-QA" model proposed by Malinowski et al. [1], utilized CNNs and LSTMs to process multimodal information and treated VQA as a sequence-to-sequence task. Subsequently, researchers explored simple feature fusion mechanisms like element-wise product, sum, and concatenation. Bilinear pooling, a feature fusion technique, has shown promise in multimodal learning tasks, especially VQA. However, it also suffers from information redundancy and suddenness issues. To tackle these problems, researchers have proposed advanced feature fusion techniques, such as Multimodal Compact Bilinear (MCB) [2], Multimodal Low-Rank Bilinear (MLB) [3], and Multimodal Factorized Bilinear (MFB) [4], to reduce model parameters and enhance robustness.

While joint embedding methods are fundamental to VQA models, they struggle with information redundancy and suddenness in feature fusion. This paper proposes a novel bilinear pooling feature fusion method that addresses these issues. It surpasses traditional joint embedding methods in capturing feature interactions, optimizing model parameters, increasing model capacity and flexibility, and enabling multi-perspective feature interaction learning.

3. Related work

3.1. Multimodal feature fusion

Multimodal feature fusion [5] is a key technology in multimodal learning. It integrates feature information from different modalities (e.g., images, text, audio) to form a more comprehensive and richer feature representation. This enhances the model's ability to understand, reason, and make decisions based on multi-source information. In Visual Question Answering (VQA) tasks, feature fusion combines visual features from images and textual features from questions, enabling the model to simultaneously utilize both types of information to generate accurate answers. Common feature fusion methods include concatenation, weighted sum, element-wise product, element-wise max, and using learning modules (e.g., MLP) to automatically learn feature fusion strategies, among others.

Early Fusion: Merge data from different modalities before feature extraction, then perform unified feature extraction and processing. This method fully utilizes the complementary information of raw data but is computationally intensive and sensitive to representation differences between modalities.

Late Fusion: Extract and process features separately for each modality, then combine the results at the decision-making stage. This approach is simple and allows for modality-specific optimization but may lose inter-modality interaction information.

Hybrid Fusion: Conduct fusion at different stages of feature extraction, combining the benefits of early and late fusion. It retains inter-modality interaction information while reducing computational complexity. This paper adopts hybrid fusion.

3.2. Bilinear Pooling and Improved Algorithms

3.2.1. Bilinear Pooling

Bilinear Pooling [6] is a technique for multimodal feature fusion, aiming to capture complex interactions between features of two different modalities. It works by mapping two feature vectors into different spaces and performing bilinear mapping to get the fused feature representation. Specifically, given image feature vectors H_{img} and text feature vectors H_{txt}, it calculates their outer product to form a matrix that contains all possible interactions between the two, thus comprehensively capturing the inter-modal correlation information.

If we denote the image feature vector as $H_{img} \in R^{d1}$ and the text feature vector as $H_{txt} \in R^{d2}$, the bilinear pooling operation can be expressed as:

$$H_{fuse} = H_{img} \cdot H_{img}{}^{T} \in R^{d1 \cdot d2} \tag{1}$$

To convert the fused feature matrix into a vector, a flattening operation is typically used, which involves unfolding the matrix into a vector.

$$H_{fuse\_vec} = vec(H_{fuse}) \in R^{d1 \cdot d2} \tag{2}$$

Bilinear pooling is highly advantageous in multimodal tasks, as it can capture higher-order interactions between the features of two modalities. Within the realm of multimodal tasks, bilinear pooling significantly enhances the model's understanding and reasoning capabilities by providing a richer set of feature interactions. It is also compatible with a variety of models.

However, the computational complexity of bilinear pooling is high, as the feature dimension it generates is the product of the dimensions of the two original features, leading to an increase in feature dimensions. Since all possible interactions are computed, some may be unimportant or noisy, resulting in information redundancy. This necessitates a large amount of data to avoid overfitting.

To tackle the above-mentioned issues of bilinear pooling, researchers have proposed a multitude of improvement schemes.

3.2.2. Low-Rank Decomposition Based Methods

Low-rank decomposition based methods assume that the result matrix of bilinear pooling has a low-rank structure, which can be approximately represented as the product of two low-dimensional matrices. In this way, the number of parameters and computational complexity can be significantly reduced.

Multimodal Compact Bilinear Pooling (MCB): By using low-rank constraints, the high-dimensional feature matrix of bilinear pooling is approximately represented as the product of two low-dimensional matrices. This greatly reduces the number of parameters and computational cost while retaining key feature interaction information.

Multimodal Low-Rank Bilinear Pooling (MLB): Low-rank decomposition is introduced into the bilinear pooling process. Images and text features are mapped to low-dimensional spaces for interaction, reducing computational complexity and improving model scalability on large datasets.

3.2.3. Feature Mapping and Transformation Based Methods

Feature mapping and transformation based methods introduce nonlinear mapping functions to map raw features into a new feature space, and then perform bilinear pooling in that space.

This enhances the model's ability to capture complex interactions while reducing information redundancy.

Multimodal Feature Transformation Bilinear Pooling (MTB): Images and text features are first nonlinearly transformed into a new feature space, and then bilinear pooling is performed to enhance the capture of complex interactions and reduce information redundancy.

Multimodal Convolutional Bilinear Pooling (MCBP): Convolutional operations are used to map and transform image and text features, followed by bilinear pooling. By leveraging the local perception and parameter sharing of convolutions, computational cost is reduced and the expressiveness of feature interactions is improved.

3.2.4. Factorization Based Methods

Factorization based methods decompose the result matrix of bilinear pooling into the product of multiple low-dimensional matrices, thereby reducing the number of parameters and computational complexity. This method assumes that feature interactions can be explained by a set of basis functions or factors.

Multimodal Factorized Bilinear Pooling (MFB): The result matrix of bilinear pooling is decomposed into the sum of outer products of multiple low-dimensional vectors. These vectors are learned to approximate original high-order interactions, reducing the number of parameters and computations.

Multimodal Partial Factorization Bilinear Pooling (MPF): Partial factorization of the feature matrix from bilinear pooling is performed, preserving the main feature components. This reduces dimensionality while minimizing information loss, improving model efficiency and performance.

4. Feature Mapping and Transformation Based Methods

The core concept of the bilinear pooling recurrent feature fusion method is to introduce a novel bilinear pooling feature fusion approach. It integrates a recurrent structure with multiple independently learned linear layers to capture the complex interactions between image and text features. We name it Multimodal Recurrent Bilinear Pooling. Here’s a detailed implementation:

4.1. Model structure design

Image feature extraction: Use a pre-trained image feature extraction model (ResNet-152)[7] to extract image features x_v.

Text feature extraction: Use a pre-trained text feature extraction model (BERT) [8]to extract text features x_q.

Feature transformation: Apply dropout to the image and text features separately to prevent overfitting. Use linear transformations to map them into the same feature space and introduce non-linearity, enabling the model to learn more complex feature representations. Obtain the processed image features y_v and text features y_q.

Recurrent structure: The processed features enter the recurrent structure. In each iteration, the image features y_v and text features y_q are transformed by separate linear layers from their respective module lists (list_linear_hv for image, list_linear_hq for text), obtaining transformed image features z_v and text features z_q. Perform bilinear pooling and add the result to the list list x_m.

Loop check: After bilinear pooling, check if the number of iterations has reached the set value R. If not, select another linear layer from list_linear_hv and list_linear_hq for the next iteration.

Feature fusion: Stack and flatten the outputs from the bilinear pooling results in list x_m to obtain the final fused feature.

The model structure diagram is shown in Fig. 1.

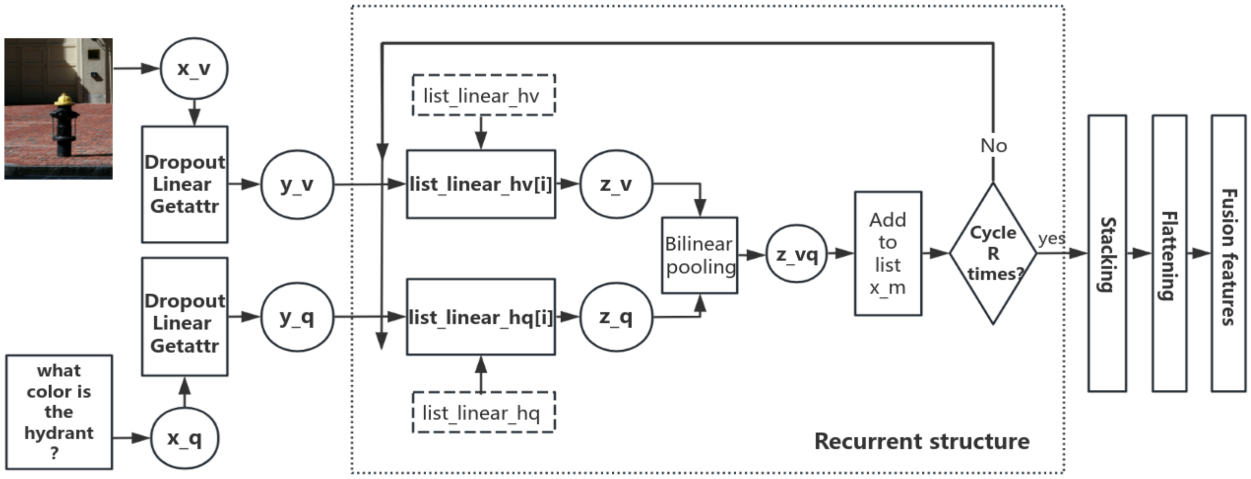

Fig. 1 The model structure diagram

This strategy combines bilinear pooling with multiple independently learned linear layers, effectively capturing complex interactions between image and text features. It not only enhances the model's understanding of multimodal data but also, through its recurrent structure, allows the model to use different linear transformations in each iteration. This increases the model's capacity and flexibility, enabling it to capture a wider variety of feature interactions. By exploring feature interactions from multiple perspectives, the model achieves deeper and broader feature fusion, significantly boosting performance in multimodal tasks.

5. Experimental Design and Results

5.1. Experimental dataset

This paper employs two datasets: VQA_v2 (Visual Question Answering 2.0) and OKVQA (Outside Knowledge Visual Question Answering). VQA_v2 comprises over 110,000 images, each paired with three questions, yielding approximately 330,000 question-answer pairs. These questions, sourced mainly from the COCO dataset, vary in type, including descriptive, inferential, and commonsense questions, with open-ended answers ranging from words to sentences. To minimize reliance on spurious correlations in training data, the dataset ensures diverse answers to the same question across images.As shown in Fig. 2.

OKVQA, on the other hand, assesses the model's ability to answer complex questions requiring external knowledge. It contains around 14,000 questions designed to prompt reasoning based on both image content and external knowledge. While VQA_v2 focuses on general question answering, OKVQA emphasizes the use of external knowledge. Together, these datasets provide a comprehensive evaluation of the model's performance across tasks of varying difficulty and type.As shown in Fig. 3.

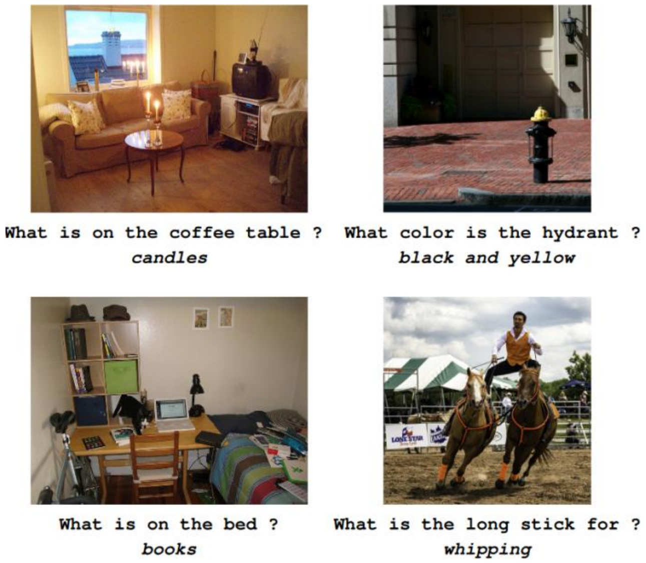
What is on the coffee table ?
candles
What color is the hydrant ?
black and yellow
What is on the bed ?
books
What is the long stick for ?
whipping

Fig. 2 VQA_v2

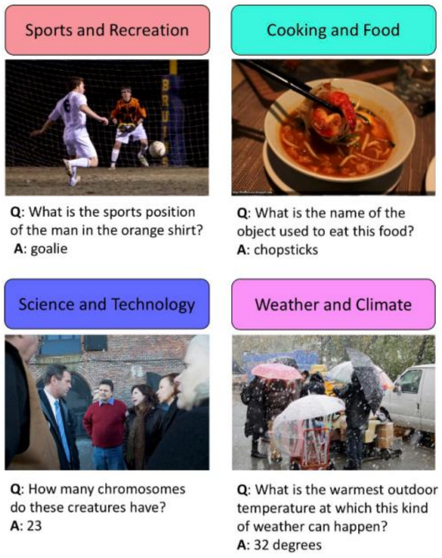
Sports and Recreation
Q: What is the sports position of the man in the orange shirt?
A: goalie
Cooking and Food
Q: What is the name of the object used to eat this food?
A: chopsticks
Science and Technology
Q: How many chromosomes do these creatures have?
A: 23
Weather and Climate
Q: What is the warmest outdoor temperature at which this kind of weather can happen?
A: 32 degrees

Fig. 3 OKVQA

5.2. Evaluation

To evaluate the model's effectiveness, we use an evaluation metric where, for each (image, question) pair, the model generates an answer. The score for this predicted answer is calculated based on its similarity to the answers provided by 10 annotators. The formula for calculating the score of each predicted answer is as follows:

$$Acc = \min(1, \frac{\# humans\ \ that\ provided\ \ that\ answer}{3}) \tag{3}$$

This means:

If the predicted answer occurs in fewer than 3 of the 10 annotators' answers, the accuracy is the number ofccu orrences divided by 3.If it occurs 3 or more times, the accuracy is 1.

5.3. Comparative experiment

We prepared the following algorithms to validate the effectiveness of our proposed model: Multimodal Bilinear Pooling(MBP)

Multimodal Low-rank Bilinear Pooling(MLB)

Multimodal Feature Transformation Bilinear Pooling (MTB)

Multimodal Factorized Bilinear Pooling (MFB)

Multimodal Recurrent Bilinear Pooling(MRB*)

The experimental results are as follows:

Table 1 Scheme Comparing

	VQA_v2 ACC	OKVQA ACC
MBP	0.5695	0.3748
MLB	0.6188	0.3389
MTB	0.5068	0.3803
MFB	0.5943	0.3293
MRB*	**0.6233**	**0.4171**

Analysis of the Influence of Cycle Count R：

Table 2 The Influence of Different Parameters R on the Results

	VQA_v2 ACC	OKVQA ACC
MRB(R=1)	0.5527	0.3233
MRB(R=10)	0.6078	0.3924
MRB*(R=20)	**0.6233**	**0.4171**
MRB(R=100)	0.5174	0.1805

As shown in Table 1, the Multimodal Recurrent Bilinear Pooling model (MRB) proposed in this paper achieved optimal performance on both benchmark datasets. Specifically, MRB attained an accuracy of 62.33% on the VQA_v2 dataset, which is 0.45 percentage points higher than the second-best MLB model; on the more challenging OKVQA dataset, MRB significantly outperformed other comparison methods with an accuracy of 41.71%, surpassing the second-ranked MTB model by 3.68 percentage points.

Table 2 further validates the crucial role of the recurrent structure: as the number of recurrent iterations R increased from 1 to 20, the model's performance on VQA_v2 and OKVQA consistently improved, reaching the optimal values of 62.33% and 41.71%, respectively; however, when R was increased to 100, the performance dropped significantly. This indicates that appropriate recurrent iterations can achieve multi-angle feature interaction through independently learned linear transformations, effectively alleviating the information redundancy inherent in traditional bilinear pooling and enhancing model capacity; but excessive iterations introduce noise and lead to overfitting. These results fully demonstrate that the recurrent structure, while maintaining parameter efficiency, significantly enhances the model's robustness and knowledge integration capabilities in complex visual question answering tasks.

6. Conclusion

This paper proposes a visual question answering feature fusion method based on bilinear pooling and recurrent feature fusion. By introducing a recurrent structure and multiple independently learned linear layers, it explores complex interactions between image and text features from various angles. This effectively addresses information redundancy and suddenness issues in traditional bilinear pooling, significantly enhancing the model's accuracy in understanding and answering complex visual questions.In addition, we also combined the recurrent structure with existing improvement methods, and the experimental results further demonstrated that the recurrent structure can effectively enhance the performance of these improvement methods. This indicates that the recurrent structure is not only suitable for basic bilinear pooling techniques, but can also be integrated with other advanced feature fusion techniques, with broad applicability and universality.

Experimental results show that the proposed method achieves remarkable performance improvements on both the VQA_v2 and OKVQA datasets, particularly in handling complex questions requiring external knowledge, where it demonstrates stronger reasoning abilities and higher answer accuracy.

Future research will focus on further optimizing the model structure, exploring its application in other multimodal tasks, and combining it with external knowledge bases to enhance performance in complex scenarios.

Overall, the innovative method presented in this paper offers new insights and effective solutions for multimodal feature fusion, advancing technology in visual question answering and related fields.

References

[1] Kiros R, Zhu Y, Salakhutdinov RR, Zemel R, Urtasun R, Torralba A, Fidler S. Skip-thought vectors. In: Proc. of the Advances in Neural Information Processing Systems. 2015. 3294−3302.

[2] Fukui A, Park DH, Yang D, Rohrbach A, Darrell T, Rohrbach M. Multimodal compact bilinear pooling for visual question answering and visual grounding. In: Proc. of the 2016 Conf. on Empirical Methods in Natural Language Processing. 2016. 457−468.

[3] Kim JH, On KW, Lim W, Kim J, Ha J, Zhang B. Hadamard product for low-rank bilinear pooling. In: Proc. of the Int'l Conf. on Learning Representations. 2017.

[4] Proc. of the IEEE Int'l Conf on Computer Vision. 2017. 1821−1830. [doi: 10.1109/ICCV.2017.202]

[5] Fei Zhao, Chengcui Zhang, and Baocheng Geng. 2024. Deep Multimodal Data Fusion. ACM Comput. Surv. 56, 9, Article 216 (September 2024), 36 pages. https://doi.org/10.1145/3649447

[6] Kim J H , Jun J , Zhang B T .Bilinear Attention Networks[J]. 2018.DOI:10.48550/arXiv.1805.07932.

[7] He K., Zhang X., Ren S., et al. Deep residual learning for image recognition[C]. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2016: 770-778.

[8] Devlin J., Chang M.W., Lee K., et al. Bert: Pre-training of deep bidirectional transformers for language understanding[C]. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Stroudsburg: ACL, 2019: 4171−4186.