Knowledge-Enhanced Visual Question Answering with Multi-modal Joint Guidance

Jianfeng Wang, Anda Zhang, Huifang Du, Haofeng Wang, Wenqiang Zhang
Academy for Engineering and Technology, Fudan University
College of Design and Innovation, Tongji University
Shanghai, China

ABSTRACT

Visual Question Answering (VQA) can facilitate social convenience, which needs to study complex joint reasoning in the visual and language over external knowledge. Recently, Knowledge-Based VQA has attracted the attention of researchers. There are many sources of external knowledge, including visual, textual, and commonsense knowledge, which can effectively improve the reasoning ability of the VQA model. However, introducing different knowledge sources increases the probability of retrieving irrelevant facts and generating noise and further impacts the model's performance. Existing approaches use contrast and prompt learning, visual matrices, density retrieval, etc., to address the noise but bring complex processes. Furthermore, the knowledge representation in these approaches is limited to specific knowledge forms, such as the triple of the knowledge graphs. To address the challenges, we propose a multi-modal joint-guided (MMJG) external knowledge introduction method. The method is to select more relevant external knowledge to the current question through the attention of multi-modal information. Unlike any existing method, our approach learns an adaptive selection module to select external knowledge that is more relevant to the question. Our approach is not specific to a particular knowledge form. The comparison and ablation experiments on the benchmark dataset show that our method achieves better results and demonstrates that our method is more effective.

CCS CONCEPTS

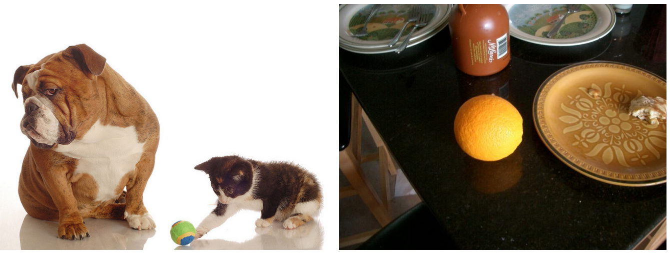
• **Computing methodologies** → **Computer vision**; **Natural language processing**; *Knowledge representation and reasoning*.

KEYWORDS

Visual question answering; multi-modal information; knowledge enhancing; self-attention mechanism

ACM Reference Format:
Jianfeng Wang, Anda Zhang, Huifang Du, Haofeng Wang, Wenqiang Zhang. 2022. Knowledge-Enhanced Visual Question Answering with Multi-modal Joint Guidance. In *Proceedings of In the 11th International Joint Conference on Knowledge Graphs (IJCKG '22).* ACM, New York, NY, USA, 6 pages. https://doi.org/XXXXXXX.XXXXXXX

IJCKG '22, October 27–29, 2022, Hangzhou, China
© 2022 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/XXXXXXX.XXXXXXX



Question: Which animal in the picture can climb a tree?
Answer: cat
Related knowledge: cats can climb trees

Question: What kind of object can be eaten on the table in the picture
Answer: orange
Related knowledge: oranges are edible

Figure 1: Samples of Visual Question Answering involving external knowledge.

1 INTRODUCTION

Visual **Q**uestion **A**nswering (VQA) is a typical multi-modal analysis and reasoning task that aims to understand natural images and related questions to infer the correct answers. VQA is essential to improve the convenience of social media and human-computer interaction, so it has excellent research significance and application prospects.

VQA has made remarkable strides in the past few years. Recently, there has been a trend toward knowledge-based VQA. To answer some challenging questions correctly, the model must employ visual recognition, logical reasoning, and outside knowledge including visual, textual, and commonsense. For example, as shown in Fig. 1, the model needs to obtain the external knowledge about "cats can climb trees" to correctly answer the question "Which animal in the picture can climb trees?".

Considering the above problem, some researches [1, 3, 12, 17, 18, 20, 26, 27] tend toward knowledge-based VQA, and these methods can be classified into three types: methods based on fact database, methods based on external knowledge retrieval, and methods based on a pre-training model. Firstly, the methods based on a fact database generally construct fact triples related to the samples (image and its question), such as FVQA [26], KBVQA [27] and KRVQA [1], which have rich and accurate knowledge to boost the performance of VQA. However, the disadvantage is that triples are limited to a specific form of knowledge which decreases the methods' generalizability. Secondly, the methods based on external knowledge retrieval can obtain more diverse external knowledge using more sources. At the same time, there will be large relevant facts or common sense through text or visual entities search [20], which may introduce noises without fine-grained selection. Thirdly, the approaches based on pre-training include LXMERT [23], ViLBERT [13], and ConceptBert [3]. Pre-trained models can implicitly provide useful prior knowledge for target tasks. However, it is computationally expensive and inefficient for external knowledge learning [12] because the model requires many occurrences of knowledge in the corpus.

To tackle the above challenges, we propose a knowledge-enhanced method based on joint guidance of multi-modal information, which comprehensively considers the joint supervision of visual and textual features in the process of knowledge introduction. It makes the process more accurate and effective. The main idea is to fuse the extracted visual and textual features, using the attention mechanism [24] to guide the introduction of external knowledge of the question entity, and perform VQA reasoning through the knowledge-enhanced questions. In short, our contributions to this paper include:

1) We proposed a knowledge-enhanced VQA method based on multi-modal guidance that uses the joint information of visual and text modalities to guide the introduction of external knowledge. It provides helpful knowledge to answer the question and further avoid adding noise.

2) We design a knowledge-aware attention module. Unlike other methods that apply the specific knowledge form [12, 20], this module can introduce a variety of external knowledge elegantly, such as triplets and passages, without being limited to specific forms of external knowledge.

3) The extensive experimental results show that the knowledge-enhanced method with multi-modal guidance proposed in this paper can improve the performance of the VQA model on the related dataset and outperform the existing methods in many indicators, which proves the feasibility and effectiveness of the method.

2 RELATED WORK

This section will introduce the related research on VQA and the VQA based on external knowledge.

2.1 Visual Question Answering

Malinowski first proposed the related research on VQA in 2014. By combining the latest technologies of natural language processing and computer vision, the computer can automatically answer questions about images. Since this research task was proposed, a large number of VQA datasets have emerged, such as Visual Genome [9], GQA [7], VQA2.0 [4] and OK-VQA [18]. Researchers have also explored many methods related to VQA tasks. Early work mainly focused on multi-modal feature fusion, such as MCB [2]. MFH [31]. Since not all image information is related to questions and answers, irrelevant information should be filtered out in the model inference process. Therefore, improving information extraction and filtering will help the model focus on more relevant information to the task. Most existing methods learn the importance of different parts of the image for question answering based on attention. Common attention-based models are: SAN [29], MCAN [30], and HCAN [14]. At the same time, VQA based on external knowledge has gradually attracted the attention of researchers, which can improve the model's generalization ability by introducing external knowledge.

2.2 Knowledge-based VQA

Knowledge-based VQA requires answering questions with external knowledge in addition to the content of images and questions. For example, some questions need to be answered by using external knowledge other than images and questions, including some common sense and fact-related questions, as shown in Fig. 1.

Researchers have explored knowledge-based VQA methods and tasks. Wang et al. [26] proposed the FVQA dataset and related methods, which parse the question into triples for knowledge base query. Wu et al. [28] used Doc2Vec to feed external knowledge into LSTM for encoding, which enables the model to use external knowledge more flexibly. Marino et al. [17] combined tacit knowledge and explicit symbolic knowledge information. Luo et al. [15] obtained external knowledge through semi-supervised information retrieval.

3 METHODOLOGY

In this section, we specifically elaborate on the model we proposed: Knowledge-Enhanced VQA under multi-modal Joint Guidance. The overall architecture of our model is shown in Figure 2; the model consists of two modules. First, we introduce our knowledge introduction based on multi-modal joint guidance (MMJG), then we describe the reasoning module of knowledge-based VQA, which is inspired by LCGN [5] to reason based on scene graphs over outside knowledge.

Knowledge-based Visual Question Answering can be formulated as follows: Given a picture I and a question Q based on the image in natural language, the VQA model needs to obtain an answer A by integrating visual features, question features, and external knowledge K required. Our model is trained end-to-end, which differs from current multiple-stage methods. The following subsections show more details of the model.

3.1 Knowledge Enhancing by MMJG

We propose a multi-modal joint guidance module that focuses on piloting the filtering method for selecting valuable external knowledge while incorporating external knowledge into the question with significant weights, which learns by attention-based approaches.

3.1.1 Multi-Modal Feature Extraction.

Question Features. We first convert the question into a lower-dimensional feature Q using Global Vectors (GloVe) [19] word embedding. Gloves embedding is an unsupervised learning algorithm that maps words into meaningful space through training.

Visual Features. We choose the VGG [21] network model to extract image features, which uses the weights pre-trained on large databases such as ImageNet [10]. VGG has two structures: VGG16 and VGG19, with different network depths. Since the feature extraction effects of the two models are not much different [21], this paper selects the VGG16 model with fewer network layers. We use the feature extraction network module, composed mainly of five convolution modules and five pooling layers. After the feature extraction module, we connect an AvgPooling (AP) in the feature extraction process. We first cut and zoom the original image to obtain a $3 \times 448 \times 448$ image feature representation as to the input of the feature extraction model. Then obtain a $7 \times 7 \times 512$ dimension feature through the feature extraction module of the pre-trained

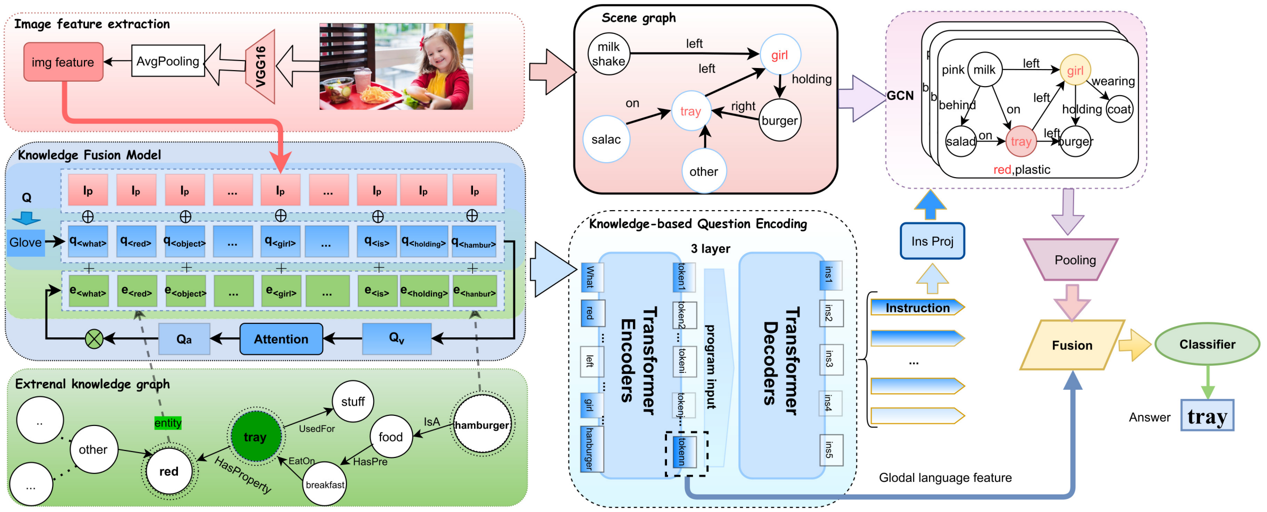


Figure 2: Model architecture. The left part shows our knowledge introduction module of multi-modal joint guidance (MMJG), and the right part is the visual question answering reasoning module based on knowledge-enhanced.

VGG model. Finally, flatten it into a one-dimensional vector I_v with a dimension of 4096, representing the image's visual feature, as shown in the top right of Fig.2.

3.1.2 Knowledge Enhancing.

Multi-modal Fusion Features. To embed multi-modal features into a joint feature, we first map the extracted image features $\boldsymbol{I_v}$ through a Multi-Layer Perceptron (MLP) into a vector $\boldsymbol{I_p} = MLP(\boldsymbol{I_v})$. Then we concatenate the question features to learn the joint multi-modal representations $\boldsymbol{v_{QI\,i}}$, which is calculated by:

$$\boldsymbol{v_{QI\,i}} = \boldsymbol{q_i} \oplus \boldsymbol{I_p}, \boldsymbol{i} \in [1, n] \tag{1}$$

where $\boldsymbol{q_i}$ is the question entity representation, $\boldsymbol{I_p}$ is the image feature representation, and n is the total number of tokens for the question.

Knowledge Fusion. As the Knowledge Fusion Model shows on the left of Fig.2, considering different parts of a visual question that would relate to different external knowledge, we leverage the attention mechanism to dynamically assign greater weight to knowledge instances that are more relevant to the sample. The weight values are as follows:

$$Q_v = [\boldsymbol{v_{QI1}}, \boldsymbol{v_{QI2}}, ..., \boldsymbol{v_{QIn}}] \tag{2}$$

$$Q_a = softmax(FCNet_{w2}(tanh(FCNet_{w1}(Q_v)))) \tag{3}$$

where Q_v is the feature vector of joint multi-modal representations. $FCNet_{w2}$ and $FCNet_{w1}$ is non-linear fully-connected network, $tanh$ is a nonlinear activation function, and Q_a is the corresponding attention coefficients. Unlike the knowledge base used in FVQA [26], we use ConceptNet [22] as an external knowledge base, which is a knowledge graph constructed from several different sources and contains more than 21 million edges as well as more than 8 million nodes.

In this paper, we use the knowledge representation $\boldsymbol{e_{kg}}$ generated by Malaviya [16]. The embedded entity features preserve the knowledge information of nodes and local neighbourhoods. Moreover, the attention score calculation in the attention model has not directly involved the representation of external knowledge. However, it will be adjusted by the answer model's prediction result due to the knowledge selection. Therefore, our model is not restricted to a specific knowledge representation and can be used for various manifestations, such as knowledge triples or documents. The fusion of external knowledge with attention to different parts of the question is shown below:

$$\boldsymbol{q_{k\,i}} = MLP(\boldsymbol{e_{kg}}) * q_{ai} + \boldsymbol{q_i} \tag{4}$$

where the q_{ai} is the attention score for token $\boldsymbol{q_i}$ of the question from Q_a, and the $\boldsymbol{q_{k\,i}}$ represents the token after knowledge enhancement.

Our model assigns the importance of knowledge by multiplying the attention score q_{ai} with the knowledge embedding vector $\boldsymbol{e_{kg}}$, which then be used to enhance the question by adding moderate external knowledge to the question token $\boldsymbol{q_i}$.

3.2 Knowledge-based Reason VQA

We use a graph reasoning module to perform joint reasoning over knowledge and scene graphs during answer prediction, which is trained using an end-to-end model, Moreover, the scene graphs are provided by annotations describing the related objects and attributes in an image in GQA datasets, in which each image is associated with the scene graph from Visual Genome [9].

Graph Reasoning Module.

Referring to the LCGN [5] general VQA reasoning framework, its convolutional graph neural network is used as the VQA reasoning module in our method. Specifically, the graph network is built on the visual entities in the scene graph and constructs a context representation for the objects in the visual scene to support relation reasoning.

The convolutional graph network is a kind of network for non-fixed graph-structured data, which mainly studies the representation embedding of graph nodes. This paper uses the graph attention convolutional network (GAT) [25], which differs from the basic GCN in that the network learns different weights on neighbor nodes by self-attention mechanism in the process of aggregating neighbor information. For nodes in the graph, the neighbor node's feature information is aggregated in each layer of the convolutional network, so it is necessary to calculate the attention score of each neighbor node:

$$ne_{ij} = \alpha(\boldsymbol{Wh_i}, \boldsymbol{Wh_j}) \tag{5}$$

where ne_{ij} denotes the importance coefficient of neighbor node j, α is the attention value function, a single-layer feedforward neural network in our model, and $\boldsymbol{h_i}$ denotes the hidden layer feature vector of the nodes in the graph.

In our model, the graph nodes are initialized by embedding the object's text description and related attributes in the scene graph, and we get the embedding by using GloVe [19].

Since each node in the graph has multiple neighbor nodes, the attention coefficients of all neighbor nodes need to be normalized, as shown in the following formula:

$$na_{ij} = \frac{\exp(ne_{ij})}{\sum_{k\epsilon N_i} \exp(e_{ik})} \tag{6}$$

where N_i denotes the set of neighboring nodes of node i, na_{ij} represents the importance coefficient after normalization. Then each node feature is weighted and summed by the attention coefficient to obtain the updated graph node features:

$$\boldsymbol{h}_{\boldsymbol{i}}^{'} = \sigma(\sum_{k\epsilon N_i} na_{ij} \boldsymbol{Wh_i}) \tag{7}$$

where σ is the nonlinear activation function and $\boldsymbol{h}_{\boldsymbol{i}}^{'}$ denotes the graph node representation that aggregates information about the neighborhood nodes.

Answering Module. As shown in the right of Fig. 2, the model first encodes the knowledge-enhanced question. We use the Transformer encoder to encode the question Q_k, get the M vector of encoding instructions, and then perform the graph convolution. In the product process, we connect the instruction vector with the edge node features in the scene graph to make it worthwhile for inference.

We perform maximum pooling on the feature vectors of all graph nodes in the last layer of graph convolution to combine the inference representation vector and then send it to the VQA classifier with the question features. Last, the model takes the predicted answer through the highest probability as the forecast result.

4 EXPERIMENT

In this section, we first describe the dataset and metrics that we used in our experiments. Next, we train our model with the proposed objective and compare its performance to several baselines. Experiments conducted on the GQA dataset show that our method outperforms comparative approaches, and an ablation study demonstrates the effectiveness of our model.

4.1 Dataset and Metrics

Hudson developed the GQA dataset at Stanford University [7]. Each image in the dataset corresponds to a scene graph, constructed mainly by referring to the Visual Genome [9] dataset. Each question is associated with a scene graph. The dataset contains more than 110K scene graphs, up to 20M pictures with authentic images from the Visual Genome dataset, more than 1 million balanced question-answer pairs, and over 1000 answer tokens, making it the largest VQA dataset to date. At the same time, the dataset has its evaluation metrics, and multi-dimensional evaluation metrics are proposed, mainly variants of accuracy, to measure the performance of relevant visual question answering models, as shown in Table 1.

Table 1: Description of evaluation indicators.

Evaluation	Description
Consistency	A metric for the level of consistency in responses across different questions.
Validity	Measures whether the model gives valid answers.
Plausibility	Measures whether the model responses are reasonable in the real world or not make sense.
Accuracy	Standard accuracy.

4.2 Performance Evaluation

We evaluate our method on the GQA dataset. We use the official balanced train set to train the models and the balanced validation set to evaluate the newly proposed GQA dataset v2, divided by 88% and 12%. The Adam optimizer was used for optimization with a learning rate of 1e-4. We train the model for 100 epochs with batch size 128 on four NVIDIA RTX2080 cards. We tune all hyperparameters by cross-experimenting.

Table 2 shows our experimental results, including performance on different types of questions and various evaluation metrics. We compared our method with the existing visual question answering methods, including GCN [8], LCGN [5] and GINE [6] that applied to the GQA dataset. We also reproduced and tested the method of GraphVQA [11] based on the official dataset. We can see that our method help improve model performance, especially in the binary classification of questions, which outperforms the base method by 2 percentage points. It shows that the external knowledge introduced by our method helps improve the model's predictive ability on questions related to content judgment.

Our model also outperforms by 1 percentage point in the open class, which means that our method is more efficient for open-ended questions. In terms of consistency, effectiveness, and rationality, our model also performs better than the previous methods, which reflects that our method helps to improve the rationality and effectiveness of the visual question and answer model prediction. At the same time, our model outperforms the previous method in overall accuracy by 1 percentage point, further proving our method's effectiveness.

Fig. 3 shows the subdivision accuracy of question types. Our model has been significantly improved compared to the baseline in the accuracy of selection, comparison, and verification, which is enough to show that our model has a better ability for questions that need outside knowledge.

Table 2: Experimental results for the GQA dataset, with indicators in percentages. Where * represents reproduced results.

Method	Binary	Open	Consistency	Validity	Plausibility	Accuracy
GCN	86.84	84.63	90.21	95.51	94.44	85.70
LCGN	90.57	88.43	93.88	95.40	93.89	88.43
GINE	92.36	88.56	94.79	95.44	94.39	90.38
GraphVQA*	94.20	92.60	97.81	95.46	94.97	93.38
Ours	**96.42**	**93.37**	**98.45**	**95.51**	**95.14**	**94.84**

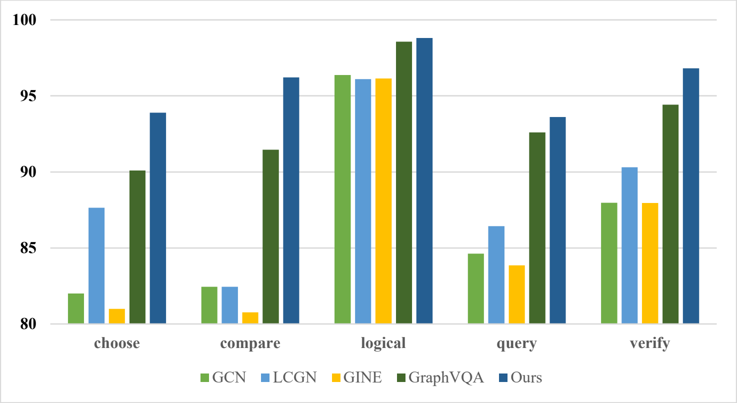


Figure 3: Subdivision accuracy of the question type, with the ordinate as a percentage.

4.3 Ablation Study

In order to verify the effectiveness of our method, we conducted ablation experiments. We tested the knowledge introduction methods based on the model's text modal guidance and multi-modal guidance. We compared them with the benchmark model to verify the role of multi-mode joint information in the process of knowledge introduction.

Table 3: Knowledge introduction Ablation Experiment. All numbers are in percentages.

Method	Accuracy
Baseline	93.38
text modal guidance	93.46
multi-modal guidance	**94.83**

Table 3 shows that the method based on multi-modal joint guidance improves the accuracy of visual question answering by nearly 1.5 percentage points. Compared with the guidance method based on text mode, the performance of multi-modal joint guidance is generally better, proving that multi-modal information plays an essential role in the process of knowledge introduction.

5 CONCLUSION

In this paper, we propose a knowledge-enhanced visual question answering method based on multi-modal joint guidance. The goal is to guide the knowledge introduction process through joint visual and text modal information based on the attention mechanism. The model can adaptively select the most helpful knowledge for the question. At the same time, our method is not restricted to a specific knowledge presentation. The experimental results of our model on the GQA dataset show the effectiveness of our method, which has been significantly improved in various indicators. We will analyze the effect of our knowledge introduction method on different datasets to verify the universality of our method in the future,

REFERENCES

[1] Qingxing Cao, Bailin Li, Xiaodan Liang, Keze Wang, and Liang Lin. 2021. Knowledge-routed visual question reasoning: Challenges for deep representation embedding. *IEEE Transactions on Neural Networks and Learning Systems* (2021).

[2] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847* (2016).

[3] François Gardères, Maryam Ziaeefard, Baptiste Abeloos, and Freddy Lecue. 2020. Conceptbert: Concept-aware representation for visual question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. 489–498.

[4] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6904–6913.

[5] Ronghang Hu, Anna Rohrbach, Trevor Darrell, and Kate Saenko. 2019. Language-conditioned graph networks for relational reasoning. In *Proceedings of the IEEE/CVF international conference on computer vision*. 10294–10303.

[6] Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec. 2019. Strategies for pre-training graph neural networks. *arXiv preprint arXiv:1905.12265* (2019).

[7] Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 6700–6709.

[8] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).

[9] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision* 123, 1 (2017), 32–73.

[10] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2017. ImageNet classification with deep convolutional neural networks. *Commun. ACM* 60, 6 (2017), 84–90. https://doi.org/10.1145/3065386

[11] Weixin Liang, Yanhao Jiang, and Zixuan Liu. 2021. GraghVQA: language-guided graph neural networks for graph-based visual question answering. *arXiv preprint arXiv:2104.10283* (2021).

[12] Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2020. K-bert: Enabling language representation with knowledge graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 2901–2908.

[13] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems* 32 (2019).

[14] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. *Advances in neural information processing systems* 29 (2016).

[15] Man Luo, Yankai Zeng, Pratyay Banerjee, and Chitta Baral. 2021. Weakly-supervised visual-retriever-reader for knowledge-based question answering. *arXiv preprint arXiv:2109.04014* (2021).

[16] Chaitanya Malaviya, Chandra Bhagavatula, Antoine Bosselut, and Yejin Choi. 2020. Commonsense knowledge base completion with structural and semantic context. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 2925–2933.

[17] Kenneth Marino, Xinlei Chen, Devi Parikh, Abhinav Gupta, and Marcus Rohrbach. 2021. Krisp: Integrating implicit and symbolic knowledge for open-domain knowledge-based vqa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14111–14121.

[18] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*. 3195–3204.

[19] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[20] Chen Qu, Hamed Zamani, Liu Yang, W Bruce Croft, and Erik Learned-Miller. 2021. Passage Retrieval for Outside-Knowledge Visual Question Answering. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1753–1757.

[21] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[22] Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-first AAAI conference on artificial intelligence*.

[23] Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490* (2019).

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[25] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903* (2017).

[26] Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. 2017. Fvqa: Fact-based visual question answering. *IEEE transactions on pattern analysis and machine intelligence* 40, 10 (2017), 2413–2427.

[27] Peng Wang, Qi Wu, Chunhua Shen, Anton van den Hengel, and Anthony Dick. 2015. Explicit knowledge-based reasoning for visual question answering. *arXiv preprint arXiv:1511.02570* (2015).

[28] Qi Wu, Peng Wang, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. 2016. 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*. 4622–4630.

[29] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. 2016. Stacked attention networks for image question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 21–29.

[30] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. 2019. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 6281–6290.

[31] Zhou Yu, Jun Yu, Chenchao Xiang, Jianping Fan, and Dacheng Tao. 2018. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. *IEEE transactions on neural networks and learning systems* 29, 12 (2018), 5947–5959.