Visual Question Answering based on multimodal triplet knowledge accumuation

Fengjuan Wang[1,2], Gaoyun An[1,2*]
[1]Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China
[2]Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing 100044, China
21112009@bjtu.edu.cn, gyan@bjtu.edu.cn

2022 16th IEEE International Conference on Signal Processing (ICSP) | 978-1-6654-6056-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICSP56322.2022.9965282

***Abstract*—Knowledge-based visual Question Answer needs to be associated with external knowledge to better understand. Existing solutions generally acquire relevant knowledge from pure text knowledge bases, but these knowledge bases only contain some facts for simple questions and answers, lacking visual content for deep understanding. This paper proposes an explicit triplet to represent multimodal knowledge, which associates visual objects and factual answers with implicit relationships. The header is obtained by calculating the Euclidean distance between the regional feature and the text token. To narrow the heterogeneity gap, we propose three target losses of learning triplet representation from the perspective of complementarity, using TransH, topological relations, and semantic space to calculate loss respectively. By adopting unique learning strategies, we gradually accumulate multimodal knowledge in specific fields and predict answers. On two knowledge-based datasets: ok-vqa and kr-vqa, the performance is higher than that of the latest method respectively. The experimental results prove the validity of the proposed model.**

***Index Terms*—Visual Question Answer, Multimodal knowledge represent, Explicit triplet.**

I. Introduction

Visual question answering tasks generally require the use of external knowledge bases to finds the relevant knowledge content of accumulated images and questions to answer the questions. Although great success has been achieved in the VQA task, kb-vqa is more challenging for the model.

Most recent work has focused on acquiring knowledge from the relevant structured knowledge graph and unstructured knowledge integration, such as Conceptnet [9] and Visual Genome [**?**]. Figure 1 (a) in addition to the understanding of objects, when people are asked to distinguish simple scenes, such as ”can you guess where that place is?” The explicit visual knowledge in our brain is very important. Therefore, how to accumulate complex, understandable, and highly relevant multimodal knowledge in VQA scenarios is an important issue, which is worthy of in-depth study.

The latest knowledge graph is all about to expand visual content and text facts at the same time, so as to form a more comprehensible and more enhanced knowledge graph [8]. Existing methods can generally be divided into two approaches, one is to structure visual features and questions, and the other is to tag entities through multimodal [14]. However, this multimodal knowledge graph still represents knowledge by first-order predicates in essence, which cannot to understand complex multi-level relationships. For example, the relationship between ”motorcycle” and ”Kawasaki” in Figure 1 (a).

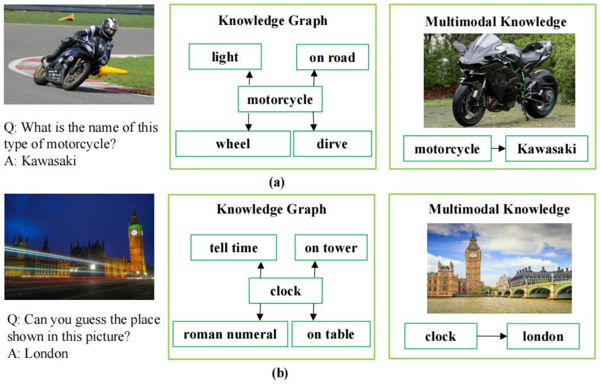

Fig. 1. An illustration of our motivation.

This paper propose three uniquely useful target loss functions, and refine the entity representation by comparing the positive and negative triples. Learning strategies combining pre-training and fine-tuning is adopted to gradually accumulate comprehensible multimodal knowledge from VQA samples outside and inside for interpretable reasoning.The innovation points of this article mainly include the following two points:

(1) the model forms multimodal facts through explicit triples. this paper uses Euclidean distance to obtain the required head entities by calculating the Euclidean distance of regional features and text tokens.

(2) We use the unique learning strategies to accumulate comprehensible knowledge within and outside to form a multimodal knowledge base. Using multiple loss functions to calculate the predicted answers of the header and related entities, and support automatic knowledge association. [11].

II. Related work

Most of the recent work is firstly based on searching relevant and important knowledge facts from external knowledge base, then use knowledge atlas to make explicit reasoning of the model, or embeds implicit knowledge into images and questions to classify answers [5]. These methods will rely on target tags to search for external knowledge, which will inevitably introduce a lot of irrelevant knowledge and noise, and will certainly cause a series of errors. This paper

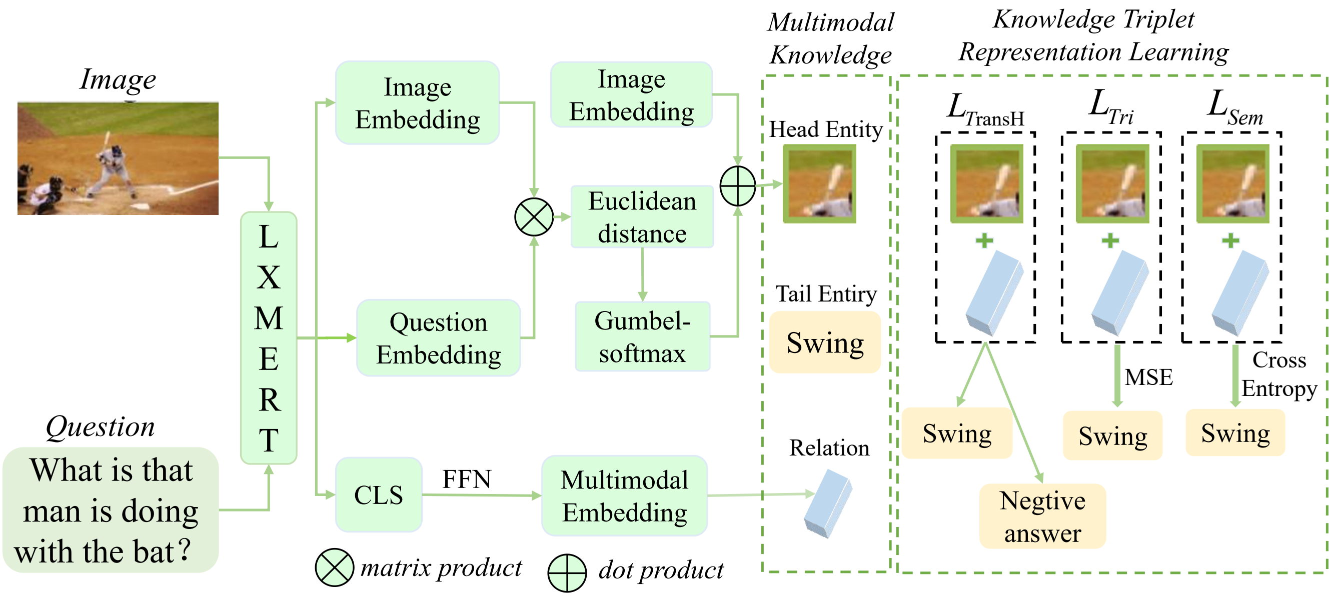

Fig. 2. The overall architecture of our model.

proposes a multimodal model based on knowledge learning and accumulation.which includes interpretable triple knowledge representation learning, Multimodal knowledge graph: the latest multi-modal knowledge graph with strong understanding aims to closely assemble visual content and text facts to automatically form a multi-modal knowledge graph with strong correlation and comprehensibility. [8]. Mainstream approach is to first structured representation of images and texts into an understandable structure graph and mark the available entities across multimodel. The important problem lies in the process of intra-model relationship extraction and cross-model entity connection, specifically learning knowledge from structured text and visual contents, and maintaining the triple structure of entity alignment. In essence, these methods still express knowledge by first-order predicates described in natural language, and they can not including multi-level complex and useful implicit relation.

III. Methodology

Given a picture and a question related to the picture, the purpose of the knowledge base of visual question answer task is to predict the corresponding answer that requires additional practical knowledge beyond the content provided by pictures and text.This paper takes the accumulated multimodal knowledge as external knowledge and infers the answer directly. Fig.2 gives a detailed description of the model. First,based on the unstructured knowledge graph, text embedding and visual embedding are extracted from the pre-trained language model LXMERT. Then three losses are proposed to learn triple embedding, which accurately describes the visual content of the question (head embedding), the fact answer of the question expectation (tail embedding), and the implicit relationship between them (relationship embedding). By using the data outside and inside the domain for training, the model accumulates extensive multimodal knowledge through training and learning, and selects the fact most closely related to the answer as the answer prediction.

A. Multimodal triplet

In the task of visual question answer, we form a multimodal triple of complex and indispensable facts, namely, header, relationship and tailer. that is (h, r, t), where h contains the visual content related to the question, t indicates the true answer to the question given, and r describes the implicit close relationship between visual features and text contents. The structured representation of triples consists of the following four steps:

- Image feature extraction and problem encoding. Because the pre-trained visual language model is very effective in modeling the implicit correlation, the paper first uses the pre-trained language model to encode the problem and image features to obtain the corresponding embedding representation.
- Header Entity Extraction. header entity is defined as the required visual object most relevant to the problem and its context in the image. Therefore, the correlation between each object in the related visual features and textual facts is evaluated by calculating Euclidean distance. we choose an object as the visual content most relevant to the problem. then obtain an approximate one-hot classification distribution. According to the problem centered object information collected, the head entity is obtained.

$$h = FFN(\sum_{i=1}^{k} a_i v_i) \tag{1}$$

TABLE I
STATE-OF-ART RESULTS COMPARISION ON OK-VQA DATESET

Method	Knowledge resources	Accuracy
MUTAN+AN [12]	Wikipedia	25.61
Mucko [20]	ConceptNet	29.20
GRUC [17]	ConceptNet	29.87
KM [18]	Multimodal knowledge from OK-VQA	31.32
ViLBERT [10]	-	31.35
LXMERT [15]	-	32.04
ConceptBert [5]	ConceptNet	33.66
KRISP [16]	DBpedia+ConceptNet+VisualGenome+haspartKB	38.90
Knowledge is Power [19]	YAGO3	39.42
MuKEA [4]	Knowledge from VQA 2.0 and OK-VQA	42.59
Ours	Knowledge from VQA 2.0 and OK-VQA	**43.06**

where a_i is one-hot categorical distribution, v_i is the visual embeddings, FFN represents a feedward network.

- Relation Extraction. The relationship in multimodal knowledge graph is defined as the close and complex implicit relationship between visual objects and text facts. LXMERT extracts the cross-model representation from the [CLS] tag and relational embedding is get by sending it to the FFN layer.
- Tailer Entity Extraction. The answer is severed as tail entity in the relevant image question samples, which reveals the important information of the visual object related to the question. we keep learning and accumulating in the multi-modal knowledge base, and finally predict the tail entity with the highest score as the answer.

B. Triplet Represents Learning

Since each triplet extracted contains semantic space information of different modes and used three loss functions with unique design to uniformly learn the structured representation of triplets. These three losse functions limit the representation of triplets from a complementary perspective, and Triplet TransH Loss preserves the embedded structure by comparing different triplets. Triplet consistency loss forces the image and question embeddings in the triplet to keep strict and consistent topological relationship for more accurate information and finally semantic consistency loss map the text embedding to the common semantic space to obtain more fair comparison among multimodes.

Triplet TransH Loss. Given an image-question pair, h and r represents head entity and tail entity respectively, each positive tail $t+$ and each negative tail $t-$.The tail entity t corresponds to a hyperplane, and the h and r in the entity space are projected onto this hyperplane through matrix mapping. The projection method is in the form of matrix multiplication. Depending on the corresponding entity, the projected vectors are:

$$h_{\perp} = h - w_t^T h w_t \tag{2}$$

$$r_{\perp} = r - w_t^T r w_t \tag{3}$$

The final scoring function is expressed by the following formula:

$$f_t(h,r) = ||h - w_t^T h w_t + d_r - r - w_t^T r w_t||_2^2 \tag{4}$$

The trained loss function adopts a negative sampling method that minimizes the score for correct triples and maximizes the score for incorrect triples:

$$L_{TransH} = \sum_{(h,r,t)\in\triangle} \sum_{(h',r',t')\in\triangle'} [\gamma + f_t(h,r) - f_t(h',r')]_+ \tag{5}$$

Triplet Consistency Loss. In order to further promote the embedding to meet the strict topological relationship, Triplet Consistency Loss applies mean square error to constrain the representation at the top of each positive triple as:

$$L_{Tri} = MSE(h + r, t_+) \tag{6}$$

Semantic Consistency Loss. In order to narrow the heterogeneous gap between the header entity and relation entity, this paper selected the negative log likelihood loss as the ground truth tail:

$$P(t_+) = sofymax((T)(h+r))L_{sem} = -log(P(t_+)) \tag{7}$$

$$L_{sem} = -log(P(t_+)) \tag{8}$$

where P is the prediction probability of the true value, and the final loss function is expressed by the following formula:

$$L = L_{TransH} + T_{Tri} + T_{Sem} \tag{9}$$

C. Knowledge accumulation and answer prediction

This paper has its own unique training strategy:(1) First, conduct pre-training on the VQA 2.0 [6] dataset to accumulate understandable multimodal knowledge as external knowledge. (2) Then fine-tune the external knowledge dataset to obtain a model with better generalization and predict the answer.

IV. Experiment

Experiments were conducted on two commonly used visual question and answer datasets: OK-VQA [12] and KRVQA [3]. It is based on the multi-step reasoning ability of the external knowledge evaluation model. This paper uses top-1 precision for comparison.

Table 1 shows the comparison results with on OK-VQA, including the method based on knowledge graph [12] [19] [20], the hybrid method based on multi-source knowledge [7], the pre-training method based on implicit knowledge [10] and the method based on multimodal knowledge [18]. Our model is superior to the the existing state model [4].Compared with most models that follow the "knowledge retrieval and reading" process, our model effectively avoids cascading errors.

TABLE II
STATE-OF-ART RESULTS COMPARISION ON KRVQA

Method	**KB-related**		**Accuracy**
	one-step	**two-step**	
Q-type [2]	0.09	0.33	8.12
LSTM [2]	0.43	0.74	8.81
FiLM [13]	6.27	7.19	16.89
MCAN [17]	12.28	13.34	20.52
UpDn [1]	8.16	13.97	21.85
Mukea [4]	6.14	18.28	27.38
Ours	-	-	**27.42**

The experimental results on KRVQA: In Table 2, we compared the model with the VQA model [17] [13] and the knowledge-based model [2]. Problems unrelated to the knowledge base only require basic visual information, while problems related to the knowledge base require factual knowledge in the knowledge base. Our model is superior to the existing model.

V. Conclusion

This paper discusses knowledge visual Question Answer from a new perspective. First, it proposes a triplet related to visual features and question answers, Through three unique loss functions and learning strategy, the model continuously accumulates common knowledge facts on vqa2.0 dataset as an external knowledge base, and questions are predicted based on the external knowledge base and relevant knowledge searched on ok-vqa. It provides a new idea for the knowledge based visual question answering model. The performance of the model on KB-VQA dataset is better than the latest technology.

References

[1] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. 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*, pages 6077–6086, 2018.

[2] Q. Cao, B. Li, X. Liang, K. Wang, and L. Lin. Knowledge-routed visual question reasoning: Challenges for deep representation embedding. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[3] Q. Cao, B. Li, X. Liang, K. Wang, and L. Lin. Knowledge-routed visual question reasoning: Challenges for deep representation embedding. *IEEE Trans. Neural Networks Learn. Syst.*, 33(7):2758–2767, 2022.

[4] Y. Ding, J. Yu, B. Liu, Y. Hu, M. Cui, and Q. Wu. Mukea: Multimodal knowledge extraction and accumulation for knowledge-based visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5089–5098, 2022.

[5] F. Gardères, M. Ziaeefard, B. Abeloos, and F. Lecue. Conceptbert: Concept-aware representation for visual question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 489–498, 2020.

[6] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. 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*, pages 6904–6913, 2017.

[7] G. Li, X. Wang, and W. Zhu. Boosting visual question answering with context-aware knowledge aggregation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1227–1235, 2020.

[8] M. Li, A. Zareian, Y. Lin, X. Pan, S. Whitehead, B. Chen, B. Wu, H. Ji, S.-F. Chang, C. Voss, et al. Gaia: A fine-grained multimedia knowledge extraction system. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 77–86, 2020.

[9] H. Liu and P. Singh. Conceptnet—a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226, 2004.

[10] J. Lu, D. Batra, D. Parikh, and S. Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

[11] K. Marino, X. Chen, D. Parikh, A. Gupta, and M. Rohrbach. 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*, pages 14111–14121, 2021.

[12] K. Marino, M. Rastegari, A. Farhadi, and R. Mottaghi. OK-VQA: A visual question answering benchmark requiring external knowledge. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 3195–3204. Computer Vision Foundation / IEEE, 2019.

[13] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[14] R. Sun, X. Cao, Y. Zhao, J. Wan, K. Zhou, F. Zhang, Z. Wang, and K. Zheng. Multi-modal knowledge graphs for recommender systems. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pages 1405–1414, 2020.

[15] H. Tan and M. Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

[16] J. Yu, Z. Zhu, Y. Wang, W. Zhang, Y. Hu, and J. Tan. Cross-modal knowledge reasoning for knowledge-based visual question answering. *Pattern Recognition*, 108:107563, 2020.

[17] Z. Yu, J. Yu, Y. Cui, D. Tao, and Q. Tian. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6281–6290, 2019.

[18] W. Zheng, L. Yan, C. Gou, and F.-Y. Wang. Km4: visual reasoning via knowledge embedding memory model with mutual modulation. *Information Fusion*, 67:14–28, 2021.

[19] W. Zheng, L. Yan, C. Gou, and F.-Y. Wang. Knowledge is power: Hierarchical-knowledge embedded meta-learning for visual reasoning in artistic domains. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 2360–2368, 2021.

[20] Z. Zhu, J. Yu, Y. Wang, Y. Sun, Y. Hu, and Q. Wu. Mucko: multi-layer cross-modal knowledge reasoning for fact-based visual question answering. *arXiv preprint arXiv:2006.09073*, 2020.