MKEAH: Multimodal Knowledge Extraction and Accumulation Based on Hyperplane Embedding for Knowledge-Based Visual Question Answering

Heng Zhang, Zhihua Wei, Guanming Liu, Ruibin Mu, Rui Wang, Chuan Bao Liu, Aiquan Yuan, Guodong Cao and Ning Hu

August 22, 2023

CGI2023 manuscript No.
(will be inserted by the editor)

MKEAH: Multimodal Knowledge Extraction and Accumulation Based on Hyperplane Embedding for Knowledge-based Visual Question Answering

Abstract External knowledge representations play an essential role in knowledge-based visual question and answering to better understand complex scenarios in the open world. Recent entity-relationship embedding approaches are deficient in some of representing complex relations, resulting in a lack of topic-related knowledge but the redundancy of topic-irrelevant information. To this end, we propose MKEAH to represent Multimodal Knowledge Extraction and Accumulation on Hyperplanes. To ensure that the length of the feature vectors projected to the hyperplane compares equally and to filter out enough topic-irrelevant information, two losses are proposed to learn the triplet representations from the complementary views: range loss and orthogonal loss. In order to interpret the capability of extracting topic-related knowledge, we present Topic Similarity (TS) between topic and entity-relation. Experimental results demonstrate the effectiveness of hyperplane embedding for knowledge representation in knowledge-based visual question answering. Our model outperforms the state-of-the-art methods by 2.12% and 3.24%, respectively, on two challenging knowledge-required datasets: OK-VQA and KRVQA. The obvious advantages of our model on TS shows that using hyperplane embedding to represent multimodal knowledge can improve the ability of the model to extract topic-related knowledge.

Keywords Knowledge-based Visual Question Answering · Hyperplane · Topic-related

Address(es) of author(s) should be given

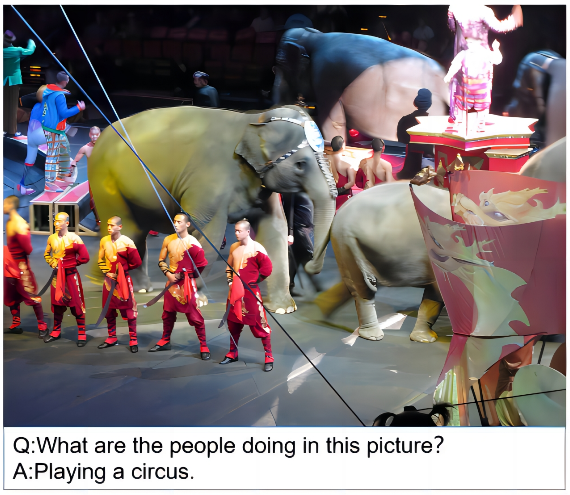

Q:What are the people doing in this picture?
A:Playing a circus.

Fig. 1 An illustration of our motivation. The questions in the picture must be answered with some outside knowledge.

1 Introduction

Advances in deep learning promote the grow-th of visual question and answering tasks [24–26]. Knowledge-based visual question and answering (KB-VQA) [13] requires associating external knowledge to realize open cross-modal scene understanding. Some questions require the model to understand a certain amount of common sense, i.e., external knowledge. As shown in Figure 1, the model relies on outside knowledge to determine what activity people are doing. How to represent external knowledge and make the model understand external knowledge is a significant challenge for the knowledge-based visual question and answering task.

There are many complex relationships in the real world, such as one-to-many, many-to-one, reflexive, many-to-many, and so on. The recent visual question and answering works do not work well in representing these relationships. Also, they lack the ability to extract topic-related knowledge. In a visual question-and-answer task, the related information is often only part of the image [8], while the other part is irrelevant to the topic. Representing this content directly as a header entity affects the capability of the model to catch real-world knowledge.

In this paper, we propose MKEAH, a novel representation of knowledge in KB-VQA. The core mechanism of MKEAH is to apply the hyperplane to mine the higher-order logical relationship between knowledge representation and questions, filtering out information irrelevant to the question topic. Specifically, we first propose a hyperplane transformation embedding method to represent the triplet in the multimodal knowledge graph [3, 15, 21], where the head entity embedding comes from the visual object, the tail entity embedding corresponds to the ground truth, and relational embedding [20] represents an implicit association between the head and tail entities. To help hyperplane embedding representation learn topic-related information and filter topic-irrelevant information, We propose two loss functions: range loss and orthogonal loss. We then propose a novel metric of the ability of the model to extract topic-related information, which is conducive to further improve the capability of represent real-world knowledge.

The main contributions of this work are summarized as follows.

1. A hyperplane embedding model is proposed to improve the capability of the model to extract topic-related knowledge.
2. Two hyperplane embedding loss functions, i.e., scale loss and orthogonal loss, are proposed to help hyperplane learn topic-related knowledge and topic-irrelevant information.
3. A corresponding evaluation metric TS is proposed. After comparing our model with the SOTA model in this metric, it can be found that our model has more advantages. We outperformed SOTA methods by 2.12% and 3.24%, respectively, on two challenging knowledge requirement datasets: OK-VQA [34] and KRVQA [6].

2 Related Work

2.1 Knowledge-based visual question and answering

Most recent work [4, 5, 9] for Knowledge-based visual question and answer are based on constructing triplets in the original space, lacking of the ability to represents complex relations in high-order logic. Narasimhan and Schwing [37] propose to retrieve relevant facts from a knowledge base. Wang et al.[38] desige a system to find the mappings from the question to a query triplet. Yang et al. [17] uses an end-to-end multimodal knowledge representation [16] learning framework, which first models the inexpressible multimodal facts by explicit triplets and provides complementary knowledge with the existing knowledge graphs [30] and unstructured knowledge bases. Wu et al. [39] uses retrieved knowledge for answer validation rather than for producing the answer based on a three-stage framework. Graph embedding [48] is proposed to describe some complex relation structures in real world, but it also brings high computational and spatial costs.

2.2 Multimodal Knowledge Graph

The aim of the emerging multimodal knowledge graph is to create a more comprehensive knowledge graph by linking image content and text facts. Common approaches involve converting images and text into structured representations, followed by cross-modal event/entity processing. However, a major challenge is to extract relationships within each modality and establish connections between entities across different modalities. Several methods have been proposed to address this challenge. Li et al. [41] developed a model that learns from structured text and visual data and maintains triplets to ensure entity alignment. Kannan et al. [15] used RDF [42] knowledge graphs to represent multimodal information based on graph alignment, but lacked the ability to capture multimodal correlation. Another approach is to link entities in an existing knowledge graph directly with relevant images, as demonstrated by Pezes-hkpour et al. [40], who expanded the representation of YAGO [36] entities by including images. However, all of these methods are limited by the use of first-order predicate knowledge representation described in natural language, and thus cannot effectively model higher-order complex relationships.

2.3 Embedding Models

Bordes et al. [43] uses relation embedding r to associate the two embedded entities in a triplet (h, r, t), and confirms the advantage of such a structure. Bordes et al. [44] introduces two independent projections to the entities in a relation. Jenatton et al. [45] models second-order correlations between entity embeddings by a quad-ratic form. In Single Layer Model [46], a structure is proposed to solve the problem of nonlinear transformation in neural networks. NTN [46] applies nonlinear transformation to the second-order correlation transformation, based on the Single Layer Model.

3 Methodology

In this section we introduce the mechanism of our model MKEAH. Given an image I and a question Q, the goal of the KB-VQA task is to predict an answer A, supported by knowledge of the outside world beyond the question and answer itself. We accumulate the multimodal knowledge of the triplet formation as external knowledge and deduce the answer directly in an end-to-end mode. Figure 2 shows a detailed illustration of our model. We first propose MKEAH to represent Multimodal Knowledge Extraction and Accumulation on Hyperplanes. Then two losses are proposed to learn the triplet representations from the complementary views and ensure that our knowledge representation is suitable for the new hyperplane. Through the training of both out-domain and in-domain data [7], our model extracts a broad range of multimodal knowledge and effectively bridges the most suitable facts to generate accurate answers.

3.1 Hyperplane Embedding Triplet Extraction

To overcome TransE's poor ability in establishing reflexive, one-to-many, many-to-one or many-to-many complex relationships and extract topic-related knowledge, we propose a model that learns different distributed representations when entities involve different relations. As shown in Figure 2, the model MKEAH extracts different entity- relation representations on the hyperplanes. For a relation r, we place the relation specific translation vector d_r in the relation specific hyperplane w_r (normal vector) rather than in the same space where the entity is embedded. Specifically, for triplets (h, r, t), the embeddings h and t are first projected into the hyperplanes w_r.

3.2 Hyperplane Embedding Knowledge Triplets Representation Learning

3.2.1 Triplet Transmission Loss

We apply TransH-like target loss as structure preserving constraint in multi-modal scenario, inspired by the knowledge embedding method TransH [2]. Given an image-question pair, let A+ and A- represent the set of positive and negative answers, respectively. Let $h_\perp$ and $t_\perp$ represent the corresponding head and tail entities embedded in the hyperplane. We can get $h_\perp$ and $t_\perp$as follow:

$$\begin{aligned} h_\perp &= h - w_r^\top h w_r, \\ t_\perp &= t - w_r^\top t w_r \end{aligned} \tag{1}$$

In the ideal case, the sum of the head entity and the relationship representation is as close to the tail entity as possible, if a triplet satisfies this property, we call it golden triplet. Inspired by the evaluation method of TransH[2], the score function we use to determines whether a triplet is golden triplet is like:

$$f_r\left(h, t\right) = \left\|h_\perp + d_r - t_\perp\right\|_2^2 \tag{2}$$

The closer the value of $f_r\left(h, t\right)$ is to 0, the more the triplet is an ideal golden triplet.

The certain boundary γ is used for ensuring that the distance between positive and negative samples can arrive a certain standard. Then we represent triplet transmission loss as:

$$\mathcal{L}_{TransH} = \sum_{t^+ \in A^+} \sum_{t^- \in A^-} \left[f_r\left(h, t^+\right) + \gamma - f_r\left(h, t^-\right)\right]_+ \tag{3}$$

3.2.2 Hyperplane Embedding Triplet Consistency Loss

The triplet transmission loss has a disadvantage that, once the absolute value of the distance between positive pairs minus the distance between negative pairs arrives γ, the process of knowledge extraction and accumulation stops. To further push the hyperplane embedding to satisfy the golden triplet structure, we apply the mean square error (MSE) criterion to push the hyperplane embedded triplets forward the standard of gold triplets:

$$\mathcal{L}_{hyper\_tri} = MSE\left(h_\perp + r_\perp, t^+\right) \tag{4}$$

3.2.3 Hyperplane Embedding Semantic Consistency Loss

Like MuKEA [17], We use semantic consistency loss to narrow the heterogeneous gap between tail entity and head entity and relation. In order to ensure that the model chooses ground-truth tail, we use negative logarithmic likelihood loss:

$$\mathcal{L}_{hyper\_sem} = P\left(t^+\right) = softmax\left(\left(T\right)^T\left(h_\perp + d_r\right)\right) \tag{5}$$

where T in the bottom left in T^T means the look-up table and T in the upper right in T^T means transpose operation.

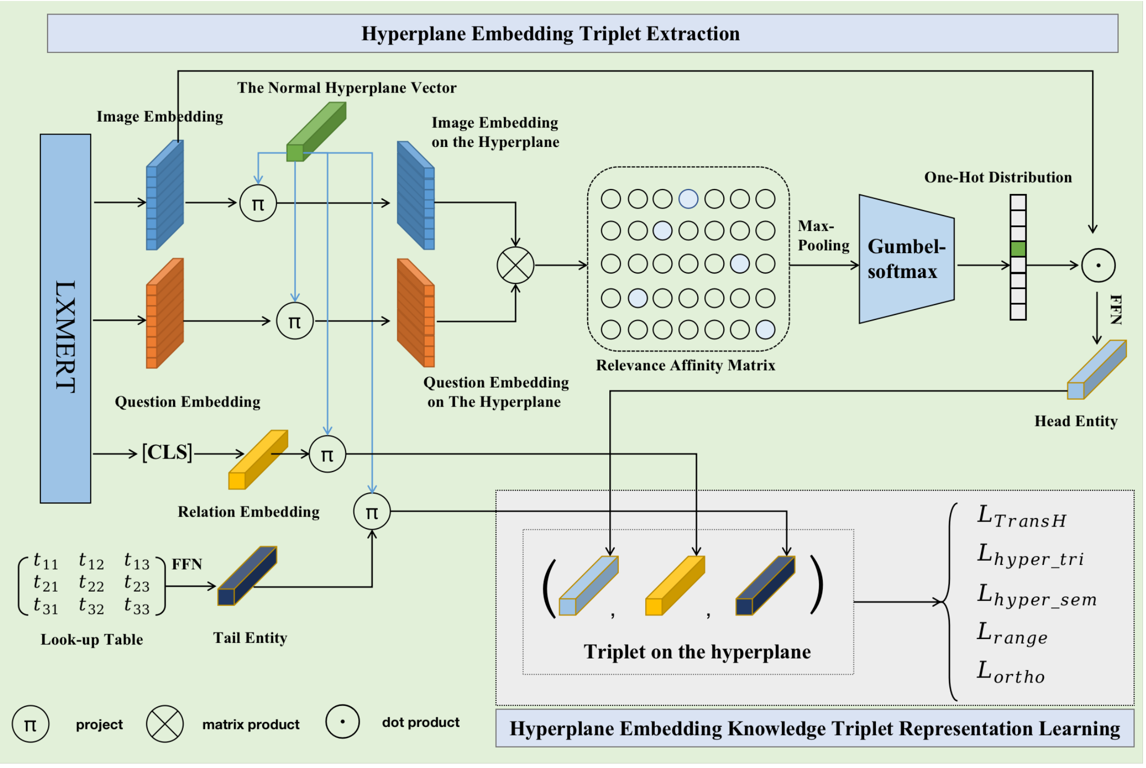


Fig. 2 An overview of our model MKEAH. The model contains two modules: Hyperplane Embedding Triplet Extraction and Hyperplane Embedding Knowledge Triplet Representation Learning. The normal hyperplane vector plays an essential role in the transformation of knowledge representation embedding

3.2.4 Range Loss

In order to ensure that the length of the feature vector projected onto the hyperplane is kept within the unit vector range, we convert the length constraint inequality $\|concat\,(h_\perp, r_\perp)\|_2 \leq 1$ into a part of the loss function:

$$\mathcal{L}_{range} = \sum_{h,r\in A_+} \left[\|concat\,(h_\perp, r_\perp)\|_2^2 - 1 \right]_+ \tag{6}$$

3.2.5 Orthogonal Loss

In order to filter out enough topic-irrelevant information and guarantee the translation vector d_r is in the hyperplane, we convert the orthogonality [11] constraint inequality$\left|w_r^\top d_r\right| / \|d_r\|_2 \leq \varepsilon$ into a part of the loss function:

$$\mathcal{L}_{Ortho} = \sum_{r\in R} \left[\frac{\left(w_r^\top d_r\right)^2}{\|d_r\|_2^2} - \varepsilon^2 \right]_+ \tag{7}$$

In total, the final loss is defined as:

$$\begin{aligned} \mathcal{L} =& \mathcal{L}_{TransH} + \mathcal{L}_{hyper\_tri} + \mathcal{L}_{hyper\_sem} \\ &+ C\left(\mathcal{L}_{range} + \mathcal{L}_{Ortho}\right) \end{aligned} \tag{8}$$

where C is a hyperparameter representing the weight of range loss and orthogonal loss.

3.3 Knowledge accumulation and prediction

We first extract and accumulate knowledge of our model using VQA 2.0, which filters the samples that do not provide factual knowledge answers. Then we use datasets from the KB-VQA tasks, such as OK-VQA and KRVQA, to fine-adjust for model extraction and accumulation of multimodal knowledge for more complex scenarios.

In the inference phase, we want to make sure that the predicted answer is as close to the tail entity $t_{i\perp}$ as possible in the hyperplane. Given an image and a problem, we input them into the network MKEAH and obtain the embedding of header entities and relationships in the hyperplane. We calculate the distance between $h_{inf} + r_{inf}$ and each tail entity T in the hyperplane in the lookup table $t_{i\perp}$, and select the tail entity with the smallest distance as the predictive answer.

In order to measure the ability of the model to extract knowledge related to the topic, we put forword the Topic Similarity as:

$$q_{\perp} = q - w_r^{\top} q w_r \tag{9}$$

$$TS = mean \left(\sum_{h,r \in A_{+}, q \in Q} \cos\left(h_{\perp} + r_{\perp}, q_{\perp}\right) \right) \tag{10}$$

4 Experiments

Datasets and evaluation metrics We conduct extensive experiments on two datasets: OK-VQA and KR-VQA. OK-VQA aims to provide diverse, difficult, and large-scale problems that promote VQA models in reasoning and accumulating knowledge beyond image content. KRVQA was proposed to address annotator bias and avoid superficial overfitting correlations between questions and answers. KRVQA aims to cut through the shortcut learning utilized by current deep embedded models and push the boundaries of knowledge-based reasoning for visual problems. We use top-1 accuracy for a fair comparison.

4.1 Experimental Configuration

For all experiments, all models are trained on 4 NVIDIA A100. We construct knowledge triplets using annotated answers filtering out low knowledge density samples from datasets that require external knowledge. We treat all samples in a batch with different answers from the positive samples as negative samples for the triplet ranking loss. Our model is trained by AdamW optimizer with 300 epochs, where the batch size is 128, and the learning rate is set to 1×10^{-5} and 1×10^{-4} in the pre-training and fine-tuning stage, respectively. The margin is set to 1.0. The hyperparameter C is set to 2.0.

4.2 Most advanced comparison

Comparison on OK-VQA. In Table 1, we present a comparison of our results with state-of-the-art models. These include approaches based on knowledge graphs, unstructured knowledge, hybrid multi-source knowledge, pre-training [28] with implicit knowledge, and multimodal knowledge. Additionally, we also compare our results with those of traditional VQA methods.

Our MKEAH model consistently beats all current techniques and outperforms the state-of-the-art model by 2.12%. In contrast to most models that represent knowledge in the original space, our model represents knowledge embedded in the hyperplane, reducing the interference of topic-irrelevant information on answers. In addition, our model outperforms the pre-training model by 12.67% because our model captures the implicit association between text and image and the multimodal knowledge of higher order logic rather than the knowledge of visual and verbal first-order logic co-occurrences in the pre-training framework. KM4 [35] also uses multimodal knowledge to associate images with entities in existing knowledge graphs but still lacks knowledge of higher-order complex relationships and is 13.39% lower than MKEAH.

Comparison on KR-VQA. In Table 2, we compare MKEAH with traditional VQA models. A "KB-not-related" question represents only basic visual knowledge, while a "KB-related" question represents fact knowledge in a knowledge base. Our approach outperforms previous models, achieving a stunning 3.24 percent improvement on the overall metric, compared with the best model. MKEAH obtains 1.56% improvement on average over the 'KB-related' questions, indicating that linking low-level visual content with high-level semantics also has important practical implications for visual-only questions, MKEAH performs worse than certain models on two-step reasoning type 3 questions because the responses are largely relations, but MKEAH commonly employs factual entities as tail entities for accumulation and prediction.

4.3 Model analysis and ablation

In Table 3, we evaluate the contribution of each loss function, hyperplane embedding, triplet structures and knowledge accumulation strategy. The OK-VQA dataset is used in the experiment.

The impact of each loss function. In models '2-6', we evaluate the effect of each loss function on the performance. The accuracy of removing $\mathcal{L}_{hyper\_tri}$ and $\mathcal{L}_{hyper\_sem}$ respectively decreases by 3.40% and 4.59% while removing $\mathcal{L}_{TransH}$ results in a significant decrease in model '2'. Because $\mathcal{L}_{TransH}$ preserves the embedded structure of the basic triplets in the hyperplane, which has a greater impact than other loss functions. The performance of models '5' and '6' is inferior to MuKEA, which indicates the importance of $\mathcal{L}_{hyper\_tri}$ and $\mathcal{L}_{hyper\_sem}$ for knowledge representation to adapt to hyperplane embedding. This is because the lengths of the feature vectors projected onto the hyperplane are inconsistent and topic-irrelevant information is not sufficiently filtered, which weaken the knowledge representation ability of hyperplane feature vectors.

The impact of hyperplane embedding. In models '7-9', we use the embedding of head entities, relations, and tail entities in the hyperplane. Their performance

Method	Knowledge Resources	Accurary
ArticleNet (AN)[34]	Wikepedia	5.28
Q-only[34]	-	14.93
BAN[10]	-	25.17
BAN+AN[34]	Wikepedia	25.61
BAN+KG-AUG[13]	Wikepedia+ConceptNet	26.71
MUTAN[22]	-	26.41
MUTAN+AN[34]	Wikepedia	27.84
Mucko[23]	ConceptNet	29.20
GRUC[31]	ConceptNet	29.87
KM^4[35]	multimodal knowledge from OK-VQA	31.32
ViLBERT[12]	-	31.35
LXMERT[14]	-	32.04
KRISP (w/o mm pre.)[33]	DBpedia+ConceptNet+VisualGenome + haspartKB	32.31
KRISP (w/ mm pre.)[33]	DBpedia+ConceptNet+VisualGenome + haspartKB	38.90
ConceptBert[32]	ConceptNet	33.60
Knowledge is Power[47]	YAGO3	39.24
MuKEA[17]	multimodal knowledge from VQA 2.0 and OK-VQA	42.59
MKEAH	multimodal knowledge from VQA 2.0 and OK-VQA	44.71

Table 1 State-of-the-art comparison on OK-VQA dataset.

Method	KB-not-related							KB-related					Overall
	one-step			two-step				one-step	two-step				
	0	1	2	3	4	5	6	2	3	4	5	6	
Q-type[1]	36.19	2.78	8.21	33.18	35.97	3.66	8.06	0.09	0.00	0.18	0.06	0.33	8.12
LSTM[1]	45.98	2.79	2.75	43.26	40.67	2.62	1.72	0.43	0.00	0.52	1.65	0.74	8.81
FiLM[18]	52.42	21.35	18.50	45.23	42.36	21.32	15.44	6.27	5.48	4.37	4.41	7.19	16.89
MFH[19]	43.74	28.28	27.49	38.71	36.48	20.77	21.01	12.97	5.10	6.05	5.02	14.38	19.55
UpDn[29]	56.42	29.89	28.63	49.69	43.87	24.71	21.28	11.07	8.16	7.09	5.37	13.97	21.85
MCAN[27]	49.60	27.67	25.76	39.69	37.92	21.22	18.63	12.28	9.35	9.22	5.23	13.34	20.52
+knowldge retrieval[1]	51.32	27.14	25.69	41.23	38.86	23.25	21.15	13.59	9.84	9.24	5.51	13.89	21.30
MuKEA[17]	59.12	44.88	37.36	52.47	48.08	35.63	31.61	17.62	6.14	9.85	6.22	18.28	27.38
MKEAH	60.34	46.23	40.12	54.21	50.26	37.23	33.72	19.18	8.23	9.91	7.13	20.12	30.62

Table 2 State-of-the-art comparison on KRVQA dataset.The numbers in the third row mean different types of questions.

Method	Accurary
1.MKEAH	44.71
2.w/o $\mathcal{L}_{TransH}$	27.57
3.w/o $\mathcal{L}_{hyper\_tri}$	41.31
4.w/o $\mathcal{L}_{hyper\_sem}$	40.12
5.w/o $\mathcal{L}_{range}$	42.23
6.w/o $\mathcal{L}_{Ortho}$	40.16
7.head entity w/o embedding on hyperplane	41.28
8.relation entity w/o embedding on hyperplane	41.34
9.tail entity w/o embedding on hyperplane	42.16
10.w/o h	40.92
11.w/o r	40.14
12.w/o LXMERT	34.29

Table 3 Ablation of key components in MuKEA on OK-VQA

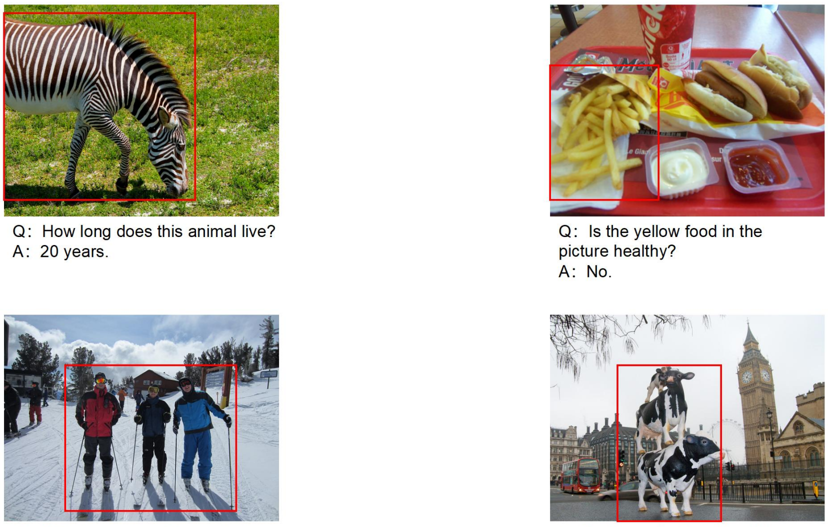
decrease 3.49%, 3.37%, and 2.55%, respectively. This proves the effectiveness of hyperplane embedding to further improve the modeling ability of the external world, and if there is no hyperplane embedding of either head entity or relational entity, the advantages of hyperplane embedding will be nullified.

The effects of triplet structures. In models '10-11', We remove the head entity and the tail entity on hyperplane respectively. The performance drops 3.79% and 4.57% accordingly, proving the effectiveness of triplet structure in hyperplane embedding for knowledge-based visual question-answering tasks.

The influence of prior knowledge accumulated in the pre-trained LXMERT. It can be seen that the accuracy drops 10.42% without pre-training in models '12'. This is caused by the fact that both the head entity and the relation representation depend on contextual information from pre-trained knowledge. Such a dependence still exists in hyperplane embedding.

4.4 Validation on Topic Similarity

In Table 4, we compare our model and MuKEA on topic similarity on OK-VQA. Our model outperforms MuKEA by 28.55, which indicates that the ability of our model to extract knowledge related to the topic has been significantly improved.



Q: How long does this animal live?
A: 20 years.

Q: Is the yellow food in the picture healthy?
A: No.

Q: Which country is best suited for the activities in the picture?
A: Austria

Q: Are the animals in the picture real?
A: No

Fig. 3 Visualization of the predicted answers of MuKEA. For MKEAH, the red box in the image shows the head entity.

Method	TS(Topic Similarity)
MuKEA	154.81
MKEAH	183.36

Table 4 Comparison on Topic Similarity on OK-VQA

4.5 Qualitative Analysis

From the case study in Figure 3, we conclude that our model is interpretable by visualizing the visual information on which the prediction is based: (1) The first line of examples requires the model to accumulate some external knowledge to answer, such as “the average age of zebras” and “whether French fries are a healthy food”, and the model needs to construct the correct triplet representation in order to answer the correct relevant questions. (2) We believe that the example in the second line requires the model to have a higher level of logical reasoning ability. For example, the model needs to judge that the activity in the text is skiing through the characters and environment, which itself requires some external knowledge. Then it also needs to use external knowledge to answer which country is more suitable for skiing. For the model to judge whether the animal model in the city is a real animal, it requires the ability of the model to distinguish the real animal from the fake animal, which also requires a more complex logical reasoning process.

4.6 Limitation Analysis

MKEAH’s entity-relationship representation is still conducted in public space, which results in poor performance in situations where entities and relationships have completely different semantic characteristics in the real world, such as entities being Los Angeles and the United States, respectively, while relationships represent geographic dependencies. In order to make a further breakthrough in the knowledge representation ability of the knowledge-based visual question answering network, more powerful knowledge representation methods can be studied in the following work.

5 Conclusion

This paper proposes a novel approach to represent knowledge in KB-VQA. The proposed model, MKEAH, extracts higher-order logical relationship between the knowledge representation and questions and filters out topic-irrelevant information, by applying the hyperplane. We propose a hyperplane transformation embedding method to represent the triplet in the multimodal knowledge graph , where the head entity embedding comes from the visual object, the tail entity embedding corresponds to the ground truth, and relational embedding represents an implicit association between the head and tail entities. Then We propose two loss functions: range loss and orthogonal loss, in order to help hyperplane embedding representation learn topic-related information and filter topic-irrelevant information. Additionally, we introduce a novel metric for evaluating the model's capability to extract topic-related information. Our study further confirms the importance of knowledge representation embedding methods to enhance the ability of models to capture complex information in the real world.

6 Acknowledgements

The work is partially supported by the National Nature Science Foundation of China (No. 61976160,61906137, 61976158, 62076184, 62076182) and Shanghai Science and Technology Plan Project (No.21DZ1204800) and Technology research plan project of Ministry of Public and Security (Grant No.2020JSYJD01).

References

1. Cao, Qingxing and Li, Bailin, et al, Knowledge-routed visual question reasoning: Challenges for deep representation embedding, IEEE Transactions on Neural Networks and Learning Systems, 33, 2758–2767(2021)
2. Wang Z, Zhang J, Feng J, et al, Knowledge graph embedding by translating on hyperplanes, Proceedings of the AAAI conference on artificial intelligence, 28(2014)
3. Zheng, Wenfeng and Yin, Lirong, et al, Knowledge base graph embedding module design for Visual question answering model, Pattern recognition, 120, 108153, Elsevier(2021)
4. Boukhers, Zeyd and Hartmann, Timo, et al, COIN: Counterfactual Image Generation for Visual Question Answering Interpretation,MDPI,22,2245(2022)
5. Walmer, Matthew and Sikka, Karan, et al, Dual-key multimodal backdoors for visual question answering, Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition, 15375–15385(2022)
6. Cao, Qingxing and Li, Bailin, et al, Knowledge-routed visual question reasoning: Challenges for deep representation embedding, IEEE Transactions on Neural Networks and Learning Systems, 33, 2758–2767(2024)
7. Gao, Peng and Jiang, Zhengkai, et al, Dynamic fusion with intra-and inter-modality attention flow for visual question answering, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,6639–6648(2019)
8. Goyal, Yash and Khot, Teja, Govind, et al, Making the v in vqa matter: Elevating the role of image understanding in visual question answering, Proceedings of the IEEE conference on computer vision and pattern recognition,6904–6913(2017)
9. Liang, Weixin and Jiang, Yanhao, et al, GraphVQA: Language-Guided Graph Neural Networks for Scene Graph Question Answering,NAACL-HLT,79(2021)
10. Kim, Jin-Hwa and Jun, Jaehyun and Zhang, Byoung-Tak, Bilinear attention networks, Advances in neural information processing systems, 31(2018)
11. Ranasinghe K, Naseer M, Hayat M, et al, Orthogonal projection loss, Proceedings of the IEEE/CVF International Conference on Computer Vision,12333-12343(2021)
12. Lu, Jiasen and Batra, Dhruv, et al, Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, Advances in neural information processing systems,32(2019)
13. Li, Guohao and Wang, Xin and Zhu, Wenwu, Boosting visual question answering with context-aware knowledge aggregation, Proceedings of the 28th ACM International Conference on Multimedia, 1227–1235(2020)
14. Tan, Hao and Bansal, Mohit, Lxmert: Learning cross-modality encoder representations from transformers, arXiv preprint arXiv:1908.07490(2019)
15. Kannan, Amar Viswanathan and Fradkin, Dmitriy, et al, Multimodal knowledge graph for deep learning papers and code, Proceedings of the 29th ACM International Conference on Information & Knowledge Management,417–3420(2020)
16. Nian, Fudong and Bao, Bing-Kun, et al, Multi-modal knowledge representation learning via webly-supervised relationships mining,Proceedings of the 25th ACM international conference on Multimedia,411–419(2017)
17. Ding, Yang and Yu, Jing, et al, Mukea: Multimodal knowledge extraction and accumulation for knowledge-based visual question answering, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 5089–5098(2022)
18. Perez, Ethan and Strub, Florian, et al, Film: Visual reasoning with a general conditioning layer, Proceedings of the AAAI Conference on Artificial Intelligence, 32(2018)
19. Yu, Zhou and Yu, Jun, et al, Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering, IEEE transactions on neural networks and learning systems, 29, 5947–5959(2018)
20. Kang D, Kwon H, Min J, et al, Relational embedding for few-shot classification, Proceedings of the IEEE/CVF International Conference on Computer Vision, 8822-8833(2021)
21. Mousselly-Sergieh, Hatem and Botschen, Teresa, et al, A multimodal translation-based approach for knowledge graph representation learning, Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics, 225–234(2018)
22. Ben-Younes, Hedi and Cadene, Rémi, Mutan: Multimodal tucker fusion for visual question answeringI, Proceedings of the IEEE international conference on computer vision, 2612–2620(2017)
23. Zhu, Zihao and Yu, Jing, et al, Mucko: multi-layer cross-modal knowledge reasoning for fact-based visual question answering, In Proceedings of the International Joint Conference on Artificial Intelligence, 1097–1103(2020)
24. Teney, Damien and Anderson, Peter, et al, Tips and tricks for visual question answering: Learnings from the 2017 challenge, Proceedings of the IEEE conference on computer vision and pattern recognition, 4223–4232(2018)
25. Antol, Stanislaw and Agrawal, Aishwarya, et al, Vqa: Visual question answering, Proceedings of the IEEE international conference on computer vision,2425–2433(2015)
26. Manmadhan, Sruthy and Kovoor, Binsu C, et al, Visual question answering: a state-of-the-art review,Artificial Intelligence Review,53,5705–5745(2020)
27. Yu, Zhou and Yu, Jun, et al, Deep modular co-attention networks for visual question answering, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 6281–6290(2019)
28. Yang, Zhengyuan and Lu, Yijuan, et al, Tap: Text-aware pre-training for text-vqa and text-caption, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,8751–8761(2021)

29. Anderson, Peter and He, Xiaodong, et al, Bottom-up and top-down attention for image captioning and visual question answering, Proceedings of the IEEE conference on computer vision and pattern recognition, 6077–6086(2018)
30. Lin, Yankai and Liu, Zhiyuan, et al, Learning entity and relation embeddings for knowledge graph completion, Proceedings of the AAAI conference on artificial intelligence, 29(2015)
31. Yu, Jing and Zhu, Zihao, et al, Cross-modal knowledge reasoning for knowledge-based visual question answering, 108, 107563, Elsevier(2020)
32. Gardères, François and Ziaeefard, Maryam, et al, Conceptbert: Concept-aware representation for visual question answering, Findings of the Association for Computational Linguistics: EMNLP 2020, 489–498(2020)
33. Marino, Kenneth and Chen, Xinlei, et al, Krisp: Integrating implicit and symbolic knowledge for open-domain knowledge-based vqa, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 14111–14121(2021)
34. Marino, Kenneth and Rastegari, et al, Ok-vqa: A visual question answering benchmark requiring external knowledge, Proceedings of the IEEE/cvf conference on computer vision and pattern recognition, 3195–3204(2019)
35. Zheng, Wenbo and Yan, Lan, et al, Km4: Visual reasoning via knowledge embedding memory model with mutual modulation, Information Fusion, 67, 14–28(2021)
36. Rebele, Thomas and Suchanek, Fabian, et al, YAGO: A multilingual knowledge base from wikipedia, wordnet, and geonames, The Semantic Web–ISWC 2016: 15th International Semantic Web Conference, Kobe, Japan, October 17–21, 2016, Proceedings, Part II 15, 177–185, Springer (2016)
37. Narasimhan, Medhini and Schwing, Alexander G, Straight to the facts: Learning knowledge base retrieval for factual visual question answering, Proceedings of the European conference on computer vision (ECCV), 451–468(2018)
38. Wang, Peng and Wu, Qi, et al, Fvqa: Fact-based visual question answering, Proceedings of the European conference on computer vision (ECCV), 451-468(2018)
39. Wu, Jialin and Lu, Jiasen, et al, Multi-modal answer validation for knowledge-based vqa, Proceedings of the AAAI Conference on Artificial Intelligence,36, 2712–2721(2022)
40. Pezeshkpour, Pouya and Chen, et al, Embedding multimodal relational data for knowledge base completion, In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 3208–3218(2018)
41. Li, Manling and Zareian, Alireza, et al, Gaia: A fine-grained multimedia knowledge extraction system, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 77–86(2020)
42. Manola, Frank and Miller, Eric, et al, RDF primer, W3C recommendation, 10, 6, Citeseer(2004)
43. Bordes, Antoine and Usunier, Nicolas, et al, Translating embeddings for modeling multi-relational data, Advances in neural information processing systems, 26(2013)
44. Bordes, Antoine and Weston, Jason, et al, Learning structured embeddings of knowledge bases, Proceedings of the AAAI conference on artificial intelligence, 25, 301–306(2011)
45. Jenatton, Rodolphe and Roux, Nicolas, et al, A latent factor model for highly multi-relational data, Advances in neural information processing systems, 25(2012)
46. Socher, Richard and Chen, Danqi, et al, Reasoning with neural tensor networks for knowledge base completion, Advances in neural information processing systems, 26(2013)
47. Zheng, Wenbo and Yan, Lan, et al, Knowledge is power: Hierarchical-knowledge embedded meta-learning for visual reasoning in artistic domains, Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, 2360–2368(2021)
48. Cai, Hongyun and Zheng, Vincent W, et al, A comprehensive survey of graph embedding: Problems, techniques, and applications, IEEE transactions on knowledge and data engineering, 30, 1616–1637(2018)