Zero-Shot Visual Question Answering Using Knowledge Graph

Zhuo Chen[1,2], Jiaoyan Chen[3], Yuxia Geng[1,2], Jeff Z. Pan[4], Zonggang Yuan[5], and Huajun Chen[1,2(✉)]

[1] College of Computer Science and Hangzhou Innovation Center, Zhejiang University, Hangzhou, China
{zhuo.chen,gengyx,huajunsir}@zju.edu.cn
[2] AZFT Joint Lab for Knowledge Engine, Hangzhou, China
[3] Department of Computer Science, University of Oxford, Oxford, UK
jiaoyan.chen@cs.ox.ac.uk
[4] School of Informatics, The University of Edinburgh, Edinburgh, UK
[5] NAIE CTO Office, Huawei Technologies Co., Ltd., Shenzhen, China
yuanzonggang@huawei.com
https://knowledge-representation.org/j.z.pan/

Abstract. Incorporating external knowledge to Visual Question Answering (VQA) has become a vital practical need. Existing methods mostly adopt pipeline approaches with different components for knowledge matching and extraction, feature learning, etc. However, such pipeline approaches suffer when some component does not perform well, which leads to error cascading and poor overall performance. Furthermore, the majority of existing approaches ignore the answer bias issue—many answers may have never appeared during training (i.e., unseen answers) in real-word application. To bridge these gaps, in this paper, we propose a Zero-shot VQA algorithm using knowledge graph and a mask-based learning mechanism for better incorporating external knowledge, and present new answer-based Zero-shot VQA splits for the F-VQA dataset. Experiments show that our method can achieve state-of-the-art performance in Zero-shot VQA with unseen answers, meanwhile dramatically augment existing end-to-end models on the normal F-VQA task.

Keywords: Visual Question Answering · Zero-shot learning · Knowledge graph

1 Introduction

Visual Question Answering (VQA) is to answer natural language questions according to given images. It plays an important role in many applications such as advertising and personal assistants to the visually impaired. It has been widely investigated with promising results achieved due to the development of image and natural language processing techniques. However, most of the current solutions still cannot address the open-world scene understanding where the answer

A. Hotho et al. (Eds.): ISWC 2021, LNCS 12922, pp. 146–162, 2021.
https://doi.org/10.1007/978-3-030-88361-4_9

is not directly contained in the image but comes from or relies on external knowledge. Considering the question "Q1: Normally you play this game with your?" in Fig. 1, some additional knowledge is indispensable since that the answer "*dog*" cannot be found out with the content in the image alone.

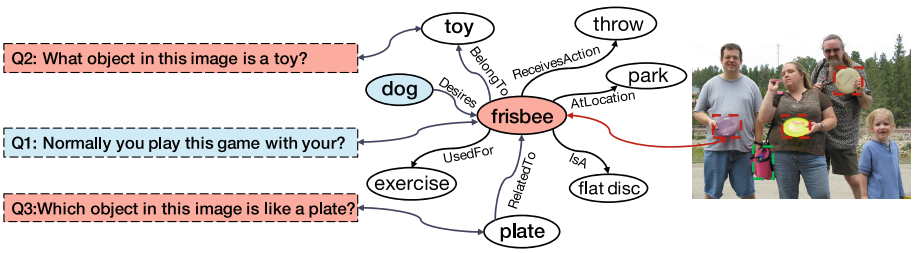


Fig. 1. VQA Examples. Q1: the answer is outside the image and question; Q2 and Q3: the answers are within the images or questions but require additional knowledge.

Some VQA methods have been developed to utilize external knowledge for open-world scene understanding. For example, Marino et al. [16] extensively utilize unstructured text information from the Web as external information but fail to address the noise (irrelevant information) in the text. Wang et al. [27] first extract visual concepts from images and then link them to an external knowledge graph (KG). The corresponding questions can then be transformed into a series of queries to the KG (e.g., SPARQL queries) to retrieve answers. Zhu et al. [31] instead construct a multi-modal heterogeneous graph by incorporating the spatial relationships and descriptive semantic relationships between visual concepts, as well as supporting facts retrieved from KGs, and then apply a modality-aware graph convolutional network to infer the answer. However, the performance of all these methods would be dramatically impacted if one module of the pipeline does not perform that well (a.k.a. error cascading [7]). Although some end-to-end models such as [2,13] have been proposed to avoid error cascading, they are still quite preliminary, especially on utilizing external knowledge, with worse performance than the pipeline methods on many VQA tasks.

Another important issue raised in VQA is the dependence on labeled training data, i.e., the model is trained by a dataset of (question, image, answer) tuples, and generalizes to answer questions about objects and situations that have already been presented in the training set. However, for new types of questions or answers, and objects newly emerge in images, there is a need for collecting labeled tuples and training the model from the scratch. Targeting such a limitation, Zero-shot VQA (ZS-VQA), which aims to predict with objects, questions or answers that have never appeared in training samples, has been proposed. Teney et al. [25] address questions that include new words; while [9,22] address images that contain new objects. However, all of these VQA methods still focus on the closed-world scene understanding without considering unseen answers and rarely make full use of KG. In this paper, we utilize KG to study VQA with open-world scene understanding,

which requires external knowledge to answer the question, and ZS-VQA, especially the sub-task that addresses new answers.

In this paper, we present a ZS-VQA algorithm using KG and a mask-based learning mechanism, and at the same time propose a new Zero-shot Fact VQA (ZS-F-VQA) dataset which is to evaluate ZS-VQA for unseen answers. Firstly, we learn three different feature mapping spaces separately, which are semantic space about relations, object space about support entities, and knowledge space about answers. Each of them is used to align the joint embedding of image-question pair (I-Q pair) with corresponding target. Via the combination between all those chosen supporting entities and relations, masks are decided according to a mapping table which contains all triplets in a fact KG, which guides the alignment process of unseen answer prediction. Specially, the marks can be used as hard masks or soft masks, depending on the VQA tasks. Hard marks are used in ZS-VQA tasks; e.g., with the ZS-F-VQA dataset, our method achieves state-of-the-art performance and far superior (30–40%) to other methods. On the other hand, soft marks are used in standard VQA tasks; e.g., with the F-VQA dataset, our method achieves a stable improvement (6–9%) on baseline end-to-end method and well alleviates the error cascading problem of pipeline models. To sum up, the main contributions are summarized below:

- We propose a robust ZS-VQA algorithm using KGs,[1] which adjusts answer prediction score via masking based on the alignments between supporting entities/relations and fusion I-Q pairs in two feature spaces.
- We define a new ZS-VQA problem which requires external knowledge and considers unseen answers. Accordingly, we develop a ZS-F-VQA dataset for evaluation.
- Our KG-based ZS-VQA algorithm is quite flexible. It can successfully address both normal VQA tasks that rely on external knowledge and ZS-VQA tasks, and can be directly used to augment existing end-to-end models.

2 Related Work

2.1 Visual Question Answering

Traditional VQA Methods. Since proposed in 2015 by [3], a few VQA methods, which apply multi-modal feature fusion between question and image for final answer decision, have been proposed. Various attention mechanisms [2,29] are adopted to refine specific regions of the image for corresponding question meanwhile to make the prediction process interpretable. Graph-based approaches such as [6] combine multi-modal information and enhance the interaction among significant entities in texts and images.

Knowledge-Based VQA. Utilizing symbolic knowledge is a straight forward solution to augment VQA. To study incorporating external knowledge with VQA, datasets such as F-VQA [27], OK-VQA [16] and KVQA [23] have been

[1] Our code and data are available at https://github.com/China-UK-ZSL/ZS-F-VQA.

developed. Each question in F-VQA refers to a specific fact triple in relevant KG like ConceptNet. While OK-VQA is manually marked without a guided KG as reference which leads to its difficulty. KVQA targets at world knowledge where questions are about the relationship between characteristics.

To incorporate such external knowledge, [26,27] generate SPARQL queries for querying the constructed sub-KG according to I-Q pairs. [17,18,28,31] extract entities from image and question to get related concepts from KG for answer prediction. Marino et al. [16] take unstructured knowledge on the Web to enhance the semantic representation of I-Q joint feature. All of the above methods utilize pipeline approaches to narrow the answer scope, but they are often ad-hoc, which limits their deployment and generalization to new datasets. Most importantly, the errors will be magnified during running since each module usually has no ability to correct previous modules' errors. End-to-end model like [2,13] are more general and can avoid error cascading, but they are still preliminary, especially in addressing VQA problems which require external knowledge.

Different from these approaches, our proposed framework leverages the advantages of both end-to-end and pipeline approaches. We improve the model transferability meanwhile effectively avoid the error cascading (see our case study as illustrated in Fig. 5), making it quite general to different tasks and very robust with promising performance achieved.

2.2 Zero-Shot VQA

Machine learning often follows a closed world assumption where classes to predict all have training samples. However, the real-world is not completely closed and it is impractical to always annotate sufficient samples to re-train the model for new classes. Targeting such a limitation, zero-shot learning (ZSL) is proposed to handle these novel classes without seeing their training samples (i.e., unseen classes) [5,10]. Teney et al. [25] first propose Zero-shot VQA (ZS-VQA) and introduce novel concepts on language semantic side, where a test sample is regarded as unseen if there is at least one novel word in its question or answer. Ramakrishnan et al. [22] incorporate prior knowledge into model through pre-training with unstructured external data (from both visual and semantic level). Farazi et al. [9] reformulate ZS-VQA as a transfer learning task that applies closely seen instances (I-Q pairs) for reasoning about unseen concepts. A major limitation of these approaches is that they seldom consider the imbalance and low resources problem regarding the answer itself. Open-answer VQA requires models to select answer with the highest activation from fixed possible K answer categories, but the model cannot tackle unseen answers because answers are isolated with no specific meaning. Besides, VQA is defined as a classification problem without utilizing enough semantic information of the answer. Agrawal et al. [1] propose a new setting for VQA where the test question-answer pairs are compositionally novel compared to training question-answer pairs. Some methods [12,24] try to align answer with I-Q joint embedding through feature representation for realizing unseen answer prediction or simply for concatenating their representation as the input of a fully connected layer for score prediction [25]. However, all of

them are powerless to answer those I-Q pairs that require external knowledge, and the relevance among answers are still not strong enough with insufficient external information. The ZS-VQA method proposed in this paper incorporates richer and more relevant knowledge by using KGs, through which the existing common sense is well utilized and more accurate answers are often given.

3 Preliminaries

Visual Question Answering (VQA) and Zero-Shot VQA. A VQA task is to provide a correct answer a given an image i paired with a question q. Following the open-answer VQA setting defined in [12], let a be a member of the answer pool $\mathcal{A} = \{a_1, ..., a_n\}$, the candidates of which are the top K (e.g. 500) most frequent answers of the whole dataset. A dataset is represented by a set of distinctive triplets $\mathcal{D} = \{(i, q, a)|i \in \mathcal{I}, q \in \mathcal{Q}, a \in \mathcal{A}\}$ where $\mathcal{I}$ and $\mathcal{Q}$ are respectively image and question sets. A testing dataset is denoted as $\mathcal{D}_{te}$ with each triplet (i, q, a) not belonging to training dataset $\mathcal{D}_{tr}$. We denote $\mathcal{D}_{tr}^{zsl} = \{(i, q, a)|i \in \mathcal{I}, q \in \mathcal{Q}, a \in \mathcal{A}_s\}$ and $\mathcal{D}_{te}^{zsl} = \{(i, q, a)|i \in \mathcal{I}, q \in \mathcal{Q}, a \in \mathcal{A}_u\}$, where $\mathcal{A}_s$ and $\mathcal{A}_u$ respectively denote the seen answer set and the unseen answer set with $\mathcal{A}_u \cap \mathcal{A}_s = \emptyset$. ZS-VQA is much harder than normal VQA, since information in the image and question is insufficient for answers that have never appeared in the training samples. Specifically, we study two settings at testing stage of ZS-VQA: one is the standard ZSL, where the candidates answers of a testing sample (i, q, a) are $\mathcal{A}_u$, while the other is the generalized ZSL (GZSL) with $\mathcal{A}_u \cup \mathcal{A}_s$ as candidates answers during testing. It should be noted that regular VQA only predicts with seen answers, while VQA under the GZSL setting predicts with both seen and unseen answers.

Knowledge Graph (KG). KGs have been widely used in knowledge representation and knowledge management [19,20]. The KG we used is a subset of three KGs (DBpedia, ConceptNet, WebChild) selected by Wang et al. [27] (in the form of RDF[2] triple). It is used to establish the prior knowledge connection, which includes a set of answer nodes and concept (tool) nodes to enrich the relationships among answers. Besides, different relations (edges) are applied to represent the fact graph (triples).

Taking Fig. 1 as an example, all (i, q) pairs could be divided into two categories according to their answer sources: 1) Those answers which are outside the images and questions. Such as the answer "dog" to question "Q1: Normally you play this game with your?", the data source of the answer here is the external KG which contains the triple <frisbee, BelongTo, toy> for QA support. 2) Those answers that can be found in images or questions. In this situation, there are often more than one object in image/question for screening through some implicit common sense relations (e.g., "Q2: Which object in this image is like a plate?" targets at finding the correct object related to plate). Then, one fact triple (e.g. <plate, RelatedTo, frisbee>) could play the role of answer guidance.

[2] https://www.w3.org/TR/2014/REC-rdf11-mt-20140225/.

4 Methodology

4.1 Main Idea

Our main idea is motivated by two deficiencies in current knowledge-based VQA approaches. Firstly, in those methods it is common to build intermediate models and involve KG queries in a pipeline way, which leads to error cascading and poor generalization. Secondly, most of them define VQA as a classification problem which does not utilize enough knowledge of the answers, and fails to predict unseen answers or to transfer across datasets whose candidate answers have little or no overlap. For example, as shown in Fig. 1, if concept "frisbee" has not appeared in training set, traditional VQA will fail to recognize it in testing phase for answer out-of-vocabulary (OOV) problem. While other method [12] which takes answer semantics into account has lost the relation information: "Desires" came from entity "dog", or "RelatedTo" came from entity "plate".

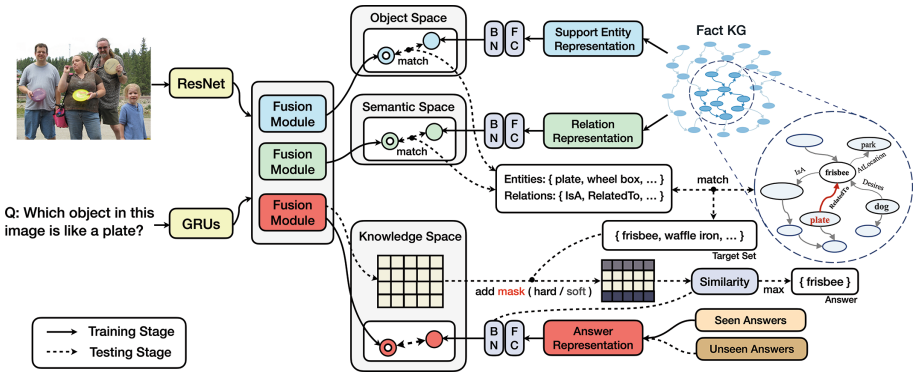


Fig. 2. An overview of our framework.

By utilizing semantics embedding feature as answer representation, we convert VQA from a classification task into a mapping task. After parameter learning, the distribution of the joint embedding between question and image can partly get close to answer's one with shadow knowledge included in. We call it the knowledge space about answers. Besides, we independently define two other feature spaces: semantic space about relations and object space about support entities. Semantic space aims to project (i, q) joint feature into a relation according to the semantic information in triplets, while object space targets at establishing relevant connection between (i, q) and a support entity (a.k.a. entity on KG). They play the role for answer guidance when combined together (see Sect. 4.2 for detail). In order to overcome those limitations proposed in Sect. 2.1, we provide a soft/hard mask method in this situation to effectively enhance alignment process meanwhile alleviating error cascading.

4.2 Establishment of Multiple Feature Spaces

Following [12], we establish connection between an answer and its corresponding (i, q) pair via projecting them into a common feature space and get close to each other. Firstly, a fusion feature extractor $F_{\boldsymbol{\theta}}(i, q)$ between q and i is leveraged to combine multimodal information. Meanwhile, we define $G_{\phi}(a)$ as the representation of answer a. We follow the probabilistic model of compatibility (PMC) drawn from [12] and add loss temperature τ for better optimization:

$$P(a \mid i_n, q_n) = \frac{\exp\left(F_{\boldsymbol{\theta}}(i_n, q_n)^{\top} G_{\phi}(a)/\tau\right)}{\sum_{a' \in \mathcal{A}} \exp\left(F_{\boldsymbol{\theta}}(i_n, q_n)^{\top} G_{\phi}(a')/\tau\right)} \tag{1}$$

where $\mathcal{A}$ denotes $\mathcal{A}_u$ when the setting is standard ZSL else remain the same, and a is the correct answer of (i_n, q_n). For learning the parameters to maximize the likelihood in overall PMC model, we employ following loss function:

$$\ell_a = -\sum_{n}^{N} \sum_{b \in \mathcal{A}} \alpha(a, b) \log P(b \mid i_n, q_n) \tag{2}$$

where weighting function $\alpha(a, b)$ measures how much the predicted answer b can contribute to the objective function. A nature design is $\alpha(a, b) = \mathbb{I}[a = b]$, where $\mathbb{I}[.]$ denotes binary indicator function, taking value of 1 if the condition is true else 0 for false. During testing, given the learned $F_{\boldsymbol{\theta}}(i, q)$ and $G_{\phi}(a)$, we can apply following decision rule to predict the answer $\hat{a}$ to (i, q) pair:

$$\hat{a} = \arg\max_{a \in \mathcal{A}} F_{\boldsymbol{\theta}}(i, q)^{\top} G_{\boldsymbol{\phi}}(a) \tag{3}$$

Like the results shown in Sect. 5.3, the above feature projection process could learn shallow knowledge in VQA which requires external knowledge. However, it performs not well since network is not sufficient to model abundant prior knowledge with small amount of training data (see data statistics in Table 1).

Matching the elements in images or questions to KG entities in an explicit [27] or implicit [9] way can augment the model with knowledge to well address the open-world scene understanding problem (see links in Fig. 1 toy example). In our method, the alignment between image/question and KG is implicitly done by multiple feature spaces rather than simply object detection. We leverage another two feature spaces for answer revising:

1) **Semantic space** focuses on the language information within (i, q), which works as a guidance toward the projection of triplet relations r in KG. In particular, the signal of q is more crucial than i in this part.
2) Compared with traditional image classification which identifies the correct category of a given image, the **object space** is more likely a feature space about support entity classifier which simultaneously observes images and texts for salient feature. Specifically, the alignment between support entity e embedding and (i, q) joint embedding avoids the direct learning of complex knowledge, meanwhile acts on the subsequent answer mask process together with the prediction relations r obtained in semantic space.

Similarly, we define their embedding function as $G_{\phi\star}(r)$, $G_{\phi\diamond}(e)$ and the corresponding (i, q) joint embedding function as $F_{\boldsymbol{\theta}\star}(i, q)$, $F_{\boldsymbol{\theta}\diamond}(i, q)$ for distinction. Other formulas and probability calculation methods remain the same as answer such as loss function ℓ_r and ℓ_e, which are model's overall optimization goal together with ℓ_a. The parameters in these three pairs of models are independent except for the frozen input embedding layers.

Pre-trained word vector contains the latent semantics in real-world natural language. In order to get the initialized representation of the answer, relation and support entity, we employ GloVe embedding [21] meanwhile compare other answer representation like KG embedding [4] or ConceptNet embedding [15] (see Sect. 5.4 for detail).

Besides, different surface forms (e.g., mice & mouse) should be considered for the same meaning. [12] takes advantage of the weighting function $\alpha(a, b)$ with WUPS score, which is reliant on semantic similarity scores between a and b. We find that it works well with singular and plural disambiguation (e.g. WUPS (*dog*, *dogs*) ≈ 0.929), but fails in many cases of tense disambiguation (e.g., WUPS (*cook*, *cooking*) ≈ 0.125, WUPS (*play*, *played*) ≈ 0.182). So we apply NLTK tools (e.g., WordNetLemmatizer) to achieve more accurate word split and Minimum Edit Distance (MED) for concept disambiguation.

4.3 Answer Mask via Knowledge

Masking is widely used in language model pre-training for improving machine's understanding of the text. Two examples are masking part of the words in the training corpus (e.g. BERT [8]) and masking common sense concepts (e.g. AMS [30]). But they rarely consider the direct constraint of knowledge in prediction results, ignoring that human beings know how to make reasonable decision under the guidance of existing prior knowledge. Different from all these methods, we propose an answer masking strategy for VQA.

With the learned $F_{\boldsymbol{\theta}\star}$ and $F_{\boldsymbol{\theta}\diamond}$, we get the disjoint fusion embedding in two independent feature spaces, which are respectively taken as the basis for subsequent entity and relation matching: For a given (i, q) pair, vector similarity Sim is calculated via $F_{\boldsymbol{\theta}\star}(i, q)^\top G_{\boldsymbol{\phi}\star}(r_n)$ for relation, and $F_{\boldsymbol{\theta}\diamond}(i, q)^\top G_{\boldsymbol{\phi}\diamond}(e_n)$ for support entity. Those e and r, which correspond to the top-k Sim value, separately constitute the candidate set $\mathcal{C}_{ent}$ and $\mathcal{C}_{rel}$ where k is distinguished with k_r and k_e. Then target set $\mathcal{C}_{tar}$ is collected as follows:

$$\mathcal{C}_{tar} = \{t \mid (\exists(t, r, e) \vee \exists(e, r, t)) \wedge r \in \mathcal{C}_{rel} \wedge e \in \mathcal{C}_{ent}\} \tag{4}$$

$\mathcal{C}_{tar}$ contributes to the masking strategy on all answers $a_n \in \mathcal{A}$ via:

$$sim((i, q), a_n) = \begin{cases} (F_{\boldsymbol{\theta}}(i, q)^\top G_{\boldsymbol{\phi}}(a_n))/\tau & \text{if } a_n \in \mathcal{C}_{tar} \\ (F_{\boldsymbol{\theta}}(i, q)^\top G_{\boldsymbol{\phi}}(a_n))/\tau + s & \text{otherwise} \end{cases} \tag{5}$$

where s represents the score for masking which is the mainly distinction between hard mask and soft mask (see Sect. 5.4 for detail). Soft score greatly reduces the error cascading caused by the pipeline method through the whole model, which

will be discussed in Sect. 5.5. Meanwhile, the significance of hard mask comes from its superior performance in ZSL setting as shown in Sect. 5.3. Finally, the predicted answer $\hat{a}$ to the (i, q) pair is identified as:

$$\hat{a} = \arg\max_{a \in \mathcal{A}} sim((i, q), a) \tag{6}$$

It should be noted that candidate targets cannot just be regarded as the candidate answers due to the existence of soft mask, which revises the answer probability rather than simply limits answer's range. Moreover, as mentioned in Sect. 5.4 and 5.4, k and s mentioned above are hyper parameters which can cause various influence toward the result.

5 Experiments

We validate our approach for both normal VQA and ZS-VQA with ZSL/GZSL settings. In addition to the overall results, we conduct ablation studies for analyzing the impact of: 1) different factors in answer embedding; 2) the mask score; and 3) different hyper parameters (e.g. k_e, k_r). Finally, we evaluate its advantage on data transferability and mitigating error cascading.

5.1 Datasets and Metrics

F-VQA. As a standard publicly available VQA benchmark which requires external knowledge, F-VQA [27] consists of 2,190 images, 5,286 questions and a KG of 193,449 facts. Each (i, q, a) in this dataset is supported by a corresponding common sense fact triple extracted from public structured databases (e.g., ConceptNet, DBPedia, and WebChild). The KG has 101K entities and 1833 relations in total, 833 entities are used as answer nodes. In order to achieve parallel comparison, we maintain the coincide experiment setting with [18,27] to use standard dataset setting which contains 5 splits (by images), and prescribe candidate answers to the top-500 (%94.30 to entire as our check) for experiments. The over all data statistics after disambiguation are shown in Table 1.

ZS-F-VQA. The ZS-F-VQA dataset is a new split of the F-VQA dataset for zero-shot problem. Firstly we obtain the original train/test split of F-VQA dataset and combine them together to filter out the triples whose answers appear in top-500 according to its occurrence frequency. Next, we randomly divide this set of answers into new training split (a.k.a. seen) $\mathcal{A}_s$ and testing split (a.k.a. unseen) $\mathcal{A}_u$ at the ratio of 1:1. With reference to F-VQA standard dataset, the division process is repeated 5 times. For each (i, q, a) triplet in original F-VQA dataset, it is divided into training set if $a \in \mathcal{A}_s$. Else it is divided into testing set. The data statistics are shown in Table 1, where #class represents the number of data after deduplicated and #instance represents the number of samples. We denote "Overlap" as the intersection size of element sets within training and testing triples. Note that the overlap of answer instance between training and testing set in F-VQA are 2565 compared to 0 in ZS-F-VQA.

Table 1. The detailed data statistics. Average number of (i, q, a) in each train/test split in F-VQA is 2757/2735 compared to 2732/2760 of ZS-F-VQA.

#class	Images	Question	Answer	Support entity
	Train/Test/Overlap	Train/Test/Overlap	Train/Test/Overlap	Train/Test/Overlap
F-VQA	1059/1064/**0**	2431/2409/573	387/401/288	1695/1668.8/312
ZS-F-VQA	1297/1312/486	2384/2380/264	250/250/**0**	1578/1477/86
#instance	Overlap	Overlap	Overlap	Overlap
F-VQA	**0**	1372	2565	312
ZS-F-VQA	990	814	**0**	218

Evaluation Metrics. We measure the performance by accuracy and choose $Hit@1$, $Hit@3$, $Hit@10$ here together with MRR/MR to judge the comprehensive performance of model. $Hit@X$ indicates that the correct answer ranks in the top-k predicted answer sorted by probability. Mean Reciprocal Rank (MRR) measure the average reciprocal values of correct predicted answers compared to Mean Rank (MR). All the results we report are averaged across all splits.

5.2 Implementation Details

Fusion Model. We employ several models to parameterize the fusion function $F_{\boldsymbol{\theta}}$. We follow [12] to employ the Multi-layer Perceptron (MLP) and Stacked Attention Network (SAN) [29] as the representation of grid based visual fusion model. Meanwhile, we choose Up-Down (Bottom-Up and Top-Down Attention) [2] and Bilinear Attention Networks (BAN) [13] to measure the impact of bottom-up issue on external knowledge VQA problem. Moreover, we directly compare with [27] in some baselines like Qqmaping [27] Hie [14] under identical setting. Among all these methods, SAN is chosen as the base feature extractor $F_{\boldsymbol{\theta}}$ of our framework for its better performance(see Fig. 2).

Visual Features. To get i_n, we extract visual features from the layer 4 output of ResNet-152 ($14 \times 14 \times 2048$ tensor) pre-trained on ImageNet. Meanwhile applying ResNet-101-based Faster R-CNN pre-trained on COCO dataset to get bottom-up image region features. The object number per image is fixed into 36 with 1024 output dimensional feature.

Text Features. Each word in question and answer is represented by its 300-dimension GloVe [21] vector. The sequence of embedded words in question (average length is 9.5) is then fed into Bi-GRU for each time step. We have also tried to embed answer with GRU but find that it mostly leads to overfitting since the training set is not huge enough and average answer length is merely 1.2. So we simply represent the answer by averaging its word embedding.

During training, we utilize Adam optimizer with the mini-batch size as 128. Dropout and batch normalization are adopted to stabilize the training. We use a gradual learning rate warm up ($2.5 \times (epoch+1) \times 5 \times 10^{-4}$) for the first 7 epochs, decay it at the rate of 0.7 for every 3 epochs for epochs 14 to 47, and remain the same in rest epochs. Meanwhile, the loss temperature τ is set to 0.01 and

early stopping is used where *patience* is equal to 30. The model is trained offline, and thus the training time usually does not impact the method's application. In prediction, we currently consider 500 candidate answers for each testing sample. This makes the computation for evaluation affordable.

5.3 Overall Results

Table 2. The overall results (% for $Hit@K$) on standard F-VQA datasets (TOP-500). † denotes that the model is modified under a mapping-based setting (i.e., remove the last classifier layer of the (i, q) fusion network), which contrasts with traditional classifier-based approach.

Methods	$Hit@1$	$Hit@3$	$Hit@10$	MRR	MR
Hie-Q+I [14]	33.70	50.00	64.08	–	–
MLP	34.12	52.26	69.11	–	–
Up-Down [2]	34.81	50.13	64.37	–	–
Up-Down†	40.91	57.47	72.74	–	–
SAN [29]	41.62	58.17	72.69	–	–
Hie-Q+I+Pre [14]	43.14	59.44	72.20	–	–
BAN [13]	44.02	58.92	71.34	–	–
BAN†	45.95	63.36	78.12	–	–
MLP†	47.55	66.76	81.55	–	–
SAN†	49.27	67.30	81.79	0.605	14.75
top-1-Qqmaping [27]	52.56	59.72	60.58	–	–
top-3-Qqmaping [27]	56.91	64.65	65.54	–	–
Our method (*soft mask score* = 10)					
$k_r = 3,\ k_e = 1$	**58.27**	75.2	86.4	0.683	11.72
$k_r = 3,\ k_e = 3$	57.42	**76.51**	87.53	**0.685**	10.51
$k_r = 3,\ k_e = 5$	53.84	74.88	**88.49**	0.661	9.58
$k_r = 5,\ k_e = 10$	54.02	74.53	88.03	0.660	**9.17**

Results on F-VQA. To demonstrate the effectiveness of our model under generalized VQA condition, we conduct experiments under standard F-VQA dataset. Results in Table 2 gives an overview of the comprehensive evaluation for some representative approaches over this datasets. It is interesting that the Up-Down and BAN behave worse than SAN, which may be caused by overfitting of the model due to more parameters and limited training data (less than 3000). But among all those settings, the results demonstrate that our models all outperform corresponding classifier-based or mapping-based models to varying degrees. The stable improvement (compare with SAN†) achieved by our model indicates that adding our method to other end-to-end framework under generalized knowledge VQA setting could also lead to stable performance improvement. Most importantly, our proposed KG-based framework is independent of fusion model, which makes it possible for multi-scene migration and multi-model replacement.

Results on ZS-F-VQA. We report the prediction results under the standard ZSL setting and GZSL setting in Table 3. Considering that the traditional classifier-based VQA model fail to work on ZS-VQA since there is no overlap of answer label between the testing set and training set (see Table 1 for detail), we simply skip these methods here. We set larger parameters k under ZSL/GZSL setting to mitigate the strong constraint on answer candidate caused by hard mask. From the overall view, the performance of our model has been significantly improved on the basis of SAN† model.

Most importantly, the models obtain the state-of-the-art performance under respective indicators with various parameter settings. Take the result in GZSL setting as an example, our method achieve 29.39% improvement for *hit*@1 (from 0.22% to 29.39%), 44.17% for *hit*@3 and 75.34% for *hit*@3. We have similar observations when the setting transforms into standard ZSL. To sum up, these observations demonstrate the effectiveness of the model in the ZSL/GZSL scenario, but it also reflects model's dependence on trade off between k_r and k_e (this will be discussed in Sect. 5.4).

Table 3. The overall results (% for $Hit@K$) on ZS-F-VQA datasets under the setting of ZSL/GZSL.

Methods	GZSL					ZSL				
	Hit@1	*Hit*@3	*Hit*@10	*MRR*	*MR*	*Hit*@1	*Hit*@3	*Hit*@10	*MRR*	*MR*
Up-Down†	0.00	2.67	16.48	–	–	13.88	25.87	45.15	–	–
BAN†	0.22	4.18	18.55	–	–	13.14	26.92	46.90	–	–
MLP†	0.07	4.07	27.40	–	–	18.84	37.85	59.88	–	–
SAN†	0.11	6.27	31.66	0.093	48.18	20.42	37.20	62.24	0.337	19.14
Our method (*hard mask score* = 100)										
$k_r = 25,\ k_e = 1$	**29.39**	43.71	62.17	**0.401**	29.52	46.87	62	78.14	0.572	12.22
$k_r = 15,\ k_e = 3$	12.22	**50.44**	73.10	0.339	22.2	**50.51**	70.44	84.24	**0.625**	9.27
$k_r = 15,\ k_e = 5$	6.69	42.91	**75.34**	0.293	20.61	49.11	**71.17**	86.06	0.622	8.6
$k_r = 25,\ k_e = 15$	1.96	24.8	72.85	0.208	18.72	40.21	67.04	**88.51**	0.563	7.68
$k_r = 25,\ k_e = 25$	1.19	18.81	66.97	0.180	**18.14**	35.87	61.86	88.09	0.528	**7.3**

5.4 Ablation Studies

Table 4. The impact of different answer embedding toward model performance (%) on standard F-VQA datasets (TOP-500). $x(a)$, $h(a)$ and $v(a)$ respectively denote KGE, ConceptNet embedding, and original GloVe embedding. CLS is classifier-based method.

Methods	*Hit*@1	*Hit*@3	*Hit*@10
CLS	38.64	54.87	69.38
$v(a)$	**46.32**	**63.96**	**78.44**
$x(a)$	44.13	59.94	71.94
$h(a)$	45.62	62.99	77.34
$v(a) + h(a)$	45.86	63.67	78.43
$v(a) + h(a) + x(a)$	45.18	62.95	77.14

Choice of Answer Embedding. To compare the influence of answer embedding in feature projection performance, we define $g_\phi(a) = g_\phi(\mathcal{C}[x(a); h(a); v(a)])$ in this part where $\mathcal{C}$ denotes simple concatenate function. Specially, $x(a)$, $h(a)$ and $v(a)$ respectively denotes KG embedding (KGE), ConceptNet embedding [15], and original GloVe embedding. This KGE technique can be used to complete the KG with missing entities or links, meanwhile produce the embedding of nodes and links as their representations. Specially, we adopt TransE [4] as $x(.)$ and train it on our KG. As for $h(a)$, we utilize the BERT-based node representations generated by a pre-trained common sense embedding model [15], which exploits the structural and semantic context of nodes within a large scale common sense KG. As the result shown in Table 4, when work independently, word2vec representation (78.44%) of answers exceed graph based methods (71.94% for KGE and 77.34% for ConceptNet Embedding in $Hit@10$) in performance even though they contain more information. We guess that when the size of the dataset is small, the complexity of neural network limits model's sensitivity to the input representation. So finally we simply choose GloVe as the initial representation of all inputs.

Impact of Mask Score. In this part we mainly discuss the effect of mask score on ZS-F-VQA and F-VQA which is reflected by $hit@1$ (Left), $hit@3$ (Middle) and $hit@10$ (Right) accuracy as shown in Fig. 3. Caused by the sparsity of high-dimensional space vector, the value of $F_{\boldsymbol{\theta}\diamond}(i,q)^\top G_{\boldsymbol{\phi}\diamond}(f_n))$ is quite small as we observing on experiment. This is also another reasons why we define τ for the scale-up of vector similarity (in addition to accelerating model convergence). Considering that $sim((i,q),a_n)$ distributes from 145 to 232, we simply take 100 as the dividing line of score between hard mask and soft mask which is big enough for correcting an incorrect answer into a correct one in testing stage. As shown in

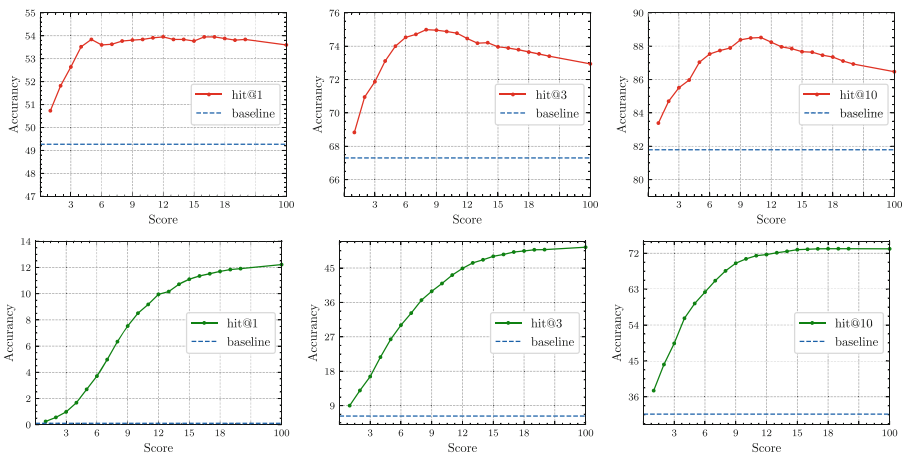


Fig. 3. Impact of mask score in standard F-VQA ($k_r = 3$, $k_e = 10$) under generalized setting (Up), and ZS-F-VQA ($k_r = 15$, $k_e = 3$) under GZSL setting (Down).

Fig. 3, the result gaps between soft mask (i.e., low score) and hard mask (i.e., high score) are completely different in ZSL and GZSL VQA scenarios. We consider following reasons: 1) Firstly, do not try to rely on network to model complex common sense knowledge when data is scarce: When applied to ZS-F-VQA, we notice that model merely learns prior shallow knowledge representation and poor transfer capabilities for unseen answers (see Sect. 5.5). In this case, the strong guiding capability of additional knowledge makes a great contribution to answer prediction. 2) Secondly, if the training samples are sufficient, the error cascading caused by pipeline mode may become the restriction of model performance: When applied to standard F-VQA, the model itself already has high confidence in correct answer and external knowledge should appropriately relax the constraint. We observe that overly strong guidance (i.e., hard mask) becomes a burden at this moment, so soft mask is in demand as a soft constraint. This reflects the necessity of defining different mask.

Impact of Support Entity and Relation. As shown in Fig. 4, we notice that *hit*@1 and *hit*@10 cannot simultaneously achieve the best, despite that the model can always exceed the baseline a lot with different k. This phenomenon is plausible since that the more restrictive target candidate set is, the more likely it succeed predicting answer in a smaller range, with the cost of missing some other true answers due to the error prediction of support entity/relation. The contrast between MRR and MR well reflects this view (see Table 3).

5.5 Interpretability

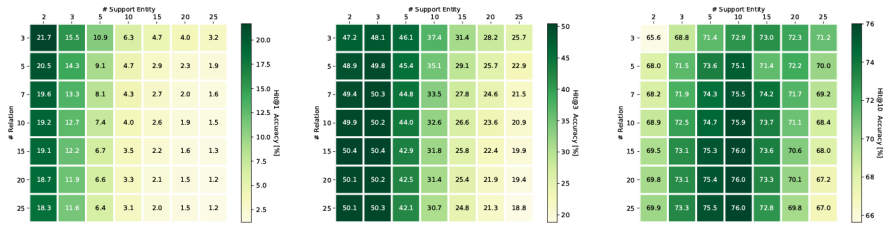


Fig. 4. Impact of #support entity (k_e) and #relation (k_r) on GZSL setting.

To further validate the effectiveness of our knowledge-based ZS-VQA model, we visualize the output and intermediate process of our method compared to best baseline model SAN† [12]. Figure 5 (Up) shows the detected support entities, relations, and answers for four examples in ZS-F-VQA dataset together with answer predicted by SAN† and the groudtruth one. It indicates that normal models tend to align answer directly with meaning content in question/image (e.g. bicycle in Case 3) or match the seen answers (e.g. airplane in case 4), which is a lazy way of learning accompanied by overfitting. To some extend, the difficult common sense knowledge stored in structured data is utilized to playing

a guiding role here. Our method can also be generalized to predict multiple answers since the probabilistic model can calculate scores for all candidates to select the top-K answers (see answer "tv" in Case 2 of Fig. 5).

Question		What thing does the animal in this image have as a part?	Which object in this image could perform a screen	which object in this image is faster than bike?	Which object in this image is related to fly?
Image					
Our Model	Support Entity	zebra, elephant, zoo, giraffe	screen capture, screen, computer display, display image	fast than bicycle, bike, ride, fast than car	fly, catch fly, attempt to fly, learn to fly
	Relation	has a, part of, is a	capable of, used for, related to	slow, expensive, efficient	related to, belong to, specific
	Answer	stripe, horse, string ✓	computer, tv, keyboard ✓	car, bicycle, traffic light ✓	dragonfly, bird, airplane ✓
SAN†	Answer	string, water, ocean ✗	mouse, hand, lamp ✗	bicycle, traffic light, airplane ✗	airplane, kite, fly ✗
Ground Truth		stripe	computer	car	dragonfly

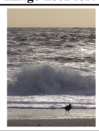
Question		Which object in this image is a cartilaginous fish?	Which object in this image is related to drive?	What is the place in this image used for?
Image				
Our Model	Support Entity	fish, eat fish, fish tank, carp, crab	drive, disk drive, drive on, drive lorry, drive only on track	ocean, sandy, lake, ocean beach, sand
	Relation	is a, belong to, related to	related to, specific, used for	used for, related to, capable of
	Answer	ray, jellyfish, lobster, turtle, sea, fish ✓	horse, cattle, car, cart, train, bicycle ✓	store boat, sail boat, swim, ski, swimming, life preserver ✓
SAN†	Answer	ray, turtle, fish, frog, jellyfish, lobster ✓	horse, cart, cow, sheep, cattle, camel ✓	life preserver, travel across water, sea, desert, ocean , store boat ✗
Ground Truth		grass	cattle	swim

Fig. 5. Cases under GZSL VQA (Up) and Generalized VQA (Down) setting.

Our method also works well under generalized VQA setting as illustrated in Fig. 5 (Down). For those simpler answers, it can increase the probability (e.g. Case 6) for correct prediction. More importantly, distinguish from the hard mask (dark shadows) in ZSL setting, the soft mask strategy here effectively alleviates error cascading which reduces the influence from previous model's error (e.g. failed prediction on support entity lead to the error mask on Case 5).

6 Conclusion

We propose a Zero-shot VQA model via knowledge graph for addressing the problem of exploiting external knowledge for Zero-shot VQA. The crucial factor to the success of our method is to consider both the knowledge contained in the answer itself and the external common sense knowledge from knowledge graphs. Meanwhile we convert VQA from a traditional classification task to a

mapping-based alignment task for addressing unseen answer prediction. Experiments on multiple models support our claim that this method can not only achieve outstanding performance in zero-shot scenarios but also make steady progress at different end-to-end models on the general VQA task. Next we will further investigate KG construction and KG embedding methods for more robust but compact semantics for addressing ZS-VQA. Moreover, we will release and improve the ZS-VQA codes and data, in conjunction with K-ZSL [11].

Acknowledgments. This work is funded by 2018YFB1402800/NSFCU19B2027 /NSFC91846204. Jiaoyan Chen is founded by the SIRIUS Centre for Scalable Data Access (Research Council of Norway) and Samsung Research UK.

References

1. Agrawal, A., Kembhavi, A., Batra, D., Parikh, D.: C-VQA: a compositional split of the visual question answering (VQA) v1.0 dataset. CoRR arXiv:1704.08243 (2017)
2. Anderson, P., et al.: Bottom-up and top-down attention for image captioning and visual question answering. In: CVPR, pp. 6077–6086 (2018)
3. Antol, S., et al.: VQA: visual question answering. In: ICCV, pp. 2425–2433 (2015)
4. Bordes, A., Usunier, N., García-Durán, A., Weston, J., Yakhnenko, O.: Translating embeddings for modeling multi-relational data. In: NIPS, pp. 2787–2795 (2013)
5. Chen, J., Geng, Y., Chen, Z., Horrocks, I., Pan, J.Z., Chen, H.: Knowledge-aware zero-shot learning: survey and perspective. In: IJCAI Survey Track (2021)
6. Chen, L., Gan, Z., Cheng, Y., Li, L., Carin, L., Liu, J.: Graph optimal transport for cross-domain alignment. In: ICML, vol. 119, pp. 1542–1553 (2020)
7. Chen, W., Zha, H., Chen, Z., Xiong, W., Wang, H., Wang, W.Y.: HybridQA: a dataset of multi-hop question answering over tabular and textual data. In: EMNLP, pp. 1026–1036 (2020)
8. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: NAACL, pp. 4171–4186 (2019)
9. Farazi, M.R., Khan, S.H., Barnes, N.: From known to the unknown: transferring knowledge to answer questions about novel visual and semantic concepts. Image Vis. Comput. **103**, 103985 (2020)
10. Geng, Y., et al.: OntoZSL: ontology-enhanced zero-shot learning. In: WWW, pp. 3325–3336 (2021)
11. Geng, Y., Chen, J., Chen, Z., Pan, J.Z., Yuan, Z., Chen, H.: K-ZSL: resources for knowledge-driven zero-shot learning. CoRR arXiv:2106.15047 (2021)
12. Hu, H., Chao, W., Sha, F.: Learning answer embeddings for visual question answering. In: CVPR, pp. 5428–5436 (2018)
13. Kim, J., Jun, J., Zhang, B.: Bilinear attention networks. In: NeurIPS, pp. 1571–1581 (2018)
14. Lu, J., Yang, J., Batra, D., Parikh, D.: Hierarchical question-image co-attention for visual question answering. In: NIPS, pp. 289–297 (2016)
15. Malaviya, C., Bhagavatula, C., Bosselut, A., Choi, Y.: Commonsense knowledge base completion with structural and semantic context. In: AAAI, pp. 2925–2933 (2020)
16. Marino, K., Rastegari, M., Farhadi, A., Mottaghi, R.: OK-VQA: A visual question answering benchmark requiring external knowledge. In: CVPR, pp. 3195–3204 (2019)

17. Narasimhan, M., Lazebnik, S., Schwing, A.G.: Out of the box: reasoning with graph convolution nets for factual visual question answering. In: NeurIPS, pp. 2659–2670 (2018)
18. Narasimhan, M., Schwing, A.G.: Straight to the facts: learning knowledge base retrieval for factual visual question answering. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11212, pp. 460–477. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01237-3_28
19. Pan, J., et al. (eds.): Reasoning Web: Logical Foundation of Knowledge Graph Construction and Querying Answering. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-49493-7
20. Pan, J., Vetere, G., Gomez-Perez, J., Wu, H. (eds.): Exploiting Linked Data and Knowledge Graphs for Large Organisations. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-45654-6
21. Pennington, J., Socher, R., Manning, C.D.: GloVe: global vectors for word representation. In: EMNLP, pp. 1532–1543 (2014)
22. Ramakrishnan, S.K., Pal, A., Sharma, G., Mittal, A.: An empirical evaluation of visual question answering for novel objects. In: CVPR, pp. 7312–7321 (2017)
23. Shah, S., Mishra, A., Yadati, N., Talukdar, P.P.: KVQA: knowledge-aware visual question answering. In: AAAI, pp. 8876–8884 (2019)
24. Shevchenko, V., Teney, D., Dick, A.R., van den Hengel, A.: Visual question answering with prior class semantics. CoRR arXiv:2005.01239 (2020)
25. Teney, D., van den Hengel, A.: Zero-shot visual question answering. CoRR arXiv:1611.05546 (2016)
26. Wang, P., Wu, Q., Shen, C., Dick, A.R., van den Hengel, A.: Explicit knowledge-based reasoning for visual question answering. In: IJCAI, pp. 1290–1296 (2017)
27. Wang, P., Wu, Q., Shen, C., Dick, A.R., van den Hengel, A.: FVQA: fact-based visual question answering. IEEE TPAMI **40**(10), 2413–2427 (2018)
28. Wu, Q., Wang, P., Shen, C., Dick, A.R., van den Hengel, A.: Ask me anything: free-form visual question answering based on knowledge from external sources. In: CVPR, pp. 4622–4630 (2016)
29. Yang, Z., He, X., Gao, J., Deng, L., Smola, A.J.: Stacked attention networks for image question answering. In: CVPR, pp. 21–29 (2016)
30. Ye, Z., Chen, Q., Wang, W., Ling, Z.: Align, mask and select: a simple method for incorporating commonsense knowledge into language representation models. CoRR arXiv:1908.06725 (2019)
31. Zhu, Z., Yu, J., Wang, Y., Sun, Y., Hu, Y., Wu, Q.: Mucko: multi-layer cross-modal knowledge reasoning for fact-based visual question answering. In: IJCAI, pp. 1097–1103 (2020)