Query and Attention Augmentation for Knowledge-Based Explainable Reasoning

Yifeng Zhang, Ming Jiang, Qi Zhao
University of Minnesota
{zhan6987, mjiang}@umn.edu, qzhao@cs.umn.edu

Abstract

Explainable visual question answering (VQA) models have been developed with neural modules and query-based knowledge incorporation to answer knowledge-requiring questions. Yet, most reasoning methods cannot effectively generate queries or incorporate external knowledge during the reasoning process, which may lead to suboptimal results. To bridge this research gap, we present Query and Attention Augmentation, a general approach that augments neural module networks to jointly reason about visual and external knowledge. To take both knowledge sources into account during reasoning, it parses the input question into a functional program with queries augmented through a novel reinforcement learning method, and jointly directs augmented attention to visual and external knowledge based on intermediate reasoning results. With extensive experiments on multiple VQA datasets, our method demonstrates significant performance, explainability, and generalizability over state-of-the-art models in answering questions requiring different extents of knowledge. Our source code is available at https://github.com/SuperJohnZhang/QAA.

1. Introduction

Reasoning about knowledge is essential for general intelligent behavior [32]. Humans have the innate ability to acquire and incorporate concepts from multiple knowledge sources, yet to simulate this mechanism with machine intelligence is nontrivial. Visual question answering (VQA) is a typical task that requires both knowledge acquisition and knowledge reasoning abilities. A desirable VQA system should understand both inputs (*i.e.,* image and question) and perform cross-modal reasoning by seeking supporting logic and evidence that lead to a reasonable answer.

Most VQA methods learn to answer questions based on the statistical correlations between the multi-modal inputs and the answer [5, 10, 11]. Studies have shown that such implicit data-driven methods tend to exploit language pri-

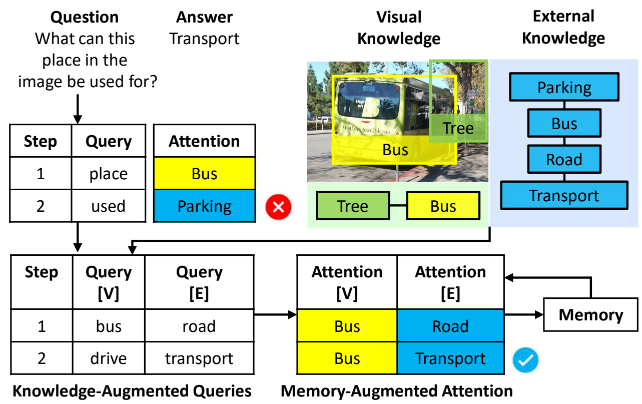

Figure 1. Based on neural module networks and explicit knowledge representation, we develop knowledge-augmented queries and memory-augmented attention to jointly reason about both the visual [V] and external [E] knowledge. These query and attention augmentation methods generalize explainable visual reasoning models to better answer knowledge-requiring questions.

ors to achieve high performance, instead of reasoning based on logic and evidence [33]. To perform multimodal reasoning, recent studies have leveraged neural modules networks (NMNs) [1] that explicitly model the multi-step reasoning process [15, 17, 43]. They parse the input question into a functional program and dynamically assemble a network of explainable neural modules to execute the program. They not only achieve remarkable performances in VQA but also provide step-by-step explanations to help understand the reasoning process behind the predicted answer [17, 33].

NMNs are commonly developed on datasets of synthetic and structured questions, such as CLEVR [19] and GQA [16], which are limited in generalization. To answer more general VQA questions while maintaining explainability, several studies have incorporated external knowledge based on explicit scene graph modeling [7, 45] or implicit feature enrichment [22, 26, 42]. They query external concepts from knowledge bases, integrate the acquired external knowledge with the observed visual knowledge, and finally conduct the reasoning on the integrated knowledge space [7, 41, 45]. Such approaches result in a loose integration between knowledge and reasoning, which may be suboptimal when dealing with complex reasoning problems. In this work, we propose Query and Attention Augmentation, an NMN-based explainable visual reasoning method that answers knowledge-requiring questions by jointly reasoning about both visual knowledge (*i.e.,* the visual features) and external knowledge (*i.e.,* the semantic embeddings of external concepts). Different from previous methods that incorporate knowledge prior to the reasoning process, it tightly couples knowledge incorporation with reasoning, which addresses two major research gaps:

First, previous methods generate functional program based only on the input question, without considering the visual or external information. As shown in Fig. 1, to answer the question "What can this place in the image be used for?", they may generate a program of two functions: 1) recognizing the place and 2) finding its usage. Two input tokens (*e.g., place* and *used*) can be extracted from the question and used as queries to guide the model's attention and reasoning. Since they are extracted from the question only, the queries may be less relevant to the context and result in a wrong answer (*e.g., Parking*). In this work, we propose to augment these question-based queries with visual and external knowledge, so that they can be more specific and relevant. For example, as shown in Fig. 1, after the augmentation, two sets of queries are generated to guide the reasoning of visual knowledge (*e.g., bus* and *drive*) and external knowledge (*e.g., road* and *transport*), respectively. Compared with the original queries, they guide the attention of NMNs more directly to find the answer.

Second, previous methods typically acquire and incorporate external knowledge as supporting features prior to reasoning [7, 41, 45]. However, during multi-step visual reasoning, the reasoning context is dynamically updated throughout the process, where additional knowledge may need to be acquired and understood along the way. To enable this ability, we propose to jointly reason about the visual and external knowledge and use a novel memory-augmented attention method to integrate their intermediate results for reasoning, so the knowledge is integrated during the reasoning process instead of only at the beginning of it. As shown in Fig. 1, jointly directing attention to important visual knowledge (*e.g., Bus*) and external knowledge (*e.g., Road*, *Transport*) throughout the reasoning process can help NMNs make better use of both knowledge sources to find the correct answer (*e.g., Transport*).

In sum, by addressing these challenges, our proposed method allows NMNs to accurately direct attention to important features in both visual and external knowledge and answer knowledge-requiring questions. The contributions of this work are summarized as follows:

1. To the best of our knowledge, this work is the first attempt to jointly reason about visual knowledge and external knowledge based on neural module networks.

2. With reinforcement learning, we generate knowledge-augmented queries to incorporate visual and external knowledge into the functional program.

3. By sharing intermediate results between the two knowledge sources with memory-augmented attention, we enable external knowledge incorporation throughout the reasoning process.

4. Our extensive experiments on multiple VQA datasets demonstrate the effectiveness, generalizability, and explainability of the proposed method.

2. Background: neural module networks

In general, NMNs perform explainable visual reasoning in two steps: they generate a program from the input question by composing a sequence of predefined functions and execute the program by implementing each function using small neural networks (*i.e.,* neural modules). They are typically designed with the following components:

Knowledge representation. Preprocessing the visual and semantic inputs into high-level knowledge representation allows visual reasoning models to focus on learning to reason about knowledge rather than the direct correlation between the input features and answers. NMNs typically encode the visual input into pixel-based [14, 15], region-based [7,34], or graph-based [17,33] features. In this work, we extract the high-level semantics and relationships from the visual input and external knowledge bases, and explicitly organize such knowledge as structured representations (*i.e.,* scene graphs and knowledge graphs). We generate scene graphs with VC-Tree [36] to represent objects and their relationships. The external knowledge graph is constructed from the ConceptNet [23], Visual Genome [21] and WordNet [9] following the KI-Net method [45].

Program generation. Learning to map the free-form natural language input into the structured functional program is a challenging task, due to the variability of real-world questions and the absence of explicit program supervision. Most NMNs design a program generator following the encoder-decoder architecture, to convert a sequence of word embeddings into a sequence of parameterized functions. For example, the StackNMN [14] uses a bidirectional LSTM [12] to predict specific modules and their textual parameters, and NSM [17] generates more general queries. These simplified approaches may work well when the question has a regular grammatical structure and it only considers the in-domain knowledge from the training data. For free-from questions involving out-of-domain knowledge, the conventional end-to-end data-driven approach may not correctly understand the question, which leads to the degradation in the visual reasoning performance. To address this challenge, instead of relying on the question itself, we generate knowledge-augmented queries by taking the visual and external knowl-

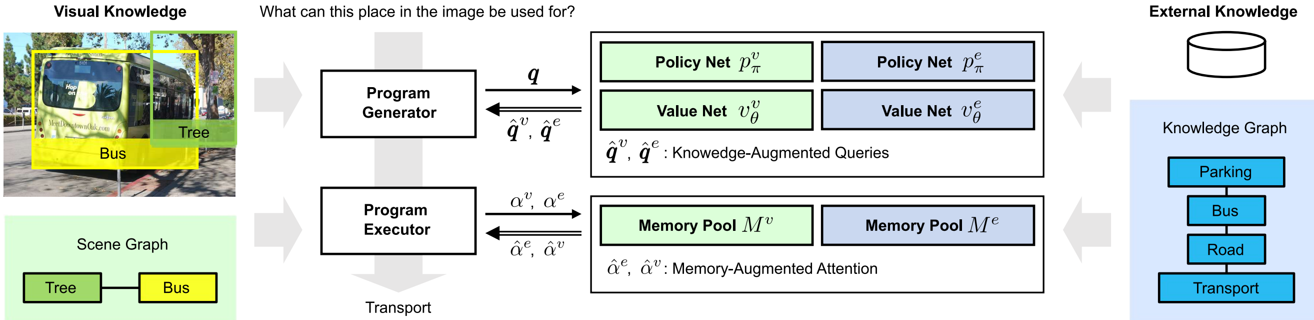

Figure 2. The overview of our method. First, it represents the visual input as a scene graph and the external information as a knowledge graph. Second, a program generator parses the question to predict the functional program and its corresponding queries $\boldsymbol{q}$. Next, two reinforcement learning agents augment $\boldsymbol{q}$ with the visual and external knowledge, resulting in the augmented queries $\hat{\boldsymbol{q}}^v$ and $\hat{\boldsymbol{q}}^e$, respectively. Further, they are used as parameters for the program executor to allocate attention (*i.e.,* α^v and α^e). Based on the memorized intermediate results M^v and M^e, it computes the augmented attention vectors $\hat{\alpha}^v_t$ and $\hat{\alpha}^e_t$, which jointly consider both knowledge sources to better allocate attention. Finally, it predicts the answer based on the attended features.

edge into account (see Sec. 3.1).

Program execution. NMNs dynamically assemble the neural modules into a complete network, to execute the generated program and output an answer to the input question. These modules play different roles during the program execution: querying the relevant knowledge by allocating or re-allocating attention to the input features (*e.g., attend*, *relate*), recognizing the attended features (*e.g., describe*), or performing numeric (*e.g., exist*, *count*, *compare*) or logical (*e.g., and*, *or*, *not*) operations, *etc.* Though previous studies have explored the incorporation of external knowledge in VQA tasks, they typically encode knowledge as supporting features to enrich the visual features before the reasoning process [7, 41, 45]. Different from existing NMNs that do not explicitly query external knowledge during the program execution, we enable NMNs to concurrently query, memorize, and share information across both knowledge sources with memory-augmented attention (see Sec. 3.2).

3. Methodology

The goal of this work is to develop an explainable NMN method that answers questions based on the supporting evidence acquired from visual and external knowledge. The key differentiating factor of our method is its ability to interact with the two knowledge sources during the generation and execution of the program. The novelty lies in two major components: 1. it augments the generated queries with knowledge from the visual input and the external knowledge base and 2. jointly allocates attention to both the visual and external knowledge and augments the attention based on information sharing supported by memorized intermediate results. Fig. 2 summarizes how this is achieved. In this section, we describe the main components of our method: knowledge-based query augmentation and memory-based attention augmentation. For further details, please refer to the Supplementary Materials.

3.1. Knowledge-based query augmentation

NMN-based methods typically adopt an encoder-decoder network to generate a sequence of reasoning functions and their corresponding queries (*i.e.,* parameters) from the input question. Based on the queries generated by an existing method (*e.g.,* NSM [17]), we propose a reinforcement learning method that generates knowledge-augmented queries for each knowledge source (*i.e.,* visual or external knowledge). Specifically, at each reasoning step, we learn query augmentation agents to select the most plausible queries from a vocabulary of relevant semantic concepts. Different from conventional query expansion methods [13, 31, 40], we adopt reinforcement learning [27, 30] to learn the agents, which allows us to efficiently choose the optimal queries from the large amount (≈100K) of semantic concepts and to optimize the network parameters in the end-to-end VQA training.

Query vocabularies. NMNs typically select queries from a vocabulary of semantic concepts and use their semantic embeddings for explainable reasoning. For example, NSM [17] builds a vocabulary using three categories of semantics in the training dataset: object identities, attributes, and relationships. In our method, to include out-of-domain knowledge from external databases (*e.g.,* ConceptNet [23], Visual Genome [21], WordNet [9]), we select queries from a sample-specific vocabulary of relevant concepts extracted from the external knowledge graph.

Specifically, we represent the functional program as a sequence of T executable neural modules with queries q_t ($t = 1, \ldots, T$), which is generated by an existing NMN method [33]. For each step t, we create a vocabulary C_t with its items semantically relevant to the query q_t:

$$\forall c^i_t \in C_t,\ d(c^i_t, q_t) \leq L_d, \tag{1}$$

where c^i_t is a semantic concept obtained from the external knowledge graph, and $d(\cdot, \cdot)$ measures the graph distance (*i.e.,* length of the shortest path) between the input concepts in the knowledge graph. We represent the vocabulary as an ordered list sorted by each item's distance to q_t. The maximal distance L_d controls the size of the vocabulary.

Query augmentation agents. Instead of greedily seeking the most relevant queries from the vocabularies, we formulate the query selection as a decision-making process and design reinforcement learning agents to optimize the selection. In particular, we design a visual-knowledge agent and an external-knowledge agent and reward them for selecting complementary queries that guide the reasoning about the visual and external knowledge, respectively.

Specifically, our goal is to find the optimal queries $\hat{\boldsymbol{q}} = [\hat{q}_1, \ldots, \hat{q}_T]$ that are relevant to not only the question but also the visual and external knowledge. At each step t, each agent predicts the next query by selecting a query from the vocabulary, *i.e.,* $\hat{q}_{t+1} \in C_{t+1}$. It observes the current state $s_t = [\hat{q}_1, \ldots, \hat{q}_{t-1}]$ consisting of the predicted queries so far. The environment $\mathcal{E}_t$ includes the visual features V, the vocabulary C_t, the input queries $[q_1, \ldots, q_t]$.

The **policy network** p_π predicts the probability for the agent to select a query as the next output, $p_\pi(\hat{q}_t | s_t, \mathcal{E}_t)$. As shown in Fig. 3, following a basic encoder-decoder framework [24], we use a CNN-based encoder to extract visual features h^v and an LSTM-based language encoder to embed the vocabulary C_t into a semantic vector h^c_t. The features h^v and h^c_t are concatenated and fed into another LSTM encoder, while an LSTM-based decoder integrates the encoded features $\boldsymbol{u}$ with the input queries $q_1, \ldots, q_t$ to predict the policy at time t. Based on the policy, the query with the highest probability is selected as the output $\hat{q}_t$ and are fed back to the decoder in the next step as q_{t+1}.

The **value network** v_θ approximates a value function v_p that predicts the total reward r from the observed state s_t, assuming that the decision making process is following a policy p. It serves as an evaluation of the state s_t. As shown in Fig. 3, it encodes the augmented queries $[\hat{q}_1, \ldots, \hat{q}_t]$ with a LSTM model, and predicts the total reward r based on the LSTM output h^s, the visual features h^v, and the semantic features h^c_t, using a multi-layer perceptron (MLP).

Reward definition. A well-defined reward for the optimization of query augmentation is important. With a goal of making the augmented queries (*i.e.,* $\hat{\boldsymbol{q}}^v = [q^v_1, \ldots, q^v_T$ and $\hat{\boldsymbol{q}}^e = [q^e_1, \ldots, q^e_T)$ relevant to the question and their corresponding knowledge (*i.e.,* visual knowledge and external knowledge), we define a specific reward function for each of the two agents. These functions compute rewards based on the queries (*i.e.,* $\hat{\boldsymbol{q}}^v$ or $\hat{\boldsymbol{q}}^e$), visual features h^v, semantic features $h^c = [h^c_1, \ldots h^c_T]$, and the ground-truth answer y.

First, we use a pre-trained visual-semantic embedding model [30] to project these features into a joint embedding space. Let $\delta(\cdot, \cdot)$ indicate the cosine similarity measure and g^{qv}, g^{qe}, g^v, g^c, g^y indicate the embeddings of $\hat{\boldsymbol{q}}^v$, $\hat{\boldsymbol{q}}^e$, h^v, h^c, y, respectively. We define the reward r^v of the visual-knowledge agent and reward r^e of the external-knowledge agent to enforce the generated queries to focus on complementary yet relevant aspects of the knowledge:

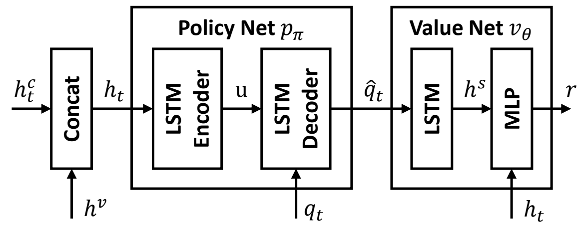

Figure 3. Each query augmentation agent consists of a policy network and a value network. The policy network predicts the augmented queries $\hat{\boldsymbol{q}}$ from the visual feature h^v, the semantic vector h^c, and the base queries $\boldsymbol{q}$. The value network evaluates the policy and predicts the total reward r.

$$r^v = \delta(g^{qv}, g^y) + \eta^v \delta(g^{qv}, g^v), \tag{2}$$

$$r^e = \delta(g^{qe}, g^y) + \eta^e \delta(g^{qe}, g^c), \tag{3}$$

where η^v and η^e balance the weights of the corresponding terms. Higher values of these hyperparameters encourage the two agents to generate more distinct queries. These rewards allow the two agents to generate complementary queries based on different knowledge sources. For each agent, the policy network and the value network are jointly optimized to approximate the total reward.

Training. We use deep reinforcement learning with our proposed reward to learn the policy and value network. Following [30], we train the networks in two steps:

First, following the common practice [28,35,38], we pretrain the policy network and the value network using supervised learning to initialize them with plausible parameters. We supervise the policy network with the base queries $\boldsymbol{q}$ and the cross-entropy loss $L_p = -\sum_{t=1}^{T} \log p_\pi(\hat{q}_t | s_t, \mathcal{E}_t)$. We supervise the value network with corresponding total final reward r and the mean squared loss $L_v = ||v_\theta(s_t) - r||^2$.

After pretraining, we jointly train the policy network and the value network with reinforcement learning. The training follows an actor-critic approach [20]. Note that both agents are trained with different rewards for maximizing their embedding relevance to the visual ($r = r^v$) and external knowledge ($r = r^e$), respectively. With Monte Carlo tree search (MCTS) [6], the two agents can output augmented queries $\hat{\boldsymbol{q}}^v$, $\hat{\boldsymbol{q}}^e$ that will be used to execute the program. The augmented queries allow neural modules to concurrently reason about the visual and external knowledge.

3.2. Memory-based attention augmentation

NMNs adopt attention mechanisms to highlight important knowledge for reasoning. Despite that different NMNs (*e.g.,* NSM [17] and XNM [33]) implement their neural

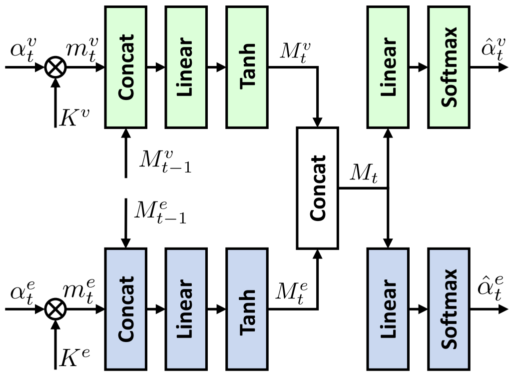

Figure 4. The proposed attention augmentation method processes the visual and external knowledge features K^v and K^e with the original attention vectors α_t^v and α_t^e to predict the memory-augmented attention vectors $\hat{\alpha}_t^v$ and $\hat{\alpha}_t^e$.

modules in different ways, their intermediate attention outputs can be similarly represented as a sequence of normalized weight vectors. We adapt existing NMN methods so that each module processes two queries and produces two attention vectors. Each attention vector is guided by the corresponding queries and further augmented with memorized intermediate results, which enables NMNs to accurately attend to both visual knowledge and external knowledge in the reasoning process.

Independent attention allocation. To jointly reason about the visual and external knowledge, each neural module is adapted to process a pair of input queries $q_t = [q_t^v, q_t^e]$ concurrently, and computes the corresponding attention vectors $\alpha_t = [\alpha_t^v, \alpha_t^e]$ to obtain the attended features from the visual scene graph and the external knowledge graph, denoted with superscripts v and e, respectively.

Memory update. We apply the attention mechanism to the features K^v, K^e, and obtain the attended features $m_t^v = \alpha_t^v K^v$, $m_t^e = \alpha_t^e K^e$ for each neural module. As the attended features from each source can serve as a piece of evidence to support the reasoning of its counterpart (see Fig. 1), enabling information sharing across the two knowledge sources will potentially improve the model's reasoning performance. Therefore, inspired by Memory Networks [39] and related studies [18], we develop two separate memories $M_t = [M_t^v, M_t^e]$ to store and retrieve the intermediate features. Specifically, to memorize the features for future queries, we append them to the end of the memories, and further encode the memories with a linear layer:

$$M_t^v = tanh(W_m^v[M_{t-1}^v, m_t^v]), \tag{4}$$
$$M_t^e = tanh(W_m^e[M_{t-1}^e, m_t^e]), \tag{5}$$

where W_m^v and W_m^e are trainable parameters.

Attention augmentation. Given the memories M_t, we augment the attention vectors with the memorized features:

$$\hat{\alpha}_t^v = softmax(W^v M_t), \tag{6}$$
$$\hat{\alpha}_t^e = softmax(W^e M_t), \tag{7}$$

where $\hat{\alpha}_t^v, \hat{\alpha}_t^e$ are the augmented attention vectors, and W^v, W^e are trainable parameters. By augmenting the attention with both memories, our method jointly considers both knowledge sources when allocating attention, to better localize the relevant features during the reasoning process.

4. Experiments and results

We demonstrate our method with experiments on OK-VQA [26], FVQA [37], GQA [16] and VQA v2 [3] datasets. It outperforms the state-of-the-art visual reasoning models, demonstrating its ability to answer both knowledge-requiring questions and general questions with explainable reasoning. Ablation studies display how the two augmentation methods independently and jointly contribute to the improvements of the reasoning performance. Quantitative and qualitative results show that the incorporation of external knowledge during the program generation and execution stages significantly improves visual reasoning performance.

4.1. Experimental settings

Datasets. We conduct extensive experiments to evaluate the proposed method on four different VQA datasets. The OK-VQA [26] and FVQA [37] are general VQA datasets specifically designed for questions requiring commonsense and factual knowledge to answer. In particular, FVQA offers ground-truth factual knowledge that can be used to support the training and evaluation of knowledge-based VQA models. The GQA [16] dataset focuses on compositional reasoning with 1.7M structured questions. The VQA v2 [3] dataset is a general VQA dataset that contains 1.1M questions, each annotated with 10 ground-truth answers. With these complementary datasets, we comprehensively evaluate the effectiveness and generalizability of our method.

Training and evaluation. We train NMNs on the training set of datasets and evaluate them on the corresponding validation set. The training of our method consists of three stages: first, we pretrain a baseline model (*e.g.,* NSM [17] or XNM [33]) under the conventional VQA setting. Next, we generate the functional program with the pretrained model and independently train the two query augmentation agents by optimizing their total rewards. Finally, we augment the program queries with these agents, execute the augmented program with memory-augmented attention, and fine-tune the entire network. For a fair comparison, we adapt XNM's *Find* module so that its inputs are similar to NSM's queries. Since few comparable NMNs perform knowledge-based reasoning, we focus our evaluation on the comparisons with

Method	OK-VQA	FVQA	GQA	VQA v2
FVQA [37]	–	64.65	–	–
OutOfBox [29]	–	65.80	–	–
KVQA [46]	29.03	–	–	–
KAN [44]	–	66.39	–	67.42
XNM [33]	25.61	63.74	62.04	64.72
+ AN [26]	25.98	64.11	62.14	65.54
+ KI-Net [45]	26.47	64.42	62.38	64.78
+ Ours	26.52	65.46	63.07	65.92
NSM [17]	26.79	64.08	63.17	65.77
+ AN [26]	27.14	64.73	63.39	66.83
+ KI-Net [45]	28.45	65.12	63.48	65.93
+ Ours	**29.24**	**68.74**	**63.82**	**67.69**

Table 1. Quantitative comparison with state-of-the-art models.

Method	OK-VQA	FVQA	GQA	VQA v2
XNM [33]	25.61	63.74	62.04	64.72
+ MA	26.24	64.78	62.27	65.37
+ KQ (V-Only)	26.10	64.33	62.32	65.21
+ KQ (E-Only)	25.87	64.29	62.48	65.07
+ KQ	26.38	65.09	62.74	65.53
+ QE	25.81	65.18	62.89	65.52
+ Ours	26.52	65.46	63.07	65.92
NSM [17]	26.79	64.08	63.17	65.77
+ MA	27.91	64.92	63.28	65.74
+ KQ (V-Only)	28.23	65.47	63.24	65.97
+ KQ (E-Only)	27.86	65.24	63.23	65.89
+ KQ	28.42	66.39	63.31	66.45
+ QE	28.37	65.94	63.04	66.28
+ Ours	**29.24**	**68.74**	**63.82**	**67.69**

Table 2. Results of different components (*i.e.,* KQ and MA).

the baseline method AN [26] and the state-of-the-art KI-Net [45]: the former enriches the visual features with the language embedding of external concepts and the latter explicitly incorporates knowledge by adding external nodes to the scene graph. We demonstrate the generalizability of our method by applying it to two NMN-based reasoning models: XNM [33] and NSM [17]. For a fair comparison, all compared models are trained and evaluated under the same single-model setting, without ensemble or language pretraining.

Implementation details. In our experiments, each query is represented as a semantic embedding with dimensionality $d_p = 300$. The dimensionality of visual features h^v, semantic features h^c_t, hidden state of the value network as well as memories M^v, M^e are also set to 300. Based on ablation studies (see Supplementary Materials), we set the hyperparameters $\eta^v = 0.6$, $\eta^e = 0.8$, and $L_d = 3$.

4.2. Performance evaluation

We present the quantitative results of our method compared with state-of-the-art knowledge-based visual reasoning methods, including non-NMN methods [29, 37, 44, 46] and different knowledge incorporation approaches [26, 45] applied to the XNM [33] and NSM [17] models.

Comparison with non-NMN methods. The first panel of Tab. 1 presents the performance (*i.e.,* answer accuracy in percentage) of several non-NMN methods [29, 37, 44, 46]. The FVQA [37] generalizes VQA models with feature-based external knowledge enrichment. OutOfBox [29] leverages graph convolution networks to encode the high-level factual semantics and achieves higher performance on the FVQA dataset. KVQA [46] and KAN [44] leverage multi-modal attention to better attend to the necessary visual or factual features. Regardless of their attention mechanisms or feature integration methods, they all focus on the learning of statistical correlations and incorporate external knowledge in a single feature enrichment step. Differently, our method leverages external knowledge throughout the entire process of multi-step structured reasoning. It not only achieves higher performances, but also offers better explainability because of the nature of NMN methods.

Comparison with other knowledge incorporation methods based on NMNs. In the second and third panels, Tab. 1 also shows that our method outperforms the compared AN [26] and KI-Net [45] methods on the two baseline models (*i.e.,* XNM [33] and NSM [17]). Based on NSM, it achieves the highest accuracy on all datasets, especially for questions that can only be answered with external knowledge (*e.g.,* OK-VQA and FVQA), which suggests that our method can better query relevant knowledge from external knowledge and use the external knowledge for reasoning. Though questions in GQA and VQA v2 do not require as much external knowledge, our method still outperforms AN [26] and KI-Net [45]. On the GQA dataset, our improvements over the XNM are more significant, because the XNM's baseline performance is limited by its more explainable but restricted semantic definition of neural modules. The performance improvements on the GQA dataset show our effective utilization of external knowledge.

Contributions of query and attention augmentation. Tab. 2 compares the contributions of knowledge-augmented queries (KQ) and memory-augmented attention (MA). On top of each baseline, we independently apply KQ or MA, and compare their results with the full model. Specifically, the "+ MA" models use the base queries to allocate attention in both the scene graph and the knowledge graph, with the help of MA. Differently, the "+ KQ" models generate two sets of knowledge-aware queries to independently reason about each source without MA. The results in Tab. 2 suggest that the KQ and MA can independently contribute to the VQA performance. They also help NMNs better exploit external knowledge in visual reasoning with a positive joint effect. An interesting observation is that MA contributes

Method	VT	BCP	OMC	SR	CF	GHLC	PEL	PA	ST	WC	Other
KVQA [46]	**27.53**	24.17	21.56	**35.72**	28.20	25.44	**25.38**	30.97	24.35	42.76	25.76
XNM [33]	26.84	21.86	18.22	33.02	23.93	23.83	20.79	24.81	21.43	42.64	24.39
+ AN [26]	25.41	21.39	20.24	33.52	24.68	23.15	20.59	25.09	22.79	43.58	24.72
+ KI-Net [45]	25.74	21.93	20.72	33.69	24.80	23.61	19.83	25.06	22.54	43.08	24.12
+ Ours	25.31	22.04	19.67	33.45	25.37	25.16	21.42	25.29	23.73	44.89	24.98
NSM [17]	27.12	22.54	19.07	33.22	26.78	23.47	20.54	26.73	21.55	37.92	23.13
+ AN [26]	27.17	22.69	20.06	33.76	27.25	24.36	21.63	28.91	21.98	38.96	24.06
+ KI-Net [45]	27.36	22.98	20.51	34.37	27.94	24.85	22.69	30.74	22.79	40.82	24.78
+ Ours	27.49	**24.84**	**21.78**	35.50	**28.39**	**25.87**	25.11	**31.06**	**24.51**	**44.86**	**25.36**

Table 3. Evaluation results of methods with external knowledge on specific question topics in the OK-VQA validation set.

Method	**OK-VQA**
NMN [2]	24.63
NS-VQA [43]	25.79
NSM [17]	27.91
NS-CL [25]	27.42
NSM + KQ (Ours)	**29.24**

Table 4. Comparison between KQ and state-of-the-art program generators. MA is applied to all the compared methods.

Dataset	Visual Genome	ConceptNet	WordNet	All
OK-VQA	28.73	28.59	28.26	**29.24**

Table 5. Results of NSM + Ours with different knowledge bases.

significantly to the performance of the XNM model, which suggests that our MA method can effectively improve the XNM's original attention mechanism that may fail to select important knowledge.

Effectiveness of query augmentation. To demonstrate the effectiveness of our reinforcement learning approach for query augmentation, we compare it with a standard query expansion ("+ QE") method based on the cosine similarity of semantic embedding [4]. Tab. 2 shows that our method outperforms query expansion significantly on the NSM baselines, especially for less structured questions (*e.g.,* OK-VQA and VQA v2). For the XNM baselines, the augmented queries of KQ are less effective in directing the attention shift of more specific modules. To evaluate the effectiveness of each agent in KQ, we reason about both knowledge sources with only one set of queries (*i.e.,* either V-Only or E-Only). Comparing the two knowledge sources, we observe that visual knowledge is more effective than external knowledge during query augmentation, and the combination of both further improves the performance. It suggests that both agents can augment queries with complementary knowledge to jointly improve the reasoning performance.

Topic-specific results. Tab. 3 presents experimental results regarding the 11 question topics of the OK-VQA dataset that requires external knowledge. Compared with KVQA [46] and state-of-the-art NMN-based methods, our method demonstrates its advantages on most of the topics. It significantly improves the performance of XNM and NSM on topics requiring a broader search through the knowledge graph, such as Science and Technology (ST), Plants and Animal (PA), Weather and Climate (WC). Its performance gain is less significant on Vehicle and Transportation (VT), Objects Material and Clothing (OMC), and Sports and Recreation (SR) and People and Everyday Life (PEL) because the knowledge area of these topics is relatively narrow.

Comparison between KQ and common program generators. We further evaluate the performance of KQ against the program generators of several common NMN to validate the necessity of query augmentation with visual and external knowledge. Since existing generators all generate a single sequence of queries, we duplicate the sequence and pass it to the neural modules to reason about both knowledge sources with MA. Tab. 4 compares the performance of these methods on the OK-VQA dataset. NSM [17] and NS-VQA [43] leverage LSTM-based models and rely on signals from answers to weakly supervise the program generation, while NMN [2] applies Standford Parser [8] to retrieve and convert sentence dependency to program layouts and queries. Differently, NS-CL [25] leverages a reinforcement learning method to train the generator, but still only considers the question information. Our knowledge-augmented queries outperform all the compared program generators.

Comparison of knowledge bases. Tab. 5 compares the effects of different knowledge bases. Our method achieves a significant performance improvement when combining Visual Genome, ConceptNet, and WordNet, suggesting the complementary nature of the three knowledge bases.

4.3. Qualitative results

Fig. 5 further demonstrates our method with qualitative results on the NSM model [17] and FVQA dataset [37]. It presents the images, questions, answers, base queries and augmented queries, and the attended visual/external knowledge (*i.e.,* relation triplets with their attention values above average). It shows that our method replaces the base queries

	(a)	(b)	(c)	(d)
Images				
Questions	Which object in this image could be used to heat a pot?	Which thing in the image do you need in a dark room?	What is the stick style instrument in the image?	What can this place in the image be used for?
Answers	**NSM+Ours:** stove **NSM+KI-Net:** oven **NSM+AN:** oven **GT:** stove	**NSM+Ours:** lamp **NSM+KI-Net:** luggage **NSM+AN:** phone **GT:** lamp	**NSM+Ours:** flute **NSM+KI-Net:** trumpet **NSM+AN:** suit **GT:** flute	**NSM+Ours:** transport **NSM+KI-Net:** parking **NSM+AN:** parking **GT:** transport
Queries	**B-Q:** 1.object, 2.heat, 3.pot **V-Q:** 1.oven, 2.heat, 3.pot **E-Q:** 1.stove, 2.boil, 3.soup	**B-Q:** 1.thing, 2.dark, 3.room **V-Q:** 1.furniture, 2.dark, 3.room **E-Q:** 1.lamp, 2.light, 3.indoor	**B-Q:** 1.stick, 2.instrument **V-Q:** 1.stick, 2.instrument **E-Q:** 1.flute, 2.music	**B-Q:** 1.place, 2.use **V-Q:** 1.bus, 2.drive **E-Q:** 1.road, 2.transport
Visual	stove-leftOf-oven	lamp-rightOf-bed luggage-rightOf-lamp pillow-topOf-bed	man-wear-suit man-play-trumpet man-play-flute	bus-leftOf-tree bus-leftOf-sign
External	pot-topOf-stove stove-capableOf-heat stove-capableOf-boil	lamp-capableOf-light light-relationOf-dark lamp-locationOf-indoor	flute-relationOf-stick flute-typeOf-instrument flute-relationOf-trumpet instrument-relationOf-music	bus-capableOf-transport road-capableOf-transport road-relationOf-place

Figure 5. Qualitative results on the FVQA dataset. Each example shows the input image, question, ground-truth (GT) answer and model predictions, base queries (B-Q) and the queries augmented with visual knowledge (V-Q) and external knowledge (E-Q), followed by the attended visual and external knowledge. Highlighted knowledge indicates the FVQA supporting fact of the question.

with more specific objects in the visual scenes (*e.g., oven vs. object*, *furniture vs. thing*) and complements with external knowledge that helps the neural modules to answer correctly. For example, in Fig. 5a, both *stove* and *oven* are capable of heating, but only *stove* can heat a *pot*. Since KI-Net and AN mainly depend on the visual semantics to choose relevant external knowledge and *pot* is absent from the scene, they fail to incorporate important external knowledge to help distinguish the two similar objects (*oven* and *stove*). Our method augments queries to include the external knowledge *boil* that is related to *heat* and *pot* and the answer *stove*. It allows neural modules to allocate memory-augmented attention to relationships of *stove*: *pot-topOf-stove*, *stove-capableOf-heat*, and *stove-capableOf-boil*, to answer correctly. Similarly, in Fig. 5b-d, our method incorporates the answers (*e.g., lamp*, *flute*, and *transport*) and their relevant external knowledge (*e.g., light*, *music*, and *road*). The augmented queries precisely correspond to the supporting facts (*i.e.,* the FVQA ground-truth knowledge) and other important external relationships. These examples show the improved performance and explainability of our method resulted from more specific queries and more accurate attention allocation.

5. Conclusion

We proposed a novel query and attention augmentation approach to explainable visual reasoning with knowledge. It leverages knowledge-augmented queries and memory-augmented attention to explicitly incorporate visual and external knowledge during the reasoning process. It allows neural module networks to concurrently interact with visual and external knowledge, bridging the research gap of explicit and explainable knowledge incorporation in visual reasoning. Our method demonstrates state-of-the-art performance in answering knowledge-requiring questions and general questions. The transparency of NMN models allows researchers to identify limitations and diagnose errors more effectively. We hope that with the proposed query and attention augmentation methods, our work will benefit the future development of more general and explainable reasoning models.

Broader impact. Most deep learning methods make decisions based on black-box models trained on large-scale datasets, which has greatly limited their interpretability or generalizability. By leveraging external knowledge bases, this work develops visual reasoning models that are less dependent on training data and thus releases the heavy workload of data annotation that requires domain knowledge. It also leverages neural module networks that explicitly define and execute reasoning operations, which improves the transparency of decision-making processes and the trustworthiness of deep learning models. This work may benefit future applications in many domains where both domain knowledge and system transparency are priorities, such as healthcare, finance, and legislation. It will encourage the development of more interpretable and generalizable AI systems and will also address concerns about ethics and fairness arising from today's data-driven systems.

Acknowledgements

This work is supported by NSF Grants 1908711 and 1849107.

References

[1] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 39–48, 2016. 1

[2] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 7

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 5

[4] Hiteshwar Kumar Azad and Akshay Deepak. Query expansion techniques for information retrieval: a survey. *CoRR*, abs/1708.00247, 2017. 7

[5] Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. Mutan: Multimodal tucker fusion for visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2612–2620, 2017. 1

[6] Cameron B Browne, Edward Powley, Daniel Whitehouse, Simon M Lucas, Peter I Cowling, Philipp Rohlfshagen, Stephen Tavener, Diego Perez, Spyridon Samothrakis, and Simon Colton. A survey of monte carlo tree search methods. *IEEE Transactions on Computational Intelligence and AI in games*, 4(1):1–43, 2012. 4

[7] Qingxing Cao, Bailin Li, Xiaodan Liang, and Liang Lin. Explainable high-order visual question reasoning: A new benchmark and knowledge-routed network. *arXiv preprint arXiv:1909.10128*, 2019. 1, 2, 3

[8] Danqi Chen and Christopher D Manning. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 740–750, 2014. 7

[9] Christiane Fellbaum. Wordnet. In *Theory and applications of ontology: computer applications*, pages 231–243. Springer, 2010. 2, 3

[10] Chuang Gan, Yandong Li, Haoxiang Li, Chen Sun, and Boqing Gong. Vqs: Linking segmentations to questions and answers for supervised attention in vqa and question-focused semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1811–1820, 2017. 1

[11] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi 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. 1

[12] Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks*, 18(5-6):602–610, 2005. 2

[13] Parth Gupta, Kalika Bali, Rafael E Banchs, Monojit Choudhury, and Paolo Rosso. Query expansion for mixed-script information retrieval. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 677–686, 2014. 3

[14] Ronghang Hu, Jacob Andreas, Trevor Darrell, and Kate Saenko. Explainable neural computation via stack neural module networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 53–69, 2018. 2

[15] Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Kate Saenko. Learning to reason: End-to-end module networks for visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 804–813, 2017. 1, 2

[16] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6700–6709, 2019. 1, 5

[17] Drew A Hudson and Christopher D Manning. Learning by abstraction: The neural state machine. *arXiv preprint arXiv:1907.03950*, 2019. 1, 2, 3, 4, 5, 6, 7

[18] Ming Jiang, Shi Chen, Jinhui Yang, and Qi Zhao. Fantastic answers and where to find them: Immersive question-directed visual attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2980–2989, 2020. 5

[19] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2901–2910, 2017. 1

[20] Vijay R Konda and John N Tsitsiklis. Actor-critic algorithms. In *Advances in neural information processing systems*, pages 1008–1014. Citeseer, 2000. 4

[21] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017. 2, 3

[22] Guohao Li, Hang Su, and Wenwu Zhu. Incorporating external knowledge to answer open-domain visual questions with dynamic memory networks. *arXiv preprint arXiv:1712.00733*, 2017. 1

[23] Hugo Liu and Push Singh. Conceptnet—a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226, 2004. 2, 3

[24] Pankaj Malhotra, Anusha Ramakrishnan, Gaurangi Anand, Lovekesh Vig, Puneet Agarwal, and Gautam Shroff. Lstm-based encoder-decoder for multi-sensor anomaly detection. *arXiv preprint arXiv:1607.00148*, 2016. 4

[25] Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B Tenenbaum, and Jiajun Wu. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. *arXiv preprint arXiv:1904.12584*, 2019. 7

[26] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering

benchmark requiring external knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3195–3204, 2019. 1, 5, 6, 7

[27] Stefan Mathe, Aleksis Pirinen, and Cristian Sminchisescu. Reinforcement learning for visual object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2894–2902, 2016. 3

[28] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pages 1928–1937. PMLR, 2016. 4

[29] Medhini Narasimhan, Svetlana Lazebnik, and Alexander G Schwing. Out of the box: Reasoning with graph convolution nets for factual visual question answering. *arXiv preprint arXiv:1811.00538*, 2018. 6

[30] Zhou Ren, Xiaoyu Wang, Ning Zhang, Xutao Lv, and Li-Jia Li. Deep reinforcement learning-based image captioning with embedding reward. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 290–298, 2017. 3, 4

[31] Dwaipayan Roy, Debjyoti Paul, Mandar Mitra, and Utpal Garain. Using word embeddings for automatic query expansion. *arXiv preprint arXiv:1606.07608*, 2016. 3

[32] Adam Santoro, David Raposo, David GT Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. A simple neural network module for relational reasoning. *arXiv preprint arXiv:1706.01427*, 2017. 1

[33] Jiaxin Shi, Hanwang Zhang, and Juanzi Li. Explainable and explicit visual reasoning over scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8376–8384, 2019. 1, 2, 3, 4, 5, 6, 7

[34] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8317–8326, 2019. 2

[35] Richard S Sutton, David A McAllester, Satinder P Singh, Yishay Mansour, et al. Policy gradient methods for reinforcement learning with function approximation. In *NIPs*, volume 99, pages 1057–1063. Citeseer, 1999. 4

[36] Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. Learning to compose dynamic tree structures for visual contexts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6619–6628, 2019. 2

[37] Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. Fvqa: Fact-based visual question answering. *IEEE transactions on pattern analysis and machine intelligence*, 40(10):2413–2427, 2017. 5, 6, 7

[38] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992. 4

[39] Caiming Xiong, Stephen Merity, and Richard Socher. Dynamic memory networks for visual and textual question answering. In *International conference on machine learning*, pages 2397–2406. PMLR, 2016. 5

[40] Semih Yagcioglu, Erkut Erdem, Aykut Erdem, and Ruket Cakıcı. A distributed representation based query expansion approach for image captioning. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 106–111, 2015. 3

[41] Xu Yang, Hanwang Zhang, and Jianfei Cai. Learning to collocate neural modules for image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4250–4260, 2019. 1, 2, 3

[42] Keren Ye and Adriana Kovashka. Advise: Symbolism and external knowledge for decoding advertisements. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 837–855, 2018. 1

[43] Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Joshua B Tenenbaum. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. *arXiv preprint arXiv:1810.02338*, 2018. 1, 7

[44] Liyang Zhang, Shuaicheng Liu, Donghao Liu, Pengpeng Zeng, Xiangpeng Li, Jingkuan Song, and Lianli Gao. Rich visual knowledge-based augmentation network for visual question answering. *IEEE Transactions on Neural Networks and Learning Systems*, 2020. 6

[45] Yifeng Zhang, Ming Jiang, and Qi Zhao. Explicit knowledge incorporation for visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2021. 1, 2, 3, 6, 7

[46] Maryam Ziaeefard and Freddy Lecue. Towards knowledge-augmented visual question answering. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1863–1873, Barcelona, Spain (Online), Dec. 2020. International Committee on Computational Linguistics. 6, 7