VQA-GNN: Reasoning with Multimodal Knowledge via Graph Neural Networks for Visual Question Answering

Yanan Wang[1]* Michihiro Yasunaga[2] Hongyu Ren[2] Shinya Wada[1] Jure Leskovec[2]
[1]KDDI Research [2]Stanford University
[1]{wa-yanan,sh-wada}@kddi.com [2]{myasu,hyren,jure}@cs.stanford.edu

Abstract

*Visual question answering (VQA) requires systems to perform concept-level reasoning by unifying unstructured (*e.g.*, the context in question and answer; "QA context") and structured (*e.g.*, knowledge graph for the QA context and scene; "concept graph") multimodal knowledge. Existing works typically combine a scene graph and a concept graph of the scene by connecting corresponding visual nodes and concept nodes, then incorporate the QA context representation to perform question answering. However, these methods only perform a unidirectional fusion from unstructured knowledge to structured knowledge, limiting their potential to capture joint reasoning over the heterogeneous modalities of knowledge. To perform more expressive reasoning, we propose VQA-GNN, a new VQA model that performs* ***bidirectional*** *fusion between unstructured and structured multimodal knowledge to obtain unified knowledge representations. Specifically, we inter-connect the scene graph and the concept graph through a super node that represents the QA context, and introduce a new multimodal GNN technique to perform inter-modal message passing for reasoning that mitigates representational gaps between modalities. On two challenging VQA tasks (VCR and GQA), our method outperforms strong baseline VQA methods by* ***3.2%*** *on VCR (Q-AR) and* ***4.6%*** *on GQA, suggesting its strength in performing concept-level reasoning. Ablation studies further demonstrate the efficacy of the bidirectional fusion and multimodal GNN method in unifying unstructured and structured multimodal knowledge.*

1. Introduction

The visual question answering (VQA) task aims to provide answers to questions about a visual scene. It is crucial in many real-world tasks including scene understanding, autonomous vehicles, search engines, and recommendation systems [1, 2, 9, 15]. To solve VQA, systems need to perform concept-level reasoning by unifying unstructured (*e.g.*, the context in question and answer; "QA context") and structured (*e.g.*, knowledge graph for the QA context and scene; "concept graph") multimodal knowledge.

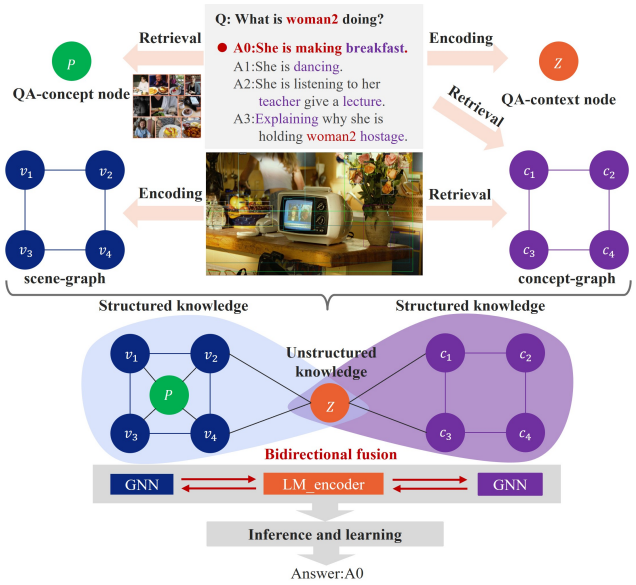


Figure 1. Overview of *VQA-GNN*. Given an image and QA sentence, we obtain unstructured knowledge (*e.g.*, QA-concept node p and QA-context node z) and structured knowledge (*e.g.*, scene-graph and concept-graph), and then unify them to perform bidirectional fusion for visual question answering.

Most of the high-performing VQA methods [6, 20, 26, 37, 49, 51, 52] pretrain a multimodal transformer model on a large-scale dataset to obtain unstructured multimodal knowledge from image and language contexts, and then finetune the pretrained model to reason on downstream tasks (*e.g.*, visual commonsense reasoning (VCR) task [50]). Existing methods (*e.g.*, SGEITL [42]) also incorporate structured knowledge into these transformer-based models by including a scene graph in the input of a pretrained multimodal transformer model. More recent meth-

* Work done while at Stanford University.

ods [27, 54] further combine the scene graph and the concept graph by inter-connecting corresponding visual nodes and concept nodes through graph neural networks (GNNs), and then incorporate the unstructured QA context representation to perform question answering. However, these methods only perform late fusion or unidirectional fusion from unstructured knowledge to structured knowledge and do not train the model to mutually aggregate information from both sides. This can limit their potential to perform joint reasoning over the heterogeneous modalities of knowledge. As unstructured knowledge and structured knowledge have complementary benefits—pretrained unstructured representations capture broader knowledge and structured representations offer scaffolds for reasoning— [48], this motivates the development of models that deeply fuse the two modalities of knowledge for visual question answering.

We propose VQA-GNN (Figure 1), a new visual question answering model performing bidirectional fusion between unstructured and structured multimodal knowledge to obtain a unified, more expressive knowledge representation. VQA-GNN extracts a scene graph from the given input image using an off-the-shelf scene graph generator [38] and then retrieves a relevant concept graph for the input image and QA context from a general knowledge graph like ConceptNet [36], obtaining a structured representation of the scene. Simultaneously, to obtain an unstructured knowledge representation for the scene, (1) we use pretrained RoBERTa [23] to encode the context in question and answer ("QA-context") as *QA-context node*, and (2) we retrieve relevant visual regions from a general scene graph VisualGenome [18] and take their mean pooled representation as a *QA-concept node*, which we connect to the scene graph. We then connect the scene graph and the concept graph through *QA-context node* to build a multimodal semantic graph.

To achieve bidirectional fusion across the multimodal semantic graph, we introduce a new multimodal GNN technique that performs inter-modal message passing. The multimodal GNN consists of two modality-specialized GNN modules, one for each modality, which perform inter-message aggregation between the *QA-context node* and nodes in structured graphs, aiming to reduce representational gaps between modalities. Meanwhile, by leveraging the robust transformer-based architecture of RoBERTa, we unfreeze and finetune the weights of the QA-context node to enable mutual information aggregation from modality-specialized GNN modules.

We evaluate VQA-GNN on two challenging VQA tasks, VCR [50] and GQA [13]. These tasks require systems to perform conceptual and compositional reasoning to answer diverse questions (*e.g.*, multiple-choice question answering and rationale selection in VCR; open-domain question answering in GQA). Our model outperforms strong baseline VQA methods [12, 42] by **3.2%** on VCR (Q-AR) and **4.6%** on GQA. Moreover, ablation studies show the efficacy of our two main techniques, bidirectional fusion and multimodal GNN message passing. On VCR, our multimodal GNN technique that reduces multimodal gaps outperforms existing works that use generic GNNs [27, 54] by **4.5%**. On GQA, bidirectional fusion outperforms a unidirectional fusion variant by **4%**. These results confirm the promise of VQA-GNN in unifying unstructured and structured multimodal knowledge for reasoning.

2. Problem Setup

This work focuses on multiple-choice and open-domain visual question answering, respectively. Each data point consists of an image c, and a natural language question q. For the multiple-choice setting, each question corresponds to a set of candidate answers $\mathcal{A}$, where only one candidate $a_{\text{correct}} \in \mathcal{A}$ is the correct answer to the question. Given a QA example $(c, q, \mathcal{A})$, we assume we have access to its relevant joint graph $\mathcal{G}^{(vcr)}$ and our goal is to identify the correct answer $a_{\text{correct}} \in \mathcal{A}$. For the open-domain setting, all questions correspond to a large set of common answer classes $\mathcal{B}$, where only one candidate $b_{\text{correct}} \in \mathcal{B}$ is the best answer to each question. Given a data example (c, q) with relevant scene graph $\mathcal{G}^{(gqa)}$, the goal is to identify $b_{\text{correct}} \in \mathcal{B}$.

3. Related Work

3.1. Multimodal transformer

VQA has emerged as one of the most popular topics in the computer vision community over the past few years [1, 2, 9, 11, 15, 25]. Existing methods for VQA [20, 26, 37, 52] employ the pretrain-and-finetune approach, where they train a multimodal transformer model on large-scale visual-language datasets, and then finetune the pretrained model on the downstream VQA datasets, *e.g.*, RESERVE-L model [51] is pretrained using 1 billion image-caption data including video frames, text, and audio. However, these methods only focus on obtaining unstructured multimodal representations by modeling implicit interactions over the visual and language domains. In contrast, our method introduces a multimodal GNN module to obtain unified knowledge representations from unstructured and structured multimodal knowledge based on explicit interactions over a well-structured multimodal semantic graph.

3.2. Structured knowledge-based VQA

Scene graph. Existing methods such as [49] introduce a scene graph prediction task to learn structured knowledge conditioned multimodal representations, and the work [42] proposes to incorporate extracted scene graph in multimodal transformer models. These works [12, 21, 30, 39] also

exploit GNNs [4,17,40,44] to incorporate unstructured QA-context knowledge into a structured scene graph for question answering. However, these methods only perform late fusion or unidirectional fusion from unstructured knowledge to structured knowledge. In contrast, our method performs bidirectional fusion to unify unstructured and structured knowledge.

Concept graph. Aiming to achieve concept-level reasoning beyond image-level recognition for visual understanding, existing works [5,7,10,19,27,28,34,35,43,46,54,55] utilize knowledge graphs (KGs) to explore how to unify commonsense knowledge [33,47,48] about background concepts of the scene. The work [45] converts the image into captions and performs GPT-3 [3] in joint knowledge retrieval and reasoning. The work [19] encodes question-related knowledge from the retrieved knowledge facts to a knowledge-aware question representation, and then performs a question and knowledge-guided graph attention operation for answer reasoning. However, structured concept knowledge relevant to the QA context is not enough to represent the background scene. We build a concept graph to cover structured and unstructured concept knowledge relevant to the QA context as well as the background scene.

Scene graph & concept graph. To enrich structured knowledge, these works [10, 43, 55] utilize GNNs to learn graph representations of the scene graph and concept graph respectively, and then perform later fusion across the QA context, scene graph and concept graph for question reasoning. However, it is insufficient to capture the interactions across different modalities for concept-level reasoning. These works [27,54] unify the scene graph and concept graph by interconnecting corresponding visual and concept nodes to capture their interactions. However, the representational gap between modalities adversely affects the performance of inter-modal message passing for capturing joint reasoning [22, 41]. Our method inter-connects the scene graph and concept graph via a QA context node and introduces a new multimodal GNN technique to mitigate representational gaps between modalities.

4. Methodology

As shown in Figure 2, given an image and its related question with an answer choice, first we build a multimodal semantic graph to unify unstructured and structured multimodal knowledge into a joint graph (§4.1). Then we propose a multimodal GNN-based bidirectional fusion method that performs inter-modal message passing to obtain node representations enhanced with unstructured and structured multimodal knowledge (§4.2). Finally, we get the pooled representations of scene-graph and concept-graph and concatenate them with the representations from the QA-context node and QA-concept node for answer prediction (§4.3).

4.1. Multimodal semantic graph

Scene-graph encoding. Given an image, we use a pretrained scene graph generator to extract a scene graph that consists of recall@20 of $(subject, predicate, object)$ triplets to represent structured image context [38], *e.g.*, $(car, behind, man)$. Then we apply a pretrained object detection model for embedding a set of scene graph nodes $\mathcal{V}^{(s)} = \{v_i\}_{i=1}^{N}$ (N indicates the maximum number of scene-graph nodes of "20") and represent $v_i^{(s)}$ with a 2048 dimensional visual feature vector [53]. We indicate the predicate edge types in the scene graph with a set of scene graph edges $\mathcal{E}^{(s)} = \{r_i^{(s)}\}_{i=1}^{D}$ (D denotes the number of edge types) and represent $r_j^{(s)}$ with a D-dimensional one-hot vector.

QA-concept node retrieval. In addition to the local image context, with an assumption that the global image context of the correct choice aligns with the local image context, we employ a pretrained sentence-BERT model to calculate the similarity between each answer choice and all descriptions of the region image within the VisualGenome dataset [18]. This process allows us to extract relevant region images that capture the global image context associated with each choice [32]. We retrieve the top 10 results and utilize the same object detector to embed them. These embeddings are averaged to obtain a QA-concept node denoted as p. Subsequently, we introduce a QA-concept edge, denoted as $r^{(p)}$, which serves to fully connect node p with node v_i.

Concept-graph retrieval. We retrieve a concept graph from the image and ConceptNet KG, a general-domain knowledge graph [36]. Our process is illustrated in Figure 3. In **Step 1**, we extract concept entities from both the image and the answer choices. Specifically, for the image, we consider the detected object names as potential contextual entities, while excluding general terms like "person" to streamline the reasoning process. For the answer choice, we ground phases if they are mentioned concepts in the ConceptNet KG, *e.g.*, "beverage" and "shop". In **Step 2-1**, we use grounded phases to retrieve their 1-hop neighbor nodes from the ConceptNet KG. In **Step 2-2**, since many concept nodes retrieved are semantically irrelevant to the answer choice, we use a word2vec model released by the spaCy library[1] to get relevance score between concept node candidates and answer choices, and prune irrelevance nodes when the relevance score is less than 0.6. As a result, given an answer choice, we can retrieve a relevance subgraph from ConceptNet KG based on the relevance score. In **Step 3**, to better comprehend concept knowledge from the image as well, in addition to linking adjacent object entities in the ConceptNet KG domain, we also combine parsed local concept entities of the image with the retrieved subgraph. For instance, considering that ConceptNet en-

[1] https://spacy.io/

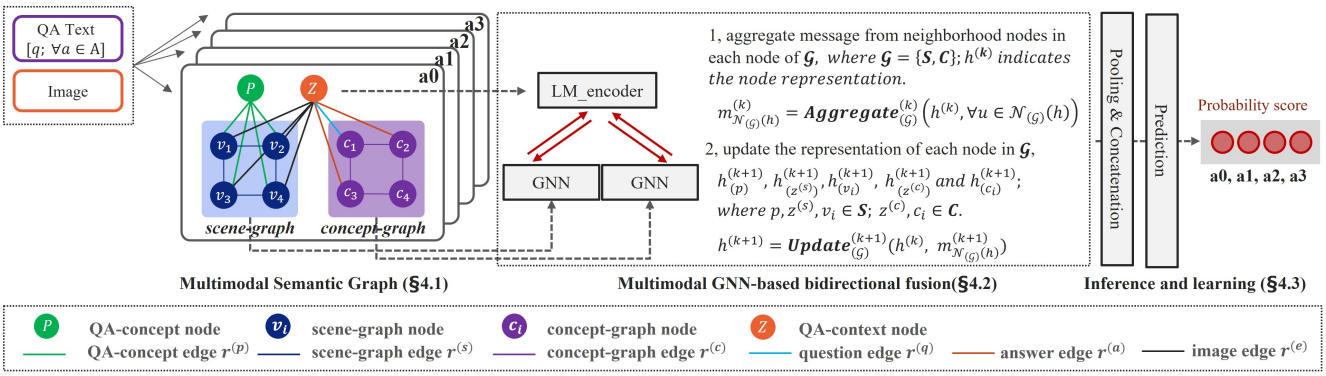


Figure 2. Reasoning procedure of VQA-GNN. We first build a multimodal semantic graph for each given image-QA pair to unify unstructured (*e.g.*, "node p" and "node z") and structured (*e.g.*, "scene-graph" and "concept-graph") multimodal knowledge (§4.1). Then we perform inter-modal message passing with a multimodal GNN-based bidirectional fusion method (§4.2) to update the representations of node z, p, v_i and c_i for $k+1$ iterations in two steps. Finally, we predict the answer with these updated various node representations (§4.3). Here, "S" and "C" indicate scene-graph and concept-graph respectively. "LM_encoder" indicates a language model used to finetune QA-context node representation, and "GNN" indicates a relation-graph neural network for iterative message passing.

compasses various types of local concept entities, if a local concept entity (*e.g.*, "bottle") is found adjacent to a retrieved entity (*e.g.*, "beverage"), we build a new knowledge triple, *e.g.*, (bottle, aclocation, beverage). Finally, we can construct a concept graph to depict the structured knowledge at the concept level. We obtain a collection of concept-graph nodes denoted as $\mathcal{V}^{(c)} = \{c_i\}_{i=1}^{N}$, where N represents the maximum number of concept-graph nodes of 60. The concept entity c_i is represented using a 1024-dimensional text feature vector as the concept entity embedding in [8]. Additionally, we initialize a set of concept-graph edges denoted as $\mathcal{E}^{(c)} = \{r_i^{(c)}\}_{i=1}^{D}$, using D-dimensional one-hot vectors, where D is the number of edge types in concept-graph.

QA-context node encoding. To construct a multimodal semantic graph, we introduce an unstructured QA-context node denoted as z to inter-connect the scene-graph and concept-graph using three additional relation types: the question edge $r^{(q)}$, the answer edge $r^{(a)}$, and the image edge $r^{(e)}$. The image edge $r^{(e)}$ fully links node z with $\mathcal{V}^{(s)}$, capturing the relationship between the QA context and relevant entities within the scene-graph. The question edge $r^{(q)}$ and answer edge $r^{(a)}$ link node z with the entities extracted from the question and the answer text, respectively, capturing the relationship between the QA context and the relevant entities within the concept-graph. As a result, we construct a multimodal semantic graph $\mathcal{G} = \{S, C\}$ to provide a joint reasoning space, which includes two sub-graphs of scene-graph S and concept-graph C, two super nodes of QA-concept node and QA-context node. Here, the QA-concept node is included in S and the QA-context is included in S and C for performing inter-modal message passing in §4.2. Especially, the QA-context node z is assigned to not only learn unstructured discriminative representations by giving

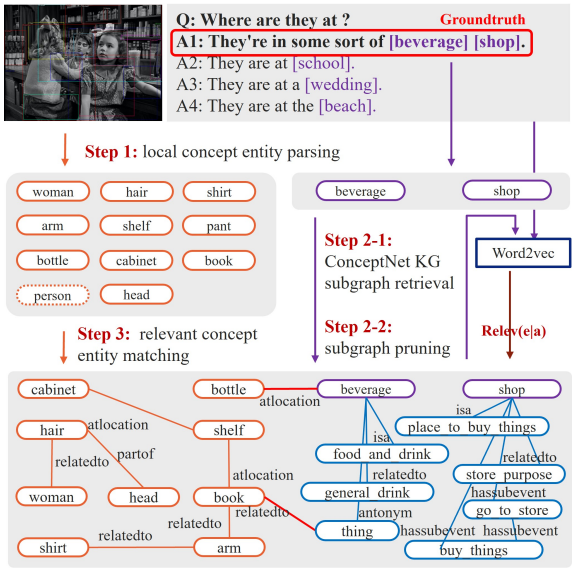


Figure 3. The process of concept-graph retrieval involves the calculation of similarity between concept-graph nodes and the answer context, denoted as $Relev(e|a)$.

a Q and A text pairs but also to incorporate structured multimodal knowledge from scene-graph and concept-graph for effective VQA. As the transformer-based method is powerful for multimodal representation learning [26, 37], we employ the RoBERTa LM [24] as the encoder of QA-context node z and finetune it with GNN modules to achieve bidirectional multimodal knowledge fusion (see Figure 2).

4.2. Multimodal GNN-based bidirectional fusion

To improve inter-modal message passing by avoiding directly aggregating neighborhood nodes that may be ini-

tialized in different modality domains, we propose a multimodal GNN-based bidirectional fusion method built by two relation-graph neural networks for scene-graph and concept-graph respectively (see §5.2.1). The relation-graph neural network is built on the Graph Attention Networks (GAT) [40] by introducing multi-relation aware message for attention-based message aggregation process to better capture multiple relation information.

The detail of the relation-graph neural network is as follows: we have five node types: $\mathcal{T} = \{\boldsymbol{Z}, \boldsymbol{P}, \boldsymbol{S}, \boldsymbol{C}\}$ in the multimodal semantic graph and they indicate QA-context node z, QA-concept node p, scene-graph node s, question node q and concept-graph node c. As relation edge representation $\boldsymbol{r}_{i,j}$ should capture relationship from node i to node j and difference of node types represents a special relation between neighborhood nodes, we first obtain node type embedding u_i, u_j and then concatenate them with edge embedding e_{ij} to generate multi-relation embedding $\boldsymbol{r}_{ij}$ from i to j by

$$\boldsymbol{r}_{ij} = f_r([e_{ij}||u_i||u_j]) \tag{1}$$

where $u_i, u_j \in \{0,1\}^{|\mathcal{T}|}$ are one-hot vectors indicating the node types of i and j, $e_{ij} \in \{0,1\}^{|\mathcal{R}|}$ is a one-hot vector indicating relation type of edge (i,j). $||$ is the concatenation operation, and $f_r : \mathbb{R}^{|\mathcal{R}|+2|\mathcal{T}|} \rightarrow \mathbb{R}^{\mathcal{D}}$ is a 2-layer MLP. Based on multi-relation embedding $\boldsymbol{r}_{ij}$, the multi-relation aware message $\boldsymbol{m}_{ij}$ from i to j is computed by

$$\boldsymbol{m}_{ij} = f_m([\boldsymbol{h}_i^{(k+1)}||\boldsymbol{r}_{ij}]) \tag{2}$$

where $f_m : \mathbb{R}^{2\mathcal{D}} \rightarrow \mathbb{R}^{\mathcal{D}}$ is a linear transformation. $h_i^{(k+1)}$ is the node representation of each node i in the graph. We then recursively updated it $k+1$ times by

$$\boldsymbol{h}_i^{(k+1)} = f_h\left(\sum_{j\in\mathcal{N}_i} \alpha_{ij}\boldsymbol{m}_{ij}\right) + \boldsymbol{h}_i^{(k)} \tag{3}$$

where $f_h : \mathbb{R}^{\mathcal{D}} \rightarrow \mathbb{R}^{\mathcal{D}}$ is 2-layer MLP with batch normalization [14]. $\mathcal{N}_i$ indicates the neighborhood of node i, α_{ij} is an attention weight to emphasize important messages passed from $\mathcal{N}_i$ to node i. We obtain $\boldsymbol{q}_i, \boldsymbol{k}_j$ by

$$\boldsymbol{q}_i = f_q(\boldsymbol{h}_i^{(k+1)}), \boldsymbol{k}_j = f_k([\boldsymbol{h}_j^{(k+1)}||\boldsymbol{r}_{ij}]) \tag{4}$$

where $f_q : \mathbb{R}^{\mathcal{D}} \rightarrow \mathbb{R}^{\mathcal{D}}$ and $f_k : \mathbb{R}^{2\mathcal{D}} \rightarrow \mathbb{R}^{\mathcal{D}}$ are linear transformations. α_{ij} is computed using the softmax function by

$$\gamma_{ij} = \frac{\boldsymbol{q}_i^T\boldsymbol{k}_j}{\sqrt{D}}, \tag{5}$$

$$\alpha_{ij} = \text{softmax}_j(\gamma_{ij}) = \frac{\exp(\gamma_{ij})}{\sum_{j'\in\mathcal{N}_i}\exp(\gamma_{ij'})} \tag{6}$$

By referring to Eq. 3, we perform message passing to update node representations in each graph in parallel by aggregating multi-relation aware messages from neighborhood nodes in each node. As a result, we obtain structured graph node representations $h_{(v_i)}^{(k+1)}$ and $h_{(c_i)}^{(k+1)}$, unstructured node representations $h_{(p)}^{(k+1)}$ and $h_{(z)}^{(k+1)}$. For node z, we update it with scene-graph and concept-graph respectively, and concatenated by

$$\boldsymbol{h}_{(z)}^{(k+1)} = f_z([\boldsymbol{h}_{(z^{(s)})}^{(k+1)}||\boldsymbol{h}_{(z^{(c)})}^{(k+1)}]) \tag{7}$$

where $f_z : \mathbb{R}^{2\mathcal{D}} \rightarrow \mathbb{R}^{\mathcal{D}}$ is a linear transformation.

4.3. Inference and Learning

To identify the correct answer $a_{correct} \in \mathcal{A}$ with a QA example $(c, q, \mathcal{A})$, we compute the probability $p(a|c,q)$ for each answer choice with its multimodal semantic knowledge from scene-graph, concept-graph, QA-context node, and QA-concept node. With various node representations on the *L-th* ($L = k+1$) layer updated by GNN modules (shown in Figure 2), we obtain pooling representations $\boldsymbol{h}_{(s)}^{(K+1)}$ and $\boldsymbol{h}_{(c)}^{(K+1)}$ of scene-graph and concept-graph and then concatenate with QA-context node and QA-concept node representations. Finally we calculate $p(a|c,q)$ by

$$\boldsymbol{h}_a^{(k+1)} = [\boldsymbol{h}_{(s)}^{(K+1)}||\boldsymbol{h}_{(c)}^{(K+1)}||\boldsymbol{h}_{(p)}^{(K+1)}||\boldsymbol{h}_{(z)}^{(K+1)}], \tag{8}$$

$$\text{logit}(a) = f_c(\boldsymbol{h}_a^{(k+1)}), \tag{9}$$

$$p(a|c,q) = \text{softmax}_a(\text{logit}(a)) \tag{10}$$

where $\text{logit}(a)$ indicates the confident score of answer choice a, $f_c : \mathbb{R}^{4\mathcal{D}} \rightarrow \mathbb{R}^1$ is a linear transformation that maps the concatenation of representations to a scale. We normalize it across all answer choices using the softmax function. For the training process, we apply the cross entropy loss to optimize the *VQA-GNN* model end-to-end.

5. Experiments

5.1. Experiment Setup

Visual Commonsense Reasoning (VCR). We evaluate VQA-GNN on VCR [50]. It contains 290k pairs of questions, answers, and rationales, over 110k unique movie scenes. VCR consists of two tasks: visual question answering (Q→A), answer justification (QA→R). Each question in the dataset is provided with four candidate answers. The goal of (Q→A) is to select the best answer, while the goal of (QA→R) is to justify the given question answer pair by picking the best rationale out of the four candidates. We joint train *VQA-GNN* on Q→A and QA→R, with a common LM encoder, the multimodal semantic graph for Q→A, a concept graph retrieved by giving question-answer pair with a rationale candidate for QA→R. We use a pretrained RoBERTa Large model to embed the QA-context node, and finetune it with the multimodal GNN for 50 epoch by using learning rates 1e-5 and 1e-4 respectively. We set the number of layers ($L = 5$) of VQA-GNN and use AdamW [16]

Model	# Image-caption in pretraining	Parameters	Structured knowledge	Test Acc.(%)		
				Q→A	QA→R	Q→AR
ViLBERT [26]	3.3M	**221M**	No	73.3	74.6	54.8
VLBERT-L [37]	3.3M	383M	No	75.8	78.4	59.7
SGEITL+VLBERT [42]	290k	$\geq$ 383M	Yes	76.0	78.0	59.6
UNITER-(B/L) [6]	9.5M	154M/378M	No	75.0/77.3	77.2/80.8	58.2/62.8
ERNIE-ViL-(B/L) [49]	3.8M	212M/533M	No	77.0/79.2	80.3/83.5	62.1/66.3
VQA-GNN (Ours)	**290k**	372M	Yes	77.9	80.0	62.8
MERLOT [52]	180M	223M	No	80.6	80.4	65.1
RESERVE-(B/L) [51]	1B	200M/644M	No	79.3/84.0	78.7/84.9	62.6/72.0
RESERVE-L + **VQA-GNN (Ours)**	1B	1B	Yes	**85.3**	**86.9**	**74.3**

Table 1. Accuracy scores for VCR test set. *VQA-GNN* outperforms *SGEITL+VLBERT* model on Q→AR metric by **3.2%**, and achieves competitive accuracy with SOTA methods, which have a close number of parameters but SOTA methods require a large amount of image caption data in pre-training process (over 13x larger than our model), *e.g.*, *"UNITER-L"*, *"ERNIE-ViL-B"*, *"RESERVE-B"*. Moreover, *"RESERVE-L+VQA-GNN"* outperforms *RESERVE-L* by **2.3%** on Q→AR metric.

optimizer to minimize the loss. We use a linear warmup of the learning rate over the *15-th* epoch, with a cosine decay thereafter to *0*.

GQA dataset. It contains open-ended questions (1.5M questions correspond to 1,842 answer tokens), along with 110K scene graphs and the semantic functional programs to offer unambiguous instructions [13]. We only use questions without giving a semantic feature program that limits the development of the model's reasoning abilities in a more practical setting. We define the question as the context node (node q) to fully connect visual and textual scene graphs (SG) respectively to structure multimodal semantic graphs. The node q is embedded with a pretrained RoBERTa large model, and we initialize object nodes' representations in visual SG using official object features, object nodes in textual SG by concatenating GloVe [31] based word embedding of the object name and attributes. Different from the training target of VCR, the goal of GQA is to classify the given image-question pair out of 1,842 answer classes. We finetune the node q with VQA-GNN for 50 epoch by using learning rates 2e-5 and 2e-4 respectively.

5.2. Performance

5.2.1 Evaluation on VCR dataset

Comparison with state-of-the-art methods. We compared *VQA-GNN* with state-of-the-art methods on the VCR test set in Table 1. Compared with the unidirectional fusion method *SGEITL+VLBERT* that can boost multimodal transformer model VLBERT by incorporating visual scene graphs, *VQA-GNN* is a multimodal GNN-based bidirectional fusion method built on the multimodal semantic graph. Both were not pretrained on the large-scale dataset. *VQA-GNN* improves SGEITL+VLBERT on the Q→AR metric by **3.2%**, and further reduces over **11M** training parameters. We think that the structured multimodal semantic graph provides much more commonsense knowledge related to QA and original image than SGEITL, and the multimodal GNN-based bidirectional fusion method works much better on unifying unstructured and structured multimodal knowledge than multimodal transformer models. Moreover, since we retrieve commonsense knowledge from structured multimodal semantic graphs directly, *VQA-GNN* is a cost-effective approach compared to multimodal transformer models that consume much GPU resources to learn commonsense knowledge with large parameters.

We also demonstrate the effectiveness of *VQA-GNN* by comparing it with state-of-the-art multimodal transformer models that were pretrained across text and images and were finetuned on the VCR dataset. As shown in Table 1, the larger image caption data and parameters, the higher performance the model can achieve. In contrast, *VQA-GNN* trained with VCR dataset with $290K$ image-caption pairs performs similarly to UNITER-L that requires over **32x** larger image-caption data than us in pretraining process. These results suggest that *VQA-GNN* obtaining structured context knowledge inferred from image-level and concept-level knowledge sources is as effective as the pretraining process for previous methods. Moreover, *VQA-GNN* can further enhance RESERVE-L performance on both Q→A and QA→R, and finally improves the score by **2.3%** on Q→AR metric. As correcting some questions requires the model to understand commonsense knowledge related to image context and have good reasoning ability, it is difficult for multimodal transformer methods including RESERVE-L. On the other hand, *VQA-GNN* not only structures a joint semantic graph to provide commonsense knowledge related to image context but also has a good reasoning ability thanks to its multimodal GNN module. Additionally, in the supplementary material, we detail the results compared

to baselines pretrained only with the VCR dataset, as well as the evaluation of different question types.

Effectiveness of the multimodal semantic graph. To further study the behavior of modules in the multimodal semantic graph, and quantitatively evaluate pretrained models used in this work (*e.g.*, RoBERTa-L, scene-graph[scene graph generator], concept-graph[conceptNet KG]), we report the performance of using different node representations in Table 2. We respectively build classification models by applying Node p and Node z to get their validation accuracy on Q→A subtask. The scene-graph structured by connecting Node p and Node z with extracted visual scene graph improves over 25% on average of these two nodes. In terms of concept-graph, it is structured by connecting Node z with retrieved conceptual triplets from ConcepNet KG, improving Node z's performance by 15.2%. We further compare *VQA-GNN* on "scene-graph + concept-graph" w/ and w/o Node p, and the result shows that including Node p can further improve the performance by 2%. We believe that the Node p representing global visual knowledge associated with the correct answer is able to pass visual commonsense knowledge to the multimodal semantic graph, and it is effective besides employing ConcepNet KG to obtain textual commonsense knowledge [48].

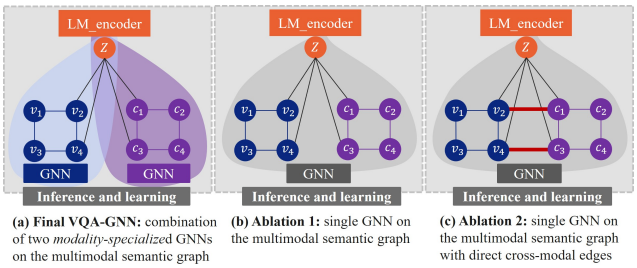
Model	Val Acc.(%) (Q→A)
Node p (Vinvl)	43.5
Node z (RoBERTa-L)	53.8
concept-graph	69.0
scene-graph	73.7
concept-graph + scene-graph (w/o node p)	75.1
concept-graph + scene-graph (w/ node p)	**77.1**

Table 2. All modules in the multimodal semantic graph help boost the final performance. Here, "scene-graph" includes node z and node p, "concept-graph" includes node z.

Model	Val Acc.(%) (Q→A)
Ablation 1 (single GNN)	73.0
Ablation 2 (single GNN w/ cross-modal edges)	70.6
VQA-GNN (two modality-specialized GNNs)	**75.1**

Table 3. Ablation 1 and Ablation 2 indicate a single GNN on the multimodal semantic graph w/o and w/ direct cross-modal edges, respectively (Figure 4). VQA-GNN with two modality-specialized GNNs on the multimodal semantic graph achieves the best score.

Analysis of the multimodal GNN method. To analyze the effect of the multimodal GNN method on mitigating the multimodal gap in performing inter-modal message passing, we compared the final VQA-GNN with two single GNNs built on multimodal semantic graphs with and without direct cross-modal edges in Figure 4. As the results of VCR validation set shown in Table 3, the final VQA-GNN



(a) Final VQA-GNN: combination of two *modality-specialized* GNNs on the multimodal semantic graph

(b) Ablation 1: single GNN on the multimodal semantic graph

(c) Ablation 2: single GNN on the multimodal semantic graph with direct cross-modal edges

Figure 4. Ablation architectures. We find that our final VQA-GNN architecture with two modality-specialized GNNs overcomes the representation gaps between modalities (§5.2.1).

built with the multimodal GNN on the multimodal semantic graph improves the accuracy of both ablative architecture by over **2%**. We believe that the multimodal GNN built by two modality-specific GNNs can effectively avoid directly aggregating nodes from scene-graph and concept-graph to alleviate the modality gap. As a result, the inter-modal message passing can be improved. We further explored the aggregation process for some node samples to demonstrate why the two ablation architectures fail to alleviate the modality gap. Here, $m^{(k)}_{\mathcal{N}(u)}$ represents the aggregated messages from all neighbors of node u at the k-th iteration.

$$m^{(k)}_{\mathcal{N}(u)} = \text{Aggregate}^{(k)}(u^{(k)}, \forall v \in \mathcal{N}(u)) \quad (11)$$

where $\mathcal{N}(u)$ denotes a set of neighborhood nodes of the node u, and k denotes the iterations of $m^{(k)}_{\mathcal{N}(u)}$.

For (c) Ablation 2 in Figure 4, we assume that node v_2 is connected with node c_1 as both represent the same notion. However, their feature vectors are distributed in different modality domains and affect the aggregation process. We show the neighborhood nodes of QA-context node z, visual node v_2 and concept node c_1 are follows:

$$\mathcal{N}(z) = \{v_2, v_4, c_1, c_3\} \quad (12)$$

$$\mathcal{N}(v_2) = \{z, v_1, v_4, c_1\}; \mathcal{N}(c_1) = \{z, c_2, c_3, v_2\} \quad (13)$$

where their neighborhood nodes include heterogeneous nodes from different modality domains.

For (b) Ablation 1 in Figure 4, the neighborhood nodes of QA-context node z, visual node v_2 and concept node c_1 are follows:

$$\mathcal{N}(z) = \{v_2, v_4, c_1, c_3\} \quad (14)$$

$$\mathcal{N}(v_2) = \{z, v_1, v_4\}; \mathcal{N}(c_1) = \{z, c_2, c_3\} \quad (15)$$

Compared with (c) Ablation 2, node c_1 and node v_2 are removed from the neighborhood nodes of v_2 and c_1 which helped improve the performance of (c) Ablation 2 by 2.4%. However, it is limited by the QA-context node z that aggregates messages across scene-graph and concept-graph.

Although QA-context node z is a pretrained LM that can be finetuned on multimodal domains, it is more difficult to adapt to two modalities (Eq. 14) than to a single modality (Eq. 16). In contrast, the multimodal GNN method is designed by introducing two GNNs for each modality. We perform aggregation for QA-context node z for each modality so that the pretrained LM is finetuned on a single modality to alleviate the modality gap. The neighborhood nodes of QA-context node z, visual node v_2 and concept node c_1 are follows:

$$\mathcal{N}(z)^{(m1)} = \{v_2, v_4\}; \mathcal{N}(z)^{(m2)} = \{c_1, c_3\} \quad (16)$$

$$\mathcal{N}(v_2) = \{z^{(m1)}, v_1, v_4\}; \mathcal{N}(c_1) = \{z^{(m2)}, c_2, c_3\} \quad (17)$$

where $m1$ and $m2$ indicate two message passing methods for each modality.

5.2.2 Evaluation on GQA dataset

Comparison with baselines. We also compared *VQA-GNN* with baseline models on GQA dataset, under the realistic setup of not using the annotated semantic functional programs (see §5.1). As the results shown in Table 4, our model achieves validation accuracy of 58.9% for visual SG and 87.9% for textual SG. Compared with SGEITL [42] and GCN [21] which are unidirectional fusion methods, our method performs bidirectional fusion to unify unstructured and structured knowledge, and improved the reasoning ability of SGEITL by **5.6%** and GCN by **2.2%**. Moreover, by inter-connecting the visual and textual SG, our method achieves validation accuracy of **90.3%** and further suggests its efficacy in performing inter-modal message passing.

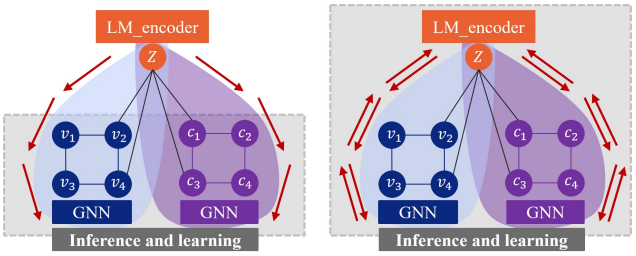
Model	Visual SG	Textual SG	Val Acc.(%)
SGEITL [42]	✓		53.3
CFR [29]	✓	✓	73.6
GCN [21]		✓	85.7
VQA-GNN	✓		58.9
		✓	87.9
	✓	✓	**90.3**

Table 4. Accuracy scores on the GQA validation set. All models are trained under the realistic setup of not using the annotated semantic functional programs.

Method	Val Acc.(%) ↑	Inference time (ms) ↓
Average pooling	62.3 (±0.40)	**5.2**
Unidirectional fusion	86.3 (±0.01)	8.6
Bidirectional fusion **(ours)**	**90.3** (±0.03)	5.5

Table 5. Ablation results on the effect of our proposed bidirectional fusion for GQA.

Ablation study on the bidirectional fusion. To fairly study the effect of bidirectional fusion for improving concept-level reasoning, we evaluated the performance of *VQA-GNN* with and without structured multimodal knowledge-enhanced question representations. We show their difference in Figure 5, compared with the unidirectional fusion, the bidirectional fusion approach is able to utilize the message aggregated from scene-graph and concept-graph in node z to predict the correct answer. It facilitates the joint reasoning ability of VQA-GNN in capturing bidirectional interactions between unstructured node z and structured multimodal semantic graph. As a result in Table 5, the bidirectional fusion approach further improved the performance of the unidirectional fusion approach by **4%**. We also compared our approach with an average pooling method that simply averages all node representations. We indeed find that this ablation performs significantly worse than others, which suggests that our approach can capture special relationship information between different nodes but average pooling cannot.



(a) Unidirectional fusion: prediction with scene-graph and concept-graph

(b) Bidirectional fusion: prediction with the multimodal semantic graph

Figure 5. Illustration of two knowledge fusion methods: our proposed bidirectional fusion v.s. the unidirectional fusion baseline.

6. Conclusion

We proposed a novel visual question answering method, *VQA-GNN*, which unifies unstructured and structured multimodal knowledge to perform joint reasoning of the scene. In the evaluation of two challenging VQA tasks (VCR and GQA), our method substantially outperforms existing models without pretraining using massive image-caption data under the same training setting, our method outperforms strong baseline VQA methods by **3.2%** on VCR (Q-AR) and **4.6%** on GQA, suggesting its strength in performing concept-level reasoning. Ablation studies further demonstrate the efficacy of the bidirectional fusion and multimodal GNN method in unifying unstructured and structured multimodal knowledge. In the next step, we will extend our work to the video domain and focus on obtaining temporal semantic knowledge to enhance the machine's reasoning ability.

Acknowledgment

We thank Rok Sosic, members of the Stanford SNAP group, as well as our anonymous reviewers for valuable feedback. We gratefully acknowledge the support of DARPA under Nos. HR00112190039 (TAMI), N660011924033 (MCS); Funai Foundation Fellowship; Microsoft Research PhD Fellowship; Masason Foundation Fellowship; Apple PhD Fellowship; ARO under Nos. W911NF-16-1-0342 (MURI), W911NF-16-1-0171 (DURIP); NSF under Nos. OAC-1835598 (CINES), OAC-1934578 (HDR), CCF-1918940 (Expeditions), IIS-2030477 (RAPID), NIH under No. R56LM013365; Stanford Data Science Initiative, Wu Tsai Neurosciences Institute, Chan Zuckerberg Biohub, Amazon, JPMorgan Chase, Docomo, Hitachi, Intel, KDDI, Toshiba, NEC, Juniper, and UnitedHealth Group.

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *ICCV*, 2015. 1, 2

[2] Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. Mutan: Multimodal tucker fusion for visual question answering. In *ICCV*, 2017. 1, 2

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. 3

[4] Dan Busbridge, Dane Sherburn, Pietro Cavallo, and Nils Y Hammerla. Relational graph attention networks. 2019. 3

[5] Qingxing Cao, Bailin Li, Xiaodan Liang, Keze Wang, and Liang Lin. Knowledge-routed visual question reasoning: Challenges for deep representation embedding. *IEEE TNNLS*, 33(7):2758–2767, 2022. 3

[6] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *ECCV*, 2020. 1, 6

[7] Yang Ding, Jing Yu, Bang Liu, Yue Hu, Mingxin Cui, and Qi Wu. Mukea: Multimodal knowledge extraction and accumulation for knowledge-based visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5089–5098, June 2022. 3

[8] Yanlin Feng, Xinyue Chen, Bill Yuchen Lin, Peifeng Wang, Jun Yan, and Xiang Ren. Scalable multi-hop relational reasoning for knowledge-aware question answering. In *EMNLP*, 2020. 4

[9] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *EMNLP*, 2016. 1, 2

[10] François Gardères, Maryam Ziaeefard, Baptiste Abeloos, and Freddy Lecue. ConceptBert: Concept-aware representation for visual question answering. In *EMNLP*, 2020. 3

[11] Xin Hong, Yanyan Lan, Liang Pang, Jiafeng Guo, and Xueqi Cheng. Transformation driven visual reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6903–6912, June 2021. 2

[12] Ronghang Hu, Anna Rohrbach, Trevor Darrell, and Kate Saenko. Language-conditioned graph networks for relational reasoning. In *ICCV*, 2019. 2

[13] 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 (CVPR)*, June 2019. 2, 6

[14] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*. JMLR.org, 2015. 5

[15] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. In *NeurIPS*, 2018. 1, 2

[16] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015. 5

[17] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016. 3

[18] 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. Springer, 2017. 2, 3

[19] Mingxiao Li and Marie-Francine Moens. Dynamic key-value memory enhanced multi-step graph reasoning for knowledge-based visual question answering. In *AAAI*, 2022. 3

[20] Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, 2020. 1, 2

[21] Weixin Liang, Yanhao Jiang, and Zixuan Liu. Graghvqa: Language-guided graph neural networks for graph-based visual question answering. *ArXiv*, abs/2104.10283, 2021. 2, 8

[22] Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *arXiv preprint arXiv:2203.02053*, 2022. 3

[23] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 2

[24] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. volume abs/1907.11692, 2019. 4

[25] Chao Lou, Wenjuan Han, Yuhuan Lin, and Zilong Zheng. Unsupervised vision-language parsing: Seamlessly bridging visual scene graphs with language structures via dependency relationships. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15607–15616, June 2022. 2

[26] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, 2019. 1, 2, 4, 6

[27] Kenneth Marino, Xinlei Chen, Devi Parikh, Abhinav Gupta, and Marcus Rohrbach. Krisp: Integrating implicit and symbolic knowledge for open-domain knowledge-based vqa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14111–14121, June 2021. 2, 3

[28] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *CVPR*, 2019. 3

[29] Binh X Nguyen, Tuong Do, Huy Tran, Erman Tjiputra, Quang D Tran, and Anh Nguyen. Coarse-to-fine reasoning for visual question answering. In *CVPR*, 2022. 8

[30] Will Norcliffe-Brown, Efstathios Vafeias, and Sarah Parisot. Learning conditioned graph structures for interpretable visual question answering. *arXiv preprint arXiv:1806.07243*, 2018. 2

[31] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *EMNLP*, 2014. 6

[32] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP*, 2019. 3

[33] Hongyu Ren, Hanjun Dai, Bo Dai, Xinyun Chen, Michihiro Yasunaga, Haitian Sun, Dale Schuurmans, Jure Leskovec, and Denny Zhou. Lego: Latent execution-guided reasoning for multi-hop question answering on knowledge graphs. In *ICML*, 2021. 3

[34] Parth Shah, Prateek Shenoy, Shivansh Bhattad, Prathamesh Bhingardive, Abir Chakraborty, and Subbarao Kambhampati. Kvqa: Knowledge-aware visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2019. 3

[35] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, and Devi Parikh. From strings to things: Knowledge-enabled vqa model that can read and reason. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019. 3

[36] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. 2017. 2, 3

[37] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. In *ICLR*, 2020. 1, 2, 4, 6

[38] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *CVPR*, 2020. 2, 3

[39] Damien Teney, Lingqiao Liu, and Anton van Den Hengel. Graph-structured representations for visual question answering. In *CVPR*, 2017. 2

[40] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph Attention Networks. In *ICLR*, 2018. 3, 5

[41] Yanan Wang, Jianming Wu, Kazuaki Furumai, Shinya Wada, and Satoshi Kurihara. Vae-based adversarial multimodal domain transfer for video-level sentiment analysis. *IEEE Access*, 10:51315–51324, 2022. 3

[42] Zhecan Wang, Haoxuan You, Liunian Harold Li, Alireza Zareian, Suji Park, Yiqing Liang, Kai-Wei Chang, and Shih-Fu Chang. Sgeitl: Scene graph enhanced image-text learning for visual commonsense reasoning. In *AAAI*, 2022. 1, 2, 6, 8

[43] Jialin Wu, Jiasen Lu, Ashish Sabharwal, and Roozbeh Mottaghi. Multi-modal answer validation for knowledge-based vqa. In *AAAI*, 2022. 3

[44] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019. 3

[45] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. An empirical study of gpt-3 for few-shot knowledge-based vqa. *ArXiv*, abs/2109.05014, 2021. 3

[46] Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Rich James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. Retrieval-augmented multimodal language modeling. In *International Conference on Machine Learning (ICML)*, 2023. 3

[47] Michihiro Yasunaga, Antoine Bosselut, Hongyu Ren, Xikun Zhang, Christopher D. Manning, Percy Liang, and Jure Leskovec. Deep bidirectional language-knowledge graph pretraining. In *Neural Information Processing Systems (NeurIPS)*, 2022. 3

[48] Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. Qa-gnn: Reasoning with language models and knowledge graphs for question answering. In *North American Chapter of the Association for Computational Linguistics (NAACL)*, 2021. 2, 3, 7

[49] Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-vil: Knowledge enhanced vision-language representations through scene graph. In *AAAI*, 2021. 1, 2, 6

[50] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *CVPR*, 2019. 1, 2, 5

[51] Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. Merlot reserve: Multimodal neural script knowledge through vision and language and sound. In *CVPR*, 2022. 1, 2, 6

[52] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. In *NeurIPS*, 2021. 1, 2, 6

[53] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Making visual representations matter in vision-language models. In *CVPR*, 2021. 3

[54] Yifeng Zhang, Ming Jiang, and Qi Zhao. Explicit knowledge incorporation for visual reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1356–1365, June 2021. 2, 3

[55] Maryam Ziaeefard and Freddy Lécué. Towards knowledge-augmented visual question answering. In *COLING*, 2020. 3