

An Empirical Study on Leveraging Scene Graphs for Visual Question Answering

Cheng Zhang¹
 zhang.7804@osu.edu

Wei-Lun Chao¹²
 weilunchao760414@gmail.com

Dong Xuan¹
 xuan.3@osu.edu

¹ The Ohio State University
 Columbus, Ohio, USA

² Cornell University
 Ithaca, New York, USA

[cs.CV] 28 Jul 2019

Abstract

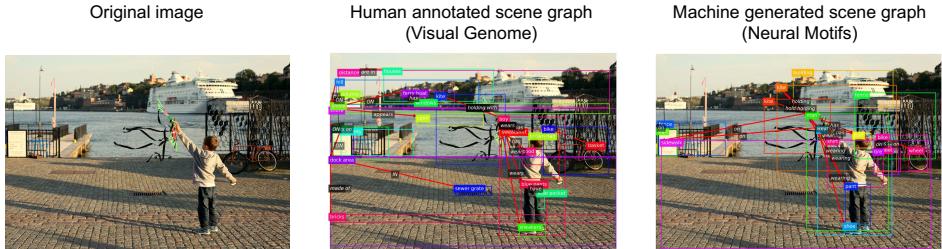
Visual question answering (Visual QA) has attracted significant attention these years. While a variety of algorithms have been proposed, most of them are built upon different combinations of image and language features as well as multi-modal attention and fusion. In this paper, we investigate an alternative approach inspired by conventional QA systems that operate on knowledge graphs. Specifically, we investigate the use of scene graphs derived from images for Visual QA: an image is abstractly represented by a graph with nodes corresponding to object entities and edges to object relationships. We adapt the recently proposed graph network (GN) to encode the scene graph and perform structured reasoning according to the input question. Our empirical studies demonstrate that scene graphs can already capture essential information of images and graph networks have the potential to outperform state-of-the-art Visual QA algorithms but with a much cleaner architecture. By analyzing the features generated by GNs we can further interpret the reasoning process, suggesting a promising direction towards explainable Visual QA.

1 Introduction

Scene understanding and reasoning has long been a core task that the computer vision community strives to advance. In recent years, we have witnessed significant improvement in many representative sub-tasks such as object recognition and detection, in which the machines' performance is on par or even surpasses humans' [23, 24], motivating the community to move toward higher-level sub-tasks such as visual captioning [1, 51, 52] and visual question answering (Visual QA) [2, 3, 11, 25].

One key factor to the recent successes is the use of neural networks [20], especially the convolutional neural network (CNN) [4] which captures the characteristics of human vision systems and the presence of objects in a scene. An object usually occupies nearby pixels and its location in the image does not change its appearance much, making CNNs, together with region proposals [5], suitable models for object recognition and detection. Indeed, when CNN [4] and R-CNN [19, 53] were introduced, we saw a leap in benchmarked performance.

Such a leap, however, has yet to be seen in high-level tasks. Taking Visual QA [3, 11, 25, 52] as an example, which takes an image and a question as inputs and outputs the



Question: What is the boy holding? **Answer:** A kite.

Figure 1: Scene graphs for Visual QA. We show the image, the human annotated graph [58], and the machine generated graph [7]. The answer can clearly be reasoned from the scene graph.

answer. In the past four years there are over a hundred of publications, but the improvement of the most advanced model over a simple multi-layer perceptron (MLP) baseline is merely $5 \sim 7\%$ [1, 12, 16]. Besides, the improvement is mostly built upon deep image and language features [24, 25, 26], multi-modal fusion [8, 18], and attention [2, 10], not a fundamental breakthrough like CNNs in object recognition.

Interestingly, if we take the visual input away from Visual QA, there has been a long history of development in question answering (QA), especially on leveraging the knowledge graphs (or bases) [8, 10, 48, 13, 15]. The basic idea is to represent the knowledge via entities and their relationships and then query the structured knowledge during testing time. In the vision community, we have also seen attempts to construct the so-called scene graph [31, 33] that can describe a visual scene in a similar way to a knowledge graph [10, 11, 15, 16, 19, 21]. Nevertheless, we have not seen a notable improvement or comprehensive analysis in exploiting the structured scene graph for Visual QA, despite the fact that several recent works have started to incorporate it [1, 10, 43, 44, 59, 10].

There are multiple possible reasons. Firstly, we may not yet have algorithms to construct high-quality scene graphs. Secondly, we may not yet have algorithms to effectively leverage scene graphs. Thirdly, perhaps we do not explicitly need scene graphs for Visual QA: either they do not offer useful information or existing algorithms have implicitly exploited them.

In this paper we aim to investigate these possible reasons. Specifically, we take advantage of the recently published graph network (GN) [8], which offers a flexible architecture to encode nodes (*e.g.*, *object entities* and *attributes*), edges (*e.g.*, *object relationships*), and global graph properties as well as perform (iterative) structured computations among them. By treating the question (and image) features as the input global graph properties and the answer as the output global properties, GNs can be directly applied to incorporate scene graphs and be learned to optimize the performance of Visual QA.

We conduct comprehensive empirical studies of GNs on the Visual Genome dataset [14, 58], which provides human annotated scene graphs. Recent algorithms on automatic scene graph generation from images [24, 45, 55, 56, 59, 12] have also focused on Visual Genome, allowing us to evaluate these algorithms using Visual QA as a down-stream task (see Fig. 1). Our experiments demonstrate that human annotated or even automatically generated scene graphs have already captured essential information for Visual QA. Moreover, applying GNs without complicated attention and fusion mechanisms shows promising results but with a much cleaner architecture. By analyzing the GN features along the scene graph we can further interpret the reasoning process, making graph networks suitable models to leverage scene graphs for Visual QA tasks.

2 Related Work

Visual question answering (Visual QA). Visual QA requires comprehending and reasoning with visual and textual information. Existing algorithms mostly adopt the pipeline that first extracts image and question features [30], followed by multi-modal fusion [8, 10] and attention [1, 4, 7] to obtain multi-modal features to infer the answer. While achieving promising results, it remains unclear if the models have been equipped with reasoning abilities or solely rely on exploiting dataset biases. Indeed, the performance gain between a simple MLP baseline and a complicated attention model is merely within 5 ~ 7% [27]. The improvement brought by newly proposed models usually lies within 0.5% [2, 8, 24].

Some new paradigms thus aim to better understand the intrinsic behavior of models [35]. One direction is to leverage object relationships for better scene understanding [32, 43, 46, 58, 65]. The other aims for interpretable Visual QA via neuro-symbolic learning [51, 62, 74], developing symbolic functional programs to test machines’ reasoning capability in Visual QA. Nevertheless, many of these works mainly experiment on synthetic data.

QA with knowledge bases. In conventional QA systems without visual inputs, exploiting knowledge bases (KBs) [9, 17, 26] to store complex structured and unstructured information so as to support combinatorial reasoning has been widely investigated [8, 11, 17, 48, 73, 75]. There are generally two kinds of KBs [10]: curated and extracted. Curated KBs, such as Freebase [9] and YAGO2 [26], extract <entity – relationship – entity> triples from knowledge sources like Wikipedia and WordNet [63]. Extracted KBs [1, 12], on the other hand, extract knowledge in the form of natural language from millions of web pages.

The scene graph [11, 38] examined in our work could be seen as a form of curated KBs that extracts triplets from images. Specifically, an image will be abstractly represented by a set of object entities and their relationships. Some Visual QA works indeed extract scene graphs followed by applying algorithms designed for reasoning on curated KBs [40, 52]. In our work we also exploit scene graphs, but apply the recently published graph networks (GNs) [8] that can easily incorporate advanced deep features for entity and relationship representations and allow end-to-end training to optimize overall performance.

Scene graphs. Lately, generating scene graphs from images has achieved notable progress [29, 44, 55, 67, 69, 77]. As an abstraction of images, scene graphs have been shown to improve image retrieval [31], generation [34], captioning [11, 27, 72], video understanding [49, 63], and human object interaction [56].

Recent works on Visual QA [2, 11, 43, 46, 59, 70] and visual reasoning [33, 59] have also begun to exploit scene graphs. However, they usually integrate multiple techniques (*e.g.*, scene graph generation, attention, multi-modal fusion) into one hybrid model to obtain state-of-the-art performance. It is thus hard to conclude if scene graphs truly contribute to the improvement. Indeed, in an early study [66] of the Visual Genome dataset, only 40% of the questions could be answered exactly with the human annotated scene graphs, assuming that the reasoning on the graphs is perfect¹. This ratio is surprisingly low considering the detailed knowledge encoded by scene graphs and the existing Visual QA models’ performance (*e.g.*, > 50% [7]).

We therefore perform a systematic study on leveraging scene graphs without applying specifically designed attention and fusion mechanisms. The results can then faithfully indicate the performance gain brought by performing structured reasoning on scene graphs.

¹In the study [56], a question is considered answerable by the scene graph if its answer exactly matches any node or relationship names or their combinations from the scene graph.

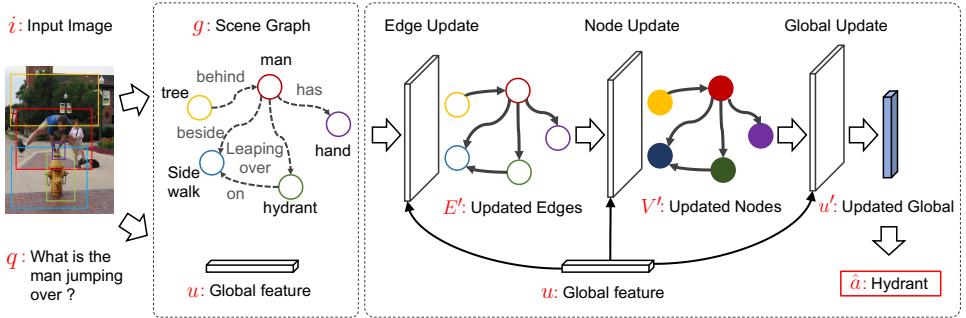


Figure 2: The GN-based Visual QA framework. The unfilled (filled) nodes, dashed (solid) edges, and unfilled (filled) cuboids denote the un-updated (updated) features. The updated global features (filled cuboid) are used to predict the answer.

3 Leveraging Scene Graphs for Visual QA

In this section we describe the graph network (GN) architecture [3] and how we apply it to reason from a scene graph according to the input image and question for Visual QA. Fig. 2 gives an illustration of the GN-based Visual QA framework. In what follows, we first define the Visual QA task and the scene graph, and then introduce a general Visual QA framework.

3.1 Problem definitions

A Visual QA model takes an image i and a related question q as inputs, and needs to output the correct answer a . In this work, we consider the multiple-choice Visual QA setting [3, 12], [28], in which the model needs to pick the correct answer a from a set of K candidate answers $\mathcal{A} = \{a, d_1, \dots, d_{K-1}\}$. d_k is called a negative answer or decoy. Nevertheless, our model can be easily extended into the open-ended setting with the answer embedding strategy [27].

We explicitly consider encoding the image i via a directed scene graph $g = (V, E)$ [31]. The node set $V = \{v_n\}_{n=1:N}$ contains N objects of the image, where each v_n records an object's properties such as its name (e.g., a car or a person), attributes (e.g., colors, materials), location, and size. The edge set $E = \{(e_m, s_m, o_m)\}_{m=1:M}$ contains M pairwise relationships between nodes, where e_m encodes the relationship name (e.g., on the top of); s_m, o_m are the indices of the subject and object nodes, respectively. For now let us assume that g is given for every i . We note that N and M may vary among different images.

A general Visual QA model thus can be formulated as follows [14, 30],

$$\hat{a} = \arg \max_{c \in \mathcal{A}} h_\theta(c, i, q, g), \quad (1)$$

where $c \in \mathcal{A}$ is a candidate answer and h_θ is a learnable scoring function on how likely c is the correct answer of (i, q, g) .

In Visual QA without the scene graph g , $h_\theta(c, i, q)$ can be modeled by an MLP with the concatenated features of (c, i, q) as the input. For example, [14, 30] represents c and q via the average word vectors and i via CNN features. Alternatively, one can factorize $h_\theta(c, i, q)$ by $\gamma_\theta(\alpha_\theta(i, q), \beta_\theta(c))$, where $\alpha_\theta(i, q)$ is the joint image-question embedding and $\beta_\theta(c)$ is the answer embedding [2]. γ_θ measures the compatibility of the two embeddings, e.g., by inner products. We denote the former as unfactorized models and the later as factorized models.

In Visual QA with scene graphs, we must define $h_\theta(c, i, q, g)$ so that it can take the nodes V and edges E into account as well as maintain the permutation invariant characteristic of a graph². Moreover, $h_\theta(c, i, q, g)$ must be able to reason from the graph g according to (c, i, q) . In the next subsection we introduce the graph network (GN), which provides a flexible computational architecture to fulfill the requirements above.

3.2 Graph networks (GNs)

The graph network (GN) proposed in [1] defines a computational block (module) that performs *graph-to-graph* mapping. That is, it takes a graph $g = (u, V, E)$ as the input and outputs an updated graph $g' = (u', V', E')$, which has the same graphical structure but different encoded information (*i.e.*, v_n and e_m will be updated). *The components u and u' encode certain global properties (features) of the graph.* For example, in Visual QA, u can be used to encode the image i and question q , while u' can be used to predict the final answer a .

There are various ways to define the updating procedure, as long as it maintains permutation invariant with respect to the graph. Here we describe a procedure that updates edges, nodes, and global features in order. We will then discuss how this updating procedure is particularly suitable for Visual QA with scene graphs.

Edge updates. The edge updating function f_θ^e is performed per edge m according to its features (encoded information) e_m , the features of the subject and object nodes v_{s_m} and v_{o_m} , and the global features u ,

$$e'_m = f_\theta^e(e_m, v_{s_m}, v_{o_m}, u). \quad (2)$$

Node updates. The subsequent node updates first aggregate the information from incoming edges of each node n ,

$$\overline{e'_n} = \phi^{e \rightarrow v}(E'_n), \quad (3)$$

where $\phi^{e \rightarrow v}$ is an aggregation function and $E'_n = \{(e'_m, s_m, o_m)\}_{o_m=n}$ is the set of incoming edges. We then apply the node updating function f_θ^v to each node n ,

$$v'_n = f_\theta^v(v_n, \overline{e'_n}, u). \quad (4)$$

Global updates. Finally, the global updating function f_θ^u is applied to update the global features. It begins with aggregating the edge and node features,

$$\overline{e'} = \phi^{e \rightarrow u}(E'), \quad (5)$$

$$\overline{v'} = \phi^{v \rightarrow u}(V'), \quad (6)$$

where $E' = \{(e'_m, s_m, o_m)\}_{m=1:M}$ and $V' = \{v'_n\}_{n=1:N}$. The updated u' is then computed by

$$u' = f_\theta^u(\overline{e'}, \overline{v'}, u). \quad (7)$$

In our studies, we assume that u, v_n, e_m (and u', v'_n, e'_m) each can be represented by a real vector. We make the following choices of aggregation and updating functions. The aggregation functions $\phi^{e \rightarrow v}, \phi^{e \rightarrow u}, \phi^{v \rightarrow u}$ should be able to take various number of inputs and must be permutation invariant with respect to the indices of their inputs. Representative

²That is, even we permute the indices of the nodes or edges, the output of h_θ should be the same.

options are element-wise average, sum, and max operations and we use element-wise average in the paper. The resulting \bar{e}_n^v , \bar{e}^l , and \bar{v}^l are thus real vectors.

The updating functions f_θ^e , f_θ^v , and f_θ^u can be any learnable modules such as MLPs and CNNs. In this paper, we apply MLPs with the concatenated features as the input. We note that, all the updating functions are optimized simultaneously; all the edges and nodes share the same updating functions f_θ^e and f_θ^v . More importantly, all the updating functions can be shared across different graphs even if they have different structures.

The resulting graph $g' = (u', V', E')$ then can be used to perform inference for tasks like Visual QA, or serve as the input to a subsequent GN block.

3.3 GN-based Visual QA with scene graphs

We apply GNs to Visual QA with scene graphs for multiple reasons. First, GNs explicitly consider the graph structures in its computations and can share the learned functions across different graphs³. Second, by encoding the question features (as well as the image and candidate answer features) into the global graph features u , GNs directly supports reasoning on graphs with appropriate choices of updating functions and procedures. Finally, GNs can be easily incorporated into Eq. (1) and learned to optimize the overall Visual QA performance. In the following we give more details.

Features. We encode q and c (a candidate answer) with averaged word vectors [32] and i with CNN features [24], following [24, 27, 30]. We obtain the scene graph of each image either via human annotation [33] or via automatic scene graph generation [2]. Ideally, every node in a graph will be provided with a node name (*e.g.*, car) and a set of attributes (*e.g.*, red, hatchback) and we represent each of them by the averaged word vectors. We then concatenate them to be the node features v_n . For edges that are provided with the relationship name (*e.g.*, on the top of), we again encode each e_m by the average word vectors. We note that more advanced visual and natural language features [26] can be applied to further improve the performance.

Learning unfactorized Visual QA models. We model the scoring function $f_\theta(c, i, q, g)$ in Eq. (1) as below,

$$f_\theta(c, i, q, g) = u' \quad \text{s.t.} \quad (u', V', E') = \text{GN}(u = [c, i, q], V, E), \quad (8)$$

where $[.]$ is a concatenation operation. That is, the output of the GN global updating function f_θ^u is a scalar indicating whether c is the correct answer of the image-question pair. We learn the parameters of the GN block to optimize the binary classification accuracy, following [31]. We note that, even if only u' is being used to predict the answer, all the three updating functions will be learned jointly according to the updating procedures in Sect. 3.2.

Learning factorized Visual QA models. We can further factorize $f_\theta(c, i, q, g)$ in Eq. (1) by $\gamma_\theta(\alpha_\theta(i, q, g), \beta_\theta(c))$ [27]. Specifically, we model $\alpha_\theta(i, q, g)$ by a GN block, $\beta_\theta(c)$ by an MLP, and γ_θ by another MLP following [15],

$$\begin{aligned} \alpha_\theta(i, q, g) &= u' \quad \text{s.t.} \quad (u', V', E') = \text{GN}(u = [i, q], V, E), \\ \beta_\theta(c) &= c' \quad \text{s.t.} \quad c' = \text{MLP}(c), \\ f_\theta(c, i, q, g) &= \gamma_\theta(u', c') = \text{MLP}([c', u', |c' - u'|, c' * u']), \end{aligned} \quad (9)$$

³Different images may have different scene graphs with varied numbers of nodes and edges.

where $*$ indicates element-wise product. The factorized model offers faster training and inference and can be used for the open-ended setting directly: the GN block will be computed just once for an image-question pair no matter how many candidate answers are considered.

4 Experiments

4.1 Setup

Visual QA dataset. We conduct experiments on the *Visual Genome (VG)* [33] dataset. VG contains 101,174 images annotated with in total 1,445,322 (i, q, a) triplets. Chao *et al.* [14] further augment each triplet with 6 auto-generated decoys (incorrect answers) and split the data into 727K/283K/433K triplets for training/validation/testing, named *qaVG*. Following [14, 27], we evaluate the accuracy of picking the correct answer from 7 candidates.

Scene graphs. We evaluate both human annotated and machine generated scene graphs. VG directly provides human-annotated graphs [33], and we obtain machine generated ones by the start-of-the-art Neural Motifs (NM) [27]. Since the released NM model is trained from a different data split of *qaVG*, we re-train the model from scratch using the training images of *qaVG*. One key difference between the ground-truth and MotifsNet graphs is that the first ones provide both object names and attributes for nodes, while the later only provide names.

Baseline methods. We compare with the unfactorized MLP models [24, 30] and the factorized fPMC(MLP) and fPMC(SAN) [27] models. SAN stands for stacked attention networks [70]. We note that, while the MLP model is extremely simple, it is only outperformed by fPMC(SAN) with layers of attentions by 5%.

Variants of our models. We denote the unfactorized GN model in Eq. (8) as **U-GN** and the factorized GN model in Eq. (9) as **F-GN**. We compare different combinations of input global features u (*i.e.*, $[c, i, q]$, $[c, q]$ for U-GN and $[i, q]$, $[q]$ for F-GN) with or without image features. We also compare different node features, with or without attributes. Finally, we consider removing edges (*i.e.*, all nodes are isolated) or even removing the whole graph (*i.e.*, Visual QA using the global features only).

Implementation details. We have briefly described the features in Sect. 3.3. We apply 300-dimensional word vectors [52] to represent q , c (a candidate answer), node names, node attributes, and edge (relationship) names. If a node has multiple attributes, we again average their features. We apply ResNet-152 [24] to extract the 2,048-dimensional image features. *We perform ℓ_2 normalization to each features before inputting them to the updating functions.*

For the GN block, we implement each updating function by a one-hidden-layer MLP whose architecture is a fully-connected (FC) layer followed by batch normalization [29], ReLU, Dropout (0.5) [60], and another FC layer. The hidden layer is 8,192-dimensional. For f_θ^v and f_θ^e , we keep the output size the same as the input. For f_θ^u , U-GN has a 1-dimensional output while F-GN has a 300-dimensional output. For F-GN, β_θ and γ_θ are implemented by the same MLP architecture as above with a 8,192-dimensional hidden layer. β_θ has a 300-dimensional output while γ_θ has a 1-dimensional output.

We learn both U-GN and F-GN using the binary classification objective (*i.e.*, if c is the correct answer or not). We use the Adam optimizer [37] with an initial learning rate 10^{-3} and a batch size 100. We divide the learning rate by 10 whenever the validation accuracy decreases. We train for at most 30 epochs and pick the best model via the validation accuracy.

Table 1: Visual QA accuracy (%) on qaVG with unfactorized and factorized models. Input: global features. NM: neural motifs [■] graphs. For U-GN and F-GN, we always consider the relationship names on edges except for the no graph case. We note that fPMC(SAN) uses LSTM [□] to encode questions and fPMC(SAN★) additionally uses it to encode answers.

(a) Unfactorized models			(b) Factorized models		
Methods	Input	Accuracy	Methods	Input	Accuracy
MLP [□]	i, q, c	58.5	fPMC(MLP) [□]	i, q, c	57.7
HieCoAtt [□]	i, q, c	57.5	fPMC(SAN) [□]	i, q, c	62.6
Attnntion [□]	i, q, c	60.1	fPMC(SAN★) [□]	i, q, c	63.4
U-GN (No graphs)			F-GN (No graphs)		
-	q, c	43.3	-	q, c	44.8
-	i, q, c	58.3	-	i, q, c	59.4
U-GN (NM graphs)			F-GN (NM graphs)		
Name	q, c	57.9	Name	q, c	57.6
Name	i, q, c	60.5	Name	i, q, c	60.0
U-GN (VG graphs)			F-GN (VG graphs)		
Name	q, c	60.5	Name	q, c	60.1
Name	i, q, c	61.9	Name	i, q, c	60.7
Name + Attr	q, c	62.2	Name + Attr	q, c	61.9
Name + Attr	i, q, c	62.6	Name + Attr	i, q, c	62.5

4.2 Main results

We summarize the main results in Table 1, in which we compare our U-GN and F-GN models to existing models on multiple-choice Visual QA. Both U-GN (NM graphs) and F-GN (VG graphs) outperform existing unfactorized methods, while F-GN (VG graphs) is outperformed by fPMC(SAN★) by 1%. We note that fPMC(SAN) uses LSTM [□] to encode questions and fPMC(SAN★) additionally uses it to encode answers. We thus expect that our GN models to have improved performance with better language features.

We then analyze different variants of our models. Comparing U-GN (No graphs) to U-GN (VG graphs) and U-GN (NM graphs), we clearly see the benefit brought by leveraging scene graphs. Specifically, even without the image features (*i.e.*, input = q, c), U-GN (NM graphs) and U-GN (VG graphs) can already be on par or even surpass U-GN (No graphs) with input image features, suggesting that scene graphs, even automatically generated by machines, indeed encode essential visual information for Visual QA.

By further analyzing U-GN (VG graphs) and U-GN (NM graphs), we found that including node attributes always improve the performance, especially when no image features are used. We suggest that future scene graph generation should also predict node attributes.

We observe similar trends when applying the F-GN model.

GNs without edges. We study the case where edges are removed and nodes are isolated: no edge updates are performed; node and global updates do not consider edge features. The result with F-GN (VG graphs) is 61.3%, compared to 62.5% in Table 1, justifying the need to take the relationships between nodes into account for Visual QA.

Stacked GNs. We investigate stacking multiple GN blocks for performance improvement. We stack two GN blocks: the updated edges and nodes of the first GN block are served as the inputs to the second one. The intuition is that the messages will pass through the scene graph with one more step to support complicated reasoning. When paired with stacked GNs, F-GN (VG graphs) achieves a gain of +0.4% in overall accuracy compared to 62.5%.

Table 2: Visual QA accuracy (%) on different question types. VG: Visual Genome [15] graphs. NM: neural motifs [17] graphs. NG: no graphs. N: node names. A: node attributes.

Question type		What	Color	Where	Number	How	Who	When	Why	Overall
Percentage		(46%)	(14%)	(17%)	(8%)	(3%)	(5%)	(4%)	(3%)	(100%)
U-GN										
NG	+ (q, c)	40.3	50.6	36.2	52.0	41.1	37.6	83.2	39.5	43.3
NG	+ (i, q, c)	57.8	59.5	59.1	55.5	45.4	56.6	84.6	48.3	58.3
NM(N)	+ (i, q, c)	59.4	58.2	60.3	63.4	54.3	66.6	85.3	48.1	60.5
VG(N)	+ (q, c)	61.6	54.0	62.4	58.6	45.9	63.9	83.2	50.3	60.5
VG(N)	+ (i, q, c)	61.1	61.4	62.3	59.4	54.3	67.5	85.3	48.9	61.9
VG(N, A)	+ (i, q, c)	61.4	63.8	62.6	61.5	54.8	67.5	84.8	49.6	62.6
F-GN										
NG	+ (q, c)	41.4	52.6	38.7	53.4	42.2	39.2	83.4	40.2	44.8
NG	+ (i, q, c)	58.7	61.0	60.4	57.4	47.1	57.6	85.8	49.8	59.4
NM(N)	+ (i, q, c)	58.7	60.8	60.4	60.1	47.2	61.8	84.8	49.0	60.0
VG(N)	+ (q, c)	60.9	53.6	62.0	58.1	46.2	63.3	83.7	50.9	60.1
VG(N)	+ (i, q, c)	60.2	60.4	61.8	58.5	47.4	63.8	85.1	49.6	60.7
VG(N, A)	+ (i, q, c)	61.0	64.4	62.4	58.8	48.2	64.2	85.6	51.2	62.5

4.3 Analysis

We provide detailed results on *qaVG* with different question types in Table 2. We found that without input image features, scene graphs with node names significantly improve all question types but “when”, which needs holistic visual features. Even with image features, scene graphs with node names can still largely improve the “what”, “who”, and “number” types; the former two take advantage of node names while the later takes advantage of number of nodes. Adding the node attributes specifically benefits the “color” type. Overall, the VG graphs are of higher quality than NM graphs except for the “number” type. We surmise that well-trained object detectors adopted in [17] can capture smaller objects easily while human annotators might focus on salient ones.

Qualitative results. Fig. 3 shows several qualitative examples. Specifically, we only show nodes and edges that have higher ℓ_2 norms after updates. We see that GN-based Visual QA models (U-GN) can implicitly attend to nodes and edges related to the questions, revealing the underlying reasoning process.

5 Conclusion

In this paper we investigate if scene graphs can facilitate Visual QA. We apply the graph network (GN) that can naturally encode information on graphs and perform structured reasoning. Our experimental results demonstrate that scene graphs, even automatically generated by machines, can definitively benefit Visual QA if paired with appropriate models like GNs. Specifically, leveraging scene graphs largely increases the Visual QA accuracy on questions related to counting, object presence and attributes, and multi-object relationships. We expect that the GN-based model can be further improved by incorporating image features on nodes as well as advanced multi-modal fusion and attention mechanisms.

Acknowledgements. The computational resources are supported by the Ohio Supercomputer Center (PAS1510) [16].



Figure 3: Qualitative results (better viewed in color). We show the original scene graphs (full) and the filtered ones by removing updated nodes and edges with smaller ℓ_2 norms. Correct answers are in green and incorrect predictions are in red. VG: Visual Genome [☒] graphs. NM: neural motifs [☒] graphs. GN-based Visual QA models can implicitly attend to nodes and edges that are related to the questions (e.g., *kite* and *holding* in (a), *tracks* in (b), *racing* and *riding* in (c), and *snow* and *ground* in (d)). The failure case with NM graphs in (e) is likely due to that no node attributes are provided. The challenge of Visual QA in (f) is the visual common sense: how to connect a coarse-grained term (e.g., *equipment*) with fine-grained terms (e.g., *laptop*, *audio device* and so on). Zoom in for details.

References

- [1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *ECCV*, 2016.
- [2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018.
- [3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. VQA: Visual question answering. In *ICCV*, 2015.
- [4] Michele Banko, Michael J Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. Open information extraction from the web. In *IJCAI*, 2007.
- [5] Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018.
- [6] Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. MUTAN: Multimodal tucker fusion for visual question answering. In *ICCV*, 2017.
- [7] Hedi Ben-Younes, Rémi Cadene, Nicolas Thome, and Matthieu Cord. Block: Bilinear superdiagonal fusion for visual question answering and visual relationship detection. *arXiv preprint arXiv:1902.00038*, 2019.
- [8] Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. Semantic parsing on freebase from question-answer pairs. In *EMNLP*, 2013.
- [9] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *SIGMOD*, 2008.
- [10] Antoine Bordes, Sumit Chopra, and Jason Weston. Question answering with subgraph embeddings. In *EMNLP*, 2014.
- [11] Remi Cadene, Hedi Ben-younes, Matthieu Cord, and Nicolas Thome. MUREL: Multimodal relational reasoning for visual question answering. In *CVPR*, 2019.
- [12] Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R Hruschka, and Tom M Mitchell. Toward an architecture for never-ending language learning. In *AAAI*, 2010.
- [13] Ohio Supercomputer Center. Ohio supercomputer center, 1987. URL <http://osc.edu/ark:/19495/f5s1ph73>.
- [14] Wei-Lun Chao, Hexiang Hu, and Fei Sha. Being negative but constructively: Lessons learnt from creating better visual question answering datasets. In *NAACL*, 2018.
- [15] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. Supervised learning of universal sentence representations from natural language inference data. In *EMNLP*, 2017.
- [16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019.
- [17] Anthony Fader, Luke Zettlemoyer, and Oren Etzioni. Open question answering over curated and extracted knowledge bases. In *KDD*, 2014.
- [18] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multi-modal compact bilinear pooling for visual question answering and visual grounding. In *EMNLP*, 2016.
- [19] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.
- [20] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.
- [21] 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 *CVPR*, 2017.

- [22] Abhinav Gupta and Larry S Davis. Beyond nouns: Exploiting prepositions and comparative adjectives for learning visual classifiers. In *ECCV*, 2008.
- [23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*, 2015.
- [24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.
- [25] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.
- [26] Johannes Hoffart, Fabian M Suchanek, Klaus Berberich, Edwin Lewis-Kelham, Gerard De Melo, and Gerhard Weikum. Yago2: exploring and querying world knowledge in time, space, context, and many languages. In *WWW*, 2011.
- [27] Hexiang Hu, Wei-Lun Chao, and Fei Sha. Learning answer embeddings for visual question answering. In *CVPR*, 2018.
- [28] Drew A Hudson and Christopher D Manning. GQA: A new dataset for compositional question answering over real-world images. *arXiv preprint arXiv:1902.09506*, 2019.
- [29] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015.
- [30] Allan Jabri, Armand Joulin, and Laurens Van Der Maaten. Revisiting visual question answering baselines. In *ECCV*, 2016.
- [31] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *CVPR*, 2015.
- [32] 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 *CVPR*, 2017.
- [33] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Judy Hoffman, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Inferring and executing programs for visual reasoning. In *CVPR*, 2017.
- [34] Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image generation from scene graphs. In *CVPR*, 2018.
- [35] Kushal Kafle and Christopher Kanan. An analysis of visual question answering algorithms. In *ICCV*, 2017.
- [36] Keizo Kato, Yin Li, and Abhinav Gupta. Compositional learning for human object interaction. In *ECCV*, 2018.
- [37] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.
- [38] 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.
- [39] Ranjay Krishna, Ines Chami, Michael Bernstein, and Li Fei-Fei. Referring relationships. In *CVPR*, 2018.
- [40] Jayant Krishnamurthy and Thomas Kollar. Jointly learning to parse and perceive: Connecting natural language to the physical world. *Transactions of the Association for Computational Linguistics*, 1:193–206, 2013.
- [41] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.
- [42] Yann LeCun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series. *The Handbook of Brain Theory and Neural Networks*, 3361(10):1995, 1995.
- [43] Linjie Li, Zhe Gan, Yu Cheng, and Jingjing Liu. Relation-aware graph attention network for visual question answering. *arXiv preprint arXiv:1903.12314*, 2019.

- [44] Yikang Li, Wanli Ouyang, Bolei Zhou, Kun Wang, and Xiaogang Wang. Scene graph generation from objects, phrases and region captions. In *ICCV*, 2017.
- [45] Yikang Li, Wanli Ouyang, Bolei Zhou, Jianping Shi, Chao Zhang, and Xiaogang Wang. Factorizable net: an efficient subgraph-based framework for scene graph generation. In *ECCV*, 2018.
- [46] Yuanzhi Liang, Yalong Bai, Wei Zhang, Xueming Qian, Li Zhu, and Tao Mei. Rethinking visual relationships for high-level image understanding. *arXiv preprint arXiv:1902.00313*, 2019.
- [47] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *NIPS*, 2016.
- [48] Denis Lukovnikov, Asja Fischer, Jens Lehmann, and Sören Auer. Neural network-based question answering over knowledge graphs on word and character level. In *WWW*, 2017.
- [49] Chih-Yao Ma, Asim Kadav, Iain Melvin, Zsolt Kira, Ghassan AlRegib, and Hans Peter Graf. Attend and interact: Higher-order object interactions for video understanding. In *CVPR*, 2018.
- [50] 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. In *ICLR*, 2019.
- [51] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. Deep captioning with multi-modal recurrent neural networks (m-rnn). In *ICLR*, 2015.
- [52] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, 2013.
- [53] George A Miller. WordNet: A lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.
- [54] Jonghwan Mun, Kimin Lee, Jinwoo Shin, and Bohyung Han. Learning to specialize with knowledge distillation for visual question answering. In *NeurIPS*, 2018.
- [55] Alejandro Newell and Jia Deng. Pixels to graphs by associative embedding. In *NIPS*, 2017.
- [56] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*, 2015.
- [57] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.
- [58] Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. A simple neural network module for relational reasoning. In *NIPS*, 2017.
- [59] Jiaxin Shi, Hanwang Zhang, and Juanzi Li. Explainable and explicit visual reasoning over scene graphs. In *CVPR*, 2019.
- [60] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.
- [61] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International Journal of Computer Vision*, 104(2):154–171, 2013.
- [62] Ramakrishna Vedantam, Karan Desai, Stefan Lee, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Probabilistic neural-symbolic models for interpretable visual question answering. *arXiv preprint arXiv:1902.07864*, 2019.
- [63] Paul Vicol, Makarand Tapaswi, Lluis Castrejon, and Sanja Fidler. Moviegraphs: Towards understanding human-centric situations from videos. In *CVPR*, 2018.
- [64] 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, 2018.

- [65] Yu-Siang Wang, Chenxi Liu, Xiaohui Zeng, and Alan Yuille. Scene graph parsing as dependency parsing. In *NAACL*, 2018.
- [66] Qi Wu, Damien Teney, Peng Wang, Chunhua Shen, Anthony Dick, and Anton van den Hengel. Visual question answering: A survey of methods and datasets. *Computer Vision and Image Understanding*, 163:21–40, 2017.
- [67] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *CVPR*, 2017.
- [68] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015.
- [69] Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. Graph R-CNN for scene graph generation. In *ECCV*, 2018.
- [70] Zhuoqian Yang, Jing Yu, Chenghao Yang, Zengchang Qin, and Yue Hu. Multi-modal learning with prior visual relation reasoning. *arXiv preprint arXiv:1812.09681*, 2018.
- [71] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *CVPR*, 2016.
- [72] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In *ECCV*, 2018.
- [73] Xuchen Yao and Benjamin Van Durme. Information extraction over structured data: Question answering with freebase. In *ACL*, 2014.
- [74] Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Josh Tenenbaum. Neural-symbolic VQA: Disentangling reasoning from vision and language understanding. In *NeurIPS*, 2018.
- [75] Scott Wen-tau Yih, Ming-Wei Chang, Xiaodong He, and Jianfeng Gao. Semantic parsing via staged query graph generation: Question answering with knowledge base. In *ACL*, 2015.
- [76] Zhou Yu, Jun Yu, Chenchao Xiang, Jianping Fan, and Dacheng Tao. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. *IEEE Transactions on Neural Networks and Learning Systems*, (99):1–13, 2018.
- [77] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *CVPR*, 2018.
- [78] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7W: Grounded question answering in images. In *CVPR*, 2016.

Supplementary Material

In this Supplementary Material, we provide details omitted in the main text.

- Section A: Implementation details (Sect. 4.1 of the main text).
- Section B: Additional experimental results (Sect. 4.2 and Sect. 4.3 of the main text).

A Implementation Details

In this section, we provide more details about the configuration of scene graph generation, scene graph encoding, the stacked GN model, and the corresponding training procedures.

A.1 Configuration of scene graph generation

Human annotated scene graph. We leverage the ground truth scene labels of the Visual Genome (VG) dataset [33] as our human annotated scene graphs⁴.

Machine generated scene graph. The commonly used data split of scene graph generation research is different from *qaVG* [24]⁵. Thus we retrain the start-of-the-art Neural Motifs (NM) [27] model only using the training images from *qaVG*. Specifically, we retrain both object detector and relationship classifier of the NM to ensure the model never uses testing images for training. We run the well-trained NM model on *qaVG* to obtain the machine generated scene graph by removing entities (<0.2) and relationships (<0.1) with small probabilities. We show the scene graph detection performance in Table 3.

Table 3: Scene graph detection accuracy (%) using recall@K metrics. The released NM model [27] is evaluated on the split of [62]. Our retrained NM model is evaluated on the split of *qaVG* [24]. We can see that the machine generated scene graph from the retrained NM model achieves satisfied performance.

Models	R@20	R@50	R@100
Released [27]	21.7	27.3	30.5
Retrained (ours)	21.5	27.5	30.6

A.2 Scene graph encoding

As mentioned in Sect. 4.1 of the main text, we extract corresponding features to represent nodes, edges, and global. To represent image i , we extract the activations from the penultimate layer of the ResNet-152 [24] pretrained on ImageNet [50] and average them to obtain a 2,048-dimentional feature representation. The question q , candidate c , node names, node attributes, and edge names are represented as the average word to vector [50] embeddings. Specifically, we remove punctuation, change all characters to lowercases, and convert all integer numbers within [0, 10] to words before computing word to vector. We use ‘UNK’ to represent out-of-vocabulary (OOV) word. Finally, the NM scene graph can be represented as

⁴The VG scene graphs are obtained from <https://visualgenome.org/> without any modifications.

⁵The *qaVG* follows the same data split of Visual Genome dataset for Visual QA task.

300-dimentional nodes and 300-dimentional edges embeddings, and the VG scene graph has 600-dimentional nodes and 300-dimentional edges representations. To enable better generalization on unseen datasets, we fix all the visual and language features in our experiments. All individual features are ℓ_2 normalized before concatenation.

A.3 Stacked GNs

We provide more details about stacked GNs. As mentioned in Sect. 4.2 of the main paper, the overall accuracy could be improved with respect to the number of GN blocks. Here, we propose one design choice of the stacked GNs as below,

$$\begin{aligned}(u, V', E') &= \text{GN}_{f_\theta^e, f_\theta^v}(u, V, E), \\ (u', V'', E'') &= \text{GN}_{f_\theta^e, f_\theta^v, f_\theta^u}(u, V', E'),\end{aligned}\tag{10}$$

where the first GN block only performs edge and node updating, and the updated properties will be served as the inputs of the second one. Finally, the resulting global feature u' can be used to perform inference in Visual QA. Similarly, multiple GN blocks can be stacked in such manner. We expect varied designs of multi-layer GN could be proposed for performance improvement, such as jointly learning global feature within the latent GN blocks and making GN blocks recurrent.

A.4 Optimization

For all above models, we train for at most 30 epochs using stochastic gradient optimization with Adam [37]. The initial learning rate is 10^{-3} , which is divided by 10 after M epochs. We turn the M on the validation set and choose the best model via validation performance.

Within each mini-batch, we sample 100 (i, q, a) triplets. In order to prevent unbalanced training⁶, we follow the sampling strategy suggested by Chao *et al.* [14]. We randomly choose to use QoU-decoys or IoU-decoys for each triplet as negative samples when training. Then the binary classifier is trained on top of the target and 3 decoys for each triplet. That is, 100 triplets in the each mini-batch related to 400 samples with binary labels. In the testing stage, we evaluate the performance of picking the correct answer from all 7 candidates.

B Additional Results

In this section, we provide more qualitative results. In Fig. 4, 5, 6, 7, 8, we show the original scene graphs (full) and the filtered ones by removing updated nodes and edges with smaller ℓ_2 norms. Correct answers are in green and incorrect predictions are in red. VG: Visual Genome [38] graphs. NM: neural motifs [27] graphs.

⁶In *qaVG* [14], each (i, q) pair contains 3 QoU-decoys (incorrect answers), 3 IoU-decoys (incorrect answers), and 1 target (correct answer). The machine tends to predict the dominant label if the training is performed among all samples.

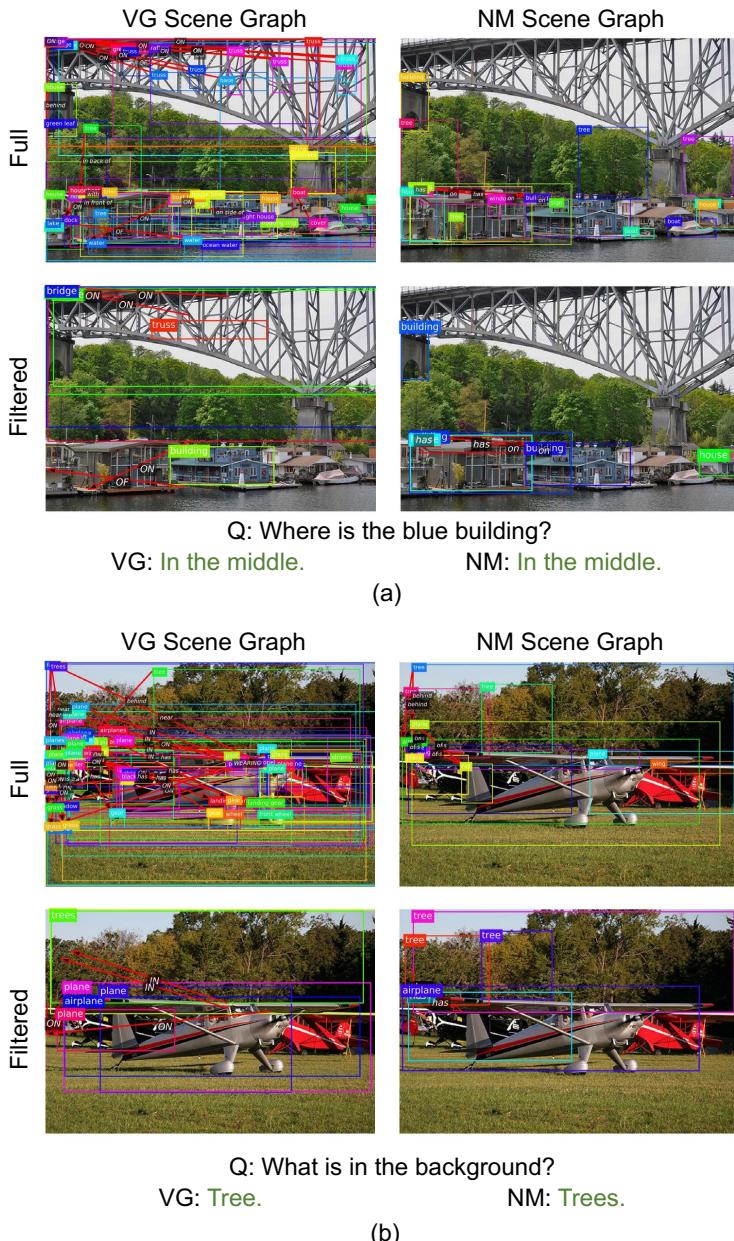
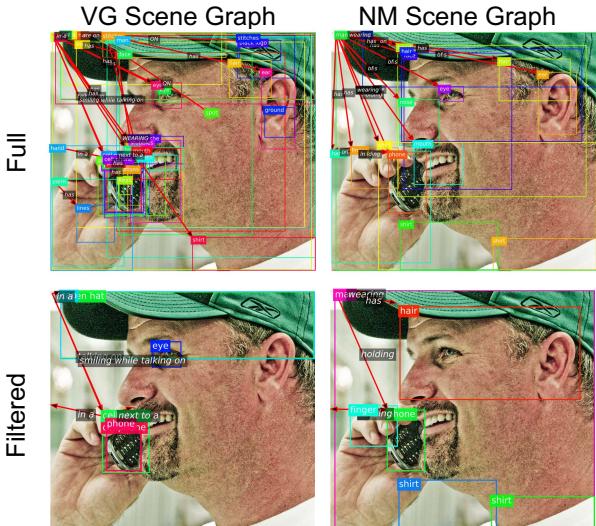


Figure 4: Qualitative results. (a) Both VG and NM graphs can attend to the *building in the middle*. (b) The models can attend to the *trees* behind the foreground objects.

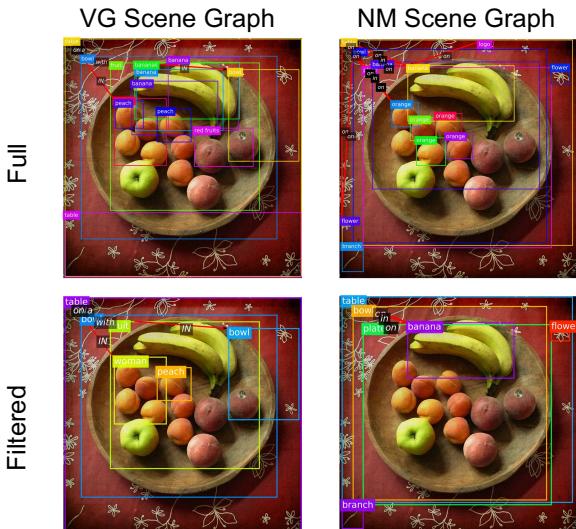


Q: What hand is the man using to hold phone?

VG: Right hand.

NM: Right hand.

(a)



Q: Where are the pieces of fruit?

VG: On the plate.

NM: On the plate.

(b)

Figure 5: Qualitative results. (a) Both two scene graphs show the location and relationships of the phone. (b) Both VG and NM graphs attend to the location of the fruits.

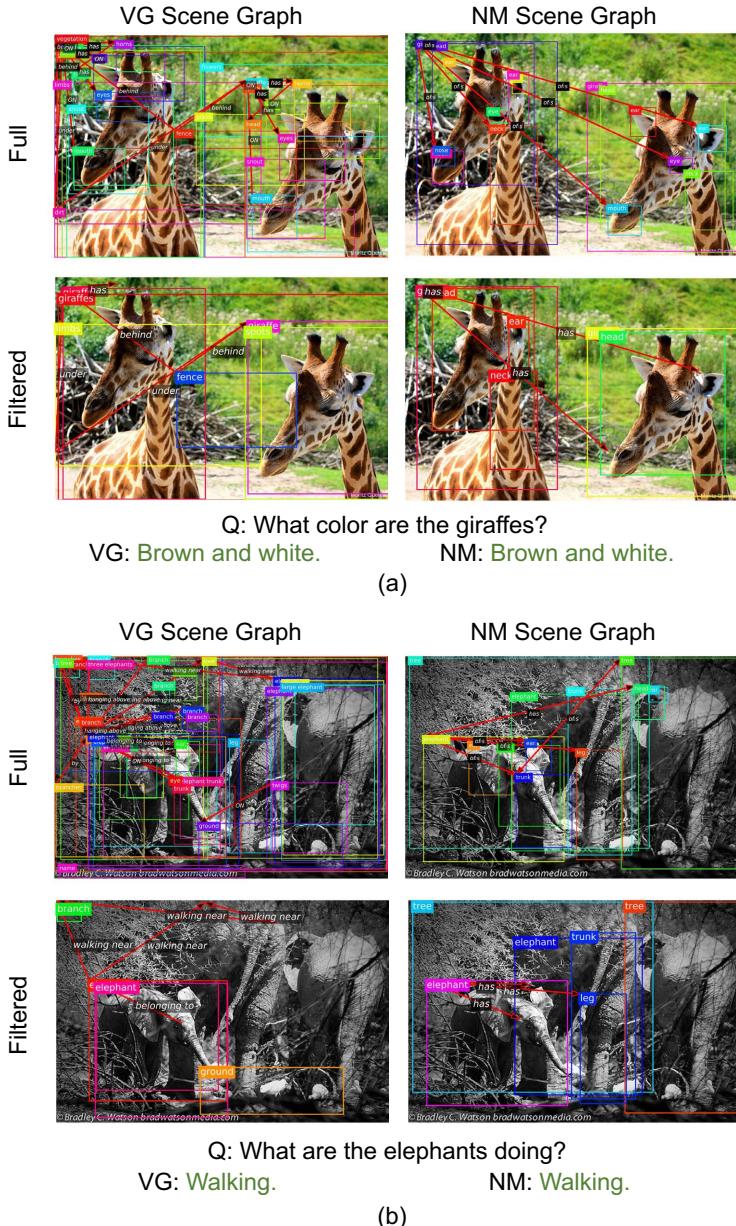


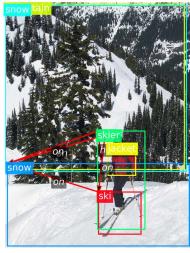
Figure 6: Qualitative results. (a) Both VG and NM graphs attend to the giraffes. Even without color attributes, the NM scene graph may learn the colors through overall image features. (b) VG graph clearly captures *walking near* and NM graph attends to *elephants*.

VG Scene Graph NM Scene Graph

Full



Filtered



Q: How many skis?

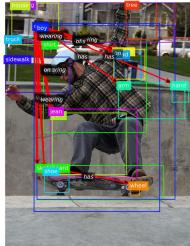
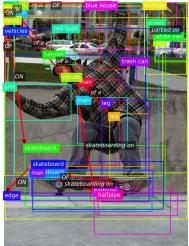
VG: Two.

NM: Two.

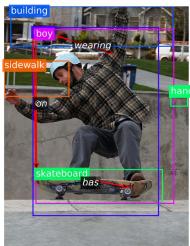
(a)

VG Scene Graph NM Scene Graph

Full



Filtered

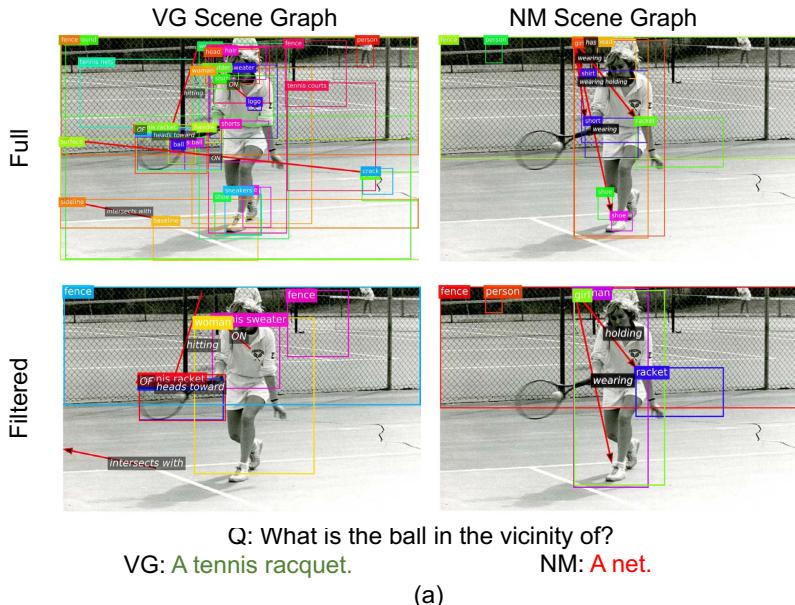


Q: What is the man on?

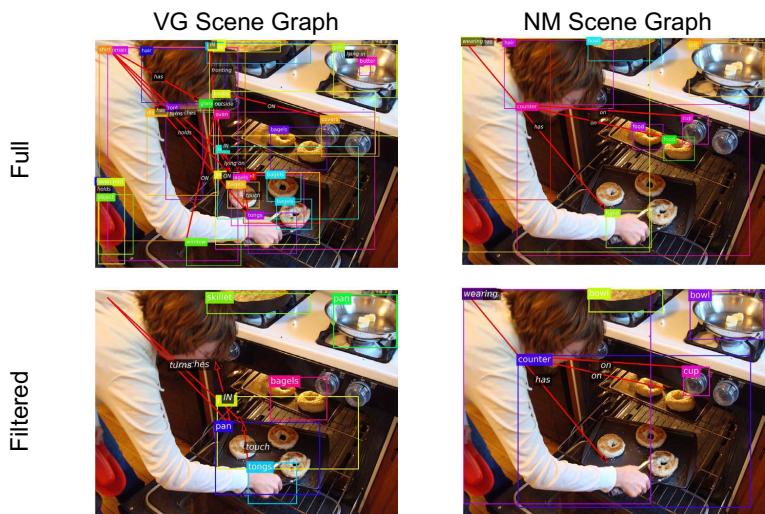
VG: A skateboard. NM: A skateboard.

(b)

Figure 7: Qualitative results. (a) Both two graphs capture *two skis* from the full scene graphs. (b) VG graph attends to the relationship *skateboarding on* and NM graph attends to *boy on skateboard*.



(a)



(b)

Figure 8: Failure cases. (a) VG scene graph attends to *heads toward* and *tennis racket* to predict the correct answer while NM graph fails because of lacking such high level semantic annotations. (b) The model needs to understand the interaction between the person and the food inside the oven, which is a hard case.