

Graph Reasoning Networks for Visual Question Answering

Dalu Guo, Chang Xu, Dacheng Tao

UBTECH Sydney AI Centre, School of Computer Science, FEIT,
University of Sydney, Darlingtown, NSW 2008, Australia

{dguo8417@uni., c.xu@, dacheng.tao@}sydney.edu.au

Abstract

The interaction between language and visual information has been emphasized in visual question answering (VQA) with the help of attention mechanism. However, the relationship between words in question has been underestimated, which makes it hard to answer questions that involve the relationship between multiple entities, such as comparison and counting. In this paper, we develop the graph reasoning networks to tackle this problem. Two kinds of graphs are investigated, namely inter-graph and intra-graph. The inter-graph transfers features of the detected objects to their related query words, enabling the output nodes to have both semantic and factual information. The intra-graph exchanges information between these output nodes from inter-graph to amplify implicit yet important relationship between objects. These two kinds of graphs cooperate with each other, and thus our resulting model can reason the relationship and dependence between objects, which leads to realization of multi-step reasoning. Experimental results on the GQA v1.1 dataset demonstrate the reasoning ability of our method to handle compositional questions about real-world images. We achieve state-of-the-art performance, boosting accuracy to 57.04%. On the VQA 2.0 dataset, we also receive a promising improvement on overall accuracy, especially on counting problem.

1. Introduction

The developments in computer vision and natural language processing enable the machine to deal with complicated tasks that require the integration and understanding of vision and language, e.g. image captioning [1], visual grounding [44], visual question answering (VQA) [2] and visual dialog [6]. Compared with image captioning that is to simply describe the topic of an image, VQA needs a complex reasoning process. Visual grounding aims to locate the related objects in the image, but VQA takes a further step to convert this information into human language. In addition, VQA is the basic and vital component in visual dialog. Con-

sidering the challenges and significance of VQA, increasing research attention has been attracted to this task.

Given an input image and a question, representative VQA models, e.g. Stacked Attention Networks (SAN) [41], Multimodal Compact Bilinear pooling (MCB) [9] and Multimodal Low-rank Bilinear Attention Networks (MLB) [19], first generate grid image features by ResNet [12] and represent the question as the last hidden state of Long Short-Term Memory (LSTM) [13], and then attend to the image features based on question vector to ground the target objects; the question vector and attention weighted image features are finally projected into a unified embedding for answer prediction. Bilinear Attention Networks (BAN) [18] notice that these methods neglect the interaction between words in the question and objects in the image and propose to build a co-attention considering each pair of multi-modal channels. However, there still lacks a full exploitation of the interactions between words in questions and a systemic exploration of the connections between questions and images. Although the text attention has been learned through Multi-layer Perceptron (MLP) in Multimodal Factorized Bilinear pooling (MFB) [42] or memory vector in Dual Attention Networks (DAN) [26], the difficulty remains over how to answer complex questions, like comparing the relative position of two entities or counting the queried objects in Figure 1. The networks should first view the image to locate one fruit, such as the tomato, and then compare its position to other fruits before deciding which one is on the left. Regarding the comparison operation, the correct answer may not be decided at once, since the orange is to the left of the tomato but is to the right of the apple. The networks therefore have to extract more clues from the question-image pair and conduct comparison for multiple times.

In this paper, we develop graph reasoning networks for visual question answering. Besides investigating visual attention map between words in the question and objects in the image, we highlight the importance of exploiting intra-relationships between words in the question and exploring the inter-relationships between the question and image for a complex reasoning. Two graphs are established to formu-

late these two kinds of relationships. The inter-graph focuses on exploring visual features of the image to their related textual features for joint embeddings, which links the semantic information of words with factual information of image. The intra-graph exploits information across different joint embeddings in term of words, which amplifies the implicit yet important relationships between objects. Given these two graphs cooperating with each other, the resulting VQA model is able to reason complex and compositional questions.

We conduct experiments on both GQA [15] v1.1 dataset and VQA v2.0 dataset [10]. On the validation dataset of GQA, our one-layer graph networks boost the accuracy by 1.55% compared with BAN, and graphs of multiple layers show advantages on multi-step reasoning for long and complex questions, evidenced by total 2.27% improvement. A similar enhancement can also be found on the validation dataset of VQA. On the test-dev dataset of GQA and VQA, our model achieves state-of-the-art performance, increasing the overall accuracy to 57.04% and 71.0% respectively.

2. Related Work

In this section, we will first introduce the related researches on VQA and then the graph networks on both text-based and visual-based tasks.

Visual Question Answering (VQA): VQA is a task to answer the given question based on the input image. The question is usually embedded into a vector with LSTM [13], and the image is represented by the fixed-size grid features from ResNet [12]. Recently, [1] focuses on bottom-up attention of image features and proposes a set of salient image regions with natural expression and additional attributes detected by Faster-RCNN [32]. Furthermore, its training set contains 1,600 object classes and 400 attribute classes, larger than the original 80 object classes.

Based on the fusion methods of the two features, we can classify VQA models into two categories: early fusion models and later fusion models. Early fusion models try to fine-tune the image classification network with the intervention of the question, they insert the question embedding into the batch normalization layer [7, 31] to propose MODERN architecture. These models have less risk of over-fitting because of affecting less than 1% parameters in the pre-trained classification network. And later fusion models mainly concentrate on how to fuse the two different distributed features. [16, 2] represent the answer vector by concatenating the question vector and global image vector. However, not all visual information is relevant to the input query, which can result in much noise during reasoning. Therefore, attention map is learned to filter out useful image features, such as SAN [41], which learns the visual attention through multi-steps, and DAN [26], which learns visual and textual attention respectively via the memory vector. But the linear

combination is not enough to represent the joint embedding of two features from different distributions. Thus, bilinear methods like MLB [19] and MFB [42] are proposed, they project the two features into a common low-rank space followed by element-wise production. Furthermore, BAN [18] learns the textual and visual attention simultaneously, which builds a graph mapping from the detected objects of the image to the words of the question.

However, sometimes, the information given in the image is not enough to infer the right answer, common sense is required in the external knowledge-based models. [38] builds the FVQA dataset based on DBpedia [3], ConceptNet [25] and WebChild [35]. [28] queries the triplet (visual concept, relation, attribute) in this dataset to score the retrieved facts. And [27] builds a relation graph based on the retrieved facts regarding the visual concept and attribute as nodes and relation as links to exchange information.

Graph Neural Network (GNN): GNN is used to build the relationship between nodes like social network, citation link [11], knowledge graph [21], protein-protein interaction [37], etc.. It could overcome the limitation of Euclidean distance between each node in the inputs and involve more context information from neighbors. In text-based tasks, such as machine translation and sequence tagging, GNN breaks the sequence restriction between each word and learns the graph weight by attention mechanism [36, 8], which makes it model longer sequence more easily than LSTM and gated recurrent neural network (GRU) [5], since each node is directly linked with others by learned weights instead of hidden state and gates. Moreover, the learned graph weight, which implies dependencies between nodes, can be easily explained and transferred to other tasks for pre-trained weights [40]. In the image-based tasks, GNN gathers information from all grids [39] other than surroundings whose size is limited by the receptive fields of convolution neural networks (CNNs), and it can aggregate features over coordinate space to compute complex dependence [4]. Besides modeling the relationship between homogeneous inputs, GNN can also work in the tasks of multi-modal inputs such as VQA. The entities retrieved from external facts exchange information through the graph with multiple hops to predict the answer [27]. And the graph enhances the interpretation of network to reason the relationship among detected objects in the image filtered by the question [29].

3. Preliminaries

The goal of VQA task is to answer the given question Q based on image I . With the object-detector Faster-RCNN [32, 1], we convert the input image I into object features $V = (v_1, \dots, v_n)$ with $v_i \in R^D$, where n is the number of detected objects, and D is the feature dimension. The question (w_1, \dots, w_m) is a sequence of m words. It can

be encoded using LSTM to $Q = (q_1, \dots, q_m)$, where $q_i = \text{LSTM}(w_i)$ and $q_i \in R^C$.

BAN [18] is introduced to reduce both input channels simultaneously and obtain a unified representation of question features Q and image features V . It first calculates a bilinear attention map $G \in R^{m \times n}$ between Q and V , conditioned on which to generate the joint embedding z as follows:

$$z = \text{BAN}(Q, V, G). \quad (1)$$

The attention map G is defined as:

$$G = \text{softmax}\left(\left((\mathbf{1} \cdot \mathbf{p}^\top) \circ \sigma(Q^\top \mathbf{U}')\right) \sigma(\mathbf{V}'^\top V)\right), \quad (2)$$

where $\mathbf{U}' \in R^{C \times K'}$, $\mathbf{V}' \in R^{D \times K'}$, $\mathbf{p} \in R^{K'}$ are variables to be learned, $\mathbf{1} \in R^m$ is a vector with all elements equal to 1, K' denotes the shared embedding size, σ is the sigmoid activation function, and \circ is Hadamard product (element-wise multiplication). Then the k -th element value of joint embedding is given by:

$$z_k = \sigma(Q^\top \mathbf{U})_k^\top G \sigma(V^\top \mathbf{V})_k, \quad (3)$$

where $\mathbf{U} \in R^{C \times K}$, $\mathbf{V} \in R^{D \times K}$ are the parameters to be optimized, $(Q^\top \mathbf{U})_k \in R^m$ is the k column of $Q^\top \mathbf{U}$, and $(V^\top \mathbf{V})_k \in R^n$ is the k column of $V^\top \mathbf{V}$.

After that, we input z to a classifier such as MLP to calculate the score p_i for answer $a_i \in A$ and choose the highest one as the predicted answer, where A is the answer set.

4. Graph Reasoning Networks

Given the calculation of z_k , Eq.(1) can be rewritten as:

$$Z'^\top = \text{BAN}'(Q, V, G) = \sigma(Q^\top \mathbf{U}) \circ G \sigma(V^\top \mathbf{V}), \quad (4)$$

$$z = Z' \cdot \mathbf{1}, \quad (5)$$

where $Z' \in R^{K \times m}$ and $\mathbf{1} \in R^m$, which has a close connection with graph attention networks [36, 37]. Given the matrix of outputs in [36, 37] as:

$$\text{Attention}(\mathbb{Q}, \mathbb{K}, \mathbb{V}) = \text{softmax}(\mathbb{Q}\mathbb{K}^\top)\mathbb{V}, \quad (6)$$

where \mathbb{Q} , \mathbb{K} and \mathbb{V} denote the query, keys and values respectively, we can easily illustrate Eq.(4) from the perspective of graph. The attention map G in Eq.(4) is equivalent to graph $\text{softmax}(\mathbb{Q}\mathbb{K}^\top)$ in Eq.(6), and $\sigma(V^\top \mathbf{V})$ is the value \mathbb{V} . Looking into the definition of attention map in Eq.(2), the map G implies how much information should flow from the nodes V to nodes Q . $(\mathbf{1} \cdot \mathbf{p}^\top) \circ \sigma(Q^\top \mathbf{U}')$ and $\sigma(V^\top \mathbf{V})$ correspond to query \mathbb{Q} and key \mathbb{K} in Eq.(6) respectively. Considering that, Eq.(6) only takes single type of inputs, while VQA models need to consider the multi-modal inputs (i.e. image and question). An additional Hadamard product of $\sigma(Q^\top \mathbf{U})$ and $G\sigma(V^\top \mathbf{V})$ is thus included in Eq.(4)

to generate output nodes (z'_1, \dots, z'_m) , where $z'_i \in R^K$. Finally, the joint embedding z represents the whole graph by summarizing of all nodes in Z' in Eq.(5).

Eqs.(4) and (5) provide an elegant approach to investigate the relationship between question features Q and image features V . However, a simply summarization over columns of Z' in Eq.(5) cannot fully address the connections between joint embeddings (z'_1, \dots, z'_m) corresponding to words. Given the question and image in Figure 1, BAN (i.e. Eqs.(4) and (5)) can locate a variety of fruits in the image, but it is confused about the relative position of each fruit from others by mixing all the information (i.e. the summarization in Eq.(5)). Hence, we are motivated to develop graph reasoning networks, as shown in Figure 1, which has two graphs for each layer, i.e. inter-graph and intra-graph. The inter-graph learns to build the relationship between words and objects and generates their joint embeddings, while the intra-graph will update joint embeddings by exploiting their interactions. And we find that the right answer may not be decided at once, therefore we stack our graphs to make the words interact with the image as well as words themselves for multiple times.

4.1. Inter-Graph

The major target of inter-graph is to locate the objects related to semantic information of each word in the question. Beginning with Eq.(4), we have a multi-layer multi-glimpse extension.

The i -th layer of inter-graph takes $o_{i-1} \in R^{C \times m}$, which are the output nodes from the previous layer, and image features $V \in R^{D \times n}$ as its input nodes and builds the graph $G_i^e \in R^{m \times n \times g^e}$ following Eq.(2) between o_{i-1} and V , where g^e is the number of glimpses and the j -th graph attention is computed as:

$$G_{i,j}^e = \text{softmax}\left(\left((\mathbf{1} \cdot \mathbf{p}_{i,j}^{e\top}) \circ \sigma(o_{i-1}^\top \mathbf{U}_{i,j}^{e\top})\right) \sigma(V^\top \mathbf{V}_{i,j}^{e\top})\right). \quad (7)$$

Layer i has its own $\mathbf{U}_{i,j}^{e\top}$, $\mathbf{V}_{i,j}^{e\top}$ and $\mathbf{p}_{i,j}^e$, and the parameters $\mathbf{U}_{i,j}^{e\top}$ and $\mathbf{V}_{i,j}^{e\top}$ are shared among glimpses except for $\mathbf{p}_{i,j}^e$. After learning the graph attention, we use the Eq.(4) to generate the joint embeddings as:

$$\begin{aligned} h_{i,j}^\top &= \text{BAN}_{i,j}^{e\top}(o_{i-1}, V, G_{i,j}^e) \\ &= \sigma(o_{i-1}^\top \mathbf{U}_{i,j}^{e\top}) \circ G_{i,j}^e \sigma(V^\top \mathbf{V}_{i,j}^{e\top}). \end{aligned} \quad (8)$$

Instead of concatenation or summarization of joint embeddings from each glimpse, we follow BAN to use the residual form to integrate previous learned joint embeddings, then Eq.(8) becomes:

$$h_{i,j} = W_{i,j}^e \text{BAN}_{i,j}^{e\top}(h_{i,j-1}, V, G_{i,j}^e)^\top + h_{i,j-1}, \quad (9)$$

where $h_{i,0} = o_{i-1}$, and $W_{i,j}^e \in R^{C \times K}$ projects the joint embeddings to the same dimension of o_{i-1} . As we mentioned above, the first layer of inter-graph takes Q as o_0 to

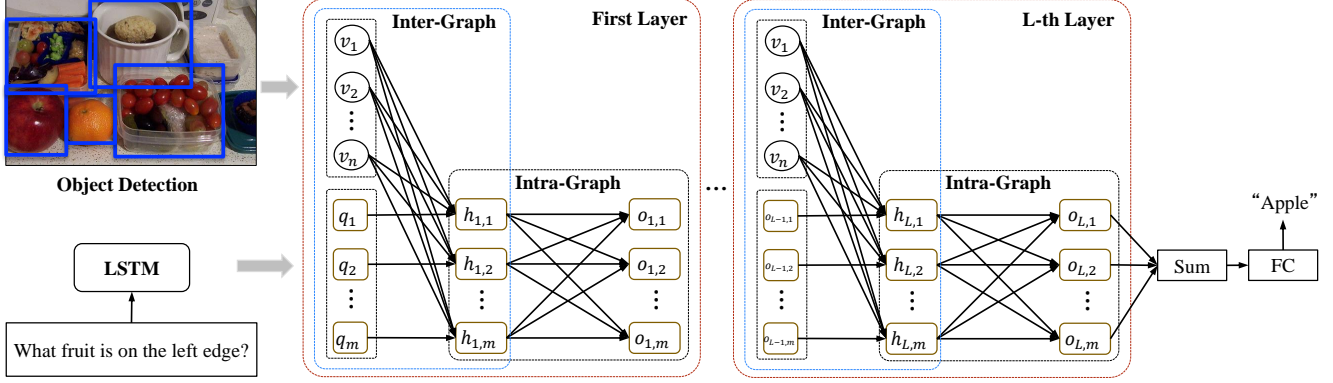


Figure 1. Architecture of our model. The inter-graph builds the relationship between words and objects, and the intra-graph learns the relationship between joint embeddings in term of words. The two graphs cooperate with each other to predict the answer.

locate queried objects, and higher layer takes the outputs of previous layer’s graph as o_{i-1} to involve more visual information related to prior knowledge. The final joint embeddings $h_{i,..,g^e}$ is rewritten as h_i for abbreviation and handled to intra-graph for further processing.

4.2. Intra-Graph

Instead of summarizing outputs from the inter-graph in Eq.(5), the intra-graph exchanges information between each node in h_i , which can learn context from other nodes.

We use the bilinear attention map to build the graph. The graph map $G_i^r \in R^{m \times m \times g^r}$ at layer i is computed following Eq.(7) with g^r glimpses, based on which, nodes of the intra-graph at glimpse j gather information from others and are represented as Eq.(9):

$$o_{i,..,j} = W_{i,j}^r \text{BAN}_{i,j}^r(o_{i,..,j-1}, h_i, G_{i,j}^r)^\top + o_{i,..,j-1}, \quad (10)$$

where $W_{i,j}^r \in R^{C \times K}$ and $o_{i,..,0} = h_i$. The outputs of intra-graph o_i , abbreviated version of $o_{i,..,g^r}$, can be utilized to answer the question or regarded as textual inputs for inter-graph of next layer to view the image again to involve more clues for complex reasoning.

After stacking L layers of the inter-graph and the intra-graph, we summarize all the nodes of o_L over channel dimension to represent the whole graph and pass it to a two-layer MLP for classification:

$$p = W^{a'} \sigma(W^a o_L \cdot \mathbb{1}), \quad (11)$$

where $W^a \in R^{C \times C}$, $W^{a'} \in R^{|\Sigma| \times C}$, and $|\Sigma|$ is the size of answer vocabulary.

5. Experiments

In this section, we evaluate our graph reasoning networks on both GQA v1.1 [15] and VQA 2.0 dataset [2, 10]. We first introduce the two datasets and then describe our implementation details and results, and finally the qualitative analysis.

5.1. Dataset

VQA v2.0 dataset: The dataset was built based on the MSCOCO images [24], which contains 1.1M questions asked by humans and ten human annotated answers for each question. Compared with v1.0 dataset, it emphasizes the visual understanding by reducing the text bias. The answers of training and validation dataset are published for training model, while those of test-dev, test-standard and test-challenge dataset are unknown. We train our model with different settings on training dataset and evaluate accuracy on validation dataset by the tools from [2]. We pick the best model to train it on training and validation dataset with extra data from Visual Genome [22], reporting results on test-server.

GQA dataset: However, the VQA dataset pays less attention to reasoning, since 19.5% of its questions have relations, 8% have spatial reasoning questions and only 3% have compositional questions. The GQA dataset makes much effort on generating questions that need multi-step reasoning and balancing the answer distributions to overcome the question-condition biases. It takes four steps to construct the datasets, i.e. cleaning and consolidating the scene graphs linked to each image in Visual Genome, traversing the graphs to gather information about objects and relations and generating diverse set of questions, reducing biases in answer distribution to get a balanced dataset, and providing functional representations for these questions. About 94% of its questions need multi-step reasons, and 51% query the relationship between two objects. The dataset has 22M questions over 110K images, which makes the original training and validation dataset very large. Thus these questions are re-sampled based on the answer distribution into a balanced training dataset and a balanced validation dataset. Besides the standard accuracy metric, four new metrics give insight into different capabilities of models: 1) consistency, the answer for the current question should not contradict with answers of previous ques-

Model	Binary	Open	Consistency	Plausibility	Validity	Distribution	Accuracy
LSTM [2]	61.90	22.69	68.68	87.30	96.39	17.93	41.07
LSTM-CNN [2]	63.26	31.80	74.57	84.25	96.02	7.46	46.55
BottomUp [1]	66.64	34.83	78.71	84.57	96.18	5.98	49.70
MAC [14]	71.23	38.91	81.59	84.48	96.16	5.34	54.06
Human	91.20	87.40	98.40	97.20	98.90	0.00	89.30
GRNs (ours)	74.93	41.24	87.41	84.68	96.14	5.76	57.04

Table 1. Accuracy of our single model on GQA test2019 dataset, it is trained on balanced training dataset and balanced validation dataset of GQA v1.1 with object features from <https://cs.stanford.edu/people/doradad/gqa/download.html>.

Model	Overall	Yes/no	Number	Other	Test-std
MCB [9]	61.96	78.41	38.81	53.23	62.27
Bottom-Up [1]	65.32	81.82	44.21	56.05	65.67
Counter [45]	68.09	83.14	51.62	58.97	68.41
MFH+Bottom-Up [43]	68.76	84.27	49.56	59.89	-
BAN [18]	69.52	85.31	50.93	60.26	-
BAN+Glove [18]	69.66	85.46	50.66	60.60	-
BAN+Glove+Counter [18]	70.04	85.42	54.04	60.52	70.35
GRNs+Glove (ours)	70.97	87.03	53.56	61.18	71.12
GRNs+Glove+Counter (ours)	71.00	86.62	56.31	60.95	-

Table 2. Accuracy of single model on VQA 2.0 test-dev and test-standard dataset, it is trained on training, validation splits and Visual Genome dataset.

tions based on the same image; 2) validity and plausibility, whether the given answer is in the scope of question and reasonable; 3) distributions, overall matching distribution between true answers and given answers using Chi-square distribution [23]; Of these metrics, a higher score is better for consistency, validity, plausibility, but a lower score is better for distribution. We adjust our model on the balanced training dataset and report results on the balanced validation dataset. Then we train our model on the balanced training dataset and the balanced validation dataset and test results on the test-server.

5.2. Implementation Details

For VQA dataset, we construct the answer vocabulary by restricting to words that appears in training and validation dataset more than 8 times, resulting in $|\Sigma| = 3,129$. We then truncate or pad the length of question m to 15 words. The input dimension of LSTM is 600, 300 of which is learned by our model and another 300 is pre-trained using GloVe vector [30], and the output dimension C is 1024. We get object features from BottomUp [1] with $D = 2048$, and the object number n is fixed to 100. The joint embedding size K is set to 1024 and $K' = K \times 3$ to increase the capacity of attention. In order to save memory in each layer to stack our network for multiple times, we shrink the glimpse number from 8 in BAN [18] to $g^e = g^r = 4$. Weight Normalization [33] and Dropout [34] ($p = 0.2$) are added after each linear mapping to stable the output and prevent from over-fitting. Due to the fact that there might exist multiple correct answers for a question, the binary cross entropy loss

L is calculated as:

$$L = - \sum_{i=1}^{|\Sigma|} (y_i \log \sigma(p_i) + (1 - y_i) \log(1 - \sigma(p_i))), \quad (12)$$

where $y_i = \min(\frac{\text{humans that provided answer } a_i}{3}, 1)$. Adamax [20], a variant of Adam, is used to optimize our model. The initial learning rate is 0.001 and grows by 0.001 every epoch until reaching 0.004 for warm start, keeps constant until the eleventh epoch and decays by 1/4 every two epochs to 0.00025. The batch size is 128.

We follow the same settings as VQA for the GQA dataset, except constructing answer vocabulary on the balanced training dataset and the balanced validation dataset making $|\Sigma| = 1567$, extending the question length m to 22 to fit the long description of complex questions, using the object features downloaded from the official website since BottomUp was trained on images contained in the GQA validation set and thus may give false improvement in scores, object number n varying from 36 to 100, and changing to softmax cross entropy loss because there is only one right answer for each question:

$$L = -y^\top \log \text{softmax}(p), \quad (13)$$

where $y_i = 1$ if answer a_i is the right one, otherwise $y_i = 0$.

5.3. Comparison with State-of-the-Art

In Table 1, we compare our graph reasoning networks (GRNs) with others on the GQA dataset: **LSTM** predicts answer only based on the question; **LSTM-CNN** linearly combines global image features and question features to get the answer; **BottomUp** is the winner of VQA Challenge 2017, which is the first to use detected object features instead of grid features; **MAC** is state-of-the-art of CLEVER dataset [17], which decomposes problems into a series of successively inferred reasoning steps to accomplish the final task. Our method significantly outperforms others, increasing the overall accuracy by 7.3% compared with BottomUp and 3.0% compared with MAC. In the detailed metrics, our model boosts the consistency from 81.59% to 87.54%, which demonstrates that our method predicts the answer with better understanding of the meaning of input image

Model	Model Size	Accuracy
BAN-4 \times 1	44.8M	61.95
BAN-4 \times 2	79.4M	62.60
BAN-4 \times 3	115.8M	61.98
Inter \times 1	32.9M	61.88
Inter \times 1 + Intra \times 1	51.8M	63.50
Inter \times 1 + Intra \times 2	70.7M	63.80
Inter \times 1 + Intra \times 3	89.6M	63.60
(Inter + Intra) \times 2	96.9M	64.07
(Inter + Intra) \times 3	142.1M	64.22

Table 3. Accuracy on the balanced validation dataset of GQA. BAN-4 represents the BAN model with four glimpses, Inter for inter-graph and Intra for intra-graph. The number after multiplication operation denotes the layer number of that module.

than other entries, not simply based on learned question bias or random guessing.

Then, we evaluate our model on VQA 2.0 test-dev dataset, it also achieves state-of-the-art. As shown in Table 2, overall accuracy of our *GRNs+Glove* model is 1.31% higher than *BAN+Glove*, nearly 3.0% on number metric. It can be explained that the counting task is a kind of relation among objects, which tries to find similar objects in latter layers with objects grounded by previous layers. And the extra counter module [45] in our *GRNs+Glove+Counter* model makes little gain on overall accuracy, since it might increase the counting ability but disturb our reasoning graphs leading drop in other metrics. Thus, we choose *GRNs+Glove* as our best single model to evaluate it on the test-standard dataset as its interpretability.

5.4. Ablation Study

We conduct several ablation studies to verify the contribution of each module in our graph reasoning networks. The first three lines in Table 3 show the accuracy of *BAN* on the balanced validation dataset of GQA. It can be seen that simply stacking the module of *BAN* can improve the accuracy to some extent (0.65% in the two-layer model) compared with the one-layer model, but the performance drops back in the three-layer model. Although the two-layer *BAN* model might gain more object information related to the global representation in Eq.(5) without exchanging context information, it is not clear about the relationship between entities in question. In contrast, our one-layer model (Inter \times 1 + Intra \times 1) gains an accuracy 1.62% higher than one-layer inter-graph (Inter \times 1) and 1.55% higher than one-layer *BAN*, proving the effect of our proposed intra-graph. Note that the accuracy of our one-layer inter-graph is comparable to that of one-layer *BAN*, though a summarization of nodes in Eqs.(4) and (5) can achieve decent performance, there is rare exchange between these nodes. However, if

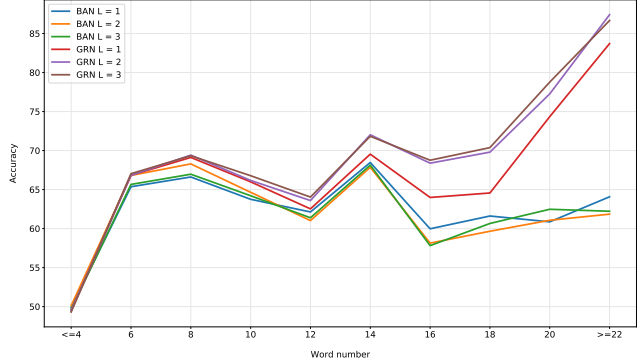


Figure 2. Accuracy on varied lengths of questions from GQA balanced validation dataset. The accuracy on y -axis is evaluated on questions with word number of x -axis.

Model	Layer	VQA Score
BAN-4	1	65.81
BAN-4	2	66.16
BAN-4	3	66.27
GRNs	1	66.43
GRNs	2	66.87
GRNs	3	67.21
GRNs	4	67.06

Table 4. Score on VQA 2.0 validation dataset. GRNs are one-layer inter-graph following by one-layer intra-graph.

we only stack the intra-graph for multiple times (Inter \times 1 + Intra \times 2 and Inter \times 1 + Intra \times 3), the performance remains still, this might be caused by that the intra-graph can only propagate the information already learned by inter-graph but cannot involve more factual information in image required to answer the questions. With stacking three layers of our graphs ((Inter + Intra) \times 3), our model achieves 64.22% on accuracy, which is chosen as the best model.

Furthermore, we investigate the accuracy curves on questions with varied lengths to show the ability of our model on multi-step reasoning in Figure 2. Due to different question numbers in each length, there is variation on the curves. Our GRNs (with layer = 1,2,3) outperform BANs (with layer = 1,2,3) in simple questions whose lengths are less than or equal to 14 words. As the questions become more complex, the accuracy of *BAN* models drops nearly by 10%, which shows their weakness on multi-step reasoning. Even though the accuracy of our model with one-layer also drops by 5%, the models with multi-layers remedy this loss and perform stably on these questions.

However, the questions of GQA are generated by following some patterns, which makes them easy to be recognized, thus we also conduct the experiments on the VQA 2.0 datasets, whose questions are naturally expressed, and it has larger answer vocabulary. In Table 4, our one-layer model also achieves higher score than *BAN* with multi-layers, proving its reasoning ability on varieties of ques-

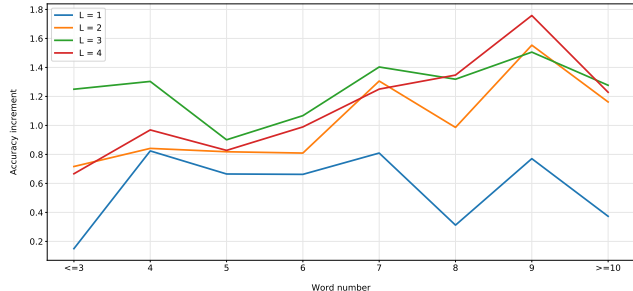


Figure 3. Score increment of our models (with layer=1,2,3,4) compared with one-layer BAN model on VQA 2.0 validation dataset.

tions. The score of our model increases gradually with the growth of layer and reaches the peak at 67.21% in the three-layer model.

What interests us is why the performance drops at four-layer, thus we show the absolute increment of score of our models compared with single-layer *BAN* to amplify the discrepancy of different layers in Figure 3. The one-layer model does not perform as well as other three models for long questions due to its shallow graphs. With more layers, our GRNs become better at long questions and achieve 1.7% increase at word number of nine. Comparing our three-layer model and four-layer model, the former one works better in short questions (word number < 8) which take 79% of all questions, while the latter one has a higher score in long questions, this may explain the performance drop. This also inspires us to design a network in the future to classify the questions to fit different layers of graphs.

5.5. Qualitative Analysis

To visualize the effects of each module in our graph reasoning networks, we present the learned attention maps of the inter-graph and the intra-graph in each layer to show how the networks work. Given the question ‘what fruit is on the left edge?’ in Figure 4, the inter-graph of the first layer attends kinds of objects in the input image, while the intra-graph broadcasts learned fruit information to other words and chooses ‘tomato’ as the answer, probably because the amount of ‘tomato’ is the biggest among detected fruits. The inter-graph of the second-layer picks ‘orange’ that is to the left of ‘tomato’, and the intra-graph keeps collecting ‘fruit’ and ‘edge’ information. In the third layer, the inter-graph locates ‘apple’ that is on the left edge, and every word in the intra-graph pays its attention to the ‘edge’ information to predict the answer.

In Figure 5, we shows the answer predicted by *BAN* and our models with one layer, two layers and three layers. In the first image of the top row, *BAN* cannot correctly answer the question because the entities of ‘young girl’ and ‘bag’ learn their positions respectively, but they do not know each other’s information, while our proposed intra-graph exchanges such positional information to make it possible

to compare the relative direction of the two entities. A similar question can also be found in first image of the bottom row, our model approaches the correct answer step by step as the layer of the graph increases. Moreover, our model can find the implicit relationship between objects, even when the sheep are far away from the dog in the second image of the top row, as well as abstract scene in the second image (five circles represents Olympics) and third image (many trees composes forest) of the bottom row. Furthermore, our model finely discriminate the highly overlapped objects, such as two sheep in the second image and the rope in the fourth image of the top row, it is possibly because the intra-graph undertakes some burden from original graph of *BAN*, which results in that the inter-graph spares more energy on learning details in the image.

6. Conclusions

In this paper, we develop graph reasoning networks composed of layers of inter-graph and intra-graph, the inter-graph learns the relationship between words in the question and objects in the image and generate the joint embeddings of them, while the intra-graph gets the relationship between these joint embeddings in term of words to exchange context information, which is first taken into consideration by our model for VQA problem. By stacking our graphs, the compositional questions that involve relation between multiple entities can be better understood and correctly answered based on the image. Our method achieves state-of-the-art performance on both GQA v1.1 and VQA v2.0 test server, and the ablation studies show that our networks significantly outperform *BAN* on a variety of questions, particularly on the long and complex ones.

References

- [1] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, volume 3, page 6, 2018. 1, 2, 5
- [2] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 1, 2, 4, 5
- [3] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer, 2007. 2
- [4] Y. Chen, M. Rohrbach, Z. Yan, S. Yan, J. Feng, and Y. Kalanidis. Graph-based global reasoning networks. *arXiv preprint arXiv:1811.12814*, 2018. 2
- [5] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014. 2
- [6] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. Moura, D. Parikh, and D. Batra. Visual dialog. In *Proceed-*

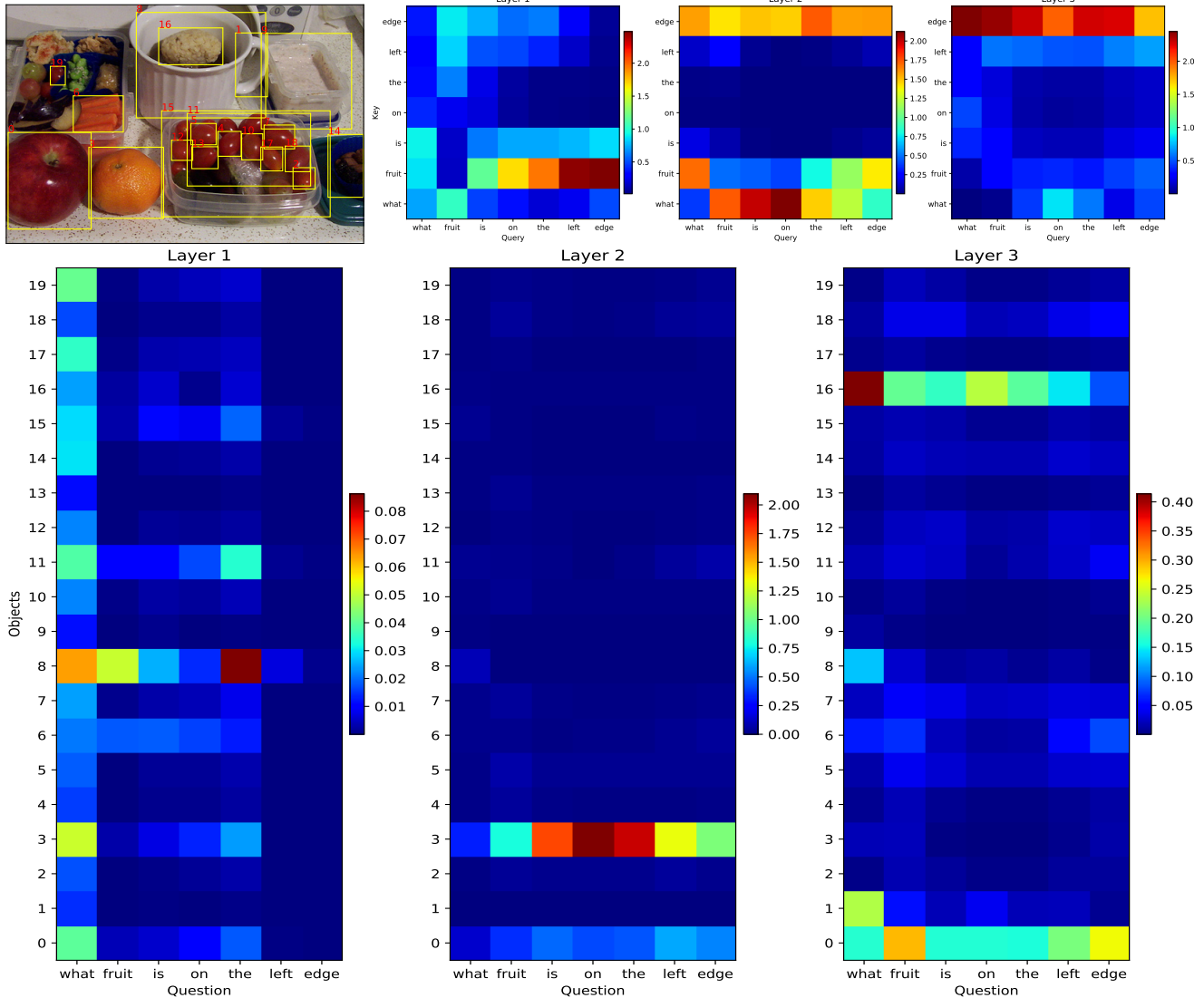


Figure 4. Visualization of attention maps for our networks. The attention maps for each graph are summed at each layer to briefly show the attended objects and words. The first image at top shows the bounding boxes of detected objects, and others for graph attentions between words. The images at bottom show the graph attentions between words and objects. The predicted answers are tomato, apple and orange respectively for one-layer, two-layer and three-layer model of our graph reasoning networks.

- ings of the *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, 2017. 1
- [7] H. De Vries, F. Strub, J. Mary, H. Larochelle, O. Pietquin, and A. C. Courville. Modulating early visual processing by language. In *Advances in Neural Information Processing Systems*, pages 6594–6604, 2017. 2
- [8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2
- [9] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*, 2016. 1, 5
- [10] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913, 2017. 2, 4
- [11] W. Hamilton, Z. Ying, and J. Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, pages 1024–1034, 2017. 2
- [12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 2
- [13] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 1, 2
- [14] D. A. Hudson and C. D. Manning. Compositional at-



Figure 5. Examples illustrate the answers predicted by BAN and our graph models. BAN, L1, L2, L3 denote the answers predicted by BAN, one-layer, two-layer and three-layer of our model respectively.

- tention networks for machine reasoning. *arXiv preprint arXiv:1803.03067*, 2018. **5**
- [15] D. A. Hudson and C. D. Manning. Gqa: a new dataset for compositional question answering over real-world images. *arXiv preprint arXiv:1902.09506*, 2019. **2, 4**
- [16] A. Jabri, A. Joulin, and L. van der Maaten. Revisiting visual question answering baselines. In *European conference on computer vision*, pages 727–739. Springer, 2016. **2**
- [17] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. Lawrence Zitnick, and R. Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2901–2910, 2017. **5**
- [18] J.-H. Kim, J. Jun, and B.-T. Zhang. Bilinear attention networks. *arXiv preprint arXiv:1805.07932*, 2018. **1, 2, 3, 5**
- [19] J.-H. Kim, K.-W. On, W. Lim, J. Kim, J.-W. Ha, and B.-T. Zhang. Hadamard product for low-rank bilinear pooling. *arXiv preprint arXiv:1610.04325*, 2016. **1, 2**
- [20] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. **5**
- [21] T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016. **2**
- [22] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al. Visual genome: Connecting language and vision using crowd-sourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017. **4**
- [23] H. O. Lancaster and E. Seneta. Chi-square distribution. *Encyclopedia of biostatistics*, 2, 2005. **5**
- [24] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. **4**
- [25] H. Liu and P. Singh. Conceptnet practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226, 2004. **2**
- [26] H. Nam, J.-W. Ha, and J. Kim. Dual attention networks for multimodal reasoning and matching. *arXiv preprint arXiv:1611.00471*, 2016. **1, 2**
- [27] M. Narasimhan, S. Lazebnik, and A. Schwing. Out of the box: Reasoning with graph convolution nets for factual visual question answering. In *Advances in Neural Information Processing Systems*, pages 2659–2670, 2018. **2**
- [28] M. Narasimhan and A. G. Schwing. Straight to the facts: Learning knowledge base retrieval for factual visual question answering. *arXiv preprint arXiv:1809.01124*, 2018. **2**
- [29] W. Norcliffe-Brown, E. Vafeais, and S. Parisot. Learning conditioned graph structures for interpretable visual question answering. *arXiv preprint arXiv:1806.07243*, 2018. **2**
- [30] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. **5**
- [31] E. Perez, H. De Vries, F. Strub, V. Dumoulin, and A. Courville. Learning visual reasoning without strong priors. *arXiv preprint arXiv:1707.03017*, 2017. **2**

- [32] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. [2](#)
- [33] T. Salimans and D. P. Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. In *Advances in Neural Information Processing Systems*, pages 901–909, 2016. [5](#)
- [34] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014. [5](#)
- [35] N. Tandon, G. De Melo, and G. Weikum. Acquiring comparative commonsense knowledge from the web. In *AAAI*, pages 166–172, 2014. [2](#)
- [36] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017. [2](#), [3](#)
- [37] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017. [2](#), [3](#)
- [38] P. Wang, Q. Wu, C. Shen, A. Dick, and A. van den Hengel. Fvqa: Fact-based visual question answering. *IEEE transactions on pattern analysis and machine intelligence*, 2017. [2](#)
- [39] X. Wang, R. Girshick, A. Gupta, and K. He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018. [2](#)
- [40] Z. Yang, B. Dhingra, K. He, W. W. Cohen, R. Salakhutdinov, Y. LeCun, et al. Glomo: Unsupervisedly learned relational graphs as transferable representations. *arXiv preprint arXiv:1806.05662*, 2018. [2](#)
- [41] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola. Stacked attention networks for image question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 21–29, 2016. [1](#), [2](#)
- [42] Z. Yu, J. Yu, J. Fan, and D. Tao. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. *IEEE International Conference on Computer Vision (ICCV)*, pages 1839–1848, 2017. [1](#), [2](#)
- [43] Z. Yu, J. Yu, C. Xiang, J. Fan, and D. 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. [5](#)
- [44] Z. Yu, J. Yu, C. Xiang, Z. Zhao, Q. Tian, and D. Tao. Rethinking diversified and discriminative proposal generation for visual grounding. *arXiv preprint arXiv:1805.03508*, 2018. [1](#)
- [45] Y. Zhang, J. Hare, and A. Prügel-Bennett. Learning to count objects in natural images for visual question answering. *arXiv preprint arXiv:1802.05766*, 2018. [5](#), [6](#)