Visual Question Answering with Graph Matching and Reasoning

Anonymous ICCV submission

Paper ID ****

Abstract

Visual Question Answering (VQA) is of great significance in offering people convenience: one can raise a question for details of objects, or high-level understanding about the scene, over an image. This paper proposes a novel method to address the VQA problem. In contrast to prior works, our method that targets single scene VQA, replies on graph-based techniques and involves reasoning. In a nutshell, our approach is centered on three graphs. The first graph, referred to as inference graph G_I, is constructed via learning over labeled data. The other two graphs, referred to as query graph Q and entity-attribute graph G_{EA}, are generated from natural language query Q_{nl} and image Img, that are issued from users, respectively. As G_{EA} often does not take sufficient information to answer Q, we develop techniques to infer missing information of G_{EA} with G_I. Based on G_{EA} and Q, we provide techniques to find matches of Q in G_{EA}, as the answer of Q_{nl} in Img. Unlike commonly used VQA methods that are based on end-to-end neural networks, our graph-based method shows well-designed reasoning capability, and thus is highly interpretable. We also create a dataset on soccer match (Soccer-VQA) with rich annotations. The experimental results show that our approach outperforms the state-of-the-art method and has high potential for future investigation.

1. Introduction

In recent years, visual query answering (VQA) has received significant attention [16, 21, 9] as it involves multidisciplinary research, *e.g.* natural language understanding, visual information retrieving and multi-modal reasoning. The task of VQA is to find an answer to a query Q_{nl} based on the content of an image. There are a variety of applications of VQA, *e.g.* surveillance video understanding, visual commentator robot, *etc*. Solving VQA problems usually requires high level reasoning from the content of an image.

Example 1: ADD AN EXAMPLE! □

This example suggests that we leverage graph-based method to resolve the VQA problem. While to do this, several questions have to be settled. (1) How to represent image and query with graphs? (2) How to infer crucial information when G_{EA} constructed from image is insufficient? (3) How to find answers from graphs with G_{EA}?

Contributions. In contrast to a majority of deep learning based VQA techniques, which lacks of necessary reasoning and thus performs poorly, our approach divides VQA tasks into three parts, and incorporates reinforcement learning and reasoning for each subtask. The main contributions of the paper are as follow.

(1) We propose new approaches for visual tasks based on reinforcement learning. Given an image and a question, our approach only identifies those objects that are related to users' questions rather than the complete set of objects along with their attributes. This substantially improves performance of object detection

(2) We propose approaches to answering visual questions with graph-based techniques. More specifically, we first construct an entity-attribute graph from a given image; we then train a classifier to infer missing information that are crucial for answering queries; we finally provide methods to answer queries with graph pattern matching.

3) We conduct extensive experimental studies to verify the performance of our method. We find that X, Y, and Z.

2. Related Work

We categorize related work into following three parts.

Visual query answering. Current VQA approaches are mainly based on deep neural works. [34] introduces a spatial attention mechanism similar to the model for image captioning. Instead of computing the attention vector iteratively, [30] obtains a global spatial attention weights vector which is then used to generate a new image embedding. [33] proposed to model the visual attention as a multivariate distribution over a grid-structured conditional random field on image regions, thus multiple regions can be selected at the same time. This attention mechanism is called structured multivariate attention in [33]. There has been many other improvements to the standard deep learning method, *e.g.* [8] utilized Multimodal Compact Bilinear (MCB) pooling to efficiently and expressively combine multimodal features. Another interesting idea is the implementation of Neural Module Networks [1, 11], which decomposes queries into their linguistic substructures, and uses these structures to dynamically instantiate module networks. [24] proposed to build graph over scene objects and question words. The visual graph is similar to ours, but the query graph differs. Note that the method [24] proposed is still a neural network based method as the structured representations are fed into a recurrent network to form the final embedding and the answer is again inferred by a classifier.

Environment Exploration in Visual Field. Reinforcement driven information acquisition is wildly applied in traditional vision domain, like visual object detection [17], face detection [10] and image classification [18]. In visual and language domain, relevant work like [23] achieves image captioning with Embedding Reward. Reinforcement learning preserves ability to effectively select preferred actions, which benefits the system decomposing the problem into a few sub-tasks.

Graph-based query answering. Query answering has been extensively studied for graph data. In a nutshell, this work includes two aspects: query understanding, and query evaluation. We next review previous work on two aspects.

(1) Queries expressed with natural languages are very user-friendly, but nontrivial to understand. Typically, they need to be structured before issuing over *e.g.* search engine, knowledge graph, since structured queries are more expressive. There exist a host of works that based on query logs, human interaction and neural network, respectively. [20] leverages query logs to train a classifier, based on which structured queries are generated. [32] propose an approach to generate the structured queries through talking between the data (*i.e.* the knowledge graph) and the user. [31] introduced how to generate a core inferential chain from a query with convolutional neural networks. As we only cope with a set of fixed queries, hence, we defer the topic of query understanding to another paper, and focus primarily on the query evaluation.

(2) To evaluate queries on graphs, a typical method is graph pattern matching. There has been a host of work on graph pattern matching, *e.g.* techniques for finding exact matches [5, 26], inexact matches [35, 25], and evaluating SPARQL queries on RDF data [27]. Our work differs from the prior work in the following: (1) we integrate arithmetical and set operations in the query graph, and (2) we develop technique to infer missing values for query answering.

3. Overview of the Approach

We start from representations of images and questions, followed by the overview of our approach.

3.1. Representation of Images and Questions

We use the same representations as [29]. To make the paper self-contained, we cite them as follows (rephrased).

Entity-Attribute Graphs. Entities are typically defined as objects or concepts that exist in the real world. An entity often carries attributes, that describe features of the entity.

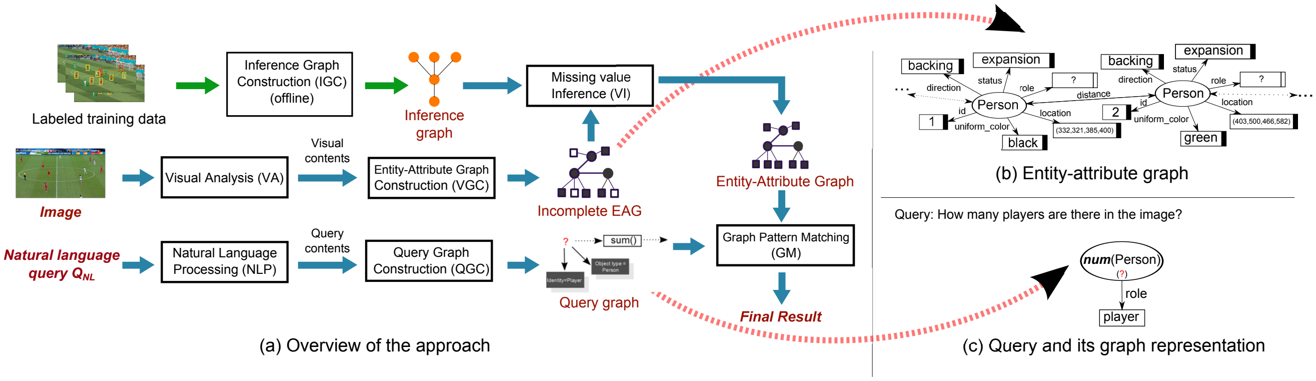
Assume a set $\mathcal{E}$ of entities, a set $\mathcal{D}$ of values, a set $\mathcal{P}$ of predicates indicating attributes of entities and a set Θ of types. Each entity e in $\mathcal{E}$ has a *unique ID* and a *type* in Θ.

An *entity-attribute* graph, denoted as EAG, is a set of triples $t = (s, p, o)$, where *subject* s is an entity in $\mathcal{E}$, p is a *predicate* in $\mathcal{P}$, and *object* o is either an entity in $\mathcal{E}$ or a value d in $\mathcal{D}$. It can be represented as a directed edge-labeled graph $G_{EA} = (V, E)$, such that (a) V is the set of nodes consisting of s and o for each triple $t = (s, p, o)$; and (b) there is an edge in E from s to o labeled by p for each triple $t = (s, p, o)$.

An image can be represented as an EAG with detected objects and obvious attributes. This can be achieved via a few visual tasks. While EAG generated directly after image processing is often incomplete, *i.e.* it may miss some crucial information to answer queries. We hence refer to *entity-attribute graphs* with incomplete information as *incomplete entity-attribute graphs*, and associate nodes with white rectangles, to indicate the missing value of an entity or attribute in EAG. Figure **??**(b) is an *incomplete entity-attribute graph*, in which square nodes representing person roles are associated with white rectangle.

Query Graphs. A query graph $Q(u_o)$ is a set of triples (s_Q, p_Q, o_Q), where s_Q is either a variable z or a function $f(z)$ taking z as parameter, o_Q is one of a value d or z or $f(z)$, and p_Q is a predicate in $\mathcal{P}$. Here function $f(z)$ is defined by users, and variable z has one of three forms: (a) *entity variable* y, to map to an entity, (b) *value variable* $y*$, to map to a value, and (c) *wildcard* $\_y$, to map to an entity. Here s_Q can be either y or $\_y$, while o_Q can be y, $y*$ or $\_y$. Entity variables and wildcard carry a *type*, denoting the type of entities they represent.

A query graph can also be represented as a graph such that two variables are represented as the same node if they have the same name of y, $y*$ or $\_y$; similarly for functions $f(z)$ and values d. We assume *w.l.o.g.* that $Q(x)$ is connected, *i.e.* there exists an undirected path between u_o and each node in $Q(u_o)$. In particular, u_o is a designated node in $Q(u_o)$, denoting the query focus and labeled by "?". Take Fig. 1(c) as example. It depicts a query graph that is generated from query "*How many players are there in the im-*



(a) Overview of the approach

(b) Entity-attribute graph

(c) Query and its graph representation

Figure 1: Overview of our approach, Entity Attribute Graph and Queries

age?". Note that the "query focus" u_o carries a function num () that calculates the total number of *person* entities with *role* "player".

3.2. Approach Overview

Along the same lines as representations for images and questions, and graph pattern matching for question answering, raised in [29], we propose a comprehensive approach as modeling of the VQA problem.

Figure 1(a) presents the overview of our approach. As can be seen, our approach revolves around three graphs: entity-attribute graph, query graph and inference graph. The generation of entity-attribute graph G_{EA} follows three steps. Module VA conducts the first step, *i.e.* image processing, and outputs all the detected objects along with their attributes. Using visual contents produced in step one, module VGA constructs an *incomplete* EAG. In the last step, module VI takes inference graph and *incomplete* EAG as inputs, infer missing information with G_I, and outputs an updated EAG for query answering. The inference graph G_I is used to infer missing values of an *incomplete* EAG. and constructed by module IGC over training data. As is query-independent, G_I is constructed offline, which warrants the efficiency of our approach. As the other part of input, natural language query Q_{nl} needs to be structured for query evaluation. To this end, Q_{NL} is first parsed via our NLP module, and then structured by module QGC. After $Q(u_o)$ and G_{EA} are generated, our approach employs module GM for matching computation, and returns final result.

As some modules employ existing techniques, to emphasize our novelty, we will elaborate modules VA and VGA in Section **??**, modules IGC and VI in Section **??**, and module GM in Section **??** with more details.

4. Query Oriented Visual Tasks

In this section, we introduce how we do visual tasks that are in connection with questions.

4.1. Visual Processing

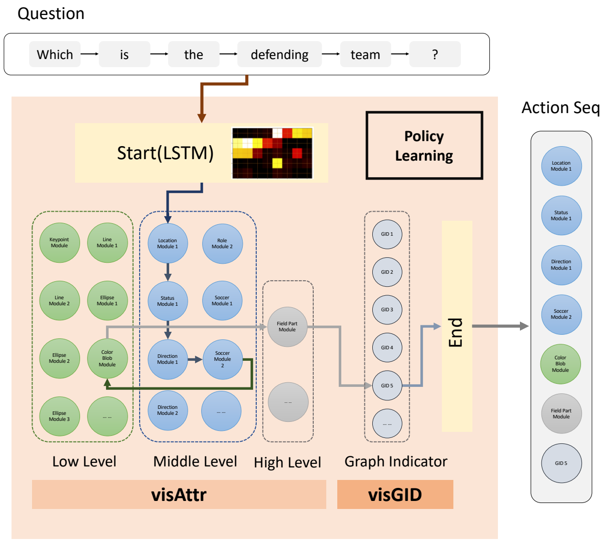


Figure 2: Overview of our approach

In our approach, we build a structure which selects subtasks to form a policy which is based on queries. For instance, with the input *which is the defending team?*, the system first predicts the corresponding visual action sequence, which are *Human Module*, *Gesture Module*, *Direction Module*, *Soccer Module*, *Color Blob Module*, *Field Part Module* and *Graph Indicator*. Guided by such sequence, the image features are then extracted by operating relevant vision tasks. An overview is shown in Figure 2.

4.1.1 Multi-layer LSTM with Attention

The task here is to predict the most suitable action modules sequence a by given questions Q and preference pre. We form the problem of seeking effective answering strategy of question Q and preference pre as a sequence-to-sequence learning problem with attention mechanism. Inspired by [3], we input word feature of questions w_i^q, $i \in \|Q\|$ into a LSTM network which is regarded as an encoder and output h_i as the hidden state for ith word in the question. By adding soft attention, the context vector c_i is calculated by the following equations.

$$c_i = \sum\nolimits_{j=1}^{\|Q\|} a_{ij} h_{ij} \tag{1}$$

$$a_{ij} = \frac{exp(e_{ij})}{\sum_{k=1}^{\|Q\|} exp(e_{ik})} \tag{2}$$

$$e_{ij} = a(s_{j-1}, h_i) \tag{3}$$

where h_i and s_j are hidden states of encoder and decoder stage, respectively. Here a_{ij} is the attention weights, with higher a_{ij} in (i, j) pair, the more attention will pay in this correlation, thus the jth output action a_i module will be more influenced by the ith input word w_i^q in question. The decode part is similar as traditional recurrent neural networks (RNN). The following steps demonstrates decoding to get the joint distribution of action module sequence $\mathsf{A} = [\mathsf{a}_1, \ldots, \mathsf{a}_t]$.

$$p(\mathsf{A}|Q) = \Pi_{t \in \|\mathsf{A}\|}\, p(\mathsf{a}_t | \{\mathsf{a}_1, \ldots, \mathsf{a}_t\}, c_i, Q) \tag{4}$$

$$p(\mathsf{a}_t | \{\mathsf{a}_1, \ldots, \mathsf{a}_t\}, c_i, Q) = g(\mathsf{a}_{t-1}, s_t, c_i, Q) \tag{5}$$

$$s_t = f(s_{t-1}, \mathsf{a}_{t-1}, c_i) \tag{6}$$

where $g(\cdot)$ is a nonlinear function which outputs the probability of action module $\mathsf{a_t}$. The probability distribution $p(\mathsf{A}|Q)$ is used to predict a maximum probability action module sequence by beam search during testing time.

Guided by this action sequence $[\mathsf{a}_1, \mathsf{a}_2, \mathsf{a}_3, \ldots, \mathsf{a}_n]$, actions are selected from the following visual task pool, and comes into next session, visual task selection (VTS).

4.1.2 Visual Task Selection

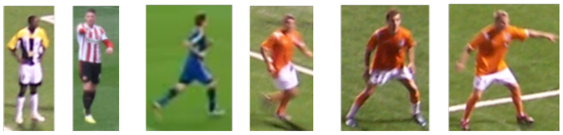
VTS is guided by the question feature, selecting target visual tasks from the visual task pool by Monte Carlo learning. The task pool is demonstrated in Table 1.

The pool is constructed by two parts, the first one is *visAttr* which aims to discover the attribute of people, soccer, field and scene [29], while the other one *visGId* is an indicator showing the current question belongs to which graph type. There are three levels of *visAttr*: low, middle and high, which represents different difficulty degree of the vision tasks.

For each vision task, the approach is not fixed, one task can be achieved by different methods with variance in time and accuracy. For instance, object detection based methods, like Faster R-CNN [22], R-FCN [6], SSD [14] or skeleton keypoints detection based method like [4] and [28] can be implemented as a set of status module methods, because all of them is able to localize and distinguish people who are moving, standing or with an expansion gesture (Figure 3). Thus, the system is not only able to determine whether resembles a vision task module into action module sequence $[\mathsf{a}_1, \mathsf{a}_2, \mathsf{a}_3, \ldots, \mathsf{a}_n]$, it can also select one specific approach under a vision task module, based on question and preference. For the preference, it is clarified in the next section.

visAttr	Level	Sub-task	Descriptions
	Low	Keypoint Module	To get detailed information of the soccer field $F_{keypoint}$.
		Line Module	
		Ellipse Module	
		Color Blob Module	To detect blobs of different colors among a region(whole soccer field or a small bounding box).
	Middle	Location Module	To detect the location of the person $P_{location}$.
		Status Module	To get the person's gesture P_{status} (standing, moving, expansion).
		Direction Module	To get whether a person is facing the goal or not $P_{direction}$.
		Uniform Module	To get the uniform color of the person $P_{uniform}$.
		Soccer Module	To detect the location of the soccer $S_{location}$.
	High	Field Part Module	To detect which part of the soccer field is there in the image F_{part} .
visGID		Graph ID Indicator	To indicate which type of graph will be used in the following process.

Table 1: Visual Task Pool



(a) Standing (b) Moving (c) Expansion

Figure 3: Person status.

4.1.3 Time and Accuracy Term

To better balance the accuracy and inference time for a given application, we proposed time and accuracy terms in loss function during training process.

$$L_{\tau\alpha}(\theta) = \ell(\theta, \mathsf{A}|I, Q) + \gamma \sum_{i \in \|\mathsf{A}\|} \alpha(\mathsf{a_i}) + (1-\gamma) \sum_{i \in \|\mathsf{A}\|} \tau(\mathsf{a_i}) \tag{7}$$

where $\tau(\cdot)$ and $\alpha(\cdot)$represent the pre-tested inference time and inference accuracy, action module sequence A samples from joint distribution $p(\mathsf{A}|Q)$, and here $\ell(\cdot)$ is the softmax loss over the predict score. For the preference term γ, it ranges from 0 to 1, which represents the preference over time and accuracy.

4.1.4 Monte Carlo Methods

The task now becomes a policy learning problem. Given a question and preference, output a policy containing a sequence of actions $[\mathsf{a}_1, \mathsf{a}_2, \mathsf{a}_3, \ldots, \mathsf{a}_n]$. There is no ground truth for each steps, but only a final reward indicates that whether the answer is correct based on current policy. We involve the concept of Monte Carlo Methods to learn the policy which guides the vision tasks, and such policy network requires an extra reward value in loss.

$$L_{policy}(\theta) = \sum\nolimits_{i \in \|A\|} log \pi(\mathsf{a}_i|Q, \theta)\ell(Q, \mathsf{A}) \tag{8}$$

where $\mathsf{a_i}$ is the action will take, based on current status. $\pi(\cdot)$ is the policy function that maps status to actions, here, the policy is the probability of outputing next action module $\mathsf{a_i}$ based on current status. And $\ell(\cdot)$ here is the softmax loss based on the whole action module sequence $[\mathsf{a}_1, \mathsf{a}_2, \mathsf{a}_3, \ldots, \mathsf{a}_n]$. Since all actions are discrete, which leads to non-differentiable, and back-propagation cannot be used. Policy gradient [13] is used here during training. The object function now becomes the combination of policy gradient loss $L_{policy}(\theta)$ with the time-accuracy-balanced loss $L_{\tau\alpha}$, and optimize it by backpropagation for $L_{\tau\alpha}$, while policy gradient for $L_{policy}(\theta)$.

4.2. Construction of EAG

After objects that are related to questions are identified, we construct a graph structure, denoted as EAG, along the same line as [29].

Example 2: ADD AN EXAMPLE TO ILLUSTRATE PROGRESS IF NECESSARY! □

5. Reasoning

According to our observation, an *incomplete* EAG isn't well satisfying of answering the query because of the insufficient attributes. To infer the hidden attributes, an inference graph is constructed accordingly. we briefly introduce the construction below.

5.1. Construction of Inference Graph

To take advantage of the prior information and increase the generalization ability of the proposed model, our inference graph is constructed using Bayesian network. Mathematically, Bayesian network [7] can be described by a pair $\mathfrak{B} =< \mathcal{G}, \Theta_\mathcal{G} >$. Here, the notation $\mathcal{G}$ is a directed acyclic graph, of which the i-th vertex corresponds to a random variable X_i, and the edge between two connected vertexes indicates the dependency. Additionally, the second item $\Theta_\mathcal{G}$ is a set of parameters used to quantify the dependencies in $\mathcal{G}$. Denoted by $\text{Pa}(X_i)$ the attributes of the parents of X_i, the parameter of X_i is represented by $\theta_{X_i|\text{Pa}(X_i)} = P_\mathfrak{B}(X_i|\text{Pa}(X_i))$. With the notations above, the joint probability distribution of Bayesian network is given by:

$$P_\mathfrak{B}(X_1, \cdots, X_n) = \prod_{i=1}^{n} P_\mathfrak{B}(X_i|\text{Pa}(X_i)) = \prod_{i=1}^{n} \theta_{X_i|\text{Pa}(X_i)} \tag{9}$$

In our inference graph, the role of Bayesian network is to predict the object class when given the attributes $\{X_i\}_{i=1}^n$ as input. In the sense of probability, the object class is also a variable [12]. Defined by $X_0 = Y$ the class variable, the network now has one extra vertex X_0. In order to infer the class attribute, and according to the Bayesian rule, the target network is given by:

$$\begin{aligned} P_\mathfrak{B}(Y|X) &= \frac{P_\mathfrak{B}(Y)P_\mathfrak{B}(X|Y)}{P_\mathfrak{B}(X)} \\ &= \frac{\theta_{Y|\text{Pa}(X_0)} \prod_{i=1}^n \theta_{X_i|Y,\text{Pa}(X_i)}}{\sum_{y' \in \mathcal{Y}} \theta_{y'|\text{Pa}(X_0)} \prod_{i=1}^n \theta_{X_i|y',\text{Pa}(X_i)}} \end{aligned} \tag{10}$$

where $\mathcal{Y}$ is the set of classes.

5.2. Structure Learning of the Inference Graph

In the context of Naïve Bayes, the structure of $P_\mathfrak{B}(Y|X)$ is simplified by taking the class variable as the root, and all attributes are conditionally independent when taking the class as a condition [19]. As a consequence, the attribute class can be explicitly inferred by:

$$P_\mathfrak{B}(Y|X) = c \cdot \theta_Y \prod_{i=1}^{n} \theta_{X_i|Y} \tag{11}$$

where c is a constant that makes the calculation being a distribution: $c = \sum_{y' \in \mathcal{Y}} \theta_{y'} \prod_{i=1}^n \theta_{X_i|y'}$.

Note from Eq.(11) that Naïve Bayes simplifies the complexity of Bayesian network. As can be validated by the experimental results, the simple model works excellently to our problem.

6. Experimental Studies

In this section, we conducted two sets of experiments to evaluate (1) the performance of our visual processing module, (2) the accuracy of our inference module, and (3) the overall performance of our approach.

Experimental Setting.

DataSet. We used two datasets: (1) Soccer dataset that we annotated; and (2) X dataset from []. We extracted images with subjects of golf and tennis (report Statistics about the dataset). We split Soccer (resp. X) data into two parts:*I* (one third) and *II* (two thirds), and used *II* as training data, and *I* as testing data.

Queries. We used two sets of questions: (1) the set of questions given in Table 2 for Soccer dataset; and (2) another set of questions listed in Table **??** for X dataset.

(We enlarge the training question scale from 7 into 28, so learning correlation between question and answer does not work at this time. For question details, please refer questionset.txt.rtf.)

Id	Question	Difficulty
Q_{nl_1}	Who is holding the soccer?	Easy
Q_{nl_2}	What is the uniform color of the referee?	Easy
Q_{nl_3}	Is there any referee in the image?	Easy
Q_{nl_4}	Which team does the goalkeeper belong to?	Medium
Q_{nl_5}	Who is the defending team?	Medium
Q_{nl_6}	Which part of the field are the players being now?	Hard
Q_{nl_7}	How many players are there in the image?	Hard
Q_{nl_8}	Is this image about corner kick? (If not, just list the correct one.)	??
Q_{nl_9}	Is this image about penalty kick? (If not, just list the correct one.)	??
$Q_{nl_{10}}$	Is this image about kick off? (If not, just list the correct one.)	??

Table 2: A set of questions

6.1. Performance of Visual Task Selecting Policy

To test the validity of reinforcement learning of selecting visual task modules, we test the inference time over accuracy with the state-of-art [2] [15] whichi is shown in Figure **??**.

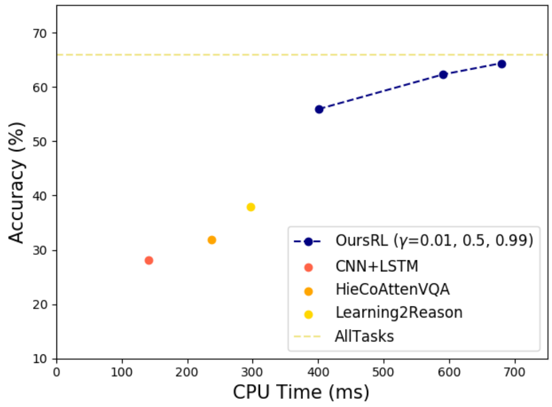
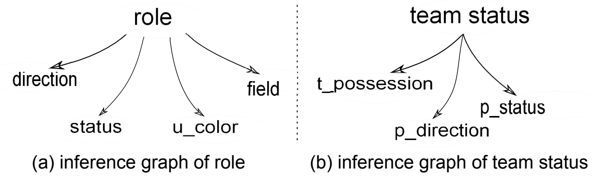


Figure 4: Inference Time and Accuracy



(a) inference graph of role (b) inference graph of team status

Figure 5: Inference Graphs

Here to test the generalization, we enlarge the training set by more various question with same meaning. For instance, the original question of Q_{nl5} is *"Who is the defending team?"*, we add three more similar question asking *Who is attacking team?*, *"What is the uniform color of the defending team?"* and *"What is the uniform color of the attacking team?"*. Unlike state-of-art methods answering questions in [29], adding generalization and variation in question would not dramatically change the performance, it is because the structure is not fixed, all the visual task selection is query oriented. For [11], even though the network is not fixed, the answering part is based on neural network, and essentially it also learns the statically correlation, which leads to the weakness in logical reasoning.

6.2. Effectiveness of Inference

Accuracy of Role. Only report results with Reinforcement learning

Accuracy of Team-Status. Only report results with Reinforcement learning

Accuracy of Kick-Off. SHOW INFERENCE GRAPH AND RESULT TABLE!

Accuracy of Penalty Kick. SHOW INFERENCE GRAPH AND RESULT TABLE!

Accuracy of Corner Kick. SHOW INFERENCE GRAPH AND RESULT TABLE!

Accuracy of Attacking Free Kick. SHOW INFERENCE GRAPH AND RESULT TABLE!

Accuracy of Balls. Over new dataset. SHOW INFERENCE GRAPH AND RESULT TABLE!

6.3. Overall Performance

We use the same question setting as [29], compared the following state-of-the-art methods: [2],and [11] with our method (γ=0.99), the overall performance is shown in Table 3.

	CNN+LSTM	HieCoAtten	Learn2Reason	Ours
Q_{nl1}	44.23	43.62	31.12	64.16
Q_{nl2}	71.31	77.66	9.4	47.43
Q_{nl3}	74.58	83.78	83.21	70.02
Q_{nl4}	40.48	39.29	51.92	62.14
Q_{nl5}	49.19	49.90	30.78	93.33
Q_{nl6}	20.56	18.70	30.0	62.13
Q_{nl7}	11.08	12.63	36.69	47.45
Avg.	46.40	49.11	51.08	65.97

Table 3: Accuracy comparison per query (%)

References

[1] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein. Neural module networks. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2016. 2

[2] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*, 2015. 6

[3] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. May 2016. 4

[4] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 4

[5] L. P. Cordella, P. Foggia, C. Sansone, and M. Vento. A (sub) graph isomorphism algorithm for matching large graphs. *TPAMI*, 26(10):1367–1372, 2004. 2

[6] J. Dai, Y. Li, K. He, and J. Sun. R-FCN: object detection via region-based fully convolutional networks. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 379–387, 2016. 4

[7] N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. *Machine learning*, 29(2-3):131–163, 1997. 5

[8] 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. 2

[9] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, and W. Xu. Are you talking to a machine? dataset and methods for multilingual image question. In *Advances in neural information processing systems*, pages 2296–2304, 2015. 1

[10] B. Goodrich and I. Arel. Reinforcement learning based visual attention with application to face detection. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Providence, RI, USA, June 16-21, 2012*, pages 19–24, 2012. 2

[11] R. Hu, J. Andreas, M. Rohrbach, T. Darrell, and K. Saenko. Learning to reason: End-to-end module networks for visual question answering. *CoRR, abs/1704.05526*, 3, 2017. 2, 6

[12] D. Koller, N. Friedman, and F. Bach. *Probabilistic graphical models: principles and techniques*. MIT press, 2009. 5

[13] S. Liu, Z. Zhu, N. Ye, S. Guadarrama, and K. Murphy. Improved image captioning via policy gradient optimization of spider. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 5

[14] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C. Fu, and A. C. Berg. SSD: single shot multibox detector. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I*, pages 21–37, 2016. 4

[15] J. Lu, J. Yang, D. Batra, and D. Parikh. Hierarchical question-image co-attention for visual question answering, 2016. 6

[16] M. Malinowski, M. Rohrbach, and M. Fritz. Ask your neurons: A neural-based approach to answering questions about images. In *Proceedings of the IEEE international conference on computer vision*, pages 1–9, 2015. 1

[17] S. Mathe, A. Pirinen, and C. Sminchisescu. Reinforcement learning for visual object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2894–2902, 2016. 2

[18] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu. Recurrent models of visual attention. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2204–2212, 2014. 2

[19] F. Petitjean, W. Buntine, G. I. Webb, and N. Zaidi. Accurate parameter estimation for bayesian network classifiers using hierarchical dirichlet processes. *Machine Learning*, 107(8-10):1303–1331, 2018. 5

[20] J. Pound, A. K. Hudek, I. F. Ilyas, and G. E. Weddell. Interpreting keyword queries over web knowledge bases. In *CIKM*, pages 305–314, 2012. 2

[21] M. Ren, R. Kiros, and R. Zemel. Image question answering: A visual semantic embedding model and a new dataset. *Proc. Advances in Neural Inf. Process. Syst*, 1(2):5, 2015. 1

[22] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, pages 91–99, Cambridge, MA, USA, 2015. MIT Press. 4

[23] Z. Ren, X. Wang, N. Zhang, X. Lv, and L. Li. Deep reinforcement learning-based image captioning with embedding reward. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1151–1159, 2017. 2

[24] D. Teney, L. Liu, and A. van den Hengel. Graph-structured representations for visual question answering. *arXiv preprint*, 2017. 2

[25] Y. Tian and J. M. Patel. TALE: A tool for approximate large graph matching. In *ICDE*, pages 963–972, 2008. 2

[26] J. R. Ullmann. An algorithm for subgraph isomorphism. *JACM*, 23(1):31–42, 1976. 2

[27] A. Wagner, D. T. Tran, G. Ladwig, A. Harth, and R. Studer. Top-k linked data query processing. In *ESWC*, pages 56–71, 2012. 2

[28] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *CVPR*, 2016. 4

[29] P. Xiong, H. Zhan, X. Wang, B. Sinha, and Y. Wu. Visual query answering by entity-attribute graph matching and reasoning. In *CVPR*, 2019. 2, 3, 4, 5, 6

[30] H. Xu and K. Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *European Conference on Computer Vision*, pages 451–466. Springer, 2016. 1

[31] W. Yih, M. Chang, X. He, and J. Gao. Semantic parsing via staged query graph generation: Question answering with

knowledge base. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, pages 1321–1331, 2015. 2

[32] W. Zheng, H. Cheng, L. Zou, J. X. Yu, and K. Zhao. Natural language question/answering: Let users talk with the knowledge graph. In *CIKM*, pages 217–226, 2017. 2

[33] C. Zhu, Y. Zhao, S. Huang, K. Tu, and Y. Ma. Structured attentions for visual question answering. In *Proc. IEEE Int. Conf. Comp. Vis*, volume 3, 2017. 1

[34] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4995–5004, 2016. 1

[35] L. Zou, L. Chen, and M. T. Özsu. Distancejoin: Pattern match query in a large graph database. *PVLDB*, 2(1):886–897, 2009. 2