Dynamic Language Binding in Relational Visual Reasoning

Thao Minh Le , **Vuong Le** , **Svetha Venkatesh** and **Truyen Tran**

Applied Artificial Intelligence Institute, Deakin University, Australia

{lethao,vuong.le,svetha.venkatesh,truyen.tran}@deakin.edu.au

Abstract

We present Language-binding Object Graph Network, the first neural reasoning method with dynamic relational structures across both visual and textual domains with applications in visual question answering. Relaxing the common assumption made by current models that the object predicates pre-exist and stay static, passive to the reasoning process, we propose that these dynamic predicates expand across the domain borders to include pairwise visual-linguistic object binding. In our method, these contextualized object links are actively found within each recurrent reasoning step without relying on external predicative priors. These dynamic structures reflect the conditional dual-domain object dependency given the evolving context of the reasoning through co-attention. Such discovered dynamic graphs facilitate multi-step knowledge combination and refinements that iteratively deduce the compact representation of the final answer. The effectiveness of this model is demonstrated on image question answering demonstrating favorable performance on major VQA datasets. Our method outperforms other methods in sophisticated question-answering tasks wherein multiple object relations are involved. The graph structure effectively assists the progress of training, and therefore the network learns efficiently compared to other reasoning models.

1 Introduction

Reasoning is crucial for intelligent agents wherein relevant clues from a knowledge source are retrieved and combined to solve a query, such as answering questions about an image. Human visual reasoning involves analyzing linguistic aspects of the query and continuously inter-linking them with visual objects through a series of information aggregation steps [Lake *et al.*, 2017]. Artificial reasoning engines mimic this ability by using structured representations (e.g. scene graphs) [Shi *et al.*, 2019] to discover categorical and relational information about visual objects.

In this work, we address two key abstractions: How can we extend this structure seamlessly across both visual-lingual borders? And, unlike prior work, how can we extend these
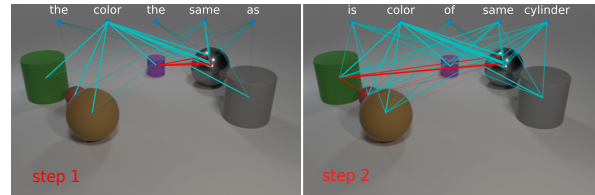

Figure 1: We aim to dynamically construct visual graphs (red edges) and linguistic-visual bindings (cyan edges (most prominent words shown)) adaptively to reasoning steps for each image-question pair.

structures to be dynamic and responsive to the reasoning process? We explore the dynamic relational structures of visual scenes that are proactively discovered within reasoning context and their adaptive connections to the components of a linguistic query to effectively answer visual questions.

Recent history observes the success of compositional reasoning which iteratively pays attention to a subset of clues in the query and simultaneously looks up a corresponding subset of facts from a static unstructured knowledge source to construct a representation related to the answer [Hudson and Manning, 2018]. Concurrently, findings in visual relational modeling show that the information in visual scenes is significantly distributed at the interconnections between semantic factors of visual objects and linguistic objects from both the image and query [Baradel *et al.*, 2018]. These observations suggest that relational structures can improve compositional reasoning [Xu *et al.*, 2020]. However, direct application of attention mechanisms on a static structuralized knowledge source [Veličković *et al.*, 2018] would miss the full advantage of compositionality. Moreover, object relations are naturally rich and multifaceted [Kim *et al.*, 2018], therefore an *a priori* defined set of semantic predicates such as visual scene graphs [Hudson and Manning, 2019a] and language grounding [Huang *et al.*, 2019] are either incomplete [Xu *et al.*, 2017], or too complicated and irrelevant to use without further pruning.

We approach this dilemma by dynamically constructing relevant object connections on-demand according to the evolving reasoning states. There are two types of connections: links that relate visual objects and links that bind visual objects in the image to linguistic counterparts in the query (See Fig. 1). Conceptually, this dynamic structure constitutes a relational working memory that temporarily links and refines concepts
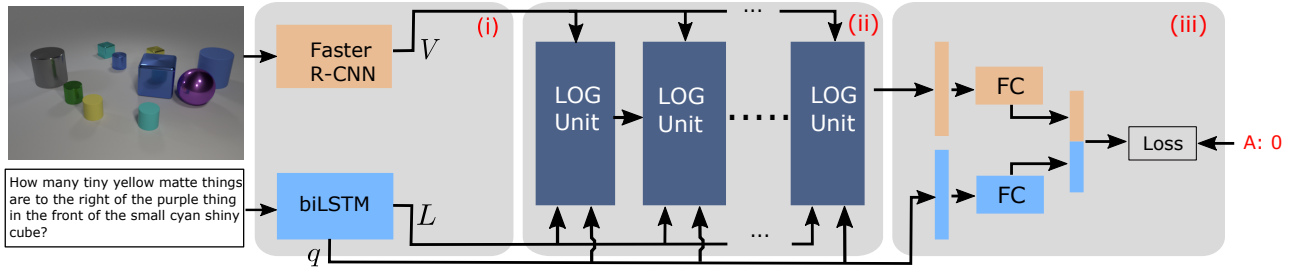
Figure 2: Overall Architecture of LOGNet. (i) Linguistic and visual representations (ii) Information refinement with LOG modules (iii) Multimodal fusion and answer prediction.

both within and across modalities. These relations are compact and readily support structural inference.

Our model, called Language-binding Object Graph Networks (LOGNet) for visual question answering (VQA), includes an iterative operation of LOG unit that uses a contextualized co-attention to identify pairs of visual objects that are temporally related. Another co-attention head is concurrently used to provide cross-domain binding between visual concepts and linguistic clues. A progressive chain of dynamic graphs is inferred by our model (see Fig. 1). These dynamic structures enable representation refinement with residual graph convolution iterations. The refined information will be added to an internal working memory progressing toward predicting the answer. The modules are interconnected through co-attention signals making the model end-to-end differentiable.

We apply our model on major VQA datasets. Both qualitative and quantitative results indicate that LOGNet has advantages over state-of-the-art methods in answering long and complex questions. Our results show superior performance even when trained on just 10% of data. These questions require complex high-order reasoning which necessitates our model's ability to dynamically couple entities to build a predicate, and then chain these predicates in the correct order. The structured representation provides guidance to the reasoning process, improving the fitness of the learning particularly with limited training data.

2 Related Work

Recent compositional reasoning research aims at either structured symbolic program execution using custom built modules [Hu *et al.*, 2017] or working through recurrent implicit reasoning steps on an unstructured representation [Perez *et al.*, 2018]. Relational structures have been demonstrated to be crucial for reasoning [Xu *et al.*, 2020]. End-to-end relational modeling considers pair-wise predicates of CNN features [Santoro *et al.*, 2017]. With reliable object detection, visual reasoning can use semantic objects as cleaner representations [Anderson *et al.*, 2018; Desta *et al.*, 2018]. When semantic or geometrical predicate labels are available, either as provided [Hudson and Manning, 2019b] or by learning [Xu *et al.*, 2017] to form semantic scene graphs, such structures can be leveraged for visual reasoning [Shi *et al.*, 2019; Li *et al.*, 2019]. In contrast to these methods, our relational graphs are not limited by the predefined predicates but liberally form them according to the reasoning context. Our model

is also different from previous question-conditioned graph construction [Norcliffe-Brown *et al.*, 2018] in the dynamic nature of the multiform graphs where only relations that are relevant emerge. Dynamic graph modeling has been considered by recurrent modeling [Palm *et al.*, 2018], and although their states transform, the graph structures stay fixed. A related idea uses language conditioned message passing to extract context-aware features [Hu *et al.*, 2019]. In contrast, LOGNet does not treat linguistic cues as a single conditioning vector, but allows them to live as a set of active objects that interact with visual objects through binding and individually contribute to the joint representation. The language binding also differentiates LOGNet from MUREL [Cadene *et al.*, 2019] where the contributions of linguistic cues to visual objects are the same though an expensive bilinear operator.

3 Language-binding Object Graph Network

The goal of a VQA task is to deduce an answer $\tilde{a}$ from an image I in response to a natural question q. Let the answer space be $\mathbb{A}$, VQA is formulated as:

$$\tilde{a} = \arg\max_{a \in \mathbb{A}} \mathcal{P}_\theta\left(a \mid q, I\right), \qquad (1)$$

where, θ is the learnable parameters of $\mathcal{P}$.

We envision VQA as a process of relational reasoning over a scene of multiple visual objects conditioned on a set of linguistic cueing objects. Crucially, a pair of co-appearing visual objects may induce multiple relations, whose nature may be unknown *a priori*, and hence must be inferred dynamically in adaptive interaction with the linguistic cues.

We present a new neural model $\mathcal{P}$ called LOGNet (See Fig. 2) to realize this vision. At the high level, for each image and query pair, LOGNet first normalizes them into two individual sets of linguistic and visual objects. Then, it performs iterative multi-step reasoning by iteratively summoning Language-binding Object Graph (LOG) units to achieve a compact multi-modal representation in a recurrent manner. This representation is finally combined with the query representation to reach the answers. We detail these steps.

3.1 Linguistic and Visual Objects

We embed words in the length-S query into 300-D vectors, which are subsequently passed through a biLSTM. The hidden states of LSTM representing the context-dependent word embeddings e_s are collected into a chain of contextual embeddings $L = \{e_s\}_{s=1}^{S} \in \mathbb{R}^{d \times S}$ and used as linguistic objects
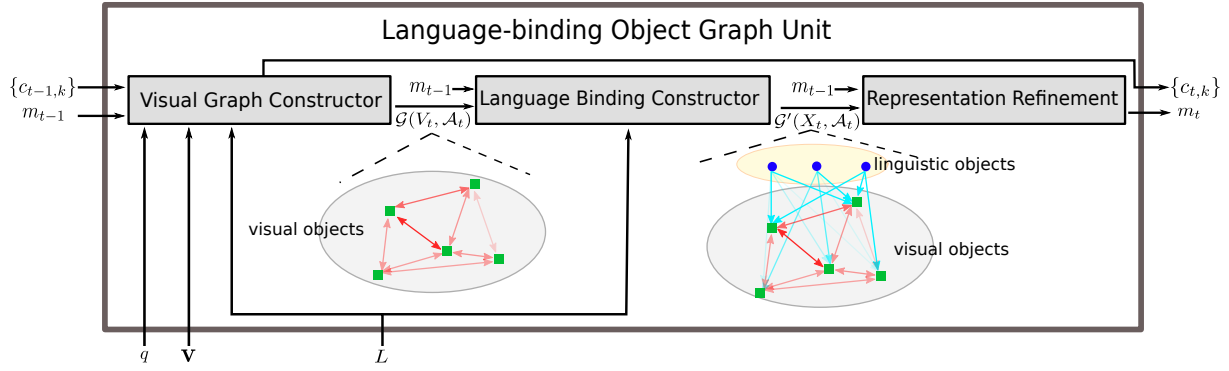
Figure 3: Language-binding Object Graph (LOG) Unit. L: linguistic objects, V: visual objects, red edges: visual graph, cyan edges: language-visual binding. The following elements are dynamic at pass t: q_t – query semantic ; $\{c_{t,k}\}$ – language-based controlling signals; m_t - working memory state.

in reasoning. We also retain the overall query semantic as $q = [\overleftarrow{e_1}; \overrightarrow{e_S}]$ which joins the final states of forward and backward LSTM passes. Unless otherwise specified, we use $[.;.]$ to denote the concatenation operator of two tensors.

The input image I is first processed into a set of appearance/spatial features $O = \{(a_i, p_i)\}_{i=1}^N$ of N regions extracted by an off-the-shelf object detection such as Faster R-CNN [Ren *et al.*, 2015]. The appearance component $a_i \in \mathbb{R}^{2048}$ are ROI pooling features and the spatial p_i are normalized coordinates of the region box [Yu *et al.*, 2017]. These features are further combined and projected by trainable linear embeddings to produce a set of visual objects $V = \{v_i\}_{i=1}^N \in \mathbb{R}^{d \times N}$. The pair (L, V) are readily used as input for a chain of LOG reasoning operations.

3.2 Language-binding Object Graph Unit
LOG is essentially a recurrent unit whose state is kept in a compact working memory m_t and a controlling signal c_t. Input of each LOG operation includes the visual and linguistic objects (V, L), and the overall query semantic q.

Each LOG consists of three submodules: (i) a *visual graph constructor* to build a context-aware weighted adjacency matrix of visual graph $\mathcal{G}_t$, (ii) a *language binding constructor* to compute the adaptive linkage between linguistic and visual objects and form a multi-modal graph $\mathcal{G}'_t$ (iii) *representation refinement* module to update object representation using the graphs. (See Fig. 3).

Visual Graph Constructor
At each LOG operation, we construct an undirected graph $\mathcal{G}_t = (V_t, \mathcal{A}_t)$ from N visual objects $V = \{v_i\}_{i=1}^N$ by finding adaptive features V_t and constructing the weighted adjacency matrix $\mathcal{A}_t$. Different from the widely used static semantic graphs [Xu *et al.*, 2017], our graph $\mathcal{G}_t$ is dynamically constructed at each reasoning step t^{th} and is modulated by the recurrent controlling signal c_t and overall linguistic cue q. This reflects the dynamic relations of objects triggered by both the question and reasoning context. In fact, this design is consistent with how human reasons. For example, looking at an image, to answer different questions, we connect different pairs of objects although their geometrical and appearance similarities were unchanged. Moreover, even at one

question, our mind traverses through multiple types of object relationships in different steps of reasoning, especially when a query contains multiple or nested relations. Let W_t denote sub-networks' weights at step t^{th}, we first augment the nodes' features as

$$V_t = W_t^v [V; m_{t-1} \odot V] + b^v. \tag{2}$$

The controlling signals $\{c_{t,k}\}$ is derived from its previous state and a step-specific query semantic q_t through a set of K attention heads $\{\alpha_{t,k}\}_{k=1}^K$ on the linguistic objects $L = \{e_s\}_{s=1}^S$:

$$c_1 = q_1, \qquad q_t = W_t^q q + b_t^q \tag{3}$$

$$q'_t = [q_t; \sum_{k=1}^K (\gamma_{t,k} * c_{t-1,k})], \qquad \sum_{k=1}^K \gamma_{t,k} = 1, \tag{4}$$

$$\alpha_{s,t,k} = \text{softmax}_s \left(W_{t,k}^\alpha (e_s \odot q'_t) \right), \tag{5}$$

$$c_{t,k} = \sum_{s=1}^S \alpha_{s,t,k} * e_s, \quad c_t = \{c_{t,k}\}, \tag{6}$$

where, $\gamma_{t,k}$ is the weights of the past controlling signals being added to the current query semantic q'_t.

While single attention can be used to guide the multi-step reasoning process [Hudson and Manning, 2018], we noticed that it tends to focus on one object attribute at a time neglecting inter-aspect relations because of the softmax operation. In VQA, multiple object attributes are usually necessary - e.g. to answer "what is the color of the small shiny object having the same shape with the cyan sphere?", the object aspects "color" and "shape" both need to be attended to. Our development of using multi-head attention enables such a goal. The controlling signals are then used to build the context modulated node description matrix of r rows, $\tilde{V}_t \in \mathbb{R}^{r \times N}$:

$$\tilde{V}_t = \text{norm} \left(W_t^{\tilde{v}} \sum_{k=1}^K (V \odot c_{t,k}) \right), \tag{7}$$

where, $norm$ is a normalization function for numerical stabilization which is softmax in our implementation.

Finally, we estimate the symmetric adjacency matrix $\mathcal{A}_t \in \mathbb{R}^{N \times N}$ by relating node features in $\tilde{V}_t$. $\mathcal{A}_t$ is a rank r symmetric matrix representing the first-order proximity in appearance and spatial features of the nodes:

$$\mathcal{A}_t = \tilde{V}_t^\top \tilde{V}_t. \tag{8}$$

The motivation behind the estimation of $\mathcal{A}_t$ is similar to recent works [Santoro *et al.*, 2017; Cadene *et al.*, 2019] on modeling *implicit* relations of visual objects, in which they do not reflect any semantic or spatial relations but indicate the probabilities of object-pair co-occurrences given a query.

Language Binding Constructor
The visual graph explored by the visual graph constructor is powerful in representing dynamic object relation albeit still lacking the two-way complementary object-level relation between visual and textual data. In one direction, visual features provide grounding to ambiguous linguistic words so that objects of the same category can be differentiated [Nagaraja *et al.*, 2016]. Imagine the question "what is the color of the cat eating the cake" in a scene with many cats visible, then appearance and spatial features will clarify the selection of the cat of interest. In the opposite direction, linguistic cues provide more precise information than visual features of segmented regions. In the previous example, the "eat" relation between "cat" and "cake" is clear from the query words and is useful to connect these two visual objects in the image. These predicative advantages are even more important in the case of higher order relationships.

Drawing inspiration from that observation, we build a multi-modal graph $\mathcal{G}'_t = (X_t, \mathcal{A}_t)$ from the constructed graph $\mathcal{G}_t = (V_t, \mathcal{A}_t)$. Each node $x_{t,i} \in X_t$ of $\mathcal{G}'_t$ is a binding of the corresponding visual node $v_{t,i}$ of $\mathcal{G}_t$ with its linguistic supplement given by the context-aware function $f_t(.)$:

$$x_{t,i} = [v_{t,i}; f_t(e_1, ..., e_S | v_{t,i})]. \tag{9}$$

Designing $f_t(.)$ is key to make this representation meaningful. In particular, we design this function as the weighted composition of contextual words $\{e_s\}_{s=1}^S$:

$$f_t(e_1, ..., e_S | v_{t,i}) = \sum_{s=1}^S \beta_{t,i,s} * e_s. \tag{10}$$

Here combination weights $\beta_{t,i,s}$ represent the cross-modality partnership between a visual object $v_{t,i}$ and a linguistic word e_s, essentially forming the contextualized pair-wise bipartite relations between the V and L.

To calculate $\beta_{t,i,s}$, we first preprocess them by modulating V with the previous memory state $\hat{V}_t = W_t^{\hat{v}}[V; m_{t-1} \odot V] + b^{\hat{v}}$ and soft classifying each word s into multiple lexical types as a weight vector z_s similar to [Yang *et al.*, 2019], $z_s = \sigma(W^{z1}(W^{z0}e_s + b^{z0}) + b^{z1})$. Subsequently, the normalized cross-modality relation weights are calculated as:

$$\beta_{t,i,s} = z_s * \text{softmax}_s(W_t^\beta(\tanh(W_t^{\hat{v}}\hat{v}_{t,i} + W_t^e e_s))). \tag{11}$$

By doing this, we allow per-object communication between the two modalities, differentiating our method from prior works where linguistic cue is reduced to a single vector for conditioning or combined with visual signal in a late fusion.

Representation Refinement
At the last step of LOG operation, we rely on the newly built multi-modal graph $\mathcal{G}'_t = (X_t, \mathcal{A}_t)$ as the structure to refine the representation of objects by employing a graph convolutional network (GCN) [Kipf and Welling, 2017] of H hidden layers. Generally, vanilla GCNs have a difficulty of stacking deep layers due to the common vanishing gradient and numerical instability. We solve this problem by borrowing the residual skip-connection trick from ResNet [He *et al.*, 2016] to create more direct gradient flow. Concretely, the refined node representation is given by:

$$R_1 = X_t, \tag{12}$$
$$F_h(R_{h-1}) = W_{h-1}^2 \rho\left(W_{h-1}^1 R_{h-1} \mathcal{A}_t + b_{h-1}\right), \tag{13}$$
$$R_h = \rho\left(R_{h-1} + F_h(R_{h-1})\right), \tag{14}$$

where, $h = 1, 2.., H$, and ρ is an activation function which is an ELU operation in our later experiments. The parameters $\left(W_{h-1}^1, W_{h-1}^2\right)$ can be optionally tied across H layers.

As we obtain the refined representation $R_{t,H} = \{r_{t,i,H}\}_{i=1}^N$ after the H refinement layers, we compute the overall final representation by smashing the graph into one single vector:

$$\tilde{x}_t = \sum_{i=0}^N \delta_{t,i} * r_{t,i,H}, \tag{15}$$

where, $\delta_{t,i} = \text{softmax}_i(W_t^\delta r_{t,i,H})$. Finally, we update LOG's working memory state:

$$m_t = W_t^m[m_{t-1}; \tilde{x}_t] + b^m. \tag{16}$$

3.3 Answer Prediction

After T passes of LOG iterations, LOGNet combines the final memory state m_T with the sequential expression q of the question by concatenation followed by a linear layer to get the final representation $J = W[m_T; q] + b$, $J \in \mathbb{R}^d$.

For answer prediction, we adopt a 2-layer multi-layer perceptron (MLP) and a batch normalization layer in between as a classifier. The network is trained using cross-entropy loss or binary cross-entropy loss according to types of questions.

4 Experiments

4.1 Datasets

We evaluate our model on multiple datasets including:

CLEVR [Johnson *et al.*, 2017a]: presents several reasoning tasks such as transitive relations and attribute comparison. We intentionally design experiments to evaluate the generalization capability of our model on various subsets of CLEVR, where most existing works fail, sampled by the number of questions.

CLEVR-Human [Johnson *et al.*, 2017b]: composes natural language question-answer pairs on images from CLEVR. Due to diverse linguistic variations, this dataset requires stronger visual reasoning ability than CLEVR.

GQA [Hudson and Manning, 2019b]: the current largest visual relational reasoning dataset providing semantic scene graphs coupled with images. Because LOGNet does not need
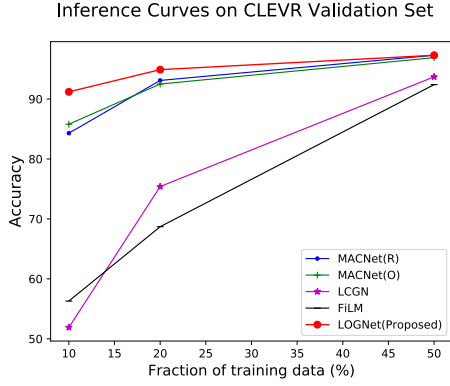
Figure 4: VQA performance on CLEVR subsets.

Method	Val. Acc. (%)
FiLM [Perez *et al.*, 2018]	56.6
MACNet(R) [Hudson and Manning, 2018]	57.4
LCGN [Hu *et al.*, 2019]	46.3
BAN [Shrestha *et al.*, 2019]	60.2
RAMEN [Shrestha *et al.*, 2019]	57.9
LOGNet	**62.5**

Table 1: Performance on CLEVR-Human.

Training size	Method	Accuracy (%)	
		val	test
Full	CNN+LSTM	49.2	46.6
	Bottom-Up	52.2	49.7
	MACNet(O)	57.5	54.1
	LCGN	63.9	56.1
	LOGNet	63.2	55.2
50%	LCGN	60.6	-
	LOGNet	61.0	-
20%	LCGN	53.2	-
	LOGNet	53.8	-

Table 2: Performance on GQA and subsets.

Method	Val. Acc. (%)
XNM	43.4
MACNet(R)	40.7
MACNet(O)	45.5
LOGNet	**46.8**

Table 3: Experiments on VQA v2 subset of long questions.

prior predicates, we ignore these static graphs using only the image and textual query as input.

VQA v2 [Goyal *et al.*, 2017]: As a large portion of questions is short and can be answered by looking for facts in images, we design experiments with a split of only long questions (>7 words). The split, hence, assesses the ability to model the relations between objects, e.g.: "What is the white substance on the left side of the plate and on top of the cake?".

4.2 Performance Against SOTAs

Our model is generally implemented with feature dimension $d = 512$, reasoning depth $T = 8$, GCN depth $H = 8$ and attention-width $K = 2$. The number of regions is $N = 14$ for CLEVR and CLEVR-Human, and 100 for GQA and 36 for VQA v2 to match with other related methods. We also match the word embeddings with others by using random vectors of a uniform distribution for CLEVR/CLEVR-Human and pretrained GloVe vectors for the other datasets.

We compare with state-of-the-art methods reporting performance as in their papers or obtained with their public code. For the better judgement of whether the improvement is from the model designs or from the use of better visual embeddings, we reimplement MACNet [Hudson and Manning, 2018] with their feature choice of ResNet - MACNet(R), and additionally try it out on our ROI pooling features - MACNet(O).

CLEVR and CLEVR-Human Dataset

Fig. 4 demonstrates the large improvement of LOGNet over SOTAs including MACNet, FiLM and LCGN particularly with limited training data. With enough data, all models converge in performance. With smaller training data, other methods struggle to generalize, while LOGNet maintains stable performance. With 10% of training data, FiLM quickly drops to

51.9%, and only 48.9% in case of LGCN, which barely surpasses the linguistic bias performance of 42.1% reported by [Johnson *et al.*, 2017a]. Behind LOGNet (91.2%), MACNet is the runner up in generalization with around 85.8%.

Our model shows significant improvement over other works on CLEVR-Human dataset (See Table 1) where language vocab is richer than the original CLEVR. We only report results without fine-tune on CLEVR for better judgment of the generalization ability. This suggests that LOGNet can better handle the linguistic variations by its advantage in modeling cross-modality interactions.

GQA

LOGNet outperforms previous works including simple fusion approaches CNN+LSTM and Bottom-Up [Anderson *et al.*, 2018], and the recent advanced multi-step inference MAC-Net. Although LOGNet achieves competitive performance as compared with LCGN on the full set, it shows its advantage in generalization and robustness against overfitting in limited data experiments (20% and 50% splits) - see Table 2.

VQA v2 - Subset of Long Questions

LOGNet is finally applied to the most difficult questions of VQA v2. Empirical results show that our model achieves favorable performance over MACNet and XNM [Shi *et al.*, 2019] on this subset. Due to the rich language vocab of human annotated datasets, the improvements are less noticeable as compared with those on synthetic datasets such as CLEVR.

4.3 Ablation Studies

We conduct ablation studies with our model on CLEVR subset of 10% training data (See Table 4). We observe consistent improvements responding to the increase in the number of reasoning steps as well as going deeper with the representation refinement process. We have tried up to $p = 16$ LOG units and $H = 16$ GCN layers in each time step, establishing a
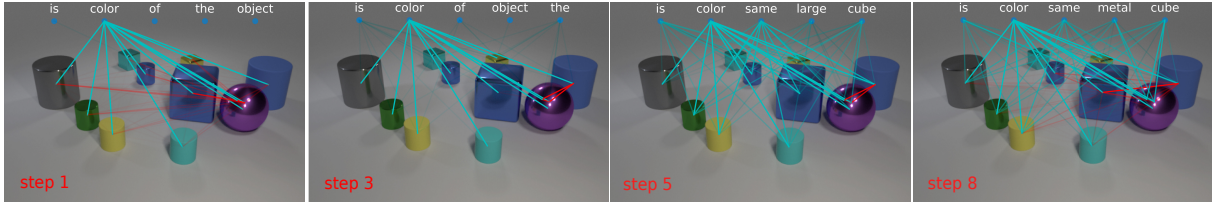
Question: Is the color of the big matte object the same as the large metal cube?
Prediction: yes   **Answer**: yes
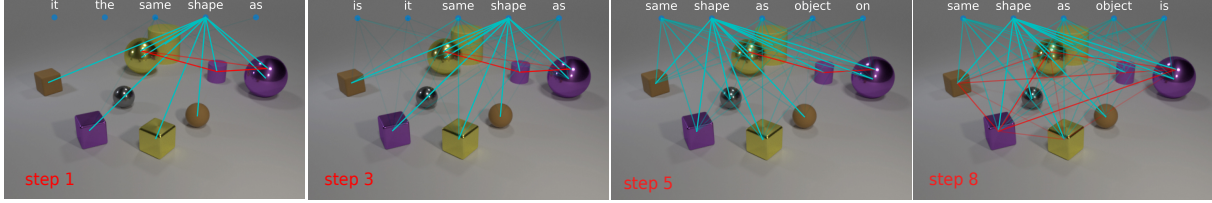
Question: There is a tiny purple rubber thing; does it have the same shape as the brown object that is on the left side of the rubber sphere?
Prediction: no   **Answer**: no

Figure 5: Chains of visual object relation (in red) with language binding (in cyan) constructed for two image-question pairs. Visual relations are found adaptively to the specific questions and reasoning stages. Language binding was sharp on key cross-modality relations at several early steps, then flats out as memory converges. Only five words included for visualization purposes. Best viewed in color.

No.	Model	Val. Acc. (%)
1	Default config. (8 LOG units, 8 GCNs)	91.2
2	w/o bounding box features	86.5
3	Graph constructor w/o previous memory	86.5
4	Graph constructor w/o language	56.2
5	Single-head attn. controlling signal	86.3
6	Rep. refinement w/ 1 GCN layers	75.9
7	Rep. refinement w/ 4 GCN layers	89.4
8	Rep. refinement w/ 12 GCN layers	91.1
9	Rep. refinement w/ 16 GCN layers	89.5
10	Language binding w/o previous memory	90.8
11	w/o language binding	89.9
12	1 LOG unit	69.0
13	4 LOG units	76.3
14	12 LOG units	91.6
15	16 LOG units	91.1

Table 4: Ablation studies - CLEVR dataset: 10% subset.

very deep reasoning process over hundreds of layers. The results strongly prove the ability to leverage recurrent cells (row 12-14) and the significance of the deep refinement layers (row 6-9). It is also clear that linguistic cue plays a crucial role in all the components of LOGNet and language binding contributes noticeably to performance (row 1 and 11).

4.4 Behavior Analysis

To understand the behavior of the dynamic graphs during LOG iterations, we visualize them for complex questions from CLEVR (see Fig. 5). As seen, the linguistic objects most selected for binding are from objects of interest or their attributes which give a hint to the model of what aspect of the visual cue to look at. Question types (e.g. yes-no/wh-question, object counting) and other function words (e.g. "the", "is", "on") are also paid much attention to. Note that as linguistic objects are outputs of LSTM passes, those of function words, such as articles and conjunctions connect nearby content words and holds their aggregated information through the LSTM operations.

Progressing through the reasoning steps, LOGNet accumulates multiple aspects of joint domain information in a compositional manner. In earlier steps when most crucial reasonings happen, it is apparent in Fig. 5 that language binding concentrates on sharp linguistic-visual relations such as from attribute and predicate words (e.g. "color", "shape", "same") to their related objects. They constitute the most principal components of the working memory. Later in the reasoning process, when the memory gets close to the convergence, the binding weights flat out as not much critical information is being added anymore. This agrees with the ablation study result in the last four rows of Table 4 where the performance raises sharply in the early steps and gradually converges.

5 Discussion

We have presented a new neural recurrent model for compositional and relational reasoning over a knowledge base with implicit intra- and inter- modality connections. Distinct from existing neural reasoning methods, our method computes dynamic dependencies *on-demand* as reasoning proceeds. Our focus is on VQA tasks, where raw visual and linguistic features are given but their relations are unknown. The experimental results demonstrated superior performance on multiple datasets even when trained on just 10% data.

The chaining of implicit relations and representation refinements in this model suggests further study (a) on the adaptive depth of refinement layers and the length of the reasoning, e.g., by considering the complexity of the scene and of the question; and (b) relationship with first-order logic inference.

References

[Anderson *et al.*, 2018] Peter Anderson, X. He, C. Buehler, D. Teney, M. Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018.

[Baradel *et al.*, 2018] Fabien Baradel, Natalia Neverova, Christian Wolf, Julien Mille, and Greg Mori. Object level visual reasoning in videos. In *ECCV*, pages 105–121, 2018.

[Cadene *et al.*, 2019] Remi Cadene, Hedi Ben-Younes, Matthieu Cord, and Nicolas Thome. Murel: Multimodal relational reasoning for visual question answering. In *CVPR*, pages 1989–1998, 2019.

[Desta *et al.*, 2018] Mikyas Desta, L. Chen, and T. Kornuta. Object-based reasoning in VQA. In *WACV*, pages 1814–1823, 2018.

[Goyal *et al.*, 2017] Yash Goyal, T. Khot, D. 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*, pages 6904–6913, 2017.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CVPR*, pages 770–778, 2016.

[Hu *et al.*, 2017] Ronghang Hu, J. Andreas, M. Rohrbach, T. Darrell, and Kate Saenko. Learning to reason: End-to-end module networks for visual question answering. In *ICCV*, pages 804–813, 2017.

[Hu *et al.*, 2019] Ronghang Hu, Anna Rohrbach, Trevor Darrell, and Kate Saenko. Language-conditioned graph networks for relational reasoning. In *ICCV*, pages 10294–10303, 2019.

[Huang *et al.*, 2019] Pingping Huang, Jianhui Huang, Yuqing Guo, Min Qiao, and Yong Zhu. Multi-grained attention with object-level grounding for visual question answering. In *ACL*, pages 3595–3600, 2019.

[Hudson and Manning, 2018] Drew A Hudson and Christopher D Manning. Compositional attention networks for machine reasoning. *ICLR*, 2018.

[Hudson and Manning, 2019a] Drew Hudson and Christopher D Manning. Learning by abstraction: The neural state machine. In *NeurIPS*, pages 5901–5914, 2019.

[Hudson and Manning, 2019b] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional qa. In *CVPR*, pages 6700–6709, 2019.

[Johnson *et al.*, 2017a] Justin Johnson, B. Hariharan, L. 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*, pages 2901–2910, 2017.

[Johnson *et al.*, 2017b] Justin Johnson, Bharath Hariharan, L. van der Maaten, Judy Hoffman, Li Fei-Fei, C.L. Zitnick, and Ross B Girshick. Inferring and executing programs for visual reasoning. In *ICCV*, pages 2989–2998, 2017.

[Kim *et al.*, 2018] Seung Wook Kim, Makarand Tapaswi, and Sanja Fidler. Visual reasoning by progressive module networks. *ICLR*, 2018.

[Kipf and Welling, 2017] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *ICLR*, 2017.

[Lake *et al.*, 2017] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *BBS*, 2017.

[Li *et al.*, 2019] Linjie Li, Zhe Gan, Yu Cheng, and Jingjing Liu. Relation-aware graph attention network for visual question answering. *ICCV*, pages 10313–10322, 2019.

[Nagaraja *et al.*, 2016] Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. Modeling context between objects for referring expression understanding. In *ECCV*, pages 792–807, 2016.

[Norcliffe-Brown *et al.*, 2018] Will Norcliffe-Brown, Stathis Vafeias, and Sarah Parisot. Learning conditioned graph structures for interpretable visual question answering. In *NeurIPS*, pages 8334–8343, 2018.

[Palm *et al.*, 2018] Rasmus Palm, U. Paquet, and O. Winther. Recurrent relational networks. In *NeurIPS*, 2018.

[Perez *et al.*, 2018] Ethan Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville. Film: Visual reasoning with a general conditioning layer. In *AAAI*, 2018.

[Ren *et al.*, 2015] Shaoqing Ren, K. He, R. Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *NIPS*, pages 91–99, 2015.

[Santoro *et al.*, 2017] Adam Santoro, David Raposo, D.G. Barrett, M. Malinowski, Razvan Pascanu, Peter Battaglia, and Tim Lillicrap. A simple neural network module for relational reasoning. In *NIPS*, pages 4967–4976, 2017.

[Shi *et al.*, 2019] Jiaxin Shi, Hanwang Zhang, and Juanzi Li. Explainable and explicit visual reasoning over scene graphs. In *CVPR*, pages 8376–8384, 2019.

[Shrestha *et al.*, 2019] Robik Shrestha, Kushal Kafle, and Christopher Kanan. Answer them all! toward universal visual question answering models. In *CVPR*, pages 10472–10481, 2019.

[Veličković *et al.*, 2018] Petar Veličković, G. Cucurull, A. Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *ICLR*, 2018.

[Xu *et al.*, 2017] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *CVPR*, pages 5410–5419, 2017.

[Xu *et al.*, 2020] Keyulu Xu, Jingling Li, Mozhi Zhang, Simon S Du, Ken-ichi Kawarabayashi, and Stefanie Jegelka. What can neural networks reason about? *ICLR*, 2020.

[Yang *et al.*, 2019] Sibei Yang, Guanbin Li, and Yizhou Yu. Dynamic graph attention for referring expression comprehension. In *ICCV*, pages 4644–4653, 2019.

[Yu *et al.*, 2017] Licheng Yu, Hao Tan, Mohit Bansal, and Tamara L Berg. A joint speaker-listener-reinforcer model for referring expressions. *CVPR*, pages 7282–7290, 2017.