Visual-Textual Alignment for Graph Inference in Visual Dialog

Tianling Jiang
School of Computer Science and Technology Soochow University
tljiang@stu.suda.edu.cn

Ji Yi
School of Computer Science and Technology Soochow University
jiyi@suda.edu.cn

Chunping Liu
School of Computer Science and Technology Soochow University
cpliu@suda.edu.cn

Hailin Shao
School of Computer Science and Technology Soochow University
20184227001@stu.suda.edu.cn

Abstract

As a conversational intelligence task, visual dialog entails answering a series of questions grounded in an image, using the dialog history as context. To generate correct answers, the comprehension of the semantic dependencies among implicit visual and textual contents is critical. Prior works usually ignored the underlying relation and failed to infer it reasonably. In this paper, we propose a Visual-Textual Alignment for Graph Inference (VTAGI) network. Compared with other approaches, it makes up the lack of structural inference in visual dialog. The whole system consists of two modules, Visual and Textual Alignment (VTA) and Visual Graph Attended by Text (VGAT). Specially, the VTA module aims at representing an image with a set of integrated visual regions and corresponding textual concepts, reflecting certain semantics. The VGAT module views the visual features with semantic information as observed nodes and each node learns the relationship with others in visual graph. We also qualitatively and quantitatively evaluate the model on VisDial v1.0 dataset, showing our VTAGI outperforms previous state-of-the-art models.

1 Introduction

Cross-modal semantic understanding has become an attractive challenge in natural language processing and computer vision, inspiring many tasks such as image captioning(Xu et al., 2015; Vinyals et al., 2015) and visual question answering (VQA)(Antol et al., 2015; Anderson et al., 2018; Shimizu et al., 2018). However, in these missions, the co-reference between vision and language is usually performed in a single round and they do not have many interactions with human over a period of time. In 2017, Das et al. introduced a continuous conversational task, visual dialog(Das et al., 2017). This task needs an AI agent to answer a sequence of questions based on visually-grounded information and contextual information from a dialog history.

Recently, a manual investigation(Kim et al., 2020) on the Visual Dialog dataset (VisDial) tried to figure out how many questions can be answered with images and how many of them need conversation history to be answered. The investigation shows that around 80% of the questions can be answered with images and about 20% of the questions need the knowledge from dialog history. Therefore, one of the key challenges in visual dialog is how to effectively utilize these underlying contents in the textual and visual information, i.e., input questions, dialog history and input image. In previous works, such as RvA(Niu et al., 2019) and DAN(Kang et al., 2019), both tended to explicitly reason over past dialog interactions by referring back to previous references, but they ignored the underlying relational structure which contributes to dialog inference. Nowadays, researchers have attempted to consider the fixed graph attention or embedding to resolve the problem with structural representations(Zheng et al., 2019; Schwartz et al., 2019). They focused on the textual modality but neglected the rich underlying information in the image. In this task, despite its significance to artificial intelligence and human-computer interaction, the agent requires understanding a series of multi-modal entities, and reasons the rich information in both vision

Proceedings of the 28th International Conference on Computational Linguistics, pages 1874–1885
Barcelona, Spain (Online), December 8-13, 2020

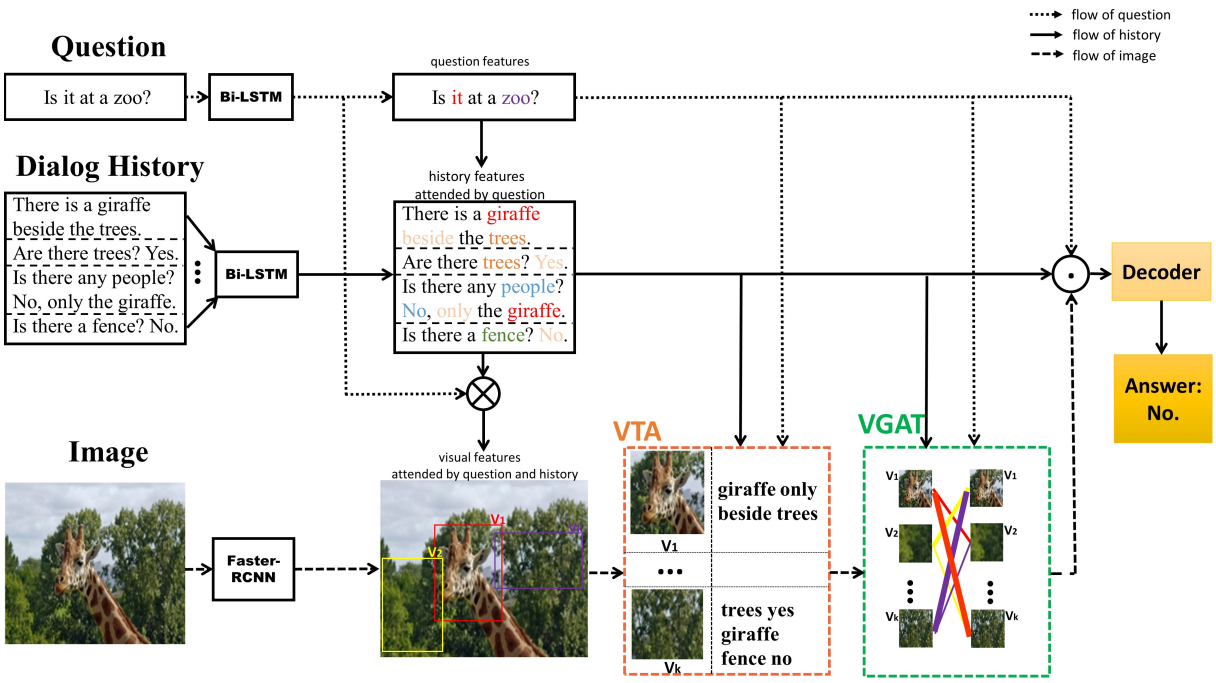

Figure 1: An overview of VTAGI. We present two modules, VTA and VGAT. VTA takes the attended visual features and attended textual features including question and history as inputs, resulting in integrated image representations reflecting semantics of certain objects in the image. VGAT aims to construct a visual graph combined with textual context. The relations among the nodes are top-5 important. For example, the most thickest link between the node v_1 and v_k indicates the most important dependencies of them. $\otimes$ and $\odot$ denote matrix multiplication and element-wise product, respectively.

and language. An ideal inference algorithm should be able to find out the underlying relational structure and give a reasonable answer based on this structure.

To address aforementioned problem, we pay more attention on visual-textual relation and propose the VTAGI in Figure 1 to explore potential information for structural inference. The agent will first obtain the question features, history features and visual features by employing different attention mechanisms. However, the semantics of the visual features and the textual concepts are usually inconsistent, and the representations of the image lack of global structural information. Thus, the VTA module is to align the visual features and global textual contents with their relevant counterparts in each image domain. As a result, the visual features contain more specific semantic information. For example, in Figure 1, the visual feature v_1 contains the semantic information of "giraffe, only, beside, trees", because each visual feature considers all contextual information. In order to infer more reasonably and connect all of the individual visual features, we design the VGAT module to construct a visual graph that shows the different relationships among various visual features. For example, considering the feature v_1 in Figure 1, the thickest link between the v_1 and v_k indicates the most important relationship between the two features. This module learns how to select other nodes related to the current node. In the last step in this process, each visual feature node in this structural module is connected to its related nodes. Through the two modules, the final visual features possess more related information and they are intra-connected in the graph, which are beneficial to inference.

2 Related Work

Visual Dialog. Most studies on the task of visual dialog introduced by Das et al.(Das et al., 2017) can be categorized into four groups. **Fusion-based Models**: late fusion (LF)(Das et al., 2017) and hierarchical recurrent network (HRE)(Das et al., 2017) directly encoded the multi-modal inputs (image, question, dialog history) and decoded the answer. **Attention-based Models**: memory network (MN)(Das

et al., 2017), history-conditioned image attention (HCIAE)(Lu et al., 2017), sequential co-attention (CoAtt)(Wu et al., 2018) and synergistic co-attention network (Sync)(Guo et al., 2019) computed attended representations of inputs. **Visual Co-reference Resolution (VCoR)-based Models**: attention memory (AMEM)(Seo et al., 2017), neural module networks (CorefNMN)(Kottur et al., 2018), recursive visual attention mechanism (RvA)(Niu et al., 2019) and dual attention network (DAN)(Kang et al., 2019) , these solutions clarified ambiguous expressions (e.g., he, she, they) in the text and focused on explicit visual co-reference resolution. **Graph-based Models** attempt to construct some structures to obtain more underlying information. Zheng et al.(Zheng et al., 2019) designed a structural inference model based on an EM-style (expectation-maximization) GNNs (graph neural networks) to conduct the textual co-reference. Schwartz et al.(Schwartz et al., 2019) proposed a factor graph mechanism and constructed the graph over all the multi-modal features. Guo et al.(Guo et al., 2020) utilized the word-level attention of question to construct a context-aware graph. The aforementioned graph-related models did not highlight the visual features and their relationships. While, in our work, which also belongs to the forth group, building a relational graph based on visual objects can contain more information.

Visual-semantic Alignment. In image captioning, Karpathy and Li(Karpathy and Fei-Fei, 2015) introduced the notion of visual-semantic alignment, which was based on a novel combination of Convolution Neural Network over image regions, bidirectional Recurrent Neural Network over sentences and a structural objective that aligned two modalities through a multi-modal embedding. In the field of VQA, some recent efforts(Nam et al., 2017; Kim et al., 2018; Nguyen and Okatani, 2018; Ben-Younes et al., 2017) have also been dedicated to studying similar alignment between image and question. To acquire integrated image representations, they normally aligned the visual features and textual concepts, which were beneficial to explore the latent relation. In this paper, we align heterogeneous modalities (image, question and history) based on distinct attention mechanisms, to make each region in the image possessing more specific and detailed contents. Especially, we make the visual features contain two levels of semantic information. By adding history features, the visual features can acquire the global textual information. And through integrating question features, visual features also get the logical contents.

Graph Neural Network. The concept of graph neural network (GNN) was first proposed by (Scarselli et al., 2008), who extended existing neural networks for processing the data represented in graph domain. GNNs have been applied in various tasks (Gu et al., 2019; Li et al., 2019; Liu et al., 2018; Wang et al., 2019; Zheng et al., 2019). The core was to combine the graphical structural representation with neural networks. The GNN follows a strategy that controls how the representation vector of a node calculated by its neighboring nodes to capture specific patterns of a graph. The neighborhood connectivity information in GNNs is unrestricted and potentially irregular, giving them greater applicability than convolutional neural networks (CNNs), which impose a fixed regular neighborhood structure. In this paper, we apply GNN to learn the relation among visual features with multi-modal contexts for inferring answers. Through this GNN, each visual feature can connect with other associated features.

3 Proposed Approach

In this section, we firstly define the visual dialog task as in Das et al.(Das et al., 2017). Formally, a visual dialog agent takes image I, question Q_t and dialog history H_t as input. Among them, the Q_t is asked in the current round t. The H_t is consist of Q&A pairs till round t-1, while in the first round it only contains the caption C about the image I. The agent is required to return an answer A_t=$\{\mathrm{A}_1^t, \mathrm{A}_2^t, \cdots, \mathrm{A}_{100}^t\}$ to the Q_t, by ranking a list of 100 candidate answers in a discriminative manner.

We will present the language features and the image features in section 3.1, followed by section 3.2 describing VTA module. Finally, the detailed information of VGAT module is provided in section 3.3.

3.1 Feature Representation

Language Features. We first embed each word in the question Q_t as W^Q=$\{\mathrm{w}_{t,1}, \mathrm{w}_{t,2}, \cdots, \mathrm{w}_{t,T}\}$ by using GloVe(Pennington et al., 2014) embeddings, where T denotes the number of tokens in Q_t. Then we use a Bi-LSTM to encode W^Q into a sequence U^Q=$\{\mathrm{q}_1^t, \mathrm{q}_2^t, \cdots, \mathrm{q}_T^t\}$. Similarly, we get the history embedding vectors as W^H and the sequence representation of U^H=$\{h_i\}_{i=0}^{t-1}$. We then adopt the attention

mechanism(Vaswani et al., 2017) of Eq.1 to obtain the question features Q by setting the inputs Q_A, K_A, V_A as U^Q. In the same way, the history features H attended by the question Q can also utilize Eq.1, but the inputs K_A and V_A come from U^H and the Q_A come from attended vectors of Q. As shown in Figure 2, the question features pay particular attention to pronouns and nouns, such as "it", "zoo". The history questions, attended by question, pay more attention to global and logical content, such as "giraffe beside trees", "No people".

$$Att(Q, K, V) = softmax(\frac{Q_A K_A^T}{\sqrt{d_k}})V_A \tag{1}$$

Visual Features. Inspired by bottom-up attention(Anderson et al., 2018), we use Faster R-CNN(Ren et al., 2015) to extract object-level image features of $\{v_1, v_2, \cdots \text{and } v_k\}$. Firstly, we fuse the question and history features by matrix multiplication of Eq.2. Then, co-attention(Lu et al., 2016) of Eq.3 is exploited to get the attended visual features V. The C is the affinity matrix, H^v is the image attention maps. In this paper, we select top-k region proposals from each image, where k is simply fixed as 36.

$$Q\_H = U^Q \otimes H \tag{2}$$

$$\begin{cases} C = tanh(Q\_H^T W_b V); \\ H^v = tanh(W_v V + (W_{Q\_H} Q\_H)C); \\ \alpha^v = softmax(W_H^T H^v); \\ V = \sum_{i=1}^{k} \alpha_i^v v_i \end{cases} \tag{3}$$

3.2 Visual and Textual Alignment

In most previous works for this area, visual features generally contain low-level visual information and are difficult to align with textual contents. The purpose of the VTA is to form accurate alignment between the visual regions and the textual words. For this purpose, the global textual concepts are introduced to compensate the lack of high-level semantic information in visual features. As shown in Figure 2, to obtain the final visual features with matched semantic features, such as v_1 matches with "giraffe only, beside trees", we deal with the history and question features successively. We adopt the attention mechanism from Vaswani et al.(Vaswani et al., 2017) to learn the correlated features in a certain domain by querying the other domain. The multi-head attention is composed of h parallel heads and each head is formulated as a scaled dot-product attention. We evaluate the alignment between visual and history features as follows:

$$Att_i(V, H) = softmax(\frac{VW_i^{Q_1}(HW_i^{K_1})^T}{\sqrt{d_k}})HW_i^{V_1}(i = 1, 2, \cdots, h) \tag{4}$$

where $V \in \mathbb{R}^{k \times d_h}$ and $H \in \mathbb{R}^{t \times d_h}$ for k visual features and t history features respectively; $W_i^{Q_1}$, $W_i^{K_1}$, $W_i^{V_1} \in \mathbb{R}^{d_h \times d_k}$ are learnable parameters of linear transformations; d_h = 256 is the size of input features and $d_k = d_h/h$ (h=8) is the size of the output features for each attention head. Results from each head are concatenated and passed through a linear transformation to construct the output:

$$V' = MultiHead(V, H) = [Att_1(V, H), Att_2(V, H), \cdots, Att_h(V, H)]W^{o_1} \tag{5}$$

the $W^{o_1} \in \mathbb{R}^{d_h \times d_k}$ is the parameter to be learned. The multi-head attention integrates t history features into k visual features. In this step, the features correspond to the global textual features. As illustrated in Figure 2, the v_k matches with "trees, yes, giraffe". Similarly, we integrate t question features into k visual features of Eq.6 and Eq.7, which make the feature v_k adding "fence, no" logical semantic information. The equations are as follows:

$$Att_i(V', Q) = softmax(\frac{V'W_i^{Q_2}(QW_i^{K_2})^T}{\sqrt{d_k}})QW_i^{V_2}(i = 1, 2, \cdots, h) \tag{6}$$

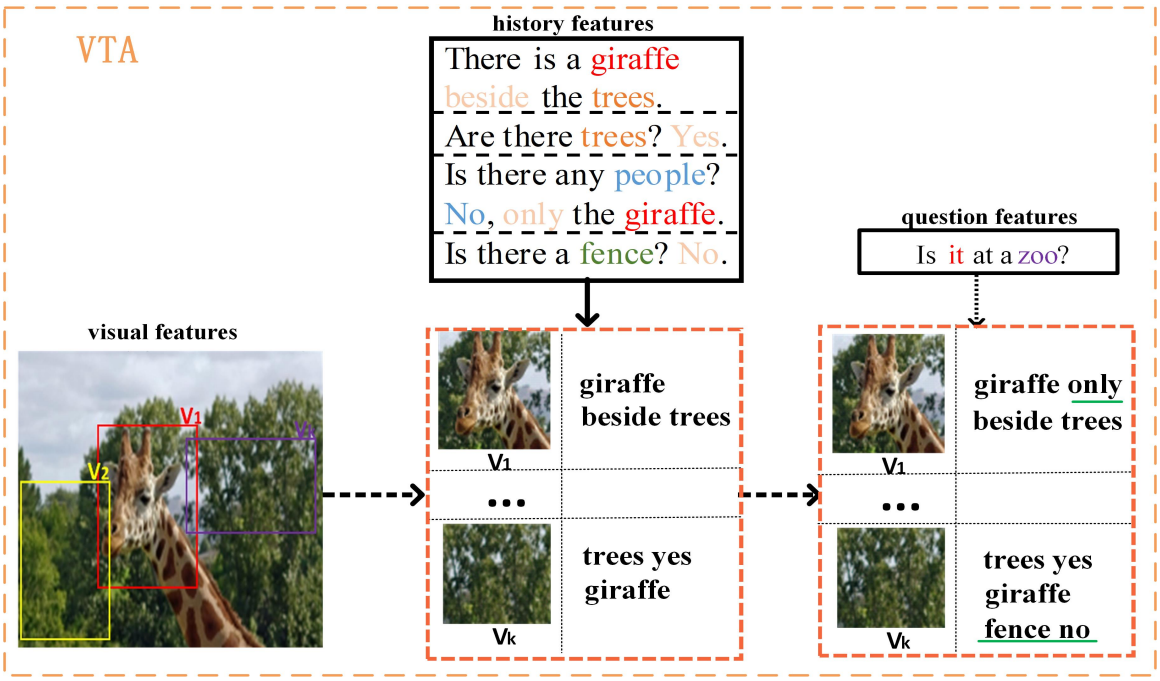

Figure 2: Visual and Textual Alignment(VTA) module. We first integrate the attended visual features and history features as inputs to get global semantic information. For example, the v_k is first matched with "trees yes, giraffe". Next, the question features will be deployed to compensate the logical semantic information for the visual features. For example, the logical semantic features ("fence no") corresponding to the v_k are added, and the logical information is underlined in green. The v_1, v_2, $\cdots$ and v_k are visual features matched with certain textual contents.

$$V_f = MultiHead(V', Q) = [Att_1(V', Q), \cdots, Att_h(V', Q)]W^{o_2} \tag{7}$$

Finally, the visual features V_f contain more consistent semantic features including global and logical contexts, resulting in the following graph construction more accurately and effectively.

3.3 Visual Graph Attended by Text

The previous VTA module ensures that the refined visual features only contain homogeneous information. For example, the visual features v_k in Fig.2 only contains related information of "fence no", but not the "only". Whereas, the visual features are independent and isolated. In order to establish the latent connection among them, we introduce a VGAT module. This module aims to build a graph which takes both visual and textual contents into account. Here, we build the visual graph by finding the visual relationships in the sentence/word-level textual information and corresponding semantics in visual features. The construction of the graph is denoted as G=$\{V_f, \varepsilon\}$, where the node v_i denotes a joint visual feature; the directed edge $\varepsilon_{i \to j}$ represents the relational dependency from node v_i to node v_j (i, j = 1, 2, $\cdots$, k).

From the Figure 3 in step S_i (i = 0, 1, $\cdots$, k), it shows that the construction of the graph has two textual operations with different colors. The graph is denoted as $G^{(i)}=\{V_f, \varepsilon^{(i)}\}$:

$$\begin{cases} G^{(i=0)} = [V_f; T_s] \\ G^{(i>0)} = [G^{(i-1)}; T_w^i] \end{cases} \tag{8}$$

where [;] is the concatenation operation; T_s is the visual-related textual feature in the sentence-level stage; T_w^i is the textual features in word-level stage. To construct the original visual graph in S_0 by introducing sentence-level information, we first calculate the question and history features attended by

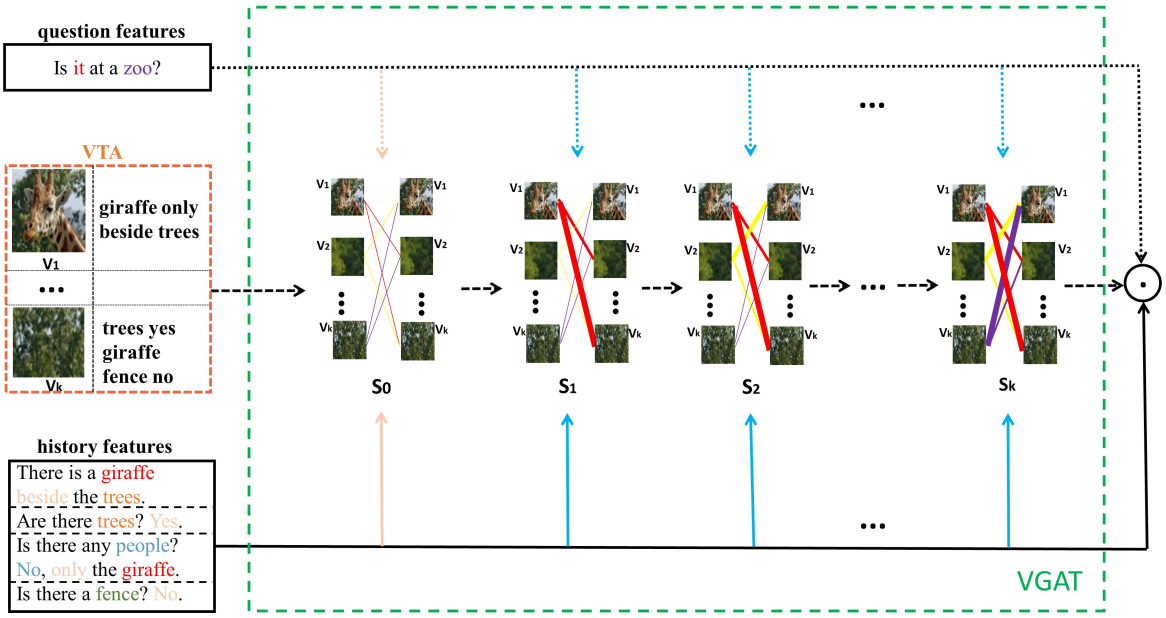

Figure 3: Visual Graph Attended by Text (VGAT) module. In two different stages, we use different colors to represent different textual operations influencing the construction of visual graph. For example, in step S_0, we construct the graph based on the visual features that are guided by sentence-level question and history features with pink link. In step S_1, to select some related neighbor visual features, we focus on the feature v_1 and integrate the word-level textual features shown in blue link. The thicker the connection between nodes, the more important the relationship between them. Finally, in step S_k, we focus on the feature v_k and the whole VGAT module is finished.

visual features that are generated from VTA module. Then concatenate the two textual features.

$$\begin{cases} Q_f = MultiHead(Q, V_f); \\ H_f = MultiHead(H, V_f); \\ \varepsilon^{(i=0)} = T_s = [Q_f; H_f] \end{cases} \tag{9}$$

The next step S_i (i>0), we adopt the word-level textual features, so the T_w^i is defined as follows:

$$\begin{cases} Z_q^{(i)} = L_2Norm(f_q^{(i)}Q); \\ \alpha_q^{(i)} = softmax(W_Q^{(i)} Z_q^{(i)}); \\ q_w^{(i)} = \sum_{j=1}^{t} \alpha_{q,j}^{(i)} w_{q_j} \end{cases} \tag{10}$$

$$\begin{cases} Z_h^{(i)} = L_2Norm(f_h^{(i)}H); \\ \alpha_h^{(i)} = softmax(W_H^{(i)} Z_h^{(i)}); \\ h_w^{(i)} = \sum_{j=0}^{t-1} \alpha_{h,j}^{(i)} w_{h_j} \end{cases} \tag{11}$$

$$\varepsilon^{(i>0)} = T_w^{(i)} = [q_w^{(i)}; h_w^{(i)}] \tag{12}$$

where $f_q^{(i)}(\cdot)$ and $f_h^{(i)}(\cdot)$ denote a two-layer MLP and $W_Q^{(i)}$ and $W_H^{(i)}$ are independently learned in the i-th step. The w_{q_j} and w_{h_j} are the component of the question embedding W^Q and history embedding W^H introduced in section 3.1.

Next, we describe the correlation among different nodes in the graph G. We define $\mathrm{A}^{(i)} \in \mathbb{R}^{k \times k}$ as the adjacency correlation matrix of the $\mathrm{G}^{(i)}$. In the matrix, the value $\mathrm{A}^{(i)}_{p \to q}$ represents the connection weight of the edge $\varepsilon^{(i)}_{p \to q}$.

$$\begin{cases} A^{(i=0)} = (W_1 G^{(i=0)})^T ((W_2 G^{(i=0)} \odot (W_3 T_s); \\ A^{(i>0)} = (W_1 A^{(i-1)})^T ((W_2 G^{(i)} \odot (W_3 T_w^{(i)})) \end{cases} \tag{13}$$

where W_1, W_2, W_3 are learnable parameters, and $\odot$ is the element-wise product.

It is a fact that there are always only a part of the detected objects in the image related to the similar textual contents. Therefore, the node at each step in the graph is required to connect with the most relevant neighbor nodes. In order to obtain a set of relevant nodes $\mathrm{R}^{(i)}$ in $\mathrm{G}^{(i)}$ (i =1, 2, $\cdots$, k), we adopt a ranking method as : $\mathrm{R}^{(i)}$ = top-5($\mathrm{A}^{(i)}$), where top-5 returns the indices of the 5 largest values in the matrix of $\mathrm{A}^{(i)}$. The $\mathrm{R}^{(i)}$ retains the most relevant nodes attributing to the final answer inference. Finally, the learning on each node in the graph not only integrates visual and textual features, but also involves context-visual relational learning. In this module, we establish links among all independent visual features.

Finally, we learn the representation of text and visual features with e_t which is fed into the discriminative decoder,

$$\begin{cases} z_g = tanh((W_g R^{(i)}) + b_g); \\ \alpha_g = softmax(P_g z_g); \\ e_{vg} = \sum_{i=0}^{k} \alpha_{g,i} R^{(i)} \end{cases} \tag{14}$$

$$e_t = tanh(W_e[Q, H, e_{vg}]) \tag{15}$$

where, W_g, b_g, W_e, P_g are learnable parameters. Q, H are attended textual features described in section 3.1, e_{vg} denotes the attended graph visual representation.

4 Experiments

4.1 Dataset and Evaluation Metrics

We evaluate the proposed approach on VisDial v1.0(Das et al., 2017), which includes additional 10k coco-like images from Flicker compared with v0.9(Das et al., 2017). The collection of dialogs on Flicker images is similar to that on MS-COCO images(Lin et al., 2014). The train, validation, test sets in v1.0 dataset contain 123k, 2k and 8k dialogs, respectively. Different from train and validation sets in v1.0 where each image is associated with a 10-round Q&A pair, the dialog in the test set has a random length within 10 rounds.

We follow(Das et al., 2017) to evaluate the response at each round. Specially, the dialog agent is given a list of 100 candidate answers, the model is expected to rank over the candidates and return a ranked list for further evaluation. The standard retrieval metrics are: mean rank evaluates the ground truth response (Mean), recall@K (K=1, 5, 10) evaluates where the ground truth is positioned in the sorted list(R@K), mean reciprocal rank evaluates the precision of the model by ranking where a ground truth answer is positioned (MRR), and normalized cumulative gain evaluates relative relevance of the predicted answers (NDCG). Higher value for R@K, MRR and NDCG is better, while lower value for Mean is better.

4.2 Quantitative Results

Comparing Methods. We compare our proposed model with the state-of-the-art approaches on VisDial v1.0 dataset. Based on the design of encoders, these methods can be grouped into: Fusion-base Models (LF and HRE(Das et al., 2017)), they fused image, question and history features at different stages; Attention-based Models (MN(Das et al., 2017) and Sync(Guo et al., 2019)), they established attention mechanisms over image, question and history; VCoR (Visual Co-reference Resolution) based Models (CorefNMN(Kottur et al., 2018), RvA(Niu et al., 2019), DAN(Kang et al., 2019) and HACAN(Yang

Table 1: Retrieval performance of our model on the VisDial v1.0 test set. In this table, there are four ways to solve the task of visual dialog. The value of NDCG, MRR and R@N, the higher the better. The lower of the value of Mean is better. Our VTAGI outperforms all other models across all metrics on the dataset.

Model	NDCG↑	MRR↑	R@1↑	R@5↑	R@10↑	Mean↓
Fusion-based Models						
LF(Das et al., 2017)	51.63	60.41	46.18	77.80	87.30	4.75
HRE(Das et al., 2017)	45.46	54.16	39.93	70.45	81.50	6.41
Attention-based Models						
MN(Das et al., 2017)	47.50	55.49	40.98	72.30	83.30	5.92
Sync(Guo et al., 2019)	57.32	62.20	47.90	80.43	89.95	4.17
VCoR-based Models						
CorefNMN(Kottur et al., 2018)	54.70	61.50	47.55	78.10	88.80	4.40
RvA(Niu et al., 2019)	55.59	63.03	49.03	80.40	89.83	4.18
DAN(Kang et al., 2019)	57.59	63.20	49.63	79.75	89.35	4.30
HACAN(Yang et al., 2019)	57.17	64.22	50.88	80.63	89.45	4.20
Graph-based Models						
GNN(Zheng et al., 2019)	52.82	61.37	47.33	77.98	87.73	4.57
FGA(Schwartz et al., 2019)	52.10	63.70	49.58	80.97	88.55	4.51
CAG(Guo et al., 2020)	56.64	63.49	49.85	80.63	90.15	4.11
Ours	**58.02**	**64.90**	**51.18**	**82.00**	**90.69**	**3.97**

et al., 2019)), they focused on explicit visual co-reference resolution based on textual features; Graph-based Models (GNN(Zheng et al., 2019), FGA(Schwartz et al., 2019) and CAG(Guo et al., 2020)), they proposed graph structure to explore more information from different modalities. The first two ways did not fully integrate textual information and image information. And the third way, the extracted information was too scattered and lacked structural guidance.

Results on VisDial v1.0. As shown in Table 1, our VTAGI outperforms the state-of-the-art method across all the metrics. We mainly compare our method with the graph-based ones. GNN(Zheng et al., 2019) constructed a graph exploring the dependencies among the textual-history. In contrast, our model builds a graph about the visual-objects integrated with question-history contexts. Compared with GNN, our model achieves 5.2% improvements on NDCG. FGA(Schwartz et al., 2019) constructed a graph, which simply combined representations of all modalities. In contrast, our method focuses more on the relationships among visual features. Compared with FGA, our model achieves about 6% improvements on NDCG. CAG(Guo et al., 2020) achieved the best performance on the metric NDCG based on the graph method for visual dialog, which designed a visual graph guided by the current question. However, our construction of visual graph guided by the different (sentence/word) levels question and history features at different stages and our method is more accurate in information extraction because of the operation of alignment. Specifically, compared with CAG(Guo et al., 2020), our result lifts NDCG from 56.64 to 58.02. In Fig. 4, we show four examples of our graphical inference. For convenience, in these examples, we only show top-2 related objects. Each example has three processes of P_1, P_2 and P_3. For example, in Fig. 4(a), the P_3 means that the dialog history already includes C, Q_1&A_1, Q_2&A_2 and the current question is Q_3. In this process, the more important relationships are between the boy and his clothes. Thus, the agent can relate the boy to his clothes, and infer the answer to Q_3 is "Yes.".

4.3 Ablation Study

In this section, we perform ablation study on VisDial v1.0 dataset with the following two model variants: Model only using VGAT module (B+VGAT) and Model only using VTA module (B+VTA). The baseline model (B) was introduced by Niu et al.(Niu et al., 2019), which proposed a novel attention mechanism RvA to capture question-relevant dialog history but ignored the structural visual inference

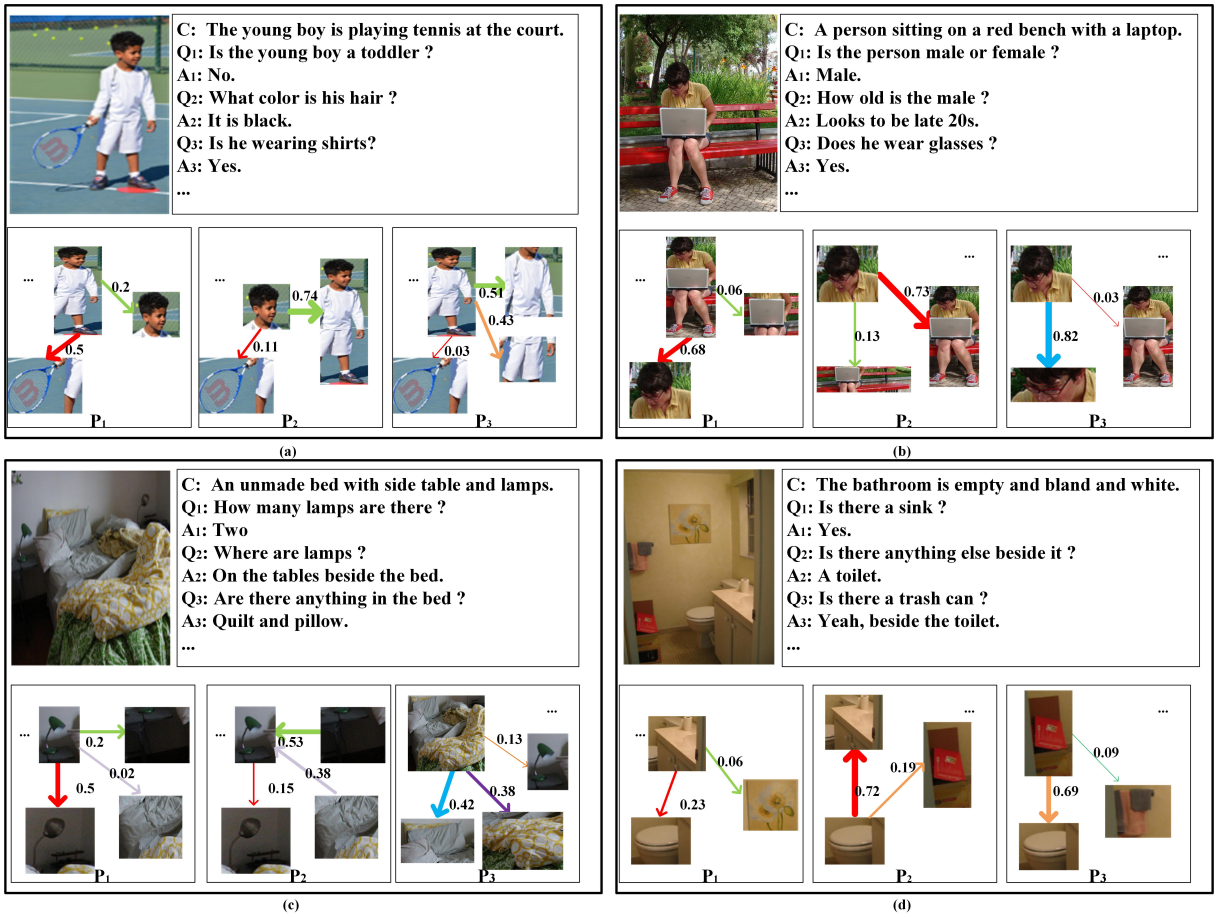

Figure 4: Visualization results of our model. It shows the relationships among various regions in the given image, and those values on links are calculated through the semantic in contextual and visual information. The higher the value, the thicker the line, the more important the relationship between the two objects. Different objects are linked with different colored lines. For example, in (c) P_3, the question Q_3 focuses on the bed in the given image. In the visual graph, we can see the object "bed" is connected to "quilt" and "pillow" with two lines having the higher weights of 0.42 and 0.38, respectively.

based on semantics. In our work, the main system not only aligns the visual and textual contents, but also constructs a visual relational features graph for effective inference. The B+VTA+VGAT(w/o question) and the B+VTA+VGAT(w/o history) confirm the importance of question and history information in VGAT module, respectively. In Table 2, B+VGAT and B+VTA improve the NDCG by about 2% respectively. Meanwhile, the combined architecture (B+VGAT+VTA) raises the NDCG from 55.59% to 58.02%.

Our ablation experiments illustrate the necessity and rationality of each part in our model. The VTA allows the visual representations to describe salient image regions with semantic perspective through the alignment between textual and visual features. This module provides more underlying information in image, thus the following module VGAT can make use of the information in both textual and visual features to learn the relationships among all features in the given image. The VGAT makes the visual features more fine-grained and correlational, and the structure of visual graph is helpful for answer inference. From the experimental results, our method is superior to the baseline and those models based on the graph method.

5 Conclusion

In this paper, we introduce Visual-Textual Alignment for Graph Inference (VTAGI) network based on graph method for the visual dialog task. Rather than relying on the visual attention maps in prior works,

Table 2: Ablation analysis of our proposed model. B means the Baseline Model. B+VGAT and B+VTA indicate the usage of different modules. And their performances are implemented on VisDial v1.0. Results highlighted in bold in the last column combine both two modules and it achieves the best performance.

Model	NDCG↑	MRR↑	R@1↑	R@5↑	R@10↑	Mean↓
B	55.59	63.03	49.03	80.40	89.83	4.18
B+VGAT	57.60	64.64	50.86	81.87	90.47	3.99
B+VTA	57.59	64.78	51.10	81.72	90.42	4.00
B+VTA+VGAT(w/o question)	57.46	64.20	50.97	81.96	90.51	4.13
B+VTA+VGAT(w/o history)	57.83	64.39	51.15	81.83	90.37	4.02
B+VGAT+VTA	**58.02**	**64.90**	**51.18**	**82.00**	**90.69**	**3.97**

VTAGI introduces alignment operation influenced by textual information and graph neural network approach. Our method is committed to obtaining more fine-grained and semantic-grounded image presentations with the help of linguistic clues. We empirically validate our proposed model on VisDial v1.0 dataset. Results show that our method is able to find and utilize underlying information for dialog inference, demonstrating its effectiveness. In future work, we aim to integrate positional relationships among visual objects by understanding the context.

Acknowledgements

Supported by National Natural Science Foundation of China (61972059, 61773272), The Natural Science Foundation of the Jiangsu Higher Education Institutions of China (19KJA230001), Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University (93K172016K08), Suzhou Key Industry Technology Innovation-Prospective Application Research Project SYG201807), the Priority Academic Program Development of Jiangsu Higher Education Institutions.

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6077–6086.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. 2017. Mutan: Multimodal tucker fusion for visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2612–2620.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 326–335.

Jiuxiang Gu, Handong Zhao, Zhe Lin, Sheng Li, Jianfei Cai, and Mingyang Ling. 2019. Scene graph generation with external knowledge and image reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1969–1978.

Dalu Guo, Chang Xu, and Dacheng Tao. 2019. Image-question-answer synergistic network for visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10434–10443.

Dan Guo, Hui Wang, Hanwang Zhang, Zheng-Jun Zha, and Meng Wang. 2020. Iterative context-aware graph inference for visual dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10055–10064.

Gi-Cheon Kang, Jaeseo Lim, and Byoung-Tak Zhang. 2019. Dual attention networks for visual reference resolution in visual dialog. *arXiv preprint arXiv:1902.09368*.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137.

Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. 2018. Bilinear attention networks. In *Advances in Neural Information Processing Systems*, pages 1564–1574.

Hyounghun Kim, Hao Tan, and Mohit Bansal. 2020. Modality-balanced models for visual dialogue. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Satwik Kottur, José MF Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. 2018. Visual coreference resolution in visual dialog using neural module networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 153–169.

Maosen Li, Siheng Chen, Xu Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. 2019. Actional-structural graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3595–3603.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Yong Liu, Ruiping Wang, Shiguang Shan, and Xilin Chen. 2018. Structure inference net: Object detection using scene-level context and instance-level relationships. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6985–6994.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. In *Advances in neural information processing systems*, pages 289–297.

Jiasen Lu, Anitha Kannan, Jianwei Yang, Devi Parikh, and Dhruv Batra. 2017. Best of both worlds: Transferring knowledge from discriminative learning to a generative visual dialog model. In *Advances in Neural Information Processing Systems*, pages 314–324.

Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. 2017. Dual attention networks for multimodal reasoning and matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 299–307.

Duy-Kien Nguyen and Takayuki Okatani. 2018. Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6087–6096.

Yulei Niu, Hanwang Zhang, Manli Zhang, Jianhong Zhang, Zhiwu Lu, and Ji-Rong Wen. 2019. Recursive visual attention in visual dialog. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99.

Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. 2008. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80.

Idan Schwartz, Seunghak Yu, Tamir Hazan, and Alexander G Schwing. 2019. Factor graph attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2039–2048.

Paul Hongsuck Seo, Andreas Lehrmann, Bohyung Han, and Leonid Sigal. 2017. Visual reference resolution using attention memory for visual dialog. In *Advances in neural information processing systems*, pages 3719–3729.

Nobuyuki Shimizu, Na Rong, and Takashi Miyazaki. 2018. Visual question answering dataset for bilingual image understanding: A study of cross-lingual transfer using attention maps. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1918–1928.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.

Peng Wang, Qi Wu, Jiewei Cao, Chunhua Shen, Lianli Gao, and Anton van den Hengel. 2019. Neighbourhood watch: Referring expression comprehension via language-guided graph attention networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1960–1968.

Qi Wu, Peng Wang, Chunhua Shen, Ian Reid, and Anton van den Hengel. 2018. Are you talking to me? reasoned visual dialog generation through adversarial learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6106–6115.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057.

Tianhao Yang, Zheng-Jun Zha, and Hanwang Zhang. 2019. Making history matter: History-advantage sequence training for visual dialog. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2561–2569.

Zilong Zheng, Wenguan Wang, Siyuan Qi, and Song-Chun Zhu. 2019. Reasoning visual dialogs with structural and partial observations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6669–6678.