Learning to Ground Visual Objects for Visual Dialog

Feilong Chen Xiuyi Chen Can Xu Daxin Jiang*
Microsoft Corporation, Beijing, China
{ivess.chan,hugheren.chan}@gmail.com
{caxu,djiang}@microsoft.com

Abstract

Visual dialog is challenging since it needs to answer a series of coherent questions based on understanding the visual environment. How to ground related visual objects is one of the key problems. Previous studies utilize the question and history to attend to the image and achieve satisfactory performance, however these methods are not sufficient to locate related visual objects without any guidance. The inappropriate grounding of visual objects prohibits the performance of visual dialog models. In this paper, we propose a novel approach to Learn to Ground visual objects for visual dialog, which employs a novel visual objects grounding mechanism where both prior and posterior distributions over visual objects are used to facilitate visual objects grounding. Specifically, a posterior distribution over visual objects is inferred from both context (history and questions) and answers, and it ensures the appropriate grounding of visual objects during the training process. Meanwhile, a prior distribution, which is inferred from context only, is used to approximate the posterior distribution so that appropriate visual objects can be grounded even without answers during the inference process. Experimental results on the VisDial v0.9 and v1.0 datasets demonstrate that our approach improves the previous strong models in both generative and discriminative settings by a significant margin.

1 Introduction

With the development of deep learning, various vision-language tasks have been introduced and attracted widespread attention, such as image captioning (Xu et al., 2015; Anderson et al., 2016, 2018; Cornia et al., 2020; Ghanimifard and Dobnik, 2019), visual question answering (Ren et al., 2015a; Gao et al., 2015; Lu et al., 2016; Anderson et al., 2018; Li et al., 2019; Huang et al., 2020) and visual dialog (Das et al., 2017; Chen et al., 2021a; Agarwal et al., 2020; Chen et al., 2021b; Qi et al., 2020). Specifically, visual dialog, which aims to hold a meaningful conversation (Chen et al., 2021c, 2020b) with a human about a given image, is a challenging task that requires models to locate related visual objects in an image and answer the current question based on the history and the located visual objects.

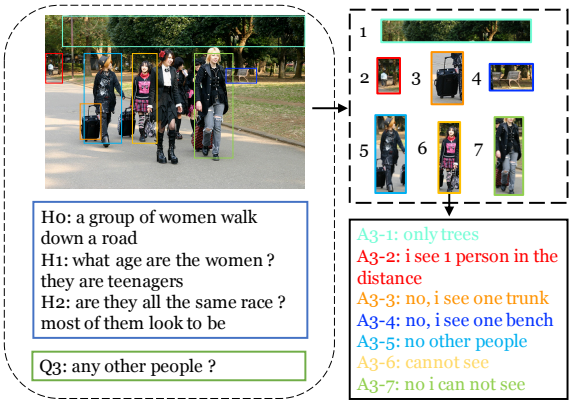

Figure 1: Comparison between different responses when focusing on different visual objects. We see that when the model focuses on wrong visual objects it makes mistakes. (Only the response A3-2 is right.)

In order to answer the question correctly, we need to accurately locate the question-related visual objects. Most existing methods utilize kinds of attention mechanism (Lu et al., 2017; Wu et al., 2018; Kottur et al., 2018; Gan et al., 2019; Guo et al., 2019b) to capture the target visual objects. ReDAN (Gan et al., 2019) and DMAM (Chen et al., 2020a) use multi-step reasoning based on dual attention to iteratively update related visual objects. DAN (Guo et al., 2019b), MCAN (Agarwal et al., 2020) and LTMI (Nguyen et al., 2020) utilize multi-head attention mechanisms to manage multi-modal intersection and obtain weight distributions. Moreover, there are some approaches (Zheng et al., 2019; Schwartz et al., 2019; Jiang et al., 2020b; Guo et al., 2020; Jiang et al., 2020a) using graph-

* Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 1081–1091
November 7–11, 2021.

based structures to capture related visual objects. FGA (Schwartz et al., 2019) realizes a factor graph attention mechanism, which constructs the graph over all the multi-modal features and estimates their interactions to ground visual objects. CAG (Guo et al., 2020) focuses on an iterative question-conditioned context-aware graph to locate related visual objects. However, the methods mentioned above obtain the prior distribution of visual objects through various interactions of questions, history and images, and finally use the prior distribution to obtain the final representation of the image. The prior distribution of visual objects is not enough to ground accurate visual objects, thus obtaining the wrong representation of the image.

In this paper, we propose a method to learn to ground visual objects in visual dialog. Specifically, we obtain the posterior distribution over visual objects by utilizing contexts and answers, while the prior distribution works without knowing answers in advance. Then we minimize the distance between the two distributions. During the training process, our model is trained to minimize the KL divergence between the prior distribution and the posterior distribution so that our model can approximate the posterior distribution accurately using the prior distribution. Then, during the inference process, the model grounds visual objects merely based on the prior distribution (i.e., without any posterior information). We show that through this process, the model can effectively learn to ground visual objects accurately and give informative and accurate responses by utilizing appropriate visual objects. We test the effectiveness of our proposed model on two large-scale datasets: VisDial v0.9 and v1.0 (Das et al., 2017). The contributions of this work are summarized as follows:

- We explore the importance of answers in grounding visual objects related to questions in visual dialog.
- We propose a novel approach to realize learning to ground visual objects in visual dialog via bridging the gap between the prior and posterior distribution over visual objects.
- We conduct extensive experiments and ablation studies on two large-scale datasets VisDial v0.9 and v1.0. Experimental results show that our approach can be used to improve previous visual dialog models in both generative and discriminative settings.

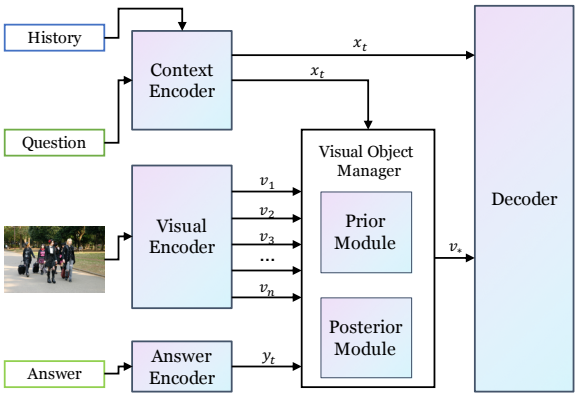

Figure 2: Architecture Overview

2 Methodology

Following Das et al. (2017), a visual dialog agent is given three inputs, i.e., an image i, history (the caption and question-answer pairs) till round $t-1$: $h = (\underbrace{Cap}_{h_0}, \underbrace{(q_1, a_1)}_{h_1}, \cdots, \underbrace{(q_{t-1}, a_{t-1})}_{h_{t-1}})$ and the current question q_t at round t, where Cap is the caption describing the image taken as h_0 and $h_1, \ldots, h_{t-1}$ are concatenations of question-answer pairs. The goal of the visual dialog agent is to generate an answer a_t to the question q_t.

2.1 Model Architecture

In this paper, we focus on training a neural visual dialog model with an effective visual objects grounding mechanism. As shown in Figure 2, we simplify existing visual dialog models into five major components:

- **The context encoder** encodes dialog history h and the current question q_t with an attention mechanism into a context vector x, and feeds it into the visual objects manager and the decoder.
- **The visual encoder** takes the image i as input and extract the image features $v = \{v_1, v_2, \ldots, v_\mu\}$ where μ denotes the number of object proposals for each image. Each object proposal is represented by a d_v-dimension feature vector.
- **The answer encoder** encodes the ground-truth answer a_t into a response vector y, and feeds it into the visual objects manager.
- **The visual object manager** consists of two sub-modules: a prior module and a posterior

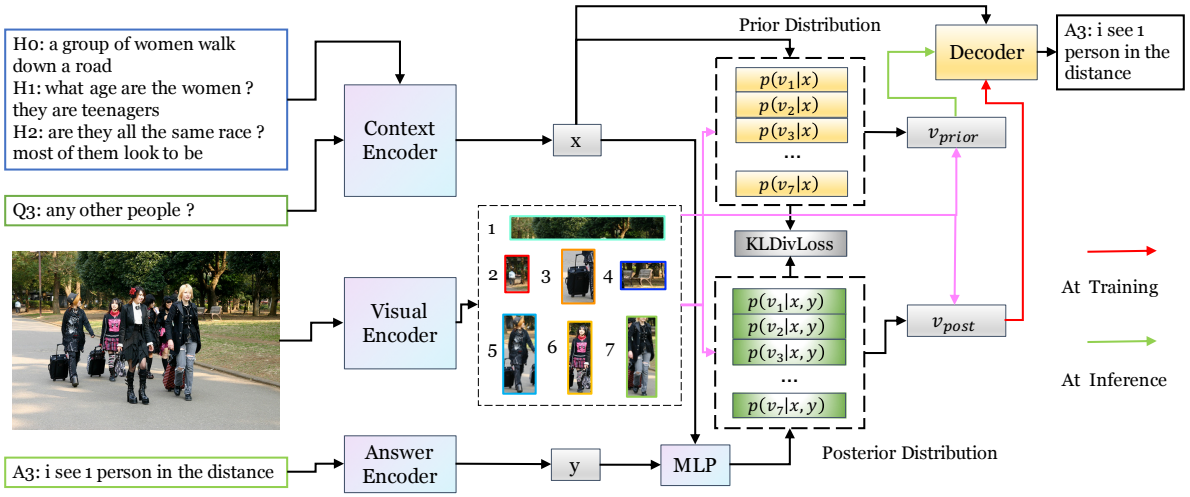

Figure 3: Framework of our Learning to Ground Visual Objects. The context encoder encodes the history and the current question into a context representation x. The visual encoder encodes the image into a region-based image features v. The answer encoder encodes the ground-truth answer into a response representation y. The visual object manager takes x, v and y as inputs, and generate a prior distribution $p(v|x)$ over visual objects and a posterior distribution $p(v|x, y)$ over visual objects, thus generating the new prior visual object features v_{prior} and the posterior visual object representation v_{post}. The decoder utilizes the context representation c and the new visual object representation (v_{post} at training or v_{prior} at inference) to generate and retrieve a response.

module. Given the previously encoded x and $v_{i=1}^{\mu}$ (and y if available), the visual object manager is responsible for deciding an appropriate distribution over visual objects and feeds the weighted visual object features v_* (together with an attention-based context vector x) into the decoder.

- **The decoder** generates and retrieves responses based on the visual object feature v_* and the attention-based context vector x.

2.2 Our Approach

When given the context vector x and the visual object features $v = \{v_1.v_2, \ldots, v_\mu\}$, and response vector y, the goal of the visual object manager is to decide an appropriate distribution D over visual objects and obtain the weighted visual object representation v_* based on the distribution D.

The visual object manager consists of two submodules: a prior module and a posterior module.

The Prior Module. The prior module aims to calculate the conditional probability distribution over μ visual objects, denoted by $\mathbf{p}(\mathbf{v}|\mathbf{x})$:

$$p(v = v_i|x) = \frac{exp(f_{cv}(x, v_i))}{\sum_{j=1}^{\mu} exp(f_{cv}(x, v_j))}, \quad (1)$$

where $f_{cv}(\cdot, \cdot)$ denotes the interaction function of the context vector x and the visual object features v_i. For example, $f_{cv}(\cdot, \cdot)$ can be the dot product, self-attention or other mechanisms to measure the association between v_i and the context vector x. A high association means that v_i is relevant to x and thus, v_i has a larger weight. Note that $p(v|x)$ is conditioned only on x and thus, it is a prior distribution over visual objects since it works without knowing the response. However, there can be different visual objects that are relevant to the contexts, and thus, it is difficult to select visual objects simply based on the prior distribution in training.

The Posterior Module. Motivated by this, in the posterior module, we define a posterior distribution over visual objects, denoted by $p(v|x, y)$, by considering both contexts and responses:

$$p(v = v_i|x, y) = \frac{exp(f_{cv}(f_{cy}(x, y), v_i))}{\sum_{j=1}^{\mu} exp(f_{cv}(f_{cy}(x, y), v_j))}, \quad (2)$$

where $f_{cy}(\cdot, \cdot)$ denotes the interaction function of x and y. For example, the $f_{cy}(\cdot, \cdot)$ can be an add operation, fully connected layer and other methods. Compared with the prior distribution, the posterior distribution is sharp since the actual visual objects used in the true response a_t can be captured.

Bridging the Gap. Clearly, the discrepancy between prior and posterior distributions introduces great challenges in training the model: it is desirable to ground visual objects based on the posterior

distribution, which, however, is unknown during inference. In this paper, we propose to approximate the posterior distribution using the prior distribution so that our model is capable of selecting appropriate visual objects even without posterior information. For this purpose, we introduce an auxiliary loss, namely the Kullback-Leibler divergence loss (KLDivLoss), to bridge the gap between the prior distribution and the posterior distribution. The KLDivLoss is defined as follows:

$$\mathcal{L}_{KL} = \sum_{i=1}^{\mu} p(v = v_i|x, y) log(\frac{p(v = v_i|x, y)}{p(v = v_i|x)}). \quad (3)$$

When minimizing KLDivLoss, the posterior distribution $p(v|x, y)$ can be regarded as labels and our model is instructed to use the prior distribution $p(v|x)$ to approximate $p(v|x, y)$ accurately. As a consequence, even when the posterior distribution is unknown in the inference process (since the actual response a_t is unknown), the prior distribution $p(v|x)$ can be effectively utilized to ground appropriate visual objects so as to generate and retrieve proper responses. To the best of our knowledge, it is the first neural model in visual dialog, which incorporates the posterior distribution as guidance, enabling accurate visual object grounding and high-quality response generation and retrieval.

3 Application of Our Approach

We take the strong baseline LTMI (Nguyen et al., 2020) as a base model to introduce our approach, which mainly consists of the following components:

Context Encoder and Answer Encoder: We use two bi-directional LSTM encoders to extract token-level representations $\mathbf{Q} \in \mathbb{R}^{\lambda \times d_q}$ and $\mathbf{y} \in \mathbb{R}^{\lambda \times d_q}$ of the question q_t and the answer a_t. We use another bi-directional LSTM encoder to extract sentence-level representations $\mathbf{H} \in \mathbb{R}^{T \times d_q}$ of the history h. λ is the length of questions and answers with paddings, T is the turn of dialog and d_q is the dimension. $\mathbf{Q}$ and $\mathbf{H}$ are fused into a context representation $\mathbf{x}$ with multi-head attention (Vaswani et al., 2017).

Visual Encoder: Similar to (Anderson et al., 2018), we extract the image features by using a pre-trained Faster RCNN (Ren et al., 2015b). We select μ object proposals for each image, where each object proposal is represented by a 2048-dimension feature vector. We transform the obtained visual region features by a multi-layer perceptron and obtain the image features $\mathbf{I} = \mathbf{I}_{I=0}^{\mu} \in \mathbb{R}^{\mu \times d_q}$.

Prior Module: We use multi-head attention (Vaswani et al., 2017) as $f_{cv}(\cdot, \cdot)$ to manage the multi-modal interaction. A cross-attention layer is firstly applied to outputs of the texutal and visual encoders:

$$\mathbf{P} = \text{softmax}(\mathbf{I}\mathbf{x}^T) \in \mathbb{R}^{\mu \times \lambda}, \quad (4)$$

$$\mathbf{I}_x = \text{CrossAttn}(\mathbf{I}, \mathbf{x}) = \mathbf{P}\mathbf{x} \in \mathbb{R}^{\mu \times d_q}, \quad (5)$$

where the softmax conducts the normalization over each column of the matrix. We convert the representation $\hat{\mathbf{I}}$ into d_q-dimension vectors $\mathbf{V}$. This conversion is performed by a simple self-attention computation as follows:

$$\mathbf{g} = \text{softmax}(\text{ReLU}(\hat{\mathbf{I}}\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2), \quad (6)$$

where $\mathbf{g} \in \mathbb{R}^{\mu \times 1}$, $\mathbf{W}_1$, $\mathbf{W}_2$, $\mathbf{b}_1$, $\mathbf{b}_2$ are learned parameters. We obtain the representation $\mathbf{V}$ as follows:

$$\mathbf{v}_{prior} = \mathbf{g}^T\hat{\mathbf{I}} \in \mathbb{R}^{d_q}. \quad (7)$$

$\mathbf{g}$ is regarded as the prior distribution over visual objects.

Posterior Module: We simply utilize the add operation as $f_{cy}(\cdot, \cdot)$ to manage the interaction of $\mathbf{x}$ and $\mathbf{y}$:

$$\mathbf{x}_y = \mathbf{x} + \mathbf{y} \quad (8)$$

We replace $\mathbf{x}$ in Eq.(6) - Eq.(8) with $\mathbf{x}_y$ and thus obtain the posterior distribution $\mathbf{G}$ and $\mathbf{v}_{post}$

Generative and Discriminative Decoder: We utilize another LSTM as our discriminative and generative decoders following the previous studies (Das et al., 2017; Nguyen et al., 2020). Receiving the representation of context, images and the candidate answers, the two decoders compute the score of each candidate answer in different ways. The objective function of the base model is to minimize the negative log-likelihood $\mathcal{L}_{\mathcal{G}}$ of answer generated for the generative decoder or the cross-entropy loss $\mathcal{L}_{\mathcal{D}}$ for the discriminative decoder. We utilize the Kullback-Leibler (KL) divergence loss to narrow the gap. The objective functions of the student are as follows:

$$\mathcal{L} = \mathcal{L}_{\mathcal{G}} + \lambda\mathcal{L}_{KL}(\mathbf{G}, \mathbf{g}), \quad (9)$$

$$\mathcal{L} = \mathcal{L}_{\mathcal{D}} + \lambda\mathcal{L}_{KL}(\mathbf{G}, \mathbf{g}), \quad (10)$$

$$\mathcal{L} = \mathcal{L}_{\mathcal{G}} + \mathcal{L}_{\mathcal{D}} + \lambda\mathcal{L}_{KL}(\mathbf{G}, \mathbf{g}), \quad (11)$$

Model	VisDial v0.9 (val)					VisDial v1.0 (val)					
	MRR ↑	R@1 ↑	R@5 ↑	R@10 ↑	Mean ↓	NDCG ↑	MRR ↑	R@1 ↑	R@5 ↑	R@10 ↑	Mean ↓
MN (Das et al., 2017)	52.59	42.29	62.85	68.88	17.06	51.86	47.99	38.18	57.54	64.32	18.60
HCIAE (Lu et al., 2017)	53.86	44.06	63.55	69.24	16.01	59.70	49.07	39.72	58.23	64.73	18.43
CorefNMN (Kottur et al., 2018)	53.50	43.66	63.54	69.93	15.69	-	-	-	-	-	-
CoAtt (Wu et al., 2018)	54.11	44.32	63.82	69.75	16.47	59.24	49.64	40.09	59.37	65.92	17.86
RvA (Niu et al., 2019)	55.43	45.37	65.27	72.97	10.71	-	-	-	-	-	-
DVAN (Guo et al., 2019b)	55.94	46.58	65.50	71.25	14.79	-	-	-	-	-	-
Primary (Guo et al., 2019a)	-	-	-	-	-	-	49.01	38.54	59.82	66.94	16.60
ReDAN (Gan et al., 2019)	-	-	-	-	-	60.47	50.02	40.27	59.93	66.78	17.40
DMRM (Chen et al., 2020a)	55.96	46.20	66.02	72.43	13.15	-	50.16	40.15	60.02	67.21	15.19
DAM (Jiang et al., 2020c)	-	-	-	-	-	60.93	50.51	40.53	60.84	67.94	16.65
VDBERT (Wang et al., 2020)$^\diamond$	55.95	46.83	65.43	72.05	13.18	-	-	-	-	-	-
KBGN (Jiang et al., 2020a)	-	-	-	-	-	60.42	50.05	40.40	60.11	66.82	17.54
LTMI (Nguyen et al., 2020)†	55.85	46.07	65.97	72.44	14.17	61.61	50.38	40.30	60.72	68.44	15.73
LTMI-LG (Ours)	56.56	46.71	66.69	73.37	13.62	63.23	51.30	41.34	61.61	69.06	15.26
LTMI-LG* (Ours)	**56.59**	**46.87**	**66.92**	**73.76**	13.35	**63.53**	**51.43**	**41.68**	**61.96**	**69.87**	**14.89**

Table 1: Main comparisons on both VisDial v0.9 and v1.0 datasets using the generative decoder. † denotes that we re-implemented the model using the released code. ⋄ denotes that the model utilizes large extra datasets for training which is unfair compared with other models. ∗ denotes that we train the model using multi-task learning. Underline indicates the highest performance among previous approaches except for pretraining-based models. Our approach improves the strong baseline a lot. (t-test, p-value<0.01)

4 Experiments

4.1 Experiment Setup

Datasets and Implementation Details. We conduct experiments on the VisDial v0.9 and v1.0 datasets (Das et al., 2017) to verify our approach. VisDial v0.9 contains 83k dialogs on COCO-train (Lu et al., 2017) and 40k dialogs on COCO-val images as the test set, for a total of 1.23M dialog question-answer pairs. VisDial v1.0 dataset is an extension of VisDial v0.9 dataset with additional 10k COCO-like images. VisDial v1.0 dataset contains 123k, 2k, and 8k images as train, validation, and test splits, respectively.

To represent image regions, we use Faster R-CNN (Ren et al., 2015b) with ResNet-101 (He et al., 2016) finetuned on the Visual Genome dataset (Krishna et al., 2017), thus obtaining a 2048-dimension feature vector for each region. Following (Nguyen et al., 2020), we detect $\mu = 100$ objects from each image. Our model is implemented based on PyTorch (Paszke et al., 2017). In experiments, we use Adam (Kingma and Ba, 2014) optimizer for training, with the mini-batch size as 32. For the choice of the learning rate, we employ the warm-up strategy (Goyal et al., 2017). Specifically, we begin with a learning rate of 0.001, the learning rate is decreased by 1/4 for every 2 epochs up to 20 epochs. We use 4 Titan-XP GPU for training. We spend about 4 hours / 1 epoch for the discriminative setting and 1 hour / 1 epoch for the generative setting. Our student model is the same as LTMI, with the total parameters 42.20M. The λ sets to 1.

Automatic Evaluation. We use a retrieval setting to evaluate individual responses at each round of a dialog, following (Das et al., 2017). Specifically, at test time, apart from the image, ground truth dialog history and the question, a list of 100-candidate answers is also given. The model is evaluated on retrieval metrics: (1) Rank of human response, (2) Existence of the human response in $top - k$ ranked responses, i.e., R@k (3) Mean Reciprocal Rank (MRR) of the human response and (4) Normalized Discounted Cumulative Gain (NDCG) for VisDial v1.0.

Human Evaluation. We randomly extract 100 samples for human evaluation (Wu et al., 2018) and then ask 3 human subjects to guess whether the last response in the dialog is human-generated or machine-generated. If at least 2 of them agree it is generated by a human, we think it is human-generated (M1). We record the percentage of responses that are evaluated better than or equal to human responses (M2), according to the human subjects' evaluation.

4.2 Main Results

Baseline methods. In our experiment, compared methods can be grouped into four types: (1) Fusion-based models: LF (Das et al., 2017) and HREA (Das et al., 2017). (2) Attention-based models: HCIAE (Lu et al., 2017), CoAtt (Wu et al., 2018), Primary (Guo et al., 2019a), ReDAN (Gan et al., 2019), CorefNMN (Kottur et al., 2018),

Model	VisDial v0.9 (val)					VisDial v1.0 (test-std)					
	MRR ↑	R@1 ↑	R@5 ↑	R@10 ↑	Mean ↓	NDCG ↑	MRR ↑	R@1 ↑	R@5 ↑	R@10 ↑	Mean ↓
ReDAN (Gan et al., 2019)	-	-	-	-	-	57.63	64.75	51.10	81.73	90.90	3.89
MCA (Agarwal et al., 2020)	-	-	-	-	-	72.73	37.68	20.67	56.67	72.12	8.89
GNN-EM (Zheng et al., 2019)	62.85	48.95	79.65	88.36	4.57	52.82	61.37	47.33	77.98	87.83	4.57
DualVD (Jiang et al., 2020b)	62.94	48.64	80.89	89.94	4.17	56.32	63.23	49.25	80.23	89.70	4.11
FGA (Schwartz et al., 2019)	65.25	51.43	82.08	89.56	4.35	56.90	66.20	52.75	82.92	91.07	3.80
CAG (Guo et al., 2020)	67.56	54.64	83.72	91.48	3.75	56.64	63.49	49.85	80.63	90.15	4.11
KBGN (Jiang et al., 2020a)	-	-	-	-	-	57.60	64.13	50.47	80.70	90.16	4.08
VisualBERT (Murahari et al., 2020)◇	-	-	-	-	-	74.47	50.74	37.95	64.13	80.00	6.28
VDBERT (Wang et al., 2020)◇	70.04	57.79	85.34	92.68	4.04	75.35	51.17	38.90	62.82	77.98	6.69
LTMI (Nguyen et al., 2020)†	66.41	53.36	82.53	90.54	4.03	60.92	60.65	47.00	77.03	87.75	4.90
LTMI-LG (Ours)	67.63	54.69	83.74	91.38	**3.75**	58.55	64.00	50.63	80.58	90.20	4.12

Table 2: Main comparisons on both VisDial v0.9 and v1.0 datasets using the discriminative decoder. ◇ denotes that the model utilizes large extra datasets for training which is unfair compared with other models. Underline indicates the highest performance among previous approaches except for the pretraining-based models. Our approach improves the strong baseline significantly. (t-test, p-value<0.01)

Model	w/o Ans	w/ Ans	with LG
LTMI (Nguyen et al., 2020)	68.6	97.1	82.1
Random	3.6	3.6	-
Human	96.7	99.3	-

Table 3: Accuracy of visual grounding with and without knowing the answer. We randomly sample 1000 samples and ask human annotators to ground the three most likely objects from the image.

Model	NDCG	MRR	R@1	R@5	R@10	Mean
LTMI	61.61	50.38	40.30	60.72	68.44	15.73
LTMI-Mean	56.66	43.64	32.59	54.66	62.91	17.59
LTMI-Random	56.89	43.79	33.01	54.47	62.76	17.76
LTMI-LG	63.23	51.30	41.34	61.61	69.06	15.26
LTMI-Human	70.10	63.96	50.74	69.29	80.02	8.12

Table 4: Effects of different visual objects distribution.

Model	NDCG	MRR	R@1	R@5	R@10	Mean
MN (Das et al., 2017)	-	60.29	46.14	77.68	87.57	4.84
HCIAE (Lu et al., 2017)	-	61.96	48.25	78.97	88.43	4.56
CoAtt (Wu et al., 2018)	-	62.77	49.38	78.99	88.49	4.56
ReDAN (Gan et al., 2019)	-	64.29	50.65	81.29	90.17	4.10
KBGN (Jiang et al., 2020a)	59.08	64.86	51.37	81.71	90.54	4.00
VDBERT (Wang et al., 2020)‡	56.20	62.25	48.16	79.57	89.01	4.31
VDBERT (Wang et al., 2020)◇	63.22	67.44	54.02	83.96	92.33	3.53
LTMI (Nguyen et al., 2020)†	62.72	62.32	48.94	78.65	87.88	4.86
LTMI-LG (Ours)	59.67	65.03	51.69	81.49	90.32	4.02

Table 5: Main comparisons on VisDial v1.0 val datasets using the discriminative decoder. ◇ denotes that the model utilizes large extra datasets for training. ‡ denotes that the model trains from scratch.

RvA (Niu et al., 2019), DVAN (Guo et al., 2019b) and DMRM (Chen et al., 2020a), DAM (Jiang et al., 2020c). (3) The pretraining model: VDBERT (Wang et al., 2020) and VisualBERT (Murahari et al., 2020). (4) Graph-based models: GNN (Zheng et al., 2019), DualVD (Jiang et al., 2020b), FGA (Schwartz et al., 2019), KBGN (Jiang et al., 2020a).

We realize our model LTMI-LG which is based on the strong baseline LTMI (Nguyen et al., 2020)[1]. LTMI is a very strong model which achieves some the-state-of-the-art results. In general, our approach brings a large improvement to the strong baseline LTMI, which shows the effectiveness of our answer-aware knowledge distillation. We use t-test and analysis of variance (ANOVA) to analyze our model and LTMI. The p-values of these two analytical methods are all less than 0.01, indicating that the results are significantly different.

As shown in Table 3, we statistic the accuracy of grounding visual objects of our LTMI-LG, which is 82.1%. Our answer-aware knowledge distillation improves the accuracy from 68.6% (LTMI) to 82.1% (LTMI-LG), gaining 13.5% improvement. As shown in Figure 5, we provide predicted answers by LTMI and our LTMI-LG. Due to the improvement of visual grounding, our approach improves the generative and retrieval results of LTMI, managing to locate visual objects more accurately, as shown in Table 4. "Mean" denotes We set the distribution of the visual objects to uniform to make all visual objects have the same weights. "Random" denotes we randomize the distribution in-batch. "Human" denotes we annotate 100 images and utilize this distribution to generate responses. The more appropriate visual objects, the better the model performance.

Generative Results. As shown in Table 1, we compare the generative performance among differ-

[1]We reproduce the result for LTMI by their official GitHub repo (https://github.com/davidnvq/visdial). We apply the default hyper-parameters as them.

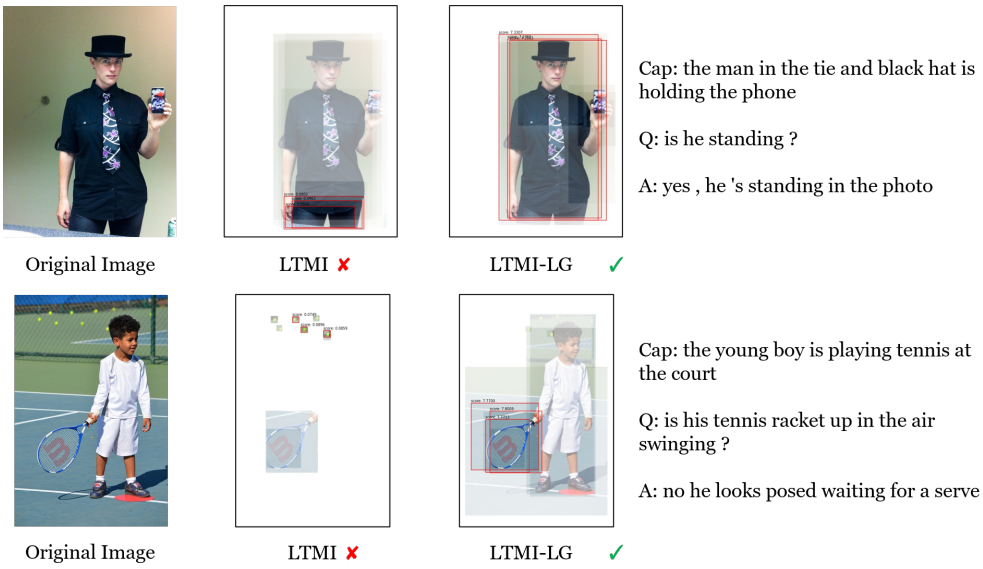

Figure 4: Visualization of attention maps generated by LTMI and our approach. Our approach grounds the related visual objects more accurately than LTMI.

ent methods on the VisDial v1.0 val and VisDial v0.9 val. With the guidance of the teacher, we train our LTMI-LG with the ability of accurately grounding visual objects. As a result, our approach improve significantly (nearly 1% on all metrics) compared with LTMI (Nguyen et al., 2020). Comparing with the state-of-the-art results of different metrics, our model improves NDCG for 61.51 to 63.53 (+1.92), MRR from 50.51 to 51.43 (+0.92), R@1 from 40.53 to 41.68 (+1.13), R@5 from 60.84 to 61.96 (+1.12), R@10 from 68.44 to 69.87 (+1.43), Mean from 15.73 to 14.89 (+0.84) on the VisDial v1.0 val. Our model also brings a large improvement to LTMI (Nguyen et al., 2020) on the VisDial v0.9 val. The performance of our model exceeds the performance of VDBERT (Wang et al., 2020)$^\diamond$ on all the metrics except Mean. We believe data is an important factor in deep learning (LeCun et al., 2015). VDBERT (Wang et al., 2020)$^\diamond$ works because it uses a lot of extra data for training. The reason why our method is effective is that we use the teacher to teach the student visual grounding, which can be regarded as a kind of data annotation.

Discriminative Results. As shown in Table 2 and Table 5, we compare our method with previous works on the VisDial v1.0 test, VisDial v0.9 val and VisDial v1.0 val. Our model improves significantly compared with LTMI (Nguyen et al., 2020), improving about +3% on MRR, R@1, R@5 and R@10 on the VisDial v1.0 test. Our approach also brings a large improvement on the Visdial v0.9 and achieves the best results on MRR, R@1 and R@5 among non-pre-trained models. In a discriminative setting, our approach performs worse than pre-training models VisualBERT (Murahari et al., 2020)$^\diamond$ and VDBERT (Wang et al., 2020)$^\diamond$ because pre-training models utilize extra large-scale datasets to train the models which are unfair compared with other models. As shown in Table 5, VDBERT‡ which trains from scratch performs worse than our LTMI-LG.

Model	NDCG	MRR	R@1	R@5	R@10	Mean
LTMI†	61.61	50.38	40.30	60.72	68.44	15.73
LG-Attn-MSE	63.03	51.14	40.91	61.78	69.43	15.08
LG-Image-MSE	62.80	51.21	41.01	**62.02**	**69.90**	**14.90**
LG-Attn-KL	**63.23**	**51.30**	**41.34**	61.61	69.06	15.26
LG-Image-KL	62.36	51.24	41.19	61.73	69.31	15.13
Attn-KL-Image-MSE	62.73	51.19	40.97	61.80	69.43	15.10

Table 6: Ablation study on VisDial v1.0 val datasets using the generative decoder.

4.3 Ablation Study

In order to transfer knowledge, we need a metric loss to measure the gap between teachers and students. In our main experiments, we utilize the kullback-leibler (KL) divergence loss to diminish the gap of the weight distribution between the student model and the teacher model, the mean squared loss to diminish the gap of the representation of images. To compare different losses, we utilize the mean squared loss for attention maps and KL loss for the representation of images as shown in Table 6. We find that KL loss is more suitable for attention distribution (better for NDCG, MRR and

a teenage girl in with long curly red hair who is on the phone

Question	Ground Truth	Generation by LTMI-LG	Generation by LTMI	Retrieval by LTMI-LG	Retrieval by LTMI
Q1: what kind of phone ?	a cell phone	looks flip phone	looks flip phone	a cell phone	ca n't tell
Q2: what color is the phone ?	it is black	black	black 's black	it is black	it 's black
Q3: what color is the girl 's shirt ?	white	white	it	white or light blue shirt	there is n't 1
Q4: does she have a coat on ?	no she does n't	no she does n't	no	no she does n't	no
Q5: is it sunny ?	yes	yes	yes it sunny	yes	yes it is

Figure 5: Examples of dialogs generated and retrieved by our model and the LTMI baseline. Our model provides answers that are more accurate than LTMI (green denotes correct answers, and red denotes wrong answers).

R@1) and MSE loss for the representation (better for R@5, R@10 and Mean). In addition, we use the attention via KL loss and the representation via MSE loss for distillation at the same time. The result is not so satisfactory and we think these two methods have some redundancy.

4.4 Case Study

As shown in Figure 4, we visualize the learned attention maps to understand the model. The colorful region means higher attention weights. We draw the bounding boxes of the first three highest scores. As shown in the top image in Figure 4, the question "is he standing ?" indicates the man's overall posture rather than the local. LTMI grounds the wrong visual objects while our model grounds the right objects. As shown in the bottom image in Figure 4, the question "is his tennis racket up in the air swinging ?" concerns the racket rather than the tennis balls. Our model grounds accurately while LTMI makes mistakes. These examples show that our LTMI-LG has learned the ability to ground visual objects via our answer-aware knowledge distillation.

4.5 Human Study

As shown in Table 7, we conduct human study to further demonstrate the effectiveness of our model. Our model achieves the highest scores both on the metric M1 and M2 compared with LTMI.

	LTMI	LTMI-LG
Method 1 (M1)	56	66
Method 2 (M2)	61	69

Table 7: Human evaluation on 1000 sampled responses on VisDial val v1.0. M1: percentage of responses which are human-generated. M2: percentage of responses evaluated better than or equal to human responses.

5 Related Work

Recent several works (Shuster et al., 2018; Liang et al., 2021; Yang et al., 2020) explore leveraging visual information to enhance dialogue models. While visual dialog models focus on the intersection of questions, history and images. How to locate the related visual objects is quite important. MN (Das et al., 2017), HCIAE (Lu et al., 2017), CorefNMN (Kottur et al., 2018), CoAtt (Wu et al., 2018), RvA (Niu et al., 2019), DVAN (Guo et al., 2019b) utilize kinds of attention mechanisms as the backbone to locate the related visual objects. VisualBERT (Murahari et al., 2020) and VDBERT (Wang et al., 2020) exploit large extra datasets to explore in visual dialog via pretraining language models. GNN-EM (Zheng et al., 2019), FGA (Schwartz et al., 2019), DualVD (Jiang et al., 2020b), CAG (Guo et al., 2020) and KBGN (Jiang et al., 2020a) utilize graph neural networks to obtain the representation of visual objects. However, most existing visual dialog models condition visual objects simply on history and questions, which we regard as a prior distribution over visual objects. In this paper, we propose an approach to learn to ground visual objects via bridge the gap between the prior distribution and the posterior distribution.

6 Conclusion

In this paper, we propose a novel approach to learn to ground visual objects for visual dialog, which employs a novel visual objects grounding mechanism where both prior and posterior distributions over visual objects are used to facilitate visual objects grounding. Experimental results on two large-scale datasets show that our approach improves the previous models by a significant margin.

References

Shubham Agarwal, Trung Bui, Joon-Young Lee, Ioannis Konstas, and Verena Rieser. 2020. History for visual dialog: Do we really need it? *arXiv preprint arXiv:2005.07493*.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. SPICE: Semantic propositional image caption evaluation. *Adaptive Behavior*, 11(4):382–398.

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.

Feilong Chen, Xiuyi Chen, Fandong Meng, Peng Li, and Jie Zhou. 2021a. Gog: Relation-aware graph-over-graph network for visual dialog. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*.

Feilong Chen, Fandong Meng, Xiuyi Chen, Peng Li, and Jie Zhou. 2021b. Multimodal incremental transformer with visual grounding for visual dialogue generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*.

Feilong Chen, Fandong Meng, Jiaming Xu, Peng Li, Bo Xu, and Jie Zhou. 2020a. DMRM: A dual-channel multi-hop reasoning model for visual dialog. *Thirty-Fourth AAAI Conference on Artificial Intelligence*.

Xiuyi Chen, Feilong Chen, Fandong Meng, Peng Li, and Jie Zhou. 2021c. Unsupervised knowledge selection for dialogue generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*.

Xiuyi Chen, Fandong Meng, Peng Li, Feilong Chen, Shuang Xu, Bo Xu, and Jie Zhou. 2020b. Bridging the gap between prior and posterior knowledge selection for knowledge-grounded dialogue generation. In *EMNLP*.

Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. 2020. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

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.

Zhe Gan, Yu Cheng, Ahmed EI Kholy, Linjie Li, Jingjing Liu, and Jianfeng Gao. 2019. Multi-step reasoning via recurrent dual attention for visual dialog. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6463–6474.

Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. 2015. Are you talking to a machine? dataset and methods for multilingual image question. In *Advances in Neural Information Processing Systems*, pages 2296–2304.

Mehdi Ghanimifard and Simon Dobnik. 2019. What goes into a word: generating image descriptions with top-down spatial knowledge. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 540–551.

Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. 2017. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*.

Dalu Guo, Chang Xu, and Dacheng Tao. 2019a. 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, and Meng Wang. 2019b. Dual visual attention network for visual dialog. pages 4989–4995.

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.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.

Qingbao Huang, Jielong Wei, Yi Cai, Changmeng Zheng, Junying Chen, Ho-fung Leung, and Qing Li. 2020. Aligned dual channel graph convolutional network for visual question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7166–7176.

Xiaoze Jiang, Siyi Du, Zengchang Qin, Yajing Sun, and Jing Yu. 2020a. KBGN: Knowledge-bridge graph network for adaptive vision-text reasoning in

visual dialogue. *Proceedings of the 28th ACM International Conference on Multimedia*.

Xiaoze Jiang, Jing Yu, Zengchang Qin, Yingying Zhuang, Xingxing Zhang, Yue Hu, and Qi Wu. 2020b. DualVD: An adaptive dual encoding model for deep visual understanding in visual dialogue. In *AAAI*, volume 1, page 5.

Xiaoze Jiang, Jing Yu, Yajing Sun, Zengchang Qin, Zihao Zhu, Yue Hu, and Qi Wu. 2020c. DAM: Deliberation, abandon and memory networks for generating detailed and non-repetitive responses in visual dialogue. *arXiv preprint arXiv:2007.03310*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Satwik Kottur, José M. F. Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. 2018. Visual coreference resolution in visual dialog using neural module networks. *ArXiv*, abs/1809.01816.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature*, 521(7553):436–444.

Linjie Li, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019. Relation-aware graph attention network for visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10313–10322.

Zujie Liang, Huang Hu, Can Xu, Chongyang Tao, Xiubo Geng, Yining Chen, Fan Liang, and Daxin Jiang. 2021. Maria: A visual experience powered conversational agent. *arXiv preprint arXiv:2105.13073*.

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.

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.

Vishvak Murahari, Dhruv Batra, Devi Parikh, and Abhishek Das. 2020. Large-scale pretraining for visual dialog: A simple state-of-the-art baseline. *Proceedings of the European Conference on Computer Vision*.

Van-Quang Nguyen, Masanori Suganuma, and Takayuki Okatani. 2020. Efficient attention mechanism for visual dialog that can handle all the interactions between multiple inputs. *Proceedings of the European Conference on Computer Vision*.

Yulei Niu, Hanwang Zhang, Manli Zhang, Jianhong Zhang, Zhiwu Lu, and Ji-Rong Wen. 2019. Recursive visual attention in visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6679–6688.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch.

Jiaxin Qi, Yulei Niu, Jianqiang Huang, and Hanwang Zhang. 2020. Two causal principles for improving visual dialog. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Mengye Ren, Ryan Kiros, and Richard Zemel. 2015a. Exploring models and data for image question answering. In *Advances in Neural Information Processing Systems*, pages 2953–2961.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015b. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, pages 91–99.

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.

Kurt Shuster, Samuel Humeau, Antoine Bordes, and Jason Weston. 2018. Image chat: Engaging grounded conversations. *arXiv preprint arXiv:1811.00945*.

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.

Yue Wang, Shafiq Joty, Michael R Lyu, Irwin King, Caiming Xiong, and Steven CH Hoi. 2020. VD-BERT: A unified vision and dialog transformer with bert. *arXiv preprint arXiv:2004.13278*.

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 *Proceedings of International Conference on Machine Learning*, pages 2048–2057.

Ze Yang, Wei Wu, Huang Hu, Can Xu, and Zhoujun Li. 2020. Open domain dialogue generation with latent images. *arXiv preprint arXiv:2004.01981*.

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.