




UED: A Unified Encoder Decoder Network for Visual Dialog

Cheng Chen and Xiaodong Gu^(✉) 

Department of Electronic Engineering, Fudan University, Shanghai 200433, China
xdgu@fudan.edu.cn

Abstract. This paper addresses the problem of visual dialog, which aims to answer multi-round questions based on the dialog history and image content. This is a challenging task because a question may be answered in relations to any previous dialog and visual clues in image. Existing methods mainly focus on discriminative setting, which design various attention mechanisms to model interaction between answer candidates and multi-modal context. Despite having impressive results with attention based model for visual dialog, a universal encoder-decoder for both answer understanding and generation remains challenging. In this paper, we propose UED, a unified framework that exploits answer candidates to jointly train discriminative and generative tasks. UED is unified in that (1) it fully exploiting the interaction between different modalities to support answer ranking and generation in a single transformer based model, and (2) it uses the answers as anchors to facilitate both two settings. We evaluate the proposed UED on the VisDial dataset, where our model outperforms the state-of-the-art.

Keywords: Visual dialog · Cross modal learning · Encoder decoder network

1 Introduction

Visual dialog is recently introduced by Abhishek et al. [2]. Compared with visual question answering, it requires the agent to communicate with human about an image in multiple rounds.

Most of the current visual dialog model focus on modeling the interaction between answer candidates, current question, previous dialog history and image. Nevertheless, the answer candidates are invisible in generative setting, how to learn a unified model that can capture such interaction for both answer ranking and generation settings is a seldom explored territory.

In this work, we formulate the interaction of all entities in discriminative setting using a pretrained transformer. As shown in Fig. 1, in discriminative setting the agent infers whether the answer candidate is the correct one with the powerful representation yielded by fully attention of each entities. Inspired by the recent success of visual and language pretraining, transformer is employed

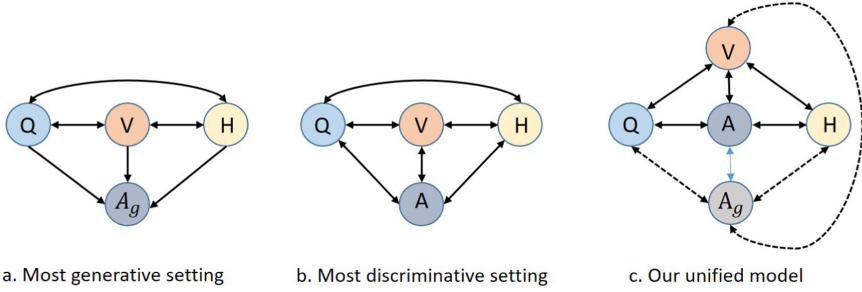


Fig. 1. Interaction flow direction illustration. Q: question, V: image, H: dialog history, A: answer candidates, A_g : generated answer.

as the encoding backbone as it has natural capability of capturing interaction between different entities from different modalities. As aforementioned, generative setting can only employ information contained in textual context and image to reconstruct the answer. As shown in Fig. 1, the interaction between generated answer and multi-modal context is unidirectional.

To leverage the discriminative clues in answer candidates for easing the difficulty of answer generation, we employ the answer candidates used in discriminative setting as anchor points to promote the bi-directional interaction between generated answer and other entities as shown in Fig. 1. Noted that, the attention flow from multi-modal to generated answer is explicit and the reverse attention is implicitly performed by anchor answer A. More specifically, a contrastive loss is devised to preserve the similarity of generated answer features and the target answer, while distinguishing other answer options. This also leads to the elegant view of how to bridge the discrepancy between discriminative and generative settings, and how to fully exploit the clues in answer candidates. The main contributions of this paper are as follows.

- (1) We introduce a unified model for visual dialog, which processes all interactions between different entities for both discriminative and generative settings.
- (2) The target answers is employed as anchor points to help both of the encoder and decoder for distinguishing the answer options with complex semantics. Compared to previous methods, the contrastive loss enables the bidirectional attention flow between all answer candidates and generated answer features to learn discriminative features for distinguishing the answers with complex semantics.
- (3) Extensive experiments were performed on visual dialog benchmark [2], and the qualitative results indicate that our model obtains reliable improvement on both tasks by unified contrastive learning.

2 Proposed Method

2.1 Problem Formulation

We first formally describe the visual dialog problem. Given a question Q_t grounded on an image I at t -th turn, as well as the previous dialog history formulated as $H_t = \{C; (Q_1; A_1), \dots, (Q_{t-1}; A_{t-1})\}$ (where C denotes the caption sentence of the image), our task aims to predict the target answer A_t by ranking a list of 100 answer candidates $\{A_t^1, A_t^2, \dots, A_t^{100}\}$ in discriminative setting or generate the required answer in generative setting.

2.2 Cross Modal Extractor Backbone

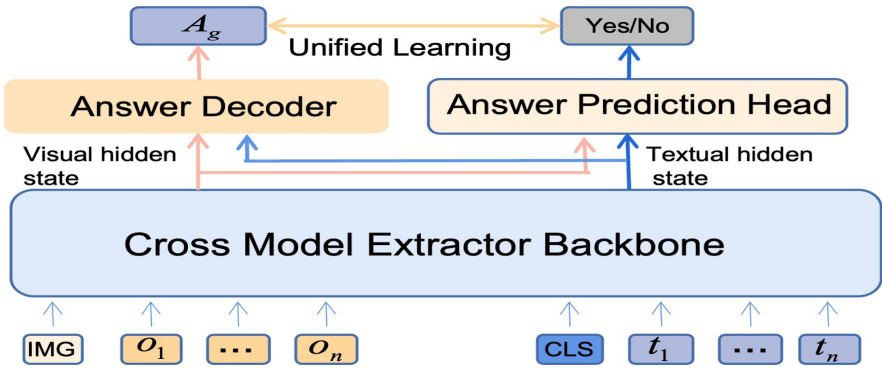


Fig. 2. The framework of our UED for unified generative and discriminative learning.

To jointly learn these the tasks in an end-to-end framework, ViLBERT [1] is adopted as the backbone network to extract cross modal features. ViLBERT is a two stream pretrained multi-modal network, which can jointly model visual and linguistic inputs by employing co-attention layers. Noted that, any pretrained multi-modal architecture can be adopted to our method. Following ViLBERT, we embedded the visual and text sequence as $I = \{[IMG]O_1, \dots, O_n\}$ and $D = \{[CLS]C[SEP]Q_1[SEP]A_1, \dots, Q_t[SEP]A_t[SEP]\}$, here I is object features extracted by Faster R-CNN. We feed the two sequences into ViLBERT and obtain output textual hidden state and visual hidden state as:

$$D_h, I_h = \text{ViLBERT}(D, I), \quad (1)$$

where $D_h = \{d_1, \dots, d_t\}$ and $I_h = \{i_1, \dots, i_n\}$.

As ViLBERT contains multiple transformer blocks and cross attention blocks, the yielded feature D_h and I_h contains deep fused cross modal features.

2.3 Unified Learning of Two Tasks

Given the learned cross modal features D_h and I_h , we rank the answer candidates through the Next Sentence Prediction (NSP) loss.

The NSP loss is trained to predict 1 when the target answer A_t is appended, and 0 when a negative answer A_n sampled from other answer candidates is appended to it. To autoregressively generate an answer, we also train UED with the textual input with answer mask:

$$D_g = \{[CLS]C[SEP]Q_1[SEP]A_1, \dots, Q_t[SEP][MASK]\}, \quad (2)$$

where the answer tokens is replaced by the special [MASK] token to make it blind to encoder. The hidden state yielded by ViLBERT D_g and I_h are fed to the decoder to generate the answer A_g .

To model the cross-impact and interaction between the two tasks, we enable the task-specific answer representations interact with each other via contrastive training. Specifically, the answer representations in discriminative setting is divided into two part, where the target answer representations A_p is regarded as positive query feature, and the negative answer representations together with all answer options in other dialog within a mini batch is regarded as negative key features $A_n = \{A_{n1}, \dots, A_{nn}\}$.

As decoder aims to generate target answer, the answer features A_g generated by it requires to semantically correspond to A_p . To encourage the decoder interact with all other answer information and optimize the two tasks simultaneously, we leverage the target answer as anchor and define a contrastive loss to transfer useful mutual information between two tasks. The contrastive loss is thus defined as:

$$L_c = \frac{\exp(A_p \cdot A_g / \tau)}{\sum_{i=0}^{n-1} \exp(A_p \cdot A_{ni} / \tau)}, \quad (3)$$

where τ is a temperature parameter.

2.4 Visually Grounded Training Objectives

During the training of UED, We use two visually grounded training objectives masked language modeling (MLM) and next sentence prediction (NSP) to supervise the cross modal extractor backbone ViLBERT.

Similar to MLM in BERT, 10% tokens in textual input and 15% tokens in visual input are randomly masked out and replaced with a special token [MASK]. The model is required to recover them based not only on the surrounding tokens and the cross modal clues:

$$L_{mlm} = -E_{(D,I) \sim T} \log P(W_m | D_{\setminus m}, I_{\setminus m}), \quad (4)$$

where W_m is the masked tokens and T refers to the training set.

The NSP loss is implemented as:

$$L_{nsp} = -E_{(D,I) \sim T} \log P(y | N(D, I)), \quad (5)$$

where $y \in \{0, 1\}$ serves as the supervision label, and $N(\cdot)$ is the binary next sentence prediction head to predict the probability based on the dot product of [CLS] representation in text features and [IMG] representation in image features.

For the generative setting, the decoder is required to reconstruct the sequential target answer tokens depending on all the dialog context and input image. The loss is defined as maximum log-likelihood loss:

$$L_g = -E_{(D,I) \sim T} \log P(A|D \setminus A, I), \quad (6)$$

The overall objective is expressed as:

$$L_{ued} = L_{mlm} + L_{nsp} + \alpha L_g + L_c, \quad (7)$$

where $\alpha = 0.05$ is the weighting parameter.

3 Experiments

3.1 Dataset

The VisDial v1.0 dataset is used in our experiments. It consists of 123,287 images in the training set, 2064 in the validation set, and 8,000 images in the testing set. Each image is associated with a caption sentence and 10 question answer pairs. For each round of question answer pair, 100 answer candidates are given.

3.2 Evaluation Metric

Following previous works [2,3], the ranking metrics like Recall@K (K=1, 5, 10), Mean Reciprocal Rank (MRR), and Mean Rank is adopted. Since the 2018 VisDial challenge releases the dense annotations of each answer option’s relevance degree, normalized discounted cumulative gain (NDCG) that penalizes the lowranked answer options with high relevance is also used.

3.3 Implementation Details

We use ViLBERT base as the backbone, which has 12 layers of transformer blocks with each block having a hidden state size of 768 and 12 attention heads. The decoder consists of 12 layers of transformer blocks, each block has hidden size of 1024 and 16 attention heads. The max text sequence length is 256. We train on 8 V100 GPUs with a batch size of 120 for 20 epochs. The Adam optimizer with initial learning rates of 2e-4 is adopted. A linear decay learning rate schedule with warm up is employed to train the model.

3.4 Comparison to State-of-the-Art Methods

We compare our method with recently published methods, including MN [2], FGA [3], CoAtt [4], HCIAE [5], ReDAN [6], LTMI [7], VDBERT [8], DAN [9], Synergistic [10], GNN [11]. Tables 1, and Table 2 summarize the results on the

aforementioned benchmark. Follow previous works [2, 4], comparison of the generative setting is performed on val split of the dataset. We select here MN, CoAtt, HCIAE, and ReDAN for comparison of generative setting, as their performances of both settings in all metrics are available in the literature. Among all evaluation metrics, our UED significantly outperforms other models, even including some ensemble variants such as Synergistic and ReDAN. Notably, our model significantly surpasses the state-of-arts by more than 1 points absolute improvements under the metrics Recall@1 in both discriminative and generative settings. Moreover, the performance improvements under strict ranking metrics are more obvious (e.g., Recall@1, MRR).

As aforementioned, UED supports ranking the answer candidates and generating answer in a single pass, the two tasks are jointly trained by unified contrastive loss. As the results show, the generative setting surpasses the state of art by a large margin, which indicates the contrastive loss enables the decoder to perceive more discriminative information from the rich answer candidates and our model is able to perform well in both task.

Table 1. Performance comparisons of discriminative setting on the test-std split of VisDial v1.0 dataset. The top 1 results are highlighted by **bold**.

Methods	R@1↑	R@5↑	R@10↑	NDCG↑	MRR↑	Mean↓
MN	40.98%	72.30%	83.30%	47.50	55.49	5.92
FGA	49.58%	80.97%	88.55%	52.10	63.70	4.51
GNN	47.33%	77.98%	87.83%	52.82	61.37	4.57
MN-Att	42.42%	74.00%	84.35%	49.58	56.90	5.59
ReDAN	42.45%	64.68%	75.68%	64.47	53.73	6.64
LTMI	50.20%	80.68%	90.35%	59.03	64.08	4.05
VDBERT	51.63%	82.23%	90.68%	59.96	65.44	3.90
DAN	49.63%	79.75%	89.35%	57.59	63.20	4.30
Synergistic	47.90%	80.43%	89.95%	57.32	62.20	4.17
Ours – UED	51.73%	82.42%	91.13%	60.22	65.86	3.78

Table 2. Performance comparisons of generative setting on the val-std split of VisDial v1.0 dataset. The top 1 results are highlighted by **bold**.

Methods	R@1 ↑	R@5↑	R@10↑	NDCG↑	MRR↑	Mean↓
MN	38.01%	57.49%	64.08%	56.99	47.83	18.76
CoAtt	40.09%	59.37%	65.92%	59.24	49.64	17.86
HCIAE	39.72%	58.23%	64.73%	59.70	49.07	18.43
ReDAN	40.27%	59.93%	66.78%	60.47	50.02	17.40
Ours – UED	41.89%	61.07%	67.12%	61.21	51.11	17.12

3.5 Ablation Studies

In this section, we perform ablation studies to evaluate the effects of different training settings. We first remove the decoder used for generative setting, and the results are shown in row 1. Comparing row 1 and row 4, it can be observed that training generative task brings improvements to ranking task.

In row 2, we vary the setting of decoder size. Specifically, a light decoder which has 8 layers of transformer blocks with each block having a hidden state size of 768 and 16 attention heads is adopted. The results shows that decoder size has little impact to the results. The reason is that decoder is not pretrained on large dataset.

The main characteristic of UED is the unified contrastive loss, which combines all answer candidates and generated answer to learn more useful clues. To study the impact of the contrastive loss alone, we train our UED without it and report the result in row 3. Comparing to the full model with contrastive loss (row 4), row 3 gets worse performance across the ranking metrics, which further verifies the effectiveness of contrastive loss. The full model UED gets highest results in all metrics.

Table 3. Ablation studies on the VisDial v1.0 dataset

Row	Methods	R@1↑	R@5↑	R@10↑	NDCG↑	MRR↑	Mean↓
1	UED-w/o-decoder	53.58%	83.86%	91.93%	60.02	64.79	3.83
2	UED-lidecoder	53.92%	84.08%	92.08%	60.86	64.97	3.81
3	UED-w/o- L_c	53.78%	83.98%	92.06%	60.02	64.88	3.88
4	Ours – UED	54.08%	84.32%	92.31%	61.06	65.48	3.71

3.6 Qualitative Result

We illustrate some qualitative examples of our UED in Fig. 3. Evidently, training with contrastive loss can produce more accurate result. Unified training of two tasks helps our model distinguish the target answer from the answers with similar semantics with the ground truth answer. It is very difficult for the model to predict the answer without proper reference to the visual information. As



Q1: Is there any person in the scene?
 A1: Yes, there is one.
 Q2: What is he doing?
 A2: He is playing with the cat and dog. (GT)

- | | |
|--|--|
| <ol style="list-style-type: none"> 1. He is playing with the cat. 2. He is playing with the cat and dog. (GT) 3. Not that i can see. 4. I cannot tell. | <ol style="list-style-type: none"> 1. He is playing with the cat and dog. (GT) 2. Not that i can see. 3. He is playing with the cat. 4. I cannot tell. |
|--|--|

Base Model

W/Unified Learning

Fig. 3. The effects of unified learning of two tasks in our UED.

our model exploits rich information from all answer candidates and generated answer. It performs better than the baseline.

4 Conclusion

In this paper, we study the problem of visual dialog. A unified transformer model UED that exploits the answers as anchor points to jointly train discriminative and generative tasks. UED is capable of modeling all the interactions between all answer candidates and the generated answer to supervise the training of two tasks via simple unified contrastive learning. Moreover, it can rank or generate answers seamlessly in one single pass, and the training of two tasks is simultaneous. Experiments on visual dialog benchmark show the effectiveness of the proposed model, and more extensive ablation studies further confirm the correlation between two tasks and reveal that modeling the relations explicitly by employing answers as anchor points can improve their performance.

Acknowledgments. This work was supported in part by National Natural Science Foundation of China under grant 61771145.

References

1. Lu, J., Batra, D., Parikh, D., Lee, S.: ViLBERT: pretraining task-agnostic visiolinguistic representations for vision and-language tasks. In: 2019 Advance in Neural Information Processing Systems (NIPS), pp. 524–534. MIT Press, Vancouver, CA (2019)
2. Das, A., et al.: Visual dialog. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 326–335. IEEE, Honolulu, HI (2017)
3. Schwartz, I., Yu, S., Hazan, T., Schwing, A.G.: Factor graph attention. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2039–2048. IEEE, Long Beach, CA (2019)
4. Wu, Q., Wang, P., Shen, C., Reid, I., van den Hengel, A.: Are you talking to me? Reasoned visual dialog generation through adversarial learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6106–6115. IEEE, Salt Lake City, Utah (2018)
5. Lu, J., Kannan, A., Yang, J., Parikh, D., Batra, D.: Best of both worlds: transferring knowledge from discriminative learning to a generative visual dialog model. In: Advances in Neural Information Processing Systems, pp. 314–324. MIT press, California, USA (2017)
6. Gan, Z., Cheng, Y., Kholy, A.E., Li, L., Liu, J., Gao, J.: Multi-step reasoning via recurrent dual attention for visual dialog. In: Proceedings of the Conference of the Association for Computational Linguistics, pp. 6463–6474. ACL, Florence, ITA (2019)
7. Nguyen, V.-Q., Suganuma, M., Okatani, T.: Efficient attention mechanism for visual dialog that can handle all the interactions between multiple inputs. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12369, pp. 223–240. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58586-0_14

8. Wang, Y., Joty, S., Lyu, M.R., King, I., Xiong, C., Hoi, S.C.: VD-BERT: a unified vision and dialog transformer with BERT. In: 2020 ACL Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 3325–3338. ACL (2020)
9. Kang, G.C., Lim, J., Zhang, B.T.: Dual attention networks for visual reference resolution in visual dialog. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 2024–2033. ACL, Hong Kong (2019)
10. Guo, D., Xu, C., Tao, D.: Image-question-answer synergistic network for visual dialog. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 10434–10443. IEEE, Long Beach, CA (2019)
11. Zheng, Z., Wang, W., Qi, S., Zhu, S.C.: Reasoning visual dialogs with structural and partial observations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6669–6678. IEEE, Long Beach, CA (2019)