IMPROVING CROSS-MODAL UNDERSTANDING IN VISUAL DIALOG VIA CONTRASTIVE LEARNING

Feilong Chen[1,2], *Xiuyi Chen*[1,3], *Shuang Xu*[1], *Bo Xu*[1,2,3]

[1]Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2]School of Future Technology, [3]University of Chinese Academy of Sciences, Beijing, China
{chenfeilong2018, chenxiuyi2017, shuang.xu, xubo}@ia.ac.cn

ABSTRACT

Visual Dialog is a challenging vision-language task since the visual dialog agent needs to answer a series of questions after reasoning over both the image content and dialog history. Though existing methods try to deal with the cross-modal understanding in visual dialog, they are still not enough in ranking candidate answers based on their understanding of visual and textual contexts. In this paper, we analyze the cross-modal understanding in visual dialog based on the vision-language pre-training model VD-BERT and propose a novel approach to improve the cross-modal understanding for visual dialog, named ICMU. ICMU enhances cross-modal understanding by distinguishing different pulled inputs (i.e. pulled images, questions or answers) based on four-way contrastive learning. In addition, ICMU exploits the single-turn visual question answering to enhance the visual dialog model's cross-modal understanding to handle a multi-turn visually-grounded conversation. Experiments show that the proposed approach improves the visual dialog model's cross-modal understanding and brings satisfactory gain to the VisDial dataset.

Index Terms— Visual Dialog, Cross-modal Understanding, Contrastive Learning

1. INTRODUCTION

Recently, with the rise of pre-trained models [2], researchers have begun to explore vision-and-language task [3, 4, 5] with pre-trained models [1]. Specifically, visual dialog [6, 7, 8, 9], which aims to hold a meaningful conversation with a human about a given image, is a challenging task that requires models have sufficient cross-modal understanding based on both visual and textual context to answer the current question.

One way to gain sufficient cross-modal understanding is through utilizing kinds of attention mechanism [10, 11, 12]. ReDAN [13] and DMAM [14] use multi-step reasoning based on dual attention to learn cross-modal understanding. DAN [15], MCAN [7] and LTMI [16] utilize multi-head attention mechanisms to manage multi-modal intersection. Moreover, there are some approaches [17, 18, 19, 20, 21] using graph-based structures to learn cross-modal understanding.

However, the approaches mentioned above do not utilize pre-trained models, which have a strong power to deal with vision-and-language tasks. Visdial-BERT [22] and VD-BERT [1] take advantage of the pre-trained model to greatly improve the performance of the visual dialog task. As shown in Figure 1, the SOTA model VD-BERT often makes mistakes and usually ranks the wrong answers first. VD-BERT does not have enough cross-modal understanding capabilities, so that it often scores unrelated wrong answers very high, such as the top 1 candidate answer "no" to the question Q4 "is the food in his mouth ?" shown in Figure 1.

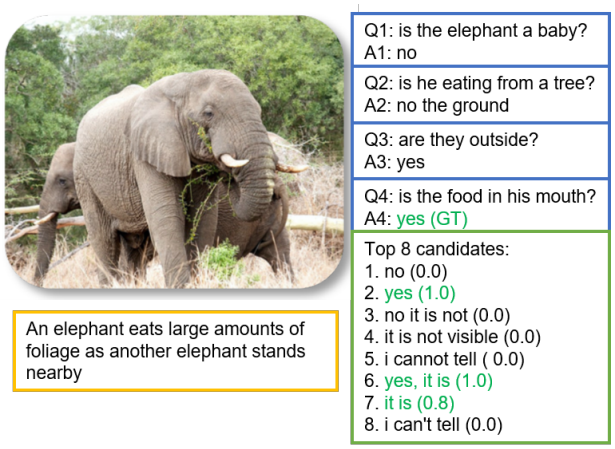

Fig. 1. A motivating example of cross-modal understanding of VD-BERT [1]. We show the candidates ranking results of VD-VBERT based on its cross-modal understanding. It can be seen that in the first 8 candidates, wrong answers account for most of them, and the ranking results of correct answers are not so good.

In this paper, we propose a novel approach to improve the cross-modal understanding for visual dialog, named ICMU. ICMU enhances cross-modal understanding by distinguishing different pulled inputs (i.e. pulled images, questions or answers) based on four-way contrastive learning. What's more, ICMU exploits the single-turn visual question answering to enhance the visual dialog model's cross-modal understanding to handle a multi-turn visually-grounded conversation. Experiments show that the proposed approach improves the visual dialog model's cross-modal understanding and brings satisfactory gain on the VisDial dataset [5]. The contributions of this work are summarized as follows:

- We propose a novel approach ICMU, including 4-way contrastive learning and enhancing by utilizing VQA, to improve the cross-modal understanding based on vision-and-language pre-trained models for visual dialog.
- We conduct extensive experiments and ablation studies on the large-scale datasets VisDial v1.0. Experimental results show that our approach improves the visual dialog model's cross-modal understanding and brings satisfactory gain.

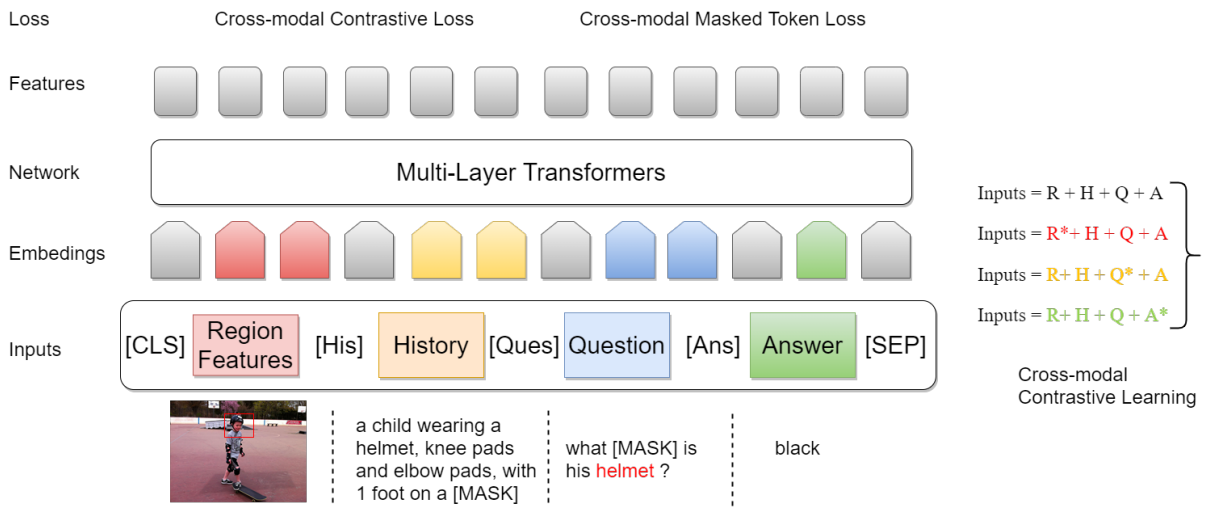

Fig. 2. The Framework of our ICMU. * indicates the pulled inputs.

2. METHODOLOGY

In this section, we first formally describe the visual dialog task. Given a current question Q_t with an image I at t-th turn, as well as its dialog history $H_t = \{C, (Q_1, A_1), ..., (Q_{t-1}, A_{t-1})\}$ (where C denotes the image caption), the dialog model is required to predict its answer A_t by ranking a list of 100 answer candidates $\{\hat{A}_t^1, \hat{A}_t^2, ..., \hat{A}_t^{100}\}$.

Figure 2 shows the overview of our approach. First, we employ a unified vision-dialog Transformer to encode both the image and dialog history, where we append an answer candidate $\hat{A}_t$ in the input to model their interactions in an early fusion manner. Next, we adopt cross-modal masked token loss and cross-modal contrastive loss to train the model for effective cross-modal understanding in visual dialog. In addition, we exploit the single-turn visual question answering to enhance the visual dialog model's cross-modal understanding to handle a multi-turn visually-grounded conversation.

2.1. Vision-Dialog Transformer

2.1.1. Visual Features.

Given an image I, we employ Faster R-CNN [23] pre-trained on Visual Genome [24] to extract the object-level vision features $R_I = \{o_1, ..., o_k\}$, where each object feature o_i is a 2048-d Region-of-Interest (RoI) feature. k is fixed to 36 in our setting. In addition, we adopt normalized bounding box coordinates as the spatial location due to disorder of visual objects. Specifically, we define the location information by constructing a 5-d vector: $p_i = (\frac{x_1}{W}, \frac{y_1}{H}, \frac{x_2}{W}, \frac{y_2}{H}, \frac{(x_2-x_1)(y_2-y_1)}{WH})$, where (x_1, y_1) and (x_2, y_2) are the coordinates of the bottom-left and top-right corner of the i-th object, W and H respectively denote the width and height of the input image, and the last element is the relative area of the object. We also extend p_i with its class id and confidence score for a richer representation to 7-d vector.

2.1.2. Textual Features.

For the textual features, we pack all the textual elements (the history, question and answer candidate) into a long sequence and employ WordPiece tokenizer [25] to split it into a word sequence $\mathbf{w}$, where each word is embedded with an absolute positional code following [26].

2.1.3. Cross-Modality Encoding.

Like a most vision-and-language transformers, we integrate the image objects with language elements into a whole input sequence. As shown in Figure 2, we use some special tokens to segment different elements in the input sequence. We use `[CLS]` to denote the beginning of the sequence, and `[SEP]` to separate the two modalities. Moreover, we utilize a special token `[HIS]` to denote *end of turn* [27], which informs the model when the dialog turn ends. And we use `[Ques]` and `[Ans]` to segment the current question and the answer candidate. As such, we prepare the input sequence into the format as $\mathbf{x} = (\texttt{[CLS]}, o_1, ..., o_k, \texttt{[SEP]}, C, \texttt{[His]}, Q_1 A_1, \texttt{[His]}, ..., \texttt{[Ques]}, Q_t, \texttt{[Ans]}, \hat{A}_t, \texttt{[SEP]})$. Finally, We combine each input token embedding with its position embedding and segment embedding (0 or 1, indicating whether it is image or text) and then perform layer normalization [28].

2.1.4. Transformer Backbone.

We utilize transformer encoder as the Transformer backbone to handle cross-modal understanding. Formally, we denote the embedded vision-language inputs as $\mathbf{H}^0 = [\mathbf{e}_1, ..., \mathbf{e}_{|\mathbf{x}|}]$ and then encode them into multiple levels of cross-modal representations $\mathbf{H}^l = [\mathbf{h}_1^l, ..., \mathbf{h}_{|\mathbf{x}|}^l]$ using L-stacked Transformer blocks, where the l-th Transformer block is denoted as $\mathbf{H}^l = \text{Transformer}(\mathbf{H}^{l-1}), l \in [1, L]$. Specifically, the cross-modal representations $\mathbf{H}^l$ is calculated by using the multi-head self-attention [29] as follows:

$$\mathbf{Q} = \mathbf{H}^{l-1}\mathbf{W}_l^Q, \mathbf{K} = \mathbf{H}^{l-1}\mathbf{W}_l^K, \mathbf{V} = \mathbf{H}^{l-1}\mathbf{W}_l^V, \quad (1)$$

$$\mathbf{M}_{ij} = \begin{cases} 0, & \text{allow to attend,} \\ -\infty, & \text{prevent from attending,} \end{cases} \quad (2)$$

$$\mathbf{A}_l = \text{softmax}(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}} + \mathbf{M})\mathbf{V}, \quad (3)$$

where $\mathbf{W}_l^Q, \mathbf{W}_l^K, \mathbf{W}_l^V \in \mathbb{R}^{d_h \times d_k}$ are learnable weights for computing the queries, keys, and values respectively, and $\mathbf{M} \in \mathbb{R}^{|\mathbf{x}| \times |\mathbf{x}|}$ is the self-attention mask that determines whether tokens from two

Model	NDCG	MRR	R@1	R@5	R@10	Mean
ReDAN	57.63	64.75	51.10	81.73	90.90	3.89
GNN-EM	52.82	61.37	47.33	77.98	87.83	4.57
DualVD	56.32	63.23	49.25	80.23	89.70	4.11
FGA	56.90	66.20	52.75	82.92	91.07	3.80
CAG	56.64	63.49	49.85	80.63	90.15	4.11
KBGN	57.60	64.13	50.47	80.70	90.16	4.08
LG	58.55	64.00	50.63	80.58	90.20	4.12
GoG	60.38	63.13	49.88	79.65	89.05	4.39
VD-BERT	59.96	65.44	51.63	82.23	90.68	3.90
ICMU (Ours)	**61.30**	**66.82**	**53.50**	**83.05**	**92.05**	**3.59**

Table 1. Main comparisons on VisDial v1.0 test datasets (online). Our approach improves the strong baseline significantly. (t-test, p-value<0.01)

sources can attend each other. Then $\mathbf{A}_l$ is passed into a feedforward layer to compute $\mathbf{H}^l$ for the next layer:

$$\mathbf{H}^l = \mathrm{FFN}(\mathbf{A}_l) \tag{4}$$

2.2. Cross-Modal Training Objectives

To make the model learn cross-modal understanding, we use two *cross-modal* training losses—cross-modal masked token loss and cross-modal contrastive loss:

$$\mathcal{L} = \mathcal{L}_{CMTL} + \mathcal{L}_{CCL4}, \tag{5}$$

where $\mathcal{L}_{CMTL}$ is the cross-modal masked token loss and $\mathcal{L}_{CCL4}$ is a novel 4-way contrastive loss.

2.2.1. Cross-modal Masked Token Loss

At each iteration, we randomly mask each input token with probability 15% and replace the masked one with a special token `[MASK]`. The model is then required to recover them based not only on the surrounding tokens $\mathbf{w}_{\setminus m}$ but also on the image I by minimizing the negative log-likelihood:

$$\mathcal{L}_{CMTL} = -E_{(I,\mathbf{w})\sim D} \log P(w_m|\mathbf{w}_{\setminus m}, I), \tag{6}$$

where w_m refers to the masked token and D denotes the training set.

2.2.2. Cross-modal Contrastive Loss

As shown in Figure 2, to compute contrastive losses, for each input quartette $X = (I, H, Q, A)$, we construct three types of negative (unmatched) quartettes, where I denotes the image, H denotes the history, Q denotes the question, A denotes the answer. The first one is the polluted image (I^*, H, Q, A), the second is the polluted question (I, H, Q^*, A) and the final one is the polluted answer (I, H, Q, A^*), where $*$ denotes the polluted input. Since the encoding of `[CLS]` can be viewed as a representation of the quartette $X = (I, H, Q, A)$, we apply a fully-connected (FC) layer on top of it as a 4-way classifier $f(\cdot)$ to predict whether the quartette is matched (c = 0), contains a polluted $I^*(c = 1)$, or contains a polluted $Q^*(c = 2)$ or contains a polluted $A^*(c = 3)$. The 4-way contrastive loss is defined as

$$\mathcal{L}_{CCL4} = -E_{(I,H,Q,A;c)\sim D} \log P(c|f(I, H, Q, A)), \tag{7}$$

where the datasets $I, H, Q, A \in D$ contains 50% matched quartettes, and the three negatives evenly divide the remaining 50% in the training set.

Model	NDCG	MRR	R@1	R@5	R@10	Mean
MN	-	60.29	46.14	77.68	87.57	4.84
HCIAE	-	61.96	48.25	78.97	88.43	4.56
CoAtt	-	62.77	49.38	78.99	88.49	4.56
ReDAN	-	64.29	50.65	81.29	90.17	4.10
KBGN	59.08	64.86	51.37	81.71	90.54	4.00
LG	59.67	65.03	51.69	81.49	90.32	4.02
GoG	63.15	62.68	49.46	78.77	87.87	4.81
VisDial-BERT	62.64	67.86	54.54	84.34	92.36	3.44
VD-BERT	63.22	67.44	54.02	83.96	92.33	3.53
ICMU (Ours)	**64.30**	**69.14**	**56.80**	**85.09**	**93.42**	**3.37**

Table 2. Main comparisons on VisDial v1.0 val datasets. Our approach improves the strong baseline significantly. (t-test, p-value<0.01)

Model	NDCG	MRR	R@1	R@5	R@10	Mean
ICMU	**64.30**	**69.14**	**56.80**	**85.09**	**93.42**	**3.37**
- VQA	63.32	67.62	54.50	84.10	92.90	3.44
- CL	63.34	67.90	54.82	84.35	92.43	3.52

Table 3. Ablation study on VisDial v1.0 val datasets. "VQA" denotes enhancing by utilizing VQA. "CL" denotes the 4-way contrastive learning.

2.3. Using VQA to Enhance Visual Dialog

Although VQA is single-turn, VQA models and visual dialog models require similar cross-modal understanding capabilities. We use VQA to enhance visual dialogue. We exploit the training and val split of VQA v2.0 dataset, which contains the same images as VisDial v1.0 train split. As there is no caption for the image in VQA v2.0, we use VisDial v1.0 to construct a caption for each image in the VQA v2.0. Thus each input from VQA v2.0 can be defined as (I, C, Q, A), where I denotes the image, C denotes the constructed caption, Q denotes the question, A denotes the answer. We let the history H be null.

3. EXPERIMENTS

3.1. Experiment Setup

3.1.1. Datasets and Implementation Details.

We evaluate our model on the VisDial v1.0 datasets [30]. Specifically, v1.0 contains a training set of 123287 images, a validation set of 2048 images and a testing set (hosted blindly in the task organizers' server) of 8,000 images. Each image is associated with one caption and 10 question-answer pairs. For each question, it is paired with a list of 100 answer candidates, one of which is regarded as the correct answer. VQA v2.0 contains the same 123287 images as VisDial v.10 but different question-answer pairs.

We use $\text{BERT}_{\text{BASE}}$ as the backbone, which consists of 12 Transformer blocks, each with 12 attention heads and a hidden state dimensions of 768. We use Adam [31] with an initial learning rate of $3e{-}5$ and a batch size of 80 to train our model. A linear learning rate decay schedule with a warmup of 0.1 is employed. We first train our model for 20 epochs on a cluster of 4 A100 GPUs with 40G memory using CMTL and CCL4 losses (with equal coefficients). Here we only utilize one previous dialog turn for training efficiency. After that, we train for another 15 epochs only using CCL4 losses. Dur-

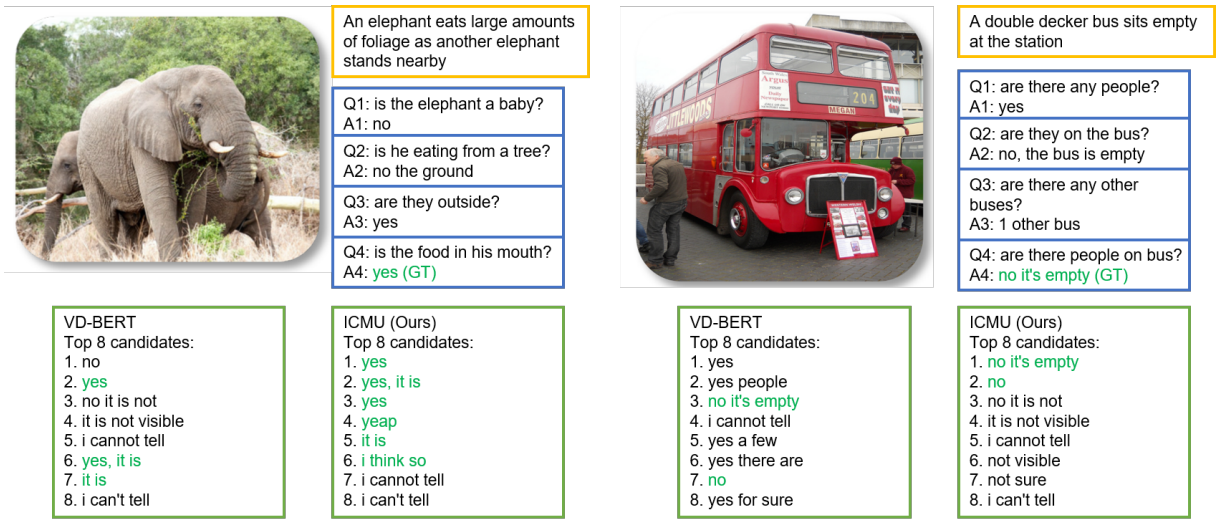

Fig. 3. Case study.

ing inference, we rank the answer candidates according to the class score $c = 0$ of the CCL4 loss.

3.1.2. Automatic Evaluation

We use a retrieval setting to evaluate individual responses at each round of a dialog, following [5]. 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) Mean Rank of human response (Mean ↓), (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.

3.2. Main Results

3.2.1. Baseline Methods

We compare our method with the following baseline methods: (1) Attention-based models: HCIAE [10], CoAtt [11], ReDAN [13], LG [32]. (2) The pretraining model: VD-BERT [1] and VisDial-BERT [22]. (4) Graph-based models: GNN-EM [17], DualVD [19], FGA [18], GoG [6], KBGN [21].

3.2.2. Results

Performance on the benchmarks VisDial is shown in Table 1 and Table 2. From the results on VisDial v1.0 test shown in Table 1, we can observe that: (1) ICMU outperforms previous works on all metrics and obtains R@1 at 53.50%, beating the previous method VD-BERT by 1.47%, which shows that ICMU can select the standard ground-truth more accurate. (2) Comparing the performance of ICMU and model VD-BERT on NDCG, ICMU beats the pre-trained model VD-BERT by 1.34%. This shows the superiority of our proposed method to understand cross-modal information at a fine-grained level. Note that NGCG is invariant to the order of options with identical relevance and to the order of options outside of the top K, where K is the number of answers marked as correct by at least one annotator. (3) Our approach is not only more accurate (R@1, Mean), but also better than previous models on multi-modal semantic understanding (NDCG).

From the results on VisDial v1.0 val shown in Table 2, we can get the same observations. From the ablation study on VisDial v1.0 val shown in Table 3, we can observe that: (1) Both cross-modal contrastive learning and enhancement by VQA bring satisfactory improvements. (2) cross-modal contrastive learning and enhancement by VQA can get along with each other and further improve the performance of the model.

3.2.3. Case Study

As shown in Figure 3, we provide two samples to analyze the cross-modal understanding of VD-BERT and ICMU. As shown in the left half of Figure 3, for Q4 "`Does he have food in his mouth?`", there are many reasonable answers to this question. VD-BERT ranks the opposite answer ``no'' first, and many reasonable answers "`yes, it is, it is`" are ranked lower. As shown in the right half of Figure 3, for Q4 "`are there people on bus?`", ICMU outperforms the VD-BERT. This shows that ICMU learns better cross-modal understanding than VD-BERT due to CCL4 and enhancing by VQA.

4. CONCLUSION

In this paper, we propose a novel approach to improve the cross-modal understanding for visual dialog, named ICMU. ICMU enhances the cross-modal understanding in visual dialog by distinguishing different pulled inputs based on 4-way contrastive learning. In addition, ICMU exploits the single-turn visual question answering to enhance the visual dialog model's cross-modal understanding. Experiments show that the proposed approach improves the visual dialog model's cross-modal understanding and brings satisfactory gain to the VisDial dataset.

5. ACKNOWLEDGEMENT

This work was supported by the National Key R&D Program of China under Grant No.2018YFB1005104 and the Key Research Program of the Chinese Academy of Sciences under Grant No.ZDBS-SSW-JSC006 and Strategic Priority Research Program of the Chinese Academy of Sciences under Grant No.XDA27030300.

6. REFERENCES

[1] Yue Wang, Shafiq Joty, et al., "VD-BERT: A unified vision and dialog transformer with bert," *arXiv preprint arXiv:2004.13278*, 2020.

[2] Xiaoqi Jiao, Yichun Yin, et al., "Tinybert: Distilling bert for natural language understanding," *arXiv preprint arXiv:1909.10351*, 2019.

[3] Mengye Ren, Ryan Kiros, and Richard Zemel, "Exploring models and data for image question answering," in *Advances in Neural Information Processing Systems*, 2015, pp. 2953–2961.

[4] Kelvin Xu, Jimmy Ba, et al., "Show, attend and tell: Neural image caption generation with visual attention," in *ICML*, 2015, pp. 2048–2057.

[5] Abhishek Das, Satwik Kottur, et al., "Visual dialog," in *CVPR*, 2017, pp. 326–335.

[6] Feilong Chen, Xiuyi Chen, et al., "Gog: Relation-aware graph-over-graph network for visual dialog," in *Findings of ACL*, 2021.

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

[8] Feilong Chen, Fandong Meng, Xiuyi Chen, Peng Li, and Jie Zhou, "Multimodal incremental transformer with visual grounding for visual dialogue generation," in *Findings of ACL*, 2021.

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

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

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

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

[13] Zhe Gan, Yu Cheng, Ahmed EI Kholy, Linjie Li, Jingjing Liu, and Jianfeng Gao, "Multi-step reasoning via recurrent dual attention for visual dialog," in *ACL*, 2019, pp. 6463–6474.

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

[15] Dan Guo, Hui Wang, and Meng Wang, "Dual visual attention network for visual dialog," pp. 4989–4995, 2019.

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

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

[18] Idan Schwartz, Seunghak Yu, Tamir Hazan, and Alexander G Schwing, "Factor graph attention," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2039–2048.

[19] Xiaoze Jiang, Jing Yu, Zengchang Qin, Yingying Zhuang, Xingxing Zhang, Yue Hu, and Qi Wu, "DualVD: An adaptive dual encoding model for deep visual understanding in visual dialogue.," in *AAAI*, 2020, vol. 1, p. 5.

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

[21] Xiaoze Jiang, Siyi Du, Zengchang Qin, Yajing Sun, and Jing Yu, "KBGN: Knowledge-bridge graph network for adaptive vision-text reasoning in visual dialogue," *Proceedings of the 28th ACM International Conference on Multimedia*, 2020.

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

[23] Shaoqing Ren, Kaiming He, et al., "Faster R-CNN: towards real-time object detection with region proposal networks," in *NeurIPS*, 2015, pp. 91–99.

[24] Ranjay Krishna, Yuke Zhu, et al., "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *IJCV*, vol. 123, no. 1, pp. 32–73, 2017.

[25] Yonghui Wu, Mike Schuster, et al., "Google's neural machine translation system: Bridging the gap between human and machine translation," *CoRR*, vol. abs/1609.08144, 2016.

[26] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT*, 2019, pp. 4171–4186.

[27] Taesun Whang, Dongyub Lee, Chanhee Lee, Kisu Yang, Dongsuk Oh, and Heuiseok Lim, "Domain adaptive training BERT for response selection," *CoRR*, vol. abs/1908.04812, 2019.

[28] Lei Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton, "Layer normalization," *CoRR*, vol. abs/1607.06450, 2016.

[29] Ashish Vaswani, Noam Shazeer, et al., "Attention is all you need," in *NeurIPS*, 2017, pp. 5998–6008.

[30] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra, "Visual dialog," in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 2017, pp. 1080–1089.

[31] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[32] Feilong Chen, Xiuyi Chen, Can Xu, and Daxin Jiang, "Learning to ground visual objects for visual dialog," *arXiv preprint arXiv:2109.06013*, 2021.