VU-BERT: A UNIFIED FRAMEWORK FOR VISUAL DIALOG

Tong Ye[1,2], Shijing Si[1], Jianzong Wang[1,], Rui Wang[3], Ning Cheng[1], Jing Xiao[1]*

[1]Ping An Technology (Shenzhen) Co., Ltd.
[2]University of Science and Technology of China
[3]Duke University

ABSTRACT

The visual dialog task attempts to train an agent to answer multi-turn questions given an image, which requires the deep understanding of interactions between the image and dialog history. Existing researches tend to employ the modality-specific modules to model the interactions, which might be troublesome to use. To fill in this gap, we propose a unified framework for image-text joint embedding, named VU-BERT, and apply patch projection to obtain vision embedding firstly in visual dialog tasks to simplify the model. The model is trained over two tasks: masked language modeling and next utterance retrieval. These tasks help in learning visual concepts, utterances dependence, and the relationships between these two modalities. Finally, our VU-BERT achieves competitive performance (0.7287 NDCG scores) on VisDial v1.0 Datasets.

Index Terms— Multi-Modal, Visual Dialog, Patch Embedding, Transformer

1. INTRODUCTION

Interactions between humans and the world generally involve joint perception of information in miscellaneous modalities, *e.g.*, image, text and audio. On the contrary, most of the well-known machine learning models are modality-specific like BERT [1, 2] or ResNet [3], agnostic to the information from interweaving of different modalities. So motivated, multimodal tasks have recently gained increasing popularity, especially in the fields of vision and language. At present, popular visual and language tasks include Visual Caption (VC) [4, 5], Visual Grounding [6, 7], Visual Question Answering (VQA) [4, 7, 8] and Visual Dialog (VD) [9, 10, 11]. VQA attempts to predict a correct answer to questions given some background texts and images. The task of VD is firstly proposed by [12], where a dialog agent needs to answer a series of questions grounded by an image. Compared to VQA, VD is more challenging as the model needs not only to present high-level semantic understanding according to image and dialog history, but also to process and reason through the longer contexts than VQA.

To address VD tasks, existing works generally encode the input contexts and images separately with modality-specific modules, then combine the information from the two sources with a variety of cross-modal attention mechanisms, as in DVAN [13], DualVD [14], VisDial-BERT [15], *etc*. Different from them, our VU-BERT leverages a single-stream transformer-based model to learn interactions, which jointly encodes information from the images and text without modality-specific modules. Most closely to our VU-BERT is VD-BERT [11], which also models interactions by a unified transformer. However, VD-BERT must concatenate every answer candidate with the input and go through a forward pass of the entire model, which requires high computing resources and cannot scale to a large number of candidates like ours. Besides, our VU-BERT adapts patch embedding firstly in VD tasks while all of prior models process images complicatedly and inefficiently by ResNet [3] or Faster R-CNN [16].

To explore the interactions between images and dialog history, our framework utilizes BERT [1] architecture, and trains the model over two objectives: masked language modeling and next utterance retrieval. Different from previous work, we adopt a tractable visual embedding introduced in ViT [17], which yields significant reduction on latency time and parameter scale [4]. The key contributions of our study can be summarized as follows:

- We propose a unified framework VU-BERT for effective vision and dialog representations to predict the next utterance by selecting or generating responses. And we achieve competitive performance on visual dialog tasks.

- Different from existing works [9, 10, 11], we attempt to obtain visual embedding by patch projection to the VD tasks, which leads to a simple architecture without any region features or deep convolution operations.

- We conduct extensive experiments to verify the effectiveness of our framework. According to the ablation studies, we found that different dialog turns affect the performance, which indicates that it is not suitable to apply models specific to VQA or VC tasks directly to VD tasks.

---

*Corresponding author: Jianzong Wang, jzwang@188.com.
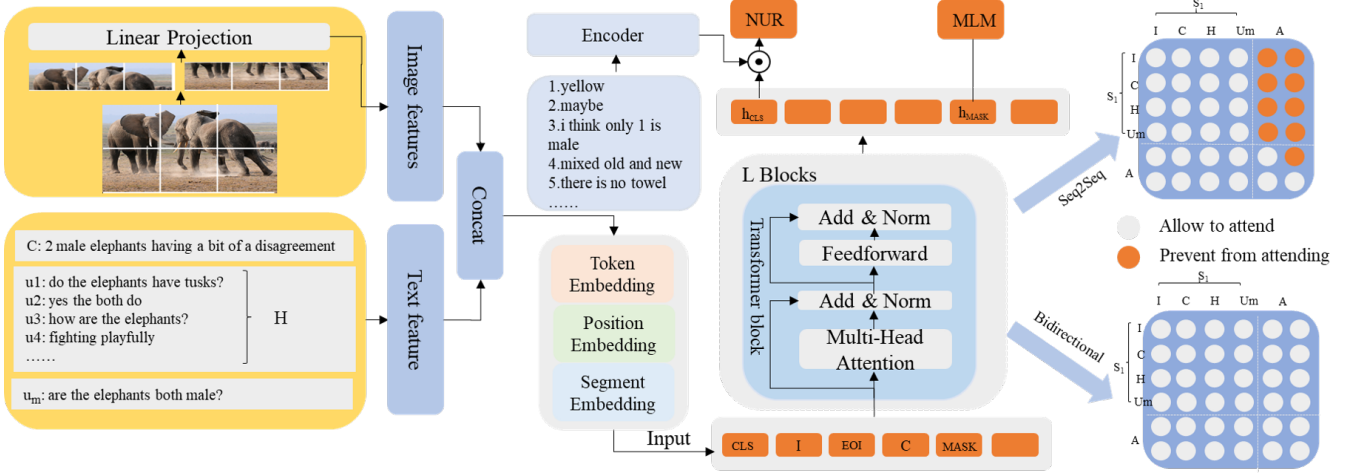
Fig. 1. Overview of our VU-BERT. The image I is split to patches and linearly projected to obtain patch embedding. The input consists of image and text embedding via summing up its position embedding and segment embedding. We keep the input as (I, C, H, U_m) and the response candidates A respectively passing through a transformer encoder.

2. METHODOLOGY

In this section, we discuss our visual dialog model VU-BERT, including two novel approaches aimed at capturing better representations of vision and utterances as shown in Fig 1. We first employ a unified vision-dialog Transformer to encode both the image and dialog history. Then, we adopt visually grounded masked language modeling for effective interactions between visual and textual features and next utterance retrieval for enhancing utterances dependence to train the model. Next, we describe the components of this model in detail.

2.1. Embeddding

Dialog Embedding. Given a m-th turn dialog represented by a list of utterances $U = [u_1, \ldots, u_m]$ grounded in the image I, an utterance is first split into words $u_i = [w_1, \ldots, w_n]$ with length of n by the same WordPiece tokenizer [18]. We follow BERT and each sub-word is obtained via summing up its word embedding, position embedding and segment embedding.

Vision Embedding. Different from extracting object-level features as vision features [9, 10, 11], we use ViT[17] as a backbone network to process images, which is faster than object detectors. According to ViT, Patch Embedding will split the height and width of the input image $I \in R^{H \times W \times C}$ to $N = \frac{HW}{P^2}$ patches according P, and then flatten and reshape the patches to $v \in R^{N \times (P^2 \times C)}$ through linear transformation. Similar to word embedding in BERT, we add segment embedding and position embedding to patch embedding as visual embedding where all the patches share the same segment embedding.

2.2. Network Architecture

VU-BERT, which unifies the multi-model architecture into a single model, is described in Fig.1. We prepare the input sequence as $x = \{[CLS], v_1, \ldots, v_t, [EOI], C, [SEP], u_1, [SEP], \ldots, u_m, [END]\}$. The special token $[CLS]$ and $[END]$ denote the beginning and the end of the sequence respectively and $[SEP]$ separates the different utterances. In addition, we add the token $[EOI]$ to indicate the end of image input. Moreover, we combine the input sequence with its position embedding and segment embedding. Then we encode the sequences into cross-modal representations $H = [h_{cls}, h_1, \ldots, h_{end}]$ using the framework of Transformer. According to the seq2seq self-attention masks [19, 20], we mask out all next-turn utterance tokens by setting $-\infty$ in M to get auto-regressive attentions while keeping tokens in the image I, caption C, dialog history H and current utterance U_m visible, which can recursively generate an answer by predicting the masked tokens. Similarly, we train VU-BERT with the NUR loss using bidirectional self-attention masks to select a correct response from response candidates.

2.3. Training Objectives

To obtain a strong and general representation for the model, we train VU-BERT with the following two losses:

2.3.1. Masked Language Modeling (MLM)

Similar to MLM in BERT, each token in utterance is randomly masked out with 15% probability and replaced with a special token [MASK]. The model is trained to predict the masked words, based on the unmasked words and the visual features. The task drives the network to not only model the

6688

Table 1. Performance of discriminative/generation models on VisDial v0.9 dataset. Higher is better for MRR and recall@k, while lower is better for Mean.

Model	MRR ↑	R@1 ↑	R@5 ↑	R@10 ↑	Mean ↓
LF	0.5807/0.5199	43.82/41.83	74.68/61.78	84.07/67.59	5.78/17.07
HRE	0.5846/0.5237	44.67/42.29	74.50/62.18	84.22/67.92	5.72/17.07
DVAN	0.6675/0.5594	53.62/46.58	82.85/65.50	90.72/71.25	3.93/14.79
FGA	0.6525/0.4138	51.43/27.42	82.08/56.33	89.56/71.32	4.35/**9.10**
VD-BERT	**0.7004/0.5595**	**55.79/46.83**	**85.34/65.43**	**92.68/72.05**	**4.04**/13.18
VU-BERT	0.6333/0.5403	48.71/44.50	81.03/62.60	89.10/71.70	4.19/12.49

dependencies in sentence words, but also to align the visual and utterance contents. We denote the visual embedding as $v = v_1, \ldots, v_t$ and the mask indices as k, where $w_{\backslash k}$ represents as $w_{\backslash k} = \{w_1, \ldots, w_{k-1}, [MASK], w_{k+1}, \ldots, w_n\}$. The parameters are learned to minimize the cross-entropy loss computed using the predicted tokens and the original tokens. Then the MLM loss function is shown in Eq (1):

$$L_{mlm} = -E_{(w,v)\sim D} \log P(w_k | w_{\backslash k}, v) \qquad (1)$$

2.3.2. Next Utterance Retrieval (NUR)

Given the utterances $[u_1, \ldots, u_{m-1}]$, the aim of NUR is to select the correct next utterance u_m from a set of N candidate responses. For this task, we need to respectively encode entire dialog context and each candidate response by using BERT, which allows for the precomputation of all possible candidates and can scale to millions of candidates. A list of candidate responses can be created automatically, with the positive sample being the next utterance in the dialog history, and making the assumption that randomly sampled utterances from different dialogues are likely to be negatives. We calculate the probability of u_i being the true next utterance and minimize the cross-entropy loss of the next utterance, which is analogous to language modeling. The NUR loss function is:

$$L_{nur} = -\log p(u_i | h_{cls}, C)$$
$$= -\log \left(\frac{\exp(h_{cls} * c_i)}{\exp(h_{cls} * c_i) + \sum \exp(h_{cls} * c'_i)} \right) \quad (2)$$

Finally, we define the total loss as $L = \alpha L_{mlm} + \beta L_{nur}$. It's worth noting that we use the NUR task for effective vision and dialog fusion rather than NSP or ITM tasks. Because these tasks are not only weaker than seq2seq or bidirectional LM but also computationally expensive [19].

3. EXPERIMENT

Datasets. We train VU-BERT on the large visual dialog datasets VisDial v0.9 and v1.0. Specifically, the v1.0 contains 123,287 images for training, 2,064 images for validation and 8,000 images for test, which is a re-organization of train and validation splits from v0.9 to form the new train. Each image for training and validation 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. In addition, visdial v1.0 provides dense annotations that specify real-valued relevance scores for the 100 answer candidates.

Baseline. We compare VU-BERT against a variety of single and ensemble baselines in VD tasks, including FGA [9], VD-BERT [11], LF [12], HRE [12], DVAN [13], DualVD [14], VisDial-BERT [15] and CAG [21], which yield prominent performance from the leadboard or publication.

Experimental Setup. We use $BERT_{BASE}$ as the backbone, which consists of 12 Transformer blocks, each with 12 attention heads and a hidden state dimension of 768. We set dropout probability to 0.1 and the learning rate with Adam optimizer to 3e-4 and use GELU as activation function. In our experiments, we split each 224×224 image into a 7×7 grid of image patches, where each patch is 32×32. We first train our model for 20 epochs with a batchsize of 32 on a cluster of 8 Tesla V100 with 16G memory using MLM and NUR losses.

We evaluate VU-BERT on VisDial v0.9 and v1.0 by using the ranking metrics including **Recall@K** ($K \in \{1, 5, 10\}$): the percentage of the ground truth answer option in top-k ranked responses, **Mean**: the average rank of the ground truth answer option, and **MRR** (Mean Reciprocal Rank): reciprocal rank of the ground truth answer option. Due to the acquisition of dense annotation, we add the **NDCG** (Normalized Discounted Cumulative Gain), which penalizes the lower rank of answers with high relevance, to evaluate each answer candidate for v1.0.

4. RESULTS AND ANALYSES

4.1. Results

On VisDial v0.9. We show both discriminative and generative results on v0.9 with the state-of-the-art baselines from previous published results and leaderboard in Table 1. VU-BERT yields comparable results with the state of the art in the discriminative and generative settings, for which the MRR is 0.6333 and 0.5403 respectively. This validates the effectiveness of our model using a unified Transformer encoder.

on VisDial v1.0. As shown in Tabel 2, we compare our results on VisDial v1.0 datasets. We can found that VU-BERT

Table 2. Performance of discriminative models on VisDial v1.0 dataset. Higher is better for MRR, recall@k and NDCG, while lower is better for Mean.

Model	MRR↑	R@1↑	R@5↑	R@10↑	Mean↓	NDCG↑
NMN	0.5880	44.15	76.88	86.88	4.81	0.5810
GNN	0.6137	47.33	77.98	87.83	4.57	0.5282
DualVD	0.6323	49.25	80.23	89.70	**4.11**	0.5632
CAG	0.6349	**49.85**	80.63	**90.15**	**4.11**	0.5664
LF	0.5542	40.95	72.45	82.83	5.95	0.4531
FGA	**0.6370**	49.58	**80.97**	88.55	4.51	0.5210
VD-BERT	0.5117	38.90	62.82	77.98	6.69	**0.7454**
VisDial-BERT	0.5074	37.95	64.13	80.00	6.53	0.7447
VU-BERT	0.4909	33.60	67.20	81.60	6.12	0.7287

Table 3. The parameters for image extraction models

Image model	VGG	Fast R-CNN	Patch
Method	LF;HRE; NMN; GNN; DVAN; FGA	DualVD; CAG; VD-BERT; VisDial-BERT	VU-BERT
# Parameters	138.36M	41.76M	**0.59M**

outperforms most of the model across various metrics. Compared to VD-BERT, VU-BERT results in 4% improvement in R@5, 3% improvement in R@10 and 4% reduction in Mean while in 2% reduction in NDCG and in 2% reduction in MRR. This indicates that VU-BERT can achieve comparable performance even without deep convolution.

Obviously, VU-BERT is a simpler unified model without any region features or deep convolution operations. We commission the unified transformer module to process image and text simultaneously in place of an additional image extractor like Fast R-CNN, which leads to significant time and parameter efficiency [4]. We notice that VU-BERT encodes candidate sets independently rather than integrates each candidate response at the input layer. Therefore, VU-BERT loses the early interaction between the input and the candidates, which coincidentally agrees with the results of some metrics lower than VD-BERT. In spite of a slight decline on metrics, VU-BERT can easily extend the candidate sets to a large amount, which facilitates adaptation to other tasks.

Parameters. In order to present the lightweight VU-BERT, we compare the Patch Projection with two image extraction models commonly used in baseline methods: VGG and Fast R-CNN in Table 3. We can observe that the Patch Projection is much lighter when capturing the image feature. Then we selected several structurally similar models, which deal with text features by transformers, to compare the number of parameters in Table 4. Obviously, VU-BERT is much lighter than existing transformer-based models.

4.2. Ablation Study

In Fig 2, we represent how dialog history affects VD tasks. VU-BERT is trained with different dialog turns with image

Table 4. The parameters of transformer-based architecture

Model	Visdial-Bert	VD-Bert	VU-Bert
# Parameters	315.26M	174.51M	**133.94M**
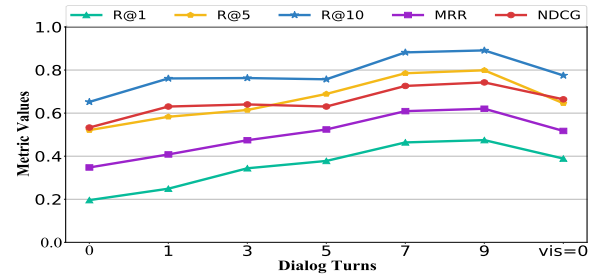
Fig. 2. Ablation study. $vis = 0$ denotes the input includes 9 dialog turns with no image while others denote the different turns with images.

input and test in v1.0 validation set according to the dense annotation. We found that the metrics with only one turn or no dialog history is lower than other multi-turn dialog history. That is to say directly employing the models specific to the VQA and VC for VD tasks will not achieve better results. Also, more does not necessarily mean better for the dialog history. When turns > 7, the metric values of Recall@K, MRR and NDCG gradually become flat, which may be caused by the limitation of input length. Besides, we compare the performance with the image input or not. As shown in Fig 2, the result of $vis = 0$ is obviously lower when the dialog turns are both 9, which proves that VU-BERT takes advantage of multi-modal information to improve performance.

5. CONCLUSION

In this paper, we present a simple visual dialog architecture VU-BERT with the effective fusion of two modalities. VU-BERT is comparable to competitors on VD tasks which are heavily equipped with convolution networks like Faster R-CNN. Besides, we explore the effect of dialog turns and found VQA and VC models are not suitable applied directly in VD tasks. This paper is an initial attempt to explore patch embedding adapted to VD tasks and we ask for future work on VD to focus more on the light model and deep interactions between images and dialog history with no convolution.

6. ACKNOWLEDGMENT

7. REFERENCES

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

[2] Shijing Si, Rui Wang, Jedrek Wosik, Hao Zhang, David Dov, Guoyin Wang, and Lawrence Carin, "Students need more attention: Bert-based attention model for small data with application to automatic patient message triage," in *Machine Learning for Healthcare Conference*. PMLR, 2020, pp. 436–456.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[4] Wonjae Kim, Bokyung Son, and Ildoo Kim, "Vilt: Vision-and-language transformer without convolution or region supervision," in *International Conference on Machine Learning*, 2021, pp. 5583–5594.

[5] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, et al., "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*, 2015, pp. 2048–2057.

[6] Yongfei Liu, Bo Wan, Xiaodan Zhu, and Xuming He, "Learning cross-modal context graph for visual grounding," in *Proceedings of the AAAI*, 2020, vol. 34, pp. 11645–11652.

[7] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," in *Proceedings of EMNLP*, 2016, pp. 457–468.

[8] Zongwen Bai, Ying Li, Marcin Woźniak, Meili Zhou, and Di Li, "Decomvqanet: Decomposing visual question answering deep network via tensor decomposition and regression," *Pattern Recognition*, vol. 110, pp. 107538, 2021.

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

[10] Yashar Deldjoo, Johanne R Trippas, and Hamed Zamani, "Towards multi-modal conversational information seeking," in *Proceedings of SIGIR*, 2021, vol. 21.

[11] Yue Wang, Shafiq R. Joty, Michael R. Lyu, Irwin King, Caiming Xiong, and Steven C. H. Hoi, "VD-BERT: A unified vision and dialog transformer with BERT," in *Proceedings of EMNLP*, 2020, pp. 3325–3338.

[12] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, et al., "Visual dialog," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 326–335.

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

[14] 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 *Proceedings of the AAAI*, 2020, vol. 34, pp. 11125–11132.

[15] Vishvak Murahari, Dhruv Batra, Devi Parikh, and Abhishek Das, "Large-scale pretraining for visual dialog: A simple state-of-the-art baseline," in *European Conference on Computer Vision*. Springer, 2020, pp. 336–352.

[16] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, pp. 91–99, 2015.

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *9th International Conference on Learning Representations*, 2021.

[18] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, et al., "Google's neural machine translation system: Bridging the gap between human and machine translation," *arXiv preprint arXiv:1609.08144*, 2016.

[19] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao, "Unified vision-language pre-training for image captioning and vqa," in *Proceedings of the AAAI*, 2020, vol. 34, pp. 13041–13049.

[20] Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, et al., "Unified language model pre-training for natural language understanding and generation," in *Advances in Neural Information Processing Systems*, 2019, pp. 13042–13054.

[21] 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.