Generative Self-training Improves Pre-training for Visual Dialog

Gi-Cheon Kang [1] Sungdong Kim [2 *] Jin-Hwa Kim [2 *] Donghyun Kwak [3 *] Byoung-Tak Zhang [1]

Abstract

Visual dialog (VisDial) is a task of answering a series of questions grounded in an image, using the dialog history as context. Prior work has trained the dialog models solely on VisDial data via supervised learning or leveraged pre-training on related vision-and-language datasets. This paper presents a semi-supervised learning approach for VisDial, called *Generative Self-Training* (GST), to enhance the pre-training. Specifically, GST generates synthetic dialog data for unlabeled images via multimodal conditional text generation and trains the dialog model on the synthetic and the original VisDial data. Moreover, we also propose perplexity-based data selection and multimodal consistency regularization for robust training of the synthetic data. Evaluation on VisDial v1.0 dataset shows that GST improves the pre-training and achieves new state-of-the-art results.

1. Introduction

Recently, there has been extensive research towards developing visually-grounded dialog systems (Das et al., 2017; De Vries et al., 2017; Kottur et al., 2019; Kim et al., 2019) due to their significance in many real-world applications (*e.g.,* helping visually impaired person). Notably, Visual Dialog (VisDial) (Das et al., 2017) has provided a testbed for studying such systems, where a dialog agent should answer a *sequence* of image-grounded questions. For instance, the agent is expected to answer the open-ended question like *"What color is it?"*. This task requires a holistic understanding of visual information, linguistic semantics in context (*e.g.,* it), and most importantly, the grounding of these two.

Most of the previous approaches in VisDial (Lu et al., 2017; Kottur et al., 2018; Gan et al., 2019; Kang et al., 2019; Nguyen et al., 2020) have trained the dialog agents solely on VisDial data via supervised learning. More recent studies (Murahari et al., 2020; Wang et al., 2020b; Chen et al., 2022) have employed self-supervised pre-trained models such as BERT (Devlin et al., 2019) or ViLBERT (Lu et al., 2019) and finetuned them on VisDial data. This *pretrain-then-transfer* learning strategy has shown impressive results by transferring knowledge successfully from the models pre-trained on large-scale data sources (Sharma et al., 2018; Antol et al., 2015; Zhu et al., 2015) to VisDial.

Our research question is the following: *How can the dialog agent expand its knowledge beyond what it can acquire via self-supervised pre-training on the provided datasets?* Some recent studies have shown that semi-supervised learning and pre-training are complementary in image classification (Zoph et al., 2020) and text classification (Du et al., 2021). Inspired by the studies, we consider semi-supervised learning (SSL) as a way to address the above question. Let us assume that large amounts of unlabeled images are available. SSL for VisDial can be applied to generate synthetic conversations for the unlabeled images and train the agent with the synthetic data. However, there are two critical challenges for this approach. First, the target output for VisDial (*i.e.,* multi-turn visual QA data) is more complex than that of the studies (Zoph et al., 2020; Du et al., 2021). Specifically, they have addressed the classification problems, yielding class probabilities as pseudo labels (Lee et al., 2013). In contrast, SSL for VisDial should generate a sequence of pseudo queries (*i.e.,* visual questions) and pseudo labels (*i.e.,* corresponding answers) in *natural language* to train the answering agent. It further indicates that the target output should be generated while considering the visual inputs and the sequential nature of the conversation. Next, even if SSL yields synthetic dialogs via text generation, there may be noise such as generating incorrect answers to given contexts. A robust training method is required to leverage such noisy synthetic dialog datasets.

In this paper, we study the above challenges in the context of SSL, especially self-training (Zoph et al., 2020; Du et al., 2021; Xie et al., 2020b; He et al., 2020; Karamanolakis et al., 2021), where a teacher model trained on labeled data predicts the pseudo labels for unlabeled data. Then, a student model jointly learns on both the labeled and the pseudo-labeled datasets. Unlike existing studies in self-training that have mainly studied discriminative tasks such as im-

[*]Equal contribution [1]Seoul National University [2]NAVER AI Lab [3]NAVER CLOVA. Correspondence to: Byoung-Tak Zhang <btzhang@snu.ac.kr>.

First Workshop of Pre-training: Perspectives, Pitfalls, and Paths Forward at ICML 2022, Baltimore, Maryland, USA, PMLR 162, 2022.

age classification (Li et al., 2019; Zoph et al., 2020; Sohn et al., 2020) or text classification (Du et al., 2021; Karamanolakis et al., 2021), we extend the idea of self-training to the task of multimodal conditional text generation. To this end, we propose a new learning strategy, called *Generative Self-Training* (GST), that artificially generates multi-turn visual QA data and utilizes the synthetic data for training. GST first trains the teacher model (answerer) and the visual question generation model (questioner) using VisDial data. It then retrieves a set of unlabeled images from a Web image dataset, Conceptual 12M (Changpinyo et al., 2021). Next, the questioner and the teacher alternately generate a series of visual QA pairs for the retrieved images. Finally, the student is trained on the synthetic and the original VisDial data. We also propose perplexity-based data selection (PPL) and multimodal consistency regularization (MCR) to effectively train the student with the noisy dialog data. PPL is to selectively utilize the answers whose perplexity of the teacher is below a threshold. MCR encourages the student to yield consistent predictions even when the perturbed multimodal inputs are given. Our key contributions are three-fold:

- We introduce Generative Self-Training (GST) to study the effect of semi-supervised learning on top of the pre-training method.
- We show the efficacy of the perplexity-based data selection (PPL) and the multimodal consistency regularization (MCR) when training the synthetic data.
- Finally, we validate GST on VisDial v1.0 dataset, and GST achieves new state-of-the-art results on both datasets. We also show that GST and self-supervised pre-training are complementary in VisDial.

2. Approach

2.1. Preliminaries

Self-Training. We have a labeled dataset $L = \{(x_n, y_n)\}_{n=1}^{N}$ and an unlabeled dataset $U = \{\tilde{x}_m\}_{m=1}^{M}$. Typically, self-training first trains a teacher model $P_{\mathcal{T}}$ on the labeled dataset L. The teacher then predicts the pseudo label $\tilde{y}$ for the unlabeled data $\tilde{x} \sim U$, constructing the pseudo-labeled dataset $\tilde{L} = \{(\tilde{x}_m, \tilde{y}_m)\}_{m=1}^{M}$. Finally, a student model $P_{\mathcal{S}}$ is trained on $L \cup \tilde{L}$. Many variants have been studied on this setup: (1) selecting the subset of the pseudo-labeled dataset (He et al., 2020; Xie et al., 2020b; Sohn et al., 2020), (2) adding noise to inputs (Zoph et al., 2020; He et al., 2020), and (3) iterating the above setup multiple times (He et al., 2020; Xie et al., 2020b).

Visual Dialog. The visual dialog (VisDial) dataset (Das et al., 2017) contains an image v and a visually-grounded dialog $d = \{\underbrace{c}_{d_0}, \underbrace{(q_1, a_1^{gt})}_{d_1}, \cdots, \underbrace{(q_T, a_T^{gt})}_{d_T}\}$ where c denotes an image caption. T is the number of rounds for each dialog. At round t, a dialog agent is given a triplet $(v, d_{<t}, q_t)$ as context, consisting of the image, the dialog history, and a visual question. $d_{<t}$ denotes all dialog rounds before t-th round. The agent should predict a ground-truth answer a_t^{gt}.

2.2. Generative Self-Training (GST)

Overview. An overview of GST is shown in Figure 1. We have a human-labeled VisDial dataset $L = \{(v_n, d_n)\}_{n=1}^{N}$ where v_n is a given image and each dialog $d_n = \{\underbrace{c_n}_{d_{n,0}}, \underbrace{(q_{n,1}, a_{n,1}^{gt})}_{d_{n,1}}, \cdots, \underbrace{(q_{n,T}, a_{n,T}^{gt})}_{d_{n,T}}\}$ consists of an image caption c and T rounds of QA pairs. In the following, we omit the superscript gt in the ground-truth answer for brevity. GST first trains a teacher $P_{\mathcal{T}}$ and a questioner $P_{\mathcal{Q}}$ with the labeled dataset L via supervised learning. It then retrieves unlabeled images $U = \{\tilde{v}_m\}_{m=1}^{M}$ from the image dataset using a simple outlier detection model. Next, the questioner and the teacher generate the visually-grounded dialog $\tilde{d}$ for the unlabeled image $\tilde{v}$ via multimodal conditional text generation, finally yielding a synthetic dialog dataset $\tilde{L} = \{(\tilde{v}_m, \tilde{d}_m)\}_{m=1}^{M}$. We call this dataset the *machine VisDial* data to distinguish it from the human-labeled VisDial dataset (Das et al., 2017) (short for the *human VisDial* data). Finally, a student $P_{\mathcal{S}}$ is trained on the human and the machine VisDial data by applying perplexity-based data selection (PPL) and multimodal consistency regularization (MCR) to the machine VisDial data.

Teacher & questioner training. GST first trains the answer generator, the teacher model $P_{\mathcal{T}}$, on the human VisDial dataset. Specifically, the teacher is optimized by minimizing the negative log-likelihood of the ground-truth answer, given the context $c_{n,t} \triangleq (v_n, d_{n,<t}, q_{n,t})$, consisting of the image, the dialog history, and the question. Formally,

$$\mathcal{L}_{Teacher} = -\frac{1}{NT} \sum_{n=1}^{N} \sum_{t=1}^{T} \log P_{\mathcal{T}}(a_{n,t} | c_{n,t}) \qquad (1)$$

where N and T denote the number of data tuples in human VisDial data and dialog rounds, respectively. Similarly to the teacher, the questioner is trained to generate the question at round t for n-th dialog data, given the image and the dialog history until round $t-1$ (*i.e.,* $P_{\mathcal{Q}}(q_{n,t} | v_n, d_{n,<t})$). Both the teacher and the questioner are based on encoder-decoder architecture, where an encoder aggregates the context, and a decoder generates the target sentence. We implement the models by integrating a pretrained vision-and-language encoder, VisDial-BERT (Murahari et al., 2020), with the transformer decoder (Rothe et al., 2020).

Unlabeled image retrieval. GST selects in-domain image data from the Conceptual 12M dataset (Changpinyo et al., 2021) with an out-of-distribution (OOD) detection

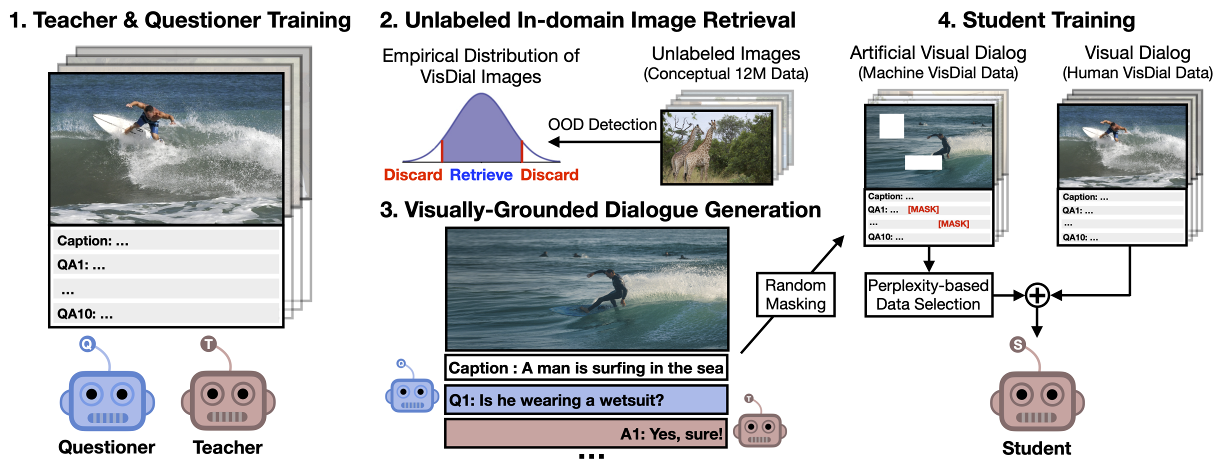


Figure 1. An overview of Generative Self-Training (GST).

model. Specifically, we extract the D-dimensional feature vector for each image in the human VisDial dataset by using the Vision Transformer (ViT) (Dosovitskiy et al., 2021), yielding a feature matrix for the entire images $\mathbf{X} = (X_1, \cdots, X_N)^\top \in \mathbb{R}^{N\times D}$. Based on the feature matrix, we build the multivariate normal distribution whose dimension is D, *i.e.,* $\mathbf{X} \sim \mathcal{N}_D(\mu, \Sigma)$. We regard this normal distribution as the empirical distribution of the human VisDial images and perform the OOD detection by identifying the probability of each feature vector for the unlabeled image. Consequently, the top-M unlabeled images are retrieved out of 12 million Web images.

Visually-grounded dialog generation. Given the retrieved images $U = \{\tilde{v}_m\}_{m=1}^M$, our goal is to generate the visually-grounded dialogs $\{\tilde{d}_m\}_{m=1}^M$ where each dialog $\tilde{d}$ consists of the image caption and T rounds of QA pairs. In an actual implementation, we use the image captions in the Conceptual 12M dataset (Changpinyo et al., 2021) and thus do not generate the captions. The QA pairs are sequentially generated. Specifically, the image $\tilde{v}$, the caption $\tilde{c}$, and the generated QA pairs until round $t-1$ are used as inputs when the questioner generates the question at round t. After then, the teacher produces the answer $\tilde{a}_t$ based on the image $\tilde{v}$, the dialog history $\tilde{d}_{<t}$, and the question $\tilde{q}_t$. Finally, GST produces the machine VisDial dataset $\tilde{L} = \{(\tilde{v}_m, \tilde{d}_m)\}_{m=1}^M$.

Student training with noisy data. As shown in Figure 1, the student $P_\mathcal{S}$ is trained on the combination of the machine and the human VisDial data. According to the studies (Xie et al., 2020b; He et al., 2020) in self-training, selectively utilizing the samples in the pseudo-labeled dataset is a common strategy. To this end, we introduce a simple data selection method called perplexity-based data selection (PPL) to utilize the answers whose perplexity of the teacher is below a certain threshold. Perplexity is defined as the exponentiated average negative log-likelihood of a sequence; the lower, the better. Furthermore, inspired by the consistency regularization (Xie et al., 2020a; Sohn et al., 2020) widely utilized in recent SSL algorithms, we also propose the multimodal consistency regularization (MCR) which encourages the student to yield the predictions similar to the teacher's predictions even when the student is provided with perturbed multimodal inputs. Finally, we design a loss function:

$$\mathcal{L}_{Student} = -\frac{1}{NT}\sum_{n=1}^{N}\sum_{t=1}^{T}\log P_\mathcal{S}(a_{n,t}|c_{n,t})$$
$$-\frac{1}{MT}\sum_{m=1}^{M}\sum_{t=1}^{T}\mathbb{1}(\mathrm{PPL}(\tilde{a}_{m,t})<\tau)\log \underbrace{P_\mathcal{S}(\tilde{a}_{m,t}|\,\mathcal{M}(\tilde{c}_{m,t}))}_{\mathrm{MCR}} \tag{2}$$

where M, $\mathbb{1}$, and τ denote the number of data tuples in machine VisDial data, indicator function, and selection threshold, respectively. $\tilde{c}_{m,t} \triangleq (\tilde{v}_m, \tilde{d}_{m,<t}, \tilde{q}_{m,t})$ denotes the context for the machine VisDial data. The loss function is the sum of the losses for the machine and the human VisDial data. PPL and MCR are applied to computing the loss of the machine VisDial data. PPL is used in the indicator function above, selecting the synthetic answers whose perplexity of the teacher is below the threshold τ. It implies that the unselected answers are ignored during training. Next, $\mathcal{M}$ denotes the stochastic function for MCR that injects perturbations to the input space of the student. Inspired by ViLBERT (Lu et al., 2019), we implement the stochastic function by randomly masking 15% of image regions and word tokens. Specifically, masked image regions have their image features zeroed out, and the masked word tokens are replaced with a special `[MASK]` token. The intuition behind MCR is minimizing the distance between the *perturbed* (*i.e.,* masked) predictions from the student and the *unperturbed* predictions (*i.e.,* $\tilde{a}_{m,t}$) from the teacher. The student can be robust to the input noise with MCR. We believe MCR

Table 1. Ablation study on the VisDial v1.0 validation split. PT, PPL, and MCR denote pretraining, perplexity-based data selection, and multimodal consistency regularization, respectively. 5x denotes the student model used the machine VisDial data, five times larger than human VisDial data, *i.e.,* $|M| = 5 \times |N|$.

Model	NDCG	MRR
Teacher	64.50	52.06
Student (w/o PT, 5x)	63.17	51.76
Student (w/o PPL, 5x)	64.01	52.23
Student (w/o MCR, 5x)	64.21	52.50
Student (w/o MCR and PPL, 5x)	63.81	52.16
Student (5x)	**65.05**	**52.53**

improves the generalization abilities of the student, and PPL encourages the student to maintain a low entropy (*i.e.,* confident) in noisy data training.

3. Experiments

3.1. Evaluation Protocol

We follow the standard evaluation protocol established in the work (Das et al., 2017). Specifically, the visual dialog models have been evaluated by the retrieval-based evaluation metrics: mean reciprocal rank (MRR), recall@k (R@k), mean rank (Mean), and normalized discounted cumulative gain (NDCG). Specifically, all dialogs in VisDial contain a list of 100 answer candidates for each visual question and there is one ground-truth answer in the answer candidates. The model sorts the answer candidates by the log-likelihood scores and then is evaluated by the four different metrics. MRR, R@k, and Mean consider the rank of the single ground-truth answer, while NDCG[1] considers all relevant answers from the 100-answers list by using the densely annotated relevance scores for all answer candidates. The community regards MRR and NDCG as primary metrics.

3.2. Ablation Study

We perform an ablation study to illustrate the effect of each component in GST. The machine VisDial data whose size is five times larger than the human VisDial data is utilized to train the student. We report the performance of ablative models: Student w/o PT, Student w/o PPL, Student w/o MCR, Student w/o MCR and PPL. Student w/o PT is the model that does not leverage the pre-trained weights of the VisDial-BERT model. Student w/o PPL denotes the model that utilizes all generated QA pairs without applying the perplexity-based data selection. Student w/o MCR does not inject noises to the inputs of the student model.

[1]https://visualdialog.org/challenge/2019#evaluation

Table 2. Comparison with the state-of-the-art generative models on both VisDial v1.0 validation dataset. ↑ indicates higher is better. ↓ indicates lower is better. In this experiment, we scale up the size of the machine VisDial data $|M|$ to 30 $\times |N|$.

	VisDial v1.0 (val)					
Model	NDCG↑	MRR↑	R@1↑	R@5↑	R@10↑	Mean↓
MN	51.86	47.99	38.18	57.54	64.32	18.60
HCIAE	59.70	49.07	39.72	58.23	64.73	18.43
CoAtt	59.24	49.64	40.09	59.37	65.92	17.86
Primary	-	49.01	38.54	59.82	66.94	16.60
ReDAN	60.47	50.02	40.27	59.93	66.78	17.40
LTMI	63.58	50.74	40.44	61.61	69.71	14.93
MITVG	61.47	51.14	41.03	61.25	68.49	**14.37**
UTC	63.86	<u>52.22</u>	<u>42.56</u>	62.40	69.51	15.67
Teacher	<u>64.50</u>	52.06	42.04	<u>62.92</u>	<u>71.06</u>	<u>14.54</u>
Student	**65.06**	**52.84**	**42.74**	**63.66**	**71.30**	14.60

In Table 1, we observe all components play a significant role in boosting the performance. Especially, we develop the teacher, the questioner, and the student models on top of VisDial-BERT (Murahari et al., 2020) which leverages vision-and-language pre-training (Lu et al., 2019). Thus, the teacher can be understood as a typical model that follows the pretrain-then-transfer learning strategy mentioned in the introduction, whereas the student leverages both pre-training and GST. By comparing the student, the student without pre-training, and the teacher in Table 1, we identify that self-supervised pre-training and GST are complementary.

3.3. Comparison with State-of-the-Art

We compare GST with the state-of-the-art approaches on the validation set of the VisDial v1.0 dataset, consisting of UTC (Chen et al., 2022), MITVG (Chen et al., 2021), LTMI (Nguyen et al., 2020), ReDAN (Gan et al., 2019), Primary (Guo et al., 2019), CoAtt (Wu et al., 2018), HCIAE (Lu et al., 2017), and MN (Das et al., 2017). In Table 2, GST significantly outperforms the state-of-the-art methods, including UTC (Chen et al., 2022), on all evaluation metrics except for Mean. GST improves NDCG 1.20% (63.86 → 65.06) and MRR 0.62% (52.22 → 52.84).

3.4. Is GST helpful when human-labeled data is scarce?

We investigate this question to identify the effect of GST in the low-data regime. We assume that only a small subset of the human VisDial data (1%, 5%, 10%, 20%, and 30%) is available. Therefore, the size of the human-labeled data is $0.01N$, $0.05N$, $0.1N$, $0.2N$, and $0.3N$, respectively. We first train the teacher and the questioner on such scarce data, and then these two agents generate a new machine VisDial data for unlabeled images in the Conceptual 12M dataset (Changpinyo et al., 2021) with size N. The student is then trained on the newly generated machine VisDial data

Table 3. Results of GST in the low-data regime. We report NDCG and MRR of the teacher and the student models on the VisDial v1.0 val split. We assume that a small subset of the human VisDial data (1%, 5%, 10%, 20%, and 30%) with the size of $0.01N$, $0.05N$, $0.1N$, $0.2N$, and $0.3N$ is available.

	NDCG / MRR				
Model	$0.01N$	$0.05N$	$0.1N$	$0.2N$	$0.3N$
Teacher	27.64 / 40.31	50.04 / 44.37	54.46 / 45.89	57.14 / 48.11	60.67 / 49.87
Student	36.99 / 41.29 (+9.35 / +0.98)	54.20 / 45.83 (+4.16 / +1.46)	57.26 / 47.40 (+2.80 / +1.51)	59.74 / 49.33 (+2.60 / +1.22)	61.59 / 50.60 (+0.92 / +0.73)

Table 4. Results of GST in the high-data regime. We report the NDCG and MRR of the teacher and the student models on VisDial v1.0 val split. Based on the full human VisDial data, the student utilizes a subset of the machine VisDial data (1, 5, 10, 20, and 30 out of 30 data chunks).

	NDCG / MRR				
Model	1x	5x	10x	20x	30x
Teacher	64.50 / 52.06	64.50 / 52.06	64.50 / 52.06	64.50 / 52.06	64.50 / 52.06
Student	64.38 / 52.14 (-0.12 / +0.08)	65.05 / 52.53 (+0.55 / +0.47)	64.90 / 52.65 (+0.40 / +0.59)	64.94 / 52.81 (+0.44 / +0.75)	65.06 / 52.84 (+0.56 / +0.78)

and the small amount of the human VisDial data. Note that PPL and MCR are still applied in this experiment. In Table 3, GST yields huge improvements on both metrics, especially NDCG, boosting up to 9.35 absolute points compared with the teacher. We observe that the smaller the amount of human-labeled data, the larger the performance gap between the teacher and the student on NDCG. It implies that GST is helpful, especially when human-labeled data is scarce. We think the results in the low-data regime are particularly remarkable in other dialog-based tasks (Thomason et al., 2020; Alamri et al., 2019; Rashkin et al., 2019; Li et al., 2017b) since many tasks provide the human dialog data less than $0.3N$ (*i.e.,* 360k QA pairs).

3.5. Scaling up the size of the machine VisDial data.

We also conduct experiments on the high-data regime where the entire human VisDial data and the machine VisDial data with different number data chunks (1, 5, 10, 20, and 30 out of 30 data chunks) are used to train the student. In Table 4, the student shows an increase in performance compared with the teacher except when using one data chunk. Notably, we observe significant gains in the student when increasing from one to five data chunks (*i.e.,* 1x $\rightarrow$ 5x). NDCG seems to be saturated when using more than five data chunks, but MRR increases monotonically. The student leveraging the entire machine VisDial data (x30) shows the best performance on average.

4. Conclusion

We propose a semi-supervised learning approach for VisDial, called GST, that explicitly generates a synthetic visual dialog dataset for unlabeled images via multimodal conditional text generation. We also present the perplexity-based data selection and the multimodal consistency regularization to effectively leverage the generated dialogs. The experiments quantitatively support the effectiveness of our proposed method. Above all, we validate that GST can further improve self-supervised pre-training approach for VisDial.

Acknowledgements. The authors thank Minjoon Jung, Yu-Jung Heo, Woosuk Choi, Seongho Choi, and Seongjun Jung for helpful comments. This work was partly supported by the Institute of Information & Communications Technology Planning & Evaluation (2015-0-00310-SW.StarLab/20%, 2019-0-01371-BabyMind/20%, 2021-0-02068-AIHub/10%, 2021-0-01343-GSAI/10%, 2022-0-00951-LBA/20%, 2022-0-00166-PICA/20%) grant funded by the Korean government.

References

Alamri, H., Cartillier, V., Das, A., Wang, J., Cherian, A., Essa, I., Batra, D., Marks, T. K., Hori, C., Anderson, P., et al. Audio visual scene-aware dialog. In *CVPR*, 2019.

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., and Zhang, L. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018.

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., and Parikh, D. Vqa: Visual question answering. In *ICCV*, 2015.

Bahdanau, D., Cho, K., and Bengio, Y. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2014.

Changpinyo, S., Sharma, P., Ding, N., and Soricut, R. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, 2021.

Chen, C., Zhu, Y., Tan, Z., Cheng, Q., Jiang, X., Liu, Q., and Gu, X. Utc: A unified transformer with inter-task contrastive learning for visual dialog. In *CVPR*, 2022.

Chen, F., Meng, F., Chen, X., Li, P., and Zhou, J. Multimodal incremental transformer with visual grounding for visual dialogue generation. In *ACL*, 2021.

Das, A., Kottur, S., Gupta, K., Singh, A., Yadav, D., Moura, J. M., Parikh, D., and Batra, D. Visual dialog. In *CVPR*, 2017.

De Vries, H., Strub, F., Chandar, S., Pietquin, O., Larochelle, H., and Courville, A. Guesswhat?! visual object discovery through multi-modal dialogue. In *CVPR*, 2017.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

Du, J., Grave, E., Gunel, B., Chaudhary, V., Celebi, O., Auli, M., Stoyanov, V., and Conneau, A. Self-training improves pre-training for natural language understanding. In *NAACL*, 2021.

Fan, A., Lewis, M., and Dauphin, Y. Hierarchical neural story generation. In *ACL*, 2018a.

Fan, Z., Wei, Z., Li, P., Lan, Y., and Huang, X. A question type driven framework to diversify visual question generation. In *IJCAI*, 2018b.

Gan, Z., Cheng, Y., Kholy, A. E., Li, L., Liu, J., and Gao, J. Multi-step reasoning via recurrent dual attention for visual dialog. In *ACL*, 2019.

Guo, D., Xu, C., and Tao, D. Image-question-answer synergistic network for visual dialog. In *CVPR*, 2019.

He, J., Gu, J., Shen, J., and Ranzato, M. Revisiting self-training for neural sequence generation. In *ICLR*, 2020.

Holtzman, A., Buys, J., Forbes, M., Bosselut, A., Golub, D., and Choi, Y. Learning to write with cooperative discriminators. In *ACL*, 2018.

Huang, X., Qi, J., Sun, Y., and Zhang, R. Mala: Cross-domain dialogue generation with action learning. In *AAAI*, 2020.

Iashin, V. and Rahtu, E. Multi-modal dense video captioning. In *CVPR Workshops*, 2020.

Jain, U., Zhang, Z., and Schwing, A. G. Creativity: Generating diverse questions using variational autoencoders. In *CVPR*, 2017.

Kai, S., Wu, L., Tang, S., Zhuang, Y., Ding, Z., Xiao, Y., Long, B., et al. Learning to generate visual questions with noisy supervision. In *NeurIPS*, 2021.

Kang, G.-C., Lim, J., and Zhang, B.-T. Dual attention networks for visual reference resolution in visual dialog. In *EMNLP*, 2019.

Kang, G.-C., Park, J., Lee, H., Zhang, B.-T., and Kim, J.-H. Reasoning visual dialog with sparse graph learning and knowledge transfer. In *EMNLP*, 2021.

Karamanolakis, G., Mukherjee, S., Zheng, G., and Awadallah, A. H. Self-training with weak supervision. In *NAACL*, 2021.

Kim, J.-H., Jun, J., and Zhang, B.-T. Bilinear attention networks. In *NeurIPS*, volume 31, 2018.

Kim, J.-H., Kitaev, N., Chen, X., Rohrbach, M., Zhang, B.-T., Tian, Y., Batra, D., and Parikh, D. Codraw: Collaborative drawing as a testbed for grounded goal-driven communication. In *ACL*, 2019.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *ICLR*, 2014.

Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. M. Opennmt: Open-source toolkit for neural machine translation. In *ACL*, 2017.

Kottur, S., Moura, J. M., Parikh, D., Batra, D., and Rohrbach, M. Visual coreference resolution in visual dialog using neural module networks. In *ECCV*, 2018.

Kottur, S., Moura, J. M. F., Parikh, D., Batra, D., and Rohrbach, M. Clevr-dialog: A diagnostic dataset for multi-round reasoning in visual dialog. In *NAACL*, 2019.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. In *ICCV*, 2017.

Krishna, R., Bernstein, M., and Fei-Fei, L. Information maximizing visual question generation. In *CVPR*, 2019.

Le, H., Sahoo, D., Chen, N. F., and Hoi, S. C. Multimodal transformer networks for end-to-end video-grounded dialogue systems. In *ACL*, 2019.

Lee, D.-H. et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICML Workshop on challenges in representation learning*, 2013.

Li, J., Monroe, W., Ritter, A., Galley, M., Gao, J., and Jurafsky, D. Deep reinforcement learning for dialogue generation. In *EMNLP*, 2016.

Li, J., Monroe, W., Shi, T., Jean, S., Ritter, A., and Jurafsky, D. Adversarial learning for neural dialogue generation. In *EMNLP*, 2017a.

Li, J., Li, D., Xiong, C., and Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*, 2022.

Li, X., Sun, Q., Liu, Y., Zhou, Q., Zheng, S., Chua, T.-S., and Schiele, B. Learning to self-train for semi-supervised few-shot classification. In *NeurIPS*, volume 32, 2019.

Li, Y., Su, H., Shen, X., Li, W., Cao, Z., and Niu, S. Dailydialog: A manually labelled multi-turn dialogue dataset. In *IJCNLP*, 2017b.

Liu, F., Xiang, T., Hospedales, T. M., Yang, W., and Sun, C. ivqa: Inverse visual question answering. In *CVPR*, 2018.

Lu, J., Kannan, A., Yang, J., Parikh, D., and Batra, D. Best of both worlds: Transferring knowledge from discriminative learning to a generative visual dialog model. In *NIPS*, 2017.

Lu, J., Batra, D., Parikh, D., and Lee, S. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, 2019.

Mehri, S. and Eskenazi, M. Unsupervised evaluation of interactive dialog with dialogpt. In *SIGDIAL*, 2020.

Murahari, V., Batra, D., Parikh, D., and Das, A. Large-scale pretraining for visual dialog: A simple state-of-the-art baseline. In *ECCV*, 2020.

Nguyen, V.-Q., Suganuma, M., and Okatani, T. Efficient attention mechanism for visual dialog that can handle all the interactions between multiple inputs. In *ECCV*, 2020.

Niu, Y., Zhang, H., Zhang, M., Zhang, J., Lu, Z., and Wen, J.-R. Recursive visual attention in visual dialog. In *CVPR*, 2019.

Pan, Y., Yao, T., Li, H., and Mei, T. Video captioning with transferred semantic attributes. In *CVPR*, 2017.

Patro, B. N., Kumar, S., Kurmi, V. K., and Namboodiri, V. P. Multimodal differential network for visual question generation. In *EMNLP*, 2018.

Paulus, R., Xiong, C., and Socher, R. A deep reinforced model for abstractive summarization. In *ICLR*, 2018.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Rashkin, H., Smith, E. M., Li, M., and Boureau, Y.-L. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *ACL*, 2019.

Ren, S., He, K., Girshick, R., and Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015.

Rothe, S., Narayan, S., and Severyn, A. Leveraging pre-trained checkpoints for sequence generation tasks. In *Transactions of the Association for Computational Linguistics*, 2020.

Saleh, A., Jaques, N., Ghandeharioun, A., Shen, J., and Picard, R. Hierarchical reinforcement learning for open-domain dialog. In *AAAI*, 2020.

Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., and Monfardini, G. The graph neural network model. In *IEEE Transactions on Neural Networks*. IEEE, 2008.

Schwartz, I., Yu, S., Hazan, T., and Schwing, A. G. Factor graph attention. In *CVPR*, 2019.

See, A., Roller, S., Kiela, D., and Weston, J. What makes a good conversation? how controllable attributes affect human judgments. In *NAACL*, 2019.

Seo, P. H., Lehrmann, A., Han, B., and Sigal, L. Visual reference resolution using attention memory for visual dialog. In *NIPS*, 2017.

Serban, I., Sordoni, A., Lowe, R., Charlin, L., Pineau, J., Courville, A., and Bengio, Y. A hierarchical latent variable encoder-decoder model for generating dialogues. In *AAAI*, 2017.

Shang, L., Lu, Z., and Li, H. Neural responding machine for short-text conversation. In *ACL*, 2015.

Sharma, P., Ding, N., Goodman, S., and Soricut, R. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018.

Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C. A., Cubuk, E. D., Kurakin, A., and Li, C.-L. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *NeurIPS*, 2020.

Thomason, J., Murray, M., Cakmak, M., and Zettlemoyer, L. Vision-and-dialog navigation. In *CoRL*, 2020.

Wang, K., Tian, J., Wang, R., Quan, X., and Yu, J. Multi-domain dialogue acts and response co-generation. In *ACL*, 2020a.

Wang, Y., Joty, S., Lyu, M. R., King, I., Xiong, C., and Hoi, S. C. Vd-bert: A unified vision and dialog transformer with bert. In *EMNLP*, 2020b.

Wu, Q., Wang, P., Shen, C., Reid, I., and Van Den Hengel, A. Are you talking to me? reasoned visual dialog generation through adversarial learning. In *CVPR*, 2018.

Xie, Q., Dai, Z., Hovy, E., Luong, T., and Le, Q. Unsupervised data augmentation for consistency training. In *NeurIPS*, 2020a.

Xie, Q., Luong, M.-T., Hovy, E., and Le, Q. V. Self-training with noisy student improves imagenet classification. In *CVPR*, 2020b.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., and Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015.

Yang, A., Miech, A., Sivic, J., Laptev, I., and Schmid, C. Just ask: Learning to answer questions from millions of narrated videos. In *ICCV*, 2021.

Zhang, Y., Sun, S., Galley, M., Chen, Y.-C., Brockett, C., Gao, X., Gao, J., Liu, J., and Dolan, B. Dialogpt: Large-scale generative pre-training for conversational response generation. In *ACL*, 2020.

Zheng, Z., Wang, W., Qi, S., and Zhu, S.-C. Reasoning visual dialogs with structural and partial observations. In *CVPR*, 2019.

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., and Fidler, S. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *ICCV*, 2015.

Zoph, B., Ghiasi, G., Lin, T.-Y., Cui, Y., Liu, H., Cubuk, E. D., and Le, Q. Rethinking pre-training and self-training. In *NeurIPS*, 2020.

A. Related Work

Visual dialog. Visual Dialog (VisDial) (Das et al., 2017) has been proposed as an extended version of Visual Question Answering (VQA) (Antol et al., 2015; Anderson et al., 2018; Kim et al., 2018), where a dialog agent should answer a series of interdependent questions using an image and the dialog history. Prior work has developed a variety attention mechanisms (Lu et al., 2017; Seo et al., 2017; Wu et al., 2018; Kottur et al., 2018; Niu et al., 2019; Schwartz et al., 2019; Guo et al., 2019; Gan et al., 2019; Kang et al., 2019; Nguyen et al., 2020) considering the interactions among the image, dialog history, and question. Some studies (Zheng et al., 2019; Kang et al., 2021) have attempted to discover the semantic structures of the dialog in the context of graph neural networks (Scarselli et al., 2008) using the soft attention mechanisms (Bahdanau et al., 2014). From the learning algorithm perspective, all of them have relied on supervised learning on VisDial data. More recently, a line of research (Murahari et al., 2020; Wang et al., 2020b; Chen et al., 2022) has employed self-supervised pre-training to leverage the knowledge of related vision-and-language datasets (Sharma et al., 2018; Antol et al., 2015; Zhu et al., 2015). However, our approach is based on semi-supervised learning and produces the task-specific data (*i.e.,* visual dialogs) for unlabeled images to train the dialog agent.

Sequence generation in vision-and-language tasks. Many studies have generated natural language for the visual inputs such as image captioning (Xu et al., 2015; Anderson et al., 2018), video captioning (Iashin & Rahtu, 2020; Pan et al., 2017), visual question generation (VQG) (Kai et al., 2021; Krishna et al., 2019; Fan et al., 2018b; Liu et al., 2018; Patro et al., 2018; Jain et al., 2017), visual dialog (VisDial) (Das et al., 2017; Gan et al., 2019), and video dialog (Alamri et al., 2019; Le et al., 2019). Furthermore, recent studies (Yang et al., 2021; Li et al., 2022) have produced text data for vision-and-language pre-training. GST is similar to these studies in that the model generates the text data, but our focus is on studying the effect of semi-supervised learning (SSL) on top of such pre-training approaches. To the best of our knowledge, GST is the first approach to show the efficacy of SSL throughout a wide range of visual QA tasks.

Neural dialog generation. In NLP literature, extensive studies have been conducted regarding neural dialogue generation for both open-domain dialogue (Zhang et al., 2020; Shang et al., 2015; Li et al., 2016; Serban et al., 2017; Saleh et al., 2020; Li et al., 2017a) and task-oriented dialogue (Wang et al., 2020a; Huang et al., 2020). Our approach is similar to neural dialogue generation in that the model should generate a corresponding response based on the dialog history and the current utterance. However, we aim to produce *visually-grounded* dialogs, and thus the image-groundedness of the question and the semantic correctness of the answer are important. On the other hand, neural dialogue generation considers many different aspects: specificity, response-relatedness (See et al., 2019), interestingness, fluency (Mehri & Eskenazi, 2020), and diversity (Li et al., 2016).

B. Implementation Details

As mentioned in Section 2.2, we integrate the vision-and-language encoder for a discriminative task (*i.e.,* VisDial-BERT (Murahari et al., 2020)) with the transformer decoder for sequence generation (*i.e.,* $\mathrm{BERT}_{\mathrm{BASE}}$ (Rothe et al., 2020)) to train the teacher, the questioner, and the student. The encoder architecture is based on the vision-and-language pre-training model, ViLBERT (Lu et al., 2019). The decoder has 12 layers of transformer blocks, with each block having 12 attention heads and a hidden size of 768. The maximum sequence length of the encoder and the decoder is 256 and 25, respectively. We extract the feature vectors of the input images by using the Faster R-CNN (Ren et al., 2015; Anderson et al., 2018) pre-trained on Visual Genome (Krishna et al., 2017). The number of bounding boxes for each image is fixed to 36. We set the threshold for PPL τ to 50. We train on one A100 GPU with a batch size of 72 for 70 epochs. Training time takes about 3 days. We use the Adam optimizer (Kingma & Ba, 2014) with an initial learning rate 1e-5. The learning rate is warmed up to 2e-5 until 10k iterations and linearly decays to 1e-5. In visually-grounded dialog generation, the questioner and the teacher decode the sequences using the top-k sampling (Fan et al., 2018a; Holtzman et al., 2018; Radford et al., 2019) with $k = 7$ and the temperature of 0.7. We use the top-k sampling since its computation is cheap yielding accurate and diverse sequences. Furthermore, we apply the 4-gram penalty (Paulus et al., 2018; Klein et al., 2017) when generating visual questions to ensure that no 4-gram appears twice in the questions for each dialog.

C. Qualitative Results and Analysis

For qualitative analysis, we visualize the generated conversations from our models and the ones from human. We excerpt the human conversation from the VisDial v1.0 validation dataset, and the questioner and the student generate the machine conversation using the image and the caption in the validation data. As in Figure 2, diverse visual questions and correct

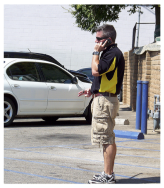
Image & Caption	Human Conversation	Machine Conversation
 A man is talking on a cell phone in a parking lot	D1 : Is it daytime? Yes D2 : Is this is a busy full parking lot? No D3 : Is the man dressed for work? No D4 : What is he wearing? Polo shirt and shorts D5 : Are there other people around? No D6 : Do you see any cars? Yes D7 : Are there a lot of cars? No D8 : How many cars do you think there are? 3 D9 : What colors are the cars? White and black D10 : Is the man wearing sunglasses? Yes	D1 : Is the man wearing a hat? No D2 : Is the man dressed for cold or warm weather? Warm weather D3 : Is the man talking on a cell phone? Yes he is D4 : Is it a flip phone? Yes it is D5 : What color is the phone? Black and silver D6 : Is the man holding it in his hand? Yes it is D7 : Are there any cars in the parking lot? Yes D8 : Is he waiting for someone to get on the car? No just standing D9 : Is the man looking at the camera? No he is looking off to the side D10 : Are there any buildings in the background? \\ There is a brick building behind him
 A person stand in front of a brown horse	D1 : Is the horse wearing a saddle? No D2 : Is it a large horse? Yes D3 : Is the person male or female? Female D4 : About what age is she? She is about 25 D5 : Is she wearing riding gear? Yes D6 : Is this a country setting? Yes it is D7 : Do you see a barn? No D8 : Does the horse look well cared for? Yes it do D9 : Is there a blanket on the horse? No D10 : Is the weather nice? Yes	D1 : Is the horse large? Yes D2 : Is it facing the camera? No D3 : Is the person a man or a woman? Woman D4 : Does the horse have a saddle? No she does not D5 : Is it sunny? It is hard to tell but it is daylight D6 : Is the woman wearing a hat? Yes she is D7 : What color is the woman's hair? It appears to be brown D8 : Does the woman have long hair? I cannot tell D9 : What color is her shirt? She is wearing a black jacket D10 : Are there any trees? There are no trees
 A woman rides her bike on a sidewalk next to a fence	D1 : Is it daytime? Yes D2 : Is the sun out? Yes D3 : How old is the woman? 24 D4 : Is she alone? Yes D5 : Is she wearing sunglasses? No D6 : Is she carrying anything else with her? No D7 : Is she wearing sneakers? Yes D8 : What color is her shirt? Wearing a black jacket D9 : Is she smiling? No D10 : Where is she going? West on the road	D1 : Is this a color photo? Yes, it is a color photo D2 : Is it daytime? Yes D3 : What is the weather like? Looks like sunny D4 : What color is the woman's hair? Hair is brown D5 : How old does she seem? Around 35 D6 : Is she wearing a helmet? No helmet seen D7 : Is she wearing jeans? No D8 : Is she wearing glasses? No D9 : What color is her bike? **Bike is blue in color** D10 : Are there any cars? Yes, i see 1 car

Figure 2. A visualization of human conversation and machine conversation on the VisDial v1.0 validation images. We excerpt the human conversation from the validation dataset, and the questioner and the student generate the machine conversation. The red-colored text denotes an incorrect answer.

answers are generated in the machine conversation. For instance, from D3 to D6 in the first example, the question generation model asks about "a cell phone," which first appears in the caption, whereas the human questioner deals with different topics. The student responds appropriately to the given questions. However, we also identify that the student sometimes fails to generate correct answers (*i.e.,* the red-colored text in D9 of the last example), which shows the importance of more precise vision-and-language grounding.