Think Beyond Words: Exploring Context-Relevant Visual Commonsense for Diverse Dialogue Generation

Yiting Liu[1,2], **Liang Li**[1]*, **Beichen Zhang**[2], **Qingming Huang**[1,2],
[1] Key Laboratory of Intelligent Information Processing,
Institute of Computing Technology, Chinese Academy of Sciences (ICT/CAS)
[2] University of Chinese Academy of Sciences
{liuyiting21s,liang.li}@ict.ac.cn
beichen.zhang@vipl.ict.ac.cn , qmhuang@ucas.ac.cn

Abstract

Commonsense knowledge has been widely considered for building intelligent open-domain dialogue agents, aiming to generate meaningful and diverse responses. Previous works in this field usually lack the ability to effectively obtain and utilize auxiliary commonsense from the external visual world. In this paper, we argue that exploiting logical information in images related to context can be effective to enrich and steer the generation process. In view of this, we propose VICTOR, a context-relevant VIsual Commonsense enhanced dialogue generaTOR for generating coherent and informative responses. To obtain the associated visual commonsense, we devise a novel approach that expands topic words on the knowledge graph and maps them into daily scenarios. During the generation, the model adopts multimodal fusion mechanism to integrate visual and textual information, and adaptively combine their decoding distributions for better response generation. The experimental results on two public datasets show that our proposed method outperforms the latest competitive methods in terms of coherence and diversity.

Figure 1: Examples from two pure language dialogue datasets, where the underlined green part is the output that needs to be generated[1]. The hidden visual memory, which contains associative commonsense and needs to be explored, can be essential for humans to make proper responses during the conversation.

1 Introduction

Building intelligent dialogue systems is a long-standing goal of artificial intelligence and has attracted increasing research attention in recent years. An ideal conversation agent is supposed to generate diverse and informative responses without sacrificing their relevance to the dialogue context. To avoid general and dull dialogue generation (Li et al., 2016), some approaches modify model architecture to manipulate latent variables and target distributions (Lin et al., 2020; Wang et al., 2021), yet these works limit themselves to original conversations without considering useful auxiliary information.

Another series of solutions augment the training corpus with extra information like emotions or personality (Mazare et al., 2018; Song et al., 2019). Following this line, works like Su et al. (2020); Majumder et al. (2021) introduce more general non-conversation text like forum comments and stories to help generate richer responses. However, these works only consider information stored in pure text, ignoring the grounding information from the external visual world, which is essential for generating really meaningful language (Harnad, 1990; Bisk et al., 2020).

As shown in Figure 1, it is natural that when making a conversation, we do not only focus on current context. We also expand or transition the topics by using associative memory gained from the physi-

*Corresponding author: Liang Li

[1]The slight difference in forms between these two datasets will be discussed in section 4.1

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 3106 - 3117
December 7-11, 2022

cal world, so that the chat can be more engaging and last longer. In this work, we introduce visual commonsense as the logical semantic information stored in visual scenes from daily life. Considering that images representing everyday scenarios are typically logical and grounded in commonsense, it is reasonable to introduce them into open-domain conversation as additional information. Liang et al. (2021); Shen et al. (2021) are pioneering works that introduce visual information into the general open-domain response generation. However, these works only connect visual information by simply matching the context representation with images, without explicitly considering the topic transition of conversation. This may lead to monotonous and narrow semantics of the responses. Besides, the semantic gap between modalities makes it difficult for these methods to effectively integrate visual features. Furthermore, they ignore the balance between the contributions of two modalities in the decoding stage.

To alleviate the above issues, we present VICTOR, a context-relevant visual commonsense enhanced dialogue generator, which consists of three components: visual commonsense retriever, multimodal fusion block, and self-adaptive response generator. The visual commonsense retriever first extracts concept words from context. Then, in order to acquire explicit commonsense knowledge, it explores related concepts by multi-hop searching on knowledge graphs. Each of these related concepts will be considered globally and mapped into the corresponding images, which then produce captions to narrow the semantic gap. In this way, we obtain visual commonsense with rich associative semantic information.

To facilitate diverse dialogue, our multimodal fusion block incorporates auxiliary visual knowledge at each decoding step. It encodes visual commonsense with a transformer block and utilizes a co-attention mechanism to fuse two modalities. The response generator is based on GPT-2 model. It takes knowledge pairs gained from knowledge graphs as guidance to encourage consistent responses with relevant topics. Finally, at each decoding step, the generator uses soft probability to adaptively combine the distributions based on the textual and visual information. We demonstrate the effectiveness of our approach on two public datasets in comparison with various representative baselines.

Our contributions are summarized as follows:

- We present a novel approach to retrieve visual scenes based on dialogue. It expands concepts on knowledge graphs and maps them to unpaired image data, so as to acquire context-related visual commonsense with high quality.
- We propose VICTOR, a new conversation agent that fuses multimodal information to enrich and steer the generation process. It adaptively balances textual information from context and external visual commonsense, generating diverse responses while maintaining their coherence with contexts.
- We conduct extensive experiments on two open-domain dialogue datasets. The results show the effectiveness of our proposed method, and verify the potential of exploiting multimodal information for intelligent conversation agents.

2 Related Work

2.1 Controllable dialogue response generation

The goal of open-domain dialogue systems is to establish engaging conversations with users. To satisfy the human need for communication and affection, an ideal conversation agent always has a higher requirement in consistency, semantics and diversity (Huang et al., 2020). Therefore, constraints on conversation attributes like persona (Mazare et al., 2018; Zhang et al., 2018) and sentiment (Song et al., 2019; Shen and Feng, 2020), and external non-conversation data like documents and knowledge base (Li et al., 2020; Majumder et al., 2020) are introduced to control the dialogue response and improve the interactivity of the conversation model. However, most of these works use additional constraints or guiding information in the form of pure text, neglecting the rich commonsense knowledge stored in the visual scene.

2.2 Multimodal open-domain dialogue

Along with the thriving of multimodal learning for tasks like captioning (Tu et al., 2022; Li et al., 2022) and entity mapping (Li et al., 2018; Liu et al., 2022), the use of visual information for improving language tasks has also shown great potential in areas such as machine translation (Caglayan et al., 2019; Fang and Feng, 2022) and semantic parsing (Shi et al., 2019; Kojima et al., 2020). However, its exploration for enhancing dialogue generation is still limited.

Early attempts on this issue assume the conversation to be grounded on a given image (Mostafazadeh et al., 2017; Shuster et al., 2020). Yang et al. (2021) tries to recover the latent image of the conversation using conditional variational auto-encoding framework (Sohn et al., 2015). Recent researches (Liang et al., 2021; Shen et al., 2021) have taken it a step further by matching context with extra image data. Distinct from these existing works, our method expands original topics from context by searching from commonsense knowledge base, and uses corresponding images to explore valid visual information for response generation.

3 The Proposed Method

In this section, we first introduce our task formulation for open-domain dialogue generation with visual commonsense, and then illustrate the three main components of our proposed VICTOR model.

3.1 Task Formulation

Let $\mathcal{D}_T = \{(C_1, R_1), (C_2, R_2), \ldots, (C_n, R_n)\}$ denotes the parallel conversational corpus[2]. C_i is the context and R_i is the corresponding response. $\mathcal{D}_I$ denotes our collected image data. We assume that for each dialogue context C_i, we can find an image subset $V_i = \{v_{i1}, v_{i2}, \ldots, v_{im}\}$ containing visual commonsense to assist the response generation, where $V_i \subseteq \mathcal{D}_I$. Thus our goal is to learn a generation model $P(R_i|C_i, V_i)$ from $\mathcal{D}_T$ and $\mathcal{D}_I$.

3.2 Visual Commonsense Retrieval

As shown in Figure 2, we design a static approach to retrieve related visual commonsense for each conversation context.

Concepts Expansion Since an engaging conversation requires dialogue agents to be able to pro-actively introduce new relevant topics, we expand the topic concepts by searching from ConceptNet (Speer et al.), a commonsense knowledge base. Following Ji et al. (2020), we first perform fuzzy matching with lemmatized form of surface texts to extract topic concepts from provided conversation context. After removing stopwords, we keep verbs and nouns as our original topic concepts T_o.

We consider the original concepts as the initial nodes, and iteratively search for their directed neighbours in ConceptNet for H hops (H iterations). During each hop, we preserve top N of the neighbouring concept nodes by the standard of their incoming degree. Hence, we got the expanded topic concepts $T_e = \{t_1, t_2, \ldots, t_m\}$.

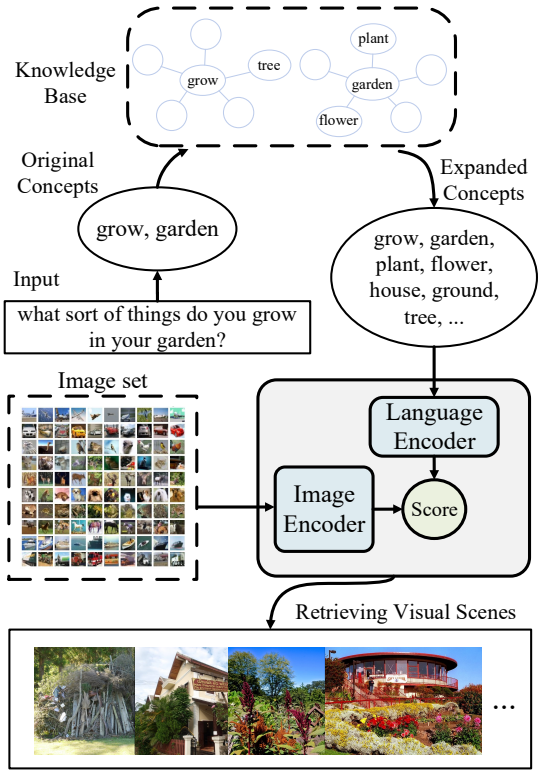


Figure 2: Retrieval process: Extracting and expanding the context concepts, and mapping them to corresponding images.

Image Mapping Our attempt is to utilize commonsense knowledge existing in the corresponding visual scenarios of the conversation topics. It is intuitive to consider the connection among chosen concepts rather than mapping them separately into visual space. Since there is no large scale aligned dialogue-image dataset available, we train our concept-image matching model from MSCOCO (Lin et al., 2014), a commonly used image-captioning dataset containing sentence-image pairs. Following Tan and Bansal (2020), we align each token in the caption s to the paired image, and perform token-level matching.

To extract feature representation of text and image, we adopt pretrained language and visual model (here we use $\text{BERT}_{\text{BASE}}$ (Devlin et al., 2018) and ResNeXt (Xie et al., 2017) respectively) to operate the encoding process. We then project the feature vectors of the two modalities into aligning space, and normalize them to norm-1 vectors of the same

[2]To illustrate our method concisely, we focus on single-turn dialogue generation here. Our approach will also work in the multi-turn setting when we use dialogue history as context.

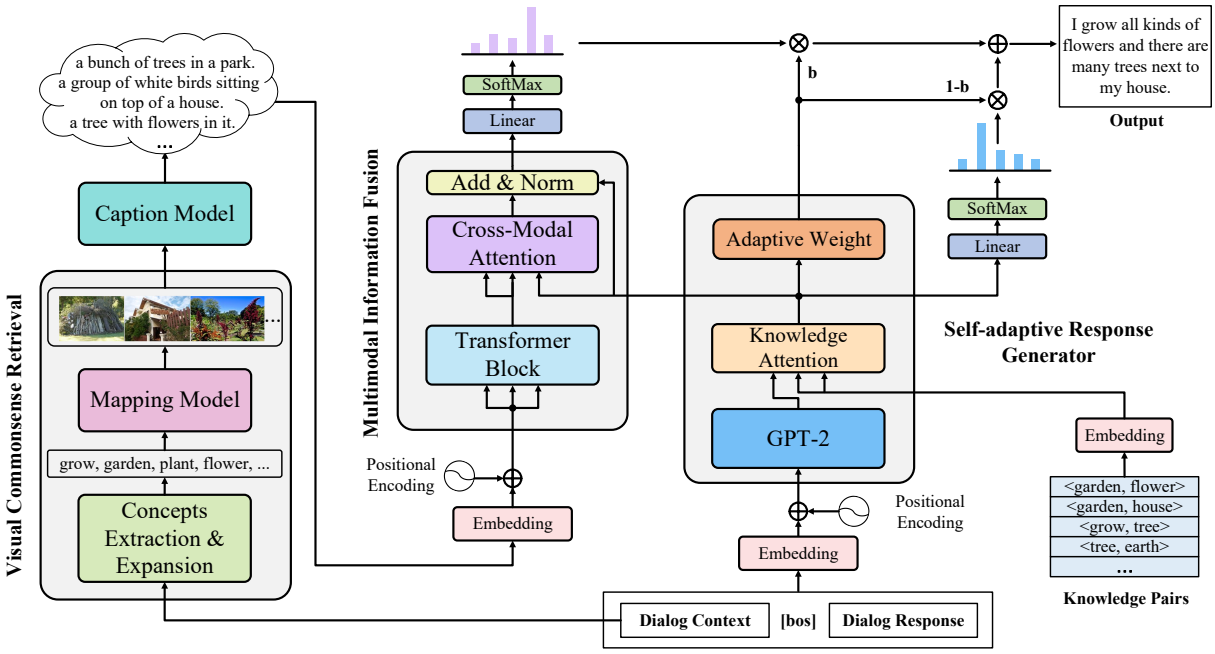


Figure 3: The overall framework of VICTOR.

dimension d:

$$H_s = f_{map}(\text{BERT}(s)) \in \mathbb{R}^{L\times d},$$
$$h_v = f_{map}(\text{ResNeXt}(v)) \in \mathbb{R}^{d} \quad (1)$$

where the mapping function $f_{map}(.)$ is a multi-layer perceptron followed by normalization function, L is the sentence length. Thus we get the aligned textual and visual representation $H_s = \{h_{si}\}$ and h_v.

The relevance score of two modalities will be measured by the inner product of their representations. Finally, hinge loss is adopted to optimize the matching model:

$$\text{score}(w_i, v) = h_{si}^{\top} h_v,$$
$$\mathcal{L}_{hinge}(s, v, v^-) = \sum_{i=1}^{L} \max\{0, \quad (2)$$
$$\alpha - \text{score}(w_i, v) + \text{score}(w_i, v^-)\}$$

where v^- is the randomly selected negative image sample, α is the margin between the similarities of a positive and a negative pair.

After training the token-image matching model, it takes the expanded topic concepts T_e to retrieve their matched images. We keep the top K images for each concept word regarding their relevance scores. Thus we get the corresponding visual scenes $V = \{v_1, v_2, \ldots, v_m\}$, which contains the desired commonsense knowledge.

3.3 Multimodal Information Fusion

A commonly-used captioning model[3] pretrained on MSCOCO dataset is adopted to caption the former retrieved image for each concept. The assumption is that caption-styled visual information is easier for the model to exploit than roughly extracted visual features. Then we concatenate these captions using token $[cap]$. Thus we get the corresponding visual commonsense $V_c = \{u_1, \ldots, u_z\}$, where z is its total length. After that, we utilize transformer block(TB) (Vaswani et al., 2017) to obtain the representation of the visual commonsense. Formally, the representation of each V_c is calculated by:

$$e_i = w_i W_{emb} + PE(i),$$
$$I_n^v = [e_1, \ldots, e_z], \quad (3)$$
$$H^v = \text{TB}(I_n^v, I_n^v, I_n^v),$$

where $W_{emb} \in \mathbb{R}^{d_{voc}\times d_h}$ is the word embedding matrix from the generator, d_{voc} is the size of the vocabulary. $\text{PE}(.)$ is the position embedding to make use of the sentence order.

Afterward, we apply the fusion module to incorporate the context information and visual knowledge, so as to determine the external information desired by current context. Formally, at each decoding step t, the response generator will produce the hidden state $\widetilde{h}_t^c$ which encodes the current context (details will be described in the next section).

[3]https://github.com/sgrvinod/a-PyTorch-Tutorial-to-Image-Captioning

We leverage the hidden state as a context query, and use multi-head attention layer to capture the correlated visual information h_t^{vc} from H^v:

$$h_t^{vc} = \text{MultiHead}(\widetilde{h}_t^c, H^v, H^v) \quad (4)$$

At the t-th decoding step, based on the extracted commonsense information, the decoding distribution over the vocabulary decided by visual knowledge can be produced by:

$$P_V(s_t|s_{<t}, V) = \text{softmax}(\text{Linear}(h_t^{vc})) \quad (5)$$

3.4 Self-adaptive Response Generator

The generation network is based on GPT-2 (Radford et al., 2019), a pretrained multi-layer transformer decoder which learns the language granularity from large amounts of open Web text data.

As shown in Figure 3, given a dialogue context C, the decoding process of each step t is as follows: By using GPT-2 model, we first obtain the hidden state h_t of the current context. To encourage the generated response to use topic knowledge, we explicitly consider the extracted concepts here. As we get the expanded concepts set by searching for neighbours on external knowledge bases earlier, we can obtain the related concepts pairs $T_{pr} = \{[t_1^{hd}, t_1^{tl}], [t_2^{hd}, t_2^{tl}], \ldots, [t_k^{hd}, t_k^{tl}]\}$, where t_i^{tl} is the tail concept found by neighbouring search from head concept t_i^{hd}. Inspired by Nie et al. (2019), we first embed the two concepts of each pair and thereafter concatenate them to obtain the related-concepts embedding. Then we use h_t to query from embedded pairs by applying single-layer multi-head attention layer, getting topic-aware $\widetilde{h}_t^c$:

$$\begin{aligned} h_t^c &= \text{GPT}(H_{\leq t}^c), \\ E^{T_{pr}} &= \text{Linear}(\text{Concat}(\{[t_i^{hd}, t_i^{tl}]\}W_{emb})), \\ \widetilde{h}_t^c &= \text{MultiHead}(h_t^c, E^{T_{pr}}, E^{T_{pr}}) \end{aligned} \quad (6)$$

the probability distribution of the t-th token decided by textual knowledge will then be computed as follows:

$$P_{LM}(s_t|s_{<t}, T_e) = \text{softmax}(\text{Linear}(\widetilde{h}_t^c)) \quad (7)$$

Since different conversation turns may require various information, it is crucial to balance the textual information from context, which constrains the direction of this conversation, and previously obtained visual knowledge, which indicates related commonsense from real world grounding. Thus we utilize a weighted average score β to decide the different levels of contribution of these two knowledge sources, for generating ideal responses. Instead of fixing a manual hyperparameter to adjust the balance, we adopt self-adaptive weight (See et al., 2017) based on the current hidden states of the context:

$$\beta_t = \sigma(\text{Linear}(\widetilde{h}_t^c)), \quad (8)$$

then we can obtain the following combined decoding distribution:

$$\begin{aligned} P(s_t|s_{<t}) = \beta_t P_{LM}(s_t|s_{<t}, T_e) \\ + (1-\beta_t) P_V(s_t|s_{<t}, V) \end{aligned} \quad (9)$$

Finally, following the standard practice of dialogue response generation, we optimize our proposed model with the cross entropy loss:

$$\mathcal{L}_{ce} = -\sum_{i=1}^{L} \log(P(s_t|s_{<t})) \quad (10)$$

4 Experimental Settings

4.1 Datasets

We conduct our experiments on two open-domain dialogue corpus, OTTers (Sevegnani et al., 2021) and DailyDialog (Li et al., 2017). OTTers is a dialogue dataset of human one-turn topic transitions. Unlike other common dialogue datasets which contain a large number of short, generic responses, each utterance in OTTers has a specific topic and is therefore more informative. OTTers is slightly different from other dialogue corpus in form: given one turn conversation $[u_a, u_b]$, where each utterance has a different topic, the goal is to generate a transition response u_t to serve as a smooth link between them. This dataset is exactly suitable for testing our model, since the response generation requires associative commonsense knowledge. During the experiments, we concatenate $[u_a, u_b]$ using separator tokens as the model inputs, and treat u_t as the outputs. To test the generalization ability of our model and make a fair comparison with other baselines, we also evaluate VICTOR on the commonly used DailyDialog dataset. The examples of both datasets are shown in Figure 1.

For image retrieval, we train our mapping model on MSCOCO dataset. We randomly sample 100K images from the Open Images dataset (Kuznetsova et al., 2020) as our candidate image set $\mathcal{D}_I$, then we retrieve images from it following section 3.2.

4.2 Comparison Methods

To demonstrate the effectiveness of our proposed model, we compare it with the following representative methods: (1) **Seq2seq**: a classic encoder-decoder framework (Sutskever et al., 2014) with global attention(Luong et al., 2015). (2) **GPT-2**: a pretrained GPT-2 model (Radford et al., 2019) fined-tuned on the task datasets. (3) **GRF**: a GPT-based generation model (Ji et al., 2020), which performs multi-hop reasoning on knowledge graphs using graph convolution network (GCN) (Kipf and Welling, 2016). (4) **GVT**: a variational transformer(Lin et al., 2020) that uses CVAE to model the discourse-level diversity with a global latent variable. (5) **AdaLabel**: an adaptive label smoothing approach(Wang et al., 2021) that diversifies dialogue generation by adaptively estimating the soft target label distribution.

Among these comparison methods, Seq2seq is a standard generation model, GPT-2 is a commonly used pretrained language model, GVT and AdaLabel are both transformer-based models for diverse dialogue generation, GRF and AdaLabel are state-of-the-art approaches for the datasets we use.

4.3 Evaluation Metrics

Automatic Evaluation We hypothesize that our proposed approach, which leverages external topic-aware visual commonsense, can increase the diversity of the generated responses, while maintaining relevance to their corresponding contexts. For fluency, we use **Perplexity** (Serban et al., 2015) to measure the confidence of the generated responses. A relatively low perplexity indicates better fluency. For relevance, we adopt widely used **BLEU** (Papineni et al., 2002) (here we use BLEU-1 and BLEU-4) and **Rouge-L** (Lin, 2004) to measure the n-gram overlaps between ground truth references and the generated responses. To measure the diversity, we report the percentage of distinct uni-grams and bi-grams (**Dist-1** and **Dist-2** respectively) (Li et al., 2016) in all generated responses.

Human Evaluation Considering that the automatic metrics are not always accurate to evaluate the responses (Liu et al., 2016), we further conduct manual evaluation following previous works Wu et al. (2021); Zou et al. (2021). Specifically, we randomly sample 200 testing pairs from each test set. Given a dialogue context, three annotators are asked to conduct pair-wise comparison between the responses generated by VICTOR and three strong baselines, including state-of-the-art methods (1200 comparisons with three baselines on two datasets in total). For each comparison, three annotators are required to compare the responses from the following perspectives: fluency, context coherence, informativeness. The annotators need to judge which response is better independently. If the two responses are both proper or inappropriate, the comparison of this pair is treated as "draw". Ultimately, we average the results of three annotators and calculate their Fleiss' kappa scores (Fleiss, 1971).

4.4 Implementation Details

During the topic-expansion, we set the number of hops $H = 2$ and preserve top $N = 5$ concepts per hop. For the retrieval model, the concatenation of the last 4 layers of BERT output and image features from ResNeXt-101-32x8d are used as embedding of each modality. We set the hidden size d of the aligning space to 256 and the hinge loss margin α to 0.5. We test the performance of retrieving different number of top-scored images for each cocept, and set $K = 1$ for its best result (see section 5.4). The pretrained captioning model is combined with a ResNet-101 encoder and a LSTM decoder.

For the generator, we base our model on gpt2-small[4] (Transformer with 12 layers, 768 hidden size, 12 heads) . The multi-head transformer block for encoding visual commonsense has the structure of 6 layers, 768 hidden size and 6 heads. To train the model, we use the Adam optimizer (Kingma and Ba, 2014) with a learning rate of 1e-6. At the inference stage, the maximum decoding length of the response is set to 40, and we adopt beam search decoding with a beam size of 3. All our experiments are implemented with PyTorch, and the entire model is trained on RTX3090 GPUs.

5 Results and Analysis

5.1 Automatic Evaluations

As shown in Table 1, our proposed model VICTOR outperforms baselines on most automatic metrics in these two datasets. In the aspect of relevance, it beats baselines in all related metrics, indicating responses generated by VICTOR can be coherent with the help of context-related knowledge. Meanwhile, enhanced by the extracted visual commonsense, VICTOR also achieves the best performance in Dist-1/2, showing it can generate diverse and informative responses. Besides, we can see

[4]https://huggingface.co/gpt2

Dataset	OTTers						DailyDialog					
Model	PPL	B-1	B-4	RG	D-1	D-2	PPL	B-1	B-4	RG	D-1	D-2
Seq2seq	52.85	13.88	1.14	14.1	6.18	15.37	47.24	12.54	2.55	22.16	6.39	25.95
GPT-2	16.37	14.36	2.30	18.72	18.03	40.48	**19.65**	16.31	2.53	21.93	7.41	23.85
GRF	17.8	17.52	2.95	18.81	21.78	47.86	19.88	16.08	3.31	22.19	11.19	35.72
GVT	37.74	14.83	0.91	13.13	18.49	48.11	34.19	22.27	9.64	22.92	6.57	36.11
AdaLabel	33.16	17.27	1.71	18.21	16.79	39.88	30.92	24.12	8.46	27.65	9.95	39.12
VICTOR	**16.29**	**21.49**	**4.82**	**20.09**	**24.41**	**56.64**	22.21	**29.15**	**14.89**	**30.21**	**14.14**	**46.47**

Table 1: Automatic evaluation results(%). The metrics Perplexity, BLEU-1/4, Rouge-L, Dist-1/2 are abbreviated as PPL, B-1/4, RG, D-1/2 respectively. The best results are highlighted in bold.

	Opponent	Win	Loss	Draw	Kappa
(a)	VIC vs. GRF	35.5%	14.2%	50.3%	0.54
	VIC vs. GVT	54.5%	7.8%	37.7%	0.43
	VIC vs. AdaLabel	61.3%	11.5%	27.2%	0.54
(b)	VIC vs. GRF	39.5%	23.5%	37.0%	0.59
	VIC vs. GVT	57.8%	15.5%	26.7%	0.56
	VIC vs. AdaLabel	42.0%	12.8%	45.2%	0.66

Table 2: Human evaluation results on (a) Otters and (b) DailyDialog datasets. VICTOR is abbreviated as VIC.

that although AdaLabel can generate relatively diverse responses, the lack of context related external knowledge prevents it from keeping high relevance to the context. This problem can be particularly acute with OTTers dataset, since most dialogues in it are topic specific. The same problem also affects GVT model, without the assistance of commonsense knowledge, it performs rather poorly on OTTers dataset. Although the GRF baseline integrates information from knowledge bases, its performance is worse than our model on both relevance and diversity. This indicates the superiority of considering commonsense information in visual scenes rather than just pure textual knowledge.

5.2 Human Evaluations

The human evaluation results are shown in Table 2. Not surprisingly, VICTOR consistently outperforms all the strong baselines and achieves significant improvements on both datasets. We also analyze the bad cases and find that the baselines still suffer from the general or irrelevant responses. The evaluation result indicates that VICTOR can generate more coherent and informative responses that are attractive to annotators. This validates the benefits of the context-relevant visual commonsense and the fusion mechanism. We also employ Fleiss' kappa scores to measure the reliability between different annotators, and results show that annotators reach a moderate agreement.

5.3 Ablation Study

To investigate the effectiveness of each part of VICTOR, we conduct ablation studies on two datasets by removing or replacing particular modules from the original model. Here, we have three variants: (1) w/o. VC: removing visual commonsense extraction and multimodal fusion block. (2) w/o. AW: removing the adaptive weight of the response generator and replacing it with a fixed weight of 0.5. (3) w. RF: replacing the caption-styled visual commonsense with ResNeXt features of the same image, which can be obtained by using the pretrained image encoder from our retrieval model.

The ablation results are shown in Table 3. We observe that without fusing visual commonsense, the performance of variant-1 drops sharply with respect to relevance and diversity metrics. The result verifies the effectiveness of integrating context-relevant visual knowledge into response generation. Besides, although variant-2 maintains a relatively high diversity, the values of relevance metrics drop largely due to the fixed balancing weight of the generator. This indicates that adaptively deciding the contribution of language and visual knowledge plays an important role in the generation process for different conversation turns. We also witness a small drop in performance of the variant-3, which uses ResNeXt features instead of the image captions as the visual commonsense source. As shown in previous researches (Jin et al., 2022; Feng et al., 2021), this phenomenon can be explained by the fact that captions of everyday scenarios, which dampen the reporting bias of the general text cor-

Dataset	OTTers						DailyDialog					
Model	PPL	B-1	B-4	RG	D-1	D-2	PPL	B-1	B-4	RG	D-1	D-2
VICTOR	**16.29**	**21.49**	**4.82**	**20.09**	**24.41**	**56.64**	**22.21**	**29.15**	**14.89**	**30.21**	**14.14**	**46.47**
w/o. VC	16.52	15.39	2.59	18.91	21.02	43.16	28.89	19.01	6.87	22.04	7.19	28.53
w/o. AW	23.18	15.92	3.3	19.16	24.2	50.75	26.36	23.23	10.47	29.88	14.01	43.61
w. RF	19.04	18.53	4.09	19.72	24.27	53.81	22.35	28.86	13.05	28.61	11.94	42.2

Table 3: Ablation study results(%) on two datasets.

K	0	1	2	3	rand_3
B-1	15.39	**21.49**	19.32	19.02	20.34
D-1	21.02	**24.41**	23.26	22.60	23.61

Table 4: Influence of the number of retrieved images on OTTers dataset(%). K means concatenating captions of top K images as visual commonsense, $K = 0$ is equivalent to not using visual information, $rand\_3$ means randomly choosing from top 3 images.

pus, are better carriers of logical commonsense and contain less noise than roughly extracted image features.

5.4 Number of images

We further study the effect of visual commonsense by varying the number of retrieved images and conducting experiments on OTTers dataset. As shown in Table 4, all the results obtained with the help of visual commonsense are better than those without, while choosing top 1 image helps achieve best performance. This can be explained that each selected image refers to key information of all core concepts, resulting in partial semantic overlap, thus additional selection of more images may introduce more unnecessary noise, which is not helpful to the generation.

5.5 Case Study

To further investigate the quality of responses generated by VICTOR, and compare the results with other baselines intuitively, we show two dialogue cases from the two datasets in Figure 4. As we can see, the retrieval process can obtain proper expanded concepts from knowledge graphs and retrieve related images. The corresponding captions with logical commonsense will then bring auxiliary visual information into the generation process. In these two cases, although all four models have generated fluent and informative responses, compared with the other three strong baselines, responses generated by VICTOR are clearly more consistent with the context and more engaging. Again, the results prove the effectiveness of exploring context-relevant visual commonsense for dialogue generation.

6 Conclusion

In this work, we propose a novel context-relevant visual commonsense enhanced approach for open domain dialogue generation. The model effectively extracts relevant visual commonsense and integrates the multimodal knowledge, and adaptively measures the contribution of different modalities, so as to produce better responses. Extensive experiments on two pure language dialogue datasets show that the proposed VICTOR model significantly outperforms previous approaches, indicating that VICTOR can generate more diverse and informative responses, while maintaining coherence with the context. For future work, we will continue to investigate the advantages of introducing external visual knowledge into the dialogue system. We notice that the current use of visual modality in this field may be too rough. Further study should focus on how to extract more specific and more necessary information from images or videos for enhancing response generation. Besides, enabling dialogue agents to handle multimodal inputs and outputs is also a relevant hot research problem.

Limitations

We discussed the limitations of this work. The proposed method trains the visual retriever and the generation model separately, which may affect the overall optimization of the system to a certain extent. Besides, limited by the performance of the retrieval modules, the extracted visual commonsense is not always an effective extension of the context content. This indicates that there is still room for improvement in the acquisition and utilization of high-quality visual knowledge for dialogue generation.

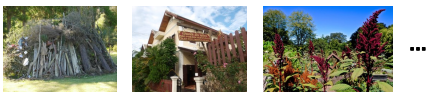
User A: I enjoy going on the sand. **User B**: _____ I love books.	**User A:** what sort of things do you grow in your garden?
Original concepts: enjoy, sand, book	**Original concepts**: garden, grow
Extended concepts: enjoy, sand, book, sun, beach, swim, read ...	**Extended concepts**: garden, grow, plant, flower, house, tree, ground ...
Pairs: enjoy-happiness, sand-beach, book-read, beach-swim ...	**Pairs**: garden-flower, garden-plant, garden-house, grow-tree ...
Images: 	**Images:** 
Captions: a group of people sitting on top of a beach. a wooden boat floating on top of water. a picture of a man reading a book. ...	**Captions**: a bunch of trees in a park. a building with a plant growing out of it. a tree with flowers in it. ...
Responses: **Ground truth**: I love reading books while sunbathing in the sand. **GRF**: I love going on the sand with my family. **GVT**: I really enjoy spending time with my family. I love to play tennis. **AdaLabel**: I have been into my parents. **VICTOR**: I enjoy going to the beach. I love to read when it is sunny outside.	**Responses:** **Ground truth**: I grow a variety of things, so that something is in bloom all year round. **GRF**: Well, I grow lots of plants. **GVT**: Quite a number. I don't think that we need anything else. **AdaLabel**: I grow up something in bloom all day. I like that. **VICTOR**: I grow all kinds of flower and there are many trees next to my house.

Figure 4: Two cases from OTTers and DailyDialog datasets respectively.

Ethics Statement

We honor and support the ACL code of Ethics. Dialogue response generation aims to build a dialogue system which better interacts with users. The generation of the responses does not involve any bias towards the participants. All datasets used in this work are from previously published works, and in our view, do not have any attached privacy or ethical issues.

Acknowledgements

This work was supported in part by the National Key R&D Program of China under Grant 2018AAA0102000, and in part by the National Natural Science Foundation of China: U21B2038, 61931008, 61732007, and CAAI-Huawei MindSpore Open Fund, Youth Innovation Promotion Association of CAS under Grant 2020108, CCF-Baidu Open Fund.

References

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, et al. 2020. Experience grounds language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735.

Ozan Caglayan, Pranava Swaroop Madhyastha, Lucia Specia, and Loïc Barrault. 2019. Probing the need for visual context in multimodal machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4159–4170.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Qingkai Fang and Yang Feng. 2022. Neural machine translation with phrase-level universal visual representations. *arXiv preprint arXiv:2203.10299*.

Steven Y Feng, Kevin Lu, Zhuofu Tao, Malihe Alikhani, Teruko Mitamura, Eduard Hovy, and Varun Gangal. 2021. Retrieve, caption, generate: Visual grounding for enhancing commonsense in text generation models. *arXiv preprint arXiv:2109.03892*.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Stevan Harnad. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346.

Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. 2020. Challenges in building intelligent open-domain dialog systems. *ACM Transactions on Information Systems (TOIS)*, 38(3):1–32.

Haozhe Ji, Pei Ke, Shaohan Huang, Furu Wei, Xiaoyan Zhu, and Minlie Huang. 2020. Language generation with multi-hop reasoning on commonsense knowledge graph. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 725–736.

Woojeong Jin, Dong-Ho Lee, Chenguang Zhu, Jay Pujara, and Xiang Ren. 2022. Leveraging visual knowledge in language tasks: An empirical study on intermediate pre-training for cross-modal knowledge

transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2750–2762.

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

Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Noriyuki Kojima, Hadar Averbuch-Elor, Alexander M Rush, and Yoav Artzi. 2020. What is learned in visually grounded neural syntax acquisition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2615–2635.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. 2020. The open images dataset v4. *International Journal of Computer Vision*, 128(7):1956–1981.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and William B Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119.

Liang Li, Xingyu Gao, Jincan Deng, Yunbin Tu, Zheng-Jun Zha, and Qingming Huang. 2022. Long short-term relation transformer with global gating for video captioning. *IEEE Transactions on Image Processing*, 31:2726–2738.

Liang Li, Shuhui Wang, Shuqiang Jiang, and Qingming Huang. 2018. Attentive recurrent neural network for weak-supervised multi-label image classification. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1092–1100.

Linxiao Li, Can Xu, Wei Wu, Yufan Zhao, Xueliang Zhao, and Chongyang Tao. 2020. Zero-resource knowledge-grounded dialogue generation. *Advances in Neural Information Processing Systems*, 33:8475–8485.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995.

Zujie Liang, Huang Hu, Can Xu, Chongyang Tao, Xiubo Geng, Yining Chen, Fan Liang, and Daxin Jiang. 2021. Maria: A visual experience powered conversational agent. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5596–5611.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Zhaojiang Lin, Genta Indra Winata, Peng Xu, Zihan Liu, and Pascale Fung. 2020. Variational transformers for diverse response generation. *arXiv preprint arXiv:2003.12738*.

Chia-Wei Liu, Ryan Lowe, Iulian Vlad Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132.

Xuejing Liu, Liang Li, Shuhui Wang, Zheng-Jun Zha, Zechao Li, Qi Tian, and Qingming Huang. 2022. Entity-enhanced adaptive reconstruction network for weakly supervised referring expression grounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421.

Bodhisattwa Prasad Majumder, Taylor Berg-Kirkpatrick, Julian McAuley, and Harsh Jhamtani. 2021. Unsupervised enrichment of persona-grounded dialog with background stories. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 585–592.

Bodhisattwa Prasad Majumder, Harsh Jhamtani, Taylor Berg-Kirkpatrick, and Julian McAuley. 2020. Like hiking? you probably enjoy nature: Persona-grounded dialog with commonsense expansions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9194–9206.

Pierre-Emmanuel Mazare, Samuel Humeau, Martin Raison, and Antoine Bordes. 2018. Training millions of personalized dialogue agents. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2775–2779.

Nasrin Mostafazadeh, Chris Brockett, William B Dolan, Michel Galley, Jianfeng Gao, Georgios Spithourakis, and Lucy Vanderwende. 2017. Image-grounded conversations: Multimodal context for natural question and response generation. In *Proceedings of*

the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 462–472.

Liqiang Nie, Wenjie Wang, Richang Hong, Meng Wang, and Qi Tian. 2019. Multimodal dialog system: Generating responses via adaptive decoders. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1098–1106.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083.

Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2015. Hierarchical neural network generative models for movie dialogues. *arXiv preprint arXiv:1507.04808*, 7(8):434–441.

Karin Sevegnani, David M Howcroft, Ioannis Konstas, and Verena Rieser. 2021. Otters: One-turn topic transitions for open-domain dialogue. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2492–2504.

Lei Shen and Yang Feng. 2020. Cdl: Curriculum dual learning for emotion-controllable response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 556–566.

Lei Shen, Haolan Zhan, Xin Shen, Yonghao Song, and Xiaofang Zhao. 2021. Text is not enough: Integrating visual impressions into open-domain dialogue generation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4287–4296.

Haoyue Shi, Jiayuan Mao, Kevin Gimpel, and Karen Livescu. 2019. Visually grounded neural syntax acquisition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1842–1861.

Kurt Shuster, Samuel Humeau, Antoine Bordes, and Jason Weston. 2020. Image-chat: Engaging grounded conversations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2414–2429.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. 2015. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28.

Zhenqiao Song, Xiaoqing Zheng, Lu Liu, Mu Xu, and Xuan-Jing Huang. 2019. Generating responses with a specific emotion in dialog. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3685–3695.

Robyn Speer, Catherine Havasi, et al. Representing general relational knowledge in conceptnet 5.

Hui Su, Xiaoyu Shen, Sanqiang Zhao, Zhou Xiao, Pengwei Hu, Randy Zhong, Cheng Niu, and Jie Zhou. 2020. Diversifying dialogue generation with non-conversational text. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7087–7097.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

Hao Tan and Mohit Bansal. 2020. Vokenization: Improving language understanding with contextualized, visual-grounded supervision. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2066–2080.

Yunbin Tu, Liang Li, Li Su, Shengxiang Gao, Chenggang Yan, Zheng-Jun Zha, Zhengtao Yu, and Qingming Huang. 2022. I2transformer: Intra-and inter-relation embedding transformer for tv show captioning. *IEEE Transactions on Image Processing*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Yida Wang, Yinhe Zheng, Yong Jiang, and Minlie Huang. 2021. Diversifying dialog generation via adaptive label smoothing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3507–3520.

Sixing Wu, Ying Li, Dawei Zhang, Yang Zhou, and Zhonghai Wu. 2021. Topicka: Generating commonsense knowledge-aware dialogue responses towards the recommended topic fact. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 3766–3772.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. 2017. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500.

Ze Yang, Wei Wu, Huang Hu, Can Xu, Wei Wang, and Zhoujun Li. 2021. Open domain dialogue generation with latent images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14239–14247.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213.

Yicheng Zou, Zhihua Liu, Xingwu Hu, and Qi Zhang. 2021. Thinking clearly, talking fast: Concept-guided non-autoregressive generation for open-domain dialogue systems. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2215–2226.