Maria: A Visual Experience Powered Conversational Agent

Zujie Liang[1*†] **Huang Hu**[2†] **Can Xu**[2] **Chongyang Tao**[2]
Xiubo Geng[2] **Yining Chen**[2] **Fan Liang**[1] **Daxin Jiang**[2‡]
[1]School of Electronics and Information Technology,
Sun Yat-sen University, Guangzhou, China
[2]Microsoft STCA NLP Group, Beijing, China
[1]{liangzj9@mail2.sysu.edu.cn, isslf@mail.sysu.edu.cn}
[2]{huahu,caxu,chotao,xigeng,yinichen,djiang}@microsoft.com

Abstract

Arguably, the visual perception of conversational agents to the physical world is a key way for them to exhibit the human-like intelligence. Image-grounded conversation is thus proposed to address this challenge. Existing works focus on exploring the multimodal dialog models that ground the conversation on a given image. In this paper, we take a step further to study image-grounded conversation under a fully open-ended setting where no paired dialog and image are assumed available. Specifically, we present Maria, a neural conversation agent powered by the visual world experiences which are retrieved from a large-scale image index. Maria consists of three flexible components, *i.e.*, text-to-image retriever, visual concept detector and visual-knowledge-grounded response generator. The retriever aims to retrieve a correlated image to the dialog from an image index, while the visual concept detector extracts rich visual knowledge from the image. Then, the response generator is grounded on the extracted visual knowledge and dialog context to generate the target response. Extensive experiments demonstrate Maria outperforms previous state-of-the-art methods on automatic metrics and human evaluation, and can generate informative responses that have some visual commonsense of the physical world.

1 Introduction

Building intelligent conversational agents that can not only converse freely with human but also have the ability to perceive the physical world, has been one of the longest standing goals of natural language processing (NLP) and artificial intelligence (AI). Although the recent large-scale conversation models trained on text-only corpora, such as Meena (Adiwardana et al., 2020), Blender (Roller et al., 2020) and DialoGPT (Zhang et al., 2020), have shown the compelling performance, they are still lack of the perception ability to our physical world. A recent study (Bisk et al., 2020) points out the successful linguistic communication relies on a shared experience of the world that makes language really meaningful. The visual perception is a rich signal for modeling a vastness of experiences in the world that cannot be documented by text alone (Harnad, 1990). On the other hand, human-human conversations involve their understandings of context, the background knowledge they had, and perhaps most importantly the experiences of the world they shared, *e.g.*, what they have seen before.

Figure 1: An example of human conversations. When human-B talks about vacation on the beach of Hawaii, human-A recalls his/her past experience of playing volleyball or having BBQ on the beach.

Figure 1 shows a conversation between humans. Human-A recalls his/her past experience of playing volleyball or having BBQ on the beach when human-B talks about vacation on the beach of Hawaii. However, the association relationship between beach and volleyball (or BBQ) is hard to capture in traditional knowledge bases, such as knowledge graph. Motivated by this, we select a common word "pizza" and collect the top 17 words that mostly co-occur with "pizza" on Google

*Work performed during the internship at Microsoft.
†Equal contribution.
‡Corresponding author.

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 5596–5611
August 1–6, 2021. ©2021 Association for Computational Linguistics

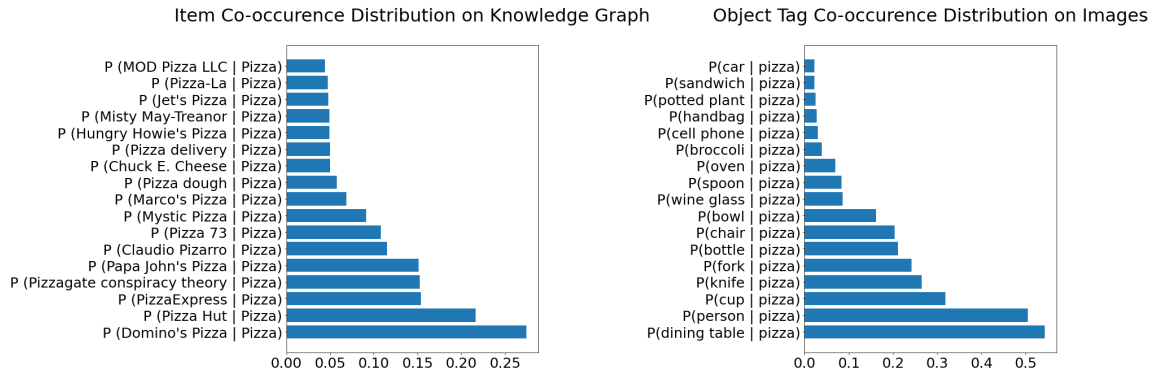


Figure 2: The word co-occurrence distribution with "pizza" on Google knowledge graph and MS-COCO images.

Knowledge Graph[1] and MS-COCO images[2] (Lin et al., 2014). As shown in Figure 2, the words co-occurring with "pizza" on knowledge graph tend to be the abstract concepts, while the co-occurrence relationship of object tags on images reflects some commonsense of our physical world, *e.g.*, "pizza" is usually on the "dining table", people usually use "knife" when eating "pizza". Interestingly, we found the "pizza" also co-occurs with "cell phone" and even "plotted plant". This indicates when people eat pizza, they sometimes would put their cell phones aside on the table, or there might exist some plotted plants in the restaurant. Thus, empowering conversational agents to have the visual perception ability about the physical world is a key way for them to exhibit the human-like intelligence.

The existing works (Mostafazadeh et al., 2017; Huber et al., 2018; Shuster et al., 2020) focus on exploring the multimodal dialog models that ground the conversation on a given image. Recently, Yang et al. (2020) propose to learn the dialog generation model with both image-grounded dialogs and textual dialogs by resorting to text-to-image synthesis techniques (Xu et al., 2018; Qiao et al., 2019) to restore a latent image for the text-only dialog. Even so, these works are still constrained by the assumption that the dialog is conducted center around a given (or synthesized) image.

In this paper, we take a step further to extend the assumption of image-grounded conversation to a fully open-ended setting where no image-dialog pairs are assumed available. Specifically, we present Maria, a neural conversational agent powered by visual world experiences which are retrieved from a pre-built image index, *e.g.*, the Open Images Dataset (Kuznetsova et al., 2018). Maria consists of three components: text-to-image retriever, visual concept detector, and visual-knowledge-grounded response generator. The retriever is responsible for retrieving a piece of visual world experiences, *e.g.*, a correlated image to the dialog from an image index. The visual concept detector utilizes the object detector from UpDown (Anderson et al., 2018) to extract the regions features (*i.e.*, bboxes) and the corresponding visual concepts (*i.e.*, tags) from the retrieval images. Hence, we can construct (*bboxes*, *tags*, *context*, *response*) 4-tuple as the training data. Finally, these constructed 4-tuples are used to train the visual-knowledge-grounded response generator, which is built on the top of a multi-layer Transformer architecture (Vaswani et al., 2017). To effectively inject the visual knowledge into the response generator, we carry out the Masked Concept Prediction and Visual Knowledge Bias besides the response generation objective. The former aims to align the semantic representations between textual words and image regions, while the latter tries to provide more visual knowledge to facilitate the dialog generation. The experimental results on Reddit Conversation Corpus (Dziri et al., 2019a) demonstrate that Maria significantly outperforms previous state-of-the-art methods, and can generate informative responses with visual commonsense of our physical world.

Overall, the contributions of this paper are summarized as follows:

- We explore the task of image-grounded dialog generation under a fully open-ended setting where no specific image-dialog pairs are assumed available, *i.e.,* zero-resource image-grounded conversation. To the best of our knowledge, this is the first work to connect dialog corpus with the unpaired image data;
- We present Maria, a neural conversational

[1] https://developers.google.com/knowledge-graph/

[2] We calculate the co-occurrence distribution of object tags from the images in MS-COCO dataset. More examples could be found in Appendices.

agent consisting of three flexible components, which can effectively capture the visual commonsense from images and accordingly generate informative and vivid responses;

- Extensive experiments on the widely used Reddit Conversation Corpus are conducted to justify the effectiveness of Maria.

2 Related Work

Vision and Language In the research of vision and language, various tasks have been extensively studied, such as image captioning (Vinyals et al., 2015; Lu et al., 2017; Hu et al., 2020), visual question answering (Antol et al., 2015; Anderson et al., 2018), visual dialog (Das et al., 2017a,b). Popular benchmark datasets in this area include MS-COCO (Lin et al., 2014), VisDial (Das et al., 2017a) and Visual Genome (Krishna et al., 2017). Visual dialog is a task to answer the questions about the factual content of the image in a multi-turn manner. Differently, image-grounded conversation studies how to reply to a dialog context and a given image with proper responses in an open-ended way.

Dialog Generation Encouraged by the success of the neural sequence-to-sequence architecture (Sutskever et al., 2014) on machine translation, end-to-end neural approaches on open-domain dialog generation (Vinyals and Le, 2015; Shang et al., 2015; Serban et al., 2016; Sordoni et al., 2015; Xing et al., 2017; Wu et al., 2018; Zhang et al., 2020; Xu et al., 2019; Adiwardana et al., 2020) have been widely studied in literature. Recently, there is an emerging trend towards grounding the dialog generation models on the external knowledge, such as knowledge graphs (Zhou et al., 2018), documents (Ghazvininejad et al., 2018; Dinan et al., 2019; Kim et al., 2020; Zhao et al., 2020a,b; Li et al., 2020) and images (Mostafazadeh et al., 2017; Shuster et al., 2020; Yang et al., 2020). Different from the previous work on knowledge-grounded conversation that connects dialogs with unpaired document knowledge (Li et al., 2020), our work lies in the research of image-grounded conversation where a response is generated with a dialog context and a given image. Existing works (Mostafazadeh et al., 2017; Shuster et al., 2020; Yang et al., 2020) in this direction assume there is a given (or synthesized) image for the dialog and explore the multimodal dialog models. In contrast to these works, we study the image-grounded conversation under a fully open-ended assumption where no paired dialog and image are assumed available, *i.e.*, zero-resource image-grounded conversation.

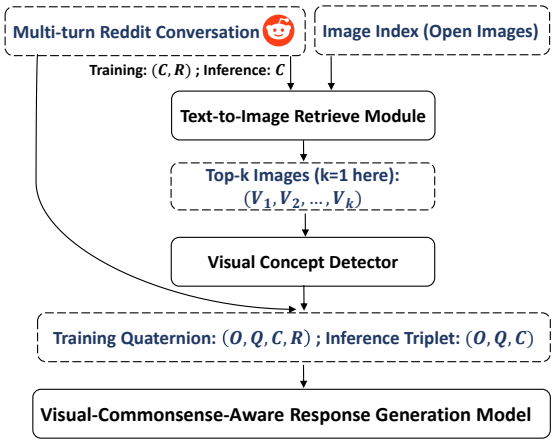


Figure 3: The flowchart of our framework. O, Q, C, R represents the image region features, extracted visual concepts, dialog context and response.

3 Problem Formalization

Suppose we have a dialog set $\mathcal{D} = \{(C_i, R_i)\}_{i=1}^{n}$, where $\forall i \in \{1, \ldots, n\}$, C_i refers to a dialog context and R_i is a response to C_i. We assume there is a set of images $\mathcal{V} = \{V_j\}_{j=1}^{m}$, where $\forall j \in \{1, \ldots, m\}$, V_j denotes an image. $\forall C \in \mathcal{D}$, we assume that there is an image V that triggered by the given dialog context C and response R. Our goal is to estimate a generation model $P(R|V, C)$ from $\mathcal{D}$ and $\mathcal{V}$. Thus, given a new dialog context C associated with an image V, the model can generate a response R according to $P(R|V, C)$.

4 Methodology

To learn such a generation model $P(R|V, C)$, we need to tackle several challenges: (1) How to bridge the gap between unpaired dialog corpus and image data; (2) After obtaining the correlated images, how to extract the detailed visual features and concepts; (3) How to effectively inject the visual knowledge into response generator and enable it to generate responses that are visual-knowledge-grounded. Figure 3 illustrates the framework of our approach. We first build a large-scale image dataset and leverage a cross-modal matching model to retrieve a correlated image using the content of the dialog. Then an off-the-shelf object detector is applied to extracting the object features and visual concepts from the retrieval image. Finally, the response generator is trained to generate the target response conditioned

on the context, extracted object features, and visual concepts. In the rest of this section, we will elaborate these three modules.

4.1 Text-to-Image Retriever

In this section, we develop a retrieval model that assigns each dialog with a correlated image V. Specifically, we train a text-to-image matching model from image captioning dataset and utilize it to construct the (C, R, V) triple data.

Modeling To improve the efficiency of cross-modal retrieval model on large-scale dialog corpus and image dataset, we adopt a two-tower architecture (Lu et al., 2019) to accelerate the retrieval process where the image features can be pre-extracted offline. The model takes a sentence T and an image V as input, and predicts the relevance score $s(T, V)$ between the sentence and the image. We use a text encoder and an image encoder to produce the representations of T and V, respectively. The text encoder is a pre-trained BERT-base model (Devlin et al., 2019) and we use the hidden state of special token `[CLS]` as the embedding of T:

$$\boldsymbol{e_t} = BERT(T) \tag{1}$$

Then a Multi-Layer Perceptron (MLP) projects the sentence embedding into the cross-modal space. We follow Tan and Bansal (2020) to perform L2-normalization on the last output features, by which we can simplify the nearest neighbor search problem in the euclidean space to the Maximum Inner Product problem (Mussmann and Ermon, 2016):

$$\boldsymbol{f_t}(T) = \frac{H_t(\boldsymbol{e_t})}{\|H_t(\boldsymbol{e_t})\|} \tag{2}$$

Similarly, the image encoder is composed of a pre-trained ResNeXt backbone (Xie et al., 2017) and a MLP with L2 normalization:

$$\boldsymbol{f_v}(V) = \frac{H_v(\boldsymbol{e_v})}{\|H_v(\boldsymbol{e_v})\|},\ \boldsymbol{e_v} = ResNeXt(V) \tag{3}$$

Thus, we define the relevance score $s(T, V)$ as an inner product of the language feature representation $f_t(T)$ and image feature representation $f_v(V)$:

$$s(T, V) = \boldsymbol{f_t}(T)^\top \boldsymbol{f_v}(V) \tag{4}$$

Training We train the cross-modal matching model on MS-COCO image captioning dataset (Lin et al., 2014), where each image is paired with 5 sentences describing its visual content. The model is optimized by minimizing the hinge loss so that the relevance score $s(T, V)$ of the positive image-sentence pair can be larger than the negative pair $s(T, V^-)$ by at least a margin M:

$$\mathcal{L}_{hinge}(T, V, V^-) = \sum_{i=1}^{l} \max\{0, M - s(T, V) + s(T, V^-)\} \tag{5}$$

Inference Given the trained retrieval model, we can now assign each dialog with a correlated image V. To ensure the diversity and richness of the retrieval results, we fetch 500,000 images from the large-scale Open Images dataset (Kuznetsova et al., 2018) as our image set $\mathcal{V}$. The image $V_i \in \mathcal{V}$ with the maximum relevance score is paired with the given dialog $(C_i, R_i) \in \mathcal{D}$. Note that for the dialog in the training set, we use both the context C and response R are concatenated as the query for retrieval (*i.e.*, $T = (C, R)$), which is beneficial to retrieving an image with the related visual knowledge. On the other hand, for the validation/test set of the dialog corpus, the query is only the context (*i.e.*, $T = C$) so as to keep consistent with the real-world setting where the response is unavailable and need to be generated at inference.

4.2 Visual Concept Detector

Given the correlated image V_i to the dialog as the visual clue, we can now extract the visual knowledge from it. One naive approach is to utilize the CNN-based models to extract the latent image features. However, this approach does not consider the fine-grained representation modeling for images, which is crucial for the dialog model to understand the local visual features in images. To address this issue, we adopt an object detection model (Anderson et al., 2018) pre-trained on Visual Genome (Krishna et al., 2017) to extract a set of salient object features $O = \{\mathbf{o}_k\}_{k=1}^K$, where each object feature $\mathbf{o}_k$ is a 2048-dimensional vector. These features represent the images at the level of objects and other salient regions, which has proven to be vital in many high-level image understanding tasks. Besides, the same detector is used to extract a set of visual concepts $Q = \{q_m\}_{m=1}^K$, where each concept q_m is the high-precision textual label of the visual region, *e.g.*, "sunset", "melon", etc. In this manner, we simultaneously obtain the fine-grained image representations and the necessary visual concepts for the subsequent dialog generation.

4.3 Visual-Knowledge-Grounded Response Generator

In this section, we propose a unified architecture to effectively inject a set of region features and corresponding visual concepts into the response generation model. In following parts, we describe the model design and training objectives in detail.

4.3.1 Model Architecture

Figure 4 shows the architecture of our response generation model, which is a multi-layer transformer network for both bidirectional vision/context (O, Q, C) encoding, and unidirectional response R decoding, via the flexible self-attention masks inspired by (Dong et al., 2019).

4.3.2 Input Representation

For each token, the final input representation to the multi-layer transformer network is the element-wise summation of four kinds of embeddings, including token-level, turn-level, position-level, and segment-level. Then, we concatenate all the input representations to one sequence for model training.

Token-Level The token-level embeddings are the concatenation of (O_w, Q_w, C_w, R_w), which denote the token embedding sequence of visual objects, visual concepts, contexts and response respectively. Note that O_w is the object embedding transformed by a linear layer into the same dimension as word embedding.

Turn-Level Since the dialog is multi-turn, we encode this turn order with a relative turn embedding (Bao et al., 2020). Specifically, the turn number is counted from the last utterance of the dialogue to the beginning. Note that as for the tokens corresponding to O and Q, we simply set them the same as the first utterance of C.

Position-Level Positional embedding encodes the signal of the token order in the total input sequence, which is the same as positional encoding of the original transformer (Vaswani et al., 2017).

Segment-Level Segment embedding is employed to differentiate which segment the token is in, *i.e.*, O, Q, C or R.

4.3.3 Masked Concept Prediction

Due to the inherent gap between visual modality and textual modality, directly optimizing the model by response generation objective may result in the insufficient utilization of the visual knowledge. To align the semantic representations of two modalities, we devise Masked Concept Prediction (MCP) objective. 15% of the visual concepts are randomly replaced with `[MASK]` tokens in each training instance, which need to be predicted by the model. However, one problem still remains, *i.e.*, the visual concepts have no specific order when extracting from images. In other words, we need to model MCP as a matching problem of set, which does not need to consider the order of predicted concepts when there are more than two concepts masked out simultaneously. To tackle this, inspired by Hu et al. (2020), we adopt the Hungarian Matching Loss (Stewart et al., 2016; Carion et al., 2020) to estimate an optimal mapping α so that the prediction for each masked position is assigned one of the target concepts. Here we denote the set of all input as $X = (O, Q, C, R)$, the set of the bidirectional self-attention part of X as $B = (O, Q, C)$, the set of masked concepts as $\hat{Q}$, the set of unmasked tokens as $B \backslash \hat{Q}$, and the prediction probabilities of the corresponding representations in the final layer of transformer as $H = \{h_i\}_{i=1}^{m}$ where h_i is the probability distribution of the i-th masked position. Hence, the MCP loss can be defined as:

$$\mathcal{L}_{\mathrm{MCP}}(\hat{Q}, H, \alpha) = - \sum_{q_{\alpha(i)} \in \hat{Q}} \log h_i \left(q_{\alpha(i)} \mid B \backslash \hat{Q} \right) \quad (6)$$

where $\alpha(i)$ is the index of the target concept assigned to the i-th prediction. When predicting a masked concept, the model will have to resort to visual region features, dialog contexts and other unmasked visual concepts. This would help the model to align the cross-modal representations between text and visual regions.

4.3.4 Masked Response Prediction

Encouraged by the success of UniLM (Dong et al., 2019) in Seq2Seq tasks, we adopt the Masked Response Prediction (MRP) objective to model the response generation. During training, 70% of the tokens in R are randomly masked with the special token `[MASK]`. The model is optimized to recover the masked tokens. The masked response tokens and other unmasked tokens in the whole input sequence can be denoted as $\hat{R}$ and $X \backslash \hat{R}$, respectively. Suppose that p_i is the conditional probability distribution of the i-th token in R, the MRP loss is the Negative Log-Likelihood (NLL) of the masked

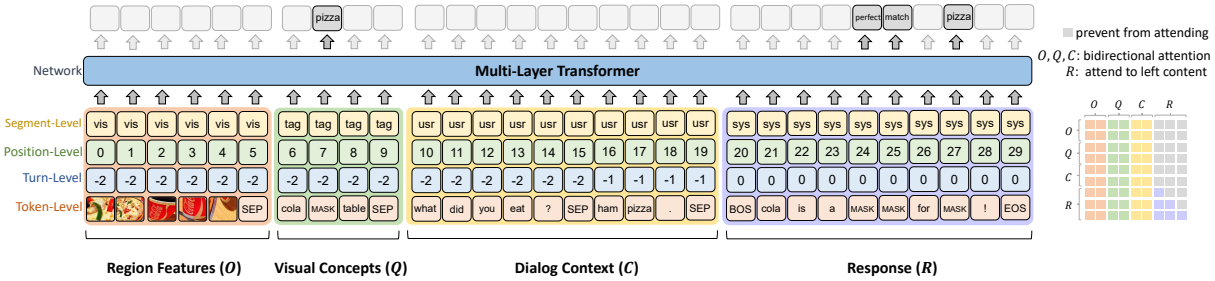


Figure 4: The overview of the response generation model. There are four kinds of inputs, *i.e.*, image region features O, extracted visual concepts Q, dialog context C and response R. The self-attention mask in R is unidirectional, *i.e.*, can only attend to the left context, while the self-attention mask in other segments is bidirectional.

response tokens as follow:

$$\mathcal{L}_{\text{MRP}}(X, \hat{R}) = -\sum_{w_i \in \hat{R}} \log p_i \left(w_i \mid X \backslash \hat{R} \right) \quad (7)$$

Note that the self-attention mask in R is left-to-right, but the rest are bidirectional. In other words, the tokens in O, Q and C can attend to each other from both directions, while the tokens in R can attend all tokens in O, Q, C and the leftward tokens in R including itself. MRP implicitly encourages the model to generate responses by learning the relationship among all input tokens.

For decoding, we first encode the image regions, visual concepts, dialog contexts, and a special token `[BOS]` as input. Then the model starts the generation by feeding a `[MASK]` token and samples a word from the predicted distribution over vocabulary. Then, the `[MASK]` token is replaced by the generated token and a new `[MASK]` is appended to the input sequence for next word prediction. The generation process terminates when the model predicts `[EOS]` token or reaches the pre-defined maximum length.

Visual Knowledge Bias Normally, the top projection layer of generation model produces a probability distribution over the vocabulary:

$$\boldsymbol{p} = softmax(We^r + b), \quad (8)$$

where the $e^r \in \mathbb{R}^d$, $W \in \mathbb{R}^{|V| \times d}$ and $b \in \mathbb{R}^{|V|}$ are the last output of the transformer network, weight and bias parameters of the decoding head, respectively. $|V|$ denotes the vocabulary size. So far, the visual world knowledge is introduced into the response generation model by the shared-parameter self-attention layers. To further inject the visual knowledge into the generation model, we design a simple but effective strategy, namely Visual Knowledge Bias (VKB). Concretely, an additional visual vocabulary bias b_q is first calculated as follow:

$$b_q = F_q(e^q_{avg}) \quad (9)$$

where $F_q : \mathbb{R}^d \rightarrow \mathbb{R}^{|V|}$ is a projection layer. e^q_{avg} denotes the average pooling on all hidden representations of visual concepts, *i.e.*, $e^q_{avg} = AvgPooling(E^q)$ where $E^q = (e^q_1, ..., e^q_K)$. Then, we mask non-visual-concept tokens in the vocabulary and the masked vocabulary bias $\hat{b_q} \in \mathbb{R}^{|V|}$ is added to the top layer of generation model to get the final distribution over vocabulary:

$$\hat{\boldsymbol{p}} = softmax(We^r + b + \hat{b_q}) \quad (10)$$

We leverage this final vocabulary distribution to calculate the MRP loss in Eq. 7 to optimize the model. This visual knowledge bias would encourage the model to generate more visual knowledge related tokens in the response.

To sum up, the final objective of our response generation model is to minimize the integrated loss:

$$\mathcal{L} = \mathcal{L}_{\text{MRP}} + \mathcal{L}_{\text{MCP}} \quad (11)$$

5 Experimental Setup

5.1 Datasets

To evaluate the performance of Maria, we conduct comprehensive experiments on the Reddit dataset released by Yang et al. (2020), which is a large-scale and high-quality multi-turn conversations extracted from Reddit Conversation Corpus (Dziri et al., 2019b). Each dialog has 3 to 5 utterances, and the training/validation/test set has 1M/20K/20K dialogs respectively.

We train and validate the retrieval model using the Karpathy's split[3] of the MS-COCO image captioning data, where the images are split into

[3] https://cs.stanford.edu/people/karpathy/deepimagesent

113.2K/5K/5K samples as training/validation/test set, respectively. After the retrieval model is trained, we fetch 500K images from the Open Images dataset as the image index, and then retrieve images from it by dialog context and response to construct the training data for response generator.

5.2 Evaluation Metrics

Both automatic metrics and human evaluation are employed to assess the performance of Maria and baselines. Automatic metrics include: (1) **Fluency**: perplexity (PPL) measures the confidence of the generated responses; (2) **Relevance**: BLEU-1 (Papineni et al., 2002), Rouge-L (Lin, 2004), and we follow Serban et al. (2017) to utilize Embedding Average cosine similarity, Vector Extrema cosine similarity, and Embedding Greedy Matching score. All this metrics are calculated by running the public NLG evaluation script[4]; (3) **Diversity**: Distinct-1 (Dist-1) and Distinct-2 (Dist-2) (Li et al., 2016) are defined as the number of distinct uni-grams or bi-grams divided by the total amount of words.

In human evaluation, we randomly select 100 dialogue contexts and the corresponding generated responses for Maria and compared baselines. Three human annotators are asked to score the response quality on a scale of $\{0, 1, 2\}$ from three aspects, including **Fluency**, **Relevance** and **Richness**. The higher score means the better. Since each response receives 3 scores on each aspect, we report the average scores over annotators and responses. The inter-annotator agreement is measured by Fleiss' Kappa(Fleiss and Cohen, 1973).

5.3 Implementation Details

For the retrieval model, ResNeXt-101-32x8d feature is used as the visual embedding, while the concatenation of the last 4 layers of BERT's outputs is used as the textual embedding. Both embeddings are then respectively fed into an MLP composed of three layers of size (1024, 1024, 512). When training the retrieval model, we set the margin $M = 0.5$ for the hinge loss, and only tune the parameters of both MLPs while freezing the parameters of ResNeXt and BERT. The total training epoch is 20. At inference, the FAISS (Johnson et al., 2019) library is utilized to accelerate the inner product search by batch processing. We use the off-the-shelf object detector from UpDown (Anderson et al., 2018) to extract top-k (k=36) image region features and the corresponding visual concepts. The detector is a Faster R-CNN (Ren et al., 2015) model trained on the Visual Genome dataset (Krishna et al., 2017).

For the response generation model, we set the number of transformer layers $L = 12$ and the hidden embedding dimension $D = 768$. Besides, the network parameters are initialized by UniLM. The maximum sequence lengths of context and response are set to 110 and 40, respectively. The sequence lengths of region features and concept tokens are both set to 36. The batch size is 64. We use the Adam Optimizer (Kingma and Ba, 2015) with a learning rate 3e-5 to train the response generation model. The training is conducted on 4 Nvidia Tesla P40 24G GPU cards for 20 epochs.

5.4 Baselines

We compare the following baselines in the experiments: (1) **Seq2Seq**: A standard Sequence to Seqence model with attention mechanism (Bahdanau et al., 2015). (2) **HRED**: A Hierarchical Recurrent Encoder-Decoder neural network (Serban et al., 2016). (3) **VHRED**: A variation of HRED that introduces latent variables into the generation (Serban et al., 2017). (4) **ReCoSa**: A hierarchical transformer-based model (Zhang et al., 2019) that achieves the state-of-the-art performance on benchmarks of dialog generation. (5) **ImgVAE**: A dialog generation model (Yang et al., 2020) that is trained on both textual dialogs and image-grounded dialogs by recovering a latent image behind the textual dialog within a conditional variational auto-encoding framework. (6) **DialoGPT**: An open-domain dialog model (Zhang et al., 2020) that fine-tunes GPT-2 (Radford et al., 2019) on massive Reddit data. Since DialoGPT is a dialog generation model trained on the text-only corpus, we introduce it as an auxiliary baseline. For a fair comparison, we choose the same model size (L=12,D=768) of DialoGPT (117M) as our model.

6 Experimental Results

6.1 Automatic and Human Evaluations

We summarize the experimental results of automatic evaluations in Table 1. Maria achieves the substantial performance improvements over baselines on all metrics except for the comparison to DialoGPT. Especially, Maria significantly surpasses ImgVAE on Dist-1/2, which indicates introducing richer visual knowledge, *i.e.*, image region features

[4] https://github.com/Maluuba/nlg-eval

Model	PPL	BLEU-1	Rouge-L	Average	Extrema	Greedy	Dist-1	Dist-2
Seq2Seq (Bahdanau et al., 2015)	77.27	12.21	10.81	78.38	40.06	62.64	0.53	1.96
HRED (Serban et al., 2016)	84.02	11.68	11.29	75.54	37.49	60.41	0.89	3.21
VHRED (Serban et al., 2017)	78.01	12.22	11.82	75.57	39.24	62.07	0.87	3.49
ReCoSa (Zhang et al., 2019)	71.75	12.75	11.75	79.84	42.29	63.02	0.66	3.83
ImgVAE (Yang et al., 2020)	72.06	12.58	12.05	79.95	42.38	63.55	1.52	6.34
DialoGPT (Zhang et al., 2020)	**36.03**	5.87	5.20	77.80	35.40	58.39	**10.41**	**49.86**
Maria	54.38	**14.21**	**13.02**	**82.54**	**44.14**	**65.98**	8.44	33.35
Maria (*w/o* MCP)	66.71	13.91	11.60	81.59	41.06	64.10	8.36	31.80
Maria (*w/o* VKB)	65.51	12.76	11.76	82.49	40.22	64.49	7.15	29.44
Maria (*w/o* VKB & MCP)	62.64	11.50	10.45	77.52	41.27	61.00	6.92	28.53
Maria (*w/o* images)	64.75	10.70	9.15	78.89	39.88	62.39	6.88	28.01
Maria (*w/o* concepts)	69.24	11.43	10.61	**82.96**	41.02	65.07	4.56	16.44
Maria (*w/o* images & concepts)	69.50	10.75	8.34	80.62	41.15	64.25	3.69	10.11

Table 1: Evaluation results of generated responses on the test set. Numbers in bold denote that the improvement over the best performing baseline is statistically significant. Numbers with underline refer to the best results except for the comparison to DialoGPT (Zhang et al., 2020).

Model	Fulency	Relevance	Richness	Kappa
ImgVAE	1.79	0.58	0.67	0.67
DialoGPT	**1.93**	0.92	**1.20**	0.59
Maria	1.89	**1.06**	0.97	0.62

Table 2: Human evaluation results.

and the corresponding visual concepts, is beneficial to generating more diverse and informative responses. This also reflects in human evaluation of Table 2 that the richness score of Maria is higher than that of ImgVAE. Besides, in terms of relevance metrics including BLEU-1, Rouge-L, Average, Extrema and Greedy, Maria outperforms all baselines and even performs better than DialoGPT. This indicates introducing the extra visual knowledge related to dialog context can further force the model to produce more relevant responses.

On the other hand, the discrepancy of data distributions between the training data (*i.e.*, ImageChat (Shuster et al., 2020) dataset) and test data (*i.e.*, Reddit conversation dataset) of the text-to-image synthesis model in ImgVAE limits its performance in practice. Besides, constrained by the capability of the text-to-image synthesis model, the richness and diversity of the synthesized images are undesirable, while Maria can retrieve a variety of images from the large-scale image index. That may be the reason why ImgVAE consistently underperforms our Maria on relevance including automatic evaluation and human judgement, which also shows the superiority of the retrieval method for the zero-resource image-grounded conversation. Another observation is that Maria slightly underperforms DialoGPT on PPL and Dist-1/2. Since DialoGPT is a large-scale pre-training based dialog generation model and introduces the extra mutual information maximization objective to improve the informativeness of generated responses, which is consistent in human evaluation with respect to fluency and richness.

6.2 Ablation Study

We conduct extensive ablation experiments over different model variants and input components to better understand their relative importance to the dialog generation task. As shown in Table 1, training the simplified versions of Maria or removing any visual signals from input components leads to worse performance in terms of relevance and diversity. In particular, the results on the ablation study validate that: (1) The performance improvement of dialog generation benefits from the MCP's effectiveness in aligning the representations of text and vision; (2) When training Maria, introducing VKB can further improve the quality and diversity of generated responses; (3) Rich visual knowledge, *i.e.*, image region features and visual concepts, play a significant role in improving the performance of dialog generation. Especially, removing the visual concepts leads to a dramatic performance drop on diversity. The phenomenon is due to the lack of necessary visual concepts, Maria can not well understand the visual world knowledge when only learning from the visual features.

6.3 Case Analysis

To further investigate the quality of responses generated by Maria, we put an example of generated responses in Figure 5. As we can see from Figure 5, when the context talks about the supermarket "Aldi", Maria can retrieve a "pizza" related image and generate the informative response grounded on

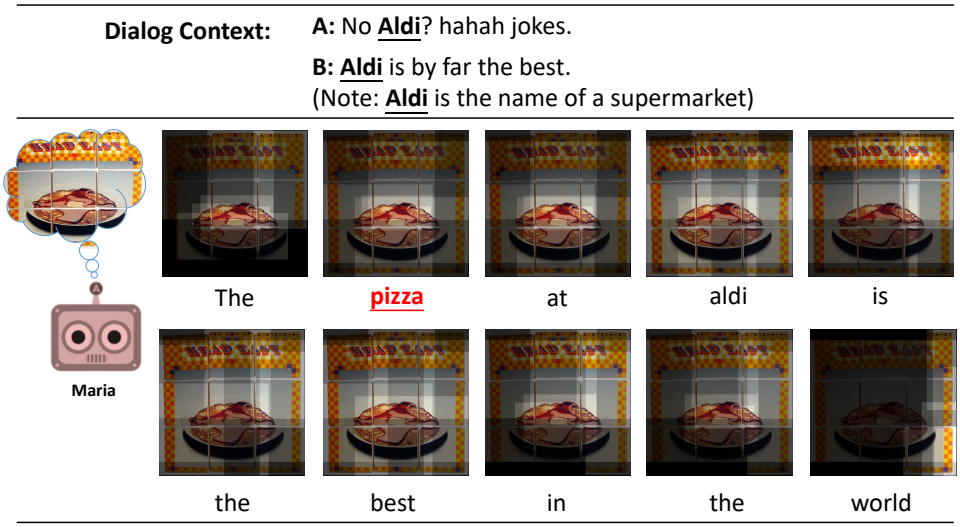


Figure 5: The visualization of attention weights on the retrieved image by Maria for an example.

it, *i.e.*, "the pizza at Aldi is the best in the world". This implies the commonsense that the supermarket usually has the pizza to sell. It is also observed that Maria pays more attention to the relevant image regions when generating the word "pizza", which demonstrates that Maria could capture useful visual knowledge from the image and subsequently leverage it to generate commonsense-aware responses. More cases are demonstrated in Appendices.

7 Conclusions

In this paper, we present Maria, a neural conversational agent powered by the visual world experiences. It is able to retrieve the visual world experiences with users and generate human-like responses with some visual commonsense. Extensive experiments demonstrate Maria achieves substantial improvements over the state-of-the-art methods in automatic and human evaluation. The future works could include: (1) Design a more precise and comprehensive image retriever to include multiple retrieval images; (2) Combining the retrieve module and dialog generation into an end-to-end model, instead of learning them individually; (3) Explore more efficient neural architectures to inject the visual knowledge into response generation.

References

Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 6077–6086. IEEE Computer Society.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: visual question answering. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2425–2433. IEEE Computer Society.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Siqi Bao, Huang He, Fan Wang, Hua Wu, and Haifeng Wang. 2020. PLATO: Pre-trained dialogue generation model with discrete latent variable. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 85–96, Online. Association for Computational Linguistics.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020. Experience grounds language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735, Online. Association for Computational Linguistics.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer.

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

Abhishek Das, Satwik Kottur, José M. F. Moura, Stefan Lee, and Dhruv Batra. 2017b. Learning cooperative visual dialog agents with deep reinforcement learning. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2970–2979. IEEE Computer Society.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13042–13054.

Nouha Dziri, Ehsan Kamalloo, Kory Mathewson, and Osmar Zaiane. 2019a. Augmenting neural response generation with context-aware topical attention. In *Proceedings of the First Workshop on NLP for Conversational AI*, pages 18–31, Florence, Italy. Association for Computational Linguistics.

Nouha Dziri, Ehsan Kamalloo, Kory Mathewson, and Osmar Zaiane. 2019b. Augmenting neural response generation with context-aware topical attention. In *Proceedings of the First Workshop on NLP for Conversational AI*, pages 18–31, Florence, Italy. Association for Computational Linguistics.

Joseph L Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 33(3):613–619.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2018. A knowledge-grounded neural conversation model. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5110–5117. AAAI Press.

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

Xiaowei Hu, Xi Yin, Kevin Lin, Lijuan Wang, Lei Zhang, Jianfeng Gao, and Zicheng Liu. 2020. Vivo: Surpassing human performance in novel object captioning with visual vocabulary pre-training. *arXiv preprint arXiv:2009.13682*.

Bernd Huber, Daniel J. McDuff, Chris Brockett, Michel Galley, and Bill Dolan. 2018. Emotional dialogue generation using image-grounded language models. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI 2018, Montreal, QC, Canada, April 21-26, 2018*, page 277. ACM.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*.

Byeongchang Kim, Jaewoo Ahn, and Gunhee Kim. 2020. Sequential latent knowledge selection for knowledge-grounded dialogue. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

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

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. 2018. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *arXiv preprint arXiv:1811.00982*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill 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, San Diego, California. Association for Computational Linguistics.

Linxiao Li, Can Xu, Wei Wu, Yufan Zhao, Xueliang Zhao, and Chongyang Tao. 2020. Zero-resource knowledge-grounded dialogue generation. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

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.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13–23.

Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 3242–3250. IEEE Computer Society.

Nasrin Mostafazadeh, Chris Brockett, Bill 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, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Stephen Mussmann and Stefano Ermon. 2016. Learning and inference via maximum inner product search. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 2587–2596. JMLR.org.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. 2019. Mirrorgan: Learning text-to-image generation by redescription. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 1505–1514. Computer Vision Foundation / IEEE.

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

Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 91–99.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M Smith, et al. 2020. Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637*.

Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C. Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pages 3776–3784. AAAI Press.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron C. Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 3295–3301. AAAI Press.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1577–1586, Beijing, China. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. In *Proceedings*

of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 196–205, Denver, Colorado. Association for Computational Linguistics.

Russell Stewart, Mykhaylo Andriluka, and Andrew Y. Ng. 2016. End-to-end people detection in crowded scenes. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2325–2333. IEEE Computer Society.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 3156–3164. IEEE Computer Society.

Yu Wu, Wei Wu, Dejian Yang, Can Xu, and Zhoujun Li. 2018. Neural response generation with dynamic vocabularies. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5594–5601. AAAI Press.

Saining Xie, Ross B. Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. 2017. Aggregated residual transformations for deep neural networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 5987–5995. IEEE Computer Society.

Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. 2017. Topic aware neural response generation. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 3351–3357. AAAI Press.

Can Xu, Wei Wu, Chongyang Tao, Huang Hu, Matt Schuerman, and Ying Wang. 2019. Neural response generation with meta-words. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5416–5426, Florence, Italy. Association for Computational Linguistics.

Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. 2018. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 1316–1324. IEEE Computer Society.

Ze Yang, Wei Wu, Huang Hu, Can Xu, and Zhoujun Li. 2020. Open domain dialogue generation with latent images. *arXiv preprint arXiv:2004.01981*.

Hainan Zhang, Yanyan Lan, Liang Pang, Jiafeng Guo, and Xueqi Cheng. 2019. ReCoSa: Detecting the relevant contexts with self-attention for multi-turn dialogue generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3721–3730, Florence, Italy. Association for Computational Linguistics.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

Xueliang Zhao, Wei Wu, Chongyang Tao, Can Xu, Dongyan Zhao, and Rui Yan. 2020a. Low-resource knowledge-grounded dialogue generation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Xueliang Zhao, Wei Wu, Can Xu, Chongyang Tao, Dongyan Zhao, and Rui Yan. 2020b. Knowledge-grounded dialogue generation with pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3377–3390, Online. Association for Computational Linguistics.

Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Commonsense knowledge aware conversation generation with graph attention. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 4623–4629. ijcai.org.

A Appendices

In this section, we show more examples of word co-occurrence distributions on Google knowledge graph and MS-COCO images. Besides, some conversation samples produced by Maria and the baselines are also presented in Section A.2.

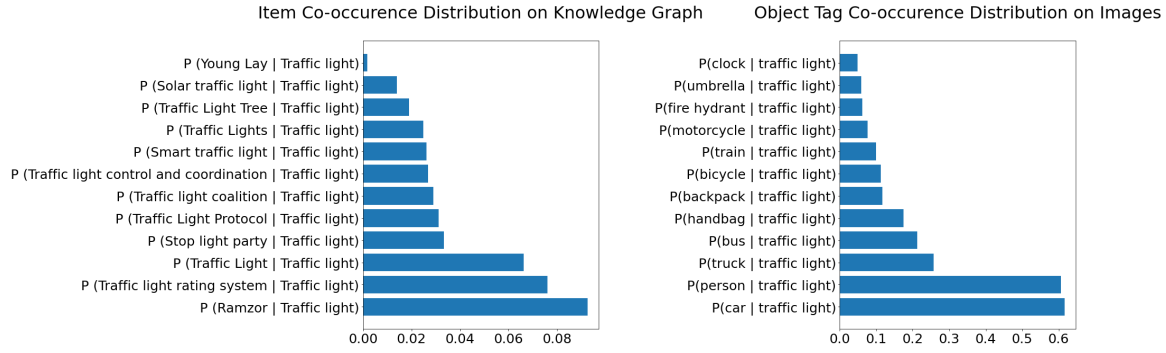
A.1 Word Co-occurrence Distribution Examples

In Figure 6, we present some supplementary examples of the word co-occurrence distribution on Google knowledge graph and MS-COCO images, including "traffic light", "bed", "book", and "pot plant". Figure 6 (a) shows the co-occurrence distributions of "traffic light" and other words on knowledge graph and images, respectively. As we can see, most of the co-occurred words with "traffic light" are the related concepts such as "smart traffic light", "traffic light protocol", "traffic light rating system", etc. While the co-occurred words on images are usually "car", "person", "truck", "bus", etc, which we often see when walking by the traffic lights. Interestingly, we found "umbrella" and "clock" also co-occurs with "traffic light" in some images. For the former, the picture we can imagine is that people were holding the "umbrellas" when they walked through a zebra crossing under the "traffic light". For the latter, the possible picture is that we can see both the "traffic light" and the "clock" on the top of a high building from a certain angle when walking on the street. Similar observations can be also seen in other examples.

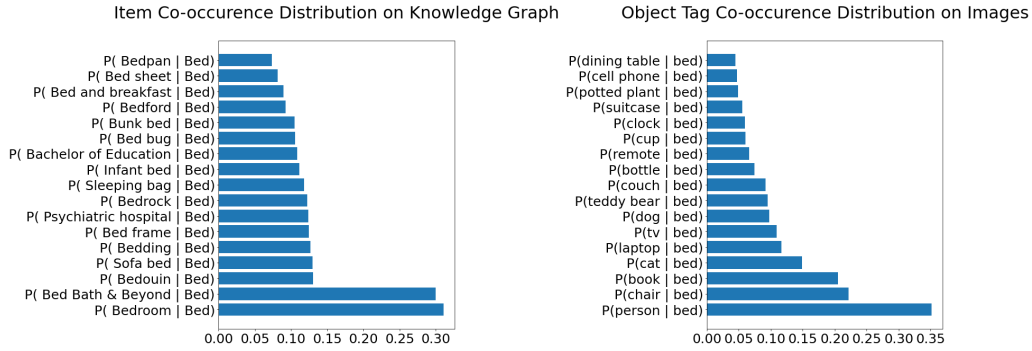
Most of the co-occurrence words on knowledge graph are logically-related concepts. However, the co-occurrence relationship of object tags on images reflects some commonsense of our physical world, which implies some pictures that we human could easily imagine. This kind of knowledge is unique and inherent in images, but it can hardly be captured in the traditional knowledge bases, such as knowledge graph.

A.2 Case Analysis

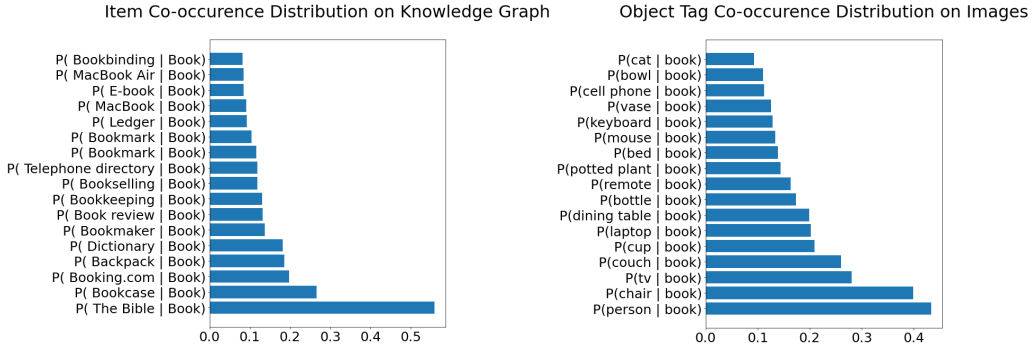
Figure 7 shows some cases from the test set of Reddit data. We observe that the responses generated by Maria are more commonsensical and vivid than those of the baseline methods, which is consistent with our automatic and human evaluation results. Interestingly, Maria is able to retrieve correlated images using the dialog contexts, which makes its response more human-like. For instance, case (a) shows that when the dialog context marvels at "the pass of the world cup", Maria recalls a football player and compliments him "the best player in the world"; case (b) shows that when the dialog context chats about the "Canada weather", Maria is aware of the fact that "Canada" is often "snowy" and then talks about "Canada" in a funny tone, "I've never been to a place that doesn't have snow"; case (c) shows that Maria understands that "swan" is sometimes "dangerous" when they are on the "beach"; case (d) shows that when the dialog context tries to guess one type of game, Maria recalls a ping-pong "ball" game and describes it; and etc.



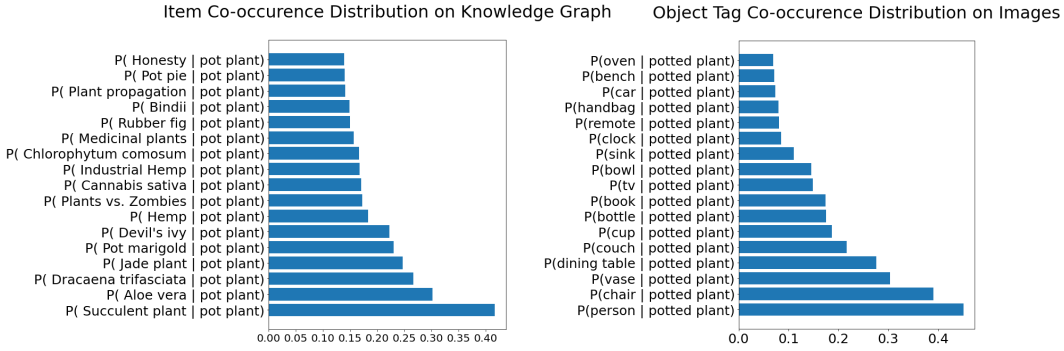
(a) traffic light



(b) bed



(c) book



(d) pot plant

Figure 6: Supplementary examples of the word co-occurrence distribution on Google knowledge graph and MS-COCO images.

Dialog Context:

A:the pass of the **world cup** (james rodriguez)

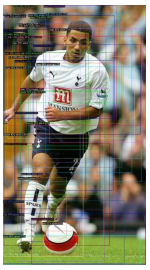
B: absoleutely gorgeous.

Baseline Response:

ImgVAE: I love the world cup.

DialoGPT: it 's not a pass of the world cup.

Maria



Detected Concepts: ball, socks, field, shirt, grass, man, shorts, ***player***, sock

Response: he's the best ***player*** in the world

(a)

Dialog Context:

A: you 've just ruined my canada weather fetish , that it would be ideal place for me to live: freezing cold winters in mild summers.

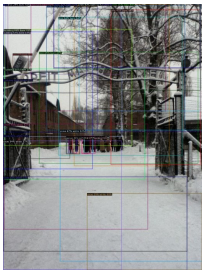
B: ya , there are parts of canada with very nice mild summers , but those same places have mild winters too.

Baseline Response:

ImgVAE: i 'm from canada , but i 've never seen a cold winters winters in cold winters winters .

DialoGPT: i 'm from the u.s. and i 've never heard of this.

Maria



Detected Concepts: person, fence, ground, tree, building, ***snow***, border, pole, gate

Response: i'm from canada and i've never been to a place that doesn't have ***snow***

(b)

Dialog Context:

A: how is a **swan** gonna break your arm?

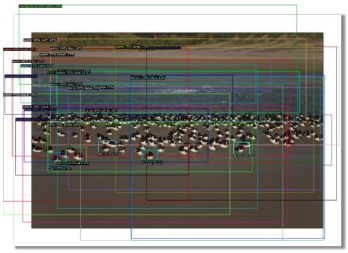
B: smack you with their wings

Baseline Response:

ImgVAE: i 'm a little nervous about this.

DialoGPT: i 'm a swan.

Maria



Detected Concepts:
duck, picture, background, land, ***beach***, border, line, ground, bird, sky, water, birds, frame

Response:
i'm not sure if you're joking or not, but i'm sure they're going to be a bit more dangerous if they are on the ***beach***

(c)

Dialog Context:

A: we all played this **game** at one point.

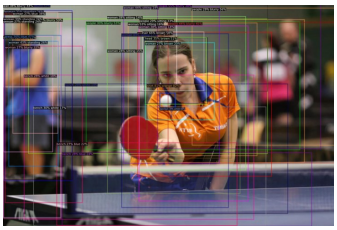
B: what game is this ?

Baseline Response:

ImgVAE: it 's a game game game , game of thrones.

DialoGPT: it 's a meme

Maria



Detected Concepts: shorts, background, heart, skateboard, woman, ***ball***, bench, counter, shirt, hair, wall, net, stands, people

Response: it's a game where you play the ***ball*** with your hands

(d)

Dialog Context:

A: what cool gadget under $200 can be life changing?

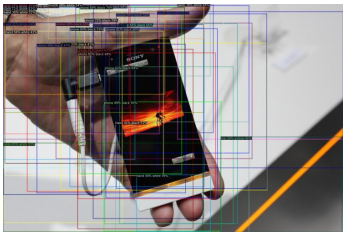
B: a smartphone.

Baseline Response:

ImgVAE: i have a smartphone and i 'm a phone.

DialoGPT: that 's not cool.

Maria



Detected Concepts: button, hand, wall, line, hair, ***phone***, table, fingers, handle, thumb

Response: i already have a ***phone***.

(e)

Dialog Context:

A: what 's the **coldest** you 've ever been?

B: almost got hypothermia near the arctic circle because my rain jacket failed to be waterproof and all my layers were soaked.

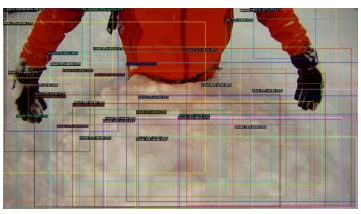
C: what kind of jacket?

Baseline Response:

ImgVAE: the one where you ski and the jacket?

DialoGPT: it 's a winter jacket.

Maria



Detected Concepts: pants, sky, ***snow***, person, clouds

Response: it's a ***snow*** jacket

(f)

Figure 7: Case Study on the Reddit data from test split.