Open Domain Dialogue Generation with Latent Images

Ze Yang [1], **Wei Wu** [2], **Huang Hu** [3], **Can Xu** [3], **Wei Wang** [4], **Zhoujun Li** [1*]

[1] State Key Lab of Software Development Environment, Beihang University, Beijing, China
[2] Meituan, Beijing, China
[3] Microsoft, Beijing, China
[4] China Resources Group, Shenzhen, China
{tobey, lizj}@buaa.edu.cn   {huahu, caxu}@microsoft.com
{wuwei19850318, ww.cs.tj}@gmail.com

Abstract

We consider grounding open domain dialogues with images. Existing work assumes that both an image and a textual context are available, but image-grounded dialogues by nature are more difficult to obtain than textual dialogues. Thus, we propose learning a response generation model with both image-grounded dialogues and textual dialogues by assuming that the visual scene information at the time of a conversation can be represented by an image, and trying to recover the latent images of the textual dialogues through text-to-image generation techniques. The likelihood of the two types of dialogues is then formulated by a response generator and an image reconstructor that are learned within a conditional variational auto-encoding framework. Empirical studies are conducted in both image-grounded conversation and text-based conversation. In the first scenario, image-grounded dialogues, especially under a low-resource setting, can be effectively augmented by textual dialogues with latent images; while in the second scenario, latent images can enrich the content of responses and at the same time keep them relevant to contexts.

Introduction

Open domain dialogue generation, due to the successful application in socialbots such as Microsoft XiaoIce (Shum, He, and Li 2018) and in virtual assistants such as Amazon Alexa (Ram et al. 2018), is emerging as a prominent research direction in conversational AI. Benefiting from the advances of neural sequence modeling (Sutskever, Vinyals, and Le 2014; Vaswani et al. 2017), existing work has achieved promising results on response quality (Zhang et al. 2019, 2018a; Xu et al. 2019), but often makes use of only textual contexts in response generation. Human conversations, on the other hand, could be grounded by more than one kind of perception. In addition to the context of conversation, people also respond according to the scene information including what they hear (e.g., voice or music) and what they see (e.g., images or videos). Hence, there is a clear trend in the research of open domain dialogues that text-based unimodal conversation is moving to perceptually-grounded multimodal conversation (Mostafazadeh et al. 2017; Chu, Li, and Fidler 2018; Le et al. 2019; Hori et al. 2019).
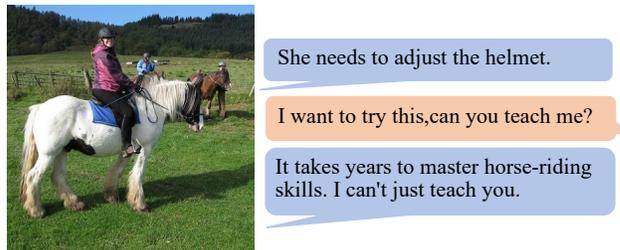
Figure 1: An example of image-grounded dialogue from Image-Chat.

We consider grounding open domain dialogues with their visual scene information which could be represented by images. Existing work (Shuster et al. 2020; Huber et al. 2018) formalizes the problem as response generation (or selection) with both a given image and several turns of conversation history as contexts, and focuses on benchmarking a combination of state-of-the-art neural structures in image modeling and dialogue modeling with either crowd-sourced data (Shuster et al. 2020) or selected data from social media (Huber et al. 2018). Figure 1 shows an example of image-grounded dialogues used in the existing work (Shuster et al. 2020). While these works provide test beds for future studies, the scale of data (e.g., a few hundred thousand triples) could hinder further progress, due to the expensive and exhausting nature of human effort and the fact that one has to abandon the large-scale textual dialogue data as their background images have not been explicitly recorded or they are naturally formed regardless of the visual scene. Motivated by this, we propose leveraging both multimodal data (i.e., image-context-response triples) and large scale of unimodal data (i.e., textual dialogues) for image-grounded response generation. The key assumption is that the visual background behind a textual conversation can be represented by a latent image, and we try to recover the latent image from text to integrate the textual dialogue data into image-grounded conversation. Advantages of our approach are two-fold: (1) for image-grounded conversation where an image is given first, textual dialogues with latent visual variables can augment the multimodal data, and help alleviate the data sparsity issue; and (2) for text-based conversation where only textual dialogues are available, the latent variables can pro-

vide visual signals for response generation, and help suppress "safe responses" (Li et al. 2015).

Our model consists of a response generator and an image reconstructor. The former synthesizes a response with both an image representation and a textual context representation as conditions and is shared when the image is explicitly given or latent variable; while the latter infers the latent image for a textual context. Challenges then include how to define the two models and how to effectively learn them from both multimodal and unimodal data. Encouraged by the recent progress on text-to-image synthesis, we define the image reconstructor within an attentional generative adversarial network framework (Xu et al. 2018; Qiao et al. 2019) where GAN-based image generation starts from a text representation and a random noise, and grows from a small scale to a large scale by letting sub-regions in each scale attend to words of the text. The response generator is defined within an encoder-decoder framework where attentive modules in the decoder involve both attention to the textual context and attention to sub-regions of the image. Considering that an inferred image could contain noise, and words in an open domain response may not relate to the image all the time, we design a gate in response decoding to control the contribution from the visual signals in prediction of each word. The two models are jointly learned from both multimodal data and unimodal data within a conditional variational auto-encoding (CVAE) framework, where through pre-training the image reconstructor on the multimodal data and fixing it in learning, we can circumvent the intractable KL term in the evidence lower bound and approximate the bound with the random noise in the image generation in a way similar to the reparameterization trick (Kingma and Welling 2013). By these means, the learning approach not only unifies dialogue generation and image generation, but also extends the commonly used CVAE from plain and uninterpretable variables to visually structured variables.

We test the proposed approach in both image-grounded conversation and text-based conversation. For the first scenario, we exploit the image-chat data published in (Shuster et al. 2020), and check if the model learned using both multimodal and unimodal data can improve upon the state-of-the-art model learned solely from the multimodal data, especially when the multimodal data is small in scale. For the second scenario, we leverage the Reddit Conversation Corpus published by (Dziri et al. 2018), and examine if latent images can provide useful signals for response generation. Evaluation results indicate that the proposed model can significantly outperform state-of-the-art models in terms of response quality in Scenario I and response informativeness in Scenario II.

Our contributions are three-fold: (1) proposal of image-grounded dialogue generation with both multimodal and unimodal data; (2) unifying text-to-image generation and image-grounded dialogue generation within a conditional variational auto-encoding framework; and (3) empirical verification of the effectiveness of the proposed approach in both image-grounded conversation and text-based conversation.

Methodology

Problem Formalization

Suppose that we have an image-grounded dialogue set $\mathcal{D}_I = \{(I_i, C_i, Y_i)\}_{i=1}^n$, where the i-th triple (I_i, C_i, Y_i) consists of an image I_i, a textual context $C_i = (u_{i,1}, u_{i,2}, \ldots, u_{i,l})$ with $u_{i,j}$ the j-th utterance, and a response Y_i. Besides, we further assume that there is a textual dialogue set $\mathcal{D}_T = \{(C_i, Y_i)\}_{i=1}^N$, where C_i and Y_i refer to a context and a response respectively. The goal is to learn two probability distributions $P(Y|I, C)$ and $P(Y|C)$ with both $\mathcal{D}_I$ and $\mathcal{D}_T$, and thus given a new pair (I, C) in image-grounded conversation and a new context C in text-based conversation, one can generate responses according to $P(Y|I, C)$ and $P(Y|C)$ respectively. $\forall C \in \mathcal{D}_T$, we assume the visual scene information at the time of conversation can be represented by a latent variable z. Then, $P(Y|C)$ is factorized as $P(Y|z, C) \cdot P(z|C)$. Here, by encoding an explicit image I and a latent image z in the same way, we can define $P(Y|I, C)$ and $P(Y|z, C)$ with one model. Thus, in the later part, we use $P(Y|z, C)$ and $P(Y|I, C)$ interchangeably.

Learning Objective

We learn $P(Y|z, C)$ and $P(Y|C)$ by maximizing the log-likelihood of $\mathcal{D}_I$ and $\mathcal{D}_T$ which is given by

$$\mathcal{J} = \underbrace{\sum_{(I,C,Y)\in\mathcal{D}_I} \log P(Y|C, I)}_{\mathcal{J}_I} + \underbrace{\sum_{(C,Y)\in\mathcal{D}_T} \log P(Y|C)}_{\mathcal{J}_T}. \quad (1)$$

While the first term $\mathcal{J}_I$ can be directly optimized through stochastic gradient descent, the problem is the second term $\mathcal{J}_T$, since $P(Y|C) = \int P(Y|z, C)P(z|C)\mathrm{d}z$ is often intractable. Thus, we employ the conditional variational auto-encoding (CVAE) framework (Sohn, Yan, and Lee 2015), and obtain the evidence lower bound (ELBO) of $\mathcal{J}_T$ as:

$$\mathcal{L}_T = -\mathrm{KL}[Q(z|C, Y)||P(z|C)] \\ + \mathbb{E}_{z\sim Q(z|C,Y)}[\log P(Y|z, C)] \leq \log P(Y|C). \quad (2)$$

where $\mathrm{KL}[\cdot||\cdot]$ refers to Kullback-Leibler divergence, and $Q(z|C, Y)$ is the posterior distribution of image generation. In CVAE, $\mathbb{E}_{z\sim Q(z|C,Y)}[\log P(Y|z, C)]$ is often approximated by sampling with $Q(z|C, Y)$ reparameterized using a deterministic function $g(C, Y, \epsilon)$ to reduce variance (Kingma and Welling 2013). Formally, $\mathbb{E}_{z\sim Q(z|C,Y)}[\log P(Y|z, C)]$ can be approximated as

$$\sum_{i=1}^L \frac{1}{L} \log P(Y|g(C, Y, \epsilon_i), C), \; \epsilon_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad (3)$$

where $\mathcal{N}(\mathbf{0}, \mathbf{I})$ denotes a normal distribution. Since the latent variable z represents an image, $g(C, Y, \epsilon)$ can be understood as reconstructing an image from (C, Y) (with a random noise ϵ). Without loss of generality, we define $g(\mathcal{T}, \epsilon)$ as an image reconstructor based on text $\mathcal{T}$. When $\mathcal{T} = (C, Y)$, $Q(z|C, Y)$ is defined by $\int_{\epsilon \in \Omega(z)} f(\epsilon|\mathbf{0}, \mathbf{I})\mathrm{d}\epsilon$, where $\Omega(z) = \{\epsilon|g(C, Y, \epsilon) = z\}$ and $f(\epsilon|\mathbf{0}, \mathbf{I})$ is the density of $\mathcal{N}(\mathbf{0}, \mathbf{I})$; when $\mathcal{T} = C$, $P(z|C) = \int_{\epsilon \in \Omega'(z)} f(\epsilon|\mathbf{0}, \mathbf{I})\mathrm{d}\epsilon$, where $\Omega'(z) = \{\epsilon|g(C, \epsilon) = z\}$.
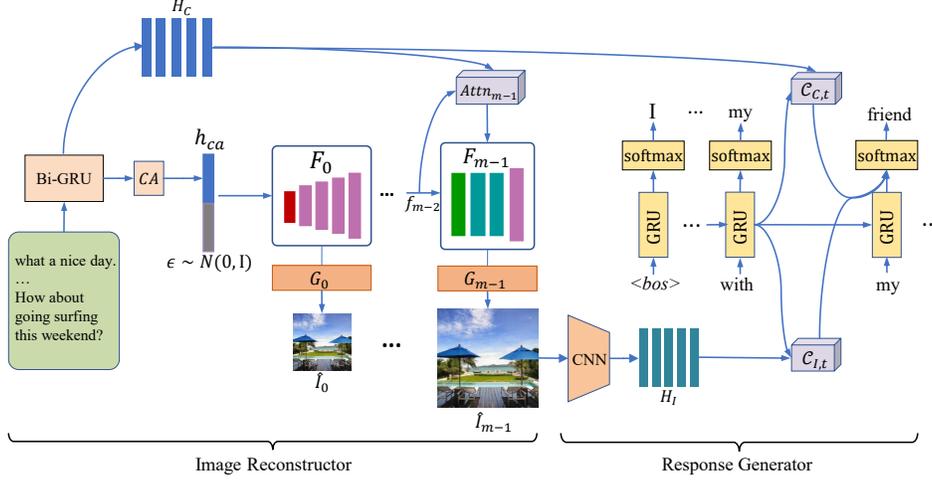
Figure 2: Architecture of our model.

Image Reconstructor $g(\mathcal{T}, \epsilon)$

The image reconstructor $g(\mathcal{T}, \epsilon)$ generates an image from text $\mathcal{T}$ and a gaussian random noise ϵ, and thus can be naturally modeled with GANs (Goodfellow et al. 2014) that represent the state-of-the-art technique in text-to-image (T2I) generation.

The left part of Figure 2 illustrates the architecture of $g(\mathcal{T}, \epsilon)$. As a premise, $\mathcal{T} = (w_1, \ldots, w_i, \ldots, w_L)$ is first transformed to $H_{\mathcal{T}} = (h_{w_1}, \ldots, h_{w_i}, \ldots, h_{w_L})$ by a bidirectional recurrent neural network with gated recurrent units (BiGRUs) (Cho et al. 2014), where $h_{w_i} \in \mathbb{R}^{d_1}$ is the hidden representation of the i-th token w_i, and L is the length of $\mathcal{T}$. Then, h_{w_L} is converted into a conditioning vector h_{ca} by the conditioning augmentation algorithm (Zhang et al. 2017) as input of image generation. We then construct a stacked attentional generative network that allows multi-stage refinement in image generation. The network consists of m attentional visual refiners $\{F_0, \cdots, F_{m-1}\}$ and m corresponding image generators $\{G_0, \cdots, G_{m-1}\}$ that generate an image for $\mathcal{T}$ from a small scale to a large scale. The generation process can be formulated as

$$
\begin{aligned}
f_0 &= F_0([\epsilon, h_{ca}]), \\
f_i &= F_i([f_{i-1}, \text{Attn}_i(H_{\mathcal{T}}, f_{i-1})]), \quad i \in \{1, \cdots, m-1\}, \\
\hat{I}_i &= G_i(f_i), \quad i \in \{0, \cdots, m-1\}
\end{aligned}
\tag{4}
$$

where $[\cdot, \cdot]$ represents a concatenation operation, $f_i \in \mathbb{R}^{d_2 \times N_i}$ denotes the image feature matrix of N_i sub-regions which is then fed into G_i to generate an image $\hat{I}_i \in \mathbb{R}^{3 \times N_i}$, and $\text{Attn}_i(\cdot, \cdot)$ is an attention module that encourages the sub-region feature to focus on certain words during generation. Specifically, $\forall i \in \{1, \ldots, m-1\}$, $\text{Attn}_i(H_{\mathcal{T}}, f_{i-1}) = (U_i H_{\mathcal{T}}) \text{softmax}(f_{i-1}^\top (U_i H_{\mathcal{T}}))^\top$, where $U_i \in \mathbb{R}^{d_2 \times d_1}$ is a parameter that maps $H_{\mathcal{T}}$ to the semantic space of f_{i-1}. F_0 consists of a fully connected layer and three upsampling layers, and $\forall i \in \{1, \cdots, m-1\}$, F_i consists of two residual blocks followed by an upsampling layer. $\forall i \in \{0, \cdots, m-1\}$, generator G_i is composed of a 3×3 convolutional layer

with $\tanh$ activation. The objective of learning is given by

$$
\mathcal{L}_G = \sum_{i=0}^{m-1} \mathcal{L}_{G_i}, \tag{5}
$$

where $\mathcal{L}_{G_i}$ is the adversarial loss of G_i which is defined as

$$
\mathcal{L}_{G_i} = -\mathbb{E}_{\hat{I}_i \sim P_{G_i}}[\log D_i(\hat{I}_i)] - \mathbb{E}_{\hat{I}_i \sim P_{G_i}}[\log D_i(\hat{I}_i, h_{w_L})]. \tag{6}
$$

In Equation (6), D_i is the discriminator corresponding to the generator G_i. The first term is the realism adversarial loss by which G_i tries to fool D_i with a generated image, and the second term is the text-image semantic consistency adversarial loss which determines if the generated image is consistent with the text condition. D_i is alternatively trained with G_i under an objective given by

$$
\begin{aligned}
\mathcal{L}_{D_i} = &- \mathbb{E}_{I_i \sim P_{data_i}}[\log D_i(I_i)] \\
&- \mathbb{E}_{\hat{I}_i \sim P_{G_i}}[\log(1 - D_i(\hat{I}_i))] \\
&- \mathbb{E}_{I_i \sim P_{data_i}}[\log D_i(I_i, h_{w_L})] \\
&- \mathbb{E}_{\hat{I}_i \sim P_{G_i}}[\log(1 - D_i(\hat{I}_i, h_{w_L}))],
\end{aligned}
\tag{7}
$$

where I_i a real image re-scaled to adapt to D_i. Note that we do not include the DAMSM loss (Xu et al. 2018) and the STREAM loss (Qiao et al. 2019) in the objective, since we find that they increase the cost of learning but do not make much difference in response generation.

In our experiments, the image reconstructor $g(\mathcal{T}, \epsilon)$ is pre-trained with data $\{(I, [C, Y])\}_{i=1}^n \cup \{(I, C)\}_{i=1}^n$ constructed from $\mathcal{D}_I$. After pre-training, instead of fine-tuning $g(\mathcal{T}, \epsilon)$ by optimizing $\mathcal{L}_T$, we fix the parameters of the model, that is, the parameters of $Q(z|C, Y)$ and $P(z|C)$ are also fixed. Thus $\text{KL}[Q(z|C, Y)||P(z|C)]$ becomes a constant and the learning objective now could be rewritten as

$$
\begin{aligned}
\mathcal{J} = &\sum_{(I, C, Y) \in \mathcal{D}_I} \log P(Y|C, I) \\
&+ \sum_{(C, Y) \in \mathcal{D}_T} \sum_{i=1}^L \frac{1}{L} \log P(Y|C, g(C, Y, \epsilon_i)).
\end{aligned}
\tag{8}
$$

Fixing $g(\mathcal{T}, \epsilon)$ may make the ELBO of $\mathcal{J}_T$ even looser, but it can let us circumvent the intractable KL term when $g(\mathcal{T}, \epsilon)$ is defined by a complicated non-linear function. In experiments, we find that a well pre-trained $g(\mathcal{T}, \epsilon)$ can already infer reasonable images for contexts, and thus aids response generation in both image-grounded conversation and text-based conversation. It is interesting to note that since $g(\mathcal{T}, \epsilon)$ is learned with GAN, the learning approach defined by Equation (8) in general falls in a (conditional) adversarial auto-encoding framework (Makhzani et al. 2015; Zhao et al. 2017).

Response Generator $P(Y|I, C)$

The right part of Figure 2 shows the architecture of the response generator $P(Y|I, C)$ ($P(Y|z, C)$). The model consists of a context encoder, an image encoder, and a response decoder. The context encoder flattens the context C by concatenating the utterances as a sequence of words of length L, and transforms C into a feature matrix $\mathrm{H}_C \in \mathbb{R}^{d_1 \times L}$ through a BiGRU shared with the text encoder of the image reconstructor. The image encoder is a convolutional neural network (CNN) built upon the Inception-V3 model (Szegedy et al. 2016) pre-trained on ImageNet (Deng et al. 2009). We rescale an image (either I or z) to be 299×299 pixels, and then feed the image to the encoder to extract region features $\tilde{\mathrm{H}}_I \in \mathbb{R}^{d_3 \times N_I}$ where N_I is the number of sub-regions. $\tilde{\mathrm{H}}_I$ is finally mapped to the space of H_C by $\mathrm{H}_I = W_I \tilde{\mathrm{H}}_I$ with $W_I \in \mathbb{R}^{d_1 \times d_3}$ parameter matrix. Parameters from the Inception-V3 model are fixed during learning.

The decoder predicts a response word by word through attending to both the context feature matrix H_C and the image feature matrix H_I. At step t, the hidden state of the decoder is calculated by

$$h_t = \mathrm{GRU}(h_{t-1}, e_{y_{t-1}}), \tag{9}$$

where $e_{y_{t-1}} \in \mathbb{R}^{d_4}$ is the embedding of the word generated at step $t-1$, and $h_{t-1} \in \mathbb{R}^{d_1}$ refers to the hidden state at step $t-1$ with $h_0 = h_{w_L}$. Then, when predicting the t-th word of the response, the decoder calculates the probability $P(y_t|y_{1:t-1}, C, I)$ by

$$P(y_t|y_{1:t-1}, C, I) = \mathrm{softmax}(W_o[h_t, \mathcal{C}_t] + b), \tag{10}$$

where $W_o \in \mathbb{R}^{|V| \times 2d_1}$ and $b \in \mathbb{R}^{|V|}$ are parameters, $|V|$ is the size of the response vocabulary, and $\mathcal{C}_t \in \mathbb{R}^{d_1}$ is a multimodal context vector defined as

$$\begin{aligned} \mathcal{C}_t &= \mathcal{C}_{C,t} + g \cdot \mathcal{C}_{I,t}, \\ g &= \sigma(W_g[h_t, \mathcal{C}_{C,t}, \mathcal{C}_{I,t}]). \end{aligned} \tag{11}$$

In Equation (11), $\mathcal{C}_{C,t} \in \mathbb{R}^{d_1}$ and $\mathcal{C}_{I,t} \in \mathbb{R}^{d_1}$ are obtained via attention over H_C and H_I respectively, i.e. $\mathcal{C}_{C,t} = \mathrm{H}_C \, \mathrm{softmax}(\mathrm{H}_C^\top h_t)$, $\mathcal{C}_{I,t} = \mathrm{H}_I \, \mathrm{softmax}(\mathrm{H}_I^\top h_t)$. $g \in (0, 1)$ is a gate that dynamically controls the contribution of $\mathcal{C}_{I,t}$ in response generation, $\sigma(\cdot)$ is the sigmoid function, and $W_g \in \mathbb{R}^{1 \times 3d_1}$ is a parameter. Here, the usage of gate g is motivated by the considerations that (1) open domain dialogues are not always related to their visual scene (e.g., the second turn in Figure 1); (2) even though the semantics of a response is grounded on the image (e.g., the first

turn in Figure 1), a large proportion of the response can still be irrelevant to the visual content (e.g., "she needs to adjust"), and (3) an inferred image z could be noisy. In these cases, a small g can block noisy signals given by $\mathcal{C}_{I,t}$. Let $Y = (y_1, \cdots, y_o)$, then the probability $P(Y|I, C)$ can be formulated as

$$P(Y|I, C) = P(y_1|C, I) \prod_{t=2}^{o} P(y_t|y_{1:t-1}, C, I). \tag{12}$$

Our approach is a generalized framework that $g(\mathcal{T}, \epsilon)$ and $P(Y|I, C)$ can be modeled by arbitrary GAN-based text to image generation models and image-grounded dialogue generation models respectively. The attentional generative adversarial network and the response generator, though effective, are just showcases.

Experiments

We test our model on two tasks: image-grounded conversation and text-based conversation. The first task requires a model to generate a response based on a textual context and a given image; while in the second task, a response is synthesized only based on the textual context.

Experimental Setup

Datasets. For image-grounded dialogue set $\mathcal{D}_I$, we choose Image-Chat data published in (Shuster et al. 2020). The dataset is made up of high-quality image-grounded open-domain dialogues collected from crowd-workers. Each dialogue consists of three turns at most based on a given image and two personalities. Since we focus on image-grounded conversation, the personality information in the data is discarded. The training/validation/test sets are split into 186,782/5,000/9,997 respectively. For the textual dialogue set $\mathcal{D}_T$, we use the Reddit Conversation Corpus[1] published by (Dziri et al. 2018) which contains more than 15M dialogues and each dialogue has at least 3 utterances. We keep $30,000$ most frequent words in the two data as a vocabulary for the text encoder and the response decoder. Other words are replaced by $\langle|\mathrm{UNK}|\rangle$. To reduce noise in the Reddit data, we remove dialogues in which more than 50% words in the response are $\langle|\mathrm{UNK}|\rangle$s, and dialogues with a response shorter than 4 words. After pre-processing, we randomly sample 1M/20K/20K dialogues as the training/validation/test set of the Reddit data. For both tasks, the last turn of dialogue in test set is used for evaluation.

Evaluation Metrics. We compare different models with both automatic metrics and human judgement. In automatic evaluation, we report *perplexity* (PPL) of ground-truth responses in test and measure quality of generated responses on both *relevance* and *informativeness*. In terms of relevance, besides BLEU-1 (Papineni et al. 2002) and Rouge-L (Lin 2004), we follow (Serban et al. 2017) and employ Embedding Average (Average), Embedding Extrema (Extrema), and Embedding Greedy (Greedy) as metrics. All the metrics are computed by scripts of a public NLG evaluation

---

[1]https://github.com/nouhadziri/THRED

project available at https://github.com/Maluuba/nlg-eval. In terms of informativeness, we follow (Li et al. 2015) and use Distinct-1 (Dist-1) and Distinct-2 (Dist-2) as metrics which are calculated as the ratios of distinct unigrams and bigrams in responses. In human evaluation, we recruit 3 well-educated native speakers as annotators to label the responses generated by each model. The annotators are required to judge the quality of each response from 3 aspects including *fluency*, *relevance* and *richness*, and assign a score from $\{0, 1, 2\}$ on each aspect which means bad, fair and good, respectively. In image-grounded conversation, relevance is judged in terms of both context and image. For each task, 500 examples are randomly sampled for annotation and each response receives 3 scores on each aspect. Average scores over annotators and responses are used as measures and the agreement among the annotators is measured by Fleiss' kappa (Fleiss and Cohen 1973).

Baselines

The following models are selected as baselines in image-grounded conversation: (1) T&I: a multimodal emotional response generation model proposed in (Huber et al. 2018), which conducts response generation based on textual features, image features, and emotional features. In our implementation, we just take account of the textural features and the image features and train the model on $\mathcal{D}_I$; (2) IMGRG: an ablation of the proposed model where the response generator is only learned with $\mathcal{D}_I$; and (3) T&I (W/ T) and IMGRG (W/ T): variants of baseline (1) and (2) which are trained with $\mathcal{D}_I \cup \mathcal{D}_T$ through patching a dummy image for each textual dialogue in $\mathcal{D}_T$.[2]

Baselines for text-based conversation include (1) SEQ2SEQ: the sequence to sequence model with attention (Bahdanau, Cho, and Bengio 2015); (2) HRED: the hierarchical recurrent encoder-decoder model proposed in (Serban et al. 2016); (3) VHRED: an extension of HRED that introduces latent variables into generation (Serban et al. 2017); and (4) RECOSA: a hierarchical transformer-based model that exhibits state-of-the-art performance on benchmarks of text-based conversation (Zhang et al. 2019). Note that to make the comparison fair, the baselines are trained with the text data in both $\mathcal{D}_I$ and $\mathcal{D}_T$.

We denote our model as IMGVAE, in which the image reconstructor is pre-trained with $\mathcal{D}_I$, and the response generator is then trained with both $\mathcal{D}_I$ and $\mathcal{D}_T$. Note that in image-grounded conversation task, all models perform response generation with the ground-truth images at inference time.

Implementation Details

In both tasks, d_1, d_2, d_3, and d_4 are set as 512, 48, 768, and 300 respectively. The image reconstructor has 2 attentional visual refiners (i.e. $m = 2$), and the number of image sub-regions N_0 and N_1 are set as 64×64 and 128×128 respectively. The dimension of ϵ and the dimension of the augmented conditioning vector are set as 100. To balance the cost effect, we check L within $\{1, 5\}$ and choose $L = 1$ in

---

[2]The RGB values of all pixels are set as (128,128,128).

our experiments. We learn all models using Adam algorithm (Kingma and Ba 2015) and the learning rates for image reconstructor and response generator are set as 1×10^{-4} and 1×10^{-3} respectively.

To stabilize adversarial training of the image reconstructor and avoid text representations being biased to image reconstruction, we pre-train the text encoder with seq2seq on the Reddit and textual part of Image-Chat training data, and fix the parameters in the learning of our model. Our model is trained on 4 Tesla 32GB P40 GPUs in a data-parallel manner with batch size 100. For the image-grounded conversation task, as there is no released code for T&I, we reproduce the model according to the framework in (Huber et al. 2018). To make fair comparisons, we set the key parameters (i.e. the dimension of embedding and hidden state, the size of vocabulary and the number of layers of encoder and decoder, etc.) to be consistent among IMGVAE and baselines. For text-based conversation task, SEQ2SEQ is implemented based on a public project at https://github.com/IBM/pytorch-seq2seq. HRED and VHRED are available at https://github.com/ctr4si/A-Hierarchical-Latent-Structure-for-Variational-Conversation-Modeling. For RECOSA, We run the code released at https://github.com/zhanghainan/ReCoSa with default settings. All models on both tasks are tuned until convergence by monitoring the PPL on the validation sets with the early stop strategy.

Evaluation Results

Table 1 reports evaluation results on automatic metrics. In image-grounded conversation, IMGVAE significantly outperforms all baseline models on most metrics. Particularly, ImgVAE outperforms T&I and IMGRG even after their training is augmented with the Reddit data. The results indicate the effectiveness of the proposed approach on leveraging both multimodal data and unimodal data for image-grounded dialogue generation. In text-based conversation, IMGVAE achieves comparable performance with the state-of-the-art deep transformer structure (i.e., RECOSA) in terms of response relevance and PPL, but improves upon informativeness of responses with large margins. This is because latent images, when properly controlled by the gate in the response generator, can enhance appearance of informative content in responses, as will be further verified by human annotations and the analysis in Discussions.

Table 2 reports human evaluation results. Basically, all models in both tasks can generate fluent and grammatical responses for most test input. In image-grounded conversation, IMGVAE outperforms all baselines in terms of context-relevance, image-relevance, and richness, which is consistent with the automatic evaluation results. In text-based conversation, IMGVAE significantly improves upon richness. which further demonstrates the effect of latent images. Besides, the information from the inferred images could enhance the understanding of context and promote to generate more relevant responses. All kappa values exceed or close to 0.6, indicating substantial agreement among the annotators.

14243

Task	Model	PPL	BLEU-1	Rouge-L	Average	Extrema	Greedy	Dist-1	Dist-2
Image-grounded	T&I	51.52	9.13	13.3	82.28	46.56	64.85	0.12	0.32
	IMGRG	51.93	12.50	14.42	85.45	49.93	67.28	0.55	1.95
	T&I (W/ T)	45.75	11.91	12.89	79.46	49.15	67.21	0.21	0.47
	IMGRG (W/ T)	46.19	13.61	14.72	84.65	**50.73**	**67.97**	0.88	3.06
	IMGVAE	**41.94**	**16.07**	**15.98**	**85.81**	49.59	67.44	**1.68**	**7.22**
	IMGVAE (W/O GATE)	43.41	15.45	15.08	85.18	49.41	67.11	1.35	5.95
Text-based	Seq2Seq	77.27	12.21	10.81	78.38	40.06	62.64	0.53	1.96
	HRED	84.02	11.68	11.29	75.54	37.49	60.41	0.89	3.21
	VHRED	78.01	12.22	11.82	75.57	39.24	62.07	0.87	3.49
	RECOSA	**71.75**	**12.75**	11.75	79.84	42.29	63.02	0.66	3.83
	IMGVAE	72.06	12.58	**12.05**	**79.95**	42.38	63.55	**1.52**	**6.34**
	IMGVAE (W/O GATE)	72.54	12.56	11.37	79.66	42.03	**63.63**	1.12	4.63

Table 1: Evaluation results on automatic metrics. Numbers in bold indicate the best performing model on the corresponding metrics.

Models	Fluency	Relevance		Richness	Kappa
		Text	Image		
Image-grounded conversation					
T&I	1.89	0.82	0.78	0.74	0.57
IMGRG	1.82	0.86	0.85	0.80	0.60
T&I (W/ T)	1.90	1.16	0.92	0.97	0.62
IMGRG (W/ T)	1.86	1.23	1.04	1.08	0.58
IMGVAE	**1.91**	**1.42**	**1.29**	**1.38**	0.65
Text-based conversation					
SEQ2SEQ	1.87	1.21	-	0.92	0.62
HRED	**1.88**	1.12	-	0.78	0.70
VHRED	1.66	1.05	-	1.10	0.61
RECOSA	1.87	1.32	-	1.12	0.63
IMGVAE	1.86	**1.48**	-	**1.47**	0.63

Table 2: Human evaluation results.

Figure 3: Performance of the models on small multimodal training data.

Discussions

In addition to the comparison with baselines, we are also curious about **Q1**: what is the performance of IMGVAE when image-grounded dialogues for training become more and more scarce? **Q2**: what content in responses is enriched by the latent images in text-based conversation? and **Q3**: what is the effect of the gate in the response generator in text-based dialogue generation?

Answer to Q1: Figure 3 illustrates the performance of IMGVAE and the baselines in terms of PPL and Rouge-L when the training size of Image-Chat is gradually halved. Note that the size of Reddit data is kept unchanged in training. We can see that when the multimodal training resource becomes more and more scarce, all baseline models suffer from dramatic performance drop. Particularly, since T&I and IMGRG count on the image-chat data, their performance drops faster than the others. This is because the baseline models, although some of them have been augmented with the textual dialogues in a trivial way, tend to overfit the small training data, and then generalize badly on the test set. On the other hand, benefiting from the large scale textual dialogues with latent images, IMGVAE exhibits robust performance in test with respect to the shrinkage of the training size of Image-Chat, and the advantage over the baselines
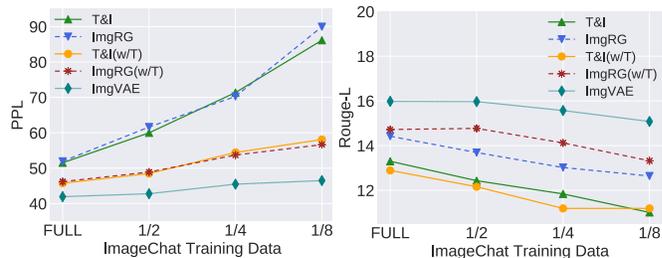
becomes bigger and bigger with the reduction of image-grounded dialogues. The results demonstrate the efficacy of the proposed method against data sparsity in low-resource image-grounded dialogue generation.

Answer to Q2: we define two new metrics *topic* and *novel-topic* in text-based conversation:

$$
\begin{aligned}
topic &= \frac{1}{|\mathcal{D}_T^t|} \sum_{(c,r)\in\mathcal{D}_T^t} \frac{|\tau(r_g)|}{l(r_g)}, \\
novel\text{-}topic &= \frac{1}{|\mathcal{D}_T^t|} \sum_{(c,r)\in\mathcal{D}_T^t} \frac{|\tau(r_g)-\tau(c)|}{l(r_g)},
\end{aligned}
\tag{13}
$$

where $\mathcal{D}_T^t$ refers to the test set of the Reddit data, (c,r) is a context-response pair, r_g is a response generated according to c, $\tau(x)$ returns a set of topical words in sentence x, $|\cdot|$ measures the size of a set, and $l(x)$ returns the length of x. We refer nouns and verbs as topical words because the topic of a dialogue is closely related to the objects and actions involved in the conversation, and recognize the POS tag of a word in a sentence with NLTK POS Tagger.[3] *topic* measures the average proportion of topical words in generated responses, while *novel-topic* further excludes topical words appearing in contexts. Table 3 gives the results on

---

[3]Tags in question include: NN, NNS, NNP, NNPS, VB, VBD, VBG, VBN, VBP, VBZ.

the two metrics. We can see that the latent images significantly enhance the ratio of topical words and the ratio of extra topical words in responses. SEQ2SEQ has a high *topic* score but the lowest *novel-topic* score, because it tends to copy words from contexts in response synthesis. *topic* and *novel-topic* for human responses in the test set are 0.398 and 0.321 respectively. This means that even though IMG-VAE can enhance appearance of informative content, it is still not so good as humans at bringing in new content and thus extending conversation, which could be a future direction for informative response generation.

Models	Seq2Seq	HRED	VHRED	ReCoSa	ImgVAE
topic	0.406	0.332	0.317	0.349	0.428
novel-topic	0.239	0.249	0.248	0.264	0.278

Table 3: Results on topical metrics.

Answer to Q3: first of all, the quantitative evaluation in Table 1 indicates that removing the gate from the response generator (i.e., IMGVAE (W/O GATE)) in general will cause performance drop on both tasks. Secondly, we find that when semantics of a textual context becomes complicated (e.g., with more nouns and verbs, and thus the topics become diverse in open chat), it is usually too challenging to recover a quality image from the context. Then, the gate shrinks, making the effect of the latent image (e.g., $\mathcal{C}_{I,t}$ in Equation (11)) fade in generation. In Figure 4, the figure on the left illustrates the distribution of average gate values of responses where the x-axis represents bins of test examples according to the number of topical words in the contexts.[4] Numbers below indicate how many test dialogues fall in the corresponding bins. We observe clear drop when the number of topical words in contexts increases. Another explanation is that a context with rich content can already provide enough information for response generation, and thus the latent image becomes marginal. Finally, we analyze the gate effect on topical words and stop words.[5] The figure on the right shows the comparison on test examples that own no more than 5 topical words in context. We find that stop words are less grounded by the latent images than topical words, even though the latent images are relatively useful on these examples.

Related Work

End-to-end open domain dialogue generation is inspired by machine translation when the vanilla sequence-to-sequence with attention architecture is applied to the task (Shang, Lu, and Li 2015; Vinyals and Le 2015) with promising results. Now, the model has been widely extended to handle the "safe response" problem (Li et al. 2015; Zhao, Zhao,

---

[4]Only 0.02% of the full 15M Reddit data do not contain a topical word. In the randomly sampled test data, the only 4 contexts without topical words are excluded from the analysis.

[5]We obtain stop words (totally 179) by NLTK toolkit available at https://github.com/nltk/nltk.
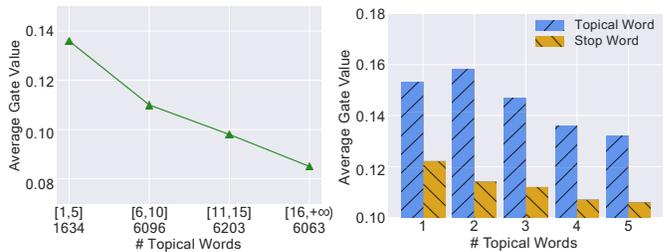

Figure 4: Analyses on the effect of the gate.

and Eskenazi 2017; Xing et al. 2017); to model conversation contexts (Serban et al. 2016, 2017; Xing et al. 2018); and to incorporate various types of knowledge for personalized (Li et al. 2016; Zhang et al. 2018b), emotional (Shuster et al. 2020; Mostafazadeh et al. 2017; Huber et al. 2018), document-grounded (Zhou, Prabhumoye, and Black 2018; Dinan et al. 2018; Zhao et al. 2020), and multimodal (Shuster et al. 2020; Mostafazadeh et al. 2017; Huber et al. 2018) conversation. This work falls in the research of multimodal open domain conversation in which a response is generated according to both a textual context and an image. To tackle the data sparsity problem, Shuster et al. (2020) pretrain the dialogue encoder on 1.7 billion textual Reddit data and achieve promising results at the fine-tuning stage. The difference we make is that through recovering the hidden image behind a textual dialogue, the image-grounded conversation and text-based conversation are unified within the CVAE framework, which not only enables the data augmentation, but the inferred images also help on the grounding of text-based dialogue generation.

Our work belongs to the interdisciplinary research of vision and language among various tasks such as image caption (Vinyals et al. 2015), visual question answering (Antol et al. 2015), visual dialog (Das et al. 2017), text to image generation (Xu et al. 2018; Qiao et al. 2019), vision-dialogue navigation (de Vries et al. 2018; Thomason et al. 2019), etc. Different from visual dialog in which models are designed to overcome the visual and contextual coreference and perform reasoning with both images and dialogue history, the challenges of image-grounded conversation lie in the lack of training data and that dialogues are not always grounded by images, etc. Rather than synthesizing an image from a caption, we consider recovering the image from a dialogue, which is encouraged by the promising results in the recent study on improving text-to-image generation by enriching caption with dialogues (Sharma et al. 2018).

Conclusions

We consider multimodal response generation with both image-grounded dialogues and textual dialogues by recovering the visual scene of a textual dialogue with an image reconstructor. The reconstructor is jointly learned with a response generator within a conditional variational autoencoding framework. Evaluation results indicate the efficacy of the proposed approach in both image-grounded conversation and text-based conversation.

Acknowledgments

References

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Lawrence Zitnick, C.; and Parikh, D. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, 2425–2433.

Bahdanau, D.; Cho, K.; and Bengio, Y. 2015. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations*.

Cho, K.; van Merriënboer, B.; Bahdanau, D.; and Bengio, Y. 2014. On the Properties of Neural Machine Translation: Encoder–Decoder Approaches. *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation* 103–111. doi:10.3115/v1/W14-4012. URL http://aclweb.org/anthology/W14-4012.

Chu, H.; Li, D.; and Fidler, S. 2018. A Face-to-Face Neural Conversation Model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7113–7121.

Das, A.; Kottur, S.; Gupta, K.; Singh, A.; Yadav, D.; Moura, J. M.; Parikh, D.; and Batra, D. 2017. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 326–335.

de Vries, H.; Shuster, K.; Batra, D.; Parikh, D.; Weston, J.; and Kiela, D. 2018. Talk the Walk: Navigating New York City through Grounded Dialogue. *CoRR* abs/1807.03367. URL http://arxiv.org/abs/1807.03367.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Dinan, E.; Roller, S.; Shuster, K.; Fan, A.; Auli, M.; and Weston, J. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. *arXiv preprint arXiv:1811.01241* .

Dziri, N.; Kamalloo, E.; Mathewson, K. W.; and Zaiane, O. 2018. Augmenting Neural Response Generation with Context-Aware Topical Attention. *arXiv preprint arXiv:1811.01063* .

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

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, 2672–2680.

Hori, C.; Alamri, H.; Wang, J.; Wichern, G.; Hori, T.; Cherian, A.; Marks, T. K.; Cartillier, V.; Lopes, R. G.; Das, A.; et al. 2019. End-to-end audio visual scene-aware dialog using multimodal attention-based video features. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2352–2356. IEEE.

Huber, B.; McDuff, D.; Brockett, C.; Galley, M.; and Dolan, B. 2018. Emotional dialogue generation using image-grounded language models. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 277. ACM.

Kingma, D. P.; and Ba, J. 2015. Adam: A method for Stochastic Optimization. In *ICLR*.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* .

Le, H.; Sahoo, D.; Chen, N.; and Hoi, S. 2019. Multimodal Transformer Networks for End-to-End Video-Grounded Dialogue Systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 5612–5623.

Li, J.; Galley, M.; Brockett, C.; Gao, J.; and Dolan, B. 2015. 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*, 110–119.

Li, J.; Galley, M.; Brockett, C.; Spithourakis, G.; Gao, J.; and Dolan, B. 2016. A Persona-Based Neural Conversation Model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 994–1003.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out* URL http://aclweb.org/anthology/W04-1013.

Makhzani, A.; Shlens, J.; Jaitly, N.; Goodfellow, I.; and Frey, B. 2015. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644* .

Mostafazadeh, N.; Brockett, C.; Dolan, B.; Galley, M.; Gao, J.; Spithourakis, G.; and Vanderwende, L. 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)*, 462–472.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. URL http://aclweb.org/anthology/P02-1040.

Qiao, T.; Zhang, J.; Xu, D.; and Tao, D. 2019. MirrorGAN: Learning Text-to-image Generation by Redescription. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1505–1514.

Ram, A.; Prasad, R.; Khatri, C.; Venkatesh, A.; Gabriel, R.; Liu, Q.; Nunn, J.; Hedayatnia, B.; Cheng, M.; Nagar, A.; et al. 2018. Conversational ai: The science behind the alexa prize. *arXiv preprint arXiv:1801.03604* .

Serban, I. V.; Sordoni, A.; Bengio, Y.; Courville, A. C.; and Pineau, J. 2016. Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models. In *AAAI*, volume 16, 3776–3784.

Serban, I. V.; Sordoni, A.; Lowe, R.; Charlin, L.; Pineau, J.; Courville, A. C.; and Bengio, Y. 2017. A Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues. In *AAAI*, 3295–3301.

Shang, L.; Lu, Z.; and Li, H. 2015. Neural Responding Machine for Short-Text Conversation. In *ACL*, 1577–1586.

Sharma, S.; Suhubdy, D.; Michalski, V.; Kahou, S. E.; and Bengio, Y. 2018. Chatpainter: Improving text to image generation using dialogue. *arXiv preprint arXiv:1802.08216* .

Shum, H.; He, X.; and Li, D. 2018. From Eliza to XiaoIce: Challenges and Opportunities with Social Chatbots. *Frontiers of Information Technology & Electronic Engineering* 19(1): 10–26. doi:10.1631/FITEE.1700826. URL https://doi.org/10.1631/FITEE.1700826.

Shuster, K.; Humeau, S.; Bordes, A.; and Weston, J. 2020. Image-Chat: Engaging Grounded Conversations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2414–2429. Online: Association for Computational Linguistics. doi:10.18653/v1/2020.acl-main.219. URL https://www.aclweb.org/anthology/2020.acl-main.219.

Sohn, K.; Yan, X.; and Lee, H. 2015. Learning Structured Output Representation Using Deep Conditional Generative Models. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, 3483–3491. Cambridge, MA, USA: MIT Press. URL http://dl.acm.org/citation.cfm?id=2969442.2969628.

Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to Sequence Learning with Neural Networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, 3104–3112. Cambridge, MA, USA: MIT Press. URL http://dl.acm.org/citation.cfm?id=2969033.2969173.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826.

Thomason, J.; Murray, M.; Cakmak, M.; and Zettlemoyer, L. 2019. Vision-and-Dialog Navigation. *CoRR* abs/1907.04957. URL http://arxiv.org/abs/1907.04957.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is All you Need. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 30*, 5998–6008. Curran Associates, Inc. URL http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf.

Vinyals, O.; and Le, Q. 2015. A Neural Conversational Model. *arXiv preprint arXiv:1506.05869* .

Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3156–3164.

Xing, C.; Wu, W.; Wu, Y.; Liu, J.; Huang, Y.; Zhou, M.; and Ma, W.-Y. 2017. Topic Aware Neural Response Generation. In *AAAI*, 3351–3357.

Xing, C.; Wu, W.; Wu, Y.; Zhou, M.; Huang, Y.; and Ma, W.-Y. 2018. Hierarchical Recurrent Attention Network for Response Generation. In *AAAI*, 5610–5617.

Xu, C.; Wu, W.; Tao, C.; Hu, H.; Schuerman, M.; and Wang, Y. 2019. Neural Response Generation with Meta-Words. *arXiv preprint arXiv:1906.06050* .

Xu, T.; Zhang, P.; Huang, Q.; Zhang, H.; Gan, Z.; Huang, X.; and He, X. 2018. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1316–1324.

Zhang, H.; Lan, Y.; Pang, L.; Guo, J.; and Cheng, X. 2019. ReCoSa: Detecting the Relevant Contexts with Self-Attention for Multi-turn Dialogue Generation. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, 3721–3730.

Zhang, H.; Xu, T.; Li, H.; Zhang, S.; Wang, X.; Huang, X.; and Metaxas, D. N. 2017. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 5907–5915.

Zhang, R.; Guo, J.; Fan, Y.; Lan, Y.; Xu, J.; and Cheng, X. 2018a. Learning to control the specificity in neural response generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, 1108–1117.

Zhang, S.; Dinan, E.; Urbanek, J.; Szlam, A.; Kiela, D.; and Weston, J. 2018b. 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)*, 2204–2213.

Zhao, J.; Kim, Y.; Zhang, K.; Rush, A. M.; and LeCun, Y. 2017. Adversarially regularized autoencoders. *arXiv preprint arXiv:1706.04223* .

Zhao, T.; Zhao, R.; and Eskenazi, M. 2017. Learning Discourse-level Diversity for Neural Dialog Models using Conditional Variational Autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, 654–664.

Zhao, X.; Wu, W.; Tao, C.; Xu, C.; Zhao, D.; and Yan, R. 2020. Low-Resource Knowledge-Grounded Dialogue Generation. *arXiv preprint arXiv:2002.10348* .

Zhou, K.; Prabhumoye, S.; and Black, A. W. 2018. A Dataset for Document Grounded Conversations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 708–713.