Dialogue Response Generation via Contrastive Latent Representation Learning

Shuyang Dai[1] **Guoyin Wang**[2] **Sunghyun Park**[2] **Sungjin Lee**[2]
[1]Duke University [2]Amazon Alexa AI
{shuyang.dai}@duke.edu
{guoyiwan, sunghyu, sungjinl}@amazon.com

Abstract

Large-scale auto-regressive models have achieved great success in dialogue response generation, with the help of Transformer layers. However, these models do not learn a representative latent space of the sentence distribution, making it hard to control the generation. Recent works have tried to learn sentence representations using Transformer-based framework, but do not model the context-response relationship embedded in the dialogue datasets. In this work, we aim to construct a robust sentence representation learning model, that is specifically designed for dialogue response generation, with Transformer-based encoder-decoder structure. An utterance-level contrastive learning is proposed, encoding predictive information in each context representation for its corresponding response. Extensive experiments are conducted to verify the robustness of the proposed representation learning mechanism. By using both reference-based and reference-free evaluation metrics, we provide detailed analysis on the generated sentences, demonstrating the effectiveness of our proposed model.

1 Introduction

Large-scale pretrained language models with Transformer-based architecture (Radford et al., 2019; Raffel et al., 2019; Zhang et al., 2019) have achieved state-of-the-art performance across a variety of natural language processing (NLP) tasks, including open-domain dialogue response generation. The majority of these models are auto-regressive (AR) language models, creating a matching mechanism between the input context and the generated response. However, such an approach often ignores the semantic meaning shared among sentences from an abstract level (Li et al., 2020), generating responses that have matching components with the context, but are not logically sound. Figure 1 demonstrates this potential problem by using some generated samples, where the response generation model generates similar patterns to the context but fail to predict semantically reasonable responses to continue the conversation.

Context (Dialogue History)
A: Hello, Miao Li, where are you going?
B: Hello, I am going to the store to buy some fruit.

Ground Truth Response
A: Oh, would you do me a favor?

Model Generated Examples
A: Oh, I see. Where are you going?
A: Hello! My name is Lin Ping, am I going to buy a beer?
A: Do you have any fruit?

Figure 1: Faulty generated sentences based on given context, via AR-based models. The provided dialogue is based on speakers 'A' and 'B'. Three model generated samples are given. Each pair of matching words/components are marked in the same color. The Ground Truth Response is provided for comparison.

In comparison, variational autoencoders (VAEs) (Kingma and Welling, 2013; Zhao et al., 2017; Gu et al., 2018) are able to explicitly model sentences in a representation space, providing control on the generated responses. While most VAE-based language models adopt shallow network such as LSTMs, it is straightforward to construct such a latent representation learning with a Transformer framework. Recent works (Li et al., 2020) aim to build a pretrained language model that uses a sentence-level VAE objective with a BERT (Devlin et al., 2018) encoder and a GPT-2 (Radford et al., 2019) decoder. Nevertheless, being designed for multiple language understanding and generation tasks, they do not utilize the enriched context-response relationship in the dialogue response generation task. (Maybe change to something like this: Nevertheless, these models are unable to utilize the enriched context-response relationship in the dialogue. This limits the model ability to capture predictive

relationship between context and response latent representations and hence constrains the response generation quality.)

In this work, we aim to design a dialogue response understanding and generation framework, that combines a higher-level sentence representation learning with a Transformer-based encoder-decoder architecture, yielding controllable response generation from an abstract level, while achieving top-tier generation quality. We adopt the approach that uses a BERT encoder and a GPT-2 decoder as the model backbone, and enrich the underlying information in the learned representation by performing utterance-level contrastive learning. Specifically, the latent representation of the context input (one or multiple utterances) is used to predict that of the response input (a single utterance). Note that our proposed contrastive loss is built based upon utterances, while the entire context input is encoded by a single context encoder.

To benefit the contrastive learning of the latent representation, we propose to use the hard negative sampling mechanism. Such an approach has been used in the computer vision domain (Robinson et al., 2020), guiding a learning method to correct its mistakes more quickly. In our case, we select the negative samples by using a pretrained context-response matching model (Cai et al., 2020). Given a context input, the responses with the top matching scores would be considered as the negative samples, and used for the contrastive objective. In terms of model architecture, we select conditional variational autoencoder (CVAE) (Zhao et al., 2017). This provides an encoder-decoder structure, in which the BERT-GPT backbone can be applied.

2 Preliminary

2.1 Contrastive Predictive Coding

The key idea of contrastive predictive coding (CPC) (Oord et al., 2018) is to learn representations that encode the underlying shared information between different parts of the high-dimensional signal, with the help of next step prediction. Given a sequence of inputs $\{x_1, x_2, \cdots, x_t, \cdots, x_T\}$ (*e.g.*, speech signal input), we obtain the latent representations $h_t = g_{\text{enc}}(x_t)$ for each x_t. A context representation $z_t = g_{\text{AR}}(h_{\leq t})$ summarizes all $\{h_1, h_2, \cdots, h_t\}$ using an auto-regressive model g_{AR}. To predict the future observations x_{t+k}, the following positive real score is considered.

$$f_k(x_{t+k}, z_t) = \exp(h_{t+k}^T W_k z_t), \quad (1)$$

where W_k is linear transformation for step k. To train the predictive model, a negative contrastive estimation (NCE) objective is used.

$$\mathcal{L}_{\text{CPC}} = -\mathbb{E}_X[\log \frac{f_k(x_{t+k}, z_t)}{\sum_{x_j \in X} f_k(x_j, z_t)}], \quad (2)$$

where $X = \{x_1, \cdots, x_N\}$ is a set of N random samples containing the positive sample x_{t+k} and $N-1$ negative samples from some distribution $p(x_{t+k})$. The negative samples aim to provide guiding on learning predictive information in the latent representations of the context inputs.

2.2 Conditional VAE

The conditional VAE (CVAE) (Zhao et al., 2017) provides a way to utilize the context feature as a conditional input to the VAE framework. Originally, VAE contains an encoder $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ and a decoder $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})$, where $\boldsymbol{x}$ is the input data and $\boldsymbol{z}$ is the latent representation. It assumes a Gaussian prior distribution $p(\boldsymbol{z})$, and by using the following objective, it approximates the expected log-likelihood.

$$\begin{aligned}\mathcal{L}_{\text{VAE}}(\boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{\phi}) = \mathbb{E}_{\boldsymbol{z}\sim q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})\right] \\ - \mathbb{D}_{\text{KL}}(q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}) \parallel p(\boldsymbol{z}))\,, \quad (3)\end{aligned}$$

where the conditional likelihood term (first) and the KL term (second) characterize reconstruction and generalization capabilities, respectively. In practice, reparameterization trick is used, in order to achieve Gaussian posterior distribution and compute the KL term in (3). Specifically, the encoder yields mean $\boldsymbol{\mu}$ and standard deviation $\boldsymbol{\sigma}$, and we can sample from such Gaussian distribution:

$$\boldsymbol{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}, \quad (4)$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. The CVAE, in comparison, has an additional condition input $\boldsymbol{c}$. In the dialogue response problem setup, the condition $\boldsymbol{c}$ corresponds to the context input, while $\boldsymbol{x}$ is the response input. Both the encoder and the decoder take an extra input $\boldsymbol{c}$, denoted as $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{c})$ and $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{c})$, respectively. For the prior, a prior network $p_{\boldsymbol{\psi}}(\boldsymbol{z}|\boldsymbol{c})$ is required, resulting in the updated objective as follow.

$$\begin{aligned}\mathcal{L}_{\text{CVAE}}(\boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\psi}) = \mathbb{E}_{\boldsymbol{z}\sim q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{c})}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{c})\right] \\ - \mathbb{D}_{\text{KL}}(q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{c}) \parallel p_{\boldsymbol{\psi}}(\boldsymbol{z}|\boldsymbol{c}))\,. \quad (5)\end{aligned}$$

During inference, we sample $\boldsymbol{z} \sim p_{\boldsymbol{\psi}}(\boldsymbol{z}|\boldsymbol{c})$ and generate using the decoder $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{c})$, while the

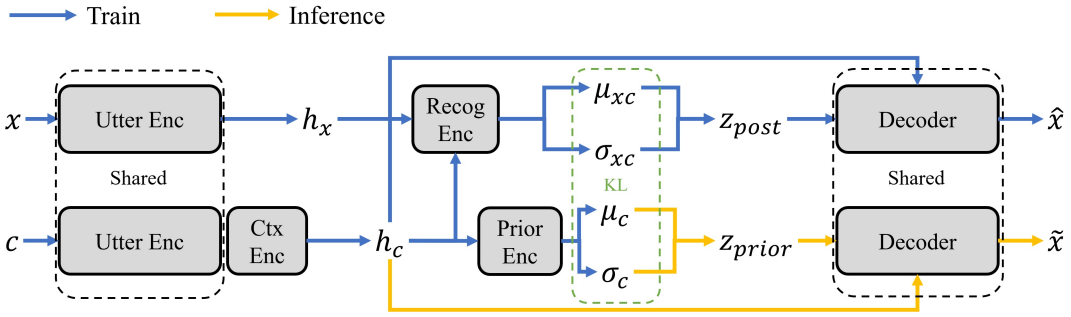

Figure 2: Conditional VAE Framework for Dialogue Response Generation with BERT Encoder and GPT-2 Decoder. The Training Flow is in Blue and the Inference Flow is in Orange.

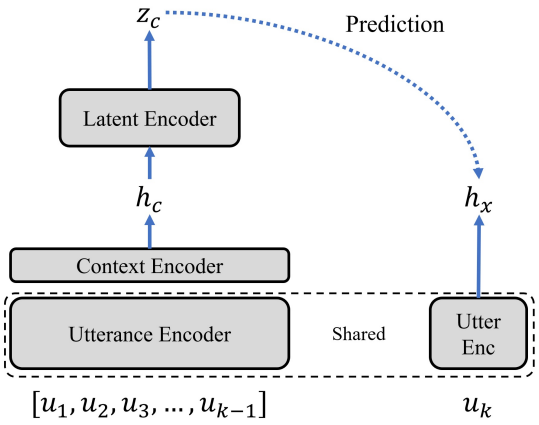

Figure 3: Applying Contrastive Learning on the Representation Framwork for Multi-turn Dialogue Response Generation Task.

response input $\boldsymbol{x}$ is not used. Similar to VAE, the reparameterization trick is used accordingly on both the posterior and prior networks.

3 Method

We include detailed model design in the following section. By introducing the utterance-level contrastive learning, our proposed model learns sentence representations that are capable of predicting future utterances. CVAE is considered as our main model architecture, where we use a BERT encoder and a GPT-2 decoder.

3.1 Representation Learning with a Transformer-Based Framework

We first introduce our model backbone, which is basically a CVAE model with a BERT encoder and a GPT-2 decoder. When encoding, we consider $\boldsymbol{h}_{[\text{CLS}]}$, the output of the special token $[CLS]$ of the last-layer, as the sentence-level representation, and obtain the latent representation by a linear mapping $e(\cdot)$ parameterized as $\boldsymbol{z} = \boldsymbol{W}_E \boldsymbol{h}_{[\text{CLS}]}$. In the CVAE framework, since we have the posterior and prior networks, we denote the former as the utterance encoder $f_u(\cdot)$, and define the latter as a combination of two encoders $f_c(f_u(\cdot))$, both of which are BERT encoders with outputs $\boldsymbol{h}_x$ and $\boldsymbol{h}_c$, respectively. Correspondingly, we have $e_r(\cdot)$ and $e_c(\cdot)$ as recognition encoder and prior encoder ($i.e.$, linear mappings), that take (h_x, h_c) and h_c as input, respectively. The outputs of the two encoder are $\boldsymbol{z}_{\text{post}} \sim q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{c})$ and $\boldsymbol{z}_{\text{prior}} \sim p_{\boldsymbol{\psi}}(\boldsymbol{z}|\boldsymbol{c})$, respectively.

When decoding, we follow the Memory approach in Optimus (Li et al., 2020). Specifically, we first compute $\boldsymbol{h}_M = \boldsymbol{W}_M \boldsymbol{z}$, where $\boldsymbol{W}_M \in \mathbb{R}^{LH}$ is a weight matrix. The output $\boldsymbol{h}_M \in \mathbb{R}^{LH}$ is separated into L vectors of length H, each of which is attended by one layer of the GPT-2 decoder $g(\cdot)$. Figure 2 shows the CVAE framework for both training and inference. Note that the outputs of the recognition encoder and the prior encoder are $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$, as the mean and standard deviation of the learnt $q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{c})$ and $p_{\boldsymbol{\psi}}(\boldsymbol{z}|\boldsymbol{c})$, respectively. As discussed in Sec 2.2, such a reparameterization trick is often used in VAE framework, where $\boldsymbol{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}$ for some $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$.

3.2 Utterance Level Contrastive Learning

With latent representation learnt, it is straightforward to adapt the contrastive learning in our model backbone. Different from speech signal data on which the original CPC objective is applied, dialogue datasets contain multi-turn utterances between two speakers for each conversation input. In this case, we consider encoding $\boldsymbol{c} = \{\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_{k-1}\}$ as a concatenation of all utterances in the context history. The output $\boldsymbol{h}_c$ is then sent to the latent encoder $e(\cdot)$, with an output $\boldsymbol{z} = e(\boldsymbol{h}_c)$ that is used to predict the sentence representation $\boldsymbol{h}_x$ of the next time step utterance $\boldsymbol{x} = \boldsymbol{u}_k$.

Figure 3 summarizes the utterance level contrastive learning procedure, with a training objective defined as follow.

$$\mathcal{L}_{\text{CL}} = -\mathbb{E}[\log \frac{\exp(\boldsymbol{h}_x \boldsymbol{W}_k \boldsymbol{z}_c)}{\sum_{\boldsymbol{x}_j \in X} \exp(\boldsymbol{h}_j \boldsymbol{W}_k \boldsymbol{z}_c)}], \quad (6)$$

where $\boldsymbol{W}_k$ is the weight matrix. Note that the proposed contrastive learning procedure can be directly applied to the CVAE model, in which $\boldsymbol{z}_c$ is basically the representation learnt by the prior encoder $e_c(\cdot)$. By combing the CVAE objective and the contrastive loss, we defined the final objective of our proposed model,

$$\mathcal{L} = \mathcal{L}_{\text{CVAE}} - \lambda \mathbb{E}_{\boldsymbol{z} \sim p_{\boldsymbol{\psi}}(\boldsymbol{z}|\boldsymbol{c})}[\log \frac{\exp(\boldsymbol{h}_x \boldsymbol{W}_k \boldsymbol{z})}{\sum_{\boldsymbol{x}_j \in X} \exp(\boldsymbol{h}_j \boldsymbol{W}_k \boldsymbol{z})}], \quad (7)$$

where λ is the hyper-parameter.

3.3 Improved CL with Hard Negative Sampling

We further improve our contrastive learning (CL) procedure by introducing hard negative sampling. The traditional CL usually uniformly selects negative samples from the data distribution. In the dialogue setup, this would be sampling random response samples $\boldsymbol{x}_j$ for a given context input $\boldsymbol{c}$. However, it is known in the computer vision domain that using negative samples, which look similar to the positive sample, would help correct the mistakes made by the encoder and learn useful information much more quickly (Robinson et al., 2020). For image datasets, one may consider adding noisy pixels, or rotating the positive images to obtain the negative ones. Similar approaches can be done in the natural language domain by using word permutation.

In our case, the multi-turn dialogue data setup allows us to further utilize the context-response relationship, and conduct hard negative sampling by using context-response matching models. Following (Cai et al., 2020), we consider training a Multi-hop Selector Network (MSN) (Yuan et al., 2019) which provides matching scores between the context and response inputs. Specifically, we construct a dialogue dataset, in which each context input $\boldsymbol{c}$ is paired with one positive response sample $\boldsymbol{x}$, and multiple randomly sample distrator response samples $\boldsymbol{x}_j$. The model is trained to predict whether a given context-response pair is positive or negative. When training our proposed CVAE model, the hard negative samples are selected based on higher matching scores, provided by the MSN model.

4 Related Work

Transformer-based representation learning for text generation. Transformer layers have been applied in various language models and achnce engaged in such a framework, it aims to generate novel ieved great results (Devlin et al., 2018; Yang et al., 2019; Keskar et al., 2019). Many are pretrained models and can be fine-tuned to adapt to various downstream tasks (Le and Mikolov, 2014; Dai and Le, 2015; Kiros et al., 2015). Recent works try to learn an interpretable representation space for the sentence encoding, using Transformer-based encoders and decoders. OPTIMUS (Li et al., 2020) provides a pretrained language model with a BERT encoder and a GPT-2 decoder. With variational inferesentences by learning a universal latent space through the pretraining. PLATO (Bao et al., 2019) also adopts BERT framework. It is specifically designed for dialogue generation, while learning a discrete latent variable for action recognition, upon which the generation is based.

Contrastive learning in NLP. Contrastive learning has been widely used with applications on various computer vision (Chopra et al., 2005; Schroff et al., 2015) and NLP problems (Mikolov et al., 2013; Logeswaran and Lee, 2018). It aims to learn a latent representation by contrasting positive and negative pairs in different setups. In the NLP domain, contrastive learning is applied in Word2Vec (Mikolov et al., 2013), a word embedding model, in which neighbouring words are predicted from context. Other than that, it is also used for sentence representation, where the positive and negative pairs are sampled as two contiguous sentences and sentences from other document, respectively (Logeswaran and Lee, 2018). CLAPS (Lee et al., 2020) is proposed recently for conditional text generation, sampling positive and negative sentences by adding perturbations to the input sequence. Being seq2seq model, CLAPS focuses on machine translation and text summarization, while our model tackles dialogue response generation, contrasting each utterance against other matching responses from the dialogue dataset.

5 Experiments

In this section, we present the experiment results of the proposed model, with comparison to sev-

	BLEU			BOW			Intra		Inter	
	R	P	F1	A	E	G	dist1	dist2	dist1	dist2
GPT-2	0.467	0.14	0.215	0.891	0.637	0.58	0.871	0.946	0.499	0.867
PLATO	0.331	0.174	0.228	0.929	0.623	0.747	0.957	0.989	0.412	0.626
iVAE	0.425	**0.248**	0.313	0.923	**0.641**	**0.784**	0.837	0.879	0.413	0.749
CVAE	0.489	0.226	0.309	0.928	0.607	0.626	**0.972**	**0.998**	**0.695**	**0.975**
CVAE + CL	**0.498**	0.239	**0.323**	**0.937**	0.621	0.64	0.936	0.972	0.661	0.948

Table 1: Reference-based Results on DailyDialog. (P: precision, R: recall, A: average, E: extreme, G: greedy). Higher BLEU and BOW Embedding indicate better quality of generated responses. Higher intra/inter-dist means better generation diversity.

	BLEU			BOW			Intra		Inter	
	R	P	F1	A	E	G	dist1	dist2	dist1	dist2
GPT-2	0.372	0.203	0.263	0.924	0.571	0.737	0.823	0.852	0.517	0.894
PLATO	0.382	0.237	0.293	0.934	**0.593**	**0.768**	0.852	0.903	0.431	0.88
GLC	**0.433**	0.243	0.311	0.928	0.543	0.717	0.774	0.879	-	-
CVAE	0.415	0.256	0.317	0.949	0.584	0.757	**0.947**	**0.995**	0.617	0.932
CVAE + CL	0.419	**0.267**	**0.326**	**0.95**	0.583	0.761	0.943	0.993	**0.618**	**0.938**

Table 2: Reference-based Results on PersonaChat.

eral baseline models across different dialogue response datasets. Detailed experimental setups are included, with analysis on both reference-based and reference-free evaluation.

5.1 Implementation Details

We implement our proposed model based on the CVAE model backbone. As discussed in Section 3.1, we apply the contrastive learning procedure (Section 3.2) on the CVAE framework, in which the context latent representation $\boldsymbol{z}_c$ is used to predict the response hidden state $\boldsymbol{h}_x$, with negative samples selected by the pretrained MSN model. Our model uses BERT encoder (12 transformer layers with hidden size 768 and self-attention heads 12), and GPT-2 decoder (12 transformer layers with hidden size 768 and self-attention heads 12).

5.2 Datasets

We mainly evaluate our model on two benchmark datasets, namely the DailyDialog dataset (Li et al., 2017) and the PersonaChat dataset (Zhang et al., 2018). The former contains 13K daily conversations for a English learner, with an average number of turns in a conversation as 8.5. The latter is a knowledge grounded conversation dataset where two participants chat naturally and try to get to know each other. A total of 11K dialogues are contained in PersonaChat with an average number of turns per conversation as 15. For both datasets, we process each utterance as the response of the previous context utterances from both speakers. Both datasets are separated into train, validation, and test subsets, with a separation ratio of 10:1:1 for DailyDialog, and 9:1:1 for PersonaChat.

5.3 Baseline Models

We select State-Of-The-Art (SOTA) models from different model categories as our baseline models. For auto-regressive model, we consider DialoGPT (Zhang et al., 2019), which is a GPT-2 based model that is specifically designed for dialogue response generation. For dialogue response generation model that uses contrastive learning, we include group-wise contrastive learning (GCL) (Cai et al., 2020), which conduct CL between target dialogue model and a pretrained reference model. PLATO (Bao et al., 2019) is another model that uses transformer-based model architecture while including a discrete latent variable to tackle the one-to-many mapping problem. Last but not the least, we compare our model to the SOTA VAE model on text generation, the implicit VAE (iVAE) (Fang et al., 2019), in which the original Gaussian assumption is replaced by an implicit distribution, learnt in an adversarial fashion.

5.4 Evaluation Metrics

In our work, we consider the traditional reference-based, as well as several newly proposed reference-

	DialogRPT				RoBERTa Eval
	updown	depth	width	score	
Grount Truth	0.403	0.537	0.591	0.319	2.812
GPT-2	0.311	0.414	0.437	0.243	**2.739**
PLATO	0.314	0.453	0.468	0.255	2.132
CVAE	0.392	0.496	0.547	0.301	1.931
CVAE + CL	**0.407**	**0.519**	**0.563**	**0.313**	2.106

Table 3: Reference-free Results on Dailydialog (H vs R: Human vs Random; H vs M: Human vs Machine). Higher scores indicate better generation quality in the corresponding categories.

	DialogRPT				RoBERTa Eval
	updown	depth	width	score	
GPT-2	0.339	0.523	0.625	0.301	1.383
PLATO	0.358	0.574	0.641	0.318	1.582
GLC	0.361	0.577	0.639	0.319	1.607
CVAE	0.367	0.577	0.647	0.322	1.678
CVAE + CL	**0.372**	**0.583**	**0.669**	**0.329**	**1.702**

Table 4: Reference-free Results on PersonaChat.

free evaluation metrics. The former basically measures the relevance between the reference response data (ground truth) and the generated response samples. In comparison, the reference-free evaluators do not need the reference responses, while computing the score based on a trained evaluation model.

We mainly consider three reference-based evaluation metrics: 1) BLEU score measures how many n-gram ($n = 4$ in our experiments) from generated response overlaps with the references. We sample 10 responses; BLEU-precision and BLEU-recall are defined as average and maximum scores. 2) Cosine similarity of bag-of-words (BOW) embedding between the generated response and the reference is considered ($e.g.$, 3 types of embedding including greedy, average, and extreme). 3) Distinct evaluates the diversity of the generation; dist-n is the ratio of unique n-grams ($n = 1, 2$) over all n-grams in the generated responses. We evaluate both within each sampled response and among all responses as intra-dist and inter-dist, respectively.

For reference-free evaluators, we consider the following two metrics. The first one is DialogRPT (Gao et al., 2020), which provides ranking scores of responses based on human evaluation on i) Width: the number of dialogues after the current turn $(\boldsymbol{c}, \boldsymbol{x})$; ii) Depth: the maximum length of the dialogue after the current turn; and iii) Updown: the number of up votes minus the number of down votes (by human raters). Another one is the RoBERTa Evaluator (Zhao et al., 2020), which is first trained on next-sentence-prediction task in an unsupervised manner, and then trained on predicting human annotated data in a semi-supervised setup. The model is trained using the RoBERTa encoder (Liu et al., 2019), on the concatenation of the input context $\boldsymbol{c}$ and the generated response $\tilde{\boldsymbol{x}}$, with output rescaled to the human rating scale.

5.5 Results

Table 1 and 2 show the reference-based results on DailyDialog and PersonaChat datasets, respectively. In terms of BLEU score, our proposed CVAE/CVAE+CL achieves better performance than the other Transformer-based models. By adding the proposed utterance level contrastive learning (CL) to CVAE, the BLEU score is improved from 0.309 to 0.318, outperforming GPT-2 and PLATO, while being comparable to iVAE. Meanwhile, CVAE-based models provide much higher generation diversity in both inter and intra-dist. Note that adding contrastive learning does slightly drop the generation diversity of CVAE, which is likely due to the additional predictive information learnt in the representation of context inputs. Overall, our proposed model CVAE+CL achieves top-tier performance in the majority of the evaluation categories, across both datasets.

Table 3 and 4 present the reference-free results on DailyDialog and PersonaChat, respectively. For

Models	Recall	Precision	F1
Proposed Model	0.498	0.239	0.323
- HNS	0.49	0.235	0.317
- CL	0.489	0.226	0.309
- TF	0.265	0.222	0.242

Table 5: Ablation Study on DailyDialog via Reference-based Evaluation Metrics.

Objective	λ	Recall	Precision	F1
$\mathcal{L}_{\text{CVAE}} + \lambda\mathcal{L}_{\text{CL}}$	0.0	0.489	0.226	0.309
	0.1	0.489	0.232	0.315
	0.2	0.493	0.232	0.316
	0.5	0.498	0.239	0.323
	1.0	0.496	0.24	0.323
	2.0	0.492	0.235	0.318

Table 6: Sensitiveness of Contrastive Learning Objective against Hyper-parameter λ. Results on DailyDialog via Reference-based Evaluation Metrics.

DailyDialog, CVAE provides better overall DialogRPT scores than the other baseline models. However, since responses are generated based on the context latent representations (instead of direct sentence inputs), CVAE does not provide matching words across context and response, reducing the RoBERTa score for the DailyDialog dataset. For PersonaChat, CVAE outperforms the other baseline models in both evaluation metrics. Overall, we see the effectiveness of adding CL to the CVAE framework, improving both DialogRPT and RoBERTa scores across the two datasets.

Ablation study is conducted based on each components in the proposed CVAE. Table 5 shows the BLEU score ($i.e.$, recall, precision, and F1) of our model on the DailyDialog dataset, in which we gradually remove one component at a time to demonstrate the effectiveness from each of them. We first remove hard negative sampling (HNS) and use random negative samples (from the same training mini batch) instead. The contrastive learning (CL) objective is then removed, where we notice a drop of 0.014 (from 0.323 to 0.309) in F1. Finally, the model performance drops drastically when we replace the transformer layers (TF) in the CVAE with LSTM ($i.e.$, original CVAE framework); F1 decreeases from 0.309 to 0.242.

We further show the sensitiveness of our proposed utterance-based contrastive loss against different values of hyper-parameter λ. By tuning λ from 0.0 to 2.0, we report the BLEU scores of each training objectives. As shown in Table 6, the model with $\lambda = 0.5$ and $\lambda = 1.0$ achieves the best overall performance, while the models with the other values of λ also provide competitive results. This demonstrates that our model is relatively insensitive to the hyper-parameter for contrastive learning objective. Meanwhile, we can see that adding the contrastive loss is improving the performance (regardless of different values of λ) when comparing the baseline CVAE ($i.e.$, $\lambda = 0.0$).

Context: - Say, Jim, how about going for a few beers after dinner?
Response: You know that is tempting but is really not good for our fitness.
All right. I'll cook for you.
I do like Thai food. Did you see much of Bangkok?
I'm so thirsty.

Context: - Are you all right? - I will be all right soon. I was terrified when I watched them fall from the wire.
Response: Don't worry. He is an acrobat.
That's all right.
And do you feel anything when I do this?
And do you have any questions?

Context: - Can you do push-ups? - Of course I can. It's a piece of cake! Believe it or not, I can do 30 push-ups a minute. - Really? I think that's impossible! - You mean 30 push-ups? - Yeah!
Response: It's easy. If you do exercise everyday, you can make it, too.
Believe me, he is the first-class.
Yeah. I'm going with Jennie.
Don't worry, Shan always screws up free throws.

Figure 4: Examples of Hard Negative Sampling. Blue Ones are Ground Truth Reference, and Red Ones are Negative Samples Selected by MSN.

Figure 4 provides examples of hard negative sentences selected by MSN. Each example case includes context sentence(s), a ground truth response sentence (reference), as well as 3 selected hard negative samples. As can be seen, the negative sentences share relevant words and contents with the input context, which provides a more efficient training for our model.

6 Conclusion

In this work, we propose a dialogue response generation model that combines the controllability of representation learning models and SOTA performance of AR-based frameworks. An efficient utterance level contrastive learning is introduced to enrich the information learned in the latent representations ($i.e.$, predictive information of utterance in the next time step). Our proposed model achieves competitive performance comparing to the SOTA models in different benchmark datasets.

References

Siqi Bao, Huang He, Fan Wang, Hua Wu, and Haifeng Wang. 2019. Plato: Pre-trained dialogue generation model with discrete latent variable. *arXiv preprint arXiv:1910.07931*.

Hengyi Cai, Hongshen Chen, Yonghao Song, Zhuoye Ding, Yongjun Bao, Weipeng Yan, and Xiaofang Zhao. 2020. Group-wise contrastive learning for neural dialogue generation. *arXiv preprint arXiv:2009.07543*.

Sumit Chopra, Raia Hadsell, and Yann LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546. IEEE.

Andrew M Dai and Quoc V Le. 2015. Semi-supervised sequence learning. *Advances in neural information processing systems*, 28:3079–3087.

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

Le Fang, Chunyuan Li, Jianfeng Gao, Wen Dong, and Changyou Chen. 2019. Implicit deep latent variable models for text generation. *arXiv preprint arXiv:1908.11527*.

Xiang Gao, Yizhe Zhang, Michel Galley, Chris Brockett, and Bill Dolan. 2020. Dialogue response ranking training with large-scale human feedback data. *arXiv preprint arXiv:2009.06978*.

Xiaodong Gu, Kyunghyun Cho, Jung-Woo Ha, and Sunghun Kim. 2018. Dialogwae: Multimodal response generation with conditional wasserstein autoencoder. *arXiv preprint arXiv:1805.12352*.

Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Ryan Kiros, Yukun Zhu, Russ R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in neural information processing systems*, pages 3294–3302.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196. PMLR.

Seanie Lee, Dong Bok Lee, and Sung Ju Hwang. 2020. Contrastive learning with adversarial perturbations for conditional text generation. *arXiv preprint arXiv:2012.07280*.

Chunyuan Li, Xiang Gao, Yuan Li, Baolin Peng, Xiujun Li, Yizhe Zhang, and Jianfeng Gao. 2020. Optimus: Organizing sentences via pre-trained modeling of a latent space. *arXiv preprint arXiv:2004.04092*.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lajanugen Logeswaran and Honglak Lee. 2018. An efficient framework for learning sentence representations. *arXiv preprint arXiv:1803.02893*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. 2020. Contrastive learning with hard negative samples. *arXiv preprint arXiv:2010.04592*.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Chunyuan Yuan, Wei Zhou, Mingming Li, Shangwen Lv, Fuqing Zhu, Jizhong Han, and Songlin Hu. 2019. Multi-hop selector network for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 111–120.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. *arXiv preprint arXiv:1703.10960*.

Tianyu Zhao, Divesh Lala, and Tatsuya Kawahara. 2020. Designing precise and robust dialogue response evaluators. *arXiv preprint arXiv:2004.04908*.