Variational Transformers for Diverse Response Generation

Zhaojiang Lin[†], Genta Indra Winata, Peng Xu, Zihan Liu, Pascale Fung

Center for Artificial Intelligence Research (CAiRE)
Department of Electronic and Computer Engineering
The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong
[†]`zlinao@connect.ust.hk`

Abstract

Despite the great promise of Transformers in many sequence modeling tasks (e.g., machine translation), their deterministic nature hinders them from generalizing to high entropy tasks such as dialogue response generation. Previous work proposes to capture the variability of dialogue responses with a recurrent neural network (RNN)-based conditional variational autoencoder (CVAE). However, the autoregressive computation of the RNN limits the training efficiency. Therefore, we propose the Variational Transformer (VT), a variational self-attentive feed-forward sequence model. The VT combines the parallelizability and global receptive field of the Transformer with the variational nature of the CVAE by incorporating stochastic latent variables into Transformers. We explore two types of the VT: 1) modeling the discourse-level diversity with a global latent variable; and 2) augmenting the Transformer decoder with a sequence of fine-grained latent variables. Then, the proposed models are evaluated on three conversational datasets with both automatic metric and human evaluation. The experimental results show that our models improve standard Transformers and other baselines in terms of diversity, semantic relevance, and human judgment.

1 Introduction

Convolutional and fully-attentional feed-forward architectures, such as Transformers (Vaswani et al., 2017), have emerged as effective alternatives to RNNs (Dehghani et al., 2018) in wide range of NLP tasks. These architectures remove the computational temporal dependency during the training and effectively address the long-standing vanishing gradients problem of recurrent models by processing all inputs simultaneously. Notably, transformers apply a fully attention strategy, where each token in the sequence is informed by other tokens via a self-attention mechanism. It acts as an effectively global receptive field across the whole sequences which absence in RNNs. Despite the powerful modeling capability of trasnformers, they often fail to model *one-to-many* [1] relation in dialogue response generation tasks (Zhao et al., 2017) due to their deterministic nature. As a result, they generate dull and generic response *(e.g., "I am not sure")*, especially with greedy and beam search, which are widely used in other sequence modeling tasks. There have been attempts to generate diverse and informative dialogue responses by incorporating latent variable(s) into the RNN encoder-decoder architecture. In particular Zhao et al. (2017) adapt a conditional variational autoencoder (CVAE) to capture discourse-level variations of dialogue, while Goyal et al. (2017) and Du et al. (2018) integrates latent variables in the hidden states of the RNN decoder. However, the inherently sequential computation of aforementioned models limit the efficiency for large scale training.

In this paper, we introduce the Variational Transformer (VT) [2] a variational self-attentive feed-forward sequence model to address the aforementioned issues. The VT combine the parallelizability and global receptive field of the transformer with the variational nature of CVAE by incorporating stochastic latent variables into transformers. We explore two types of VT: 1) Global Variational Transformer (GVT), and 2) Sequential Variational Transformer. The GVT is the extension of CVAE in Zhao et al. (2017), which modeling the discourse-level diversity with a global latent variable, While SVT, inspired by variational autoregressive models (Goyal et al., 2017; Du et al., 2018), incorporates a sequence of latent variables into decod-

---

[1] Given a similar dialogue history, there may exist many valid responses.
[2] The source code is available in `https://github.com/zlinao/Variational-Transformer`

ing process by using a novel variational decoder layer. Unlike previous approaches (Zhao et al., 2017; Goyal et al., 2017; Du et al., 2018), SVT uses *Non-causal Multi-head Attention*, which attend to future tokens for computing posterior latent variables instead of using an additional encoder.

The proposed VT architectures integrate stochastic latent variables into Transformers. The experimental results on a three conversation dataset demonstrate that our models can generate more informative and coherent responses.

2 Related work

2.1 Neural Conversational Models

Conversational systems has been widely studied (Weizenbaum et al., 1966; Wallace, 2009; Vinyals and Le, 2015; Serban et al., 2016). Compare to rule-based systems (Weizenbaum et al., 1966; Wallace, 2009), sequence-to-sequence conversation models achieve superior performance in terms of scalable training and generalization ability (Vinyals and Le, 2015). However, it has been pointed out that encoder-decoder models tend to generate generic and repetitive responses like *"I am sorry"* (Li et al., 2016a). To address this issue, there have been three main lines of work. The first is adding additional information (e.g., persona) as input to guild model generate more informative responses (Li et al., 2016b; Zhang et al., 2018). The second modifies the learning objective to promote more diverse generation (Li et al., 2016a), and the third integrates stochastic latent variables into Seq2Seq models by using the CVAE framework (Serban et al., 2017; Zhao et al., 2017). Our work comes within this third line introducing a novel model, the Variational Transformer, to improve dialogue response generation.

2.2 Conditional Variational Autoencoders

Many works have attempted to combine CVAEs with encoder-decoder architectures for sequence generation tasks. Zhang et al. (2016) propose a variational encoder-decoder model for neural machine translation, while Li et al. (2017) apply variational recurrent neural networks (VRNN) (Chung et al., 2015) for text summarization. Zhao et al. (2017) and Zhou and Wang (2018) explore incorporating meta features into CVAE framework in dialogue response generation tasks. (Goyal et al., 2017) and (Du et al., 2018) propose variational autoregressive decoders which enhanced by highly multi-modal

latent variables to capture the high variability in dialogue responses. Le et al. (2018) further augment variational autoregressive decoders with dynamic memory networks for improving generation quality. We unify the previous successful ideas of CVAE, and explore the combinations of CVAE and Transformer.

2.3 Fully Attentional Networks

Taking advantage of the parallel-in-time structure and global receptive field, Transformers (Vaswani et al., 2017) have recently been shown to achieve impressive results on various sequence modeling tasks. Based on this, several follow-up models have been presented. The Image Transformer (Parmar et al., 2018) has been proposed for image generation, while the MultiModel (Kaiser et al., 2017) integrates convolution, attention and sparsely-gated mixture-of-expert blocks into a single deep-learning model for simultaneously learning multiple tasks from various domains. Lin et al. (2019) proposed a fully attentional mixture-of-expert model (MoEL) for empathetic dialogue modeling. The Universal Transformer (Dehghani et al., 2018) incorporates the recurrent inductive bias of RNNs into the standard Transformer, and achieves better result on a wide range of algorithmic and language understanding tasks. Kaiser et al. (2018) introduce the Latent Transformer (LT) for non-autoregressive machine translation. During training, the LT first autoencodes a target sequence into a shorter sequence discrete latent variables. Then a parallel decoder decodes the target using discrete latent variables and an input sequence. Different from the LT (Kaiser et al., 2018), the VT generates continuous latent variables during the decoding process.

3 Preliminaries

3.1 Conditional Variational Autoencoder for Dialogue Generation

The CVAE framework (Sohn et al., 2015) represents a dyadic conversation via three random variables: the input condition c, including conversation context and meta features (meta features can be ignored when not available); a latent variable z; and the target response x. A CVAE can be efficiently trained with Stochastic Gradient Variational Bayes (SGVB) (Kingma and Welling, 2013) by maximizing the variational lower bound of x given

c, according to:

$$p(x|c) = \int_z p(x|z,c)p(z|c)dz. \quad (1)$$

The typical CVAE consists of a prior network $p_\theta(z|c)$, which is used to approximate $p(z|c)$, a recognition network $p_\phi(z|c,x)$, which is used to approximate posterior distribution $q(z|c,x)$, and a decoder $p_\theta(x|z,c)$, which is used to approximate $p(x|z,c)$. By assuming z follows multivariate Gaussian distribution with a diagonal co-variance matrix, the evidence lower bound (ELBO) can be written as

$$
\begin{aligned}
\mathcal{L}_{ELBO} &= \mathcal{L}_{REC} - \mathcal{L}_{KL} \\
&= \mathbf{E}_{q_\phi(z|c,x)}\left[\log p_\theta(x|z,c)\right] \\
&\quad - KL\left(q_\phi(z|c,x)\|p_\theta(z|c)\right) \\
&\leq \log p(x|c),
\end{aligned} \quad (2)
$$

where $\mathcal{L}_{REC}$ denotes the reconstruction loss and $\mathcal{L}_{KL}$ denotes the Kullback-Leibler (KL) divergence between the posterior and prior.

In dialogue generation tasks, previous works (Zhao et al., 2017; Zhou and Wang, 2018) apply RNN encoders (with GRU or LSTM cell) to encode dialogue contexts and responses separately. The condition c is represented by the concatenation of the last hidden state of the context encoder and the meta features (e.g., topic, emotion), while the response x is represented by the last hidden state of response encoder. Then the prior network $p_\theta(z|c)$ and the recognition network $p_\phi(z|c,x)$ parameterized by multi-layer perceptrons (MLPs) are applied to approximate the means and the log variances of the prior latent distribution $\mathcal{N}\left(z; \mu', \sigma'^2\mathbf{I}\right)$ and posterior latent distribution $\mathcal{N}\left(z; \mu, \sigma^2\mathbf{I}\right)$. With the reparameterization trick (Kingma and Welling, 2013), we can obtain samples of the prior latent variable (for testing) from $\mathcal{N}\left(z; \mu', \sigma'^2\mathbf{I}\right)$ and samples of the posterior latent variable (for training) from $\mathcal{N}\left(z; \mu, \sigma^2\mathbf{I}\right)$. Finally, an RNN decoder use z and c as the initial state to predicts the response x.

The *vanishing latent variable problem* (Bowman et al., 2016) is a common issue in RNN-based CVAEs. That is, the powerful autoregressive RNN decoder first learns to ignore the latent variable, and decodes the response by only condition on the previous tokens. Thus the latent variable fails to encode the meaningful information, and the CVAE deteriorates to seq2seq model. To alleviate this issue, *KL annealing* (Bowman et al., 2016) and
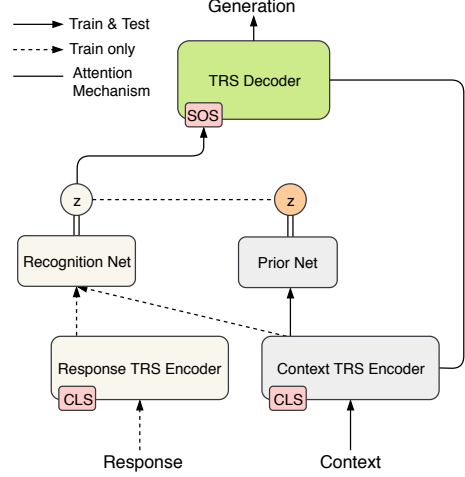

Figure 1: The Global Variational Transformer. During training, The posterior latent variable z by the posterior network is passed to the decoder, while during testing, the target response is absent, and z is replaced by the prior latent variable. The word embeddings, positional encoding, softmax layer and meta vectors are ignored for simplicity

bag-of-word loss (Zhao et al., 2017) have been proposed, and have shown effectiveness in various dialogue tasks (Zhao et al., 2017; Zhou and Wang, 2018).

3.2 CVAE with Transformer

The aforementioned RNN-based CVAE framework integrate the latent variable into the initial state of RNN decoder, while in transformer, it is more flexible to incorporate the latent variable embedding into the first input token of the decoder to generate the initial state.

The overall architecture of GVT is depicted in Figure 1. Different from RNNs, the Transformer encoder maps an input sequence of symbol representations to a sequence of contextualized representations (Vaswani et al., 2017). In order to get fixed dimension representations of the response and context, we add a special token CLS at the beginning of the input sequence as in BERT (Devlin et al., 2018), to compute the weighted sum of the output representations via self-attention. Thus the output representation of the token CLS is considered as the representation of the whole sequence. Then we introduce a recognition network and a prior network to compute the posterior latent variable and prior latent variable as in (Zhao et al., 2017; Zhou and Wang, 2018). We add the latent variable sample z and meta features m (can be ignored

when not available) into e_{SOS}, the embedding of the *start-of-sequence* token SOS:

$$e'_{SOS} = z + m + e_{SOS}. \quad (3)$$

Finally, the transformer decoder decodes the response x sequentially while attending to the new embedding e'_{SOS} of token SOS with latent information.

This design enhances the CVAE framework with the global receptive field, and each position of the GVT can directly access the latent information via the multi-head self-attention mechanism. However, we still observe that the GVT suffers the *vanishing latent variable problem* as RNN-based CVAE because the decoder can bypass the latent information by paying less attention to the SOS token. Hence, we apply the *KL annealing*, and *bag-of-word* auxiliary loss $\mathcal{L}_{bow}$ as in (Zhao et al., 2017; Zhou and Wang, 2018) to preserve the useful information of the latent variable. Therefore, the learning objective of the GVT is defined as follows:

$$\mathcal{L} = \mathcal{L}_{ELBO} + \mathcal{L}_{bow}. \quad (4)$$

4 Sequential Variational Transformer

In order to augment the capacity of the latent variable with multi-modal distributions and to better utilize the latent information, we further explore incorporating a sequence of latent variables in decoding process. We introduce Sequential Variational Transformer (SVT) with a novel variational decoder layer which generate latent variables for each position: $z = (z_1, \ldots, z_T)$. Similar to Goyal et al. (2017), we interpret the latent variables as a generation plan for the future sequence. Unlike previous CVAE models which use an extra encoder to encode the response separately (Zhao et al., 2017; Zhou and Wang, 2018) or use a backward RNN to encode the future sequence for each time step (Goyal et al., 2017; Du et al., 2018), SVT uses a *Non-causal Multi-head Attention* which leaks the future information to the recognition network for computing the posterior latent variables.

As shown in Figure 2, the SVT shares the same encoder as the standard Transformer (Vaswani et al., 2017), while its decoder consists of a variational decoder layer followed by a stack of N standard Transformer decoder layers. The variational decoder layer has two paths for computing the posterior latent variable and prior latent variable respectively. We denote them as *Posterior Path* and *Prior Path*.
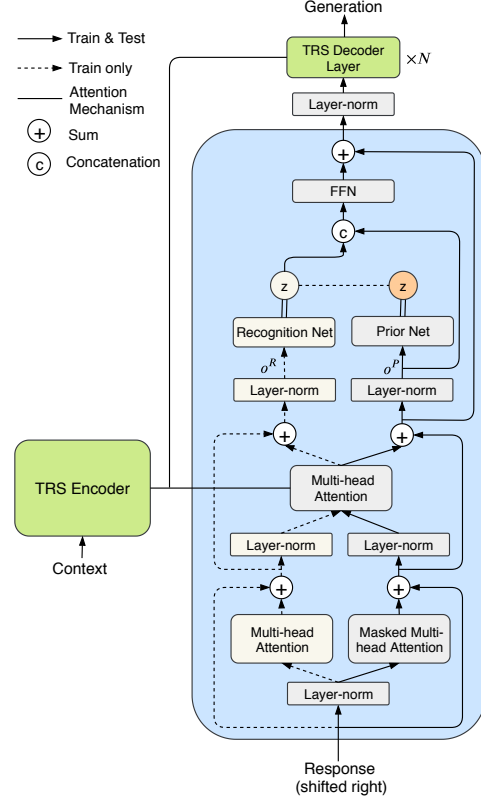

Figure 2: The Sequential Variational Transformer. During training, The posterior latent variables z by the posterior network are passed to the decoder, while during testing, the target response is absent, and z is replaced by the prior latent variables z. The word embeddings, positional encoding, softmax layer and meta vectors are ignored for simplicity

4.1 Prior Path

The *Prior Path* (solid line in Figure 2) has a masked multi-head self-attention sub-layer which performs causal attention on the shifted response, followed by a multi-head self-attention sub-layer which performs encoder-decoder multi-head attention on the context encoder. The last sub-layer is composed of a MLP prior network which approximates a sequence of prior latent variable for each position, and a Position-wise Feed-Forward Network (FFN) which fuse the latent information z with the observed information representation o^P before the prior network (shown in Figure 2). Specifically, we concatenate o^P with z as the input to the FNN, and the FNN pass the fused representation to the next layer. Same as Vaswani et al. (2017), in the variational decoder layer, each sub-layer is followed by a residual connection and layer normalization. That is, the output of each sub-layer is $LayerNorm(x + Sublayer(x))$.

We decompose the response x as $x = (x_1, \cdots, x_T)$ and the latent variable z as $z = (z_1, \ldots, z_T)$. The prior model produces latent variables at each position z_t by not only conditioning on the input condition c (the concatenation of context and meta features), but also conditioning on the **observed response tokens** $x_{1:t-1}$. By assuming z_t follows a multivariate Gaussian distribution, the prior model becomes:

$$p_\theta\left(z_t|c, x_{1:t-1}\right) = \mathcal{N}\left(z_t; \mu_t', \sigma_t'\right), \qquad (5)$$

where

$$[\mu_t', \log \sigma_t'] = MLP(o^P).$$

4.2 Posterior Path

The only difference between the *Posterior Path* (dash line in Figure 2) and *Prior Path* is that the mask is removed from the masked multi-head attention. Thus the **masked (casual) multi-head attention** become **non-casual multi-head attention**, which allows each position to attend to the subsequent positions. Then, the second multi-head attention sub-layer (shared the same weight with prior path) performs posterior attention on the encoder and passes the posterior observed information o_R to the recognition network. The recognition network produces the posterior latent variable for each position z_t as:

$$q_\phi\left(z_t|c, x\right) = \mathcal{N}\left(z_t; \mu_t, \sigma_t\right), \qquad (6)$$

where

$$[\mu_t, \log \sigma_t] = MLP(o^R).$$

During the training, the posterior path guides the learning of prior path via KL divergence constraint:

$$\mathcal{L}_{KL} = \sum_t KL\left(q_\phi(z_t|c, x)\|p_\theta(z_t|c, x_{1:t-1})\right) \qquad (7)$$

In the training phase, the posterior latent variables from Equation 6 are passed to the FFN, while in the testing phase the *Posterior Path* will be blocked and the posterior latent variables will be replaced with the prior latent variables from Equation 5.

During the decoding process, each response token x_t is generated by conditioning on observed response tokens $x_{1:t-1}$, latent variables $z_{1:t}$, and the input condition c. The decoding process of the SVT is:

$$p_\theta(x|z, c) = \prod_t p_\theta\left(x_t|z_{1:t}, x_{1:t-1}, c\right). \qquad (8)$$

4.3 Auxiliary Loss

As we expect the latent variables to be a generation plan for the future sequence, we inject such bias into latent variables by using an auxiliary loss: *Sequential-Bag-of-Word (SBOW)* which proposed by Du et al. (2018). The idea of the *SBOW* auxiliary objective is to sequentially predict the bag of succeeding target words $x_{t:T}$ by using latent variable z_t. In our case, the succeeding words prediction also leverages the observed information c and $x_{1:t-1}$. Thus the auxiliary loss at each position is computed by:

$$p_\xi(x_{t:T}|z_t, c, x_{1:t-1}) = f_{aux}(z_t, o^P) \qquad (9)$$

where f_{aux} is a feed-forward neural network with the softmax output.

4.4 Learning

The evidence lower bound (ELBO) objective of SVT is the sum of the reconstruction loss $\mathcal{L}_{REC}(t)$ and Kullback-Leibler divergence loss $\mathcal{L}_{KL}(t)$ at each position:

$$
\begin{aligned}
\mathcal{L}_{ELBO} &= \sum_t \mathcal{L}_{REC}(t) - \mathcal{L}_{KL}(t) \\
&= \sum_t \mathbf{E}_{q_\phi(z_t|c,x)}\left[\log p_\theta(x_t|z_{1:t}, x_{1:t-1}, c)\right] \\
&\quad - KL\left(q_\phi(z_t|c,x)\|p_\theta(z_t|c, x_{1:t-1})\right).
\end{aligned}
$$
$$(10)$$

We regularize the ELBO learning objective with an auxiliary loss $\mathcal{L}_{sbow}$ to enhance the expressiveness of the latent variables. Therefore, the final learning objective is formulated as follows:

$$\mathcal{L} = \mathcal{L}_{ELBO} + \mathcal{L}_{sbow}, \qquad (11)$$

where,

$$\mathcal{L}_{sbow} = \sum_t \mathbf{E}_{q_\phi(z_t|c,x)}\left[\log p_\xi(x_{t:T}|z_t, x_{1:t-1}, c)\right]. \qquad (12)$$

5 Experiments

5.1 Dataset

We evaluate the proposed models on three conversationet dataset such as MojiTalk (Zhou and Wang, 2018), PersonaChat (Zhang et al., 2018), Empathetic-Dialogues (Rashkin et al., 2019).

MojiTalk dataset consists of 596,959 post and response pairs from Twitter. Each response is labeled by one emoji which indicates the response emotion. There are 64 emoji labels in total with unbalanced distribution. We use the preprocessed data and vocabulary released from Zhou and Wang (2018) and follow the same split of train/validation/test set.

PersonaChat & Empathetic-Dialogues are one-to-one multi-turn conversation datasets. In PersonaChat (Persona), the conversations are revolve around personas which are established by four to six persona sentences. While in Empathetic-Dialogues (ED), the conversation are mostly about situation that happened to one of the speaker and another speaker is trying to understand the feeling and reply accordingly. Both datasets are about modeling social skills and the goal is to make user more engaging. Therefore, we combine the train/validation/test set of two datasets.

5.2 Baselines

We compare the proposed models with the following baselines:

Seq2Seq. An attention-based sequence-to-sequence model with the emoji vector as additional input as discribed in MojiTalk (Zhou and Wang, 2018).

CVAE. An RNN-based conditional variational autoencoder for dialogue response generation (Zhou and Wang, 2018), which uses a multivariate Gaussian latent variable to model the response and concatenate it with the last hidden state of the encoder as the initial state of the decoder. KL annealing, early stopping strategy and *bag-of-word* auxiliary loss are applied during the training. We use the implementation [3] released by Zhou and Wang (2018).

Transformer. A transformer (Vaswani et al., 2017) trained by using a Maximum Likelihood Estimation (MLE) objective and can be considered as the base model for both the GVT and SVT.

5.3 Hyper-parameters and Training Setup

We use a 4-layer Transformer as our base model. The hidden size is set to be 300 everywhere, and the word embedding is initialized with the 300-dimensional pre-trained GloVe embeddings for

---

[3]The implementation of CVAE baseline: `https://github.com/claude-zhou/MojiTalk`

both encoder and decoder. The multi-head attention sub-layers are made up of 4 attention heads each with embedding dimension 64. The size of latent variable is 300. The recognition network and the prior network are parameterized by 3-layer MLPs with 512 hidden dimension. Following the training setup of Zhou and Wang (2018), we first train our baseline transformer model with the MLE objective and use it to initialize its counterparts in both GVT and SVT. Then the models are trained end-to-end by the Adam optimizer with the initial learning rate 2×10^{-4}. KL annealing and early stopping strategy are applied as in (Zhou and Wang, 2018). In the test time, we use greedy decoding strategy for all models.

5.4 Automatic Evaluation

PPL & KLD. The evaluation metrics include Perplexity (**PPL**) and Kullback-Leibler divergence between the posterior and prior (**KLD**). A well trained model should achieve a low reconstruction and small but non-trivial KL distance (Zhao et al., 2018).

Diversity. To measure the generation diversity, we calculate **Dist-1**, **Dist-2**, and **Dist-3**, the ratio of the number of distinct n-grams (unigrams, bigrams, and trigrams) over the total number of n-grams. A higher distinct n-grams ratio indicates more diverse generation.

Embeddings Similarity. This metric computes the cosine similarity between the sentence embedding of a generated sequence and that of a ground-truth response. In our experiments, we introduce two different ways to represent sentence embeddings. The first is $\mathbf{EMB_{FT}}$ (Liu et al., 2016) that calculates the average of word embeddings in a sentence using FastText (Mikolov et al., 2018) which is trained with Common Crawl and Wikipedia data. We use FastText embeddings instead of other pre-trained word embeddings because it can handle out-of-vocabulary issue. However, representing a sentence by simply taking the average of word embeddings ignores the context information. Therefore, we propose to use a pre-trained language model BERT (Devlin et al., 2018) to compute the contextualized sentence representation. Specifically, we use a pre-trained BERT to encode a generated sentence and a ground-truth response, and average the output representation of both to obtain the sentence embeddings. We denote such contextualized sentence embedding as $\mathbf{EMB_{BERT}}$.

MojiTalk

Model	PPL	KLD	Diversity			Embeddings Similarity		Human Evaluation	
			Dist-1	Dist-2	Dist-3	EMB$_{FT}$	EMB$_{BERT}$	Coherence	Emotion
Seq2Seq	130.75	-	0.0055	0.0187	0.0347	0.738	0.594	20.67	20.67
CVAE	35.33	27.55	0.0189	0.1340	0.3640	0.751	0.613	18.33	18
Transformer	72.66	-	0.0040	0.0161	0.0324	0.741	0.596	19.67	23.33
GVT	19.71	18.15	**0.0207**	**0.1524**	**0.4064**	0.753	0.609	23	22.67
SVT	**18.96**	32.27	0.0079	0.1053	0.3654	**0.762**	**0.619**	**26**	**27.67**
Human	-	-	0.0557	0.4009	0.7697	-	-	-	-

Persona + ED

Model	PPL	KLD	Diversity			Embeddings Similarity		Human Evaluation	
			Dist-1	Dist-2	Dist-3	EMB$_{FT}$	EMB$_{BERT}$	Coherence	Engagedness
CVAE	31.32	10.01	0.0186	0.1102	0.295	**0.917**	0.666	20.67	21.33
Transformer	48.03	-	0.0058	0.0237	0.0524	0.915	0.672	24.67	24.67
GVT	18.34	19.13	0.0204	0.1406	**0.3995**	**0.917**	**0.675**	20	21.33
SVT	**17.75**	24.67	**0.0213**	**0.1521**	0.3936	0.906	0.665	**38.67**	**36.67**
Human	-	-	0.0640	0.3800	0.7070	-	-	-	-

Table 1: Results of Variational Transformer compared to baselines on automatic and human evaluations.

5.5 Human Evaluation

In the human evaluation, we prepare multiple-choice questions for human evaluators and the answers are the generation results from the five models (Seq2Seq, CVAE, Transformer, GVT, and SVT). we first randomly sample 100 dialogues and their corresponding responses from our models and the baselines. For each response, we assign three human annotators to select the most coherent (on topic) response to the context (multiple answers are allowed). In addition, annotators also need to choose the best response correlated to the given emoji label in Mojitalk and the most engaging response in PersonaChat and Empathetic-Dialogues. If there is no response that satisfies the evaluators, they can choose "all answers are bad", which means none of the answer is chosen. We compute the rate that each model is chosen to quantify generation quality regarding to the human standard.

6 Results

6.1 Quantitative Analysis

The automatic evaluation results are shown in Table 1. Transformer-based models have significantly lower perplexity compared to RNN-based models which indicate that the global receptive field performed by multi-head self-attention boost the modeling capacity. However, deterministic Seq2Seq and Transformer models tends to generate generic responses which leads to a low diversity score. Meanwhile incorporating a stochastic latent variable into both models (CVAE and GVT) promote

more diverse generation results and boost the diversity scores such as **Dist-1**, **Dist-2**, and **Dist-3**.

Compare to baseline models, the GVT achieves relatively lower reconstruction **PPL**, which suggests that the global latent variable contains rich latent information (e.g., topic) for response generation. Meanwhile, the sequential latent variables of the SVT encode fine-grained latent information and further improve the reconstruction **PPL**.

On the other hand, SVT achieves the highest score in terms of two semantic relevance-oriented metrics such as **EMB$_{FT}$** and **EMB$_{BERT}$** in MojiTalk dataset, while in the combined dataset of Persona and ED, we observe performance drop of SVT compare to other models. This is because both Persona and ED are well designed and have lower entropy than MojiTalk which collected from Twitter. We hypothesize that the sequential latent variables have no advantage in term of similarity to single, fixed "gold response" when model low entropy response. Indeed, in open domain dialogue response generation, automatic metric is not always aligned with the human judgement (Liu et al., 2016). In contrast, human evaluation result reported in Table 1 demonstrates the generations of SVT are closer to the human standard in terms of coherence, invoked emotion and engagedness.

6.2 Qualitative Analysis

Table 2 compares the generation of the proposed models with baselines given the same contexts. We observe that the Seq2Seq and vanilla transformer tend to generate generic and repetitive responses

Context	trade must 've made you mad ?
Emotion	😔
Responses	**Seq2Seq:** i 'm not sure if i 'm not sure if i 'm not sure if i 'm not sure about it **CVAE:** <unk> but i don 't think it 's been on **Transformer:** i 'm not sure i 'm not **GVT:** i 'll pass it on , she 's mad **SVT:** hell yeah bro . yeah **Ref:** i don 't wanna talk about it
Context	love the smell of a good bbq ! 😊 🍴
Emotion	😁
Responses	**Seq2Seq:** i love it ! i love it ! **CVAE:** aw you 're getting better **Transformer:** i 'm glad you like it ! **GVT:** i 'd like to know you 're very well . **SVT:** omg what kind of smell ? thanks for sharing it with a pizza in <unk> **Ref:** hahaha that sounds like friday to me pc
Context	good evening , how are you tonight ?
Responses	**CVAE:** hi , i am doing well . you ? **Transformer:** i am doing well , how are you ? **GVT:** i am good . just studying some people there . **SVT:** : i am doing well , just finished eating some ice cream . **Ref:** i am doing well . i am relaxing before bed . i work in the morning .
Context	i cook mine at home while watching one tree hill . love that show .
Responses	**CVAE:** i love to cook . i like to cook . **Transformer:** i love the outdoors . i love the outdoors . **GVT:**it is good . . . you can make some money . **SVT:** do you have any pets ? i have never watched any of tv. **Ref:** i am looking for a new job . i hate sitting still all day
Context	User: well do you work ? do you have a degree to sustain yourself ? System: i built models when i was a kid . now i sculpt and mold and carve. User: nice , i specialize in computer science degree so i mostly mold 3d images.
Responses	**CVAE:** i do not like it when you get to do the same **Transformer:** i am a teacher . i am a teacher . i am a teacher . **GVT:** me too ! my favorite is the best baker . **SVT:** i love the technology . i like to play when i get older **Ref:** i am looking for a new job . i hate sitting still all day

Table 2: Generated responses from proposed models and baseline models. The reference responses (Ref) are given.

(e.g., i am not sure) in MojiTalk due to their deterministic structure fail to capture the variability in dialogue response. By incorporating stochastic latent variables, the CVAE and GVT can generate more diverse responses, but their responses are sometimes digressive (e.g., example 5). Interestingly, GVT and SVT generalize the topic beyond the context which make the dialogue more engaging (e.g., example 4). In general, SVT is able to generate more coherent and informative responses.

7 Conclusion

This paper introduces the Variational Transformer (VT), a variational self-attentive feed-forward sequence model that combines the global receptive field of a Transformer with the variational nature of a CVAE. We propose two types of the VT: 1) the Global Variational Transformer (GVT) which incorporates a global latent variable as additional input to the transformer decoder; and 2) the Sequential Variational Transformer (SVT) which generates latent variables for each position during decoding process. Quantitative and qualitative experimental results shows that our models outperform baselines in terms of diversity, semantic relevance, and human judgment. In future work, we will utilize the pre-training language models (Radford et al., 2019) as the back-bone to strengthen the language model of the VT for better generation.

References

Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. 2016. Generating sentences from a continuous space. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*.

Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel, Aaron C Courville, and Yoshua Bengio. 2015. A recurrent latent variable model for sequential data. In *Advances in neural information processing systems*, pages 2980–2988.

Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Łukasz Kaiser. 2018. Universal transformers. *arXiv preprint arXiv:1807.03819*.

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*.

Jiachen Du, Wenjie Li, Yulan He, Ruifeng Xu, Lidong Bing, and Xuan Wang. 2018. Variational autoregressive decoder for neural response generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3154–3163.

Anirudh Goyal Alias Parth Goyal, Alessandro Sordoni, Marc-Alexandre Côté, Nan Rosemary Ke, and Yoshua Bengio. 2017. Z-forcing: Training stochastic recurrent networks. In *Advances in neural information processing systems*, pages 6713–6723.

Lukasz Kaiser, Samy Bengio, Aurko Roy, Ashish Vaswani, Niki Parmar, Jakob Uszkoreit, and Noam Shazeer. 2018. Fast decoding in sequence models using discrete latent variables. In *International Conference on Machine Learning*, pages 2395–2404.

Lukasz Kaiser, Aidan N Gomez, Noam Shazeer, Ashish Vaswani, Niki Parmar, Llion Jones, and Jakob Uszkoreit. 2017. One model to learn them all. *arXiv preprint arXiv:1706.05137*.

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

Hung Le, Truyen Tran, Thin Nguyen, and Svetha Venkatesh. 2018. Variational memory encoder-decoder. In *Advances in Neural Information Processing Systems*, pages 1508–1518.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A diversity-promoting objective function for neural conversation models. In *Proceedings of NAACL-HLT*, pages 110–119.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016b. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003.

Piji Li, Wai Lam, Lidong Bing, and Zihao Wang. 2017. Deep recurrent generative decoder for abstractive text summarization. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2091–2100.

Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 2019. Moel: Mixture of empathetic listeners. 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 121–132.

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

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. 2018. Image transformer. In *International Conference on Machine Learning*, pages 4052–4061.

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

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381.

Iulian Vlad Serban, Ryan Lowe, Laurent Charlin, and Joelle Pineau. 2016. Generative deep neural networks for dialogue: A short review. *arXiv preprint arXiv:1611.06216*.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. 2015. Learning structured output representation using deep conditional generative models. In *Advances in neural information processing systems*, pages 3483–3491.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

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

Richard S Wallace. 2009. The anatomy of alice. In *Parsing the Turing Test*, pages 181–210. Springer.

Joseph Weizenbaum et al. 1966. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.

Biao Zhang, Deyi Xiong, Hong Duan, Min Zhang, et al. 2016. Variational neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 521–530.

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

Tiancheng Zhao, Kyusong Lee, and Maxine Eskenazi. 2018. Unsupervised discrete sentence representation learning for interpretable neural dialog generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1098–1107.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 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)*, pages 654–664.

Xianda Zhou and William Yang Wang. 2018. Mojitalk: Generating emotional responses at scale. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1128–1137.