Dual-View Conditional Variational Auto-Encoder for Emotional Dialogue Generation

MEI LI, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, School of Artificial Intelligence, University of Chinese Academy of Sciences
JIAJUN ZHANG, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences,School of Artificial Intelligence, University of Chinese Academy of Sciences
XIANG LU, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, School of Artificial Intelligence, University of Chinese Academy of Sciences
CHENGQING ZONG, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences and School of Artificial Intelligence, University of Chinese Academy of Sciences

Emotional dialogue generation aims to generate appropriate responses that are content-relevant with the query and emotion-consistent with the given emotion tag. Previous work mainly focuses on incorporating emotion information into the sequence to sequence (Seq2Seq) or conditional variational auto-encoder (CVAE) models, and they usually utilize the given emotion tag as a conditional feature to influence the response generation process. However, emotion tag as a feature cannot well guarantee the emotion consistency between the response and the given emotion tag. In this paper, we propose a novel Dual-View CVAE model to explicitly model the content-relevance and emotion consistency jointly. These two views gather the emotional information and the content-relevant information from the latent distribution of responses respectively. We jointly model the dual-view via VAE to get richer and complementary information. Extensive experiments on both English and Chinese emotion dialogue datasets demonstrate the effectiveness of our proposed Dual-View CVAE model, which significantly outperforms the strong baseline models in both aspects of content relevance and emotion consistency.

CCS Concepts: • **Computing methodologies → Natural language processing**; **Natural language generation**; *Discourse, dialogue and pragmatics.*

Additional Key Words and Phrases: sentiment, dialogue, neural networks

ACM Reference Format:
Mei Li, Jiajun Zhang, Xiang Lu, and Chengqing Zong . 2018. Dual-View Conditional Variational Auto-Encoder for Emotional Dialogue Generation. *J. ACM* 37, 4, Article 111 (August 2018), 18 pages. https://doi.org/10.1145/1122445.1122456

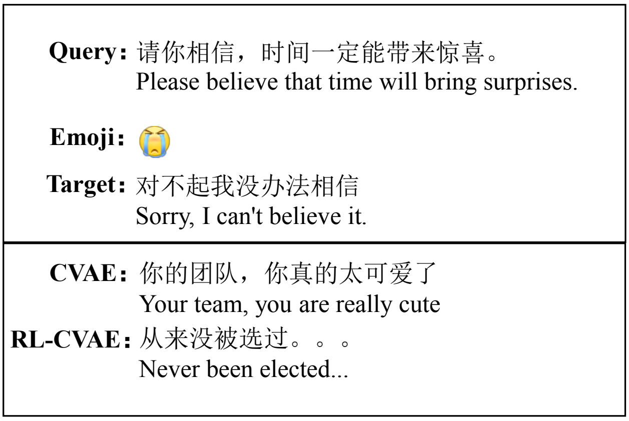
Authors' addresses: Mei Li, mei.li@nlpr.ia.ac.cn, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, School of Artificial Intelligence, University of Chinese Academy of Sciences, Intelligence Building, No. 95, Zhongguancun East Road, Haidian District, Beijing, China, 100190; Jiajun Zhang, jjzhang@nlpr.ia.ac.cn, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences,School of Artificial Intelligence, University of Chinese Academy of Sciences, Intelligence Building, No. 95, Zhongguancun East Road, Haidian District, Beijing, China, 100190; Xiang Lu, lu.xiang@nlpr.ia.ac.cn, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, School of Artificial Intelligence, University of Chinese Academy of Sciences, Intelligence Building, No. 95, Zhongguancun East Road, Haidian District, Beijing, China, 100190; Chengqing Zong, cqzong@nlpr.ia.ac.cn, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences and School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China, 100190.

0004-5411/2018/8-ART111 $15.00
https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

For a long time, building an excellent conversation system which can converse with humans naturally and intelligently has attracted wide attention in both academia and industry. In recent years, there has been a surge of interests towards designing large-scale non-task-oriented dialogue systems using sequence to sequence neural networks [27–29, 35, 44]. However, these studies mainly focus on modeling the diversity, topic and the personality of the response, neglecting the importance of the emotion in dialogue. Previous studies have shown that emotional intelligence has a crucial influence on the performance of human-computer dialogue system and is an indispensable part of successful dialogue systems [23, 26].

To alleviate this issue, several studies have already made contribution to the emotional dialogue generation task and emotional text generation [10, 12, 19, 20, 30, 33, 41, 46, 49]. Existing emotional dialogue generation methods can be mainly divided into two categories: coarse-grained and fine-grained emotion models. The coarse-grained model are based on limited emotion categories, such as three categories (e.g. positive, negative and neutral) or five categories (e.g. like, happy, sad, disgust and angry). These models are easy to train relatively as they can use a handful of existing resources in the field of sentiment analysis. However, such methods are hard to incorporate the rich emotions of humans whose emotion types are far more than three or five categories. Apart from this, human annotation is still costly. The fine-grained emotion model employs the emojis as the emotion categories. Nowadays hundreds of emoji characters widely used in social media are borrowed to represent the human emotion. These methods can avoid the defects of human annotation and bring richer emotional data without the need for annotation.



Query: 请你相信，时间一定能带来惊喜。
Please believe that time will bring surprises.

Emoji: 😭

Target: 对不起我没办法相信
Sorry, I can't believe it.

CVAE: 你的团队，你真的太可爱了
Your team, you are really cute

RL-CVAE: 从来没被选过。。。
Never been elected...

Fig. 1. An example of CVAE and RL-CVAE generations

Zhou and Wang [49] first introduces the emoji dataset crawling from Twitter into the dialogue system. They use Conditional Variational Auto-Encoder (CVAE) to model the emotional dialogue generation, and reinforced learning is further applied to enhance the emotion consistency (RL-CVAE). However, these methods are not powerful enough to guarantee the content-relevance between the response and the query as well as the emotion-consistency between the response and the given emotion tag. Zhou et al. [46] report that the simply embedding emotion information cannot produce desirable emotional responses. Although reinforcement learning is further adopted to enhance the emotion consistency, it also causes side-effects that influence the language model performance and tend to generate non-fluent responses. As an example shown in Fig 1, the CVAE model generate response that is not well consistent with the emoji and the reinforcement learning tends to generate emotionally consistent responses, but the content is less relevant. Besides, the Reinforced CVAE heavily relies on the final result of the response's emoji classification, whereas the classification accuracy is very low since many emojis are similar and difficult to distinguish from each other. Finally, we find in the experiments that when CVAE training is sufficient, the reinforcement learning model is rarely to get a better result.

To improve the emotion consistency while maintaining the content relevance, we propose a novel Dual-View CVAE (DV-CVAE) model, which explicitly takes the **content relevance view** and the **emotion consistency view** into consideration. Our intuition is that the aspects in the response that query and emoji concerned with are usually different and the given emotion tag cannot only be used as a conditional feature, but also can bring more emotional relevant information. We learn the representation distribution of the content and emotion of the response by using the Variational Auto-Encoder (VAE) with different attention models. In terms of the content relevance, we first map the query and response into two different representation distributions with VAE, and then require the content distributions of the response and the query to be as close as possible. As for the emotion consistency, we employ the intermediate result of the emoji classifier, attention weighted sentence representation, as the emotion representation of the response, and then ensure the emotion distribution distance between the response and the given emoji tag to be as small as possible. In the test phase, we sample from the content distribution and the emotional distribution separately to obtain content-related information and emotion-related information.

To test the proposed model on more scenarios in addition to the emoji dataset from Twitter, we collect a large-scale Chinese emoji conversation data from Chinese social networking platform Weibo[1]. We conduct the experiments on both Chinese and English emoji datasets, and the results show that our model performs significantly better than the strong baselines. In summary, our contributions are three-fold can listed as below:

- We propose a novel dual-view CVAE model, which explicitly models both of the content relevance and emotion consistency at the same time.
- We construct a large-scale dataset of Chinese conversation pairs[2] that naturally use fine-grained emojis, which will be publicly available.
- Extensive experiments on both Chinese and English emoji datasets demonstrate the superiority of our proposed model.

The rest of this article is organized as follows. Section 2 summarizes the related work. In section 3, we introduce the background. Our proposed model will be detailed in section 4. Section 5 provides the experiments and analysis. We finally conclude this work in section 6.

2 RELATED WORK

In recent years, the conversation system has grown rapidly in a variety of areas [28, 36, 37, 44], among which generating an emotional response has gained more and more attention [10, 16, 31, 41, 46]. Ghosh et al. [10] proposed the emotional language model, which adds emotion categories and emotional intensity into the language model and generates emotional sentences. Zhou et al. [46] controls the generation of responses with different emotions, feeds the embedded emotion category to the decoder and captures the implicit change of the internal emotional state to balance the weights between the grammar state and the emotion state, and lastly uses an external memory module to select the word explicitly. Asghar et al. [1] proposed three strategies for affective dialogue generation: (a) augment traditional word embeddings with a 3D affective space by using an external

[1] https://weibo.com

[2] https://drive.google.com/open?id=1diyfF9B1q4JsfhuxYF6KxSEJoTTysOdb

cognitively engineered affective dictionary; (b) design affective training loss functions; and (c) apply affectively diverse beam search. Xu et al. [41] adopted multi-task learning and dual-attention for generating emotional controllable response. Additionally, Wang and Wan [33] proposed a novel framework SentiGAN to address the lack of diversity and mode collapse problems. Colombo et al. [6] presented an affect-driven dialog system by modeling emotions at a word and sequence level.

However, these work are mainly based on coarse-grained emotions and it is hard to transfer to fine-grained datasets. As mentioned above, they usually use external engineered affective dictionaries. When there are hundreds of emotional categories, it is almost impossible to construct such dictionaries. In addition, the coarse-grained emotional categories are difficult to express the rich emotions of human beings. The emotional categories people used in daily life are far from three or five. Nowadays, people widely use emojis to express their rich emotion on social media platforms in daily life. These visual symbols can be used to convey emotions, underlying information and provide extra information for the semantics of sentences. Research has shown that considering emojis in sentiment analysis can help improve sentiment analysis task [2]. People have used emojis to do a lot of work in the task of sentiment analysis, such as using emoji-related representation to assist the task of emotion classification, building datasets, building emotion dictionaries, and so on [8, 24]. Besides, recent work [3, 17, 47] have investigated the relation between words and emojis, predicting which emojis are evoked by text messages.

Recently, Zhou and Wang [49] combined emoji prediction with the task of emotional response generation, by using reinforcement learning and pretrained classifier. They use a pretrained classifier to fine-tune the pretrained CVAE model with a hybrid objective. However, it is still difficult for RL-CVAE to weigh the content relevance and emotion consistency. Although the RL model improves consistency, it sacrifices the fluency of the sentences even though a hybrid objective function of reinforcement learning and variational lower bound is considered. Besides, the accuracy of the classifier is very low because the difference between some categories is actually insignificant. Directly using the result of the classifier is hard to achieve the desired effect. In contrast, we propose a method that leverages the intermediate results of the classifier to model the emotion consistency between the response and the given emoji tag. We consider both content relevance and emotion consistency in a dual-view CVAE model.

Multi-view or Dual-View methods have gained great success in many tasks [5, 39, 48]. Xie and Ma [39] leverage sentence-level and word-level features for text matching task, and use VAE to encode sentences into latent codes. Similarly, Zhou et al. [48] proposed a multi-view response selection model that also utilize word sequence view and utterance sequence view. Wu and Wu [38] proposed a simple dual decoder to model positive and negative emotions respectively, but it was difficult to construct positive and negative responses for the same sentence. And this kind of method is difficult to extend to multiple classes. Okamoto et al. [25] proposed a dual variational autoencoder for generating images corresponding to multiclass labels, which condition with latent vectors that include label information. They assume that the latent vector is a linear combination of vectors of latent space and the dual latent space. Unlike the above method, we propose a dual-view CVAE model, which model emotion view and content view by using the CVAE model separately.

3 BACKGROUND

3.1 VAE and CVAE

The Variational AutoEncoder (VAE) [7, 15] is a directed graphical model with continuous latent variables, and it is widely used in the generation task of image and natural language. Different from traditional autoencoder, the VAE encodes an input x into a probability distribution, then reconstructs the original input with a decoder network by sampling a continuous latent variable

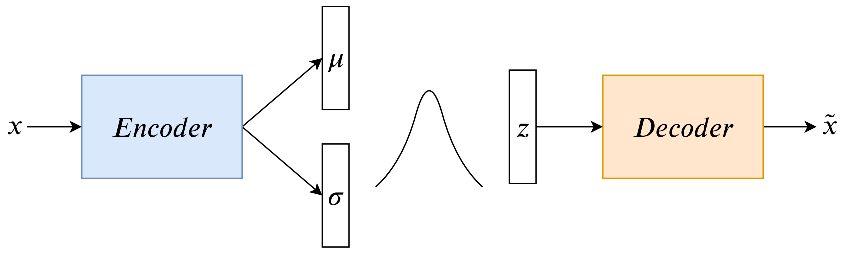


Fig. 2. The VAE architecture

z from this probability distribution, as illustrated in Fig 2. A formal description of the problem is as follows. Let x be an observation of random variable, taking values in $\mathcal{D}$. We assume that the generation of x involves a continuous latent variable z, taking values in $\mathcal{Z}$, by means of a point density $p_\theta(x, z)$, parametrized by θ. Given a set of observed data points $\mathcal{D} = \left\{x^1, \ldots, x^n\right\}$, the goal of maximum likelihood estimation is to eatimate the parameters θ that maximize the marginal log-likelihood $\mathcal{L}(\theta; \mathcal{D})$:

$$\theta^* = \arg\max_\theta \mathcal{L}(\theta; \mathcal{D}) = \sum_{i=1}^{n} \log \int_z p_\theta\left(x^i, z\right) \mathrm{d}z \tag{1}$$

Due to the integration over the latent variables, it is intractable to directly compute or differentiate the marginal log-likelihood. A common approach is to maximize a variational lower bound on the marginal log-likelihood by introducing an approximate posterior $q_\phi(z|x)$:

$$\begin{aligned} \log p_\theta(x) &\geq \mathbb{E}_{q_\phi(z|x)}\left[\log \frac{p_\theta(x, z)}{q_\phi(z|x)}\right] \\ &= \log p(x) - D_{KL}\left(q_\phi(z|x) \| p(z|x)\right) \\ &= \mathcal{L}(x; \theta, \phi) \end{aligned} \tag{2}$$

where KL denotes the Kullback-Leibler divergence. The evidence lower bound (ELBO) can be also rewritten as a minimum description length loss function:

$$\mathcal{L}(x; \theta, \phi) = \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x|z)\right] - D_{KL}\left(q_\phi(z|x) \| p_\theta(z)\right) \tag{3}$$

where the neural network with parameters ϕ, called "recognition" model, is introduced to approximate the true posterior $p_\theta(z|x)$. Another neural network with parameters θ, which is represented as $p_\theta(x|z)$, is aim to reconstruct the data. In general, we assume that $q_\phi(z|x)$ is a multivariate diagonal Gaussian distribution:

$$q_\phi(z|x) = \mathcal{N}\left(z; \mu_\phi(x), \Sigma_\phi(x)\right),$$

For particularly simple parametric forms of $q_\phi(z|x)$, one can backpropagate through the sampling process $z \sim q_\phi(z|x)$ by applying the reparametrization trick, which first samples $\epsilon \sim \mathcal{N}(0, I)$, and then computes $z = \mu_\phi(x) + \sum_\phi^{\frac{1}{2}}(x) * \epsilon$. As a result, the VAE can be trained efficiently using stochastic gradient descent (SGD). This is essential in VAE training.

The conditional VAE (CVAE), is a modification of VAE based on certain attributes, e.g. generating different human faces given skin color, gender, age and so on [42], or generating different sentences given sentiment topic and so on. The formula is as follows, the c in this formula is condition:

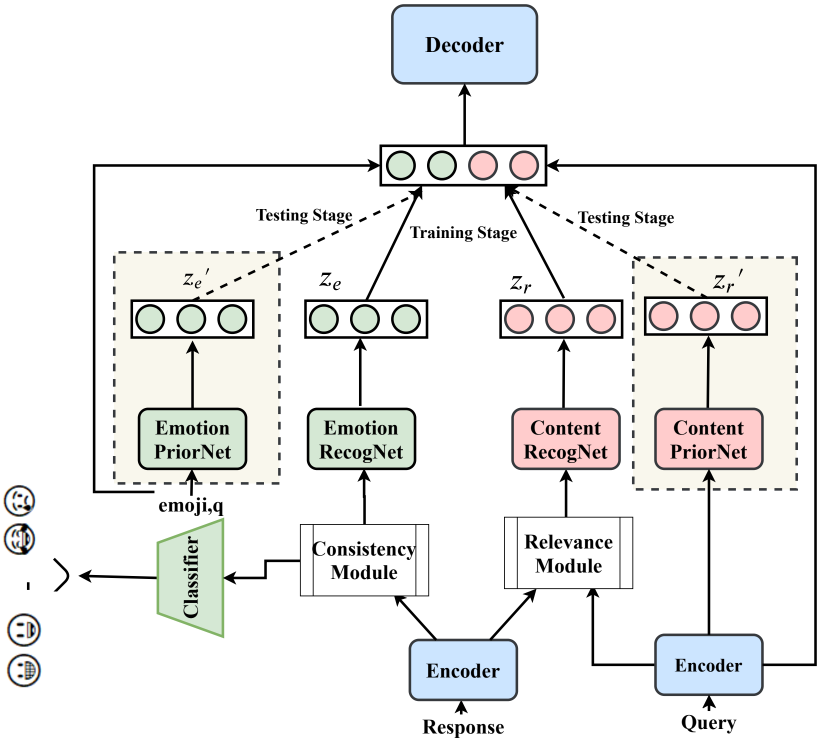


Fig. 3. The overview of the Dual-View CVAE architecture. Each utterance in query and response is encoded into a dense vector through Bi-LSTM and then mapped into the latent variables. The left green part is associate with emotion representation supervised by an emoji classifier and the right pink part is the content relevance module. The decoder predicts one token at a time using RNN.

$$\mathcal{L}_{CVAE}(x, c; \theta, \phi) = \mathbb{E}_{q_\phi(z|x,c)} [\log p_\theta(x|z, c)] - D_{KL}(q_\phi(z|x, c)||p_\theta(z|c)) \tag{4}$$

Variational encoder-decoders have shown promising results in text generation [4, 34, 43]. Straightforwardly optimizing with Equation 4 suffer from the KL-vanishing problem that the RNN part ends up explaining all the structures without making use of the latent representation. Much meaningful work has been done to alleviate this problem [11, 14, 45]. When dealing with text generation, CVAE model can generate more diverse sentences than seq2seq model. However, in emotional dialogue generation task, general CVAE model is not powerful enough to be consistent with the corresponding emotion. Our proposed method employs CVAE as the baseline to accommodate the fine-grained emotions.

4 DUAL-VIEW CVAE

In this paper, we propose a dual-view model to generate responses according to content relevance and emotion consistency simultaneously. Traditional CVAE models ignore that the query and emoji have different contributions to generating the response: the query is content-related to its response, while the emoji mainly focuses on the generation of the emotional expression. By modeling the response from these two perspectives, our proposed model can generate a content related and emotion consistent response. More specifically, we design a dual-view CVAE model by transforming the traditional CVAE's latent variable z into content relevance variable z_r and emotion consistency variable z_e. The detailed formula is as follows:

$$\begin{aligned}\mathcal{L}_{DV-CVAE}(x,c;\theta,\phi) = &-\text{KL}(q_\phi(z_r|x,c)||p_\theta(z_r|c)) \\ &-\text{KL}(q_\phi(z_e|x,c)||p_\theta(z_e|c)) \\ &+\mathbb{E}_{q_\phi(z_r,z_e|x,c)}\left[\log p_\theta(x|z_r,z_e,c)\right]\end{aligned} \tag{5}$$

where c is the context vector, corresponding to the query history and emoji tag in this model. x is the output response. The output x is generated from the distribution $p_\theta(x|z_r,z_e,c)$. z_r and z_e are latent variables sampled from content Gaussian distribution and emotion Gaussian distribution. More details about z_r and z_e are described in section 4.1 and 4.2. The network named Prior network is to approximate $p_\theta(z|c)$, used in testing time and Recognition network $q_\phi(z|x,c)$ is introduced to approximate the true posterior distribution used in training time.

Fig 3 is an overview of our model. Firstly, the two encoders, Bi-LSTM, transform query and response into fixed-size vectors respectively, by concatenating Bi-LSTM's last bidirectional hidden states. After that, the content relevance module takes query vector and all hidden states of the response encoder as input, getting the content-related response representation by calculating a weighted sum of all hidden states, where weights are query's attention over response hidden states. Then, we map this representation into Gaussian distribution and use the reparametrization trick [15] to obtain samples of z_r , which can be conducted with a recognition network (training) or prior network (testing). As for the emotion consistency module, we only take hidden states of the response as input, and obtain the emotional representation from the hidden state of an emoji classifier's output layer. Then, we encode the emotional representation into a probability distribution, and drawing samples of z_e from the learned emotional distribution with a recognition network (training) or prior network (testing). Finally, we reconstruct the response sequence with the dual-view representation via a decoder network.

4.1 Content Relevance Modeling

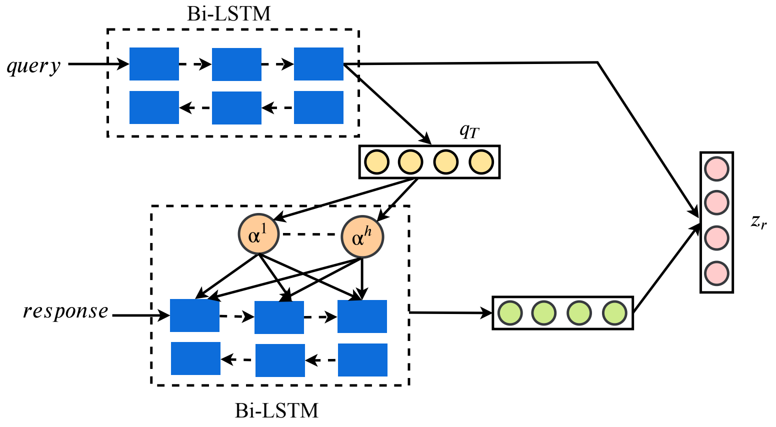


Fig. 4. Relevance module for Dual-View CVAE model

The content relevance module uses the LSTMs model with a multi-head attention mechanism [32] to model the semantics of query-response pairs and predict their relatedness. This attention mechanism is designed to focus on the phrases of a response that are connected to the query content. This mechanism allows the model to learn pair-specific representations that are more effective at predicting responses, which better contributes to the predictions of key information in the inference process. Instead of performing a single attention function, Vaswani et al. [32] found it is beneficial to capture different context with multiple individual attention functions. More specifically, multi-head attention model first transforms Q, K, and V into H subspaces, with different, learnable linear projections.

Aiming at building pairwise relevance directly, multi-head attention calculates attention weights between each pair of elements in all input. Fig 4 demonstrates an overview of our content relevance model. Query interacts with response, resulting in a weighted vector representation (the green vector) for generating content-related latent vectors(z_r). In order to contain sentence-level information, we first encode the query and response into fixed-size vectors using Bi-LSTM. The input contains the last hidden state of query q and all hidden states of response which are expressed as $x = (x_1, x_2, ..., x_n)$. Formally, for an H heads model, the h-th head can be computed as:

$$out^h = \sum_{i}^{n} \alpha_i^h (x_i W_V^h) \tag{6}$$

where W_V^h means the h-th trainable parameter matrices. α_i^h is the attention weight between the q and j-th response elements. It can be calculated as:

$$\alpha_i^h = \frac{exp(e_i^h)}{\sum_{j=0}^{n} exp(e_j^h)} \tag{7}$$

and

$$e_j^h = \frac{(qW_Q^h)(x_j W_K^h)^T}{\sqrt{d_k}} \tag{8}$$

in which W_Q^h,W_K^h are parameter matrices, d_k represents the dimensionality of its subspace. Finally, we get the concatenation of all weighted response .

$$out = [out^1, out^2, ..., out^H] \tag{9}$$

4.2 Emotional Consistency Modeling

This part mainly introduces the emotional consistency module, which is the green part on the left side of Fig 3. Fig 5 demonstrates a detail structure of our emotion consistency model. Our goal is to incorporate the emotion features into the basic CVAE model, and we first introduce self-attention described in Xu et al. [40] into CVAE. Generally speaking, the attention score partly reflects the sentiment contribution of each word in a well-trained sentiment classifier. Emotional words tend to get a higher score in sentences. Therefore we use the attention scores to denote the emotion representation in response. we encode this emotional representation into a probability distribution which is supervised by emoji tag, and drawing samples of z_e from the learned emotional distribution with a recognition network in the training time. During inference, we use the emoji vector to predict the emotion-related representation directly. In this part, we did not use the multi-head attention mechanism, because the multi-head attention mechanism has more parameters, and the redundant attention weights are a kind of noise for extracting emotional information. The experimental results show that the self-attention mechanism is better than the multi-head attention mechanism. The details of sentiment classifier are described as follows.

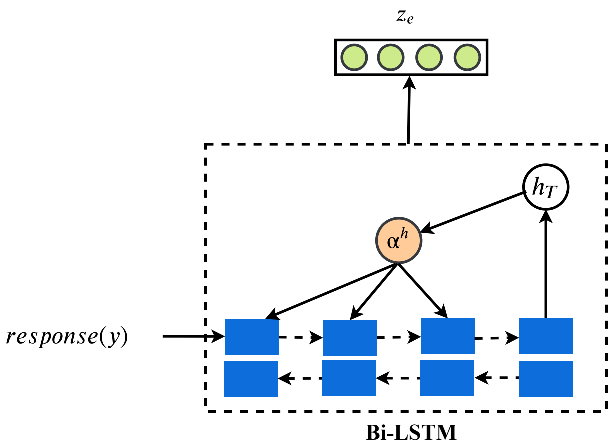


Fig. 5. Emotional Consistency module for Dual-View CVAE model

This module assumes that the last hidden state of encoder represents the information of the whole sentence, and then uses this last hidden state to compute the attention weights with all previous hidden states. This is why it is called self-attention. Finally, the weighted vectors are mapped to softmax layer to calculate the final classification distribution. The final sentence representation before softmax layer is:

$$out = \sum_{i=0}^{n} \alpha_i h_i \tag{10}$$

where h_i is the hidden state of the Bi-LSTM in the i-th time step. α_i is the attention weight computed as:

$$\alpha_i = \frac{exp(e_i)}{\sum_{i=0}^{T} exp(e_i)} \tag{11}$$

where $e_i = f(h_i, h_T)$ is an alignment module and h_T is the final hidden state of sentences which contains all information of an input sentence.

5 EXPERIMENT AND ANALYSIS

5.1 Dataset

Zhou and Wang [49] constructed a dataset by collecting data from the Twitter website, annotated as MojiTalk dataset in this paper. The MojiTalk dataset contains 64 common emojis as labels. The corpus consists of 596,959/32,600/32,600 conversation pairs for train /validation/test set. However, the emoji category of this dataset is extremely unbalanced, and the top-1 category accounts for more than 30% of the total data. Moreover, this dataset is an English dataset and no Chinese dataset is available. To test the model in more scenarios than Twitter's emoji dataset, we collected a large-scale Chinese emoji conversation data from the Weibo.

To construct a large-scale emoji-enriched post-comment conversation dataset, we first crawl hundreds of millions of post-reply conversation pairs on Sina Weibo. Weibo is a popular Twitter-like microblogging service in China, on which a user can post short messages and comments on others' post. The most popular emoji expressions in Weibo are in the form of pictures. Each picture has

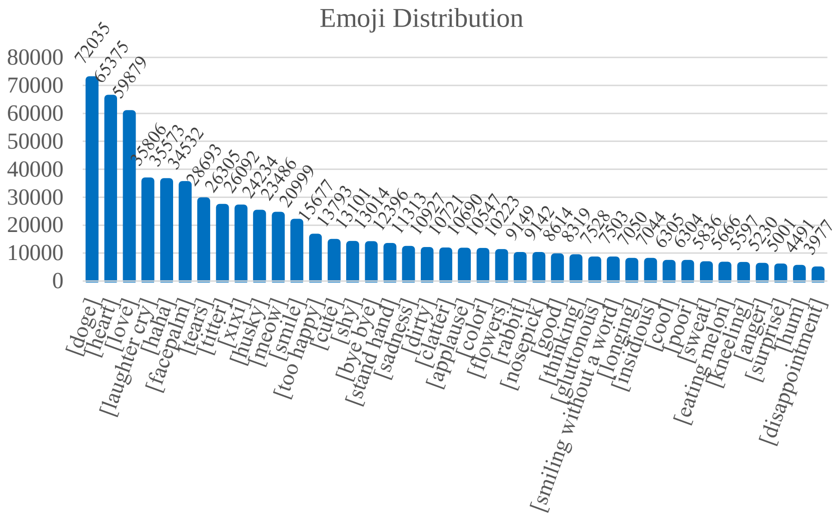


Fig. 6. Emoji distribution in Weibo dataset

a tag like "[haha]" or "[doge]", so we use the emoji tag instead of emoji Unicode. During data pre-processing, we remove those data pairs whose responses do not contain any emoji. Then we filter out potential advertisements and a large number of repetitive high-frequency responses such as "[heart][heart][heart]". We finally select 40 most frequent emoji labels for our experiments. Fig 6 shows some statistics of the dataset used in this work. Similar to the extraction method in Zhou and Wang [49], we select the emoji in the response as the tag. If there are multiple types of emoji in a response, we use the emoji with most occurrences inside the response. If there are multiple Emoji with the same frequency, then we choose emoji with lower frequency in the whole corpus. Finally, we select 698167 training data, while the validation and test data are 9966 and 5200 respectively.

We also use a high-quality coarse-grained dataset NLPCC2017 Emotional Conversation Generation (ECG) dataset[3]. This dataset has 1,119,207 post-response pairs and each post and response are annotated with 6 emotion labels (*happy, angry, sad, like, disgust, Other*).

5.2 Evaluation Measures

Automatic evaluation. Automatic evaluation of the open domain generation dialogue model is still an open research challenge [21]. We use both quantitative metrics and human judgment to evaluate the proposed models. The quantitative metrics include not only PPL (perplexity) but also the DIST-n which is a kind of metric to evaluate the degree of diversity of generated responses. PPL explicitly measures the model's ability to interpret the syntactic structure of dialogue and each utterance. Lower PPL is indicative of a better model. DIST-n calculates the percentage of the distinct n-grams in all the n-grams of the generated responses [18]. We calculated the DIST-1 and DIST-2 scores that measure the degree of the unigram and bigram diversity.

Human Evaluation. In the human judgment, we ask three judges to evaluate 100 random items, each of which consists of a query, a given emoji and two generated responses from different models. We evaluate three aspects including relevance, consistency and overall evaluation respectively. We present our results from these three aspects. In the overall aspect, we present the original query, emoji and generated responses, and the judges are asked to decide which one is a better reply to

[3] http://tcci.ccf.org.cn/conference/2017/

the original query and emoji. In the relevance aspect, we present the original query and generated responses, and the judges are asked to choose the one which is more relevant to the query. In the consistency setting, the emoji label is provided, and the judges are asked to select the better one that fits this emoji. Finally, we treat the 300 results as 300 cases and average the results.

5.3 Baselines

Seq2Seq. Sequence to sequence model has been successfully applied to a variety of tasks ranging from machine translation to speech recognition and dialogue generation. Here we use the traditional attention-enhanced seq2seq model [22] as our baseline. To model the emotion, the emoji tag is generated as the same way of the word embedding and is further reduced to smaller size through a dense layer. The small dense emotion vector together with the encoder output are fed to a decoder in the initial time. Response is generated step by step from the decoder.

Transformer. Sequence to sequence model solely on attention mechanisms. The emotion tags are encoded embeddings and add to the inputs of the decoder.

ECM a sequence-to-sequence model with the emotion category embeddings, internal and external memory mechanisms [46]. The external memory mechanisms of ECM needs an external manual dictionary of emotions. As far as we know, there is no fine-grained dictionary of emotions. So we compare with ECM model only on coarse-grained dataset and our mainly experiments is on fine-grained dataset.

CVAE. As described in section 3.1, the CVAE uses the emotion tag as a condition in the generation, which is similar to the Seq2Seq model. In order to get z, we use another same structure of encoder to encode the response into a dense vector and use it to generate z by using MLP. Based on the emotion latent vector , the encoder output of query and the emotion tag dense embedding, the decoder generates the response step by step.

RL-CVAE. Reinforced CVAE (RL-CVAE) [49] is a combination of the above CVAE model and reinforcement learning using the classifier's result as rewards. The formula is as follows:

$$\nabla \mathcal{J}^{'}(\theta) = \alpha(R - r)\nabla \sum_{t}^{|x|} logp(x_t|c, x_{1:t-1}) \tag{12}$$

$$\nabla \mathcal{L}(\theta) = \mathcal{L}_{CVAE} - \lambda \mathcal{J}^{'} \tag{13}$$

where α is a variant coefficient related to the rank of emoji label probability, the higher R ranks in all types of emoji label, the closer α to 0, the value of $\alpha \in [0,1]$. r is the baseline reward and R is the generation reward. λ is a balancing coefficient. This two-stage method is based on the pre-trained CVAE model.

5.4 Training Details

To verify the effectiveness of our approach, we conduct several experiments on two fine-grained datasets and one coarse-grained dataset. Prior studies have shown that pretraining can achieve better performance [49]. Therefore, we pre-train seq2seq together with self-attention emoji classifier. We perform training with the following hyperparameters: word embedding has size 128, bi-directional LSTM is used for the encoder, and the dimension of the LSTM hidden units set to 128. To be fair, the ECM's hidden units and embedding size are also set 128. The latent variable z in baselines CVAE and RL-CVAE has a size of 268. The relevance latent variable size is 64, and the latent variable size in the consistency module is 128. We use Adam learning [13] to update the gradient and clip the gradient to 5.0. Finally, we use the BOW loss for both emotional consistency module and content relevance module, along with KL annealing [4] to achieve the best performance. For the transformer model, we use a smaller Transformer than the base configuration [32]: 3 layers, 278-dimensional embedding, 507-dimensional inner layers, 2 heads and the dropout is 0.1.

5.5 Result and Analysis

Model	PPL	DIST-1	DIST-2
Seq2Seq	243.7	0.015	0.036
Transformer	129.87	0.063	0.159
CVAE	49.58	0.087	0.335
RL-CVAE	41.52	0.105	**0.451**
DV-CVAE	**25.70**	**0.109**	0.400

Table 1. Generation Perplexity and DIST-n, with Seq2Seq, CVAE, RL-CVAE and DV-CVAE results on Weibo dataset

Model	PPL	DIST-1	DIST-2
Seq2Seq	132.20	0.004	0.012
Transformer	81.05	0.040	0.116
CVAE	35.56	0.021	0.155
RL-CVAE	32.38	**0.024**	0.200
DV-CVAE	**23.60**	0.017	**0.225**

Table 2. Generation Perplexity and DIST-n, with Seq2Seq, CVAE, RL-CVAE and DV-CVAE results on MojiTalk dataset

Model	PPL	DIST-1	DIST-2
Seq2Seq	176.1	0.036	0.144
Transformer	93.34	0.061	0.222
ECM	88.4	0.085	0.203
CVAE	43.3	0.102	0.496
RL-CVAE	42.1	0.081	0.476
DV-CVAE	**33.7**	**0.105**	**0.517**

Table 3. Generation Perplexity and DIST-n, with Seq2Seq, CVAE, RL-CVAE and DV-CVAE results on NLPCC2017 dataset

Automatic Evaluation The quantitative evaluation results are shown in table 1, table 2 and table 3. We use the PPL to test the fluency of generated sentences. The DIST-n metric indicates the diversity of generated sentences. Intuitively a good model should achieve low perplexity and high DIST-n scores. From the results, we can see that our Dual-View CVAE outperforms baselines in terms of PPL and distinct measures. Our model achieves lower PPL on both fine-grained datasets and coarsed-grained dataset. These results demonstrate that our model maintains good fluency while generating texts with emotions. In terms of DIST-1 metric, our model performs best on Weibo dataset. Regarding to DIST-2, our model works best on MojiTalk dataset. This indicates that the response diversity generated by our model is at least on par with the baselines, but the fluency is much better than the baselines. In addition, we found that RL-CVAE is very prone to model collapse and generate meaningless continuous non-repetitive words. This result greatly increases the value of DIST-n, but the fluency of generated sentences is affected. RL-CVAE model is trained on the basis of pre-trained CVAE model, while CVAE model is trained on pre-trained Seq2Seq model. The results of CVAE model pre-training greatly affect the results of RL model. When the CVAE model converges to a certain extent, the RL model can not be improved at all. CVAE is relatively stable whose effect is better than ECM model, and our model is superior to CVAE and does not require additional pre-training beyond Seq2Seq.

Models	Weibo Data		MojiTalk Data	
	top-1	top-5	top-1	top-5
CVAE	17.2%	46.2%	29.8%	53.8%
RL-CVAE	**18.3%**	45.8%	30.5%	54.7%
DV-CVAE	14.2%	**59.9%**	**31.6%**	**56.0%**

Table 4. Emoji generation accuracy

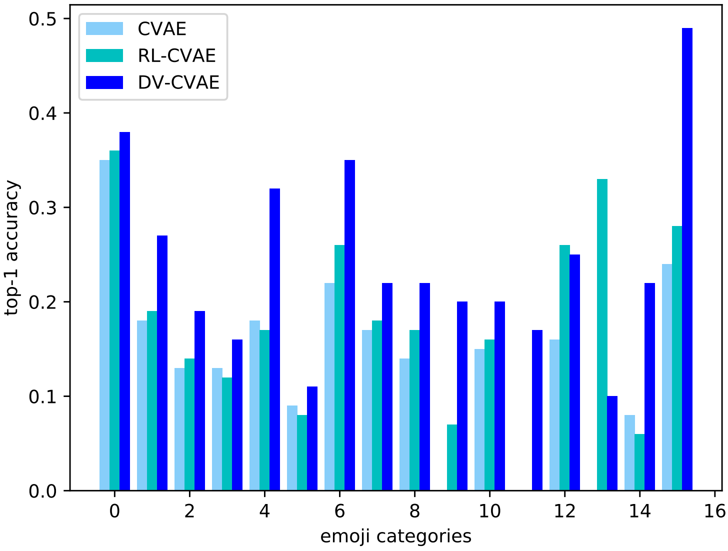


Fig. 7. Top1 emoji accuracy of the first 16 emoji labels in MojiTalk dataset.

Considering the consistency between the emotion category of the generated sentence and the given emoji label, we also performed an experiment on the classification of the generated sentences on fine-grained datasets. Since the meaning of different emojis may overlap with only a subtle difference, there are potentially multiple types of emotion in reaction to an utterance. Some automatic categorization is considered wrong, but the results of human assessment are indeed acceptable. This makes the difference of automatic and human evaluations. Specially, in Weibo dataset the “[heart]” and “[love]” which are the second and third largest categories in the dataset are similar. This afftects the top-1 accuracy of the classifier to some extent. So we also use top-5 accuracy as an automatic evaluation metric. As the results are shown in table 4, our model is also superior to baselines in emotional categories of generated sentences in MojiTalk Dataset, which further indicates that the emotional consistency module we display is effective. In Weibo dataset, our model performs better than other models on top-5 accuracy metric. In Fig 7, we drew a comparison of the accuracy of CVAE, RL-CVAE and DV-CVAE models among the top 16 emotions in the English MojiTalk dataset. In most cases, the accuracy of DV-CVAE is higher than that of the other two models.

Models	Aspect	Win	Lose	Tie
DV-CVAE VS CVAE	Overall	**48.3**%	32.6%	19.0%
	Relevance	**47.3**%	30.3%	22.3%
	Consistency	**40.7**%	33.0%	26.7%
DV-CVAE VS RL-CVAE	Overall	**47.3**%	30.7%	22.0%
	Relevance	**46.0**%	30.3%	23.7%
	Consistency	**46.0**%	31.0%	23.0%

Table 5. Results of human evaluation in Chinese Weibo dataset.

Models	Aspect	Win	Lose	Tie
DV-CVAE VS CVAE	Overall	**36.3**%	28.3%	35.3%
	Relevance	34.3%	23.6%	**42.0**%
	Consistency	36.3%	26.6%	**37.0**%
DV-CVAE VS RL-CVAE	Overall	**39.0**%	27.3%	35.3%
	Relevance	**39.0**%	25.6%	35.0%
	Consistency	**35.3**%	29.3%	**35.3**%

Table 6. Results of human evaluation in English MojiTalk dataset.

Human Evaluation At present, there is no unified and authoritative criterion to test the correlation between sentences in dialogues. Most papers use human evaluation, so we use human evaluation to test the relevance and consistency. And we calculated the Fleiss’ kappa [9] to measure the consistency among annotators. Fleiss’ Kappa values are mostly between 0.511 and 0.613, which indicate most of them are “Moderate agreement”.

As shown in table 5 and table 6, the results of human evaluation of our model are obviously better than those of baselines. Compared with the RL-CVAE model, the percentage of win in Chinese dataset is as high as 46% in relevance and consistency. In English dataset, the percentage of win is also more than 35%. In the Weibo dataset, the percentage of Win in our model exceeds that of Lose by more than 15%. At the same time, our model has obvious advantages in the English MojiTalk dataset. The percentage of Win in our model exceeds that of Lose by more than 5%. From the results of table 5 and table 6, we observe that our Dual-view CVAE model significantly outperforms the baseline methods. Notably, our model obtains particularly high scores in both relevance and consistency aspects, which indicates that our model is capable of capturing the content and emotion information. That is, our model does not treat the words in the sentence uniformly, but can better capture the key information.

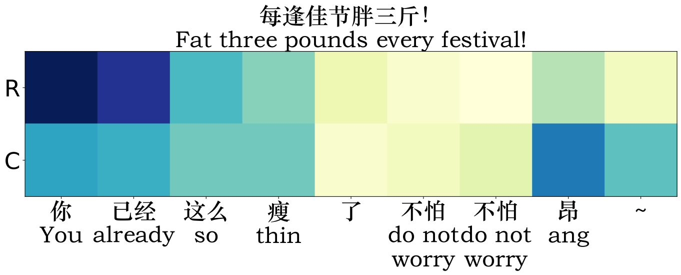


Fig. 8. Attention Weight for Relevance module and Consistency module. On the y-axis, R stands for Relevance and C stands for Consistency.

Query	我最近真的有在健身 I'm really working out these days	请你相信，时间一定能带来惊喜。 Please believe that time will bring surprises.
Emoji	🐶 [doge]	😭 [tears]
Seq2Seq	我也是 Me too	好 Ok
CVAE	我觉得你那边有点高 I think your side is a little high	你的团队，你真的太可爱了 Your team, you are really cute
RL-CVAE	哈哈哈，您对了，甜豆，还挺可爱啊 Ha ha ha, you are right, sweet bean, still pretty lovely	从来没被选过。。。 Never been elected...
DV-CVAE	运动的辰哥更性感 Sports make you sexier，Chen	不要欺骗自己啊 Don't deceive yourself.
Query	马上北京见! See you in Beijing right now!	早an Good morning
Emoji	🥺 [pitiful]	😆 [haha]
Seq2Seq	我也是 Me too	早 Good morning
CVAE	又来了这节目 Here comes the show again.	早安 Good morning
RL-CVAE	求解答! Ask for answers!	哇哈哈哈 Wa Hahahaha
DV-CVAE	哎，我在学校了!!! Hey, I'm already at school！！！	每天都被你帅醒 Everyday, I am called up by your handsome.

Table 7. Examples of generated emotional responses.

To further verify the effect of the relevance and consistency part, we use a case study to visualize the attention weights. As shown in Fig 8, the affective consistency module is more inclined to notice some words related to the expression of mood and emotion, such as “ang” in the example, which is a common modal particle. The correlation module pays more attention to the words related to the content, such as “you already so thin” in the example. The word "thin" is associated with both content and emotion, so the weight of attention in each is similar. We also show some examples of the response generation in table 7. In these examples, the assigned emoji is [doge], which represents a lonely bachelor/bachelorette or expresses pride, silence, disdain. We can see that

Query	emoji	Response
我最近真的有在健身 (I’m really working out these days)	😳 [shy]	美腻了，你是怎么做到滴 It’s so beautiful, how did you do it
	😓 [sweat]	怪不得你是瘦了 No wonder you are thin
	😡 [angry]	不准你瘦了，你是不是把人家甩辣 You are not allowed to lose weight. Do you dump others?

Table 8. The outputs of the DV-CVAE model for the same query conditioning on different emojis.

the seq2seq model tends to generate safe responses and CVAE model can generate content-related responses but cannot consistent with the emotion. The RL-CVAE model seems better, generating more emotion-related responses. However, it is hard for RL-CVAE to cover content and emotion two aspects, which is just the reason why we propose the dual-view CVAE model. Our model is more prone to generate content related to the query, such as “fitness” and “sports”, and more consistent response in terms of emotions. In addition, we also provide the results of giving different emoji conditions for the same query, as shown in table 8. From the table, we can see that emoji tags are effective in changing the emotional expression of response in our DV-CVAE model.

6 CONCLUSION AND FUTURE WORK

In this paper, we have proposed a novel dual-view conditional variational auto-encoder model to build the emotional dialogue generation system. This method explicitly models the content-relevance and emotion consistency at the same time. We evaluate our model on two large-scale datasets (Twitter and Weibo) and one coarse-grained dataset, and achieve the best results on multiple evaluation metrics. Our results demonstrate that the proposed model produces more diverse and interesting responses while improving the content relevance and emotion consistency by human evaluation.

In the future, we plan to find more powerful methods to model the distribution of query, response and emotion. In addition, we will explore the potential to automatically detect emotions during dialogue instead of emoji tag as given.

ACKNOWLEDGMENTS

The research work described in this article has been supported by the National Key Research and Development Program of China under Grant No. 2020AAA0108600.

REFERENCES

[1] Nabiha Asghar, Pascal Poupart, Jesse Hoey, Xin Jiang, and Lili Mou. 2017. Affective Neural Response Generation. 15 (2017), 154–166.

[2] Serkan Ayvaz and Mohammed O. Shiha. 2017. The Effects of Emoji in Sentiment Analysis. *International Journal of Computer and Electrical Engineering* 9, 38 (01 2017), 360–369. https://doi.org/10.17706/IJCEE.2017.9.1.360-369

[3] Francesco Barbieri, Miguel Ballesteros, and Horacio Saggion. 2017. Are Emojis Predictable? 14 (2017), 105–111.

[4] Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. 2015. Generating Sentences from a Continuous Space. *Computer Science* 21 (2015).

[5] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. 2017. Multi-View 3D Object Detection Network for Autonomous Driving. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

[6] Pierre Colombo, Wojciech Witon, Ashutosh Modi, James Kennedy, and Mubbasir Kapadia. 2019. Affect-Driven Dialog Generation. *CoRR* abs/1904.02793 (2019). arXiv:1904.02793 http://arxiv.org/abs/1904.02793

[7] Carl Doersch. 2016. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908* (2016).

[8] Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. *stat* 1050 (2017), 11.

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

[10] Sayan Ghosh, Mathieu Chollet, Eugene Laksana, Louis Philippe Morency, and Stefan Scherer. 2017. Affect-LM: A Neural Language Model for Customizable Affective Text Generation. 6 (2017), 634–642.

[11] Anirudh Goyal, Alessandro Sordoni, Marc Alexandre Côté, Nan Rosemary Ke, and Yoshua Bengio. 2017. Z-Forcing: Training Stochastic Recurrent Networks. 26 (2017).

[12] Chenyang Huang, Osmar Zaiane, Amine Trabelsi, and Nouha Dziri. 2018. Automatic Dialogue Generation with Expressed Emotions. In *NAACL*. Association for Computational Linguistics, 49–54. https://doi.org/10.18653/v1/N18-2008

[13] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *CoRR* abs/1412.6980, 39 (2014). arXiv:1412.6980 http://arxiv.org/abs/1412.6980

[14] Diederik P. Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. 2016. Improving Variational Inference with Inverse Autoregressive Flow. *arXiv preprint arXiv:1606.04934* (2016).

[15] Diederik P Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. *stat* 1050, 17 (2014), 1.

[16] Xiang Kong, Bohan Li, Graham Neubig, Eduard Hovy, and Yiming Yang. 2019. An adversarial approach to high-quality, sentiment-controlled neural dialogue generation. *arXiv preprint arXiv:1901.07129* (2019).

[17] Daniel Kopev, Atanas Atanasov, Dimitrina Zlatkova, Momchil Hardalov, Ivan Koychev, Ivelina Nikolova, and Galia Angelova. 2018. Tweety at SemEval-2018 Task 2: Predicting Emojis using Hierarchical Attention Neural Networks and Support Vector Machine. In *Proceedings of The 12th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, New Orleans, Louisiana, 497–501. https://doi.org/10.18653/v1/S18-1080

[18] Jiwei Li, Will Monroe, and Dan Jurafsky. 2016. A Simple, Fast Diverse Decoding Algorithm for Neural Generation. *CoRR* abs/1611.08562, 31 (2016). arXiv:1611.08562 http://arxiv.org/abs/1611.08562

[19] Jia Li, Xiao Sun, Xing Wei, Changliang Li, and Jianhua Tao. 2019. Reinforcement Learning Based Emotional Editing Constraint Conversation Generation. *arXiv preprint arXiv:1904.08061* (2019).

[20] Nanxing Li, Bei Liu, Zhizhong Han, Yu-Shen Liu, and Jianlong Fu. 2019. Emotion Reinforced Visual Storytelling. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*. 297–305.

[21] 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 *Empirical Methods in Natural Language Processing (EMNLP)*. 2122–2132.

[22] Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective Approaches to Attention-based Neural Machine Translation. *CoRR* abs/1508.04025 (2015). arXiv:1508.04025 http://arxiv.org/abs/1508.04025

[23] Bilyana Martinovski and David Traum. 2003. Breakdown in human-machine interaction: the error is the clue. In *Proceedings of the ISCA tutorial and research workshop on Error handling in dialogue systems*. 11–16.

[24] Petra Kralj Novak, Jasmina Smailović, Borut Sluban, and Igor Mozetič. 2015. Sentiment of Emojis. *Plos One* 10, 16 (2015).

[25] Hirono Okamoto, Masahiro Suzuki, Itto Higuchi, Shohei Ohsawa, and Yutaka Matsuo. 2019. Dual Space Learning With Variational Autoencoders. In *DGS@ICLR*.

[26] Helmut Prendinger, Junichiro Mori, and Mitsuru Ishizuka. 2005. *Using human physiology to evaluate subtle expressivity of a virtual quizmaster in a mathematical game* ☆. Number 4. Academic Press, Inc. 231–245 pages.

[27] Qiao Qian, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2017. Assigning personality/identity to a chatting machine for coherent conversation generation. *CoRR* abs/1706.02861, 37 (2017). arXiv:1706.02861 http://arxiv.org/abs/1706.02861

[28] Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2016. A Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues. 19 (2016).

[29] Louis Shao, Stephan Gouws, Denny Britz, Anna Goldie, Brian Strope, and Ray Kurzweil. 2017. Generating Long and Diverse Responses with Neural Conversation Models. *CoRR* abs/1701.03185, 32 (2017). arXiv:1701.03185 http://arxiv.org/abs/1701.03185

[30] Xiaoyu Shen, Hui Su, Yanran Li, Wenjie Li, Shuzi Niu, Yang Zhao, Akiko Aizawa, and Guoping Long. 2017. A Conditional Variational Framework for Dialog Generation. 5 (2017), 504–509.

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

[32] 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*. 6000–6010.

[33] Ke Wang and Xiaojun Wan. 2018. SentiGAN: Generating Sentimental Texts via Mixture Adversarial Networks. In *IJCAI*. 4446–4452. https://doi.org/10.24963/ijcai.2018/618

[34] Tianming Wang and Xiaojun Wan. 2019. T-CVAE: transformer-based conditioned variational autoencoder for story completion. In *IJCAI'19 Proceedings of the 28th International Joint Conference on Artificial Intelligence*. 5233–5239.

[35] Wenlin Wang, Zhe Gan, Wenqi Wang, Dinghan Shen, Jiaji Huang, Wei Ping, Sanjeev Satheesh, and Lawrence Carin. 2017. Topic Compositional Neural Language Model. *CoRR* abs/1712.09783, 27 (2017). arXiv:1712.09783 http://arxiv.org/abs/1712.09783

[36] Weikang Wang, Jiajun Zhang, Qian Li, Mei-Yuh Hwang, Chengqing Zong, and Zhifei Li. 2019. Incremental Learning from Scratch for Task-Oriented Dialogue Systems. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, Anna Korhonen, David R. Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, 3710–3720. https://doi.org/10.18653/v1/p19-1361

[37] Weikang Wang, Jiajun Zhang, Qian Li, Chengqing Zong, and Zhifei Li. 2019. Are You for Real? Detecting Identity Fraud via Dialogue Interactions. 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 2019, Hong Kong, China, November 3-7, 2019*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, 1762–1771. https://doi.org/10.18653/v1/D19-1185

[38] Xiuyu Wu and Yunfang Wu. 2019. A Simple Dual-decoder Model for Generating Response with Sentiment. *arXiv preprint arXiv:1905.06597* (2019).

[39] Zhongbin Xie and Shuai Ma. 2019. Dual-View Variational Autoencoders for Semi-Supervised Text Matching.. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*. 5306–5312.

[40] Jingjing Xu, Xu Sun, Qi Zeng, Xuancheng Ren, Xiaodong Zhang, Houfeng Wang, and Wenjie Li. 2018. Unpaired Sentiment-to-Sentiment Translation: A Cycled Reinforcement Learning Approach. In *Association for Computational Linguistics*.

[41] Weiran Xu, Xiusen Gu, and Guang Chen. 2019. Generating Emotional Controllable Response Based on Multi-Task and Dual Attention Framework. *IEEE Access* 7 (2019), 93734–93741.

[42] Xinchen Yan, Jimei Yang, Kihyuk Sohn, and Honglak Lee. 2015. Attribute2Image: Conditional Image Generation from Visual Attributes. *CoRR* abs/1512.00570 (2015). arXiv:1512.00570 http://arxiv.org/abs/1512.00570

[43] Zhao Yan, Nan Duan, Peng Chen, Ming Zhou, Jianshe Zhou, and Zhoujun Li. 2017. Building Task-Oriented Dialogue Systems for Online Shopping. 3 (2017). https://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14261

[44] Hainan Zhang, Yanyan Lan, Jiafeng Guo, Jun Xu, and Xueqi Cheng. 2018. Reinforcing Coherence for Sequence to Sequence Model in Dialogue Generation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*. International Joint Conferences on Artificial Intelligence Organization, 4567–4573. https://doi.org/10.24963/ijcai.2018/635

[45] Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning Discourse-level Diversity for Neural Dialog Models using Conditional Variational Autoencoders. 20 (2017), 654–664.

[46] Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional Chatting Machine: Emotional Conversation Generation with Internal and External Memory. 7 (2018).

[47] Liyuan Zhou, Qiongkai Xu, Hanna Suominen, and Tom Gedeon. 2018. EPUTION at SemEval-2018 Task 2: Emoji Prediction with User Adaption. In *Proceedings of The 12th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, New Orleans, Louisiana, 449–453. https://doi.org/10.18653/v1/S18-1071

[48] Xiangyang Zhou, Daxiang Dong, Hua Wu, Shiqi Zhao, Dianhai Yu, Hao Tian, Xuan Liu, and Rui Yan. 2016. Multi-view Response Selection for Human-Computer Conversation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 372–381.

[49] Xianda Zhou and William Yang Wang. 2018. MojiTalk: Generating Emotional Responses at Scale. In *Association for Computational Linguistics*. 1128–1137.