Interpreting open-domain dialogue generation by disentangling latent feature representations

Ye Wang, Jingbo Liao, Hong Yu*, Guoyin Wang, Xiaoxia Zhang and Li Liu

Chongqing Key Laboratory of Computational Intelligence,
Chongqing University of Posts and Telecommunications,
Chongqing, 400065, China.

*Corresponding author(s). E-mail(s): yuhong@cqupt.edu.cn;

Abstract

Currently end-to-end deep learning based open-domain dialogue systems remain black box models, making it easy to generate irrelevant contents with data-driven models. Specifically, latent variables are highly entangled with different semantics in the latent space due to the lack of priori knowledge to guide the training. To address this problem, this paper proposes to harness the generative model with a priori knowledge through a cognitive approach involving feature disentanglement. Particularly, the model integrates the guided-category knowledge and open-domain dialogue data for the training, leveraging the priori knowledge into the latent space, which enables the model to disentangle the latent variables. Besides, this paper proposes a new metric for open-domain dialogues, which can objectively evaluate the interpretability of the latent space distribution. Finally, this paper validates our model on different datasets and experimentally demonstrate that our model is able to generate higher quality and more interpretable dialogues than other models.

Keywords: Deep learning interpretability, Open-domain dialogue, Feature disentanglement

1 Introduction

Human-machine dialogue systems have been developed for almost 70 years, from the introduction of the Turing Test [1] to the present day. Over the years it has acquired increasing attention for its enormous commercial and research value in areas such as semantic assistants and chatbots. Although the black-box deep learning model of human-machine dialogue is continuously progressing regarding the dialogue quality, the end-to-end model still lacks interpretability, which indicates that the model is procedurally unverifiable and incomprehensible. Therefore, the range of applications is comparably limited especially when particularly high-demanding applications require manifestly decision-making logic.

Currently, several deep learning models [2–6] have been proposed to build dialogue systems, mainly divided into task-oriented and non-task-oriented dialogue application systems. Task-oriented dialogue systems are designed to accomplish specific tasks in a domain. Regarding the non-task-oriented system, it is principally an open-domain dialogue system for entertainment, which aims to generate relevant responses. In open domain dialogue systems, the traditional seq2seq generation model [7] encodes the dialogue into a fixed vector of knowledge representations, and inputting the same questions will generate the same responses. In contrast, Transformer-based dialogue generation models [8] abandon the basic paradigm of using a circular recursive structure to encode sequences and use a self-attention mechanism to compute the hidden state of a sequence. Although self-attention mechanism can better model dependencies in long sequences, transformer still does not address the problem that seq2seq cannot generate diverse text. Variational Autoencoders (VAEs) [9] encodes text sequences as probability distributions in the latent space instead of deterministic vectors, which can better model text diversity and achieve category controllability of generation. However, in practical text generation tasks, data often appear in pairs, such as questions and responses in dialogue systems. Conditional Variable Autoencoders (CVAE) [10] adds a conditional encoder to VAE to encode ground-truth responses and some conditions in dialogues, making it more suitable for text generation tasks.

The above dialogue model generation process is not interpretable, making it impossible for users to trust and use the model. The latent variables of the deep latent variable model bring the possibility of interpretability for dialogue generation. Specifically, the latent variables of deep models potentially contain richer semantic features, attracting the researchers to explore the relationship between the feature and the generated data. Therefore, deep latent variable models have additionally become an important work in the study of dialogue systems. In the deep latent variable model, the distribution of the latent space has a great influence on dialogue generation. If the distribution of the latent space is not learned correctly, the latent variable may degenerate into a constant, resulting in the deep latent variable model losing its ability to generate diverse texts. At the same time, the ordinary latent space distribution is cluttered with various semantics, leading to the inability to sample the latent variables with accurate semantics. In a semantically mixed latent space, latent variables with different semantics are chaotically distributed in the latent space. When sampling the latent variables, the model tends to sample around the center of the distribution, and it is easy to sample semantically incorrect latent variables. To solve the problem of semantic clutter, it is generally solved by disentangle the latent space [11, 12]. The semantics of the latent variables in the disentangled latent space is more explicit, and the semantics of the latent variables can be used to predict dialogue generation. In addition, the interpretability of dialogue generation can be enhanced by acquiring interpretable latent spaces.

To summarise our contributions:

- To solve the semantic entanglement problem in open-domain dialogue system for generating relevant dialogue, this paper integrates the guided-category knowledge and open-domain dialogue data for the training, which enables the model to disentangle the latent variables.
- Because hidden features in the black box of dialogue models are hard to measure, this paper proposes a new metric to objectively evaluate the interpretability of potential features. Extensive experiments are also designed to analyze the interpretability of A-CVAE using this metric.
- To validate the model in terms of interpretability, diversity and quality of text generation, adequate experiments have been performed on different datasets and comprehensive analyses have been provided.

2 Related Works

The dialogue system that relies on the seq2seq model encodes the dialogue text into a fixed vector, but it may generate generic responses due to the maximum likelihood-based optimization. On the other hand, using the Variational Autoencoder (VAE) allows encoding the dialogue as continuous latent variables, which can increase the diversity of generated dialogues. However, the semantics of these latent variables may not be clear, resulting in an uninterpretable generation process. Our work is a dialogue generation model that is based on a deep latent variable model and is closely related to the research on deep latent variable models. Kingma et al. [9] proposed VAE and applied it to the image domain. Later, Bowman et al. [13] applied VAE to the field of natural language processing. VAE-based dialogue generation approaches leverage the latent space distribution to enhance interpretability and diversity in the generated dialogues. As a result, there have been numerous efforts in using VAE for dialogue generation. Zhao et al. [14] proposed two models based on VAE, DI-VAE and DI-VST. The models discover interpretable semantics through automatic encoding or contextual prediction, thus solving the problem that traditional dialogue models can not output interpretable actions. Serban et al. [15] introduced a novel neural network structure called Latent Variational Hierarchical Recurrent Encoder-Decoder (VHRED), which incorporates latent stochastic variables that span across variable numbers of time steps.

This model is specifically designed for text generation, conditioned on long contexts, making it well-suited for generating dialogues with extended context information. Zhao et al. [16] proposed a dialogue generation model based on conditional variational autoencoders and incorporated external knowledge to improve the interpretability of the model. Gao et al. [17] introduced a discrete variable with explicit semantics in CVAE, using the semantic distance between latent variables to maintain good diversity among the sampled latent variables. Wang et al. [11] proposed the semantic-aware conditional variant autoencoder (S-CVAE) model. S-CVAE can generate diverse dialogue text by utilizing embedded classifiers and feature disentangling modules. Shi et al. [12] proposed a method that utilizes an exponential mixture distribution as a replacement for the Gaussian prior in the Variational Autoencoder framework. This allows for capturing hidden semantic relations between the mixture components and the data, resulting in a more interpretable latent space. Additionally, they decompose the evidence lower bound (ELBO) and derive a loss term that addresses the issue of mode collapse, which is a common problem in VAEs. This approach helps to mitigate mode collapse and improve the quality of generated dialogues to a certain extent. Pang et al. [18] introduced a text generation model that leverages the energy of the latent space. They incorporated both continuous latent variables and discrete categories in the prior distributions, which guides the generator to generate high-quality, diverse, and interpretable texts. By combining continuous and discrete variables in the latent space, their proposed model aims to improve the interpretability and diversity of the generated texts, allowing for more controlled and meaningful text generation. In order to better disentangle the latent space features, Higgins et al. [19] proposed beta-VAE to learn interpretable representations by introducing beta hyperparameters. Mathieu et al. [20] analyzed the characteristics of the VAE latent space encoding and explored how to solve the entangled latent space, while demonstrating that choosing different priors can yield different decompositions. Chen et al. [21] use a pre-trained classifier to guide the model to disentangle the features of interest from the latent representation.

The studies mentioned above have identified that in the latent space of deep latent variable models, dialogues with different semantics can be entangled together, leading to a distribution of semantic data scattered around a single dialogue with definite semantics. This can result in the generation of dialogues with incorrect semantics due to the randomness of sampling from the latent space. To address this issue, this paper proposes a different approach. In this paper, dialogues of different categories are trained separately using a dialogue generation task. The approximate posterior distribution of the deep latent variable model is then used as prior knowledge for the corresponding category. This prior knowledge is utilized to guide the decoupling of the latent space, ensuring that dialogues of different categories are distributed in the corresponding regions of the latent space. By disentangling the latent space, the model is better able to sample latent variables that correspond to the accurate semantic categories, resulting in the generation of high-quality, interpretable, and diverse texts. This approach differs from previous studies as it leverages category-specific prior knowledge to guide the latent space decoupling process, leading to improved dialogue generation quality.

3 Proposed Scheme

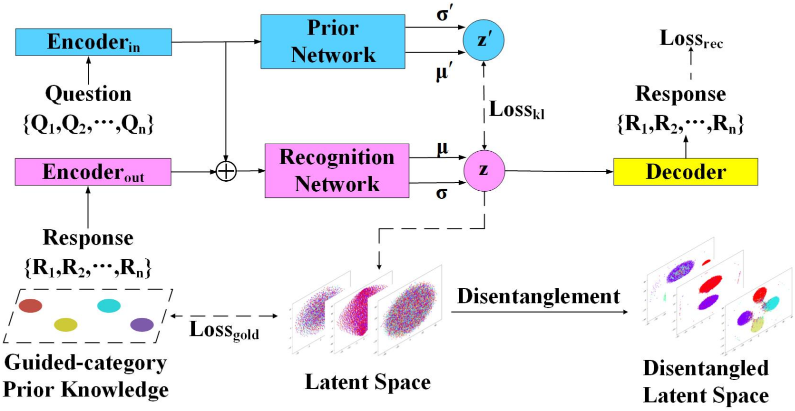


Fig. 1: A-CVAE model diagram.

3.1 Advanced Conditional Variational Autoencoders

The traditional deep latent variable model introduces latent variables to optimize the model, but the model itself remains a black box as we do not have clear understanding of what the latent variables represent. However, the interpretability of deep latent variable models can be improved if we can specify the meaning of the latent variables. In this approach, the latent variable is represented as a 200-dimensional multidimensional Gaussian distribution, with each dimension assumed to be independent. To make the latent variable distribution learn external knowledge, the Kullback-Leibler (KL) divergence between the latent variable distribution and the knowledge distribution can be minimized. By minimizing the KL divergence, the latent variable distribution approximates the external knowledge distribution, allowing the latent variables to incorporate the external knowledge. This approach enables the deep latent variable model to learn interpretable representations of the latent variables, as they are now associated with external knowledge and can be better understood in terms of their meaning and significance.

The proposed scheme is named Advanced Conditional Variational Autoencoders (A-CVAE)(Fig. 1) and aims to disentangle the latent space in a deep latent variable model by incorporating external knowledge. By introducing external knowledge, A-CVAE can ensure that different categories of dialogues are encoded into separate regions of the latent space, allowing for more accurate sampling of latent variables during the random sampling process. For example, if the external knowledge is emotion, the latent space can be disentangled so that sampled latent variables are not only related to emotion, but also specify whether the emotion is positive or negative. This enhances the interpretability of the dialogue text generation process, as the sampled latent variables can now be associated with specific categories and their corresponding semantics. To introduce external knowledge and enable disentanglement of the latent space, the CVAE model will be used to train the category prior knowledge, which will be represented as a Gaussian distribution. This category prior knowledge will then be utilized to guide the disentanglement of the latent space, ensuring that different categories of dialogues are represented in distinct regions of the latent space, leading to improved interpretability and better control over the generated dialogue texts.

Our approach is divided into two steps, representing prior knowledge and training the model using prior knowledge.

3.1.1 Module Definition.

The input in the dialogue task is divided into a question sequence $\{\mathbf{C}_1, \mathbf{C}_2, \cdots, \mathbf{C}_n\}$ denoted as C and a response sequence $\{\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n\}$ denoted as X. To apply the text to the deep model, we need to embed words into the word vector space. $e(W)$ denotes embedding the word W into the word space, and W can be either C or X. For training, we use questions and responses to sample latent variables from an approximate posterior distribution to generate dialogue text. For testing, we use only the questions to sample the latent variables from the prior distribution, thus generating the text. A-CVAE consists of five main modules, the input encoder $\mathrm{Enc}_{in}(\cdot)$, the output encoder $\mathrm{Enc}_{out}(\cdot)$, the recognition network $\mathrm{RecogNetq}_{\phi(z|X,C)}$, the prior network $\mathrm{PriorNetp}_{\ \varphi(z|C)}$ and the decoder $\mathrm{Dec}\ (\cdot)$. The input encoder is used to encode the questions in the dialogue, the output encoder is used to encode the responses in the dialogue, the recognition network is responsible for generating the mean and variance of the approximate posterior distribution, the prior network is responsible for generating the mean and variance of the prior distribution, and the decoder is used to decode the latent variables and generate the responses.

3.1.2 The priori knowledge representation.

To fuse the category knowledge into the model and then disentangle the latent space, we need to train the CVAE in the dialogue task using a single category dataset and use the latent space distribution of the CVAE as the category prior knowledge representation. First, the dataset is divided into k sub-datasets by category. Then data are randomly selected from each sub-dataset and trained on CVAE to obtain a priori knowledge representations $\mathcal{N}\left(\mu_k, \sigma_k^2 I\right)$ for each category, as shown in Eq.1. k denotes the kth category, RecogNet $\mathrm{q}_{\phi(z|X,C)}$ denotes the approximate posterior network of CVAE, h_c denotes the feature vector of the problem text, and h_x denotes the feature vector of the reference response text.

$$\mathcal{N}\left(\mu_k, \sigma_k^2 I\right) = \text{RecogNet}\, \mathrm{q}_{\phi(z|X,C)}\left(h_x, h_c\right) \tag{1}$$

The prior knowledge representation of the category is the latent space distribution of CVAE, which is a multidimensional Gaussian distribution. We name the prior knowledge representation as gold Gaussian. The gold Gaussian is used to guide the training of the model, which is approximated by the Kullback-Leibler divergence to the multidimensional Gaussian distribution in the latent space of the model during training.

3.1.3 Train.

In A-CVAE model, this paper proposes the optimized objective function by extra adding our proposed optimized loss function for disentangling the latent space distribution to the standard loss of the CVAE.
In CVAE, the goal is to fit the conditional probability distribution $P(X \mid C)$. We calculate $P(X \mid C)$ by introducing the latent variable z, as shown in Eq.2.

$$P(X \mid C) = \int_z p(z \mid C) P(X \mid z, C) dz \tag{2}$$

Since there is an integral over the continuous variable z in the middle of Eq.2, and the integral is difficult to be calculated. An approximate solution is obtained by means of variational extrapolation, and the solution is called the evidence lower bound (ELBO), as shown in Eq.3.

$$\text{ELBO} = E_{q_\phi(z|X,C)}\left[\log P_\theta(X \mid z, C)\right] - KL\left(q_\phi(z \mid X, C) \| p_\varphi(z \mid C)\right) \tag{3}$$

$E_{q_\phi(z|X,C)}\left[\log P_\theta(X \mid z, C)\right]$ is the loss function of the reconstructed response text. We use the input encoder $\text{Enc}_{in}(\cdot)$ to encode the question sequence to obtain the feature r_C of the question C, and we use the output encoder $\text{Enc}_{out}(\cdot)$ to encode the response to obtain the feature r_X of the response sequence X. Meanwhile, we use r_C and r_X as inputs to the recognition network $\text{RecogNetq}_{\phi(z|X,C)}$ to obtain the mean μ and variance σ^2 of the approximate posterior distribution. r_C is used as the input of the prior network $\text{PriorNetp}_{\ \varphi(z|C)}$ to calculate the mean μ' and variance ${\sigma'}^2$ of the prior distribution. $q_\phi(z \mid X, C)$ is the approximate posterior distribution, which is used during training. $P_\theta(X \mid z, C)$ is a conditional probability distribution for generating the response text using the latent variable z and the input condition C, which is fitted by the decoder $\text{Dec}(\cdot)$. The processes sampling the latent variables are not derivable, and to be able to optimize the model by the gradient descent algorithm, we use the reparameterization trick to make all the processes derivable, as shown in Eq.4.

$$z=\mu+\sigma * \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, I) \tag{4}$$

$KL\left(q_\phi(z \mid X, C) \| p_{\varphi}(z \mid C)\right)$ is the KL dispersion of the prior distributions $p_{\varphi}(z \mid C)$ and approximate posterior distributions $q_\phi(z \mid X, C)$. Due to the specificity of the dialogue generation task data, during testing we have only question sequences as input and no reference response sequences. Therefore, in the training and testing phases, we sample in two different multidimensional Gaussian distributions, respectively, the approximate posterior distribution $\mathcal{N}\left(\mu, \quad \sigma^2 I\right)$ and the prior distribution $\mathcal{N}\left(\mu', \quad \sigma'^2 I\right)$. To make the two distributions closer, the algorithm uses the KL distance to approximate them. The parameters μ, σ^2, μ', σ'^2 are fitted by the prior network and the recognition network, respectively. Where ϕ, φ, θ are the parameters of two Gaussian distributions.

Our proposed optimization loss is shown in Eq.5. This loss function can guide the KL distance between the approximate posterior distribution and the gold Gaussian distribution of each class to present variability in different dimensions. $\mathcal{N}\left(\mu_k, \sigma_k^2 I\right)$ is the gold Gaussian distribution obtained by pre-training the kth category. During training, the KL value is calculated by selecting the corresponding gold Gaussian distribution according to the category of the dialogue text, thus approximating the posterior distribution and the Gaussian distribution to optimize the parameters.

$$\mathcal{L}_{\text{gold}}=KL\left(\mathcal{N}\left(\mu_k, \sigma_k^2 I\right) \| q_\phi(z \mid X, C)\right) \tag{5}$$

The total loss during training of the A-CVAE model is the reconstructed expectation loss plus the KL scatter of the prior and approximate posterior distributions, plus the optimisation loss. β, λ are the kl term coefficient to avoid the KL posteriori collapse problem. If a penalty factor is not added to the KL scatter, it is easy to make the KL term become 0 during the optimization process, thus making the output of the encoder constant. Adding penalty coefficients can solve this problem. The value of the penalty coefficient is related to the number of training steps, and the periodic annealing strategy is used. As shown in Eq.6:

$$\begin{aligned}
\mathcal{L}= & -E_{q_\phi(z \mid X, C)}\left[\log P_\theta(X \mid z, C)\right] \\
& +\beta KL\left(q_\phi(z \mid X, C) \| p_{\varphi}(z \mid C)\right) \\
& +\lambda KL\left(\mathcal{N}\left(\mu_k, \sigma_k^2 I\right) \| q_\phi(z \mid X, C)\right)
\end{aligned} \tag{6}$$

Algorithm 1 A-CVAE training procedure

```
Input: Data samples (C, X) = $(C_i, X_i)_{i=1}^{N}$.
Output: Parameter($\phi$, $\varphi$, $\theta$) convergence.
Initialize the parameters ($\phi$, $\varphi$, $\theta$).
if Train then
    for iter from 1 to max_iter do
        Sampling $(C, X)$ from the training set.
        Calculate question text features $r_C \leftarrow \mathrm{Enc}_{in}(\cdot)(e(C))$.
        Calculate response text features $r_X \leftarrow \mathrm{Enc}_{out}(\cdot)(e(X))$.
        Calculate $\mu$, $\sigma \leftarrow RecogNetq_{\phi(z|X,C)}(r_X, r_C)$.
        Calculate $\mu'$, $\sigma' \leftarrow PriorNetp_{\varphi(z|C)}(r_C)$.
        Reparameterize $z = \mu + \sigma * \varepsilon$.
        Calculate generated text $\bar{X} \leftarrow$ Dec $(z)$.
        Calculate gold KL loss by Eq.5.
        Calculate variational evidence lower bound by Eq.3.
        Update $\phi$, $\varphi$ and $\theta$ by gradient ascent by Eq.6.
    end for
end if
```

3.2 Interpretability Evaluation Index for Dialog

We propose a metric named Interpretability Evaluation Index for Dialog (IEID) for objectively evaluating the interpretability of latent space distribution. Since we sample the latent space to get latent variables containing category semantics, the text generated by the decoder using latent variables will contain category information. Therefore, we train a text classifier using CVAE with reference responses from the dataset as input to the input encoder $Encoder_{in}$. Besides, the category labels is the input of the output encoder $Encoder_{out}$, and the decoder reconstructs the category labels. Since the responses are generated with latent variables containing category information, the semantics of the output contains related category information. Under this assumption, we further implement the specific classifiers to classify the generated responses and reference output, aiming to verify the correctness of the proposed hypothesis. Therefore, we define the proportion of reference responses and generated responses that are classified into the same category as an evaluation metric to evaluate the interpretability of the latent space distribution. Here below is the description of IEID: t stands for ground-truth response and g represents generated response. $Clf(\cdot)$ denotes the classifier. $Label_t$ defines the classifier takes t as the classification result of the input. $Label_g$ denotes the classifier takes g as the classification result of the input and n is the number of dialogue in the test set.

$$label_t = Clf(t), \qquad label_g = Clf(g) \tag{7}$$

$$IEID = \frac{\sum_{i=1}^{n} f\left(label_{t_i}, label_{g_i}\right)}{n} \tag{8}$$

$$f(label_{t_i}, label_{g_i}) = \begin{cases} 1 \; label_{t_i} = label_{g_i} \\ 0 \; label_{t_i} \neq label_{g_i} \end{cases} \tag{9}$$

4 Experiments

4.1 Dataset

DD dataset: DailyDialog Dataset [22] is a multi-round dialogue dataset for everyday chat scenarios, which has less noise than previous corpora. DD dataset covers several major topics of life, and is annotated with the action and emotion of each dialogue. In our experiments, we used the emotion and action labels from the dataset. Due to the uneven distribution of dialogue across categories in the dataset, we aggregated the dialogue into three categories (no emotion, negative, positive) according to the emotion category. We also decomposed the multi-round dialogue into single-round dialogue.

ED dataset: EmpatheticDialogues dataset[23] is a large-scale multi-round dialogue dataset collected on Amazon Mechanical Turk, containing 24,850 one-to-one open domain dialogue. The dataset provides labels for emotions in 32 categories. We aggregated the dialogue by emotion category into positive and negative.

For the labels in the dataset, we have two main uses: dividing the sub-dataset according to the labels and selecting the golden Gaussian to optimize the model according to the labels of the dialogue during model training.

4.2 Automatic Evaluation Metrics

We use three evaluation metrics to assess the quality of text generation, an evaluation metric to evaluate the diversity of the generated texts, and our proposed metric (IEID) to evaluate the interpretability of the hidden space distribution.

BLEU [24] is a fast, inexpensive and language-independent method for assessing the quality of text generation. This method can replace review by skilled humans when rapid or frequent evaluation is required. And BLEU calculates the co-occurrence word frequency for both sentences.

ROUGE [25] is proposed to estimate text quality by calculating co-occurrence frequencies. Unlike BLEU the words in ROUGE can be discontinuous.

METEOR [26] is made more relevant to manual discrimination by calculating the relationships between synonyms, roots and affixes.

Distinct-2 [27] is used to evaluate the diversity of text generated by the generative model. A higher value indicates that the generated text is richer, while a lower value indicates that the generated text contains a large amount of repetitive, generic and meaningless text.

4.3 Benchmark Methods

We used the same dataset on the dialogue generation task and compared it with the model below.

Seq2seq [7], breaking through the traditional fixed-size input problem framework, converts an input sequence of indefinite length into an output sequence of indefinite length. It also incorporates a Global Attention mechanism [28][29] to notice more textual information;

Transformer [8], eschewing the traditional recurrent neural network, and using the self-attention mechanism to encode and decode the text. The self-attention mechanism also allows the Transformer to encode longer sequences of text;

Vanilla VAE [9], encoding text into a continuous latent space rather than a fixed vector. Sampling the latent variables from a Gaussian distribution increases the diversity of the generated text;

Vanilla CVAE [10], which adds a conditional encoder to VAE, is able to encode the response text, sentiment and other conditions into the model. Such a structure is more suitable for dialogue generation tasks;

CVAE with VADE [30], introducing a Gaussian mixture model in the latent space, to obtain a better representation of the text through the clustering task.

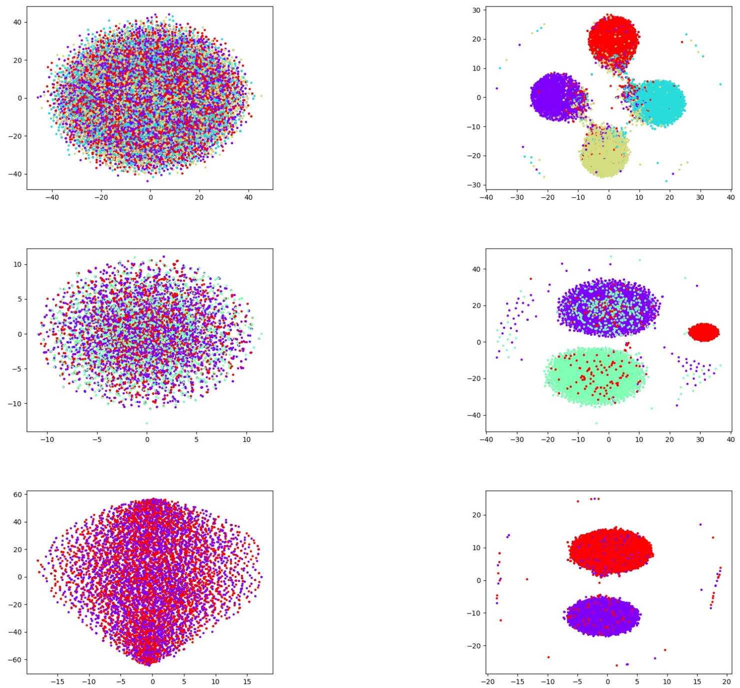
DGM-VAE [12], adds Dispersion term as penalty term to GM-VAE to solve Mode-Collapse Problem and acquires well-structured and meaningful latent space.

DCM-VAE [12], adds Dispersion term as penalty term to CM-VAE to solve Mode-Collapse Problem and acquires well-structured and meaningful latent space.

SVEBM [18], couples continuous latent variables and discrete categories to extract latent variables with category information.

4.4 Implementation Details

The words in the dataset were processed to produce a word list dictionary containing 6918 words. The dictionary contains four special tokens, $\langle PAD\rangle$, $\langle SOS\rangle$, $\langle EOS\rangle$ and $\langle UNK\rangle$. $\langle PAD\rangle$ denotes padding values, $\langle SOS\rangle$ denotes the start of sentence, $\langle EOS\rangle$ denotes the end of sentence and $\langle UNK\rangle$ denotes words that do not appear in the word list. The word embedding vector is 300 dimensions. The encoder and decoder of A-CVAE utilize LSTM [31] as the basic unit, with hidden size 300, latent variable 200 dimensions, and periodic annealing using the KL annealing algorithm to avoid KL posterior collapse. The recognition and prior networks are multilayer perceptrons with the number of hidden layer neurons set to 250. In the training phase, the model uses the



(a) Original latent space distribute

(b) Disentangled latent space distribute

Fig. 2: Visualization results of the latent space distribution for different datasets. The datasets from top to bottom are DailyDialog dataset with action label, DailyDialog dataset with emotion label, and EmpatheticDialogues dataset with emotion label.

Adam optimizer. The initial learning rate is 0.0001, the learning rate decays by 0.01 after each epoch, and the coefficient of the kl term increases to 1 after 10,000 updates.

4.5 Results and analysis

4.5.1 Disentangled result.

We downscaled the original latent space distribution of the deep latent variable model and the disentangled latent space distribution into two dimensions by t-SNE and visualized them. The original latent space represents the latent space without disentangling, and the disentangled latent space represents the A-CVAE disentangled latent space. Fig. 2 shows the visualization results. We can see that the original latent space is a miscellaneous distribution, which cannot distinguish between different classes of data in the latent space. The deep latent variable generation model obtains the latent variables by random sampling in the latent space, and it is very difficult to accurately sample the corresponding semantic latent variables in the miscellaneous latent space. Due to the randomness of sampling and the chaotic data distribution of each category in the original latent space, it is easy for the model to sample other semantic latent variables in the mixed latent space. After disentangling the latent space by our method, the data of the same category are closer to each other in the latent space, while the data of different categories are further away from each other. We can find that after disentangling, the same class of data is distributed around one same category data in the latent space. Although the sampling process is still random, we can still sample the corresponding class of data more accurately in the disentangled latent space. The comparison experiments in the next subsection also show that extracting semantically accurate latent variables in the disentangled latent space can improve the quality, interpretability, and diversity of the generated text.

Table 1: The automatic evaluation results of all compared methods on the DD dataset using action labels.

Method	BLEU	METEOR	ROUGE	dist-2
Seq2seq	10.14±0.08	10.52±0.09	30.15±0.11	0.17
Transformer	10.11±0.08	11.16±0.09	**33.09±0.11**	0.25
Vanilla VAE	10.85±0.10	11.52±0.11	30.73±0.11	0.14
Vanilla CVAE	11.50±0.11	11.95±0.12	31.12±0.11	0.14
CVAE with VADE	11.49±0.20	12.23±0.23	32.44±0.21	0.22
DCM-VAE	9.57±0.11	10.02±0.10	26.38±0.11	0.15
DGM-VAE	11.20±0.20	11.24±0.21	31.85±0.20	0.17
SVEBM	11.75±0.21	11.56±0.20	31.94±0.22	0.20
A-CVAE	**12.28±0.23**	**12.98±0.26**	32.36±0.22	**0.26**

Table 2: The automatic evaluation results of all compared methods on the DD dataset using emotion labels.

Method	BLEU	METEOR	ROUGE	dist-2
Seq2seq	11.15±0.22	11.89±0.26	30.99±0.25	0.30
Transformer	10.44±0.19	11.03±0.21	**31.59±0.25**	**0.42**
Vanilla VAE	10.01±0.17	10.48±0.19	28.09±0.21	0.18
Vanilla CVAE	11.17±0.22	11.50±0.25	29.46±0.23	0.25
CVAE with VADE	10.12±0.19	10.50±0.21	28.07±0.22	0.23
DCM-VAE	10.04±0.13	9.58±0.14	27.59±0.15	0.18
DGM-VAE	11.92±0.23	12.30±0.22	31.56±0.23	0.18
SVEBM	11.03±0.26	10.98±0.27	28.82±0.27	0.25
A-CVAE	**12.59±0.29**	**13.22±0.33**	30.72±0.28	0.28

Table 3: The automatic evaluation results of all compared methods on the Empathetic Dialogu dataset using emotion labels.

Method	BLEU	METEOR	ROUGE	dist-2
Seq2seq	6.86±0.08	8.04±0.10	31.03±0.12	0.21
Transformer	6.40±0.08	7.90±0.09	31.69±0.13	**0.41**
Vailla VAE	7.40±0.09	8.37±0.10	30.37±0.12	0.19
Vanilla CVAE	7.37±0.09	8.39±0.10	30.79±0.12	0.22
CVAE with VADE	6.94±0.09	8.09±0.10	30.17±0.12	0.14
DCM-VAE	6.59±0.08	8.10±0.11	29.07±0.11	0.18
DGM-VAE	7.58±0.10	**8.63±0.08**	**31.87±0.12**	0.17
SVEBM	7.49±0.20	7.84±0.21	27.11±0.24	0.20
A-CVAE	**7.75±0.10**	8.56±0.11	31.72±0.13	0.23

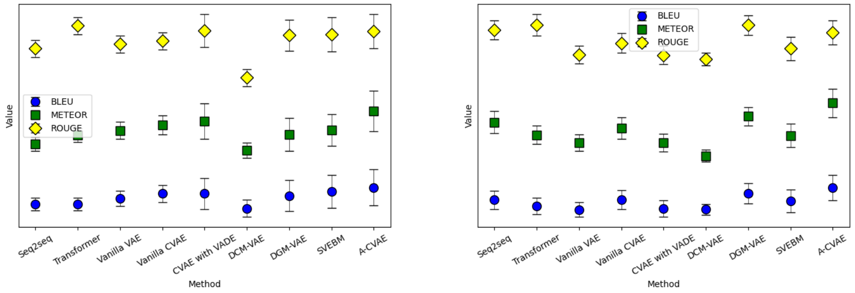
Table 4: The result of IEID.

Dataset Method	Action Of DD	Emotion of DD	Emotion of ED
Seq2seq	0.19	0.56	0.62
Transformer	0.24	**0.61**	**0.72**
Vanilla VAE	0.21	0.52	0.63
Vanilla CVAE	0.44	0.57	0.63
CVAE with VADE	0.39	0.56	0.63
DCM-VAE	0.40	0.56	0.64
DGM-VAE	0.36	0.55	0.64
SVEBM	0.42	0.60	0.68
A-CVAE	**0.44**	0.56	0.65

4.5.2 Automatic evaluation result.

Table 1, Table 2 and Table 3 show the results of the evaluation metrics for A-CVAE and the comparison models on different datasets. Table 1 shows the results on the DailyDialog dataset with action labels, Table 2 shows the results on the DailyDialog dataset with emotion labels, and Table 3 shows the results on the EmpatheticDialogues dataset with emotion labels.

The tables show that A-CVAE outperforms most comparative models on several metrics that assess the quality of text generation. This indicates that A-CVAE is able to generate high quality, diverse dialogue. In particular, A-CVAE outperforms all comparative models on the BLEU and ROUGE metrics, and is only lower than Transformer on ROUGE. With the fusion of category information, we can sample higher quality latent variables from the disentangled latent space and thus generate higher quality text. The results illustrates that external prior knowledge has an impact on text generation, and that sampling the correct category of latent variables enhances text generation. A-CVAE achieves excellent results with different datasets and different labels, indicating that A-CVAE has good generalization ability.



(a) Error bar result of action

(b) Error bar of emotion

Fig. 3: Error bar results for metrics with different generation models on DailyDialog dataset.

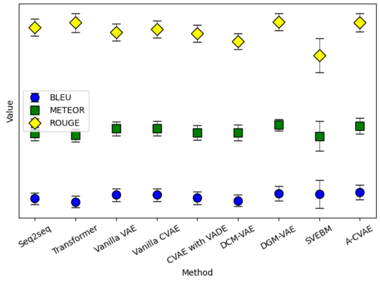


Fig. 4: Error bar results for metrics with different generation models on EmpatheticDialogues dataset.

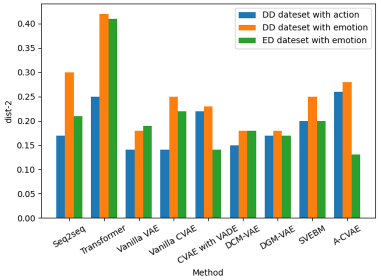


Fig. 5: Diversity assessment metric dist-2 bar chart of different models under different datasets.

Since the model generates different quality results for different texts in the dataset, we calculate the error bar between the metric value of each data in the test set and the mean value of the metric results for all data, in order to verify the stability of the text generated by the model. Since it is also important that the dialogue model can generate text stably, we calculate the error bar. Fig. 3 and Fig. 4 show the error bar results of different models for different metrics on different datasets. We can find that although A-CVAE achieves good results on several different evaluation metrics, the value of error bar is relatively large, indicating that the text generated by A-CVAE is not very stable and sometimes the quality of the generated text is poor. Meanwhile, other work on latent variable based models such as SVEBM, DGM-VAE, etc. suffer from similar problems. Going to explore how to maintain or even improve the stability of the model while increasing the complexity of the model is also a valuable research point. In our future work, we will study how to generate more stable and high-quality texts.

Fig.5 shows the results of the diversity evaluation index in the form of a bar graph. From the bar graph, we can find that A-CVAE is better than most the other comparison models except Transformer. We have analyzed the generated text and found that Transformer tends to generate some longer dialogues, which is due to Transformer's self-attention mechanism can encode longer sequences of text, so the generated text is also longer than the normal model. The metric dist-2 is used by stitching all generated texts together and then taking n words as a group, and finally calculating the proportion of non-repeated n-gram phrases to the total n-gram phrases, so the longer texts may get higher dist-2 scores.

4.5.3 IEID result.

Interpretability Evaluation Index for Dialog (IEID) is the metric we proposed in section 3.2 for evaluating the interpretability of the distribution of the latent space. IEID assesses the interpretability of the model by the category consistency of the generated text and the reference response text. Table 4 shows the IEID results for different models on different datasets. In the DailyDialog dataset with action tags, conventional models such as Seq2seq and VAE have lower IEID values, indicating that the models are not very interpretable. A-CVAE is higher than most of the comparison models, indicating that the disentangled latent space has a positive impact on text generation. While on the DailyDialog dataset with emotion labels and EmpatheticDialogues dataset A-CVAE does not differ much from the other models, after analysis, we found that it is due to the low accuracy of the classifier used in the process of computing the metrics. Moreover, the emotion tag only has three categories of no emotion, positive and negative on the DailyDialog dataset, and only two categories of positive and negative on the EmpatheticDialogues dataset, so the generated text The probability of the generated text and the reference text being classified into the same category becomes larger, resulting in higher values for both indicators, which leads to insignificant results. However, it can still be seen that A-CVAE has better interpretability than the common generative model. In addition, the DCM-VAE, DGM-VAE, and SVEBM in the comparison algorithm have improved values on IEID in some datasets compared to traditional models such as VAE and CVAE, indicating a positive y impact of interpretable latent space inventory on dialogue generation.

Table 5 gives some examples of dialogues, where *Post* represents the question, *Response* represents the reference response, *Generate* represents the model-generated response, and the content in parentheses indicates the category of the text. From these examples, it can be found that the reference response text and the generated text are classified into the same category, which indicates that the latent variables with accurate semantics can be sampled from the disentangled latent space to generate the dialogue text, enhancing the interpretability of the dialogue model.

Table 5: The examples of dialogue generate.

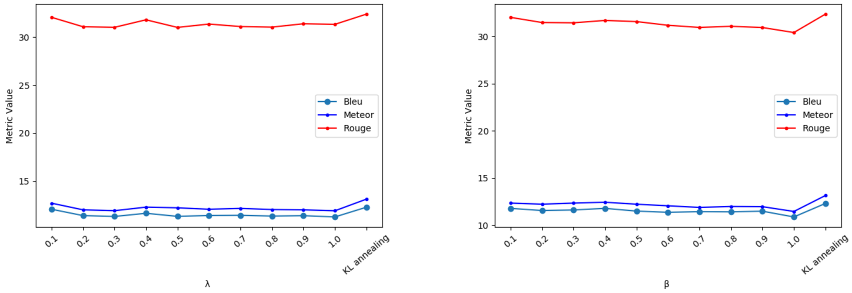
Post	What do you need? **(Questions)**
Response	I need to talk to you about that new driver you've hired. I think I am going tohave some problems working with him. **(Directives)**
Generate	I need to make an appointment, please. **(Directives)**
Post	What time do you expect him back? **(no emotion)**
Response	Sorry , I have no idea. You can call him there if you like. **(no emotion)**
Generate	I think he'll be back in about an hour at least. **(no emotion)**
Post	Sorry , I have no idea. You can call him there if you like. **(Directives)**
Response	Ok , I have the number. Bye! **(Commissive)**
Generate	Good idea. We'll get her a bottle of wine. **(Commissive)**

Table 6: The automatic evaluation results of multi-dimensional latent space on the DD dataset using action and emotion labels. Ratio represents the ratio of the dimensions of common features, action, and emotion in the latent space.

Ratio	BLEU	METEOR	ROUGE	dist-2
50:50:100	12.30±0.23	13.14±0.26	32.38±0.22	**0.25**
50:50:100(Math)	12.15±0.22	12.97±0.25	32.37±0.21	0.24
50:75:75	11.83±0.22	12.54±0.25	31.81±0.21	0.24
50:75:75(Math)	12.03±0.22	12.76±0.25	31.96±0.22	0.24
50:100:50	**12.53±0.23**	**13.36±0.26**	**33.08±0.22**	0.24
50:100:50(Math)	11.96±0.22	12.77±0.25	32.30±0.21	0.24

4.5.4 Other analysis.

In previous experiments, this paper was using each tag separately, so that the latent space is disentangled according to the action tag or the emotion tag. In this subsection, this paper uses both action tags and emotion tags to disentangle the latent space. Meanwhile, this paper abstracted a common feature which represents the part that is not disentangled according to the tag. The latent space is decomposed into three parts, the common feature which is not disentangled , the action feature which is disentangled by action tags, and the emotion feature which is disentangled by emotion tags. This paper constructs three different hidden spaces with the three parts in different proportions. In addition, when describing the motivation of our method, it is mentioned that different gold Gaussians are different between different dimensions, so the disentangling can be done by approximating the latent space to the corresponding gold Gaussians through KL distance, and finally the experiments also prove this point. The gold Gaussian is pre-trained by following different classes of data, which is essentially a multi-dimensional Gaussian distribution with several different parameters. It is natural to think that the gold Gaussian may not necessarily need to be pre-trained to get it. This paper can also construct gold Gaussian with several different parameters through mathematical methods. This paper takes the expectation of the Gaussian distribution changing in steps of 1 from 0 (including 0) to the left and right and the variance changing in steps of 1 from 0 (not including 0) to the left and right. Thus, several Gaussian distributions with different parameters are constructed as gold Gaussians. Table 6 shows the results of the experiments, where the ratio is the proportion of three different features, the labeled math represents our own constructed gold Gaussian, and the unlabeled represents the gold Gaussian trained by a single class of dataset. From the results, we can see that both the pre-trained gold Gaussian and our own constructed gold Gaussian can get better results than the baseline, which shows that the gold Gaussian can be constructed by mathematical methods without the need to get it by training. Meanwhile, we can see from the results that after dividing the latent space into 3 different parts, we can still get better results than baseline, but the results are slightly worse than A-CVAE with a single label. The main reason for this is that the data set is not balanced in terms of categories, and it is not possible to balance the action and emotion at the same time, so that the labels of both categories are equally distributed in the data set.



(a) λ parameter sensitive line chart

(b) β parameter sensitive line chart

Fig. 6: Parameter sensitivity analysis line chart.

In this study, we conducted a parameter sensitivity analysis for the coefficients λ and β in Eq.6, and the results are presented in Fig. 6. Among the methods used, KL annealing employs the KL annealing algorithm, where the current coefficient value is the ratio of current training steps to the total training steps. This algorithm is used to prevent the KL disappearance problem of the deep latent variable model, which could cause the hidden variable to degenerate to a constant. Furthermore, we fixed the coefficients separately, starting from 0.1 with a step increment of 0.1 up to 1.0. As shown in Fig. 6, KL annealing provided the best results. After fixing the parameters, there was no significant difference in the overall results. However, coefficients of 0.1 and 0.4 produced relatively good results. On the other hand, when β was fixed to 1, the worst results were obtained, but the difference in the results was not significant.

5 Conclusion

In this paper, for further disentangle the latent space, we propose the A-CVAE model, which integrates category prior knowledge and dialogue data to guide the latent feature disentangling. Specifically, the correct-semantic latent variables are sampled from the disentangled latent space, which enhances the interpretability of the model in the latent space and generate related responses. Additionally, we propose a new metric for objectively evaluating the interpretability of the latent space distribution in open-domain dialogues. The experimental results show that on the DD dataset and ED dataset, A-CVAE generates higher quality and more diverse dialogue text, providing extra interpretable latent spaces than other models. In future work, we consider how to control the sampling of latent variables in a disentangle latent space to generate controlled dialogue texts.

6 Acknowledgements

This work was partly supported by National Key R&D Program of China (2021YFF0704100), the National Natural Science Foundation of China (62136002, 61936001 and 61876027), the Science and Technology Research Program of Chongqing Municipal Education Commission (KJQN202100627 and KJQN202100629), and the National Natural Science Foundation of Chongqing (cstc2019jcyj-cxttX0002,cstc2022ycjh-bgzxm0004), respectively.

7 Compliance with ethical standards

Conflict of interest We declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.
Data Availability Statement The datasets generated during and analysed during the current study are available from the corresponding author on reasonable request.

References

[1] Moor, J.H.: The turing test: The elusive standard of artificial intelligence. Comput. Linguistics **30**(1), 115–116 (2004)

[2] Madotto, A., Wu, C., Fung, P.: Mem2seq: Effectively incorporating knowledge bases into end-to-end task-oriented dialog systems. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, pp. 1468–1478 (2018)

[3] Qin, L., Xu, X., Che, W., Zhang, Y., Liu, T.: Dynamic fusion network for multi-domain end-to-end task-oriented dialog. In: Proceedings of the

58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, pp. 6344–6354 (2020)

[4] Liu, Q., Chen, Y., Chen, B., Lou, J., Chen, Z., Zhou, B., Zhang, D.: You impress me: Dialogue generation via mutual persona perception. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, pp. 1417–1427 (2020)

[5] Gangadharaiah, R., Narayanaswamy, B.: Recursive template-based frame generation for task oriented dialog. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, pp. 2059–2064 (2020)

[6] Saha, T., Patra, A.P., Saha, S., Bhattacharyya, P.: Towards emotion-aided multi-modal dialogue act classification. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, pp. 4361–4372 (2020)

[7] Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, pp. 3104–3112 (2014)

[8] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, pp. 5998–6008 (2017)

[9] Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: 2nd International Conference on Learning Representations (2014)

[10] Sohn, K., Lee, H., Yan, X.: Learning structured output representation using deep conditional generative models. In: Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, pp. 3483–3491 (2015)

[11] Wang, Y., Liao, J., Yu, H., Leng, J.: Semantic-aware conditional variational autoencoder for one-to-many dialogue generation. Neural Computing and Applications, 1–13 (2022)

[12] Shi, W., Zhou, H., Miao, N., Li, L.: Dispersed exponential family mixture vaes for interpretable text generation. In: Proceedings of the 37th International Conference on Machine Learning, vol. 119, pp. 8840–8851 (2020)

[13] Bowman, S.R., Vilnis, L., Vinyals, O., Dai, A.M., Józefowicz, R., Bengio, S.: Generating sentences from a continuous space. In: Proceedings of the

20th SIGNLL Conference on Computational Natural Language Learning, pp. 10–21 (2016)

[14] Zhao, T., Lee, K., Eskénazi, M.: Unsupervised discrete sentence representation learning for interpretable neural dialog generation. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, pp. 1098–1107 (2018)

[15] Serban, I.V., Sordoni, A., Lowe, R., Charlin, L., Pineau, J., Courville, A.C., Bengio, Y.: A hierarchical latent variable encoder-decoder model for generating dialogues. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA, pp. 3295–3301 (2017)

[16] Zhao, T., Zhao, R., Eskénazi, M.: 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, pp. 654–664 (2017)

[17] Gao, J., Bi, W., Liu, X., Li, J., Zhou, G., Shi, S.: A discrete CVAE for response generation on short-text conversation. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pp. 1898–1908 (2019)

[18] Pang, B., Wu, Y.N.: Latent space energy-based model of symbol-vector coupling for text generation and classification. In: Proceedings of the 38th International Conference on Machine Learning, vol. 139, pp. 8359–8370 (2021)

[19] Higgins, I., Matthey, L., Pal, A., Burgess, C.P., Glorot, X., Botvinick, M.M., Mohamed, S., Lerchner, A.: beta-vae: Learning basic visual concepts with a constrained variational framework. In: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings (2017)

[20] Mathieu, E., Rainforth, T., Siddharth, N., Teh, Y.W.: Disentangling disentanglement in variational autoencoders. In: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA. Proceedings of Machine Learning Research, vol. 97, pp. 4402–4412 (2019)

[21] Chen, S., Yan, J., Su, Y., Wang, Y.F.: Representation decomposition for image manipulation and beyond. In: 2021 IEEE International Conference on Image Processing, ICIP 2021, Anchorage, AK, USA, September 19-22, 2021, pp. 1169–1173 (2021)

[22] Li, Y., Su, H., Shen, X., Li, W., Cao, Z., Niu, S.: Dailydialog: A manually labelled multi-turn dialogue dataset. In: Proceedings of the Eighth International Joint Conference on Natural Language Processing, pp. 986–995 (2017)

[23] Rashkin, H., Smith, E.M., Li, M., Boureau, Y.: Towards empathetic open-domain conversation models: A new benchmark and dataset. 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, pp. 5370–5381 (2019)

[24] Papineni, K., Roukos, S., Ward, T., Zhu, W.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311–318 (2002)

[25] Lin, C.-Y.: Rouge: A package for automatic evaluation of summaries. In: Text Summarization Branches Out, pp. 74–81 (2004)

[26] Banerjee, S., Lavie, A.: METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation And/or Summarization@ACL 2005, pp. 65–72 (2005)

[27] Li, J., Galley, M., Brockett, C., Gao, J., Dolan, B.: A diversity-promoting objective function for neural conversation models. In: The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 110–119 (2016)

[28] Luong, T., Pham, H., Manning, C.D.: Effective approaches to attention-based neural machine translation. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 1412–1421 (2015)

[29] Wang, Y., Wang, H., Zhang, X., Chaspari, T., Choe, Y., Lu, M.: An attention-aware bidirectional multi-residual recurrent neural network (abmrnn): A study about better short-term text classification. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019, pp. 3582–3586 (2019)

[30] Jiang, Z., Zheng, Y., Tan, H., Tang, B., Zhou, H.: Variational deep embedding: An unsupervised and generative approach to clustering. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, pp. 1965–1972 (2017)

[31] Wang, Y., Zhang, X., Lu, M., Wang, H., Choe, Y.: Attention augmentation with multi-residual in bidirectional LSTM. Neurocomputing **385**, 340–347 (2020)