Correlation Encoder-Decoder Model for Text Generation

Xu Zhang
Qilu University of Technology
(Shandong Academy of Sciences)
Jinan, China
xuzhang.p@foxmail.com

Yifeng Li
Department of Computer Science
Brock University
Niagara Region, Canada
yli2@brocku.ca

Xueping Peng
Australian Artificial Intelligence Institute
University of Technology Sydney
Sydney, Australia
xueping.peng@uts.edu.au

Xinxiao Qiao*
Qilu University of Technology
(Shandong Academy of Sciences)
Jinan, China
qxxyn@qlu.edu.cn

Hui Zhang
Shandong Massclouds Co.,Ltd
Jinan, China
zhang_hui@massclouds.com

Wenpeng Lu*
Qilu University of Technology
(Shandong Academy of Sciences)
Jinan, China
wenpeng.lu@qlu.edu.cn

***Abstract*—Text generation is crucial for many applications in natural language processing. With the prevalence of deep learning, the encoder-decoder architecture is dominantly adopted for this task. Accurately encoding the source information is of key importance to text generation, because the target text can be generated only when accurate and complete source information is captured by the encoder and fed into the decoder. However, most existing approaches fail to effectively encode and learn the entire source information, as some features are easy to be missed along with the encoding procedures of the encoder. Similar problems also confuse the implementation of the decoder. How to reduce the problem of information loss in the encoder-decoder model is critical for text generation. To address this issue, we propose a novel correlation encoder-decoder model, which optimizes both the encoder and the decoder to reduce the problem of information loss by enforcing them to minimize the differences between hierarchical layers by maximizing the mutual information. Experimental results on two benchmark datasets demonstrate that the proposed model substantially outperforms the existing state-of-the-art methods. Our source code is publicly available on GitHub[1].**

I. Introduction

Natural language generation (NLG) is defined originally as a task to generate understandable language expressions from some non-linguistic representation of information [1]. Later, NLG is extended to include text-to-text generation, data-to-text generation, and image-to-text generation, etc. Text generation is fundamental and crucial for many NLG tasks, such as machine translation [2], [3], generative dialogue [4], text style conversion [5].

In recent years, deep learning has achieved exceptional successes in various tasks [6]–[11], and a large number of neural models are applied in text generation, such as the sequence-to-sequence model [12], generative adversarial network [13], and variational autoencoder [14].The essential encoding module employed in these models is recurrent neural network (RNN) or its variants. Because of the vanishing gradient problem in RNN, its ability for capturing long-term dependency and feature representation is limited, which means that some key features may be missing and lost in RNN encoder.

With the model structure of attention and seq2seq fusion, attention mechanism is also used in encoding after the great breakthrough in machine translation [2], [3], [15], [16]. In order to further acquire rich semantic features representation and avoid too much semantic information missing in the encoding process, some research try to use hierarchical encoding and other methods to acquire text encoding features [14], [17]–[19]. Although the limitation of using RNN encoding has been alleviated with the development of research, there are still some problems. No matter what encoding method is used, it is inevitable that there will be a feature representation loss in the process of encoding, which will lead to information loss in the text generated after decoding which may result in changes in the whole text semantic information or make the text semantic information ambiguous.

In order to improve the performance of representation learning, Hjelm et al. have proposed the Deep InfoMax (DIM) model, which employs the mutual information (MI) criterion to maximize the correlation between its inputs and the corresponding learned high-dimensional representations. DIM leverages MI estimation for accurate representation learning and demonstrates a strong potential to achieve better performance in various downstream tasks [20]. Inspired by their work, we first focus on the encoding layer and optimize the intra-hierarchical encoding by maximizing MI between different encoding layers to reduce information loss in the encoding procedure. In addition, we carry out prior modeling of the hidden variables for the decoding layer through the

Xinxiao Qiao and Wenpeng Lu are the corresponding authors. The research work is partly supported by National Natural Science Foundation of China under Grant No.61502259 and No.11901325, Key Program of Science and Technology of Shandong Province under Grant No.2020CXGC010901 and No.2019JZZY020124.

[1]https://github.com/zhangxu90s/CED

Gaussian distribution and optimize the objective function by maximizing MI between the Gaussian prior and the encoding hidden variables, so as to reduce information loss in the decoding procedure. VAE doesn't require that the prior to be Gaussian. Theoretically, it can be an arbitrarily sophisticated distribution [21]. However, it is more difficult to implement sophisticated distributions in VAE. Compared with them, it is easy to implement Gaussian distributions. Besides, based on the mathematical facts, given the mean and variance, the probability distribution that maximizes the information entropy is Gaussian. Considering the above reasons as in other widely used VAE models, we adopt Gaussian prior in our model. In summary, this paper makes the following contributions.

- We propose a novel deep neural architecture for text generation task, **named correlation encoder-decoder model (CED)**. By maximizing the mutual information, CED learns ideal text representations and reduces information loss in the encoder-decoder model. As far as we know, this is the first work for leveraging MI to achieve end-to-end text generation.
- CED employs mutual information (MI) to maximize the correlation between its inputs and the learned high-dimensional representations to reduce information loss of the encoder. Moreover, for the decoder, CED makes prior modeling of the hidden variables for the decoding layer with a Gaussian distribution and optimizes the objective function by maximum MI between the Gaussian prior and the encoding hidden variables.

The reminder of the paper is structured as follows. Some related work is briefly reviewed in Section II. The detailed implementation of our correlation encoder-decoder model is described in Section III. Section IV reports the experiments and results. Finally, the paper is concluded in Section V.

II. Related Work

Natural language processing has many fundamentally challenging tasks such as natural language understanding and generation. With the development of pre-trained models [22], [23], significant improvements have been made in natural language understanding tasks. However, for natural language generation tasks, there is still no universal model for training. We often try to optimize the encoding and decoding modules to further improve the model. Natural language generation technology is widely applied in the key modules of various tasks, such as machine translation [2], [3], story generation [24], dialogue generation [4], and so on.

Bengio et al. attempt to apply the neural language model on the text generation task for the first time, and achieve remarkable performance [25]. Neural language model can be regarded as an extension of the n-gram model, which has a good generalization ability. Since the n-gram model is theoretically impossible to capture long-term dependencies, it is very easy to lose features during the encoding procedure. In order to solve this problem, Mikolov et al. propose a recurrent neural network (RNN) language model [26], which alleviates the problem of long-term dependencies in the encoding procedure to some extent [27]–[29]. Later, some variants of RNN are also applied to the text generation task. However, for the long texts, RNNs still cannot extract the long-term dependency of text semantics well. Some researchers attempt to utilize convolution neural network (CNN) for the encoding module. Both Gehring et al. and Dauphin et al. adopt CNNs instead of RNNs for text generation [30], [31]. CNNs can take advantage of the natural locality of the language and solve the long-term dependency by converting the global dependency to the local dependency on the higher level of the network. CNN has been proved useful in many classic NLP tasks, which can also be applied to text generation tasks through adversarial training [32]. In addition, Bahdanau et al. make a breakthrough in machine translation by combining the attention mechanism into the sequence-to-sequence structure [2]. At the same time, a series of studies have shown that the attention mechanism can effectively improve RNNs for text encoding [15]. Although the above methods solve the long-term dependency problem of RNNs to some extent, it is still challenged by the problem of feature and information loss.

In order to solve the feature and information loss during text encoding, Serban et al. extend the hierarchical recurrent encoder-decoder neural network for text generation task, and demonstrate that this model is competitive with state-of-the-art neural language models and back-off n-gram models [33]. To further optimize the encoding structure, they propose a neural generative architecture that models contexts in a hierarchical encoder-decoder framework, with latent stochastic variables that span a variable number of time steps to model complex dependencies between subsequences [17]. Hierarchical attention mechanism also has a good effect on the encoding module. Chen et al. propose a hierarchical recurrent attention network that models the hierarchy of contexts, word importance, and utterance importance in a unified framework, attending to important parts in contexts in multi-turn response generation [34]. The above methods alleviate the influence of feature representation loss on encoding to some extent but do not fundamentally solve the problem of feature loss.

In order to solve the above problem, we propose a novel deep neural architecture for text generation task, called correlation encoder-decoder model. On the foundation of hierarchical encoding, we employ maximum mutual information to optimize intra-hierarchical encoding. Besides text generation, this method of correlation encoding can be applied to various NLP tasks. In order to further improve the decoding performance, we make prior modeling on the latent variables in the decoding layer with Gaussian distribution, and further optimize the objective function by using the maximum mutual information between the Gaussian prior variable and encoded posterior variable.

III. Correlation Encoder-Decoder Model

The architecture of our CED model is shown in Fig. 1, which includes an embedding layer, an encoder layer, and a decoder layer. For reducing the information loss in the

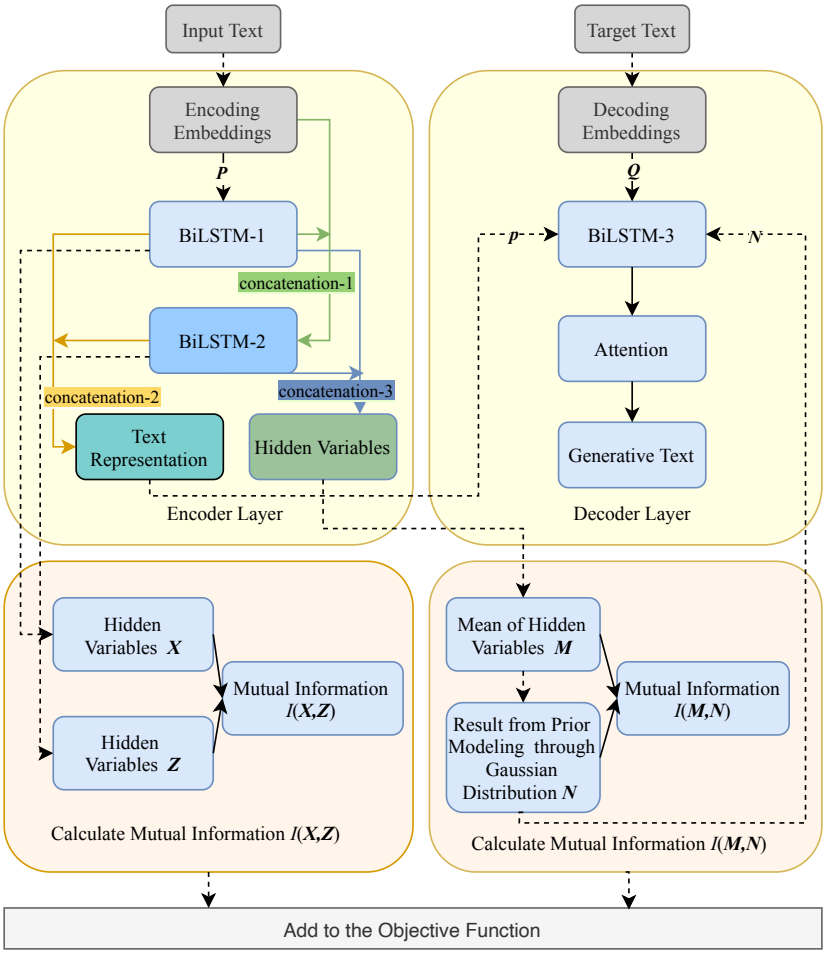

Fig. 1. Architecture of our CED model.

encoding procedure, CED employs MI to maximize the correlation of latent variables between different encoding layers. For reducing the information loss in the decoding procedure, CED makes prior modeling of the latent variables for the decoding layer with a Gaussian distribution and optimizes the objective function by maximum MI between this distribution and the encoding latent variables.

A. Encoding Layer

1) Inter-hierarchical Encoding: A pair of texts is denoted as $< P, Q >$, where P is the input text and Q is the target text. As shown in Fig. 1, we employ two BiLSTM layers as the inter-hierarchical encoding components for P. The outputs of both BiLSTM layers are concatenated as the final text representation. The operation, as formulated in Equ. (1), is applied to $\boldsymbol{P}$ which is the embedding matrix for sentence P.

$$\boldsymbol{p}, \boldsymbol{p}_{hidden} = [\mathrm{BiLSTM}([\mathrm{BiLSTM}(\boldsymbol{P}), \boldsymbol{P}]); \mathrm{BiLSTM}(\boldsymbol{P})], \tag{1}$$

where $\boldsymbol{p}$ and $\boldsymbol{p}_{hidden}$ represent the output and hidden state generated by the inter-hierarchical encoding component for P.

2) Optimize Inter-hierarchical Encoding: In order to reduce information loss during the encoding procedure, we employ MI in our CED model to further optimize the inter-hierarchical encoding.

We first compute MI between the hidden variables from the first and second BiLSTM layers. In our notation, $\boldsymbol{X}$ represents the set of encoding hidden variables in the first BiLSTM layer, $x \in \boldsymbol{X}$ denotes an instance of hidden variables; $\boldsymbol{Z}$ represents the set of encoding hidden variables in the second BiLSTM layer; $z \in \boldsymbol{Z}$ denotes an instance of hidden variables; $p(z|x)$ denotes the conditional distribution of encoding vectors generated by $\boldsymbol{X}$. So, we can represent the correlation of $\boldsymbol{X}$ and $\boldsymbol{Z}$ in the term of MI, as shown in Equ. (2).

$$I(\boldsymbol{X}, \boldsymbol{Z}) = \iint p(z|x)\tilde{p}(x) \log \frac{p(z|x)}{p(z)} dx dz, \tag{2}$$

where $\tilde{p}(x)$ is the distribution of hidden variables from the first BiLSTM layer encoding, $p(z)$ is the marginal distribution of $\boldsymbol{Z}$ as formulated in Equ. (3):

$$p(z) = \int p(z|x)\tilde{p}(x) dx. \tag{3}$$

In order to get a better feature representation, the mutual information should be as large as possible, which is described in Equ. (4).

$$p(z|x) = \max_{p(z|x)} I(\boldsymbol{X}, \boldsymbol{Z}). \tag{4}$$

The larger the mutual information is, the larger $\log \frac{p(z|x)}{p(z)}$ should be. It means that $p(z|x)$ should be as large as possible than $p(z)$. For each x, the encoding module can find a better z which is exclusive to x. Therefore the probability of $p(z|x)$ is much larger than $p(z)$.

In order to maximize the mutual information, Equ. (2) is slightly adjusted to Equ. (5).

$$\begin{aligned} I(\boldsymbol{X}, \boldsymbol{Z}) &= \iint p(z|x)\tilde{p}(x) \log \frac{p(z|x)\tilde{p}(x)}{p(z)\tilde{p}(x)} dxdz \\ &= KL(p(z|x)\tilde{p}(x) \| p(z)\tilde{p}(x)). \end{aligned} \tag{5}$$

This formulation reveals the essential meaning of mutual information. The mutual information is the KL divergence of these two distributions. Therefore, its maximization implies the maximization of the distance between $p(z|x)\tilde{p}(x)$ and $p(z)\tilde{p}(x)$.

Since KL divergence is theoretically unbounded at the upper end, if we attempt to maximize the unbounded quantity, it is unstable and potentially infinite. Therefore, we employ another measure, i.e., JS divergence, which also measures the distance between two distributions, but it has an upper bound $\frac{1}{2}\log 2$, which is defined in Equ. (6):

$$JS(P, Q) = \frac{1}{2} KL\left(P \middle\| \frac{P+Q}{2}\right) + \frac{1}{2} KL\left(Q \middle\| \frac{P+Q}{2}\right), \tag{6}$$

where P and Q are two probability distributions. Next we replace KL divergence in Equ. (6) by the the general local variational inference of f divergence (a convex function) [35] as shown below:

$$\mathcal{D}_f(P\|Q) = \int q(x) f\left(\frac{p(x)}{q(x)}\right) dx. \tag{7}$$

One approach to facilitate optimization of the JS divergence is to make use of the conjugate function $g(T(u))$ for $f(u)$, where the two functions are related by

$$f(u) = \max_{T \in f'(\mathbb{D}), \mathbb{D}=\mathbb{R}_+} \big\{T(u)u - g(T(u))\big\}, \tag{8}$$

where T is a function with a range of $f'(\mathbb{D})$ which is the derivative of f [35]. In our work, the function $T(\cdot)$ will be approximated by a multi-layer perceptron.

Combining Equ. (7) and (8), we get

$$\begin{aligned} &\mathcal{D}_f(P\|Q) = \\ &\max_T \int q(x) \left[\frac{p(x)}{q(x)} T\left(\frac{p(x)}{q(x)}\right) - g\left(T\left(\frac{p(x)}{q(x)}\right)\right)\right] dx \\ &= \max_T \int \left[p(x) \cdot T\left(\frac{p(x)}{q(x)}\right) - q(x) \cdot g\left(T\left(\frac{p(x)}{q(x)}\right)\right)\right] dx. \end{aligned} \tag{9}$$

If $T\left(\frac{p(x)}{q(x)}\right)$ is viewed as the whole $T(x)$, then one can obtain

$$\mathcal{D}_f(P\|Q) = \max_T \Big(\mathbb{E}_{x\sim p(x)}[T(x)] - \mathbb{E}_{x\sim q(x)}[g(T(x))]\Big). \tag{10}$$

Here we implement the generalized KL divergence used in JS. For JS divergence, the task thus becomes

$$\begin{aligned} JS(P, Q) = \max_T \Big(&\mathbb{E}_{x\sim p(x)}[\log \sigma(T(x))] \\ &+ \mathbb{E}_{x\sim q(x)}[\log(1 - \sigma(T(x)))]\Big). \end{aligned} \tag{11}$$

Substituting $p(z|x)\tilde{p}(x)$ and $p(z)\tilde{p}(x)$ into Equ. (11), we can obtain Equ. (12), where $\sigma(T(x, z))$ is a discriminant network. Maximizing the likelihood function is equivalent to maximizing the cross entropy. The mutual information in Equ. (5) can thus be maximized using Equ. (12).

$$\begin{aligned} &JS\big(p(z|x)\tilde{p}(x), p(z)\tilde{p}(x)\big) = \\ &\max_T \Big(\mathbb{E}_{(x,z)\sim p(z|x)\tilde{p}(x)}[\log \sigma(T(x, z))] + \\ &\mathbb{E}_{(x,z)\sim p(z)\tilde{p}(x)}[\log(1 - \sigma(T(x, z)))]\Big). \end{aligned} \tag{12}$$

B. Decoding Layer

1) Optimize Decoding: As shown in Fig. 1, for the decoding layer, we adopt the method applied in the variational encoder-decoder with deterministic attention (VED + DAttn) proposed by Bahuleyan et al. [36]. Based on the work in Section III-A2, we obtain the prior latent variable $\boldsymbol{N}$ and the posterior mean $\boldsymbol{M}$ from the encoding layer. M imitates the mean value of VAE [37]. N means that Gaussian noise is added to the encoder results, which makes that the decoder can be robust to noise. We constrain the noise to keep that it has as little influence on the subsequent decoding effect as possible without losing the robustness. We further optimize variational latent space $\boldsymbol{N}$ with maximum mutual information (MMI) before feeding it into the decoding layer. Since we have given a detailed derivation of the specific solution of MMI in Section III-A2, we do not repeat the derivation process here. The equation is as follows:

$$\begin{aligned} &JS\big(p(n|m)\tilde{p}(m), p(n)\tilde{p}(m)\big) = \\ &\max_T \Big(\mathbb{E}_{(m,n)\sim p(n|m)\tilde{p}(m)}[\log \sigma(T(m, n))] \\ &+ \mathbb{E}_{(m,n)\sim p(n)\tilde{p}(m)}[\log(1 - \sigma(T(m, n)))]\Big). \end{aligned} \tag{13}$$

We can obtain the optimized latent representation $n \in \boldsymbol{N}$ and feed it into the decoding layer. Then the decoder generates corresponding text based on a random sample n drawn from the prior distribution.

2) Decoding Layer: We can obtain the optimized text representation by calculating the inter-layer mutual information as described in Section III-A, and the optimized latent representation as explained in Section III-B, then we put the above two parts to the decoding layer.

$$\boldsymbol{Output} = \mathrm{LuongAttention}(\mathrm{BiLSTM}(\boldsymbol{Q}, \boldsymbol{p}, \boldsymbol{N})), \tag{14}$$

where $\boldsymbol{Q}$ is the target text embedding matrix; the initial state $\boldsymbol{p}$ of BiLSTM is the output vector from the encoding layer; and the latent vector $\boldsymbol{N}$ is the result from prior modeling of the encoding layer's latent variables through Gaussian distribution. The feature representation after BiLSTM decoding is sent to the LuongAttention mechanism [38].

C. Loss Function

The loss function consists of three terms. The first term is VAE's loss function, denoted as $\mathcal{V}$ [37]. The second term is the loss from the correlation encoding layer through MMI, as shown in Equ. (12). The third term is the loss from the decoding layer through MMI between the variational latent space $\boldsymbol{N}$ and the mean $\boldsymbol{M}$ of the output from the encoding layer, as shown in Equ. (13).

For a batch of data samples $Y = \{y_1, \ldots, y_n\}$ (e.g., a sentence)), the encoding layer encodes the data Y as samples $\{m_1, \ldots, m_n\}$ of a hidden random variable $\boldsymbol{M}$, and the decoding layer reconstructs Y. We integrate the above three terms (i.e. Equ. (12), (13), and VAE loss $\mathcal{V}$) together as the loss function of our proposed CED model, as shown in Equ. (15):

$$\begin{aligned} \mathcal{L} = \min_{p(z|x),T(x,z)} \Big\{ \\ \mathcal{V} - \mathbb{E}_{(x,z)\sim p(z|x)\tilde{p}(x)} JS\big(p(z|x)\tilde{p}(x), p(z)\tilde{p}(x)\big). \\ -\mathbb{E}_{(m,n)\sim p(n|m)\tilde{p}(m)} JS\big(p(n|m)\tilde{p}(m), p(n)\tilde{p}(m)\big) \Big\} \end{aligned} \tag{15}$$

IV. Experiment

A. Data Sets

- **SQuAD:** The Stanford Question Answering Dataset is proposed by Rajpurkar et al., and applies to reading comprehension tasks. We apply this data set to the text generation task, aiming to generate question-based on the paragraph. We use the same train-validation-test split as in Du et al. [39] and Bahuleyan et al. [36].
- **DailyDialog:** DailyDialog is a data set of multi-round conversations for everyday chat scenarios collected by Li et al. [40]. We use the data set reassembled by Bahuleyan et al. and their same train-validation-test split [41].

B. Training Details

For the two data sets SQuAD and DailyDialog, we use the same model framework. In case of an inconsistency of some parameters, we will introduce the parameters for each data set in the following paragraphs. All the experiments are performed at a Thinkstation P910 with dual Xeon E5-2600 processors and 192GB memory, equipped with one Nvidia 2080Ti GPU.

- **SQuAD:** In these experiments, the embedding dimension is 300, the encoding dimension is 100 and the latent space dimension is 100. The dropout rate is 0.8 in the encoding layer and the word dropout rate is 0.75 for the decoding layer. We set the batch size to 100 and train 30 epochs.
- **DailyDialog:** In these experiments, the embedding dimension is 300, the encoding dimension is 500 and the latent space dimension is 300. The dropout rate is 0.8 in the encoding layer and the word dropout rate is 0.75 for the decoding layer. We set the batch size to 128 and train 200 epochs.

C. Baseline Methods

- **SQuAD:** Du et al. [39] and Bahuleyan et al. [36] have made a lot of research progress on the generation of SQuAD data set. We use these findings as the baseline for our experiments, like DED, WED, and VED.
- **DailyDialog:** Bahuleyan et al. reconstruct the data set of DailyDialog and conduct a series of experiments on this data set [41]. We select their model as the baseline, such as DED, WED-D, VED, and WAE-S.

Our correlation encoding model (CED) is customized to four versions to evaluate performance, as follows:

- **CED**$^{-e}$: We consider the influence of the correlation from intra-hierarchical layers during encoding process. CED^{-e} reduces the encoding gap between layers by using the maximized mutual information method to reduce the feature representation loss phenomenon in the encoding process.
- **CED**$_{-d}$: We use MMI to optimize the objective function between Gaussian distribution and the mean output of encoding hidden variables, consequently improving the effect of the decoding module.
- **CED**: It is the full version of CED.

D. Results on SQuAD

As shown in Table I, we compare our model with the baselines on SQuAD. In the table, the results for model (1) were obtained by Du et al. [39] and the results for models (2-11) were acquired from Bahuleyan et al. [36].

Our approach is extended from the work of Bahuleyan et al., i.e., VED + DAttn - Sampling [36]. As shown in Table I, both the former two models (1) and (2) represent the traditional vanilla Seq2Seq model. Obviously, such a model could not extract text feature representation effectively either at the encoding or decoding level. In order to improve the performance of encoding and decoding modules, a deterministic attention mechanism and variational encoder-decoder module are used in models (3) and (4). From the table, we can see that compared with model (2), its performance can be effectively improved by using the attention mechanism and stochastic modeling in the latent space. Results for models (4) and (5) are separately obtained by applying maximum a posterior estimation and sampling in the models. Compared with the ordinary training, models (6) and (7) attempt to employ a heuristic of 2-stage training but fails to improve the performance. In models (8-11), they utilize the variational attention to further improve the modeling effect. However, compared with the corresponding models using deterministic attention, their experimental performance do not show any advantage.

In the above experiments, the best benchmark is achieved by model (5). We improve the original model (5) to realize our CED model.

Models	BLEU-1	BLEU-2	BLEU-3	BLEU-4
1. DED(w/o Attn)	31.34	13.79	7.36	4.26
2. DED(w/0 Attn) - MAP	29.31	12.42	6.55	3.61
3. DED + DAttn - MAP	30.24	14.33	8.26	4.96
4. WED + DAttn - MAP	31.02	14.57	8.49	5.02
5. WED + DAttn - Sampling	30.87	14.71	8.61	5.08
6. WED + DAttn (2-stage) - MAP	28.88	13.02	7.33	4.16
7. WED + DAttn (2-stage) - Sampling	29.25	13.21	7.45	4.25
8. VED + VAttn-0 - MAP	29.70	14.17	8.21	4.92
9. VED + VAttn-0 - Sampling	30.22	14.22	8.28	4.87
10. VED + VAttn-$\overline{h}$ - MAP	30.23	14.30	8.28	4.93
11. VED + VAttn-$\overline{h}$ - Sampling	30.47	14.35	8.39	4.96
12. CED^{-e}	32.08	15.74	9.28	5.65
13. CED^{-d}	32.24	15.78	9.33	5.72
14. CED	**32.53**	**16.02**	**9.52**	**5.86**

TABLE I
EXPERIMENTAL RESULTS ON SQUAD DATASET.

Specially, we add the maximum mutual information criterion to alleviate the information loss of feature representation, so as to improve the performance of the model. The outstanding experimental results in Table I indicate that CED can extract text features and generate text representation effectively.

Comparing the results for models (12-14), both ablation variants are inferior to the standard CED model, which demonstrates the necessity of MI mechanism in encoding and decoding layers.

E. Results on DailyDialog

In Table II, we evaluate the effectiveness of our models for dialogue generation as natural language a generation task. Other than our baselines, we compare with Bahuleyan et al. [41]. Compared to the work from Bahuleyan et al., our models significantly improve the dialogue generation task.

Compared with SQuAD data set, the best model for dialogue generation data set is CED$_{-d}$. CED is not as effective as CED$_{-d}$, mainly because the text from dialogue generation data set is relatively short. In addition, the text structure is not very neat compared with SQuAD data set, and the oral characteristics are more serious. When we consider the correlation from intra-hierarchical layers, the encoding effect is easily disturbed by these factors. Because using CED^{-e} model, we can capture more noise in the encoding process, and then affect the effect of decoding layer. Through the comparison of experimental results of CED^{-e}, CED$_{-d}$ and CED, we can see that the introduction of CED$_{-d}$ can optimize model CED^{-e} and further reduce the influence of interference factors in the encoding process.

Models	BLEU-1	BLEU-2	BLEU-3	BLEU-4
1. DED	-	3.96	-	0.85
2. VED	-	3.26	-	0.59
3. WED-D	-	4.05	-	0.98
4. WAE-S	-	3.72	-	0.69
5. CED^{-e}	17.74	6.81	3.95	2.72
6. CED^{-d}	**18.27**	**7.09**	**4.17**	**2.89**
7. CED	18.05	6.95	4.06	2.83

TABLE II
EXPERIMENTAL RESULTS ON DAILYDIALOG DATASET.

V. CONCLUSION

Precise text generation relies on effective encoder-decoder architecture. In this paper, we propose the correlation encoder-decoder model, namely CED. By maximizing the mutual information, CED optimizes both the encoder and the decoder to reduce information loss to improve the performance of text generation. In the encoding module, the difference of encoding features between layers is minimized by maximum MI, thus alleviating the problem of information loss. In the decoding module, we make prior modeling of the hidden variables with the Gaussian distribution and optimize the objective function by maximizing MI to improve the decoding quality. Extensive experiments on two datasets verify that our proposed CED model outperforms the existing state-of-the-art methods. Our method is workable and appliable for the classical variational encoder-decoder based models. However, currently, its architecture is unable to integrate with Transformer/BERT-based models [22], [42]. Therefore, we didn't compare it with these giant models. Yet, it is a very interesting question to integrate our model with Transformer/BERT-based models, which will be our future work along with scalability studies.

REFERENCES

[1] E. Reiter and R. Dale, "Building applied natural language generation systems," *Natural Language Engineering*, vol. 3, no. 1, pp. 57–87, 1997.

[2] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *3rd International Conference on Learning Representations*, 2015.

[3] J. Zhang, H. Luan, M. Sun, F. Zhai, J. Xu, M. Zhang, and Y. Liu, "Improving the transformer translation model with document-level context," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 533–542.

[4] Y. Wu, W. Wu, D. Yang, C. Xu, and Z. Li, "Neural response generation with dynamic vocabularies," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[5] C. Wu, X. Ren, F. Luo, and X. Sun, "A hierarchical reinforced sequence operation method for unsupervised text style transfer," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 4873–4883.

[6] X. Peng, Y. Cao, and Z. Niu, "Mining web access log for the personalization recommendation," in *2008 International Conference on MultiMedia and Information Technology*. IEEE, 2008, pp. 172–175.

[7] W. Lu, Y. Zhang, S. Wang, H. Huang, Q. Liu, and S. Luo, "Concept representation by learning explicit and implicit concept couplings," *IEEE Intelligent Systems*, vol. 36, no. 1, pp. 6–15, 2021.

[8] Y. Zhao, Z. Niu, X. Peng, and L. Dai, "A discretization algorithm of numerical attributes for digital library evaluation based on data mining technology," in *International Conference on Asian Digital Libraries*. Springer, 2011, pp. 70–76.

[9] Y.-J. Cao, Z.-D. Niu, K. Zhao, and X.-P. Peng, "Near duplicated web pages detection based on concept and semantic network," *Ruanjian Xuebao/Journal of Software*, vol. 22, no. 8, pp. 1816–1826, 2011.

[10] S. Wang, L. Hu, L. Cao, X. Huang, D. Lian, and W. Liu, "Attention-based transactional context embedding for next-item recommendation," in *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI 2018)*. Association for the Advancement of Artificial Intelligence, 2018, pp. 2532–2539.

[11] X. Zhang, Y. Li, W. Lu, P. Jian, and G. Zhang, "Intra-correlation encoding for chinese sentence intention matching," in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 5193–5204.

[12] Z. Li, R. Wang, K. Chen, M. Utiyama, E. Sumita, Z. Zhang, and H. Zhao, "Data-dependent gaussian prior objective for language generation," in *International Conference on Learning Representations*, 2019.

[13] Z. Liu, J. Wang, and Z. Liang, "CatGAN: Category-aware generative adversarial networks with hierarchical evolutionary learning for category text generation," *arXiv preprint arXiv:1911.06641*, 2019.

[14] Z. Shao, M. Huang, J. Wen, W. Xu *et al.*, "Long and diverse text generation with planning-based hierarchical variational model," 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, pp. 3248–3259.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[16] F. Yuan, S. Chen, L. Xu, and et al, *Research on the coordination mechanism of traditional Chinese medicine medical record data standardization and characteristic protection under big data environment*. Shandong People's Publishing House, 2021.

[17] I. V. Serban, A. Sordoni, R. Lowe, L. Charlin, J. Pineau, A. Courville, and Y. Bengio, "A hierarchical latent variable encoder-decoder model for generating dialogues," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, 2017, pp. 3295–3301.

[18] W. Lu, X. Zhang, H. Lu, and F. Li, "Deep hierarchical encoding model for sentence semantic matching," *Journal of Visual Communication and Image Representation*, vol. 71, p. 102794, 2020.

[19] X. Zhang, W. Lu, F. Li, X. Peng, and R. Zhang, "Deep feature fusion model for sentence semantic matching," *Computers, Materials and Continua*, 2019.

[20] R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, and Y. Bengio, "Learning deep representations by mutual information estimation and maximization," in *Proceedings of the International Conference on Learning Representations*, 2019.

[21] C. Doersch, "Tutorial on variational autoencoders," *arXiv preprint arXiv:1606.05908*, 2016.

[22] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019, pp. 4171–4186.

[23] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," in *Advances in Neural Information Processing Systems*, 2019, pp. 5754–5764.

[24] A. Fan, M. Lewis, and Y. Dauphin, "Strategies for structuring story generation," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 2650–2660.

[25] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, "A neural probabilistic language model," *Journal of Machine Learning Research*, pp. 1137–1155, 2003.

[26] T. Mikolov, M. Karafiát, L. Burget, J. Černockỳ, and S. Khudanpur, "Recurrent neural network based language model," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010, pp. 1045–1048.

[27] Y. Song, J. Lu, H. Lu, and G. Zhang, "Learning data streams with changing distributions and temporal dependency," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[28] H. Yu, J. Lu, A. Liu, B. Wang, R. Li, and G. Zhang, "Real-time prediction system of train carriage load based on multi-stream fuzzy learning," *IEEE Transactions on Intelligent Transportation Systems*, 2022.

[29] Y. Song, J. Lu, H. Lu, and G. Zhang, "Fuzzy clustering-based adaptive regression for drifting data streams," *IEEE Transactions on Fuzzy Systems*, vol. 28, no. 3, pp. 544–557, 2019.

[30] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *Proceedings of the 34th International Conference on Machine Learning*, 2017, p. 933–941.

[31] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin, "Convolutional sequence to sequence learning," in *Proceedings of the 34th International Conference on Machine Learning*, 2017, p. 1243–1252.

[32] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," in *Advances in Neural Information Processing Systems*, 2017, pp. 5767–5777.

[33] I. V. Serban, A. Sordoni, Y. Bengio, A. Courville, and J. Pineau, "Building end-to-end dialogue systems using generative hierarchical neural network models," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, 2016, pp. 3776–3783.

[34] C. Xing, Y. Wu, W. Wu, Y. Huang, and M. Zhou, "Hierarchical recurrent attention network for response generation," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[35] S. Nowozin, B. Cseke, and R. Tomioka, "f-GAN: Training generative neural samplers using variational divergence minimization," in *Advances in Neural Information Processing Systems*, 2016, pp. 271–279.

[36] H. Bahuleyan, L. Mou, O. Vechtomova, and P. Poupart, "Variational attention for sequence-to-sequence models," in *Proceedings of the 27th International Conference on Computational Linguistics*, 2018, pp. 1672–1682.

[37] R. Lopez, J. Regier, M. I. Jordan, and N. Yosef, "Information constraints on auto-encoding variational Bayes," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018, pp. 6117–6128.

[38] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 1412–1421.

[39] X. Du and C. Cardie, "Identifying where to focus in reading comprehension for neural question generation," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 2067–2073.

[40] Y. Li, H. Su, X. Shen, W. Li, Z. Cao, and S. Niu, "DailyDialog: A manually labelled multi-turn dialogue dataset," in *Proceedings of the Eighth International Joint Conference on Natural Language Processing*, 2017, pp. 986–995.

[41] H. Bahuleyan, L. Mou, H. Zhou, and O. Vechtomova, "Stochastic wasserstein autoencoder for probabilistic sentence generation," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019, pp. 4068–4076.

[42] Y.-H. Chan and Y.-C. Fan, "A recurrent bert-based model for question generation," in *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, 2019, pp. 154–162.