Enhancing Variational Autoencoders with Mutual Information Neural Estimation for Text Generation

Dong Qian, William K. Cheung
Department of Computer Science
Hong Kong Baptist University
{dongqian,william}@comp.hkbu.edu.hk

Abstract

While broadly applicable to many natural language processing (NLP) tasks, variational autoencoders (VAEs) are hard to train due to the posterior collapse issue where the latent variable fails to encode the input data effectively. Various approaches have been proposed to alleviate this problem to improve the capability of the VAE. In this paper, we propose to introduce a mutual information (MI) term between the input and its latent variable to regularize the objective of the VAE. Since estimating the MI in the high-dimensional space is intractable, we employ neural networks for the estimation of the MI and provide a training algorithm based on the convex duality approach. Our experimental results on three benchmark datasets demonstrate that the proposed model, compared to the state-of-the-art baselines, exhibits less posterior collapse and has comparable or better performance in language modeling and text generation. We also qualitatively evaluate the inferred latent space and show that the proposed model can generate more reasonable and diverse sentences via linear interpolation in the latent space.

1 Introduction

Deep learning architectures are parameterized by families of non-linear functions, which learn multiple levels of more abstract representations (Bengio, 2009; Bengio et al., 2013). Recently, there has been a surge of interest in deep generative models for unsupervised learning, such as variational autoencoders (VAEs) (Kingma and Welling, 2014; Rezende et al., 2014), generative adversarial networks (GANs) (Goodfellow et al., 2014), and autoregressive models (van den Oord et al., 2016). The goal is to learn a compact representation to capture the salient structure in a given highly complex high-dimensional unlabelled data so that new data with some variations can be generated. They have been widely applied to a range of NLP tasks, such as language modeling (Bowman et al., 2016; Zhao et al., 2018a), dialog generation (Zhao et al., 2017, 2018b), etc. In this paper, we focus on the VAE with recurrent neural networks as its encoder and decoder for text generation.

Recurrent neural networks (RNNs) are one of the state-of-the-art autoregressive models for text modeling (Melis et al., 2018; Trinh et al., 2018). Training RNNs involves factorizing the joint distribution over a set of random variables into a series of conditional distributions. Yet, the one-step-ahead predictions force RNNs to learn local correlations, rather than global coherence. This is insufficient to capture high-level abstractions which characterize text sequences.

One approach to circumvent this challenge is to introduce variational autoencoders (VAEs) (Bowman et al., 2016), where an encoder learns a latent variable from text sequences and a decoder takes advantage of the variable to reconstruct word-level details. The inferred latent variable in the continuous space captures the semantics and syntactics of text sequences and text generation can be performed in the latent space. Although the approach is theoretically elegant, training the VAE often suffers from the well-known *posterior collapse* issue. It tends to ignore the latent variable and the resulting model reduces to a language model. The issue is especially challenging when a powerful autoregressive decoder (e.g., RNNs) is adopted (Bowman et al., 2016). Many recent efforts have been made to address this problem by modifying the objective (Zhao et al., 2019; Dieng et al., 2019; Wang and Wang, 2019), the decoder architecture (Yang et al., 2017), and the variational inference procedure (He et al., 2019; Fu et al., 2019).

The posterior collapse issue can also be understood from the fact that word-level details tend to carry more entropy bits than semantically-relevant concepts. It is well known that most of the changes in the word level to ensure grammatical coherence

*Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing
and the 9th International Joint Conference on Natural Language Processing*, pages 4047–4057,
Hong Kong, China, November 3–7, 2019. ©2019 Association for Computational Linguistics

do not quickly change the underlying semantics. The maximum-likelihood objective prefers to capture entropy bits at the word level, rather than how well high-level semantics are encoded in the latent space. Training the VAE thus often ends up with relying solely on the autoregressive properties of text sequences, leaving the latent variable unused. Some previous studies (Chen et al., 2017; Alemi et al., 2018) have showed that optimizing the original objective of the VAE is deviated from the goal of learning a good representation.

We consider to address the posterior collapse issue by regularizing the latent space so that the inferred posterior of the latent variable is more specific to its own input, instead of only sticking to the prior. This can be done by maximizing the mutual information (MI) between the input and its latent variable. The MI not only measures how accurate the output agrees with its input, but also whether a meaningful latent variable can be learned in the latent space. However, the estimation and maximization of the MI in the high-dimensional space are difficult. A recent line of work (van den Oord et al., 2018; Belghazi et al., 2018; Hjelm et al., 2019; Veličković et al., 2019) tries to estimate and maximize the MI using the GAN-based approach, where the MI is defined as the Jensen-Shannon divergence (JSD) between the joint distribution and marginals. In this paper, we propose to add a mutual information regularization term, defined as the Kullback-Leibler (KL) divergence, into the objective of the VAE to alleviate the posterior collapse. Different from the models proposed in (Zhao et al., 2019; Dieng et al., 2019), we make effective use of the expressive power of neural networks for the tractable estimation and maximization of the MI in the high-dimensional space. Our empirical results demonstrate that the proposed model can achieve better modeling quality.

Our contributions are summarized as follows:

- We introduce a mutual information term into the objective of the VAE. The latent space can be learned by promoting the posteriors to be more specific to its own inputs, thus leading to low-redundant and diverse representations.

- We establish a lower bound on the mutual information via introducing an energy function paramterized by neural networks. An effective learning algorithm is derived for the unbiased estimation of the gradients.

- We empirically show that the proposed model improves the performance in language modeling and text generation.

2 Related Work

It has been observed that VAEs tend to suffer from the posterior collapse issue, especially when powerful autoregressive decoders are used for modeling text sequences. A common solution is to warm up the KL term in the objective of the VAE by gradually increasing the KL weight from 0 to 1 during training (Bowman et al., 2016). Other proposed methods include randomly dropping words during decoding (Bowman et al., 2016), thresholding the KL term to retain free bits (Kingma et al., 2016; Razavi et al., 2019; Pelsmaeker and Aziz, 2019), adding auxiliary objectives to ensure the effective latent variable for decoding (Goyal et al., 2017; Zhao et al., 2017; Dieng et al., 2019), imposing conditional independence assumptions on the inputs of autoregressive decoders to limit the contextual capacity (Yang et al., 2017; Semeniuta et al., 2017), replacing the Gaussian distribution with the von Mises-Fisher distribution to obtain a fixed KL term (Xu and Durrett, 2018). Also, the semi-amortized approach (Kim et al., 2018) is proposed to perform stochastic variational inference on top of amortized variational inference. Yet, an effective latent variable is learned at the cost of the computational complexity. Some recent work (He et al., 2019; Fu et al., 2019) attempts to improve the training procedure of the VAE without changing its original objective.

Another recent thread of research studies has focused on enhancing the dependency between the input and its latent variable. The skip connection was introduced in (Dieng et al., 2019) to enforce stronger links between the latent variable and the likelihood. InfoVAE (Zhao et al., 2019) introduces a mutual information term, approximated by maximum mean discrepancy, into the objective of the VAE. A similar approach was proposed in (Zhao et al., 2018b) for learning discrete representations.

Recent years have seen some attempts for unsupervised learning through the use of the mutual information. Neural estimation for the MI was proposed in (Brakel and Bengio, 2017), where an encoder and a discriminator are trained to minimize the JSD-based MI. This framework has been further improved and extended, including DeepInfoMax (Hjelm et al., 2019) for image classification and Deep Graph InfoMax (Veličković et al., 2019) for graph-structured data. Contrastive predicting coding (van den Oord et al., 2018) is an architecture to learn global representations by maximizing the mutual information between the past and future information. Other than the neural estimation approach, the Monte Carlo method can also be used to estimate the MI (Zhao et al., 2018b; Dieng et al., 2019). Yet it suffers from biased estimation and high computational cost.

3 Background

3.1 Variational Autoencoders (VAEs)

Deep generative models define a joint distribution over a set of random variables composed in multiple layers of hierarchies. The goal is to learn the model distribution $p_\theta(\mathbf{x})$ to fit the true data distribution $q(\mathbf{x})$ as well as possible. This can be done by minimizing the Kullback-Leibler (KL) divergence between two distributions, and it is equivalent to the maximum-likelihood objective:

$$\min_\theta \mathrm{KL}[q(\mathbf{x})||p_\theta(\mathbf{x})] = \max_\theta \mathbb{E}_{q(\mathbf{x})}[\log p_\theta(\mathbf{x})]. \tag{1}$$

In this paper, we focus on latent variable models, which define the marginal log-likelihood via a latent variable $\mathbf{z}$:

$$\log p_\theta(\mathbf{x}) = \log \int p_\theta(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z}. \tag{2}$$

Since exactly estimating $\log p_\theta(\mathbf{x})$ is typically intractable, the VAE instead optimizes a tractable Evidence Lower BOund (ELBO):

$$\begin{aligned}&\mathbb{E}_{q(\mathbf{x})}[\log p_\theta(\mathbf{x})] \geq \mathcal{L}_{\text{ELBO}} = \\ &\mathbb{E}_{q(\mathbf{x})}\{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \mathrm{KL}[q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})]\},\end{aligned} \tag{3}$$

where the variational posterior $q_\phi(\mathbf{z}|\mathbf{x})$ and the generative distribution $p_\theta(\mathbf{x}|\mathbf{z})$ are parameterized by neural networks (also known as the encoder and the decoder). The prior $p(\mathbf{z})$ is assumed a standard Gaussian distribution $\mathcal{N}(0, \mathbf{I})$.

The ELBO consists of a reconstruction likelihood term that ensures $q_\phi(\mathbf{z}|\mathbf{x})$ to encode *enough* information to generate $\mathbf{x}$ and a KL regularizer that ensures $q_\phi(\mathbf{z}|\mathbf{x})$ to encode *little* information in the posterior by matching it to the prior. Ideally, the encoder would embed high-level abstractions of $\mathbf{x}$ into the latent variable $\mathbf{z}$ and guide the decoder to recover low-level details based on $\mathbf{z}$.

To generate text sequences $\mathbf{x} = \{x_1, \ldots, x_T\}$ with length T, the VAE encodes holistic properties of text sequences $\mathbf{x}$ into the latent variable $\mathbf{z}$. RNNs are used for both the encoder and decoder. For encoding, the last hidden state of the encoder RNN is mapped into the latent variable $\mathbf{z}$. For decoding, $\mathbf{z}$ is fed as an additional input to the decoder RNN at each step, and then the next word x_t is generated conditional on the latent variable $\mathbf{z}$ and all preceding information $x_{<t}$.

3.2 Posterior Collapse

The ELBO will be tighter via optimizing parameters $\{\phi, \theta\}$. The optimal ELBO should be equal to the true data distribution, and thus

$$\mathcal{L}_{\text{ELBO}} \leq \mathbb{E}_{q(\mathbf{x})}[\log p_\theta(\mathbf{x})] \leq \mathbb{E}_{q(\mathbf{x})}[\log q(\mathbf{x})]. \tag{4}$$

If the autoregressive decoder $p_\theta(\mathbf{x}|\mathbf{z})$ is powerful enough to approximate well the true data distribution $q(\mathbf{x})$, the ELBO would be

$$\mathcal{L}_{\text{ELBO}} = \mathbb{E}_{q(\mathbf{x})}\{\log q(\mathbf{x}) - \mathrm{KL}[q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})]\}. \tag{5}$$

The optimal ELBO forces the variational posterior to be like the prior, regardless of how expressively $q_\phi(\mathbf{z}|\mathbf{x})$ is parameterized. The zero-forcing effect of the KL term causes two undesirable outcomes. (*i*) The latent variable $\mathbf{z}$ is independent of the input $\mathbf{x}$, meaning that it contains no information about $\mathbf{x}$. (*ii*) The reconstruction of $\mathbf{x}$ cannot benefit from the encoder at all, meaning that the VAE generates sequences without making use of the latent variable $\mathbf{z}$. This phenomenon is called *posterior collapse*.

It has been proposed in (Hoffman and Johnson, 2016) that decomposing the ELBO can produce an equivalent way to define the ELBO:

$$\begin{aligned}\mathcal{L}_{\text{ELBO}} = \mathbb{E}_{q_\phi(\mathbf{x},\mathbf{z})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \mathbb{I}_q[\mathbf{x}, \mathbf{z}] \\ -\mathrm{KL}[q_\phi(\mathbf{z})||p(\mathbf{z})],\end{aligned} \tag{6}$$

where $q_\phi(\mathbf{x}, \mathbf{z})$ denotes the *variational joint distribution* induced by the posterior $q_\phi(\mathbf{z}|\mathbf{x})$, $\mathbb{I}_q[\mathbf{x}, \mathbf{z}]$ denotes the mutual information between $\mathbf{x}$ and $\mathbf{z}$ under $q_\phi(\mathbf{x}, \mathbf{z})$, and $q_\phi(\mathbf{z}) = \frac{1}{N}\sum_{n=1}^{N} q_\phi(\mathbf{z}_n|\mathbf{x}_n)$ denotes the *aggregated posterior* (Makhzani et al., 2016). The zero-forcing effect of the KL indicates

$$\mathbb{I}_q[\mathbf{x}, \mathbf{z}] = \mathrm{KL}[q_\phi(\mathbf{z})||p(\mathbf{z})] = 0. \tag{7}$$

However, if the latent variable $\mathbf{z}$ is a good representation of $\mathbf{x}$, the mutual information between $\mathbf{x}$ and $\mathbf{z}$ should take a large value and the KL term in Equation (3) would be non-zero. As the ELBO objective contains the negative term of the mutual information, achieving higher mutual information between $\mathbf{x}$ and $\mathbf{z}$ is in fact opposite to maximizing the ELBO. This provides another way to explain the posterior collapse issue.

3.3 Mutual Information

Mutual information measures the non-linear relationship between two variables $\mathbf{x}$ and $\mathbf{y}$:

$$\mathbb{I}[\mathbf{x}, \mathbf{y}] = \mathbb{E}_{p(\mathbf{x},\mathbf{y})}\left[\log \frac{p(\mathbf{x}, \mathbf{y})}{p(\mathbf{x})p(\mathbf{y})}\right], \tag{8}$$

where the MI is minimum if two random variables are statistically independent, or maximum when two variables contain identical information. So if the MI is high, the variables are highly predictive of each other.

This hints that we can learn a stochastic encoder $q_\phi(\mathbf{z}|\mathbf{x})$ to maximize the MI between the input $\mathbf{x}$ and its latent variable $\mathbf{z}$ under the variational joint distribution $q_\phi(\mathbf{x}, \mathbf{z})$, given by

$$\max_\phi \mathbb{I}_q[\mathbf{x}, \mathbf{z}] = \max_\phi \mathbb{E}_{q_\phi(\mathbf{x},\mathbf{z})}\left[\log \frac{q_\phi(\mathbf{x}, \mathbf{z})}{q(\mathbf{x})q_\phi(\mathbf{z})}\right], \tag{9}$$

where $q_\phi(\mathbf{z}|\mathbf{x})$ and $q_\phi(\mathbf{z})$ are distributions over the latent variable. This leads to a decoder-free model for the maximization of the MI. Yet the MI is hard to compute, as $q_\phi(\mathbf{z})$ involves a mixture of the data points. The Monte Carlo approach has been used to approximate the MI in the VAE (Dieng et al., 2019), which however incurs a high computational cost and easily suffers from biased estimation.

4 Proposed Approach

4.1 Model Formulation

As shown in Equation (6), maximizing the ELBO penalizes the mutual information between the input $\mathbf{x}$ and its latent variable $\mathbf{z}$. Due to the posterior collapse issue, the latent variable $\mathbf{z}$ does not represent high-level abstractions of $\mathbf{x}$, making the representation learning even harder. We explicitly introduce a mutual information regularization term into the original ELBO objective that prefers high mutual information between $\mathbf{x}$ and $\mathbf{z}$. This encourages the model to make effective use of the latent variable and alleviate the posterior collapse issue. More formally, we arrive at the following training objective:

$$\begin{aligned}\mathcal{L}_{\text{MI-ELBO}} =& \mathbb{E}_{q(\mathbf{x})}\{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z}) \\ & - \mathrm{KL}[q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})]\} + \alpha\mathbb{I}_q[\mathbf{x}, \mathbf{z}],\end{aligned} \tag{10}$$

where the first two terms can be optimized through the reparameterization trick as in the ELBO and α is a hyperparameter, which balances the trade-off between the inference and generation. We choose to compute the mutual information under the variational joint distribution, as it focuses on the latent space and thus allows us to organize the latent space. The consequence is that the variational posteriors would be more diverse in the latent space characterizing different input sequences, while the KL regularizer restricts the posteriors to match the Gaussian prior, with less “holes” in between where the decoder cannot be trained.

There has been prior work leveraging the mutual information for improving the inference and generation in deep generative models. Some studies (Chen et al., 2016; Dieng et al., 2019) have attempted to maximize the mutual information under the *generative joint distribution* $p_\theta(\mathbf{x}, \mathbf{z})$. In this paper, we focus on explicitly maximizing the mutual information under the variational joint distribution $q_\phi(\mathbf{x}, \mathbf{z})$, which encourages the VAE to learn a useful latent variable $\mathbf{z}$ for improving word generation. Our proposed model is similar in the spirit to the InfoVAE (Zhao et al., 2019). Instead, we adopt neural networks to estimate the MI accurately in the high-dimensional space.

4.2 Mutual Information Formulation

The Jensen-Shannon divergence was employed to define the MI between two variables (Brakel and Bengio, 2017; Hjelm et al., 2019; Veličković et al., 2019). In this paper, we instead maximize the MI, defined by the KL divergence between the joint distribution and marginals. Due to the asymmetric discrepancy, the dependencies between the input and its latent variable can be enhanced. More formally, we define an energy function over the variational joint distribution to estimate the probability of each configuration between two variables (Bengio, 2009), given by:

$$q_\phi(\mathbf{x}, \mathbf{z}) = e^{f_\psi(\mathbf{x},\mathbf{z})}q(\mathbf{x})q_\phi(\mathbf{z})/Z_\psi, \tag{11}$$

where the energy function $f_\psi(\mathbf{x}, \mathbf{z})$ is parameterized by neural networks and the partition function Z_ψ is defined as $\mathbb{E}_{q(\mathbf{x})q_\phi(\mathbf{z})}[e^{f_\psi(\mathbf{x},\mathbf{z})}]$.

Based on Equations (9) and (11), a lower bound on the mutual information $\mathbb{I}_q[\mathbf{x}, \mathbf{z}]$ is given by:

$$\begin{aligned}
&\mathbb{I}_q[\mathbf{x}, \mathbf{z}] = \\
&\mathbb{E}_{q_\phi(\mathbf{x},\mathbf{z})}[f_\psi(\mathbf{x}, \mathbf{z})] - \log\Big[\mathbb{E}_{q(\mathbf{x})q_\phi(\mathbf{z})}[e^{f_\psi(\mathbf{x},\mathbf{z})}]\Big] \geq \\
&\mathbb{E}_{q_\phi(\mathbf{x},\mathbf{z})}[f_\psi(\mathbf{x}, \mathbf{z})] - \xi \cdot \mathbb{E}_{q(\mathbf{x})q_\phi(\mathbf{z})}\Big[e^{f_\psi(\mathbf{x},\mathbf{z})}\Big] + \log \xi + 1,
\end{aligned} \tag{12}$$

where the concave function $\log[x]$ is approximated by a tangent line with the scope ξ according to the Taylor expansion. The approximation is $\log[x] \leq \xi \cdot x - \log[\xi] - 1$, where different values of ξ correspond to different tangent lines. The role of the energy function $f_\psi(\mathbf{x}, \mathbf{z})$ is similar to the *discriminator* in GANs, which learns to distinguish pairs that are sampled from $q(\mathbf{x})q_\phi(\mathbf{z}|\mathbf{x})$ or pairs that are independently sampled from $q(\mathbf{x})q_\phi(\mathbf{z})$. The maximization of the lower bound suggests that the energy function learns to assign higher values to the samples from $q(\mathbf{x})q_\phi(\mathbf{z}|\mathbf{x})$, rather than those from $q(\mathbf{x})q_\phi(\mathbf{z})$, thus enhancing the dependencies.

Since after computing the gradient derivative of $\log[\cdot]$ with respect to parameters ψ, the denominator would contain the expectation of the gradients, leading to biased estimation of the full batch gradients. To solve this, we approximate $\log[\cdot]$ with a tangent line for unbiased estimation of the gradients with respect to the optimal function $f^*_\psi(\mathbf{x}, \mathbf{z})$.

In order to achieve the tighter lower bound on the MI and unbiased estimation for the gradients, we use the convex duality approach (Jordan et al., 1999). The update procedure of model parameters $\{\phi, \psi\}$ can be described as follows.

(*i*). Fixing ϕ and ψ, we optimize Equation (12) with respect to ξ, and the optimal value ξ^* can thus be obtained by

$$\begin{aligned}
\xi^* &= \underset{\xi>0}{\operatorname{argmax}}\Big\{-\xi \cdot \mathbb{E}_{q(\mathbf{x})q_\phi(\mathbf{z})}[e^{f_\psi(\mathbf{x},\mathbf{z})}] + \log \xi\Big\} \\
&= \mathbb{E}_{q(\mathbf{x})q_\phi(\mathbf{z})}\Big[e^{f_\psi(\mathbf{x},\mathbf{z})}\Big].
\end{aligned} \tag{13}$$

(*ii*). Fixing ξ, we optimize Equation (12) with respect to ϕ and ψ:

$$\max_{\phi,\psi}\Big\{\mathbb{E}_{q_\phi(\mathbf{x},\mathbf{z})}[f_\psi(\mathbf{x}, \mathbf{z})] - \xi^* \cdot \mathbb{E}_{q(\mathbf{x})q_\phi(\mathbf{z})}[e^{f_\psi(\mathbf{x},\mathbf{z})}]\Big\}. \tag{14}$$

Here, we first consider the optimal energy function $f^*_\psi(\mathbf{x}, \mathbf{z})$ for any given posterior $q_\phi(\mathbf{z}|\mathbf{x})$ and tangent scope ξ.

Proposition 1. With the fixed ϕ and ξ, the optimal energy function $f^*_\psi(\mathbf{x}, \mathbf{z})$ according to the objective in Equation (12) is given by

$$f^*_\psi(\mathbf{x}, \mathbf{z}) = \log q_\phi(\mathbf{x}, \mathbf{z}) - \log[q(\mathbf{x})q_\phi(\mathbf{z})] - \log \xi, \tag{15}$$

where $f^*_\psi(\mathbf{x}, \mathbf{z})$ becomes the *pointwise mutual information* when $\xi = 1$. This suggests that the energy function assigns zero probability to the samples independently drawn from $q(\mathbf{x})q_\phi(\mathbf{z})$.

With the optimal energy function $f^*_\psi(\mathbf{x}, \mathbf{z})$, the max-max objective in Equation (12) can be reformulated as:

$$\begin{aligned}
&C(\phi, \psi^*) = \\
&\mathbb{E}_{q_\phi(\mathbf{x},\mathbf{z})}[f^*_\psi(\mathbf{x}, \mathbf{z})] - \xi \cdot \mathbb{E}_{q(\mathbf{x})q_\phi(\mathbf{z})}\Big[e^{f^*_\psi(\mathbf{x},\mathbf{z})}\Big] + \log \xi + 1 \\
&= \mathrm{KL}[q_\phi(\mathbf{x}, \mathbf{z}) || q(\mathbf{x})q_\phi(\mathbf{z})] = \mathbb{I}_q[\mathbf{x}, \mathbf{z}].
\end{aligned} \tag{16}$$

We show that maximizing the lower bound on the MI with respect to $\{\phi, \psi, \xi\}$ is equivalent to maximizing the KL divergence between the joint distribution and their marginals. Together with Equations (10) and (16) as well as Proposition 1, the optimization objective becomes:

$$\begin{aligned}
\mathcal{L}_{\text{MI-ELBO}} = &\mathbb{E}_{q(\mathbf{x})}\{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z}) \\
&- \mathrm{KL}[q_\phi(\mathbf{z}|\mathbf{x}) || p(\mathbf{z})]\} + \alpha C(\phi, \psi^*),
\end{aligned} \tag{17}$$

where $C(\phi, \psi^*)$ is defined as a function that maximizes the lower bound on the MI with the optimal ψ^*. Then, we compute the gradients of Equation (17) with respect to $\{\phi, \theta\}$. While it is easy to compute the gradient with respect to θ, the gradient with respect to ϕ is hard to compute since $C(\phi, \psi^*)$ itself depends on ϕ. Actually, when the function $f_\psi(\mathbf{x}, \mathbf{z})$ is optimal, the expectation of the gradients becomes zero, that is

$$\mathbb{E}_{q_\phi(\mathbf{x},\mathbf{z})}[\nabla_\phi f^*_\psi(\mathbf{x}, \mathbf{z})] - \mathbb{E}_{q(\mathbf{x})q_\phi(\mathbf{z})}\Big[\nabla_\phi e^{f^*_\psi(\mathbf{x},\mathbf{z})}\Big] = 0. \tag{18}$$

In practice, we can ignore the gradients once $f_\psi(\mathbf{x}, \mathbf{z})$ achieves the optimality. We refer the proposed model as *VAE-MINE*. The overall learning procedure is summarized in Algorithm 1.

5 Experiments and Results

5.1 Datasets

We adopt three benchmark datasets, Penn Treebank (Marcus et al., 1993), Stanford Natural Language Inference (Bowman et al., 2015) and Yahoo Answers (Yang et al., 2017) to evaluate whether the inferred latent variable can give better performance in language modeling and text generation. For data pre-processing, we set the maximum

Algorithm 1: VAE with Mutual Information Neural Estimation (VAE-MINE)

```
Initialize the parameters of the encoder, decoder, energy
  function $\{\phi, \theta, \psi\}$ and tangent scope $\xi$;
repeat
    Sample mini-batch of $M$ sentences $\{\mathbf{x}_1, \ldots, \mathbf{x}_M\}$
      from the data distribution $q(\mathbf{x})$;
    Sample mini-batch of $M$ latent variables
      $\{\mathbf{z}_1, \ldots, \mathbf{z}_M\}$ from the encoder $q_\phi(\mathbf{z}|\mathbf{x})$;
    Shuffle $M$ latent variables $\{\mathbf{z}'_1, \ldots, \mathbf{z}'_M\}$;
    Update $\psi$ by ascending its stochastic gradient:
    $\frac{1}{M}\sum_{i=1}^{M}\nabla_\psi\Big[f_\psi(\mathbf{x}_i, \mathbf{z}_i) - \xi \cdot e^{f_\psi(\mathbf{x}_i, \mathbf{z}'_i)}\Big]$;
    Compute the optimal tangent scope $\xi^*$:
    $\xi^* = \frac{1}{M}\sum_{i=1}^{M} e^{f_\psi(\mathbf{x}_i, \mathbf{z}'_i)}$;
    Update $\phi$ by ascending its stochastic gradient:
    $\frac{1}{M}\sum_{i=1}^{M}\nabla_\phi[\log p_\theta(\mathbf{x}_i|\mathbf{z}_i) - \mathrm{KL}[q_\phi(\mathbf{z}_i|\mathbf{x}_i)||p(\mathbf{z})]]$
    $+\frac{1}{M}\sum_{i=1}^{M}\nabla_\phi\Big[f_\psi(\mathbf{x}_i, \mathbf{z}_i) - \xi \cdot e^{f_\psi(\mathbf{x}_i, \mathbf{z}'_i)}\Big]$;
    Update $\theta$ by ascending its stochastic gradient:
    $\frac{1}{M}\sum_{i=1}^{M}\nabla_\theta \log p_\theta(\mathbf{x}_i|\mathbf{z}_i)$;
    Perform SGD-updates for parameters $\{\phi, \theta, \psi\}$;
until convergence;
```

Dataset	#Train	#Valid	#Test	Length	Vocab
PTB	42K	3.3K	3.7K	21	10K
SNLI	1.1M	13K	13K	20	20K
Yahoo	101K	10K	10K	78	20K

Table 1: Statistics of three benchmark datasets used in the experiments. Length denotes the average length of the sentences in the dataset and Vocab denotes the vocabulary size.

length of sentences to 200 and the maximum vocabulary size to 20K across all the datasets. Statistics of three datasets are summarized in Table 1.

5.2 Experimental Setup

We implement the VAE using a one-layer unidirectional LSTM (Hochreiter and Schmidhuber, 1997) with 512 hidden units and 128-dimensional word embedding for the encoder and decoder. The latent variable is 64-dimensional for all the models. We adopt the standard Gaussian prior and the diagonal Gaussian posterior. The last hidden state of the encoder RNN is fed into an MLP to estimate the mean and variance of the Gaussian posterior. Then, we sample the latent variable **z** and feed it to the decoder RNN at each step. The initial hidden state of the decoder RNN is obtained by feeding **z** to another MLP with the $\tanh()$ activation function. In order to further regularize the decoder RNN, dropout with a keep probability of 0.5 is applied to the inputs and outputs. We adopt SGD to optimize the model with a decayed learning rate and a gradient clipping.

Our architecture adopts an encoder RNN and an energy function parameterized by neural networks to maximize the lower bound on the MI. The energy function is trained by samples from the variational joint distribution and marginals. In practice, samples from the marginals are obtained by pairing one sentence **x** and another latent variable $\mathbf{z}'$ inferred by another sentence $\mathbf{x}'$. We adopt an MLP with two 512-unit hidden layers and ReLU activation to parameterize the energy function. As the encoder RNN and energy function optimize the same MI objective, the lower layers are shared between two modules. We set the value of the hyperparameter α within $[2, 5]$ with step $= 0.5$ for VAE-MINE and select the best α^* based on the validation dataset.

We compare the proposed VAE-MINE with the state-of-the-art baselines. LSTM-LM is an unconditional language model. For VAE-0.5 (Bowman et al., 2016), we implement KL annealing by increasing the KL weight linearly from 0.1 to 1.0 in the first 10 epochs and adopt word dropout rate of 0.5 to alleviate the posterior collapse. For VAE-BOW (Zhao et al., 2017), we replace word dropout with the bag-of-word (BOW) loss. For InfoVAE (Zhao et al., 2019), CyclicalVAE (Fu et al., 2019), and LaggingVAE (He et al., 2019), we follow the reported implementations.

5.3 Performance Evaluation

Posterior Collapse. We apply three metrics to evaluate the severity of the posterior collapse: KL divergence between the variational posterior and the prior, mutual information (MI), and the number of active units (AU) of the latent variable **z**. If the KL in Equation (3) becomes zero, the posterior collapses, meaning that the variational posterior is equal to the prior. For the MI, we use the Monte Carlo approach to approximate Equation (6),

$$\mathrm{KL}[q_\phi(\mathbf{z})||p(\mathbf{z})] = \frac{1}{S}\sum_{s=1}^{S}[\log q_\phi(\mathbf{z}_s) - \log p(\mathbf{z}_s)], \tag{19}$$

where S is the sample size.

The number of active units (Burda et al., 2015) is computed by

$$\mathrm{AU} = \sum_{d=1}^{D}\mathbb{I}\{\mathrm{Cov}_{q(\mathbf{x})}(\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[z_d]) \geq \epsilon\}, \tag{20}$$

where z_d is the d-th dimension of **z** and the threshold ϵ is set to 0.01. $\mathbb{I}\{\cdot\}$ is an indicator giving 1 when its statement is true and 0 otherwise.

Forward and Reverse Perplexity. A common measure for the quality of the generated sentences is to evaluate the perplexity of a language model trained on the training or generated sentences. We fit a non-parametric Kneiser-Ney (KN) smoothed 5-gram language model (LM) (Heafield, 2011) on the sentences sampled from the model distribution $p_\theta(\mathbf{x})$ and the true data distribution $q(\mathbf{x})$. Two evaluation criteria are defined as *Forward Perplexity* (FPPL) and *Reverse Perplexity* (RPPL) (Zhao et al., 2018a; Cífka et al., 2018). Computing the FPPL is equivalent to training a KN 5-gram LM on the training sentences and reporting the perplexity on the 100K sentences generated from the model, while the RPPL involves training a KN 5-gram LM on the 100K generated sentences and reporting the perplexity on the test sentences.

5.4 Language Modeling Results

The results for language modeling are shown in Table 2. We report negative log-likelihood (NLL), KL divergence (KL), perplexity (PPL), mutual information (MI), the number of active units (AU), forward perplexity (FPPL) and reverse perplexity (RPPL).

According to Table 2, we observe that the KL value for VAE-0.5 is almost zero, indicating that the model suffers from posterior collapse. Similar situations can also be observed for InfoVAE. When the bag-of-word loss is used, the KL value increases, yet making VAE-BOW produce a larger NLL. This suggests that the latent variable does not contain useful information to reduce the reconstruction error. For VAE-MINE, we see that it can produce either comparable or better results over the evaluation metrics when compared with other state-of-the-art baselines. With the mutual information term considered for the optimization, we believe that VAE-MINE allows more reasonable correspondence patterns between the input and its inferred latent variable so as to better alleviate posterior collapse. Together with the original ELBO objective, the inferred posteriors of the different input sentences give an appropriate level of overlapping among them in the latent space, instead of sticking to the Gaussian prior. To summarize, the latent variable with non-trivial KL value helps the decoder RNN for better word-level generation.

Regarding the generation capability, we observe that VAE-MINE and LaggingVAE exhibit similar performance in terms of PPL and FPPL (sample quality) and outperform the other baselines by a significant margin. Particularly, VAE-MINE performs significantly better than others in terms of RPPL (sample diversity). We believe that this is due to the explicit use of the MI term to guide the learning process. It is consistent with our motivation that the MI-regularized objective can alleviate the posterior collapse to improve VAEs' capability of generating fluent and diverse sentences with the effective use of the latent variable. In contrast, LaggingVAE treats the MI as a stopping criterion for training the encoder RNN at the beginning of optimization. Even though the model can achieve good performance on alleviating the posterior collapse and generate fluent sentences, the generated sentences lack diversity. VAE-MINE can achieve relatively lower RPPL values, which suggests that it can achieve a good balance between sample diversity and training quality. This is not surprising that maximizing the MI would produce more diverse sentences. Also, the encoder parameters can benefit from the gradients from the energy function during training.

5.5 Reconstruction, Interpolation and Generation

We qualitatively evaluate the sentences generated using VAE-MINE as well as other baselines. The VAE variants can reconstruct input sentences by encoding them into the latent space and then decoding the means or samples drawn from the variational posteriors. Table 3 shows two sets of sentences generated using VAE-MINE with a greedy decoding based on two particular inputs. It shows that sentences involving similar underlying structures and concepts can be generated.

In order to evaluate the quality of the inferred latent space, we interpolate two points in the latent space and then decode the interpolated points for generating sentences. Table 4 shows the transitions of the generated sentences between two particular sentences in the SNLI dataset using different models. LaggingVAE turns out to be performing the worst by generating repetitive sentences. This explains why it gives large RPPL values as shown in Table 2. For VAE-MINE, we find that it can generate smoothly transitioned sentences in term of sentence semantics and syntactics via linear interpolation in the latent space. This demonstrates that the proposed VAE-MINE can learn the latent representations which smoothly fill up the

Dataset	Model	NLL (KL)	PPL	MI	AU	FPPL	RPPL
PTB	LSTM-LM	102.85 (–)	95.63	–	–	461.44	472.86
	VAE-0.5	104.06 (0.03)	100.90	0.07	2	339.32	444.81
	VAE-BOW	105.58 (4.35)	107.94	1.65	4	385.77	463.34
	InfoVAE	108.37 (0.02)	122.15	0.06	0	483.71	515.09
	CyclicalVAE	104.36 (4.93)	102.25	2.84	10	419.61	464.78
	LaggingVAE	100.93 (4.84)	87.82	2.84	**13**	**12.64**	842.02
	VAE-MINE	**100.91** (4.86)	**87.75**	**3.01**	**13**	13.34	**400.21**
SNLI	LSTM-LM	33.41 (–)	18.97	–	–	49.99	289.30
	VAE-0.5	33.55 (0.01)	19.20	0.01	0	45.04	281.15
	VAE-BOW	37.00 (13.57)	26.06	2.07	23	44.72	278.31
	InfoVAE	33.88 (0.02)	19.78	0.09	0	45.34	295.38
	CyclicalVAE	34.84 (6.39)	21.52	3.12	28	58.16	291.53
	LaggingVAE	**33.10** (1.34)	**18.03**	0.73	3	**4.02**	173.47
	VAE-MINE	33.13 (4.05)	18.09	**3.15**	**30**	6.21	**152.63**
Yahoo	LSTM-LM	348.95 (–)	78.99	–	–	461.44	472.85
	VAE-0.5	340.64 (0.63)	71.17	0.44	3	178.97	252.51
	VAE-BOW	343.39 (10.13)	73.66	1.99	7	182.95	256.85
	InfoVAE	344.84 (0.03)	74.21	0.01	1	161.93	260.77
	CyclicalVAE	335.40 (7.51)	75.56	3.29	12	193.79	256.28
	LaggingVAE	**328.69** (6.65)	**61.29**	2.99	17	**5.60**	843.56
	VAE-MINE	328.70 (6.70)	61.31	**4.34**	**18**	6.72	**185.33**

Table 2: Language modeling results on the PTB, SNLI and Yahoo datasets. We report negative log-likelihood (NLL), KL divergence (KL), perplexity (PPL), mutual information (MI), the number of active units (AU) of the latent variable **z**, forward perplexity (FPPL) and reverse perplexity (RPPL). We express the NLL as the ELBO. **Bold** numbers indicate the best performance and underlined numbers indicate the second best performance.

Input:	**the man is at home sleeping.**	**three girls are sitting at desks and appear to be working intently.**
Output1:	the man is at home sleeping.	three women are sitting at desks and looking to be working together.
Output2:	the woman is sleeping at home.	Two ladies are standing at desks and appear to be playing together.
Output3:	the woman is working at home.	three women are sitting at desks and looking to be playing together.

Table 3: Sentences generated from the variational posteriors of the latent variable based on SNLI dataset.

latent space.

Next, we compare generated sentences by sampling latent variables from the Gaussian prior and then greedy decoding. In Table 5, we observe that sentences generated from LaggingVAE are not as diverse as the ones generated from VAE-MINE.

5.6 Nearest Neighbour Analysis

To better understand the learned representations, we did a nearest-neighbour analysis by choosing a sentence from the training set, ordering the training set in terms of the cosine distance computed in the latent space, then selecting the top three with the highest similarity. Our results in Table 6 show that VAE-MINE learns a good latent variable such that novel sentences can be chosen, due to the nature of MI-regularized objective.

6 Conclusion

Variational autoencoders suffer from the posterior collapse issue. In this paper, we propose to alleviate this problem by adding a mutual information regularization into the objective of the VAE. Explicitly optimizing the proposed objective allows to encode high-level abstractions effectively. We use an energy function parameterzied by neural networks and the convex duality approach to maximize the lower bound on the MI. Experimental results on three benchmark datasets show that the proposed model VAE-MINE in general outperforms other state-of-the-art baselines and can generate more reasonable sentences via linear interpolation in the latent space. One research direction is to explore how to generate long-form text with deep generative models organized in multiple layers of latent variables, due to natural language characterizing a hierarchical structure.

Acknowledgments

The authors would like to thank the anonymous reviewers for their insightful comments.

References

Alexander A. Alemi, Ben Poole, Ian Fischer, Joshua V. Dillon, Rif A. Saurous, and Kevin Murphy. 2018. Fixing a broken ELBO. In *Proceedings of the 35th International Conference on Machine Learning*, pages 159–168.

InfoVAE	two men are sitting on a couch while they wait for their parents to work. the woman is the world war. a man telling his hand. this man is sitting on bike. man jet ski is catching colorful waves. two little girl are happily not about to play twister.
CyclicalVAE	the men are ready to fight in the streets in front of the crowd with a sheet. a woman is standing with the canister in a band. a female age matured on the stage is looking in a microscope. a man is eating a small meal at a nightclub. the performer is taking his photos by a statue. there is a crowd of people, and several women posing for a picture.
LaggingVAE	this church choir sings to the masses as they sing joyous songs from the book at a church. this church choir sings to the masses as they sing joyous songs from the book at a church. this church choir sings to the masses as they sing joyous songs from the book at a church. a woman gets picture taken in front of the masses. a woman gets picture taken in front of the masses. a woman gets picture taken in front of the masses.
VAE-MINE	a choir including three people sing and dance on the stage in front of the masses. two musicians are playing the drums and a girl sits on a piano. a young family sits while waiting a girl at the bottom of a church. a man takes a photo of the girl. a little girl has a toy and a digital camera. the woman with red shirt is smiling to the girl.

Table 4: Sentences generated by interpolating between the encodings of "**this church choir sings to the masses as they sing joyous songs from the book at a church**." and "**a woman with a green headscarf, blue shirt and a very big grin**.".

the company said it will be able to sell its n stake in the u.s. the company said it will be to pay $ n million in the debt the company said it will be able to replace its existing UNK and UNK the company said it would buy n shares of its common shares for $ n each
the stock market's plunge was a UNK shot of the company's stock market yesterday mr. UNK said the agreement has n't yet to be determined by the company but the japanese are struggling to the UNK of the u.s. government the company also said it will sell its boston corp. unit to the company's UNK group
how do i get rid of UNK UNK? i have a UNK UNK and i have a UNK UNK. i have tried UNK and UNK. how do you say "UNK"? i'm not sure what you mean. i'm not sure what you mean. how do i get a UNK? i am a UNK student in college and i want to know how to get a job in the UNK area. how do i get a job in a bank? i am a student. i have a UNK what is the best way to study in a UNK.
what is the difference between a democrat and republican? i think it is a UNK, but i do n't think UNK is a UNK. what is the best city in the world? i'm looking for a good place to start a UNK, but i'm not sure if this is the best way. how do i get a free music download? i want to download it from my computer and i want to know the name of the song. who is the best player in the nba? i think UNK is a great player.

Table 5: Qualitative comparisons on the generated sentences. First row: PTB samples generated from the Gaussian prior by LaggingVAE (upper half) and VAE-MINE (lower half). Second row: Yahoo samples generated from the Gaussian prior by LaggingVAE (upper half) and VAE-MINE (lower half).

Query	**but the japanese are struggling to the UNK of the u.s. government**
LaggingVAE	the japanese UNK openly about the u.s. public's UNK many UNK regard a u.s. presence as a desirable UNK to japanese influence the u.s. government in recent years has accused japanese companies of UNK slashing prices
VAE-MINE	the u.s. government in recent years has accused japanese companies of UNK slashing prices while the small deals are far less UNK they add to japanese UNK of the u.s. market continuing demand for dollars from japanese investors boosted the u.s. currency
Query	**a person on a horse jumps over a broken down airplane.**
LaggingVAE	a horse jockey is jumping a horse over an obstacle on a coarse. a person aquestrian horse jumping over a wooden fence. a jockey riding a horse prepares to jump over an obstacle.
VAE-MINE	a horse jockey is jumping a horse over an obstacle on a coarse. a jockey riding a horse prepares to jump over an obstacle. a small child plays with her airplane as a cat looks on.

Table 6: Sentences selected from the PTB and SNLI training set in terms of the cosine distance between the inferred latent variables.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R. Devon Hjelm. 2018. Mutual information neural estimation. In *Proceedings of the 35th International Conference on Machine Learning*, pages 530–539.

Yoshua Bengio. 2009. Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, 2(1):1–127.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642.

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew M. Dai, Rafal Jozefowicz, and Samy Bengio. 2016. Generating sentences from a continuous space. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 10–21.

Philemon Brakel and Yoshua Bengio. 2017. Maximizing independence with GANs for non-linear ICA. In *ICML Workshop on Implicit Models*.

Yuri Burda, Roger B. Grosse, and Ruslan Salakhutdinov. 2015. Importance weighted autoencoders. In *3rd International Conference on Learning Representations*.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. 2016. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2172–2180.

Xi Chen, Diederik P. Kingma, Tim Salimans, Yan Duan, Prafulla Dhariwal, John Schulman, Ilya Sutskever, and Pieter Abbeel. 2017. Variational lossy autoencoder. In *5th International Conference on Learning Representations*.

Ondřej Cífka, Aliaksei Severyn, Enrique Alfonseca, and Katja Filippova. 2018. Eval all, trust a few, do wrong to none: Comparing sentence generation models. *Computing Research Repository*, arXiv:1804.07972.

Adji B. Dieng, Yoon Kim, Alexander M. Rush, and David M. Blei. 2019. Avoiding latent variable collapse with generative skip models. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, pages 2397–2405.

Hao Fu, Chunyuan Li, Xiaodong Liu, Jianfeng Gao, Asli Çelikyilmaz, and Lawrence Carin. 2019. Cyclical annealing schedule: A simple approach to mitigating KL vanishing. In *Proceedings of the 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680.

Anirudh Goyal, Alessandro Sordoni, Marc-Alexandre Côté, Nan Rosemary Ke, and Yoshua Bengio. 2017. Z-forcing: Training stochastic recurrent networks. In *Advances in Neural Information Processing Systems*, pages 6716–6726.

Junxian He, Daniel Spokoyny, Graham Neubig, and Taylor Berg-Kirkpatrick. 2019. Lagging inference networks and posterior collapse in variational autoencoders. In *7th International Conference on Learning Representations*.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the 6th Workshop on Statistical Machine Translation*, pages 187–197.

R. Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. 2019. Learning deep representations by mutual information estimation and maximization. In *7th International Conference on Learning Representations*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Matthew D. Hoffman and Matthew J. Johnson. 2016. ELBO surgery: yet another way to carve up the variational evidence lower bound. In *NIPS Workshop in Advances in Approximate Bayesian Inference*.

Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. 1999. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233.

Yoon Kim, Sam Wiseman, Andrew C. Miller, David Sontag, and Alexander M. Rush. 2018. Semi-amortized variational autoencoders. In *Proceedings of the 35th International Conference on Machine Learning*, pages 2683–2692.

Diederik P. Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. 2016. Improving variational inference with inverse autoregressive flow. In *Advances in Neural Information Processing Systems*, pages 4743–4751.

Diederik P. Kingma and Max Welling. 2014. Auto-encoding variational Bayes. In *2nd International Conference on Learning Representations*.

Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, and Ian Goodfellow. 2016. Adversarial autoencoders. In *4th International Conference on Learning Representations*.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Gábor Melis, Chris Dyer, and Phil Blunsom. 2018. On the state of the art of evaluation in neural language models. In *6th International Conference on Learning Representations*.

Aäron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. 2016. Pixel recurrent neural networks. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 1747–1756.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *Computing Research Repository*, arXiv:1807.03748.

Tom Pelsmaeker and Wilker Aziz. 2019. Effective estimation of deep generative language models. *Computing Research Repository*, arXiv:1904.08194.

Ali Razavi, Aäron van den Oord, Ben Poole, and Oriol Vinyals. 2019. Preventing posterior collapse with δ-VAEs. In *7th International Conference on Learning Representations*.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. 2014. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31st International Conference on Machine Learning*, pages 1278–1286.

Stanislau Semeniuta, Aliaksei Severyn, and Erhardt Barth. 2017. A hybrid convolutional variational autoencoder for text generation. In *Empirical Methods in Natural Language Processing*, pages 627–637.

Trieu H. Trinh, Andrew M. Dai, Minh-Thang Luong, and Quoc V. Le. 2018. Learning longer-term dependencies in RNNs with auxiliary losses. In *Proceedings of the 35th International Conference on Machine Learning*, pages 4972–4981.

Petar Veličković, William Fedus, William L. Hamilton, Pietro Liò, Yoshua Bengio, and R. Devon Hjelm. 2019. Deep graph Infomax. In *7th International Conference on Learning Representations*.

Prince Zizhuang Wang and William Yang Wang. 2019. Riemannian normalizing flow on variational wasserstein autoencoder for text modeling. In *Proceedings of the 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

Jiacheng Xu and Greg Durrett. 2018. Spherical latent spaces for stable variational autoencoders. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4503–4513.

Zichao Yang, Zhiting Hu, Ruslan Salakhutdinov, and Taylor Berg-Kirkpatrick. 2017. Improved variational autoencoders for text modeling using dilated convolutions. In *Proceedings of the 34th International Conference on Machine Learning*, pages 3881–3890.

Junbo Zhao, Yoon Kim, Kelly Zhang, Alexander M. Rush, and Yann LeCun. 2018a. Adversarially regularized autoencoders. In *Proceedings of the 35th International Conference on Machine Learning*, pages 5897–5906.

Shengjia Zhao, Jiaming Song, and Stefano Ermon. 2019. InfoVAE: Balancing learning and inference in variational autoencoders. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*.

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

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 654–664.