COUPLE-VAE: MITIGATING THE ENCODER-DECODER INCOMPATIBILITY IN VARIATIONAL TEXT MODELING WITH COUPLED DETERMINISTIC NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

The variational autoencoder (VAE) combines latent variable models and amortized variational inference. Despite its theoretical attractiveness, the optimization of VAE for text modeling suffers from the posterior collapse problem, where the decoder ignores the latent codes, and the posterior becomes nearly identical to the prior. We demonstrate that the VAE training dynamics face the challenge of *encoder-decoder incompatibility*, in which the encoder receives scarce backpropagated gradients from the decoder, and little encoded information is passed to the decoder. We propose a model-agnostic approach, named Couple-VAE, to mitigate this issue. Specifically, we couple the VAE model with a deterministic network with the same structure, which is optimized with the reconstruction loss without any regularization (e.g., the KL divergence). To enrich the backpropagated gradients for the encoder, we share the encoder between the deterministic network and the stochastic network. To encourage nontrivial decoding signals, we propose a coupling loss that pushes the stochastic decoding signals to the deterministic ones. We conduct extensive experiments on the Penn Treebank, Yelp, and Yahoo. We apply the proposed method to various variational text modeling models with different regularization terms, posterior families, decoder architectures, and optimization strategies and observe consistently improved text modeling results in terms of probability estimation and the richness of the encoded text.[1]

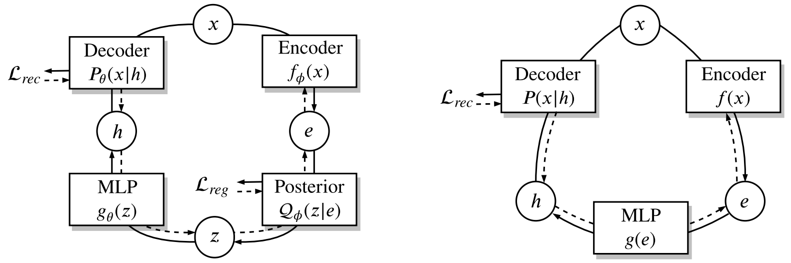
1 INTRODUCTION

The variational autoencoder (VAE) (Kingma & Welling, 2014) is a generative model that combines neural latent variables and amortized variational inference, which is capable of both estimating and sampling from the data distribution. It infers a posterior distribution for each instance with a shared inference network and optimizes the evidence lower bound (ELBO) instead of the intractable marginal log-likelihood. Given its potential to learn representations from massive text data, there has been much interest in using VAE for text modeling (Xu & Durrett, 2018; He et al., 2019).

Prior work has observed that the optimization of VAE suffers from the *posterior collapse* problem, i.e., the posterior becomes nearly identical to the prior and the decoder degenerate into a standard language model (Bowman et al., 2016; Zhao et al., 2017). A widely mentioned explanation is that a strong decoder makes the collapsed posterior a good local optimum of ELBO, and existing solutions include weakened decoders (Yang et al., 2017; Semeniuta et al., 2017), modified regularization terms (Higgins et al., 2017; Wang & Wang, 2019), alternative posterior families (Rezende & Mohamed, 2015; Davidson et al., 2018), richer prior distributions (Tomczak & Welling, 2018), improved optimization strategies (He et al., 2019), and narrowed amortization gaps (Kim et al., 2018).

In this paper, we provide a new perspective on the posterior collapse problem. In Section 3, by tracking the gradient norms of multiple model components w.r.t. the encoded text, we demonstrate the *encoder-decoder incompatibility* issue in both the forward pass and the backpropagation of VAE training, which leads to a poorly expressive encoder and an over-expressive decoder. To mitigate this issue, we propose a model-agnostic approach in Section 4, named Couple-VAE. We couple the

[1] We will make our code public.



(a) Variational autoencoder (VAE) (b) Deterministic autoencoder (DAE)

Figure 1: General frameworks of VAE and DAE for text modeling.

VAE model with a deterministic network with the same structure, which is optimized purely with the reconstruction loss without any regularization. To propagate richer gradients into the encoder, we share the encoder between the coupled networks. To learn nontrivial decoding signals, we propose a coupling loss that pushes the stochastic decoding signals to the deterministic decoding signals. Since our approach does not make any assumption on the regularization term, the posterior family, the decoder architecture, or the optimization strategy, we apply our method to various VAE-based models for text modeling. Experimental results on the Penn Treebank, Yelp, and Yahoo show that the proposed method consistently improves the performance of various VAE-based models in terms of probability estimation and the richness of the encoded information.

2 Background

2.1 Variational Inference for Text Modeling

VAE adopts a two-step view of the generative process of text data. It first samples the latent code $\boldsymbol{z}$ from the prior distribution $\mathcal{P}(\boldsymbol{z})$ and then samples the text x from the generator $P(x|\boldsymbol{z};\theta)$. By marginalizing out the latent variable, the likelihood of each datapoint is derived as

$$P(x;\theta) = \mathbb{E}_{\boldsymbol{z}\sim\mathcal{P}(\boldsymbol{z})}[P(x|\boldsymbol{z};\theta)] \tag{1}$$

As maximizing the log-likelihood with exact marginalization is usually intractable, VAE uses a variational family of posterior distributions $\mathcal{Q}(\boldsymbol{z}|x;\phi)$ and derives the evidence lower bound (ELBO)

$$\log P(x;\theta) \geq \mathbb{E}_{\boldsymbol{z}\sim\mathcal{Q}(\boldsymbol{z}|x;\phi)}[\log P(x|\boldsymbol{z};\theta)] - \mathrm{KL}[\mathcal{Q}(\boldsymbol{z}|x;\phi) \parallel \mathcal{P}(\boldsymbol{z})] \tag{2}$$

For training, as shown in Figure 1(a), the encoded text is transformed into posterior via a posterior network. A low-dimensional latent code is sampled from the posterior and then transformed into the decoding signal $\boldsymbol{h}$ with an MLP. Finally, the decoder infers the input with the decoding signal. The VAE objective can be generally interpreted as a reconstruction loss $\mathcal{L}_{rec}$ plus a regularization loss $\mathcal{L}_{reg}$, whose concrete forms can be modified, i.e.,

$$\mathcal{L} = \mathcal{L}_{rec} + \mathcal{L}_{reg} \tag{3}$$

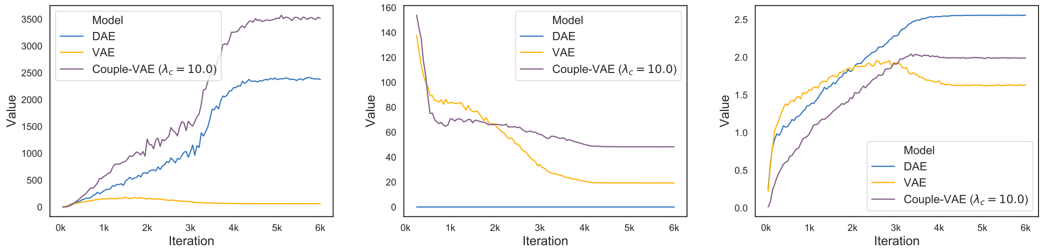
However, the optimization of the VAE objective is challenging. We usually observe a very small $\mathcal{L}_{reg}$ and a $\mathcal{L}_{rec}$ similar to a standard language model, i.e., the well-known *posterior collapse* problem.

2.2 Deterministic Autoencoders

Our new perspective on the posterior collapse problem is based on the comparison between VAE and the deterministic autoencoder (DAE, or simply, autoencoders) (Rumelhart et al., 1986; Ballard, 1987). Figure 1(b) shows a graphical overview of DAE for text modeling, which is composed of a text encoder, an optional MLP, and a text decoder. The objective for DAE is the reconstruction loss, which is empirically usually much lower than that of VAE after convergence.

3 Encoder-Decoder Incompatibility in VAE for Text Modeling

To understand the posterior collapse problem, we take a deeper look into the training dynamics of VAE. In this part, we investigate the following questions. How much backpropagated gradient does



(a) Gradient norm of the reconstruction loss w.r.t. the encoded text

(b) Gradient norm of the regularization loss w.r.t. the encoded text

(c) Gradient norm of the decoding signal w.r.t. the encoded text

Figure 2: Training dynamics of DAE, VAE, and the proposed Couple-VAE ($\lambda_c = 10.0$) on the Yelp test set. Please find the analysis in Section 3 and Section 5.5. Best viewed in color.

the encoder receive from reconstruction? How much does it receive from regularization? How much information does the decoder receive from the encoded text?

3.1 Tracking Training Dynamics

To answer the first question, we study the gradient norm of the reconstruction loss w.r.t. the encoded text, i.e., $\|\partial \mathcal{L}_{rec} / \partial \boldsymbol{e}\|_2$, which according to the chain rule is the amount of backpropagated gradient received by the encoder parameters. From Figure 2(a), we observe that for DAE it constantly increases, while for VAE it increases marginally in the early stage and then decreases continuously. It shows that the reconstruction loss actively optimizes the DAE encoder throughout the training phase, while the VAE encoder lacks backpropagated gradients after the early stage of training.

We seek the answer to the second question by studying the gradient norm of the regularization loss w.r.t. the encoded text, i.e., $\|\partial \mathcal{L}_{reg} / \partial \boldsymbol{e}\|_2$. The parameters of the posterior network consist of both *weights* and *bias*, and a larger gradient norm shows that the model relies more on the weights than the bias to optimize $\mathcal{L}_{reg}$. Figure 2(b) shows a constant decrease of the gradient norm in VAE from the 2.5K step until convergence, which shows that the posterior collapse is aggravated as the KL weight increases (KL annealing is applied from step 1K to 41K).

For the third question, we compute the normalized gradient norm of the decoding signal w.r.t. the encoded text, i.e., $\|\partial \boldsymbol{h} / \partial \boldsymbol{e}\|_F \,/\, \|\boldsymbol{h}\|_2$. As this term shows how relatively the decoding signal changes with the perturbation of the encoded text, it reflects the amount of information passed from the encoder to the decoder. Figure 2(c) shows that for DAE it constantly increases. For VAE, it at first increases even faster than DAE, then slows down, and finally decreases until convergence, indicating that the VAE decoder to some extent ignores the encoder in the late stage of training.

3.2 Encoder-Decoder Incompatibility

The above observations indicate an *incompatibility* between the encoder and the decoder in VAE training. We argue that this incompatibility is caused by the regularization on the posterior, which does not allow much encoded information to pass through and thus forces the decoder to exploit its expressive power. Thus, the decoder is usually over-expressive after convergence. The failure in using the encoded information, in turn, results in the scare backpropagated gradients from the reconstruction, resulting in a poorly expressive encoder. The poorly expressive encoder then encourages the posterior network to rely more on the bias rather than the weights, as discussed in Section 3.1.

4 Coupling Variational Text Modeling with Deterministic Networks

To mitigate the encoder-decoder incompatibility issue, we propose to couple the VAE model with a deterministic network, displayed in Figure 3. All modules in the deterministic network (upper half) share the *structure* with those in the stochastic (variational) network (lower half). Specifically, the coupled posterior network $\mathrm{Posterior}^c$ transforms the encoded text into the coupled latent code $\boldsymbol{z}^c$,

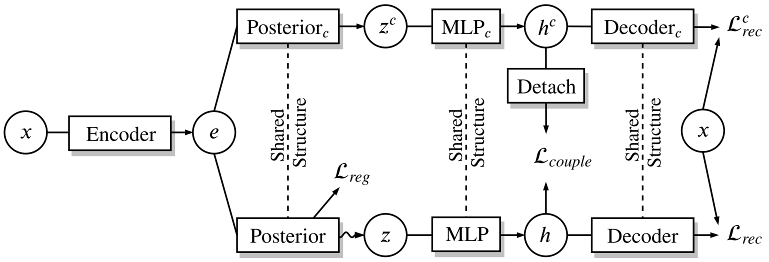


Figure 3: A graphical overview of Couple-VAE. The upper path is deterministic and optimized by the coupled reconstruction loss $\mathcal{L}^c_{rec}$, and the lower path is the VAE model optimized by the reconstruction loss $\mathcal{L}_{rec}$, the regularization loss $\mathcal{L}_{reg}$, and the coupling loss $\mathcal{L}_{couple}$.

which is mapped to the coupled decoding signal $\boldsymbol{h}^c$ by MLP^c. The coupled decoder $\mathrm{Decoder}^c$ then infers the text with $\boldsymbol{h}^c$. Whenever sampling is applied in the stochastic network, we use the predicted mean vector for the deterministic network, e.g., for the Gaussian posterior, we simply discard the predicted standard deviation vector and use the predicted mean vector for later computation. Please find details for other posterior families in Appendix B. Similar to DAE, the coupled deterministic network is optimized solely by the coupled reconstruction loss $\mathcal{L}^c_{rec}$, which is the same autoregressive cross-entropy loss as $\mathcal{L}_{rec}$. We share the encoder between the stochastic network and the deterministic network, which enriches the reconstruction gradients backpropagated to the encoder (compared with VAE) by leveraging the coupled reconstruction loss. To encourage nontrivial decoding signals, we propose a coupling loss $\mathcal{L}_{couple}$ that pushes the stochastic decoding signals to the deterministic ones. Formally, the objective of our approach is given as

$$\mathcal{L} = \mathcal{L}_{rec} + \mathcal{L}_{reg} + \mathcal{L}^c_{rec} + \lambda_c \mathcal{L}_{couple} \tag{4}$$

where λ_c is a hyperparameter, $\mathcal{L}^c_{rec}$ is the coupled reconstruction loss, and the coupling loss $\mathcal{L}_{couple}$ is essentially a distance metric between $\boldsymbol{h}$ and $\boldsymbol{h}^c$. Since the decoding signals are usually not distributed in a Euclidean space, we adopt the Rational Quadratic kernel at multiple scales, i.e.,

$$\mathcal{L}_{couple} = -\sum_s (1 + \frac{\|\boldsymbol{h} - \mathrm{Detach}(\boldsymbol{h}^c)\|^2}{s \cdot C})^{-1} \tag{5}$$

where C is a hyper-parameter, $s \in \{0.1, 0.2, 0.5, 1, 2, 5, 10\}$, and $\mathrm{Detach}(\boldsymbol{h}^c)$ prevents gradients to be propagated into $\boldsymbol{h}^c$ since we would like $\boldsymbol{h}^c$ to guide $\boldsymbol{h}$ but not the opposite. In Section 5.5, we show how Couple-VAE has improved training dynamics compared with VAE.

One would resort to the universal approximation theorem (Hornik et al., 1989) and question the necessity to share the module structures between the deterministic network and the stochastic network. Indeed, implementing the deterministic network as an MLP is theoretically adequate. However, we argue that every structure has its favored geometry of the learned manifold, on which the decoding signals are distributed. For example, the latent space learned by planar normalizing flows (Rezende & Mohamed, 2015), which has compression and expansion, and vMF-VAE (Xu & Durrett, 2018; Davidson et al., 2018), which is supported on a $(d-1)$-dimensional sphere in $\mathbb{R}^d$, may significantly influence the learned manifold of decoding signals.

5 Experiments

5.1 Experimental Setup

We conduct the experiments on three commonly used datasets for text modeling, i.e., the Penn Treebank (PTB) (Marcus et al., 1993), Yelp (Xu et al., 2016), and Yahoo. The training/validation/test splits are 42K/3370/3761 for PTB, 63K/7773/8671 for Yelp, and 100K/10K/10K for Yahoo. The vocabulary size for PTB/Yelp/Yahoo is 10K/15K/20K. We discard the sentiment labels in Yelp.

We evaluate the proposed method by applying it to previous variational text modeling models, including VAE (Kingma & Welling, 2014), β-VAE (Higgins et al., 2017), vMF-VAE (Xu & Durrett, 2018; Davidson et al., 2018) with learnable κ, CNN-VAE (Yang et al., 2017), WAE (Tolstikhin et al.,

Table 1: Language modeling results. NLL is estimated with importance sampling. PPL is based on the estimated NLL. KL and MI are approximated by their Monte Carlo estimates. *Couple-* stands for "with the coupled deterministic network". The better results between a model and the version with the coupled deterministic network are shown in **bold**. *The exact NLL is reported. †Using or modifying open-source code which does not follow our setup and evaluation. ‡Previously reported.

	PTB		Yelp		Yahoo	
	NLL (KL)	PPL	NLL (KL)	PPL	NLL (KL)	PPL
GRU-LM*	105.8 (-)	125.3	196.3 (-)	57.3	347.9 (-)	78.0
VAE	103.6 (8.6)	112.9	193.7 (7.2)	54.3	344.5 (12.4)	74.7
Couple-VAE	**103.1 (9.5)**	**110.5**	**191.2 (8.0)**	**51.6**	**342.4 (12.8)**	**72.8**
$\beta(0.8)$-VAE	103.8 (11.0)	113.9	193.8 (10.2)	54.5	344.9 (16.1)	75.1
Couple-$\beta(0.8)$-VAE	**103.3 (12.1)**	**111.5**	**191.5 (12.2)**	**51.9**	**342.8 (17.0)**	**73.2**
$\beta(1.2)$-VAE	103.7 (7.8)	113.3	193.7 (6.0)	54.3	345.3 (10.5)	75.5
Couple-$\beta(1.2)$-VAE	**102.9 (8.6)**	**109.6**	**191.2 (6.9)**	**51.6**	**342.3 (11.3)**	**72.7**
vMF-VAE	103.6 (2.0)	113.2	195.4 (0.0)	56.3	344.5 (2.5)	74.7
Couple-vMF-VAE	**103.0 (3.0)**	**110.1**	**191.2 (2.8)**	**51.6**	**342.2 (4.0)**	**72.5**
CNN-VAE	118.5 (29.6)	222.6	194.2 (12.8)	54.8	344.3 (19.7)	74.5
Couple-CNN-VAE	**118.2 (30.2)**	**219.7**	**193.9 (13.7)**	**54.6**	**343.3 (22.4)**	**73.6**
WAE	103.7 (11.0)	113.3	193.7 (10.7)	54.3	344.7 (16.6)	74.9
Couple-WAE	**103.2 (12.5)**	**110.9**	**191.3 (12.5)**	**51.7**	**343.3 (18.2)**	**73.6**
VAE-NF	103.3 (5.5)	111.3	193.9 (5.3)	54.5	344.3 (8.1)	74.5
Couple-VAE-NF	**102.6 (5.7)**	**108.1**	**191.8 (5.6)**	**52.2**	**342.6 (8.8)**	**73.0**
WAE-NF	103.4 (6.7)	111.9	194.1 (7.0)	54.7	344.3 (10.6)	74.5
Couple-WAE-NF	**102.7 (7.4)**	**108.4**	**192.1 (7.4)**	**52.5**	**342.7 (11.0)**	**73.1**
SA-VAE†	100.7 (7.7)	98.7	183.5 (3.8)	44.0	327.5 (7.2)‡	60.4‡
Lagging-VAE†	98.8 (6.0)	90.7	182.5 (1.2)	43.1	326.7 (6.0)	59.7
Couple-Lagging-VAE†	**98.7 (11.0)**	**90.4**	**182.3 (3.8)**	**42.9**	**326.2 (7.4)**	**59.3**

2018), VAE with normalizing flows (VAE-NF) (Rezende & Mohamed, 2015), WAE with normalizing flows (WAE-NF), and Lagging-VAE (He et al., 2019). We also show the result of GRU-LM (Cho et al., 2014) and SA-VAE (Kim et al., 2018). We do not apply our method to SA-VAE since it does not follow the amortized variational inference framework. We use the setups and open-source implementations of SA-VAE and Lagging-VAE. Please find more details in Appendix C.[2]

5.2 Language Modeling Results

We report negative log-likelihood (NLL), KL divergence, and perplexity as the metrics for language modeling. NLL is estimated with importance sampling, KL is approximated by its Monte Carlo estimate, and perplexity is computed based on NLL. Please find the metric details in Appendix D.

Table 1 shows the language modeling results. We find that for all models, our proposed approach achieves smaller negative log-likelihood and lower perplexity, which shows the effectiveness of our approach to improve the probability estimation capability of various VAE-based models. Larger KL divergence is also observed, showing that our approach helps address the posterior collapse problem.

5.3 Mutual Information and Reconstruction

Language modeling results only evaluate the probability estimation ability of VAE. We are also interested in how rich the latent space is. We report the mutual information (MI) between the text x and the latent code $\boldsymbol{z}$ under $\mathcal{Q}(\boldsymbol{z}|x)$, which is approximated with Monte Carlo estimation. Better

[2] We will make the code for each baseline and its coupled version public.

Table 2: Mutual information (MI) and reconstruction metrics (i.e., BLEU-1 and BLEU-2). MI is approximated by its Monte Carlo estimate. Other notations follow Table 1.

	PTB		Yelp		Yahoo	
	MI	BLEU-1/2	MI	BLEU-1/2	MI	BLEU-1/2
VAE	10.48	23.2 / 4.4	8.28	28.7 / 5.3	15.43	21.2 / 3.6
Couple-VAE	**11.99**	**23.4 / 4.5**	**9.65**	**30.4 / 5.8**	**16.44**	**23.1 / 4.1**
$\beta(0.8)$-VAE	15.43	**24.5 / 4.9**	13.52	30.6 / 6.0	24.16	24.0 / 4.3
Couple-$\beta(0.8)$-VAE	**18.13**	24.3 / 4.8	**17.69**	**32.6 / 6.6**	**28.03**	**26.4 / 4.9**
$\beta(1.2)$-VAE	9.16	22.8 / **4.3**	6.60	28.0 / 5.0	11.83	18.2 / 2.9
Couple-$\beta(1.2)$-VAE	**10.28**	**22.9** / 4.2	**7.90**	**29.8 / 5.6**	**13.51**	**22.4 / 3.8**
vMF-VAE	1.74	15.2 / 2.0	0.03	22.4 / 2.8	2.06	8.5 / 1.1
Couple-vMF-VAE	**2.37**	**16.1 / 2.3**	**2.60**	**25.1 / 4.0**	**3.37**	**10.3 / 1.4**
CNN-VAE	78.49	**32.0 / 7.8**	17.26	32.9 / 7.1	30.18	24.9 / 5.3
Couple-CNN-VAE	**80.54**	31.8 / 7.7	**19.15**	**33.4 / 7.3**	**37.62**	**26.9 / 5.9**
WAE	15.09	**24.8 / 5.1**	15.08	30.7 / 6.1	24.73	24.2 / 4.5
Couple-WAE	**18.51**	24.7 / **5.1**	**18.56**	**32.5 / 6.6**	**30.08**	**27.7 / 5.3**
VAE-NF	5.63	19.2 / **3.3**	5.64	25.6 / 4.5	8.02	13.7 / 2.1
Couple-VAE-NF	**5.86**	**19.4 / 3.3**	**6.06**	**26.3 / 4.6**	**9.14**	**15.3 / 2.5**
WAE-NF	7.18	19.7 / 3.5	7.95	26.0 / 4.6	11.43	13.8 / 2.2
Couple-WAE-NF	**8.10**	**20.7 / 3.7**	**8.53**	**27.2 / 5.0**	**12.56**	**14.9 / 2.5**
Lagging-VAE†	2.90	-	0.96	-	3.04	-
Couple-Lagging-VAE†	**3.29**	-	**2.36**	-	**3.06**	-

Table 3: The effect of the coupling weight λ_c. *Reported in the Table 1 and 2.

	PTB				Yelp			
	NLL (KL)	PPL	MI	BLEU-1/2	NLL (KL)	PPL	MI	BLEU-1/2
VAE	103.6 (8.6)	112.9	10.48	23.2 / 4.4	193.7 (7.2)	54.3	8.28	28.7 / 5.3
$\lambda_c = 0.1$*	103.1 (9.5)	110.5	11.99	23.4 / 4.5	191.2 (8.0)	51.6	9.65	30.4 / 5.8
$\lambda_c = 1.0$	103.3 (10.7)	111.4	14.32	24.0 / 4.8	191.1 (8.1)	51.5	9.92	30.5 / 5.8
$\lambda_c = 2.0$	103.2 (11.8)	111.1	16.58	24.2 / 5.0	191.7 (10.5)	52.1	14.13	31.9 / 6.2
$\lambda_c = 5.0$	103.7 (16.1)	113.2	32.78	26.5 / 5.8	191.5 (12.8)	51.9	19.77	32.8 / 6.5
$\lambda_c = 10.0$	104.7 (21.8)	118.5	44.93	29.0 / 7.0	191.8 (17.3)	52.2	27.08	34.7 / 7.2

reconstruction from the encoded text is another way to show the richness of the latent space. For each text x, we sample ten latent codes from $\mathcal{Q}(\boldsymbol{z}|x)$ and decode them with greedy search. We report the BLEU-1 and BLEU-2 scores between the reconstruction and the input. Please find the metric details in Appendix E. In Table 2, we observe that our approach improves MI on all datasets, showing that our approach helps learn a richer latent space. BLEU-1 and BLEU-2 are consistently improved on Yelp and Yahoo, but not on PTB. Given that text samples in PTB are significantly shorter than those in Yelp and Yahoo, we conjecture that it is easier for the decoder to reconstruct on PTB by exploiting its autoregressive expressiveness, even without less rich latent codes.

5.4 Analysis of the Coupling Weight

We investigate the model performance with different coupling weights, shown in Table 3. With larger coupling weight, the model achieves higher KL divergence, MI, and reconstruction metrics. It shows that by pushing the stochastic decoding signals closer to the deterministic decoding signals, we get more complex posterior distribution and latent codes that contains richer text information. Note that the best NLL does not guarantee the highest other metrics, which justifies the necessity to evaluate VAE models with multiple metrics.

Table 4: Gradient norms of the reconstruction loss, the coupled reconstruction loss, the regularization loss, and the decoding signal w.r.t. the encoded text on each test set.

		$\left\Vert\frac{\partial \mathcal{L}_{rec}}{\partial \boldsymbol{e}}\right\Vert_2$	$\left\Vert\frac{\partial \mathcal{L}^c_{rec}}{\partial \boldsymbol{e}}\right\Vert_2$	$\left\Vert\frac{\partial (\mathcal{L}_{rec}+\mathcal{L}^c_{rec})}{\partial \boldsymbol{e}}\right\Vert_2$	$\left\Vert\frac{\partial \mathcal{L}_{reg}}{\partial \boldsymbol{e}}\right\Vert_2$	$\frac{\Vert\frac{\partial \boldsymbol{h}}{\partial \boldsymbol{e}}\Vert_F}{\Vert\boldsymbol{h}\Vert_2}$
PTB	DAE	1719.8	-	-	-	3.14
	VAE	112.5	-	-	19.4	2.05
	Couple-VAE ($\lambda_c = 0.1$)	**148.5**	2109.6	2320.2	**27.7**	**2.12**
Yelp	DAE	2443.6	-	-	-	2.55
	VAE	59.7	-	-	18.8	1.62
	Couple-VAE ($\lambda_c = 0.1$)	**84.8**	3640.8	3764.7	**25.0**	**2.25**
Yahoo	DAE	4104.6	-	-	-	3.39
	VAE	257.9	-	-	52.8	2.92
	Couple-VAE ($\lambda_c = 0.1$)	**335.3**	5105.0	5615.0	**65.0**	**3.91**

5.5 Analysis of Gradient Norms

We study the three gradient norms defined in Section 3. Table 4 displays the gradient norms the models reported in the main experiments. Notably, $\|\partial\mathcal{L}^c_{rec}/\partial\boldsymbol{e}\|_2$ in Couple-VAE is even larger than $\|\partial\mathcal{L}_{rec}/\partial\boldsymbol{e}\|_2$ in DAE. It has two indications. First, the encoder indeed encodes rich information of the text. Second, compared with DAE, Couple-VAE better generalizes to the test sets, which we conjecture is due to the regularization on the posterior. Couple-VAE also has a larger $\|\partial\mathcal{L}_{reg}/\partial\boldsymbol{e}\|_2$ compared with VAE, which based on the argument in Section 3 indicates that Couple-VAE relies on the weights (which learn a better aggregated posterior) than the bias (which leads to a collapsed posterior) of the posterior network. We also observe larger $\|\partial\boldsymbol{h}/\partial\boldsymbol{e}\|_F \,/\, \|\boldsymbol{h}\|_2$ in Couple-VAE, which indicates that the decoder in Couple-VAE uses more encoded information than in VAE.

To show how Couple-VAE ameliorates the training dynamics, we also track the gradient norms of Couple-VAE (with $\lambda_c = 10.0$ for a clearer comparison), plotted along with VAE and DAE in Figure 2. The curve for Couple-VAE in Figure 2(a) stands for $\|\partial(\mathcal{L}_{rec}+\mathcal{L}^c_{rec})/\partial\boldsymbol{e}\|_2$. Please find the plots for more datasets in Appendix F. We observe that Couple-VAE receives constantly increasing backpropagated gradients from the reconstruction. In contrast to VAE, the $\|\partial\mathcal{L}_{reg}/\partial\boldsymbol{e}\|_2$ in Couple-VAE does not decrease significantly as the KL weight increases. The decrease of $\|\partial\boldsymbol{h}/\partial\boldsymbol{e}\|_F \,/\, \|\boldsymbol{h}\|_2$, which VAE suffers from, is not observed in Couple-VAE.

5.6 Diversity and Samples from the Prior Distribution

Given the generative nature of VAE, we evaluate the diversity and the quality of samples from the prior distribution. For diversity, we sample 3200 texts from the prior and report the Dist-1 and Dist-2 metrics (Li et al., 2016), which are the ratios of distinct unigrams and bigrams over all generated unigrams and bigrams. For quality, we provide the first three texts sampled from each model. Table 5 displays the diversity metrics and the sampled texts on Yelp, and results on other datasets are in Appendix G. Dist-1 and Dist-2 show that texts sampled from Couple-VAE are more diverse than those from VAE. Qualitatively, we observe that the long texts generated from VAE have more redundancies compared with Couple-VAE. Given that both models have the same latent dimension, the indication is that Couple-VAE is using the latent codes more efficiently. Given limited space, we put the interpolation for qualitative analysis in Appendix H.

6 Relation to Related Work

Bowman et al. (2016) first identify the posterior collapse problem of VAE for text modeling and propose KL annealing and word drop to alleviate the problem. Zhao et al. (2017) propose an auxiliary bag-of-words (BoW) loss to mitigate this issue. Later work that works on the posterior collapse problem mainly focuses on using less powerful decoders (Yang et al., 2017; Semeniuta et al.,

Table 5: Diversity and the first three samples from each model on Yelp. Dist-1 and Dist-2 stand for the ratios of distinct unigrams and bigrams over all generated ones. Redundancies are shown in red.

VAE	Dist-1 = 0.0062	Dist-2 = 0.0248
1. the food is good , but the food is good . i had the chicken fried steak with a side of mashed potatoes , and it was a good choice . the fries were good , but the fries were good . i had the chicken breast with a side 2. ok , so i was excited to check out this place for a while . i was in the area , and i was n't sure what to expect . i was a little disappointed with the food , but i was n't sure what to expect . i was 3. we went to the biltmore fashion park . we were seated right away , but we were seated right away . we were seated right away , but we were seated right away . we were seated right away and we were seated right away . the staff was very		
Couple-VAE ($\lambda_c = 10.0$)	Dist-1 = **0.0115**	Dist-2 = **0.0593**
1. i 'm a fan of the " asian " restaurants in the valley , and i 'm not sure what to expect , but i 'm not sure what the fuss is about . the meat is fresh and delicious . i 'm not a fan of the " skinny 2. i 'm not a fan of the fox restaurants in phoenix , but i have to say that the service is always a great experience . the atmosphere is a little dated and there is a great view of the mountains . 3. i have been here twice , and the food was good , but the service was good , but the food was good . i had a great time , but the service was great . the food was a bit pricey , but the service was a bit slow		

2017), modifying the regularization objective (Higgins et al., 2017; Bahuleyan et al., 2019; Wang & Wang, 2019), seeking alternative posterior families (Rezende & Mohamed, 2015; Xu & Durrett, 2018; Davidson et al., 2018; Xiao et al., 2018), finding richer prior distributions (Tomczak & Welling, 2018), improving optimization strategies (He et al., 2019; Fu et al., 2019), incorporating skip connections into the posterior network (Dieng et al., 2019), adopting hierarchical or autoregressive posterior distributions (Park et al., 2018; Du et al., 2018), and narrowing the amortization gap (Hjelm et al., 2016; Kim et al., 2018; Marino et al., 2018).

A model to be noted is β-VAE (Higgins et al., 2017), in which the improvement of reconstruction and the sacrifice of regularization are modeled as a trade-off via a hyperparameter. Different from their work, our approach can be viewed as multi-task learning, where the *coupled reconstruction* task helps learn richer encoded representations and the *distance minimization* task helps learn nontrivial decoding signals. Since the auxiliary tasks are locally applied to the encoded texts and the decoding signals, they do not necessarily require a sacrifice of other components of the overall objective.

He et al. (2019) demonstrate the incompatibility between the variational posterior and the true model posterior, i.e., the variational posterior sometimes lags behind the true model posterior. We investigate in another direction and provide a perspective that focuses on the encoder-decoder incompatibility, which comes from the stochasticity and over-regularization. In our experiments, we show that mitigating the encoder-decoder incompatibility can further improve the results in He et al. (2019).

Ghosh et al. (2019) propose to remove the stochasticity in VAE by directly injecting noises into a deterministic autoencoder, which empirically generates less "blurry" images. Different from their work, we still strictly follow the probability (density) estimation nature of VAE, and the deterministic network in our approach serves as an auxiliary to enrich the encoded representations and the decoding signals.

7 Conclusions

In this paper, we provide the encode-decoder incompatibility as a new perspective on the posterior collapse problem of VAE for text modeling. By tracking and comparing the gradient norms of multiple components in DAE and VAE, we demonstrate that the incompatibility exists in both the forward pass and the backpropagation during VAE training. We propose a model-agnostic approach termed Couple-VAE, which mitigates the encoder-decoder incompatibility by enriching the reconstruction gradients for the encoder and encouraging nontrivial decoding signals. Experimental results show the effectiveness of our approach to improve various VAE-based models for text modeling in terms of probability estimation and the richness of the learned latent space. The training dynamics demonstrate that our approach mitigates the encoder-decoder incompatibility compared with VAE.

REFERENCES

Hareesh Bahuleyan, Lili Mou, Hao Zhou, and Olga Vechtomova. Stochastic wasserstein autoencoder for probabilistic sentence generation. In *Proceedings of NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pp. 4068–4076, 2019.

Dana H. Ballard. Modular learning in neural networks. In *Proceedings of the 6th National Conference on Artificial Intelligence. Seattle, WA, USA, July 1987.*, pp. 279–284, 1987.

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew M. Dai, Rafal Józefowicz, and Samy Bengio. Generating sentences from a continuous space. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, CoNLL 2016, Berlin, Germany, August 11-12, 2016*, pp. 10–21, 2016.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pp. 1724–1734, 2014.

Tim R. Davidson, Luca Falorsi, Nicola De Cao, Thomas Kipf, and Jakub M. Tomczak. Hyperspherical variational auto-encoders. In *Proceedings of UAI 2018, Monterey, California, USA, August 6-10, 2018*, pp. 856–865, 2018.

Adji B. Dieng, Yoon Kim, Alexander M. Rush, and David M. Blei. Avoiding latent variable collapse with generative skip models. In *AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, volume 89 of *Proceedings of Machine Learning Research*, pp. 2397–2405, 2019.

Jiachen Du, Wenjie Li, Yulan He, Ruifeng Xu, Lidong Bing, and Xuan Wang. Variational autoregressive decoder for neural response generation. In *Proceedings of EMNLP 2018, Brussels, Belgium, October 31 - November 4, 2018*, pp. 3154–3163, 2018.

Hao Fu, Chunyuan Li, Xiaodong Liu, Jianfeng Gao, Asli Çelikyilmaz, and Lawrence Carin. Cyclical annealing schedule: A simple approach to mitigating KL vanishing. In *Proceedings of NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pp. 240–250, 2019.

Partha Ghosh, Mehdi S. M. Sajjadi, Antonio Vergari, Michael J. Black, and Bernhard Schölkopf. From variational to deterministic autoencoders. *CoRR*, abs/1903.12436, 2019.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander J. Smola. A kernel two-sample test. *J. Mach. Learn. Res.*, 13:723–773, 2012.

Junxian He, Daniel Spokoyny, Graham Neubig, and Taylor Berg-Kirkpatrick. Lagging inference networks and posterior collapse in variational autoencoders. In *ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019.

Irina Higgins, Loïc Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. β-VAE: Learning basic visual concepts with a constrained variational framework. In *ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017.

R. Devon Hjelm, Ruslan Salakhutdinov, Kyunghyun Cho, Nebojsa Jojic, Vince D. Calhoun, and Junyoung Chung. Iterative refinement of the approximate posterior for directed belief networks. In *NIPS 2016, December 5-10, 2016, Barcelona, Spain*, pp. 4691–4699, 2016.

Kurt Hornik, Maxwell B. Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989.

Yoon Kim, Sam Wiseman, Andrew C. Miller, David Sontag, and Alexander M. Rush. Semi-amortized variational autoencoders. In *ICML*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2683–2692, 2018.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In *Proceedings of NAACL-HLT 2016, San Diego California, USA, June 12-17, 2016*, pp. 110–119, 2016.

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

Joseph Marino, Yisong Yue, and Stephan Mandt. Iterative amortized inference. In *Proceedings of ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3400–3409, 2018.

Yookoon Park, Jaemin Cho, and Gunhee Kim. A hierarchical latent structure for variational conversation modeling. In *Proceedings of NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pp. 1792–1801, 2018.

Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *Proceedings of ICML 2015, Lille, France, 6-11 July 2015*, pp. 1530–1538, 2015.

DE Rumelhart, GE Hinton, and RJ Williams. Learning internal representations by error propagation. In *Parallel distributed processing: explorations in the microstructure of cognition, vol. 1*, pp. 318–362. MIT Press, 1986.

Stanislau Semeniuta, Aliaksei Severyn, and Erhardt Barth. A hybrid convolutional variational autoencoder for text generation. In *Proceedings of EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pp. 627–637, 2017.

Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1): 1929–1958, 2014.

Ilya O. Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schölkopf. Wasserstein auto-encoders. In *ICLR 2018*, 2018.

Jakub M. Tomczak and Max Welling. VAE with a vampprior. In *AISTATS 2018, 9-11 April 2018, Playa Blanca, Lanzarote, Canary Islands, Spain*, volume 84 of *Proceedings of Machine Learning Research*, pp. 1214–1223, 2018.

Prince Zizhuang Wang and William Yang Wang. Riemannian normalizing flow on variational wasserstein autoencoder for text modeling. In *Proceedings of NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pp. 284–294, 2019.

Yijun Xiao, Tiancheng Zhao, and William Yang Wang. Dirichlet variational autoencoder for text modeling. *CoRR*, abs/1811.00135, 2018.

Jiacheng Xu and Greg Durrett. Spherical latent spaces for stable variational autoencoders. In *Proceedings of EMNLP 2018, Brussels, Belgium, October 31 - November 4, 2018*, pp. 4503–4513, 2018.

Jiacheng Xu, Danlu Chen, Xipeng Qiu, and Xuanjing Huang. Cached long short-term memory neural networks for document-level sentiment classification. In *Proceedings of EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pp. 1660–1669, 2016.

Zichao Yang, Zhiting Hu, Ruslan Salakhutdinov, and Taylor Berg-Kirkpatrick. Improved variational autoencoders for text modeling using dilated convolutions. In *ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pp. 3881–3890, 2017.

Tiancheng Zhao, Ran Zhao, and Maxine Eskénazi. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pp. 654–664, 2017.

A Notations

We first introduce the notations used in the following parts. Calligraphic letters (e.g., $\mathcal{Q}_0$) denotes continuous distributions, and the corresponding lowercase letters (e.g., q_0) stands for probability density functions. Probability of the text is represented as P.

B Deterministic Networks for Different Posterior Families

In this part, we detail the forward computation of the deterministic networks for different posterior families, including multivariate Gaussian, Gaussian with normalizing flows, and von MisesFisher.

B.1 Multivariate Gaussian

For multivariate Gaussian, we compute the coupled latent code $\boldsymbol{z}^c$ as

$$\boldsymbol{z}^c = \mathbb{E}_{\boldsymbol{z}\sim\mathcal{Q}^c(\boldsymbol{z}|x)}[\boldsymbol{z}] \tag{6}$$

where $\mathcal{Q}^c(\boldsymbol{z}|x)$ is the posterior distribution learned by the coupled deterministic network. In effect, $\boldsymbol{z}$ is the mean vector predicted by the coupled posterior network $\mathrm{Posterior}^c$.

B.2 Gaussian with Normalizing Flows

We first review the background and notations of normalizing flows. An initial latent code is first sampled from an initial distribution, i.e., $\boldsymbol{z}_0 \sim \mathcal{Q}_0(\boldsymbol{z}_0|x)$. The normalizing flow is defined as a series of *reversible* transformations $f_1, \dots, f_K$, i.e.,

$$\boldsymbol{z}_k = f_k \circ \cdots \circ f_1(\boldsymbol{z}_0) \tag{7}$$

where $k = 1, \dots, K$. The evidence lower bound (ELBO) for normalizing flows is derived as

$$\begin{aligned}\log P(x) &\geq \mathbb{E}_{\boldsymbol{z}_K\sim\mathcal{Q}_K(\boldsymbol{z}_K|x)}[\log P(x|\boldsymbol{z}_K)] - \mathrm{KL}[\mathcal{Q}_K(\boldsymbol{z}_K|x) \parallel \mathcal{P}_K(\boldsymbol{z}_K)] \\ &= \mathbb{E}_{\boldsymbol{z}_0\sim\mathcal{Q}_0(\boldsymbol{z}_0|x)}[\log P(x|\boldsymbol{z}_K) - \log q_0(\boldsymbol{z}_0|x) + \sum_{k=1}^{K} \log|\det\frac{\partial f_k}{\partial \boldsymbol{z}_{k-1}}| + \log p_K(\boldsymbol{z}_K)]\end{aligned} \tag{8}$$

where $\mathcal{P}_K(\boldsymbol{z}_K)$ is the prior distribution of the transformed latent variable and the reversibility of the transformations guarantees non-zero determinants. Obviously, the optimization of the ELBO for normalizing flows requires sampling from the initial distribution; thus, we compute the coupled latent code $\boldsymbol{z}^c$ by transforming the predicted mean vector of the coupled initial distribution, i.e.,

$$\boldsymbol{z}^c = f_k^c \circ \cdots \circ f_1^c(\mathbb{E}_{\boldsymbol{z}_0\sim\mathcal{Q}_0^c(\boldsymbol{z}_0|x)}[\boldsymbol{z}_0]) \tag{9}$$

where $\mathcal{Q}_0^c(\boldsymbol{z}_0|x)$ is the coupled initial distribution and $f_1^c, \dots, f_K^c$ are the coupled transformations. Note that all modules in the deterministic network share the structure with those in the stochastic network. We do not use the posterior mean as the coupled latent code for two reasons. First, our interest is to acquire a deterministic representation that guides the stochastic network, but not necessarily the mean vector. Second, the computation of the posterior mean after the transformations is intractable.

B.3 von Mises-Fisher

The von Mises-Fisher distribution is supported on a $(d-1)$-dimensional sphere in $\mathbb{R}^d$ and parameterized by a direction parameter $\boldsymbol{\mu} \in \mathbb{R}^d$ ($\|\boldsymbol{\mu}\| = 1$) and a concentration parameter κ, both of which are mapped from the encoded text by the posterior network. The probability density function is

$$q(\boldsymbol{z}|\boldsymbol{\mu}, \kappa) = \frac{\kappa^{d/2-1} \cdot \exp(\kappa\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{z})}{(2\pi)^{d/2} I_{d/2-1}(\kappa)} \tag{10}$$

where I_v is the modified Bessel function of the first kind at order v. We use the direction parameter $\boldsymbol{\mu}$ as the coupled latent code $\boldsymbol{z}^c$. Note that we do not use the posterior mean as the coupled latent code for two reasons. First, similar to normalizing flows, our interest is a deterministic representation rather than the mean vector. Second, the posterior mean of von Mises-Fisher *never* lies on the support of the distribution, which is suboptimal to guide the stochastic network.

C Details of the Experimental Setup

The dimension of latent vectors is 32. The dimension of word embeddings is 200. The encoder and the decoder are one-layer GRUs with the hidden state size of 128 for PTB and 256 for Yelp and Yahoo. For optimization, we use Adam (Kingma & Ba, 2015) with a learning rate of 10^{-3} and $\beta_1 = 0.9$, $\beta_1 = 0.999$. The decoding signal is viewed as the first word embedding and also concatenated to the word embedding in each decoding step. After 30K steps, the learning rate is decayed by half each 2K steps. Dropout (Srivastava et al., 2014) rate is 0.2. KL-annealing (Bowman et al., 2016) is applied from step 2K to 42K (on Yelp, it is applied from step 1K to 41K for VAE, Couple-VAE, β-VAE, and Couple-β-VAE; otherwise, the KL divergence becomes very large in the early stage of training). For each 1K steps, we estimate the NLL for validation.

For normalizing flows, we use planar flows (Rezende & Mohamed, 2015) with three contiguous transformations. For WAE and WAE-NF, we use Maximum Mean Discrepancy (MMD) (Gretton et al., 2012) as the regularization term. Following Wang & Wang (2019), an additional KL regularization term with the weight $\beta = 0.8$ (also with KL-annealing) is added to WAE and WAE-NF since MMD does not guarantee the convergence of the KL divergence.

D Estimation of Language Modeling Metrics

For language modeling, we report negative log-likelihood (NLL), KL divergence, and perplexity. To get more reliable results, we make the estimation of each metric explicit. For each test sample x, NLL is estimated by importance sampling, and KL is approximated by its Monte Carlo estimate:

$$\mathrm{NLL}_x = -\log P(x) \approx -\log(\frac{1}{N}\sum_{i=1}^{N}\frac{p(\boldsymbol{z}^{(i)})P(x|\boldsymbol{z}^{(i)})}{q(\boldsymbol{z}^{(i)}|x)}) \tag{11}$$

$$\mathrm{KL}_x = \mathrm{KL}[\mathcal{Q}(\boldsymbol{z}|x) \parallel \mathcal{P}(\boldsymbol{z})] \approx \frac{1}{N}\sum_{i=1}^{N}\log\frac{q(\boldsymbol{z}^{(i)}|x)}{p(\boldsymbol{z}^{(i)})} \tag{12}$$

where $\boldsymbol{z}^{(i)} \sim \mathcal{Q}(\boldsymbol{z}|x)$ are sampled latent codes and all notations follow Eq. (2). We report the averaged NLL and KL on all test samples. Perplexity is computed based on the estimated NLL. For validation, the number of samples is $N = 10$; for evaluation, the number of samples is $N = 100$.

E Estimation of Mutual Information and Reconstruction

We report the mutual information (MI) between the text x and the latent code $\boldsymbol{z}$ under $\mathcal{Q}(\boldsymbol{z}|x)$ to investigate how much useful information is encoded. The MI component of each test sample x is approximated by Monte Carlo estimation:

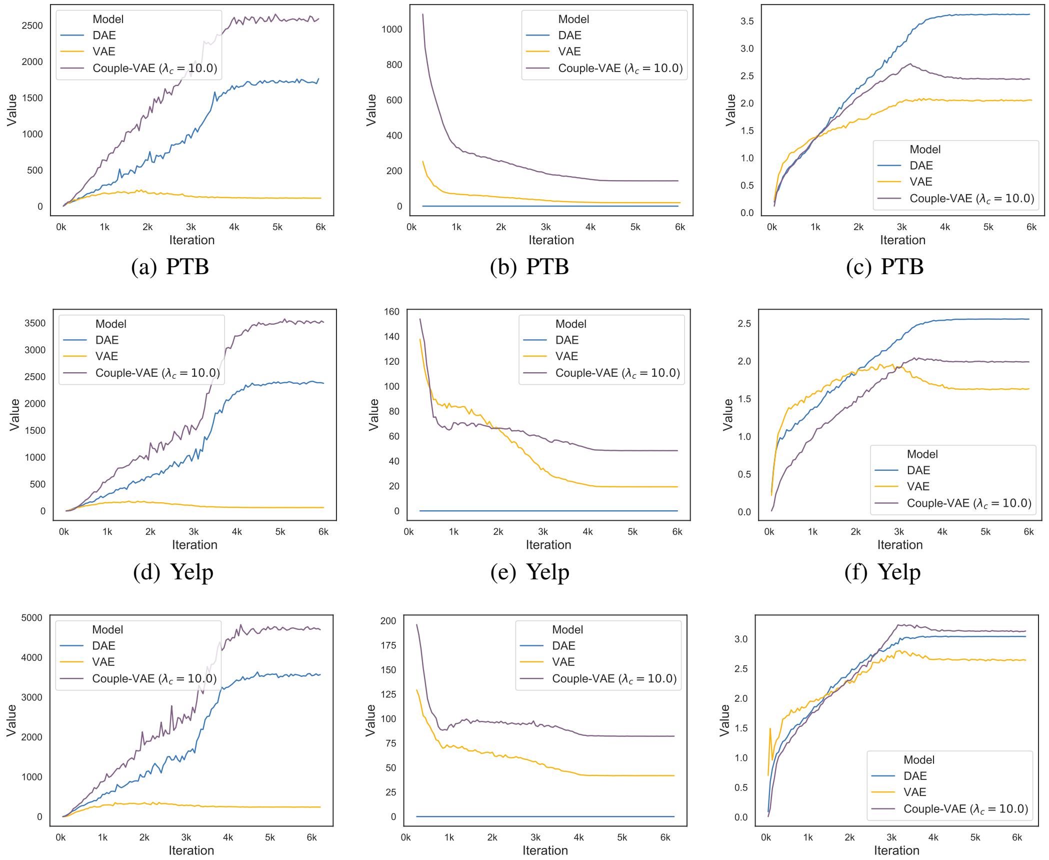
$$\mathrm{MI}_x = \mathbb{E}_{\boldsymbol{z}\sim\mathcal{Q}(\boldsymbol{z}|x)}[\log\frac{q(\boldsymbol{z}|x)}{q(\boldsymbol{z})}] \approx \frac{1}{N}\sum_{i=1}^{N}(\log q(\boldsymbol{z}^{(i)}|x) - \log q(\boldsymbol{z}^{(i)})) \tag{13}$$

where the aggregated posterior density $q(\boldsymbol{z}^{(i)})$ is approximated with its Monte Carlo estimate:

$$q(\boldsymbol{z}^{(i)}) = \mathbb{E}_x[q(\boldsymbol{z}^{(i)}|x)] \approx \frac{1}{M}\sum_{j=1}^{M}q(\boldsymbol{z}^{(i)}|x^{(j)}) \tag{14}$$

where $x^{(j)}$ are sampled from the test set. For convenience, most previous work uses the texts within each batch as the sampled $x^{(j)}$'s (which are supposed to be sampled from the entire test set). However, this convention results in a biased estimation since the $q(\boldsymbol{z}^{(i)}|x^{(i)})$ is computed when $j = i$, i.e., the text itself is always sampled when computing its MI component. We remedy it by skipping the term when $j = i$. The overall MI $= \mathbb{E}_x[\mathrm{MI}_x]$ is then estimated by averaging MI_x over all test samples. We set the numbers of samples as $N = 100$ and $M = 512$.

For reconstruction, we sample ten latent codes from the posterior of each text input and decode them with greedy search. We compute BLEU-1 and BLEU-2 between the reconstruction and the input with the Moses script.



(a) PTB (b) PTB (c) PTB

(d) Yelp (e) Yelp (f) Yelp

(g) Yahoo (h) Yahoo (i) Yahoo

Figure 4: Training dynamics of DAE, VAE, and Couple-VAE ($\lambda_c = 10.0$). (a), (d), and (g) are $\|\partial \mathcal{L}_{rec}/\partial \boldsymbol{e}\|_2$ for DAE and VAE, and $\|\partial(\mathcal{L}_{rec} + \mathcal{L}^c_{rec})/\partial \boldsymbol{e}\|_2$ for Couple-VAE. (b), (e), (h) denote $\|\partial \mathcal{L}_{reg}/\partial \boldsymbol{e}\|_2$. (c), (f), (i) stand for $\|\partial \boldsymbol{h}/\partial \boldsymbol{e}\|_F \,/\, \|\boldsymbol{h}\|_2$. Best viewed in color.

F Training Dynamics of Gradient Norms

We show the tracked gradient norms on all datasets in Figure 4. The observations are consistent with those discussed in Section 5.5.

G Diversity and Samples from the Prior Distribution

Given the limited space in the main text, we place the comprehensive evaluation of samples from the prior distribution in this part. Table 6 shows the diversity metrics and the first three samples from each model on all datasets. In line with the observations in Section 5.6, samples from Couple-VAE is more diverse than those from VAE. Moreover, more redundancies are observed in the VAE samples.

Table 6: Diversity metrics and the first three samples from each model on PTB. Redundancies (pieces of text that have appeared in the same text before) are shown in red.

VAE	Dist-1 = 0.0461	Dist-2 = 0.1636
1. but the market is a bit of the market ’s recent slide and the fed is trying to sell investors to buy back and forth between the s&p N and N 2. the company said it will be developed by a joint venture with the u.s. 3. the new york stock exchange composite index rose N to N		
Couple-VAE (λ_c = 10.0)	Dist-1 = **0.0551**	Dist-2 = **0.2446**
1. dd acquisition said it will offer to acquire N shares of lin ’s shares to be sold 2. but the u.s. would be closed at N p.m. edt in N but that was caused by lower rates 3. $ N billion in the stock market was a lot of it to be worth for each of N		
VAE (Yelp)	Dist-1 = 0.0062	Dist-2 = 0.0248
1. the food is good , but the food is good . i had the chicken fried steak with a side of mashed potatoes , and it was a good choice . the fries were good , but the fries were good . i had the chicken breast with a side 2. ok , so i was excited to check out this place for a while . i was in the area , and i was n’t sure what to expect . i was a little disappointed with the food , but i was n’t sure what to expect . i was 3. we went to the biltmore fashion park . we were seated right away , but we were seated right away . we were seated right away , but we were seated right away . we were seated right away and we were seated right away . the staff was very		
Couple-VAE ($\lambda_c = 10.0$) (Yelp)	Dist-1 = **0.0115**	Dist-2 = **0.0593**
1. i ’m a fan of the “ asian ” restaurants in the valley , and i ’m not sure what to expect , but i ’m not sure what the fuss is about . the meat is fresh and delicious . i ’m not a fan of the “ skinny 2. i ’m not a fan of the fox restaurants in phoenix , but i have to say that the service is always a great experience . the atmosphere is a little dated and there is a great view of the mountains . 3. i have been here twice , and the food was good , but the service was good , but the food was good . i had a great time , but the service was great . the food was a bit pricey , but the service was a bit slow		
VAE (Yahoo)	Dist-1 = 0.0044	Dist-2 = 0.0211
1. what is the difference between the two and the _UNK ? i am not sure what you mean , but i ’m not sure what you mean . i ’m not sure what you mean , but i ’m not sure what you mean . the answer is : 1 . the first person is the first person to be the first person to be the first person to be the first person . 2 . the first person is the first person to be the first person to be the first person . the first thing is that the person who is the best person is to be a person , and the person who is the best person to be born . the person who is not the best person is to be a person , and the person who is not the best person to be born . 2. what do you think of the song “ _UNK ” ? i ’m not sure what you ’re talking about . i ’m not sure what you ’re talking about . i ’m not sure what you ’re talking about . i ’m not sure what you ’re talking about . i ’m not sure what you ’re talking about . i ’m not sure what you ’re talking about . i ’m not sure what you ’re talking about . 3. what is the name of the song ? i heard that the song was a song called “ _UNK ” . it was a song called “ _UNK ” . it was a song called “ _UNK ” . it was a song called “ _UNK ” . it was a song called “ _UNK ” . it was a song called “ _UNK ” . it was a song called “ _UNK ” , “ _UNK ”		
Couple-VAE ($\lambda_c = 10.0$) (Yahoo)	Dist-1 = **0.0075**	Dist-2 = **0.0397**
1. if you are looking for a good wrestler , what do you think about the future ? i am not sure what i mean . i have been watching the ufc for 3 months . i have been watching the ufc and i have to be able to see what happens . 2. is it true that the war is not a hoax ? it is a myth that the _UNK of the war is not a war , but it is not possible to be able to see the war . the _UNK is not a war , but it ’s not a crime . 3. how do i get a _UNK on ebay ? ebay is free and they are free !		

Table 7: Texts generated from the interpolations of two latent codes.

VAE	Couple-VAE ($\lambda_c = 10.0$)
Text A (PTB): now those routes are n't expected to begin until jan	
they are n't expected to be completed	both sides are expected to be delivered at their contract
the new york stock exchange is scheduled to resume today	both sides are expected to be delivered at least
the new york stock exchange is scheduled to resume	both sides have been able to produce up with the current level
it is n't clear that it will be sold through its own account	it also has been used for comment
it is n't a major source of credit	it also has been working for the first time
it also has a major chunk of its assets	it also has a new drug for two years
it also has a major pharmaceutical company	it also has a $ N million defense initiative
Text B (PTB): it also has a _unk_ facility in california	

H Interpolation

A property of VAE is to match the interpolation in the latent space with the smooth transition in the text space (Bowman et al., 2016). In Table 7, we show the interpolation of VAE and Couple-VAE (with the coupling weight $\lambda_c = 10.0$) on PTB. It shows that compared with VAE, Couple-VAE has smoother transitions of subjects (*both sides* → *it*) and verbs (*are expected* → *have been* → *has been* → *has*), indicating that the information about subjects and verbs is more smoothly encoded in the latent space of Couple-VAE.