On the Encoder-Decoder Incompatibility in Variational Text Modeling and Beyond

Chen Wu[1], Prince Zizhuang Wang[2], William Yang Wang[2]
[1]Department of Foreign Languages and Literatures, Tsinghua University
[2]Department of Computer Science, University of California, Santa Barbara
`wu-c16@mails.tsinghua.edu.cn,`
`zizhuang_wang@ucsb.edu, william@cs.ucsb.edu`

Abstract

Variational autoencoders (VAEs) combine latent variables with amortized variational inference, whose optimization usually converges into a trivial local optimum termed *posterior collapse*, especially in text modeling. By tracking the optimization dynamics, we observe the *encoder-decoder incompatibility* that leads to poor parameterizations of the data manifold. We argue that the trivial local optimum may be avoided by improving the encoder and decoder parameterizations since the posterior network is part of a transition map between them. To this end, we propose Coupled-VAE, which couples a VAE model with a deterministic autoencoder with the same structure and improves the encoder and decoder parameterizations via encoder weight sharing and decoder signal matching. We apply the proposed Coupled-VAE approach to various VAE models with different regularization, posterior family, decoder structure, and optimization strategy. Experiments on benchmark datasets (i.e., PTB, Yelp, and Yahoo) show consistently improved results in terms of probability estimation and richness of the latent space. We also generalize our method to conditional language modeling and propose Coupled-**C**VAE, which largely improves the diversity of dialogue generation on the Switchboard dataset.[1]

1 Introduction

The variational autoencoder (VAE) (Kingma and Welling, 2014) is a generative model that combines neural latent variables and amortized variational inference, which is efficient in 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 (Zhao et al., 2017; Xu and 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 and Wang, 2019), alternative posterior families (Rezende and Mohamed, 2015; Davidson et al., 2018), richer prior distributions (Tomczak and Welling, 2018), improved optimization strategies (He et al., 2019), and narrowed amortization gaps (Kim et al., 2018).

In this paper, we provide a novel perspective for the posterior collapse problem. By comparing the optimization dynamics of VAE with deterministic autoencoders (DAE), we observe the incompatibility between a poorly optimized encoder and a decoder with too strong expressiveness. From the perspective of differential geometry, we show that this issue indicates poor *chart maps* from the data manifold to the *parameterizations*, which makes it difficult to learn a *transition map* between them. Since the posterior network is a part of the transition map, we argue that the posterior collapse would be mitigated with better parameterizations.

To this end, we propose the Coupled-VAE approach, which couples the VAE model with a deterministic network with the same structure. For better encoder parameterization, we share the encoder weights between the coupled networks. For better decoder parameterization, we propose a signal matching loss that pushes the stochastic decod-

---

[1]Our code is publicly available at `https://github.com/ChenWu98/Coupled-VAE`.

ing signals to the deterministic ones. Notably, our approach is model-agnostic since it does not make any assumption on the regularization term, the posterior family, the decoder architecture, or the optimization strategy. Experiments on PTB, Yelp, and Yahoo show that our method consistently improves the performance of various VAE models in terms of probability estimation and the richness of the latent space. The generalization to conditional modeling, i.e., Coupled-**C**VAE, largely improves the diversity of dialogue generation on the Switchboard dataset. Our contributions are as follows:

- We observe the encoder-decoder incompatibility in VAE and connect it to the posterior collapse problem.

- We propose the Coupled-VAE, which helps the encoder and the decoder to learn better parameterizations of the data manifold with a coupled deterministic network, via encoder weight sharing and decoder signal matching.

- Experiments on PTB, Yelp, and Yahoo show that our approach improves the performance of various VAE models in terms of probability estimation and richness of the latent space. We also generalize Coupled-VAE to conditional modeling and propose Coupled-**C**VAE, which largely improves the diversity of dialogue generation on the Switchboard dataset.

2 Background

2.1 Variational Inference for Text Modeling

The generative process of VAE is first to sample a latent code z from the prior distribution $\mathcal{P}(z)$ and then to sample the data x from $P(x|z;\theta)$ (Kingma and Ba, 2015). Since the exact marginalization of the log-likelihood is intractable, a variational family of posterior distributions $\mathcal{Q}(z|x;\phi)$ is adopted to derive the evidence lower bound (ELBO), i.e.,

$$\log P(x;\theta) \geq \mathbb{E}_{z \sim \mathcal{Q}(z|x;\phi)}[\log P(x|z;\theta)] \\ - \mathrm{KL}[\mathcal{Q}(z|x;\phi) \parallel \mathcal{P}(z)] \quad (1)$$

For training, as shown in Figure 1(a), the encoded text e is transformed into its posterior via a posterior network. A latent code is sampled and mapped to the decoding signal h. Finally, the decoder infers the input with the decoding signal. The objective can be viewed as a reconstruction loss $\mathcal{L}_{rec}$ plus a regularization loss $\mathcal{L}_{reg}$ (whose form varies), i.e.,

$$\mathcal{L} = \mathcal{L}_{rec} + \mathcal{L}_{reg} \quad (2)$$
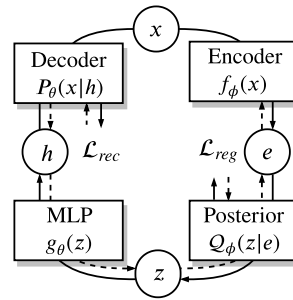
(a) Variational autoencoder
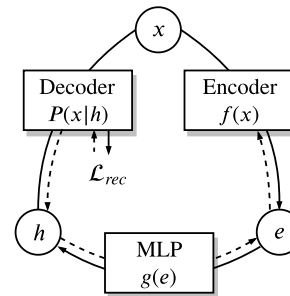
(b) Deterministic autoencoder

Figure 1: VAE and DAE for text modeling.

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

An older family of autoencoders is the deterministic autoencoder (DAE) (Rumelhart et al., 1986; Ballard, 1987). Figure 1(b) shows an overview of DAE for text modeling, which is composed of a text encoder, an optional MLP, and a text decoder. The reconstruction loss of DAE is 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. We investigate the following questions. How much backpropagated gradient does 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., $\|\partial \mathcal{L}_{rec}/\partial e\|_2$, which shows the magni-

(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

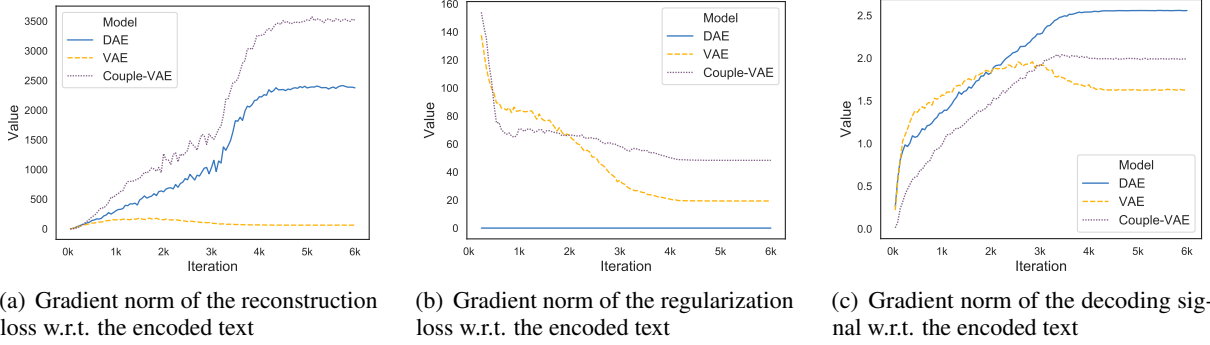(c) Gradient norm of the decoding signal w.r.t. the encoded text

Figure 2: Training dynamics of DAE, VAE, and the proposed Coupled-VAE on the Yelp test set. Please find the analysis in Section 3 and Section 5.7. Best viewed in color (yet the models are distinguished by line markers).

tude of gradients received by the encoder parameters. From Figure 2(a), we observe that it constantly increases in DAE, while in VAE it increases marginally in the early stage and then decreases continuously. It shows that the reconstruction loss actively optimizes the DAE encoder, 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., $\|\partial \mathcal{L}_{reg}/\partial e\|_2$. In a *totally* collapsed posterior, i.e., $\mathcal{Q}(z|x;\phi) = \mathcal{P}(z)$ for each x, $\|\partial \mathcal{L}_{reg}/\partial e\|_2$ would be zero. Thus, $\|\partial \mathcal{L}_{reg}/\partial e\|_2$ can show how far the posterior of each instance is from the aggregate posterior or the prior. 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.

For the third question, we compute the normalized gradient norm of the decoding signal w.r.t. the encoded text, i.e., $\|\partial h/\partial e\|_F / \|h\|_2$. 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, 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

Based on the training dynamics in Section 3.1 and the observations in previous work (Bowman et al., 2016; Zhao et al., 2017), text VAE has three features, listed as follows. First, the encoder is poorly

optimized, as shown by the low $\|\partial \mathcal{L}_{rec}/\partial e\|_2$. Second, the decoder degenerates into a powerful language model. Third, h contains less information from e in VAE than in DAE, which is indicated by the lower $\|\partial h/\partial e\|_F / \|h\|_2$. We call these features as *encoder-decoder incompatibility*.

To bridge the incompatibility and posterior collapse, we start with the *manifold hypothesis* which states that real-world data concentrates near a manifold with a lower dimensionality than the ambient space (Narayanan and Mitter, 2010; Bengio et al., 2013). In our case, we denote the manifold of text data as $\mathcal{X} \subset \bigcup_{l \in \mathbb{N}} \mathcal{V}^l$ where $\mathcal{V}$ is the vocabulary. In the language of differential geometry, the encoded text $e \in \mathcal{E} \subset \mathbb{R}^d$ and the decoding signal $h \in \mathcal{H} \subset \mathbb{R}^d$ can be viewed as the *parameterizations* (or *coordinates*) of $x \in \mathcal{X}$ under two different *charts* (or *coordinate systems*). Formally, we denote the chart maps as $\varphi_e : \mathcal{X} \to \mathcal{E}$ and $\varphi_h : \mathcal{X} \to \mathcal{H}$, which satisfy $e = \varphi_e(x)$ and $h = \varphi_h(x)$ for any $x \in \mathcal{X}$. Given the two charts, the map from $\mathcal{E}$ to $\mathcal{H}$ is called the *transition map* $\varphi_h \circ \varphi_e^{-1} : \mathcal{E} \to \mathcal{H}$ between the two charts.

In DAE, the two chart maps and the transition map between them are learned simultaneously via the single reconstruction loss, which we rewrite as

$$\mathcal{L}_{rec} = \mathbb{E}_{x \in \mathcal{X}}[\mathcal{L}(x, \varphi_h^{-1}(\varphi_h \circ \varphi_e^{-1}(\varphi_e(x))))] \quad (3)$$

where φ_e, $\varphi_h \circ \varphi_e^{-1}$, and φ_h^{-1} are modeled as the encoder, the MLP, and the decoder (strictly speaking, in text modeling, the range of φ_h^{-1} is not $\mathcal{X}$ but distributions on $\mathcal{X}$), as illustrated in Figure 3.

In VAE, as discussed before, both φ_e and φ_h inadequately parameterize the data manifold. We argue that the inadequate parameterizations make it harder to find a smooth transition map in VAE than in DAE, as shown by the lower $\|\partial h/\partial e\|_F / \|h\|_2$.
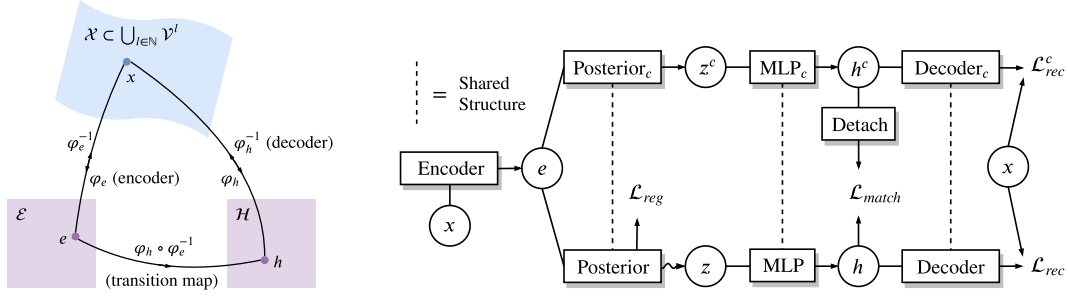
3451

Figure 3: Left: DAE and VAE interpreted as manifold parameterizations and a transition map. Right: A graphical overview of the proposed Coupled-VAE. The upper path is deterministic, and the lower path is stochastic.

Since the posterior network is a part of the transition map, it consequently seeks to map each instance to the prior (discussed in Section 3.1) rather than learning the transition map.

4 Coupling Variational and Deterministic Networks

Based on the above analysis, we argue that posterior collapse could be alleviated by learning chart maps (i.e., φ_e and φ_h) that better parameterize the data manifold. Inspired by the chart maps in DAE, we propose to couple the VAE model with a deterministic network, outlined in Figure 3. Modules with a subscript c are deterministic networks that share the *structure* with those in the stochastic network. Sampling is disabled in the deterministic network, e.g., in the case of Gaussian posterior, we 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 $\mathcal{L}_{rec}$.

To learn a well-optimized φ_e, we share the encoder between the stochastic and the deterministic networks, which leverages the rich gradients backpropagated from $\mathcal{L}^c_{rec}$. To learn better φ_h, we propose to guide φ_h with a well-learned chart map, i.e., the one characterized by Decoder^c. Thus, we introduce a signal matching loss $\mathcal{L}_{match}$ that pushes the $\boldsymbol{h}$ to $\boldsymbol{h}^c$. The objective of our approach is

$$\mathcal{L} = \mathcal{L}_{rec} + \mathcal{L}_{reg} + \lambda_r \mathcal{L}^c_{rec} + \lambda_m \mathcal{L}_{match} \quad (4)$$

where λ_r and λ_m are hyperparameters[2], $\mathcal{L}^c_{rec}$ is the coupled reconstruction loss, and the signal matching loss $\mathcal{L}_{match}$ is essentially a distance function

between $\boldsymbol{h}$ and $\boldsymbol{h}^c$. We evaluate both the Euclidean distance and the Rational Quadratic kernel[3], i.e.,

$$\mathcal{L}_{match} = \begin{cases} \|\boldsymbol{h} - \text{Detach}(\boldsymbol{h}^c)\|^2 & \text{Eucl} \\ \sum_s \frac{-s \cdot C}{s \cdot C + \|\boldsymbol{h} - \text{Detach}(\boldsymbol{h}^c)\|^2} & \text{RQ} \end{cases} \quad (5)$$

where $s \in \{0.1, 0.2, 0.5, 1, 2, 5, 10\}$, C is a hyperparameter, and Detach prevents gradients to be propagated into $\boldsymbol{h}^c$ since we would like $\boldsymbol{h}^c$ to guide $\boldsymbol{h}$ but not the opposite.

One would question the necessity of sharing the structure of the posterior network by resorting to universal approximation (Hornik et al., 1989). Specifically, a common question is: why not using an MLP as Posterior_c? We argue that each structure has a favored distribution of $\mathcal{H}$ in $\mathbb{R}^d$, so structure sharing facilitates the optimization when we are learning by gradient descent. For example, the latent space learned by planar flows (Rezende and Mohamed, 2015) has compression and expansion, and vMF-VAE (Xu and Durrett, 2018), which is supported on a sphere, may significantly influence the distribution of $\mathcal{H}$ in its ambient space $\mathbb{R}^d$.

5 Experiments

5.1 Datasets

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.

5.2 Baselines

We evaluate the proposed Coupled-VAE approach by applying it to various VAE models, which in-

---

[2]To avoid heavy hyperparameter tuning, we set $\lambda_r = 1.0$ unless otherwise specified.

[3]To avoid heavy hyperparameter tuning, we use the Rational Quadratic kernel unless otherwise specified.

	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
Coupled-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
Coupled-$\beta(0.8)$-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
Coupled-$\beta(1.2)$-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
Coupled-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
Coupled-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
Coupled-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
Coupled-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
Coupled-WAE-NF	**102.7 (7.4)**	**108.4**	**192.1 (7.4)**	**52.5**	**342.7 (11.0)**	**73.1**
CycAnn-VAE	104.2 (1.6)	116.3	192.5 (1.2)	53.0	345.4 (3.9)	75.5
Coupled-CycAnn-VAE	**103.7 (2.4)**	**113.3**	**190.8 (2.0)**	**51.1**	**342.4 (4.4)**	**72.7**
PreFB-VAE	103.4 (14.6)	111.9	190.4 (14.1)	50.7	341.4 (17.6)	71.8
Coupled-PreFB-VAE	**103.3 (15.6)**	**111.4**	**189.9 (14.4)**	**50.3**	**341.3 (17.9)**	**71.7**
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
Coupled-Lagging-VAE[†]	**98.7 (11.0)**	**90.4**	**182.3 (3.8)**	**42.9**	**326.2 (7.4)**	**59.3**

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. *Coupled-* stands for "with the coupled deterministic network". The better results in each block are shown in **bold**. *The exact NLL is reported. [†]Modifying open-source implementation which does not follow our setup and evaluation. [‡]Previously reported.

clude VAE (Kingma and Welling, 2014), β-VAE (Higgins et al., 2017), vMF-VAE (Xu and Durrett, 2018; Davidson et al., 2018) with learnable κ, CNN-VAE (Yang et al., 2017), WAE (Tolstikhin et al., 2018), VAE with normalizing flows (VAE-NF) (Rezende and Mohamed, 2015), WAE with normalizing flows (WAE-NF), VAE with cyclic annealing schedule (CycAnn-VAE) (Fu et al., 2019), VAE with encoder pretraining and the free bits objective (PreFB-VAE) (Li et al., 2019), 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 amortized variational inference. Please find more details in Appendix C and previous footnotes.

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

5.4 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 z under $\mathcal{Q}(z|x)$, which is approximated with Monte Carlo estimation. Better

	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
Coupled-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
Coupled-$\beta(0.8)$-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
Coupled-$\beta(1.2)$-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
Coupled-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
Coupled-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
Coupled-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
Coupled-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
Coupled-WAE-NF	**8.10**	**20.7 / 3.7**	**8.53**	**27.2 / 5.0**	**12.56**	**14.9 / 2.5**
CycAnn-VAE	1.55	16.3 / 2.3	1.18	22.6 / 3.2	3.09	8.3 / 1.1
Coupled-CycAnn-VAE	**2.27**	**16.7 / 2.6**	**2.01**	**23.1 / 3.4**	**3.89**	**10.9 / 1.5**
PreFB-VAE	20.6	25.5 / 5.7	20.3	33.1 / 6.8	26.2	27.2 / 5.2
Coupled-PreFB-VAE	**23.2**	**25.8 / 5.8**	**21.0**	**33.3 / 6.8**	**27.0**	**27.2 / 5.3**
Lagging-VAE[†]	2.90	-	0.96	-	3.04	-
Coupled-Lagging-VAE[†]	**3.29**	-	**2.36**	-	**3.06**	-

Table 2: Mutual information (MI) and reconstruction. [†]Modifying the open-source implementation.

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 a rich latent space.

5.5 Hyperparameter Analysis: Distance Function, λ_r, and λ_m

We investigate the effect of key hyperparameters. Results are shown in Table 3. Note that the lowest NLL does not guarantee the best other metrics, which shows the necessity to use multiple metrics for a more comprehensive evaluation.

For the distance function, we observe that the Euclidean distance (denoted as Eucl in Table 3) is more sensitive to λ_m than the Rational Quadratic kernel (denoted as RQ in Table 3).

The first and the third block in Table 3 show that, with larger λ_m, the model achieves higher KL divergence, MI, and reconstruction metrics. Our interpretation is that by pushing the stochastic decoding signals closer to the deterministic ones, we get latent codes with richer text information. We leave the analysis of $\lambda_m = 0.0$ in Section 5.6.

The second block in Table 3 shows the role of λ_r, which we interpret as follows. When λ_r is too small (e.g., 0.5), the learned parameterizations are still inadequate for a smooth transition map; when λ_r is too large (e.g., 5.0), it distracts the optimization too far away from the original objective (i.e., $\mathcal{L}_{rec} + \mathcal{L}_{reg}$). Note that $\lambda_r = 0.0$ is equivalent to removing the coupled reconstruction loss $\mathcal{L}_{rec}^c$ in Eq. (4)).

5.6 The Heterogeneous Effect of Signal Matching on Probability Estimation

In Section 5.5 we observe richer latent space (i.e., larger MI and BLEU scores) with larger λ_m. However, a richer latent space does not guarantee a better probability estimation result. Thus, in this

Dist	λ_m	λ_r	PTB				Yelp			
			NLL (KL)	PPL	MI	BLEU-1/2	NLL (KL)	PPL	MI	BLEU-1/2
RQ	0.1*	1.0	**103.1 (9.5)**	**110.5**	11.99	23.4 / 4.5	191.2 (8.0)	51.6	9.65	30.4 / 5.8
	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
	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**
RQ	0.1	0.0	104.1 (7.3)	115.3	8.60	21.0 / 3.7	191.7 (5.8)	52.1	6.40	27.7 / 5.0
		0.5	103.4 (9.2)	111.8	11.58	23.1 / 4.3	191.3 (7.8)	51.7	9.32	29.8 / 5.7
		1.0*	**103.1 (9.5)**	**110.5**	**11.99**	**23.4 / 4.5**	**191.2 (8.0)**	**51.6**	**9.65**	**30.4 / 5.8**
		5.0	**103.1 (9.1)**	110.6	11.15	22.9 / 4.4	192.9 (8.0)	53.4	9.53	30.0 / 5.8
Eucl	0.1	1.0	**103.3 (10.1)**	**111.5**	13.25	23.4 / 4.7	**191.2 (9.2)**	**51.6**	11.69	31.1 / 6.0
	1.0		103.9 (17.4)	114.5	30.52	27.7 / 6.1	192.1 (14.3)	52.5	23.14	33.8 / 6.9
	5.0		108.9 (33.3)	144.0	**98.02**	**32.0 / 8.5**	194.4 (25.0)	55.1	**61.62**	**36.8 / 8.2**
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

Table 3: Hyperparameter analysis. The best results in each block are shown in **bold**. *Reported in Table 1 and 2.

	PTB		Yelp	
	NLL	PPL	NLL	PPL
Coupled-VAE*	103.1	110.5	191.2	51.6
Coupled-VAE (λ_m=0)	**103.1**	**110.3**	**190.7**	**51.1**
Coupled-VAE-NF*	**102.6**	**108.1**	**191.8**	**52.2**
Coupled-VAE-NF (λ_m=0)	102.8	109.1	192.7	53.2
Coupled-vMF-VAE*	**103.0**	**110.1**	**191.2**	**51.6**
Coupled-vMF-VAE (λ_m=0)	104.4	117.1	193.5	54.1

Table 4: The effect of signal matching on probability estimation. * Reported in Table 1.

part, we delve deeper into whether the decoder signal matching mechanism helps improve probability estimation. We study three models of different posterior families (i.e., Coupled-VAE, Coupled-VAE-NF, and Coupled-vMF-VAE). Results are shown in Table 4, where we do not report the KL, MI, and BLEU scores because they have been shown to be improved with larger λ_m in Table 3. We observe that the effects of signal matching on probability estimation vary in different posterior families.

5.7 Is the Incompatibility Mitigated?

We study the three gradient norms defined in Section 3 on the test sets, displayed in Table 5 (for Coupled-VAE, $\lambda_m = 0.1$). Notably, $\|\partial \mathcal{L}^c_{rec}/\partial e\|_2$ in Coupled-VAE is even larger than $\|\partial \mathcal{L}_{rec}/\partial e\|_2$ in DAE. It has two indications. First, the encoder indeed encodes rich information of the text. Second, compared with DAE, Coupled-VAE better generalizes to the test sets, which we conjecture is due to the regularization on the posterior. Coupled-VAE also has a larger $\|\partial \mathcal{L}_{reg}/\partial e\|_2$ compared with VAE, which based on the argument in Section 3.1 indicates that, in Coupled-VAE, the posterior of each

instance is not similar to the prior. We also observe larger $\|\partial h/\partial e\|_F / \|h\|_2$ in Coupled-VAE, which indicates a better transition map between the two parameterizations in Coupled-VAE than in VAE.

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

5.8 Sample Diversity

We evaluate the diversity of the samples from the prior distribution. We sample 3200 texts from the prior distribution 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. Distinct-1 and Distinct-2 in Table 6 show that texts sampled from Coupled-VAE ($\lambda_m = 10.0$) are more diverse than those from VAE. Given limited space, we put several samples in Appendix G for qualitative analysis.

5.9 Interpolation

A property of VAE is to match the interpolation in the latent space with the smooth transition in the data space (Bowman et al., 2016). In Table 7, we show the interpolation of VAE and Coupled-VAE on PTB. It shows that compared with VAE, Coupled-VAE has smoother transitions of subjects

		$\|\partial\mathcal{L}_{rec}/\partial e\|_2$	$\|\partial\mathcal{L}_{rec}^c/\partial e\|_2$	$\|(\partial\mathcal{L}_{rec}+\mathcal{L}_{rec}^c)/\partial e\|_2$	$\|\partial\mathcal{L}_{reg}/\partial e\|_2$	$\|\partial h/\partial e\|_F / \|h\|_2$
PTB	DAE	1719.8	-	-	-	3.14
	VAE	112.5	-	-	19.4	2.05
	Coupled-VAE	**148.5**	2109.6	2320.2	**27.7**	**2.12**
Yelp	DAE	2443.6	-	-	-	2.55
	VAE	59.7	-	-	18.8	1.62
	Coupled-VAE	**84.8**	3640.8	3764.7	**25.0**	**2.25**
Yahoo	DAE	4104.6	-	-	-	3.39
	VAE	257.9	-	-	52.8	2.92
	Coupled-VAE	**335.3**	5105.0	5615.0	**65.0**	**3.91**

Table 5: Gradient norms defined in Section 3.1 on each test set. $\lambda_m = 0.1$.

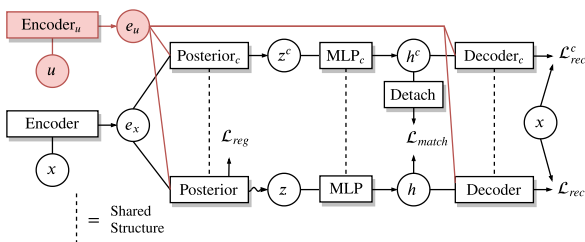|	PTB		Yelp		Yahoo	
	D-1	D-2	D-1	D-2	D-1	D-2
VAE	4.61	16.36	0.62	2.48	0.44	2.11
Coupled-VAE	**5.51**	**24.46**	**1.15**	**5.93**	**0.75**	**3.97**

Table 6: Diversity of samples from the prior.

Figure 4: A graphical overview of the generalization to Coupled-CVAE. u is the condition, encoded as e_u.

(*both sides* → *it*) and verbs (*are expected* → *have been* → *has been* → *has*), indicating that the linguistic information is more smoothly encoded in the latent space of Coupled-VAE.

5.10 Generalization to Conditional Language Modeling: Coupled-CVAE

To generalize our approach to conditional language modeling, we propose Coupled-**C**VAE. A graphical overview is displayed in Figure 4. Specifically, the (coupled) posterior network and the (coupled) decoder are additionally conditioned. The objective of Coupled-CVAE is identical to Eq. (4).

We compare Couple-CVAE with GRU encoder-decoder (Cho et al., 2014) and CVAE (Zhao et al., 2017) for dialogue generation. We use the Switchboard dataset (John and Holliman, 1993), whose training/validation/test splits are 203K/5K/5K, and the vocabulary size is 13K. For probability estimation, we report the NLL, KL, and PPL based on the gold responses. Since the key motivation of using CVAE in Zhao et al. (2017) is the diversity of responses, we sample one response for each post and report the Distinct-1 and Distinct-2 metrics over all samples. Please find more details in Appendix I.

Table 8 shows that Coupled-CVAE greatly increases the diversity of dialogue modeling, while it only slightly harms the probability estimation capability. It indicates that Coupled-CVAE better captures the one-to-many nature of conversations than CVAE and GRU encoder-decoder. We also observe that the diversity is improved with increasing λ_m, which shows that λ_m can control diversity via specifying the richness of the latent space.

6 Relation to Related Work

Bowman et al. (2016) identify the posterior collapse problem of text VAE and propose KL annealing and word drop to handle the problem. Zhao et al. (2017) propose the bag-of-words loss to mitigate this issue. Later work on this problem focuses on less powerful decoders (Yang et al., 2017; Semeniuta et al., 2017), modified regularization objective (Higgins et al., 2017; Bahuleyan et al., 2019; Wang and Wang, 2019), alternative posterior families (Rezende and Mohamed, 2015; Xu and Durrett, 2018; Davidson et al., 2018; Xiao et al., 2018), richer prior distributions (Tomczak and Welling, 2018), improved optimization (He et al., 2019) or KL annealing strategy (Fu et al., 2019), the use of skip connections (Dieng et al., 2019), 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). We provide the encoder-decoder incompatibility as a new perspective on the posterior collapse problem. Empirically, our approach can be combined with the above ones to alleviate the problem further.

VAE	Coupled-VAE ($\lambda_m = 10.0$)
Text A (sampled from 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 (sampled from PTB): it also has a _unk_ facility in california	

Table 7: Latent space interpolation.

	NLL (KL)	PPL	D-1	D-2
GRU Encoder-Decoder*	**53.9 (-)**	**41.6**	0.33	0.80
CVAE	54.0 (3.8)	41.8	0.61	2.60
Coupled-CVAE (λ_m=0.1)	54.1 (4.6)	42.2	0.71	3.18
Coupled-CVAE (λ_m=0.5)	54.2 (5.3)	42.5	0.78	3.63
Coupled-CVAE (λ_m=1.0)	54.3 (6.1)	42.7	0.86	4.10
Coupled-CVAE (λ_m=2.0)	54.6 (7.8)	43.6	**0.99**	**5.16**

Table 8: Dialogue generation. *Exact NLL is reported.

A model to be noted is β-VAE (Higgins et al., 2017), in which the reconstruction and regularization are modeled as a hyperparameterized trade-off, i.e., the improvement of one term compromises the other. Different from β-VAE, we adopt the idea of multi-task learning, i.e., the *coupled reconstruction* task helps improve the encoder chart map and the *signal matching* task helps improve the decoder chart map. Both our analysis in Section 3.2 and the empirical results show that the modeling of posterior distribution can be improved (but not necessarily compromised) with the additional tasks.

Ghosh et al. (2020) propose to substitute stochasticity with explicit and implicit regularizations, which is easier to train and empirically improves the quality of generated outputs. Different from their work, we still strictly follow the generative nature (i.e., data density estimation) of VAE, and the deterministic network in our approach serves as an auxiliary to aid the optimization.

Encoder pretraining (Li et al., 2019) initializes the text encoder and the posterior network with an autoencoding objective. Li et al. (2019) shows that encoder pretraining itself does not improve the performance of VAE, which indicates that initialization is not strong enough as an inductive bias to learn a meaningful latent space.

Given the discrete nature of text data, we highlight the two-level representation learning for text modeling: 1) the encoder and decoder parameterizations via autoencoding and 2) a transition map between the parameterizations. Notably, the transition map has large freedom. In our case, the transition map decides the amount and type of information encoded in the variational posterior, and there are other possible instances of the transition map, e.g., flow-based models (Dinh et al., 2015).

7 Conclusions

In this paper, we observe the encode-decoder incompatibility of VAE for text modeling. We bridge the incompatibility and the posterior collapse problem by viewing the encoder and the decoder as two inadequately learned chart maps from the data manifold to the parameterizations, and the posterior network as a part of the transition map between them. We couple the VAE model with a deterministic network and improve the parameterizations via encoder weight sharing and decoder signal matching. Our approach is model-agnostic and can be applied to a wide range of models in the VAE family. Experiments on benchmark datasets, i.e., PTB, Yelp, and Yahoo, show that our approach improves various VAE models in terms of probability estimation and the richness of the latent space. We also generalize Coupled-VAE to conditional language modeling and propose Coupled-CVAE. Results on Switchboard show that Coupled-CVAE largely improves diversity in dialogue generation.

Acknowledgments

References

Hareesh Bahuleyan, Lili Mou, Hao Zhou, and Olga Vechtomova. 2019. 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)*, pages 4068–4076.

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

Yoshua Bengio, Aaron C. Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1798–1828.

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew M. Dai, Rafal Józefowicz, and Samy Bengio. 2016. 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*, pages 10–21.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. 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*, pages 1724–1734.

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

Adji B. Dieng, Yoon Kim, Alexander M. Rush, and David M. Blei. 2019. 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*, pages 2397–2405.

Laurent Dinh, David Krueger, and Yoshua Bengio. 2015. NICE: non-linear independent components estimation. In *ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*.

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

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 NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 240–250.

Partha Ghosh, Mehdi S. M. Sajjadi, Antonio Vergari, Michael Black, and Bernhard Scholkopf. 2020. From variational to deterministic autoencoders. In *ICLR 2020, Addis Ababa, Ethiopia, April 30, 2020*.

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

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

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

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

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

Godfrey John and Edward Holliman. 1993. Switchboard-1 release 2 ldc97s62. Web Download. *Philadelphia: Linguistic Data Consortium*.

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

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

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

Bohan Li, Junxian He, Graham Neubig, Taylor Berg-Kirkpatrick, and Yiming Yang. 2019. A surprisingly effective fix for deep latent variable modeling of text. In *Proceedings of EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3601–3612. Association for Computational Linguistics.

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

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.

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

Hariharan Narayanan and Sanjoy K. Mitter. 2010. Sample complexity of testing the manifold hypothesis. In *NIPS 2010, December 6-9, 2010, Vancouver, British Columbia, Canada.*, pages 1786–1794.

Yookoon Park, Jaemin Cho, and Gunhee Kim. 2018. 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)*, pages 1792–1801.

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

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

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

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

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

Jakub M. Tomczak and Max Welling. 2018. 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*, pages 1214–1223.

Prince Zizhuang Wang and William Yang Wang. 2019. 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)*, pages 284–294.

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

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

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

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

Tiancheng Zhao, Ran Zhao, and Maxine Eskénazi. 2017. 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*, pages 654–664.

Appendix

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. The 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 z^c as

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

where $\mathcal{Q}^c(z|x)$ is the posterior distribution learned by the coupled deterministic network. In effect, z is the mean vector predicted by the coupled posterior network 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., $z_0 \sim Q_0(z_0|x)$. The normalizing flow is defined as a series of *reversible* transformations $f_1, \ldots, f_K$, i.e.,

$$z_k = f_k \circ \cdots \circ f_1(z_0) \quad (7)$$

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

$$
\begin{aligned}
\log P(x) &\geq \mathbb{E}_{z_K \sim Q_K(z_K|x)}[\log P(x|z_K)] \\
&\quad - \text{KL}[Q_K(z_K|x) \parallel \mathcal{P}_K(z_K)] \\
&= \mathbb{E}_{z_0 \sim Q_0(z_0|x)}\Big[\log P(x|z_K) \\
&\quad - \log q_0(z_0|x) + \log p_K(z_K) \\
&\quad + \sum_{k=1}^{K} \log |\det \frac{\partial f_k}{\partial z_{k-1}}| \Big]
\end{aligned}
\quad (8)
$$

where $\mathcal{P}_K(z_K)$ 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 z^c by transforming the predicted mean vector of the coupled initial distribution, i.e.,

$$z^c = f_k^c \circ \cdots \circ f_1^c(\mathbb{E}_{z_0 \sim Q_0^c(z_0|x)}[z_0]) \quad (9)$$

where $Q_0^c(z_0|x)$ is the coupled initial distribution and $f_1^c, \ldots, f_K^c$ 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 $\mu \in \mathbb{R}^d$ ($\|\mu\| = 1$) and a concentration parameter κ, both of which are mapped from the encoded text by the posterior network. The probability density function is

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

where I_v is the modified Bessel function of the first kind at order v. We use the direction parameter μ as the coupled latent code 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 and 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, Coupled-VAE, β-VAE, and Coupled-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 (NF), we use planar flows (Rezende and 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. 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:

$$\text{NLL}_x = -\log P(x)$$

$$\approx -\log\left(\frac{1}{N}\sum_{i=1}^{N}\frac{p(\boldsymbol{z}^{(i)})P(x|\boldsymbol{z}^{(i)})}{q(\boldsymbol{z}^{(i)}|x)}\right)$$

$$\text{KL}_x = \text{KL}[\mathcal{Q}(\boldsymbol{z}|x) \parallel \mathcal{P}(\boldsymbol{z})] \tag{11}$$

$$\approx \frac{1}{N}\sum_{i=1}^{N}\log\frac{q(\boldsymbol{z}^{(i)}|x)}{p(\boldsymbol{z}^{(i)})}$$

where $\boldsymbol{z}^{(i)} \sim \mathcal{Q}(\boldsymbol{z}|x)$ are sampled latent codes and all notations follow Eq. (1) in the main text. 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 Metrics

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:

$$\text{MI}_x = \mathbb{E}_{\boldsymbol{z}\sim\mathcal{Q}(\boldsymbol{z}|x)}\left[\log\frac{q(\boldsymbol{z}|x)}{q(\boldsymbol{z})}\right]$$

$$\approx \frac{1}{N}\sum_{i=1}^{N}(\log q(\boldsymbol{z}^{(i)}|x) - \log q(\boldsymbol{z}^{(i)})) \tag{12}$$

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

$$q(\boldsymbol{z}^{(i)}) = \mathbb{E}_x[q(\boldsymbol{z}^{(i)}|x)] \approx \frac{1}{M}\sum_{j=1}^{M}q(\boldsymbol{z}^{(i)}|x^{(j)}) \tag{13}$$

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 $= \mathbb{E}_x[\text{MI}_x]$ 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.

F   Training Dynamics of Gradient Norms

We show the tracked gradient norms on all datasets in Figure 5. The observations are consistent with those discussed in Section 5.7 in the main text.

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 9 shows the diversity metrics and the first three (thus totally random) samples from each model. Qualitatively, samples from Coupled-VAE is more diverse than those from VAE. The long texts generated from VAE have more redundancies compared with Coupled-VAE. Given that both models have the same latent dimension, it indicates that Coupled-VAE is using the latent codes more efficiently.

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 Coupled-VAE on PTB. It shows that compared with VAE, Coupled-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 Coupled-VAE.

I   Generalization to Conditional Generation: Coupled-CVAE

To generalize our approach to conditional generation, we focus on whether it can improve the CVAE model (Zhao et al., 2017) for dialogue generation. To this end, we propose the Coupled-CVAE model.

I.1   CVAE

CVAE adopts a two-step view of diverse dialogue generation. Let x be the response and y be the post (or the context). CVAE first samples the latent code $\boldsymbol{z}$ from the prior distribution $\mathcal{P}(\boldsymbol{z}|y)$ and then samples the response from the decoder $P(x|\boldsymbol{z}, y; \theta)$. Given the post y, the marginal distribution of the response x is

$$P(x|y; \theta) = \mathbb{E}_{\boldsymbol{z}\sim\mathcal{P}(\boldsymbol{z}|y)}[P(x|\boldsymbol{z}, y; \theta)] \tag{14}$$

Similar to VAE, the exact marginalization is intractable, and we derive the evidence lower bound
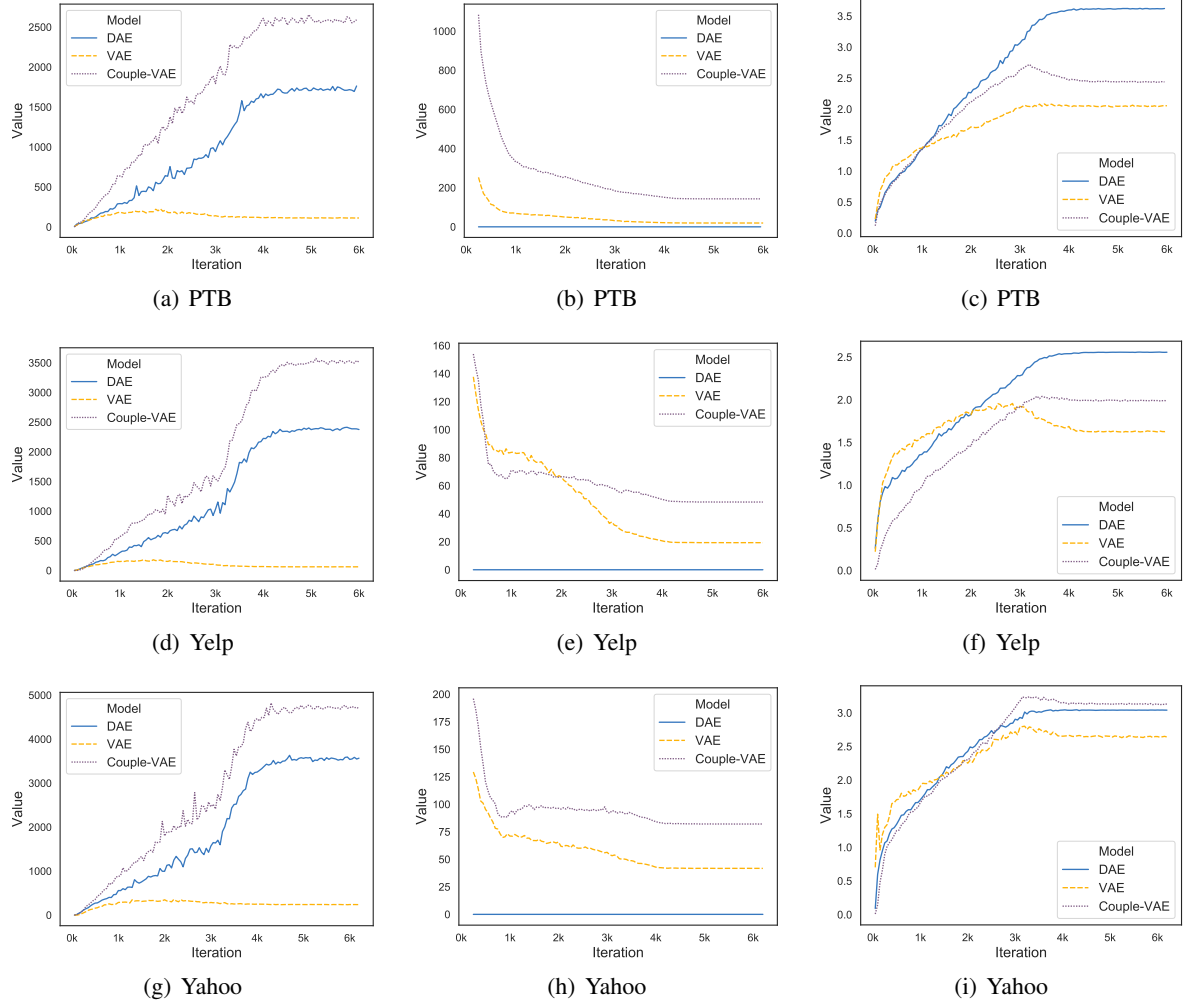
Figure 5: Training dynamics of DAE, VAE, and Coupled-VAE ($\lambda_m = 10.0$). (a), (d), and (g) are $\|\partial \mathcal{L}_{rec}/\partial e\|_2$ for DAE and VAE, and $\|\partial(\mathcal{L}_{rec} + \mathcal{L}_{rec}^c)/\partial e\|_2$ for Coupled-VAE. (b), (e), (h) denote $\|\partial \mathcal{L}_{reg}/\partial e\|_2$. (c), (f), (i) stand for $\|\partial h/\partial e\|_F / \|h\|_2$. Best viewed in color (yet the models are distinguished by line markers).

(ELBO) of CVAE as

$$\log P(x|y;\theta) \geq \mathbb{E}_{z \sim \mathcal{Q}(z|x,y;\phi)}[\log P(x|z,y;\theta)] \\ - \mathrm{KL}[\mathcal{Q}(z|x,y;\phi) \parallel \mathcal{P}(z|y)] \tag{15}$$

During training, the response and the post are encoded as e_x and e_y, respectively. The two vectors are concatenated and transformed into the posterior via the posterior network. A latent code is then sampled and mapped to a higher-dimensional h. The decoding signal in CVAE is computed by h and e_y and utilized to infer the response. Similar to VAE, the objective of CVAE can also be viewed as a reconstruction loss and a regularization term in Eq. (15).

I.2 Coupled-CVAE

As observed in Zhao et al. (2017), the CVAE model also suffers from the posterior collapse problem. We generalize our approach to the conditional setting and arrive at Coupled-CVAE. A graphical overview is displayed in Figure 4. The difference from Coupled-VAE is shown in red. Specifically, the (coupled) posterior network and the (coupled) decoder are additionally conditioned on the post representation. The objective of Coupled-CVAE is identical to Eq. (4) in the main text.

The coupled reconstruction loss $\mathcal{L}_{rec}^c$ in Coupled-CVAE has two functions. First, it improves the encoded response e_x, which is similar to Coupled-VAE. Second, it encourages h^c to encode more response information rather than the post information, which collaborates with $\mathcal{L}_{match}$ to improve the parameterization h.

3462

I.3 Dataset

We use the Switchboard dataset (John and Holliman, 1993). We split the dialogues into single-turn post-response pairs, and the number of pairs in the training/validation/test split is 203K/5K/5K. The vocabulary size is 13K.

I.4 Evaluation

For probability estimation, we report the NLL, KL, and PPL based on the gold responses. NLL, KL, and PPL are as computed in Appendix D except for the additional condition on the post. Since the key motivation of using CVAE in Zhao et al. (2017) is the response diversity, we sample one response for each post and report the Distinct-1 and Distinct-2 metrics over all test samples.

I.5 Experimental Setup

We compare our Coupled-CVAE model with two baselines: GRU encoder-decoder (Cho et al., 2014) and CVAE (Zhao et al., 2017). The detailed setup follows that of the PTB dataset in Appendix C. For each 1K steps, we estimate the NLL for validation.

I.6 Results

Experimental results of Coupled-CVAE are shown in the main text.

VAE (PTB)	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

Coupled-VAE ($\lambda_m = 10.0$) (PTB)	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

Coupled-VAE ($\lambda_m = 10.0$) (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 "

Coupled-VAE ($\lambda_m = 10.0$) (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 9: Diversity metrics and the first three samples from each model. Redundancies (pieces of text that appeared before) are shown in red.