Amortized Mixture Prior for Variational Sequence Generation

Jen-Tzung Chien
Department of Electrical and Computer Engineering
National Chiao Tung University
Hsinchu, Taiwan

Chih-Jung Tsai
Department of Electrical and Computer Engineering
National Chiao Tung University
Hsinchu, Taiwan

***Abstract*—Variational autoencoder (VAE) is a popular latent variable model for data generation. However, in natural language applications, VAE suffers from the posterior collapse in optimization procedure where the model posterior likely collapses to a standard Gaussian prior which disregards latent semantics from sequence data. The recurrent decoder accordingly generates duplicate or noninformative sequence data. To tackle this issue, this paper adopts the Gaussian mixture prior for latent variable, and simultaneously fulfills the *amortized* regularization in encoder and *skip* connection in decoder. The noise robust prior, learned from the amortized encoder, becomes semantically meaningful. The prediction of sequence samples, due to skip connection, becomes contextually precise at each time. The amortized mixture prior (AMP) is then formulated in construction of variational recurrent autoencoder (VRAE) for sequence generation. Experiments on different tasks show that AMP-VRAE can avoid the posterior collapse, learn the meaningful latent features and improve the inference and generation for semantic representation.**

***Index Terms*—Sequence generation, recurrent neural network, variational autoencoder, language model**

I. Introduction

Generative models have been emerging in the era of machine learning and deep learning in recent years where the applications of computer vision [1], [2] and natural language processing [3]–[9] have been explored. Variational autoencoder (VAE) [10]–[13], generative adversarial network [14], normalizing flow [15]–[17] and autoregressive neural network [18] have been extensively proposed with a variety of applications [19]–[21]. Among these models, VAE has the advantage of utilizing latent variables in model construction. By applying the variational inference in deep latent variable model, VAE is trained to estimate the distribution of latent variables [22] or the underlying structure of disentangled features [23] from input observations. VAE consists of an encoder as inference model and a decoder as generative model. The encoder obtains latent representation corresponding to input data while the decoder generates the synthesized data given by latent samples. Encoder and decoder are jointly trained by maximizing the evidence lower bound (ELBO) of log marginal likelihood of training data.

In spite of a great success, VAE still faces different challenges when generating sequence data [24], [25]. Previous studies [26]–[29] showed that vanilla VAE could not generate meaningful sentences. The distribution of latent variable is reduced to a standard Gaussian. The generated samples lack diversity. This phenomenon is undesirable since the variational posterior does not depend on input data. The issue of *posterior collapse* happens because the Kullback-Leibler (KL) divergence between variational posterior and standard Gaussian prior in ELBO approaches to zero. The variational posterior barely learns any information from input data. This causes meaningless latent representation. Sampling from this posterior has no difference from sampling by a standard Gaussian [30]. VAE is then reduced as an autoregressive generative model where the underlying structure of data was disregarded. To tackle this problem, the von Mises-Fisher distribution was used to replace Gaussian distribution in latent variable representation [31]. In [27], the convolutional neural network was used as hierarchical decoder which coped with this issue by restricting the receptive field in temporal convolutional network.

This paper presents a new solution to deal with the dilemma in variational sequential learning due to minimization of KL term in variational optimization. Motivated by [32], [33], we learn an informative prior by using Gaussian mixture model which encourages a flexible construction of latent space from training data. In the implementation, the encoder and decoder are further strengthened by performing the amortized regularization and skip connection, respectively. Amortized regularization leads to a smooth encoder, especially for sequence data. This smooth encoder compresses the neighboring sequences from observation space into nearby locations in latent space. Owing to the preservation of semantic information, the latent code is embedded with semantic meaning of sequence data. In addition, the skip connection from latent code to hidden state in recurrent network is performed at each time so as to enrich latent information for prediction of output sequence. Information loss is reduced during propagation of recurrent steps in decoding which leads to a desirable latent space. A set of experiments are evaluated to illustrate the proposed method.

II. Variational Neural Models

Variational neural models based on variational autoencoder (VAE) and variational recurrent autoencoder (VRAE) are first introduced. Basically, VAE is a neural machine consisting of an encoder and a decoder where the encoder acts as an inference model for latent distribution and the decoder serves as a generative model for synthesized data from latent

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

distribution. VAE is learned from a set of training signals $\mathbf{x}$ by maximizing over a model with decoder parameter θ and latent variable $\mathbf{z}$. In learning procedure, VAE introduces a variational distribution $q_\phi(\mathbf{z}|\mathbf{x})$ with encoder parameter ϕ to approximate true posterior $p(\mathbf{z}|\mathbf{x})$. The evidence lower bound (ELBO) of log marginal likelihood is then formulated for maximization as [34], [35]

$$\begin{aligned}\log p(\mathbf{x}) &= \log \int p_\theta(\mathbf{x}|\mathbf{z})p(\mathbf{z})\mathrm{d}\mathbf{z} \\ &\geq \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right] - D_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{x})\|p(\mathbf{z}))\end{aligned} \tag{1}$$

where $p(\mathbf{z}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is a standard Gaussian, and the first and second terms in RHS denote the reconstruction and regularization losses, respectively. The so-called posterior collapse happens when the regularization loss goes to zero. In [16], [20], [36], the normalizing flow was proposed to obtain flexible variational distribution. Normalizing flow utilized *invertible* transformation to assure richness in latent distribution.

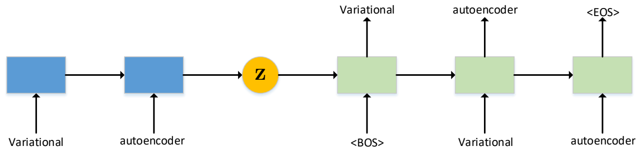


Fig. 1: Illustration for variational recurrent autoencoder.

In [26], [37], variational recurrent autoencoder (VRAE) was proposed for sequence generation where two individual recurrent neural networks (RNNs) [38] were employed in encoder and decoder. VRAE was developed for modeling of music signals and text data. In sequential learning procedure, the encoder RNN recursively characterizes time samples $\mathbf{x} = \{\mathbf{x}_t\}_{t=1}^T$ and finally infers latent distribution of $\mathbf{z}$ using the hidden state at last time $\mathbf{h}_T$. Meanwhile, the decoder RNN reconstructs the input sequence $\mathbf{x}$ recursively from the initial hidden state $\mathbf{s}_0$ and the first input <BOS>. The initial hidden state $\mathbf{s}_0$ is a mapping or embedding of latent code $\mathbf{z}$ sampled from latent distribution, and the first input usually sets as begin of sentence <BOS>. Figure 1 illustrates the recursive nature of VRAE. In practice, the teacher forcing is imposed on decoder to prevent wrong predictions recursively affecting later predictions during training procedure.

III. Amortized Mixture Prior VRAE

To deal with the difficulties in sequential learning procedure, we propose the amortized mixture prior (AMP) for VRAE and apply AMP-VRAE for sentence generation which tackles the issue of posterior collapse from three perspectives which range from *encoder* to *decoder* and *latent distribution*.

A. Amortized regularization on encoder

Traditionally, variational inference optimizes $p(\mathbf{x})$ by using individual seen samples $\mathbf{x} = \{\mathbf{x}_t\}$. This is usually impractical when a large dataset $\mathbf{x}$ is adopted. Amortized variational inference replaces the per-sample optimization over $p(\mathbf{x})$ by means of an inference model $q_\phi(\mathbf{z}|\mathbf{x})$ driven by encoder parameter ϕ [39]. In general, VAE relies on the amortized inference which accelerates the computation by amortizing the computation of optimization on each sample. An inference model is then optimized by using all data samples. In addition to this acceleration, the amortized variational inference was treated as a regularization for maximum likelihood estimation [21]. An alternative learning objective to ELBO was derived as

$$\begin{aligned}\max_\theta \Big\{ &\mathbb{E}_{\hat{p}(\mathbf{x})}\left[\log p_\theta(\mathbf{x})\right] \\ &- \min_\phi \mathbb{E}_{\hat{p}(\mathbf{x})} D_{\mathrm{KL}}\left(q_\phi(\mathbf{z}|\mathbf{x})\|p_\theta(\mathbf{z}|\mathbf{x})\right)\Big\}\end{aligned} \tag{2}$$

where a uniform distribution $\hat{p}(\mathbf{x})$ over a dataset $\mathbf{x}$ is used. The choice of variational distribution parameter ϕ or amotized inference model $q_\phi(\mathbf{z}|\mathbf{x})$, via KL minimization, actually regularizes the optimization of marginal likelihood $p_\theta(\mathbf{x})$. By injecting the noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma\mathbf{I})$, i.e. using $q_\phi(\mathbf{z}|\mathbf{x}+\boldsymbol{\epsilon})$ in Eq. (2), the *denoising VAE* is constructed to regularize the mapping or control the smoothness of an inference model. Smoothness indicates that neighboring data samples are mapped to similar locations in latent space. This property is incorporated in VRAE so that the sequence embedding can preserve semantic information, or equivalently the words with similar meanings are embedded with similar mappings.

B. Mixture prior for latent distribution

Using VAE, the prior distribution $p(\mathbf{z})$ is often assumed to be standard Gaussian. Such a naive assumption typically leads to over regularization or posterior collapse in the estimated variational distribution $q_\phi(\mathbf{z}|\mathbf{x})$, as reflected by KL term in Eq. (1). Inspired by [33], we estimate the variational prior for VRAE by using N_v *validation* sequences or sentences $\mathbf{x} = \{\mathbf{x}_n\}$

$$p_\lambda(\mathbf{z}) = \frac{1}{N_v}\sum_{n=1}^{N_v} q_\phi(\mathbf{z}|\mathbf{x}_n) \tag{3}$$

which is seen as a sentence-level mixture prior of latent variable $\mathbf{z}$. The restriction of simple prior is then relaxed. ELBO is maximized to find $p_\lambda(\mathbf{z})$. However, this prior becomes infeasible if N_v is large. In [33], pseudo inputs were introduced as additional parameters to learn the amortized prior. Nevertheless, pseudo inputs could not be used in sequential learning since the length of each sequence was not fixed. In the experiments, we estimate the amortized mixture prior by using validation sentences, which is efficient and memory saving. This prior is learned along with the variational posterior with a shared ϕ. The estimated prior leads to a multimodal and flexible latent space. This prior effectively prevents posterior collapse by learning semantic information in sequence data.

C. Skip connection on decoder

Skip connection has been widely used in deep learning, such as residual network [40] or highway network [41]. In [42], the scheme of skip connection was incorporated into VAE where the mutual information between observations $\mathbf{x}$

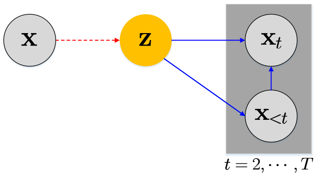


Fig. 2: Information paths for VRAE with skip connection.

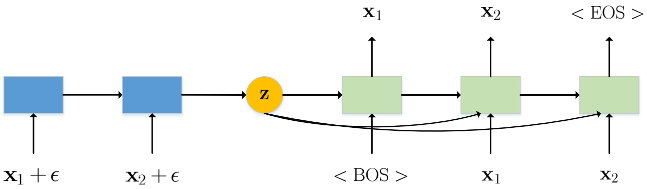


Fig. 3: Amortized mixture prior and skip connection in variational recurrent autoencoder.

and latent codes $\mathbf{z}$ was enhanced. Here, we address different perspective. As we know, RNN decoder predicts next sample $\mathbf{x}_t$ given by the previous inputs $\mathbf{x}_{<t}$. Information of data flows from encoder to decoder only through the initial hidden state $\mathbf{s}_0$. However, in training stage of VRAE, the usage of teacher forcing provides another source of information. Such a source information reduces the dependence on the latent space, and accordingly causes the issue of posterior collapse. With the skip connection, the latent code $\mathbf{z}$ is reinforced to join every prediction at different time steps. The information flow from encoder to the prediction is shortened. Figure 2 depicts how skip connection changes the generation process. By considering these three perspectives, AMP-VRAE is proposed and illustrated by an overall architecture as shown in Figure 3. The learning criterion of AMP-VRAE is therefore formulated by

$$\begin{aligned}&\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})]\\&- D_{\text{KL}}\left(q_\phi(\mathbf{z}|\mathbf{x}+\boldsymbol{\epsilon})\middle\|\frac{1}{N_v}\sum_{n=1}^{N_v} q_\phi(\mathbf{z}|\mathbf{x}_n)\right)\end{aligned} \tag{4}$$

where the noise term $\boldsymbol{\epsilon}$ is merged and the variational mixture prior is adopted in model training with skip connection.

IV. Experiments

In the experiments, different methods were evaluated by using three datasets: Penn TreeBank (PTB) [43], Yelp 2013 (Yelp) [31] and IMDB [44]. To conduct comparative study, we implemented the recently proposed methods including the amortized inference regularized (AIR) VRAE [21], the variational-mixture-of-posterior prior (VAMP) VRAE [33], and the normalizing flow (Flow) VRAE [20], [36]. There are four metrics in language modeling tasks [45]–[49] where negative log-likelihood (NLL) and perplexity (PPL) shows the ability of word generation and prediction, KL term value reflects if the model prevents the posterior collapse, and the number of active units (AU) investigates how well the inference model is active to work. While we have 32 dimensions in latent space, the models exploiting larger dimensions are considered to work better in inference procedure. An active dimension is defined to achieve its variance to be greater than 0.01.

A. Language modeling on Penn TreeBank

PTB is a benchmark dataset for evaluation of language model. In PTB, the average length of a sentence is 21.07 words. Vocabulary size is set to 8K. The proposed AMP-VRAE is compared with different models. Table I reports the results using different methods. In this comparison, AIR-VRAE employs the amortized regularization with noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, 0.5\mathbf{I})$. VAMP-VRAE utilizes the variational-mixture-of-posterior as prior estimated by validation data. Flow-VRAE uses the inverse autoregressive flow to transform the posterior. The flow comprises of ten convex combination of linear inverse autoregressive flows. Overall, AMP-VRAE obtains the best performance for generation, as it achieves the lowest NLL and PPL. It successfully prevents posterior collapse and has the most effective result with the largest KL value. For a 32 dimensional latent space, VAMP-VRAE fully exploits all the dimensions. Flow-VRAE utilizes 27 of them. AMP-VRAE reports similar AU, which is much larger than AU of baseline.

TABLE I: Evaluation of different methods using PTB.

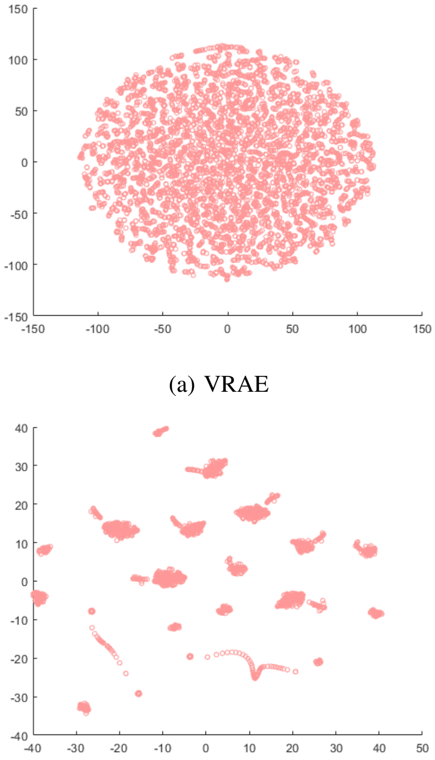
Model	**NLL**	**KL**	**PPL**	**AU**
VRAE	98.66	3.57	111.91	4
AIR-VRAE	98.30	4.10	109.96	7
VAMP-VRAE	98.83	4.17	112.79	**32**
Flow-VRAE	100.81	1.13	123.99	27
AMP-VRAE	**97.69**	**6.58**	**106.81**	25

TABLE II: Evaluation of different methods using Yelp.

Model	**NLL**	**KL**	**PPL**	**AU**
VRAE	194.76	0.96	60.40	2
AIR-VRAE	193.12	2.18	58.36	4
VAMP-VRAE	194.63	1.97	60.24	**26**
Flow-VRAE	194.79	0.49	60.45	19
AMP-VRAE	**191.80**	**5.18**	**56.76**	21

B. Language modeling on Yelp 2013

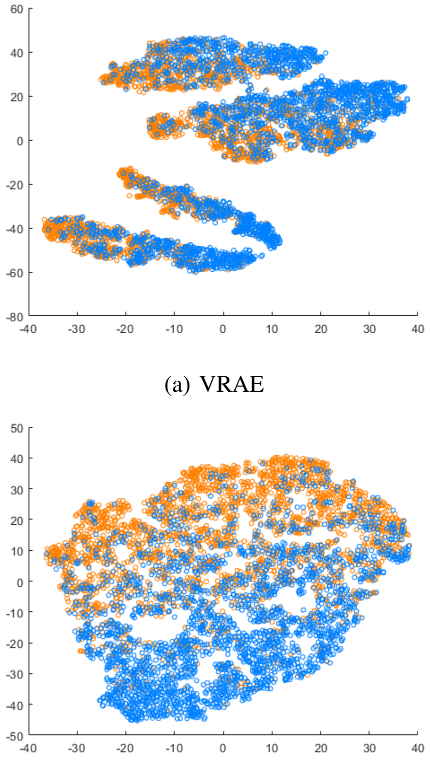
Yelp is a restaurant review dataset collected from Yelp Dataset Challenge in year 2013. There are 47.55 words in an averaged length sentence. Vocabulary size is 12K. Results are shown in Table II. AMP-VRAE achieves the lowest value on NLL and PPL, as well as highest value in KL divergence. It has 21 active units in 32 dimensions, just below 26 of VAMP-VRAE. AMP-VRAE improves VRAE and outperforms the other models in generation performance. It has competitive results in inference metric. We compare the latent space of VRAE and AMP-VRAE in Figure 4. We reduce the dimension to two with t-SNE [50] and show the global structure. VRAE has a latent space in the shape of a circle, which indicates



(a) VRAE

(b) AMP-VRAE

Fig. 4: Latent distributions of (a) VRAE and (b) AMP-VRAE by using Yelp 2013 dataset.



(a) VRAE

(b) AMP-VRAE

Fig. 5: Latent distributions of (a) VRAE and (b) AMP-VRAE by using IMDB dataset.

Gaussian distribution with diagonal covariances. AMP-VRAE, on the other hand, reports multi-modal mixture distribution, which leads to rich representation in the latent space.

TABLE III: Evaluation of different methods using IMDB.

Model	NLL	KL	PPL	AU	Accu
VRAE	387.67	1.56	141.58	2	68.04
AIR-VRAE	386.37	2.06	139.24	3	69.94
VAMP-VRAE	386.69	0.89	139.82	4	68.76
Flow-VRAE	387.86	0.85	141.92	**27**	67.94
AMP-VRAE	**386.92**	**2.32**	**140.23**	5	**71.30**

C. *Sentiment classification on IMDB*

IMDB is the movie review dataset collected from the Internet Movie Database website. IMDB contains 50K labeled data with even number of positive and negative reviews. The average length in a sentence is 78.17 words. Vocabulary size is 20K. Evaluation on sentiment classification is performed. An additional classifier is connected to the latent space. It predicts if the review is positive or negative. The classifier is jointly trained with VRAE. The accuracy is reported. Table III shows that AMP-VRAE achieves the best in most metrics. Although active units are less than those in Flow-VRAE, the accuracy is improved. We show the latent space of VRAE and AMP-VRAE in Figure 5. Orange indicates positive reviews while blue indicates negative reviews. AMP-VRAE can separate most reviews in different classes. However, VRAE has large spaces that two classes are overlapped.

V. Conclusions

We have presented the amortized mixture prior variational recurrent autoencoder for stochastic and sequential learning, and applied it in language modeling and sentiment analysis and classification. A number of strategies were proposed to improve neural sequential learning based on VRAE where different treatments in encoder, decoder and latent space were performed. First, the amortized regularization was adopted to encourage smoothing for encoder where the semantic information of an input sentence was sufficiently learned. Second, the incorporation of mixture prior in VRAE led to the richness in the estimated latent distributions. Third, the skip connection reinforced the latent code to join each prediction in the decoder at each time step for individual words. Experimental results on language modeling and sentiment classification over three tasks showed that the proposed method alleviated the issue of posterior collapse and improved the performance of VRAE in terms of inference and generation. In general, the proposed variational recurrent autoencoder is developed as a tool for semantic analysis and representation which can be extended to other natural language tasks or applications including document summarization and text classification.

References

[1] A. Van Oord, N. Kalchbrenner, and K. Kavukcuoglu, "Pixel recurrent neural networks," in *Proc. of International Conference on Machine Learning*, 2016, pp. 1747–1756.

[2] J.-T. Chien and Y.-T. Bao, "Tensor-factorized neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 5, pp. 1998–2011, 2018.

[3] Y. Miao, L. Yu, and P. Blunsom, "Neural variational inference for text processing," in *Proc. of International Conference on Machine Learning*, 2016, pp. 1727–1736.

[4] A. B. Dieng, C. Wang, J. Gao, and J. Paisley, "TopicRNN: A recurrent neural network with long-range semantic dependency," in *Proc. of International Conference on Learning Representations*, 2017.

[5] J.-T. Chien and C.-H. Lee, "Deep unfolding for topic models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 2, pp. 318–331, 2017.

[6] J.-T. Chien, "Hierarchical theme and topic modeling," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 3, pp. 565–578, 2015.

[7] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Proc. of IEEE Internationl Conference on Acoustics, Speech and Signal Processing*, 2016, pp. 4960–4964.

[8] J.-T. Chien and C.-L. Kuo, "Bayesian adversarial learning for speaker recognition," in *Proc. of IEEE Automatic Speech Recognition and Understanding Workshop*, 2019, pp. 381–388.

[9] J.-T. Chien and Y.-Y. Lyu, "Partially adversarial learning and adaptation," in *Proc. of European Signal Processing Conference*, 2019, pp. 1444–1448.

[10] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[11] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," in *Proc. of International Conference on Machine Learning*, 2014, pp. 1278–1286.

[12] J. Engel, C. Resnick, A. Roberts, S. Dieleman, M. Norouzi, D. Eck, and K. Simonyan, "Neural audio synthesis of musical notes with WaveNet autoencoders," in *Proc. of International Conference on Machine Learning*, 2017, pp. 1068–1077.

[13] I. Gulrajani, K. Kumar, F. Ahmed, A. A. Taiga, F. Visin, D. Vazquez, and A. Courville, "PixelVAE: A latent variable model for natural images," in *Proc. of International Conference on Learning Representations*, 2017.

[14] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, 2014, pp. 2672–2680.

[15] L. Dinh, D. Krueger, and Y. Bengio, "NICE: Non-linear independent components estimation," *arXiv preprint arXiv:1410.8516*, 2014.

[16] D. J. Rezende and S. Mohamed, "Variational inference with normalizing flows," *arXiv preprint arXiv:1505.05770*, 2015.

[17] L. Dinh, J. Sohl-Dickstein, and S. Bengio, "Density estimation using real NVP," *arXiv preprint arXiv:1605.08803*, 2016.

[18] A. van den Oord, N. Kalchbrenner, L. Espeholt, O. Vinyals, A. Graves, et al., "Conditional image generation with PixelCNN decoders," in *Advances in Neural Information Processing Systems*, 2016, pp. 4790–4798.

[19] D. P. Kingma and P. Dhariwal, "Glow: Generative flow with invertible 1x1 convolutions," in *Advances in Neural Information Processing Systems*, 2018, pp. 10215–10224.

[20] D. P. Kingma, T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever, and M. Welling, "Improved variational inference with inverse autoregressive flow," in *Advances in Neural Information Processing Systems*, 2016, pp. 4743–4751.

[21] R. Shu, H. H. Bui, S. Zhao, M. J. Kochenderfer, and S. Ermon, "Amortized inference regularization," in *Advances in Neural Information Processing Systems*, 2018, pp. 4393–4402.

[22] X. Chen, D. P. Kingma, T. Salimans, Y. Duan, P. Dhariwal, J. Schulman, I. Sutskever, and P. Abbeel, "Variational lossy autoencoder," in *Proc. of International Conference on Learning Representations*, 2017.

[23] E. Mathieu, T. Rainforth, S. Narayanaswamy, and Y. W. Teh, "Disentangling disentanglement," *arXiv preprint arXiv:1812.02833*, 2018.

[24] J.-T. Chien, "Deep Bayesian natural language processing," in *Proc. of Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, 2019, pp. 25–30.

[25] A. G. A. P. Goyal, A. Sordoni, M.-A. Côté, N. R. Ke, and Y. Bengio, "Z-forcing: Training stochastic recurrent networks," in *Advances in Neural Information Processing Systems*, 2017, pp. 6713–6723.

[26] S. R. Bowman, L. Vilnis, O. Vinyals, A. Dai, R. Jozefowicz, and S. Bengio, "Generating sentences from a continuous space," in *Proc. of SIGNLL Conference on Computational Natural Language Learning*, 2016, pp. 10–21.

[27] S. Semeniuta, A. Severyn, and E. Barth, "A hybrid convolutional variational autoencoder for text generation," in *Proc. of Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 627–637.

[28] J.-T. Chien and C.-W. Wang, "Variational and hierarchical recurrent autoencoder," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019, pp. 3202–3206.

[29] J.-T. Chien and C.-W. Wang, "Self attention in variational sequential learning for summarization," *Proc. of Annual Conference of International Speech Communication Association*, pp. 1318–1322, 2019.

[30] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "β-VAE: Learning basic visual concepts with a constrained variational framework," in *Proc. of International Conference on Learning Representations*, 2017.

[31] J. Xu and G. Durrett, "Spherical latent spaces for stable variational autoencoders," in *Proc. of Conference on Empirical Methods in Natural Language Processing*, 2018.

[32] T. R. Davidson, L. Falorsi, N. De Cao, T. Kipf, and J. M. Tomczak, "Hyperspherical variational auto-encoders," *arXiv preprint arXiv:1804.00891*, 2018.

[33] J. M. Tomczak and M. Welling, "VAE with a VampPrior," *arXiv preprint arXiv:1705.07120*, 2017.

[34] S. Watanabe and J.-T. Chien, *Bayesian Speech and Language Processing*, Cambridge University Press, 2015.

[35] C. J. Maddison, J. Lawson, G. Tucker, N. Heess, M. Norouzi, A. Mnih, A. Doucet, and Y. Teh, "Filtering variational objectives," in *Advances in Neural Information Processing Systems*, 2017, pp. 6576–6586.

[36] J. M. Tomczak and M. Welling, "Improving variational auto-encoders using convex combination linear inverse autoregressive flow," *arXiv preprint arXiv:1706.02326*, 2017.

[37] O. Fabius and J. R. van Amersfoort, "Variational recurrent auto-encoders," *arXiv preprint arXiv:1412.6581*, 2014.

[38] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[39] Y. Kim, S. Wiseman, A. Miller, D. Sontag, and A. Rush, "Semi-amortized variational autoencoders," in *Proc. of International Conference on Machine Learning*, 2018, pp. 2683–2692.

[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[41] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Highway networks," *arXiv preprint arXiv:1505.00387*, 2015.

[42] A. B. Dieng, Y. Kim, A. M. Rush, and D. M. Blei, "Avoiding latent variable collapse with generative skip models," *arXiv preprint arXiv:1807.04863*, 2018.

[43] T. Mikolov and G. Zweig, "Context dependent recurrent neural network language model," in *Proc. of IEEE Spoken Language Technology Workshop*, 2012, pp. 234–239.

[44] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in *Proc. of Annual Meeting of Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2011, vol. 1, pp. 142–150.

[45] J.-T. Chien, "Association pattern language modeling," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1719–1728, 2006.

[46] J.-T. Chien and C.-H. Chueh, "Latent Dirichlet language model for speech recognition," in *Proc. of IEEE Spoken Language Technology Workshop*, 2008, pp. 201–204.

[47] J.-T. Chien and Y.-C. Ku, "Bayesian recurrent neural network for language modeling," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 2, pp. 361–374, 2016.

[48] J.-T. Chien and C.-H. Chueh, "Joint acoustic and language modeling for speech recognition," *Speech Communication*, vol. 52, no. 3, pp. 223–235, 2010.

[49] J.-T. Chien, "Hierarchical Pitman-Yor-Dirichlet language model," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 8, pp. 1259–1272, 2015.

[50] L. van der Maaten and G. E. Hinton, "Visualizing data using *t*-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.