VARIATIONAL AND HIERARCHICAL RECURRENT AUTOENCODER

Jen-Tzung Chien *Chun-Wei Wang*

Department of Electrical and Computer Engineering, National Chiao Tung University, Hsinchu, Taiwan

ABSTRACT

Despite a great success in learning representation for image data, it is challenging to learn the stochastic latent features from natural language based on variational inference. The difficulty in stochastic sequential learning is due to the *posterior collapse* caused by an autoregressive decoder which is prone to be too strong to learn sufficient latent information during optimization. To compensate this weakness in learning procedure, a sophisticated latent structure is required to assure good convergence so that random features are sufficiently captured for sequential decoding. This study presents a new variational recurrent autoencoder (VRAE) for sequence reconstruction. There are two complementary encoders consisting of a long short-term memory (LSTM) and a pyramid bidirectional LSTM which are merged to discover the global and local dependencies in a hierarchical latent variable model, respectively. Experiments on Penn Treebank and Yelp 2013 demonstrate that the proposed hierarchical VRAE is able to learn the complementary representation as well as tackle the posterior collapse in stochastic sequential learning. The performance of recurrent autoencoder is substantially improved in terms of perplexity.

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

1. INTRODUCTION

In recent years, deep generative models offering the promising performance for generation of realistic data from unlabeled data have been rapidly developing for different types of technical data including image [1], speech [2, 3] and text [4, 5]. The emerging approaches, including variational autoencoder (VAE) [6], generative adversarial network [7] and autoregressive neural network [8], have achieved remarkable performance in many real-world applications [9]. One of the most successful solutions is the latent variable model based on VAE, which is a stochastic variant of autoencoder (AE) consisting of an encoder as the inference model and a decoder as the generative model. The encoder compresses the input data into a latent representation while the decoder generates synthesized samples from the latent space. The encoder and decoder parameters are jointly learned by maximizing a variational lower bound of log likelihood of training data. Despite a great success, a crucial issue in VAE is the difficulty in learning the complicated latent structure especially in presence of a large-scale set of images or with the highly structured sequential data. Given the abundant natural images for training, VAE tends to generate the blurry images in prediction. In addition, VAE in sequence generation is composed of two recurrent neural networks (RNNs) for both encoder and decoder. In practice, the RNN decoder is trained by teacher forcing where the model receives the ground truth output as input at next time during training. However, this leads to an issue in training phase where a latent loss function as a Kullback-Leibler (KL) divergence vanishes so that the latent variables are not really modeled. This problem is known as the posterior collapse [10] which widely exists in the stochastic RNN [11, 12, 13, 14, 15] where additional latent variables are introduced to represent the hidden states of RNN. In this situation, VAE is specialized as an autoregressive generative model which could not truly learn a stochastic representation. To tackle this issue, one solution [16, 17] was to weaken the capacity of decoder so as to encourage the utilization of latent variables in training procedure. Another solution [10, 18, 19] was to replace the simple Gaussian prior with a sophisticated prior for latent features. In [20], a stochastic variational inference was run to iteratively refine variational parameters. In [21], the skip connections were employed to enforce different dependencies between latent variables and observations.

In contrast to the approach which weakens the decoder, we propose a hierarchical latent variable model for stochastic sequential learning. To cope with the issue of posterior collapse in sequence generation, we strengthen the capability of an encoder by using two different networks which capture the complementary latent representations based on long short-term memory (LSTM) and pyramid bidirectional LSTM. The global and local dependencies in latent structure are characterized by a sophisticated model and sufficiently learned in stochastic generation of sequence data. A set of experiments are reported to illustrate the merit of this hierarchical model.

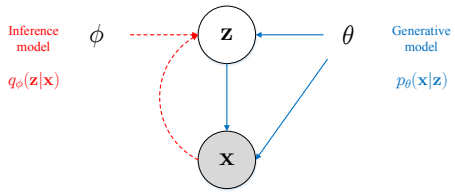
2. BACKGROUND SURVEY

Variational autoencoder (VAE) [6] was proposed to estimate the distribution of latent variable $\mathbf{z}$ and use this information to reconstruct original input signal $\mathbf{x}$. This generative model makes it possible to produce the synthesized signals and an-

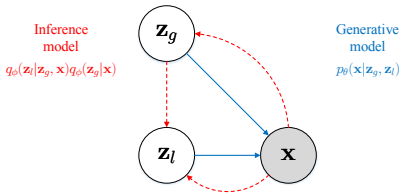
alyze the statistics of latent information in neural networks. The graphical model of VAE is depicted by Figure 1(a) which consists of an inference model for encoding and a generative model for decoding. The encoder $q_\phi(\mathbf{z}|\mathbf{x})$ with parameter ϕ and decoder $p_\theta(\mathbf{x}|\mathbf{z})$ with parameter θ are learned by maximizing a variational lower bound of log likelihood [22]

$$\mathcal{L}(\mathbf{x};\boldsymbol{\theta},\boldsymbol{\phi}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right] - D_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{x})\|p(\mathbf{z})) \tag{1}$$

where the first term reflects the negative reconstruction error due to decoder $p_\theta(\mathbf{x}|\mathbf{z})$ by using the samples $\mathbf{z}$ from encoder $q_\phi(\mathbf{z}|\mathbf{x})$ and the second term is a KL divergence to regularize the variational distribution to match with a standard Gaussian prior $p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$ where $\mathbf{I}$ is an identity matrix.



(a)



(b)

Fig. 1: Graphical representation for (a) VAE and (b) Hierarchical VRAE. Solid lines denote the generative model (decoder) $p_\theta(\mathbf{x}|\mathbf{z})$ or $p_\theta(\mathbf{x}|\mathbf{z}_g,\mathbf{z}_l)$. Dash lines denote the inference model (encoder) $q_\phi(\mathbf{z}|\mathbf{x})$ or $q_\phi(\mathbf{z}_l|\mathbf{x},\mathbf{z}_g)q_\phi(\mathbf{z}_g|\mathbf{x})$.

On the other hand, the RNN-based VAE was proposed to implement the variational recurrent autoenocder (VRAE) for stochastic representation of music and text [23, 24]. This model was composed of two RNNs for both encoder and decoder for reconstruction of a sequence data as depicted in Figure 2. The encoder (green) infers the parameter ϕ of a distribution $q_\phi(\mathbf{z}|\mathbf{x})$ over latent variable $\mathbf{z}$ using an input sequence $\mathbf{x} = \{\mathbf{x}_t\}$ with length T which is a function of final hidden state $\mathbf{h}_T$. The decoder (blue) $p_\theta(\mathbf{x}|\mathbf{z})$ uses the latent vector $\mathbf{z}$ sampled from $q_\phi(\mathbf{z}|\mathbf{x})$ to set the deterministic state $\mathbf{h}_t$ at each time t and accordingly produces the output sequence $\mathbf{y} = \{\mathbf{y}_t\}_{t=1}^T$ for reconstruction of input sequence $\mathbf{x}$.

3. VARIATIONAL AND HIERARCHICAL MODEL

VRAE suffers from the problem of posterior collapse in stochastic sequential learning where the KL term in variational lower bound tends to be vanished, $q_\phi(\mathbf{z}|\mathbf{x}) \approx p(\mathbf{z})$, due

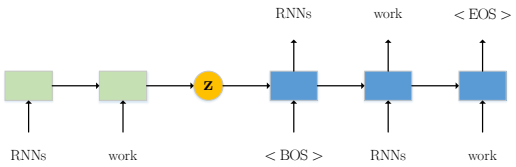


Fig. 2: Illustration for variational recurrent autoencoder.

to an autoregressive decoder for sequence generation. Latent variable $\mathbf{z}$ is then ignored so that the learning procedure likely goes to local optimum. To prevent the vanishing KL divergence, we strengthen the encoder instead of weakening the decoder as performed in [16, 17] and propose the hierarchical latent variable model with two complementary latent variables $\{\mathbf{z}_g, \mathbf{z}_l\}$ which characterize the global and local dependencies of input sequence data $\mathbf{z}$. Physical attributes are interpretable to fulfill a hierarchical VRAE where the graphical representation is depicted in Figure 1(b). First of all, the pyramid bidirectional long short-term memory (BLSTM) is introduced to carry out the variational and hierarchical model.

3.1. Pyramid bidirectional long short-term memory

The pyramid BLSTM (pBLSTM) [25] was proposed to reduce the time resolution as well as capture the complicated features $\mathbf{h}$ in latent space by using BLSTMs with a pyramid structure. A reduced length is resulted in hidden vector $\mathbf{h}_n$ from the original length T of input signal $\mathbf{x}_t$ via a hierarchy of BLSTMs. Using this pBLSTM, the hidden vector at time step n from layer l is computed by concatenating the outputs at consecutive steps with BLSTM at layer $l-1$ and feeding them together into BLSTM at layer l at time step $n-1$ as expressed by

$$\mathbf{h}_n^{(l)} = \mathrm{pBLSTM}(\mathbf{h}_{n-1}^{(l)}, [\mathbf{h}_{2n}^{(l-1)}, \mathbf{h}_{2n+1}^{(l-1)}]) \tag{2}$$

where the time resolution is reduced by a layer with factor 2. Local dependency in input sequence $\mathbf{x}$ can be captured.

3.2. Architecture and optimization

This study presents a hierarchical latent variable model for variational recurrent autoencoder (hereafter called the hierarchical VRAE) where the posterior collapse is handled for stochastic sequential learning. Figure 3 depicts the architecture of the proposed hierarchical VRAE. Each input sequence $\mathbf{x} = \{\mathbf{x}_t\}_{t=1}^T$ is reconstructed by using one global latent variable $\mathbf{z}_g$ and one local latent variable $\mathbf{z}_l$ from encoders (green). The marginal likelihood of sequence data $\mathbf{x}$ is yielded by

$$p(\mathbf{x}) = \int_{\mathbf{z}_g}\int_{\mathbf{z}_l} p_\theta(\mathbf{x}|\mathbf{z}_g,\mathbf{z}_l)p(\mathbf{z}_g)p(\mathbf{z}_l)d\mathbf{z}_l d\mathbf{z}_g. \tag{3}$$

Green and blue boxes mean the LSTM units. For model inference, we first assume Gaussian latent variables similar

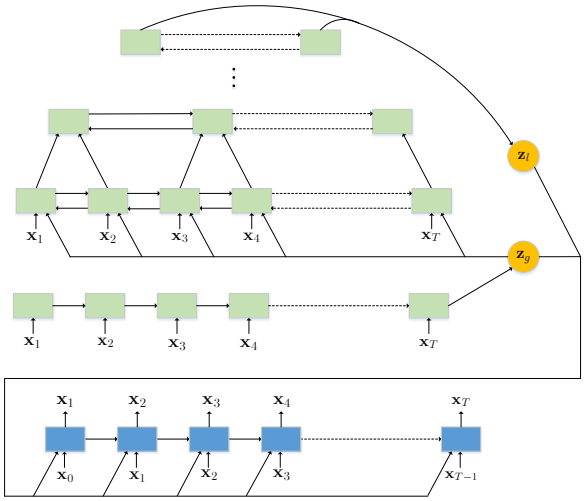


Fig. 3: Illustration for hierarchical VRAE. Input sequence $\{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$ is attended two times to find $\{\mathbf{z}_g, \mathbf{z}_l\}$.

to VAE. The likelihood function is expressed by a *decoder* (blue) or a *generative model* conditional on two latent variables

$$p_\theta(\mathbf{x}|\mathbf{z}_g, \mathbf{z}_l) = \mathcal{N}(\boldsymbol{\mu}_x, \text{diag}(\boldsymbol{\sigma}_x^2)) \tag{4}$$

based on the outputs of a neural network $f_\theta^{\text{dec}}(\cdot)$ which computes the means and variances of a diagonal Gaussian by $[\boldsymbol{\mu}_x, \boldsymbol{\sigma}_x^2] = f_\theta^{\text{dec}}(\mathbf{z}_g, \mathbf{z}_l)$. The zero-mean-unit-variance Gaussian is assumed as the prior for both latent variables $p(\mathbf{z}_g) = p(\mathbf{z}_l) = \mathcal{N}(\mathbf{0}, \mathbf{I})$. Since the true posterior distribution with latent variables $p(\mathbf{z}_g, \mathbf{z}_l|\mathbf{x})$ is intractable, the *encoder* or *inference model* $q_\phi(\mathbf{z}_g, \mathbf{z}_l|\mathbf{x})$ is introduced to approximate $p(\mathbf{z}_g, \mathbf{z}_l|\mathbf{x})$ in variational inference. The variational posterior is expressed by

$$q_\phi(\mathbf{z}_g, \mathbf{z}_l|\mathbf{x}) = q_\phi(\mathbf{z}_l|\mathbf{z}_g, \mathbf{x}) q_\phi(\mathbf{z}_g|\mathbf{x}) \tag{5}$$

where $q_\phi(\mathbf{z}_l|\mathbf{z}_g, \mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}_{\mathbf{z}_l}, \text{diag}(\boldsymbol{\sigma}_{\mathbf{z}_l}^2))$ and $q_\phi(\mathbf{z}_g|\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}_{\mathbf{z}_g}, \text{diag}(\boldsymbol{\sigma}_{\mathbf{z}_g}^2))$ are formed by using the Gaussian parameters which are calculated by neural networks $[\boldsymbol{\mu}_{\mathbf{z}_l}, \boldsymbol{\sigma}_{\mathbf{z}_l}^2] = f_{\phi_l}^{\text{enc}}(\mathbf{x}, \mathbf{z}_g)$ and $[\boldsymbol{\mu}_{\mathbf{z}_g}, \boldsymbol{\sigma}_{\mathbf{z}_g}^2] = f_{\phi_g}^{\text{enc}}(\mathbf{x})$. The global dependency in $\mathbf{x}$ is characterized by the encoder $q_\phi(\mathbf{z}_g|\mathbf{x})$ using a one-layer LSTM [24]. Latent variable $\mathbf{z}_g$ is sampled from the Gaussian distribution with the parameters calculated from the final hidden state $\mathbf{h}_T$ of LSTM based on a fully-connected neural network $f_{\phi_g}^{\text{enc}}$. It is noted that we strengthen the encoder network by incorporating a complementary latent code $\mathbf{z}_l$ which characterizes the local dependency of $\mathbf{x}$ by using a three-layer pyramid bidirectional LSTM. The inputs of pBLSTM are formed by concatenating time signal $\mathbf{x}_t$ with global variable $\mathbf{z}_g$. This pBLSTM is seen as the second encoder $q_\phi(\mathbf{z}_l|\mathbf{z}_g, \mathbf{x})$ to infer the hierarchical and local features $\mathbf{z}_l$. The latent Gaussian parameters are obtained by feeding the outputs of pBLSTM into a fully-connected network $f_{\phi_l}^{\text{enc}}$.

The variational lower bound of the proposed hierarchical VRAE can be derived from Eq. (3) to find

$$\begin{aligned}
&\log p(\mathbf{x}) \\
&\geq \int\int q_\phi(\mathbf{z}_g, \mathbf{z}_l|\mathbf{x}) \log \left(\frac{p_\theta(\mathbf{x}|\mathbf{z}_g, \mathbf{z}_l) p(\mathbf{z}_g) p(\mathbf{z}_l)}{q_\phi(\mathbf{z}_g, \mathbf{z}_l|\mathbf{x})} \right) d\mathbf{z}_g d\mathbf{z}_l \\
&= \mathbb{E}_{q_\phi(\mathbf{z}_g, \mathbf{z}_l|\mathbf{x})} \left[\log \left(\frac{p_\theta(\mathbf{x}|\mathbf{z}_g, \mathbf{z}_l) p(\mathbf{z}_g) p(\mathbf{z}_l)}{q_\phi(\mathbf{z}_l|\mathbf{z}_g, \mathbf{x}) q_\phi(\mathbf{z}_g|\mathbf{x})} \right) \right] \\
&= \mathbb{E}_{q_\phi(\mathbf{z}_g, \mathbf{z}_l|\mathbf{x})} \left[\log p_\theta(\mathbf{x}|\mathbf{z}_g, \mathbf{z}_l) \right] - D_{\text{KL}} \left(q_\phi(\mathbf{z}_g|\mathbf{x}) || p(\mathbf{z}_g) \right) \\
&\quad - \mathbb{E}_{q_\phi(\mathbf{z}_g|\mathbf{x})} \left[D_{\text{KL}} \left(q_\phi(\mathbf{z}_l|\mathbf{z}_g, \mathbf{x}) || p(\mathbf{z}_l) \right) \right] \triangleq \mathcal{L}(\mathbf{x}; \boldsymbol{\theta}, \boldsymbol{\phi})
\end{aligned} \tag{6}$$

where the first term reflects the reconstruction error, the remaining two terms denote the KL divergence obtained from latent variables $\{\mathbf{z}_g, \mathbf{z}_l\}$. The vanishing problem of KL divergence is mitigated due to the additional KL term from latent code $\mathbf{z}_l$ learned by pBLSTM. This work adopts the stochastic gradient variational Bayes estimator [6] to learn decoder $\boldsymbol{\theta}$ and encoder parameters $\boldsymbol{\phi}$ by maximizing $\mathcal{L}(\mathbf{x}; \boldsymbol{\theta}, \boldsymbol{\phi})$.

3.3. Discussion and comparison

Basically, VRAE is viewed as a variational version of autoencoder for sequence data with two regularizations. One is to consider the *stochastic* reconstruction error by integrating over the variational distribution $q_\phi(\mathbf{z}|\mathbf{x})$ while the other is to impose a normalization constraint via a KL term which encourages $q_\phi(\mathbf{z}|\mathbf{x})$ close to $\mathcal{N}(\mathbf{0}, \mathbf{I})$. KL term is critical in variational model. In early learning stage VRAE, the variational posterior $q_\phi(\mathbf{z}|\mathbf{x})$ conveys little information about input signal $\mathbf{x}$ so that VRAE tends to learn the variational posterior to be the prior $p(\mathbf{z})$. Recent researches [24, 26, 27] illustrate that VRAE or VAE with autoregressive decoder leads KL divergence to become vanished. Latent variable $\mathbf{z}$ is accordingly ignored and then could not bring stochastic information in the trained model. Comparatively, the hierarchical VRAE encodes input sequence $\mathbf{x}$ two times. This model likely infers the interpretable and complementary latent variables $\mathbf{z}_g$ and $\mathbf{z}_l$ which bring two KL terms in Eq. (6) to be nonzero and therefore stimulate the stochastic and sequential learning. In case that the posterior collapse happens to vanish KL term of $\mathbf{z}_g$, our model is reduced to a standard VRAE where the effect of $\mathbf{z}_g$ is neglected and the objective $\mathcal{L}(\mathbf{x}; \boldsymbol{\theta}, \boldsymbol{\phi})$ is simplified as $\mathbb{E}_{q_\phi(\mathbf{z}_l|\mathbf{x})} \left[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}_l) \right] - D_{\text{KL}} \left(q_\phi(\mathbf{z}_l|\mathbf{x}) || p_\theta(\mathbf{z}_l) \right)$. Namely, due to the incorporation of two latent variables, either $\mathbf{z}_l$ or $\mathbf{z}_g$ will be sufficiently inferred. As a result, the issue of posterior collapse is mitigated in the proposed hierarchical VRAE.

4. EXPERIMENTS

In the experiments, different methods were evaluated by using two datasets: Penn TreeBank (PTB) [28] and Yelp 2013 (Yelp) [19]. PTB was a benchmark dataset for language modeling [29, 30, 31]. Yelp was a review dataset collected from

Yelp Dataset Challenge in year 2013. The averaged length in a sentence was 21.1 and 47.6 words and vocabulary size was 10K and 15K in PTB and Yelp datasets, respectively.

mr. wathen who says pinkerton's had a loss of nearly $ N million in N under american brands boasts that he's made pinkerton 's profitable again
mr. <unk> said he was pleased with his estimate of N N in N and N N in N after mr. <unk>'s departure
in addition the company's <unk> business is n't being acquired by <unk>'s stock market share
in the past two months mr. <unk> said he expects to report a loss of $ N million
in the first nine months of N shares of N N and a nominal N N
the dow jones industrial average fell N points to N
in when-issued trading the notes were quoted at a price to yield N N

(a) Interpolate $\mathbf{z}_g$ and $\mathbf{z}_l$

mr. wathen who says pinkerton's had a loss of nearly $ N million in N under american brands boasts that he's made pinkerton 's profitable again
mr. <unk> said he was pleased with his estimate of N N in N and N N in N after mr. <unk>'s departure
mr. <unk> said he expects the company's earnings growth in N and N N of its common stock outstanding
in addition to the new york stock exchange yesterday the company's $ N <unk> had been sold at a share up $ N million
in the past two months the company said it expects to report a loss of $ N million or $ N a share
in the first nine months the company said it expects to report a loss of $ N million or N cents a share
in when-issued trading the notes were quoted at a price to yield N N

(b) Interpolate $\mathbf{z}_g$ but fix $\mathbf{z}_l$

Fig. 4: Linear interpolation of two sentences in top and bottom rows by using their latent variables. Synthesized sentences are shown in middle rows by interpolating (a) both $\mathbf{z}_g$ and $\mathbf{z}_l$ and (b) only $\mathbf{z}_g$. <unk> means an unknown word.

4.1. Experimental setup

Three methods were implemented and evaluated for language modeling. Baseline system was built by the LSTM language model (denoted by RNNLM) where only a single LSTM was applied for sentence reconstruction via prediction of next word at each time step. VRAE and hierarchical VRAE were carried out with encoder and decoder. All models adopted one-layer LSTM as both encoder and decoder with an embedding of size 300 and hidden units of size 256. The hierarchical VRAE additionally employed a *three-layer* pBLSTM with 256 hidden units. The dimension of hidden codes $\{\mathbf{z}_g, \mathbf{z}_l\}$ was fixed as 16. The batch size of 32 was used. All models were trained by 20 epochs. All of these models were optimized by using Adam optimizer with initial learning rate of 0.001 which was decreased by a factor of 2 every 2 epochs after epoch 10. There was a dropout layer with probability 0.5 in the input-to-hidden layer in LSTM decoder. Gradient clipping was applied with maximum norm 5. Following [24], the KL-cost annealing strategy was utilized. The initial weight of KL term was set to be 0.01 and was increased linearly to 1 over the first 10 epochs.

4.2. Experimental results

To demonstrate the latent semantics by using the hierarchical VRAE, we interpolate latent variables $\mathbf{z}_g$ and $\mathbf{z}_l$ of two sentences and use them to generate a new sentence for investigation. Figure 4 shows the synthesized sentences due to the effects of global variable $\mathbf{z}_g$ and local variable $\mathbf{z}_l$. Obviously, the global latent variable $\mathbf{z}_g$ dominates the semantics of the synthesized sentences. Next, Tables 1 and 2 compare the negative log-likelihood (NLL) and the perplexity (PPL) of test sentences using different models where PTB and Yelp are examined, respectively. In this comparison, we evaluate how different neural models perform for sequence generation. RNNLM is run via a single LSTM and does not involve stochastic learning. VRAE performs RNN-based VAE with one LSTM for encoder and the other LSTM for decoder. The best results or the lowest NLL and PPL among all models are obtained by using hierarchical VRAE. In addition to NLL and PPL, we find that larger KL divergence produces better performance. KL values due to $\mathbf{z}_g$ and $\mathbf{z}_l$ are shown. Posterior collapse less likely happen so as to learn meaningful latent representation for prediction. Hierarchical VRAE performs better than RNNLM and VRAE for sentence generation. Source codes are accessible at https://github.com/NCTUMLlab/

Model	NLL	KL ($\mathbf{z}_g$, $\mathbf{z}_l$)	PPL
RNNLM	102.27	-	132.89
VRAE	101.45	4.86	127.78
Hierarchical VRAE	**99.28**	**7.25** (4.40, 2.85)	**115.17**

Table 1: Comparison of different methods under PTB dataset.

Model	NLL	KL ($\mathbf{z}_g$, $\mathbf{z}_l$)	PPL
RNNLM	196.69	-	62.91
VRAE	196.28	2.25	62.38
Hierarchical VRAE	**192.25**	**6.44** (4.66, 1.78)	**57.30**

Table 2: Comparison of different methods under Yelp dataset.

5. CONCLUSION

We proposed a variational and hierarchical recurrent autoencoder for stochastic and sequential learning and applied it for sentence reconstruction. This model employed a LSTM and a pyramid bidirectional LSTM as the encoders to characterize global and local latent variables, respectively. Two encoders were merged to capture the complementary latent features based on two passes of encoding and reasoning. Experimental results show that the proposed method mitigated the issue of posterior collapse and improved the prediction performance for sentence generation in terms of perplexity. This model learned the meaningful latent representations based on the global and local dependencies in natural language.

6. REFERENCES

[1] 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.

[2] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[3] 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.

[4] 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.

[5] 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.

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

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

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

[9] J.-T. Chien, "Deep Bayesian learning and understanding," in *Proc. of International Conference on Computational Linguistics: Tutorial Abstracts*, 2018, pp. 13–18.

[10] A. van den Oord, O. Vinyals, et al., "Neural discrete representation learning," in *Proc. of Advances in Neural Information Processing Systems*, 2017, pp. 6306–6315.

[11] A. Goyal, A. Sordoni, M.-A. Côté, N. Ke, and Y. Bengio, "Z-forcing: Training stochastic recurrent networks," in *Proc. of Advances in Neural Information Processing Systems*, 2017, pp. 6716–6726.

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

[13] J.-T. Chien and K.-T. Kuo, "Variational recurrent neural networks for speech separation," in *Proc. Annual Conference of International Speech Communication Association*, 2017, pp. 1193–1197.

[14] J.-T. Chien and C. Shen, "Stochastic recurrent neural network for speech recognition," in *Proc. of Annual Conference of International Speech Communication Association*, 2017, pp. 1313–1317.

[15] J.-T. Chien and C.-Y. Kuo, "Stochastic markov recurrent neural network for source separation," in *Proc. of IEEE Internationl Conference on Acoustics, Speech and Signal Processing*, 2019.

[16] 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.

[17] Z. Yang, Z. Hu, R. Salakhutdinov, and T. Berg-Kirkpatrick, "Improved variational autoencoders for text modeling using dilated convolutions," in *Proc. of International Conference on Machine Learning*, 2017, pp. 3881–3890.

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

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

[20] 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.

[21] 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.

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

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

[24] 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.

[25] 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.

[26] C. K. Sønderby, T. Raiko, L. Maaløe, S. K. Sønderby, and O. Winther, "Ladder variational autoencoders," in *Proc. of Advances in Neural Information Processing Systems*, 2016, pp. 3738–3746.

[27] 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.

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

[29] 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.

[30] C.-Y. Kuo and J.-T. Chien, "Markov recurrent neural networks," in *Proc. of IEEE International Workshop on Machine Learning for Signal Processing*, 2018, pp. 1–6.

[31] 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.