SEMI-IMPLICIT STOCHASTIC RECURRENT NEURAL NETWORKS

Ehsan Hajiramezanali [†*], *Arman Hasanzadeh* [†*],
Nick Duffield [†], *Krishna Narayanan* [†], *Mingyuan Zhou* [‡], *Xiaoning Qian* [†]

[†] Texas A&M University, [‡] University of Texas at Austin
{*ehsanr, armanihm, duffieldng, krn, xqian*}*@tamu.edu, mingyuan.zhou@mccombs.utexas.edu*

ABSTRACT

Stochastic recurrent neural networks with latent random variables of complex dependency structures have shown to be more successful in modeling sequential data than deterministic deep models. However, the majority of existing methods have limited expressive power due to the Gaussian assumption of latent variables. In this paper, we advocate learning implicit latent representations using semi-implicit variational inference to further increase model flexibility. Semi-implicit stochastic recurrent neural network (SIS-RNN) is developed to enrich inferred model posteriors that may have no analytic density functions, as long as independent random samples can be generated via reparameterization. Extensive experiments in different tasks on real-world datasets show that SIS-RNN outperforms the existing methods.

***Index Terms*—** Semi-implicit variational inference, Variational auto-encoder (VAE), Recurrent neural network (RNN), Natural language processing (NLP).

1. INTRODUCTION

Deep auto-regressive models, such as recurrent neural networks (RNNs), are widely used for modeling sequential data due to their effective representation of long-term dependencies. It has been shown that inducing uncertainty in hidden states of deep auto-regressive models could drastically improve their performance in many applications such as speech modeling, text generation, sequential image modeling and dynamic graph representation learning [1, 2, 3, 4, 5, 6]. These methods integrate the variational auto-encoder (VAE) framework with deep auto-regressive models to infer stochastic latent variables, which can capture higher-level semantic abstraction (e.g. objects, speakers, or graph modules/communities) from the observed variables in a sequence (e.g. pixels, sound-waves, or partially observed dynamic graphs).

Existing stochastic recurrent models, while having different encoder and decoder structures, have restricted expressive power due to the commonly adopted Gaussian assumption on prior and posterior distributions of latent variables. The Gaussian assumption has a well-known issue in underestimating the variance of the posterior [7], which can be further amplified by mean field variational inference (MFVI). This issue is often attributed to two key factors: 1) the mismatch between the restricted representation power of the variational family Q and the complexity of the posterior to be approximated by Q; 2) the use of $\mathbf{KL}$ divergence, which is an asymmetric measure for the distance between Q and the posterior [8, 9, 10].

In this paper, we break the Gaussian assumption and propose a semi-implicit stochastic recurrent neural network (SIS-RNN) that is capable of inferring implicit posteriors for sequential data while maintaining simple optimization. Inspired by semi-implicit variational inference (SIVI) [8], we impose a semi-implicit hierarchical construction on a backbone RNN to represent the posterior distribution of stochastic recurrent layers. SIVI enables a flexible (implicit) mixing distribution for variational inference of our proposed SIS-RNN. As a result, even if the marginal of the hierarchy is not tractable, its density can be evaluated by Monte Carlo estimation. Our proposed framework is capable of modeling skewness, kurtosis, multimodality, and other characteristics that are exhibited by the posterior of latent variables but fail to be captured by the mean-field Gaussian variational family. Our experiments demonstrate the superior performance of our proposed model in sequential image modeling and language modeling on multiple real-world datasets.

2. PRELIMINARIES

2.1. Semi-implicit variational inference (SIVI)

SIVI has been proposed by [8] as a method for inferring implicit posteriors while maintaining simple optimization. SIVI assumes that the parameters of the posterior, ψ, are drawn from an implicit distribution instead of taking deterministic values. This hierarchical construction enables flexible mixture modeling and allows to have richer variational posteriors. More specifically, let $\mathbf{Z} \sim q(\mathbf{Z} \,|\, \psi)$ and $\psi \sim q_\phi(\psi)$, with ϕ denoting the distribution parameters to be inferred, and $q(\mathbf{Z} \,|\, \psi)$ be the posterior distribution. Marginalizing ψ out leads to the random variables $\mathbf{Z}$ drawn from a distribution family $\boldsymbol{\mathcal{H}}$ indexed by variational parameters ϕ, expressed as

$$\boldsymbol{\mathcal{H}} = \left\{ h_{\boldsymbol{\phi}}(\mathbf{Z}) \;:\; h_{\boldsymbol{\phi}}(\mathbf{Z}) = \int_{\psi} q(\mathbf{Z} \,|\, \psi) q_{\boldsymbol{\phi}}(\psi)\, d\psi \right\}. \quad (1)$$

* The first two authors contributed equally.

The essence of the semi-implicit formulation is that while the conditional posterior $q(\mathbf{Z} \mid \psi)$ is explicit and analytic, the marginal distribution, $h_{\boldsymbol{\phi}}(\mathbf{Z})$ is often implicit. Note that, if $q_{\boldsymbol{\phi}}$ equals a delta function, then $h_{\boldsymbol{\phi}}$ is an explicit distribution. Unlike regular variational inference that assumes independent latent dimensions, SIVI does not impose such a constraint. This enables the resulting variational distributions to model very complex multivariate distributions such as multimodal or skewed distributions, which can not be captured by vanilla variational inference due to its often restricted exponential family assumption over both prior and posterior.

2.2. Variational recurrent neural network (VRNN)

VRNN [11] combines VAE with RNN to increase the expressive power of RNN and better model variability observed in highly structured sequential data. In VRNN, in addition to hidden states of RNN, a latent random variable is used to summarize past information. More specifically, given observations $\mathbf{x}_{\leq t}$ and the stochastic variables $\mathbf{z}_{\leq t}$, model likelihood $p(\mathbf{x}_t|\mathbf{z}_{\leq t}, \mathbf{x}_{<t})$, and prior $p(\mathbf{z}_t|\mathbf{z}_{<t}, \mathbf{x}_{<t})$, we approximate the posterior $p(\mathbf{z}_t|\mathbf{x}_{\leq t}, \mathbf{z}_{<t})$ with a variational distribution $q(\mathbf{z}_t|\boldsymbol{\psi}_t)$ that is required to be explicit. We learn the variational parameters by minimizing $\mathbf{KL}(q(\mathbf{z}_t|\boldsymbol{\psi}_t)||p(\mathbf{z}_t|\mathbf{x}_{\leq t}, \mathbf{z}_{<t}))$, the $\mathbf{KL}$ divergence of $p(\mathbf{z}_t|\mathbf{x}_{\leq t}, \mathbf{z}_{<t})$ and $q(\mathbf{z}_t|\boldsymbol{\psi}_t)$. Knowing that

$$\log p(\mathbf{x}_{\leq T}) = \mathbf{ELBO} + \sum_{t=1}^{T} \mathbf{KL}(q(\mathbf{z}_t|\boldsymbol{\psi}_t)||p(\mathbf{z}_t|\mathbf{x}_{\leq t}, \mathbf{z}_{<t})),$$

with $\mathbf{ELBO} =$

$$-\mathbb{E}_{q(\mathbf{z}_{\leq T}|\mathbf{x}_{\leq T})}\left[\sum_{t=1}^{T}(\log q(\mathbf{z}_t|\boldsymbol{\psi}_t) - \log p(\mathbf{x}_{\leq t}, \mathbf{z}_{\leq t}))\right], \quad (2)$$

minimizing $\mathbf{KL}(q(\mathbf{z}_t|\boldsymbol{\psi}_t) \,||\, p(\mathbf{z}_t|\mathbf{x}_{\leq t}, \mathbf{z}_{<t}))$ is hence equivalent to maximizing the $\mathbf{ELBO}$ [11]. Note that past information, i.e. $\mathbf{x}_{\leq t}$, is transformed through RNN hidden states $\mathbf{h}_t$ as detailed below.

3. SIS-RNN

We introduce our SIS-RNN that imposes a distribution over parameters of posterior in VRNN, i.e. $\boldsymbol{\psi}_t \sim q(\boldsymbol{\psi}_t)$, instead of simply taking deterministic parameters.

Model construction. Assuming $\boldsymbol{\psi}_t \sim q_{\boldsymbol{\phi}}(\boldsymbol{\psi}_t|\mathbf{x}_{\leq t}, \mathbf{z}_{<t})$, where ϕ denotes the parameters of the distribution to be inferred, the semi-implicit variational distribution for $\mathbf{z}_t$ can be defined in a hierarchical manner as

$$\mathbf{z}_t \sim q(\mathbf{z}_t|\boldsymbol{\psi}_t), \qquad \boldsymbol{\psi}_t \sim q_{\boldsymbol{\phi}}(\boldsymbol{\psi}_t|\mathbf{x}_{\leq t}, \mathbf{z}_{<t}). \quad (3)$$

We impose an auto-regressive model to capture long-term dependency in the mixing distribution by exploiting an RNN architecture, that runs through the sequence as follows:

$$\mathbf{h}_t = f_\theta(\mathbf{x}_t, \mathbf{z}_t, \mathbf{h}_{t-1}), \quad (4)$$

where f is a deterministic non-linear transition function, and θ is the parameter set of f to infer. By coupling the observations and latent variables using the recurrence equation (4), SIS-RNN in (3) can be equivalently expressed as

$$\mathbf{z}_t \sim q(\mathbf{z}_t|\boldsymbol{\psi}_t), \qquad \boldsymbol{\psi}_t \sim q_{\boldsymbol{\phi}}(\boldsymbol{\psi}_t|\mathbf{x}_t, \mathbf{h}_{t-1}). \quad (5)$$

Marginalizing $\boldsymbol{\psi}_t$ out leads to the random variables $\mathbf{z}_t$ drawn from the distribution family $\mathcal{G}$ indexed by variational parameters ϕ, expressed as

$$\mathcal{G} = \left\{ g_{\boldsymbol{\phi}}(\mathbf{z}_t|\mathbf{x}_t, \mathbf{h}_{t-1}) = \int_{\boldsymbol{\psi}_t} q(\mathbf{z}_t|\boldsymbol{\psi}_t) q_{\boldsymbol{\phi}}(\boldsymbol{\psi}_t|\mathbf{x}_t, \mathbf{h}_{t-1}) d\boldsymbol{\psi}_t \right\} \quad (6)$$

While the variational distribution $q(\mathbf{z}_t|\boldsymbol{\psi}_t)$ is required to be explicit, there is no such a constraint on the mixing distribution $q_{\boldsymbol{\phi}}(\boldsymbol{\psi}_t|\mathbf{x}_t, \mathbf{h}_{t-1})$ and it is only required to be reparameterizable. In addition, $q(\mathbf{z}_t|\boldsymbol{\psi}_t)$ can be reparameterizable, with $\mathbf{z}_t \sim q(\mathbf{z}_t|\boldsymbol{\psi}_t)$ being generated by transforming random noise ϵ via $f(\epsilon, \boldsymbol{\psi}_t)$ or allowing the $\mathbf{ELBO}$ in (2) to be analytic. More specifically, SIS-RNN draws samples from $q_{\boldsymbol{\phi}}(\boldsymbol{\psi}_t|\mathbf{x}_t, \mathbf{h}_{t-1})$ by transforming random noise ϵ_t via a deep neural network. Specifically, assuming that conditional posterior is Gaussian, then, $q(\mathbf{z}_t|\boldsymbol{\psi}_t) \sim \mathcal{N}\left(\boldsymbol{\mu}^{(t)}_{\text{encoder}}, \text{diag}((\boldsymbol{\sigma}^{(t)}_{\text{encoder}})^2)\right)$, with $\left\{\boldsymbol{\mu}^{(t)}_{\text{encoder}}, \boldsymbol{\sigma}^{(t)}_{\text{encoder}}\right\} = \varphi^{\text{encoder}}(\mathbf{z}_t, \mathbf{h}_{t-1}, \epsilon_t)$ where φ^{encoder} is a neural network. This generally leads to an implicit distribution for $q_{\boldsymbol{\phi}}(\boldsymbol{\psi}_t|\mathbf{x}_t, \mathbf{h}_{t-1})$ due to a non-invertible transform φ^{encoder}. Therefore, the marginal variational distribution $g_{\boldsymbol{\phi}}(\mathbf{z}_t|\mathbf{x}_t, \mathbf{h}_{t-1}) \in \mathcal{G}$ is often implicit, unless $q_{\boldsymbol{\phi}}(\boldsymbol{\psi}_t|\mathbf{x}_t, \mathbf{h}_{t-1})$ is conjugate to $q(\mathbf{z}_t|\boldsymbol{\psi}_t)$.

Note that if $q_{\boldsymbol{\phi}}(\boldsymbol{\psi}_t|\mathbf{x}_t, \mathbf{h}_{t-1})$ degenerates to the delta function $\delta_{\boldsymbol{\psi}^0_t}(\boldsymbol{\psi}_t|\mathbf{x}_t, \mathbf{h}_{t-1})$, the semi-implicit variational family $\mathcal{G}$ reduces to the original $\mathcal{Q} = q(\mathbf{z}_t|\boldsymbol{\psi}^0_t)$ family, where $\mathcal{Q} \subseteq \mathcal{G}$, as discussed in [11]. Unlike MFVI that assumes independent latent variables $z^{(l)}_t$, this expansion significantly helps restore the dependencies between them if $\psi^{(l)}_t$ are not imposed to be independent of each other. Under this construction, the temporal variational distribution can be factorized as

$$q(\mathbf{z}_{\leq T}|\mathbf{x}_{\leq T}) = \prod_{t=1}^{T} g_{\boldsymbol{\phi}}(\mathbf{z}_t|\mathbf{x}_{\leq t}, \mathbf{z}_{<t}). \quad (7)$$

Instead of imposing a standard multivariate Gaussian distribution with deterministic parameters, VAE in our SIS-RNN learns the prior distribution parameters based on the hidden states in previous time steps. In particular, we can write the construction of the prior distribution adopted as follows,

$$\mathbf{z}_t \sim \mathcal{N}\left(\boldsymbol{\mu}^{(t)}_{\text{prior}}, \text{diag}((\boldsymbol{\sigma}^{(t)}_{\text{prior}})^2)\right), \quad (8)$$

where $\left\{\boldsymbol{\mu}^{(t)}_{\text{prior}}, \boldsymbol{\sigma}^{(t)}_{\text{prior}}\right\} = \varphi^{\text{prior}}(\mathbf{h}_{t-1})$ denote the parameters of the conditional prior distribution. Therefore, the generative

model can be factorized as

$$p(\mathbf{x}_{\leq T}, \mathbf{z}_{\leq T}) = \prod_{t=1}^{T} p(\mathbf{x}_t|\mathbf{z}_{\leq t}, \mathbf{x}_{<t}) p(\mathbf{z}_t|\mathbf{z}_{<t}, \mathbf{x}_{<t}) = \prod_{t=1}^{T} p(\mathbf{x}_t|\mathbf{z}_t, \mathbf{h}_{t-1}) p(\mathbf{z}_t|\mathbf{h}_{t-1}), \tag{9}$$

where the parameters of the generating distribution $p(\mathbf{x}_t|\mathbf{z}_t, \mathbf{h}_{t-1})$ can be learned using neural networks φ^{decoder}.

Learning. Since the parameters of the posterior are random variables, the ELBO goes beyond the simple VRNN and using equations (7) and (9), **ELBO** can be derived as follows:

$$\mathcal{L} = \sum_{t=1}^{T} \Big\{ \mathbb{E}_{\boldsymbol{\psi}_t \sim q_{\boldsymbol{\phi}}(\boldsymbol{\psi}_t|\mathbf{x}_t, \mathbf{h}_{t-1})} \mathbb{E}_{\mathbf{z}_t \sim q(\mathbf{z}_t|\boldsymbol{\psi}_t)} \log p(\mathbf{x}_t|\mathbf{z}_t, \mathbf{h}_{t-1}) - \mathbf{KL}\left(\mathbb{E}_{\boldsymbol{\psi}_t \sim q_{\boldsymbol{\phi}}(\boldsymbol{\psi}_t|\mathbf{x}_t, \mathbf{h}_{t-1})} q(\mathbf{z}_t|\boldsymbol{\psi}_t) \,||\, p(\mathbf{z}_t|\mathbf{h}_{t-1}) \right) \Big\}. \tag{10}$$

Direct optimization of the ELBO is not tractable [8, 9, 2, 12]. Hence to infer variational parameters of SI-VGRNN, we derive a lower bound for the ELBO as follows:

$$\begin{aligned}
\mathcal{L} &= \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z}_{\leq T}|\mathbf{x}_{\leq T})} \left[\log p(\mathbf{x}_{\leq T}, \mathbf{z}_{\leq T}) - \log q(\mathbf{z}_{\leq T}|\mathbf{x}_{\leq T}) \right] \\
&= \sum_{t=1}^{T} \mathbb{E}_{\mathbf{z}_t \sim g_{\boldsymbol{\phi}}(\mathbf{z}_t|\mathbf{x}_t, \mathbf{h}_{t-1})} \log \frac{p(\mathbf{x}_t|\mathbf{z}_t, \mathbf{h}_{t-1}) p(\mathbf{z}_t|\mathbf{h}_{t-1})}{g_{\boldsymbol{\phi}}(\mathbf{z}_t|\mathbf{x}_t, \mathbf{h}_{t-1})} \\
&= -\sum_{t=1}^{T} \Big\{ \mathbf{KL}(\mathbb{E}_{\boldsymbol{\psi}_t \sim q_{\boldsymbol{\phi}}(\boldsymbol{\psi}_t|\mathbf{x}_t, \mathbf{h}_{t-1})} q(\mathbf{z}_t|\boldsymbol{\psi}_t) \,||\, p(\mathbf{z}_t|\mathbf{h}_{t-1})) \\
&\quad + \mathbb{E}_{\boldsymbol{\psi}_t \sim q_{\boldsymbol{\phi}}(\boldsymbol{\psi}_t|\mathbf{x}_t, \mathbf{h}_{t-1})} \mathbb{E}_{\mathbf{z}_t \sim q(\mathbf{z}_t|\boldsymbol{\psi}_t)} \log p(\mathbf{x}_t|\mathbf{z}_t, \mathbf{h}_{t-1}) \Big\} \\
&\geq -\sum_{t=1}^{T} \Big\{ \mathbb{E}_{\boldsymbol{\psi}_t \sim q_{\boldsymbol{\phi}}(\boldsymbol{\psi}_t|\mathbf{x}_t, \mathbf{h}_{t-1})} \mathbf{KL}(q(\mathbf{z}_t|\boldsymbol{\psi}_t) \,||\, p(\mathbf{z}_t|\mathbf{h}_{t-1})) \\
&\quad + \mathbb{E}_{\boldsymbol{\psi}_t \sim q_{\boldsymbol{\phi}}(\boldsymbol{\psi}_t|\mathbf{x}_t, \mathbf{h}_{t-1})} \mathbb{E}_{\mathbf{z}_t \sim q(\mathbf{z}_t|\boldsymbol{\psi}_t)} \log p(\mathbf{x}_t|\mathbf{z}_t, \mathbf{h}_{t-1}) \Big\} \\
&= \sum_{t=1}^{T} \mathbb{E}_{\boldsymbol{\psi}_t \sim q_{\boldsymbol{\phi}}(\boldsymbol{\psi}_t|\mathbf{x}_t, \mathbf{h}_{t-1})} \mathbb{E}_{\mathbf{z}_t \sim q(\mathbf{z}_t|\boldsymbol{\psi}_t)} \\
&\qquad \log \left(\frac{p(\mathbf{x}_t|\mathbf{z}_t, \mathbf{h}_{t-1}) p(\mathbf{z}_t|\mathbf{h}_{t-1})}{q(\mathbf{z}_t|\boldsymbol{\psi}_t)} \right) = \underline{\mathcal{L}}.
\end{aligned} \tag{11}$$

Note that we used the following inequality from [8] to derive $\underline{\mathcal{L}}$, $\mathbb{E}_{\boldsymbol{\psi}_t} \mathbf{KL}(q(\mathbf{z}_t|\boldsymbol{\psi}_t)||p(\mathbf{z}_t) \geq \mathbf{KL}(\mathbb{E}_{\boldsymbol{\psi}_t} q(\mathbf{z}_t|\boldsymbol{\psi}_t)||p(\mathbf{z}_t))$.

While Monte Carlo estimation of $\underline{\mathcal{L}}$ only requires $q_{\boldsymbol{\phi}}(\mathbf{z}_t|\boldsymbol{\psi}_t)$ to have an analytic density function and $q_{\boldsymbol{\phi}}(\boldsymbol{\psi}_t|\mathbf{x}_t, \mathbf{h}_{t-1})$ to be convenient to sample from, $g_{\boldsymbol{\phi}}(\mathbf{z}_t|\mathbf{x}_t, \mathbf{h}_{t-1})$ is often intractable, and so the Monte Carlo estimation of the ELBO $\mathcal{L}$ is prohibited. Therefore, SIS-RNN evaluates the lower bound separately from the distribution sampling. While the combination of an explicit $q_{\boldsymbol{\phi}}(\mathbf{z}_t|\boldsymbol{\psi}_t)$ with an implicit $q_{\boldsymbol{\phi}}(\boldsymbol{\psi}_t|\mathbf{x}_t, \mathbf{h}_{t-1})$ is as powerful as needed, it is computationally tractable.

As discussed in [8], without early stopping optimization, $q_{\boldsymbol{\phi}}(\boldsymbol{\psi}_t|\mathbf{x}_t, \mathbf{h}_{t-1})$ can converge to a point mass density, making SIS-RNN degenerated to vanilla VRNN. To avoid this problem, we impose a regularization term to the lower bound $\underline{\mathcal{L}}_K = \underline{\mathcal{L}} + B_K$ as inspired by SIVI [8]:

$$B_K = \sum_{t=1}^{T} \mathbb{E}_{\boldsymbol{\psi}_t, \boldsymbol{\psi}_t^{(1)}, \ldots, \boldsymbol{\psi}_t^{(K)} \sim q_{\boldsymbol{\phi}}(\boldsymbol{\psi}_t|\mathbf{x}_t, \mathbf{h}_{t-1})} \mathbf{KL}(q(\mathbf{z}_t|\boldsymbol{\psi}_t) \,||\, \tilde{g}_K(\mathbf{z}_t|\mathbf{x}_t, \mathbf{h}_{t-1})),$$

where $\tilde{g}_K(\mathbf{z}_t|\mathbf{x}_t, \mathbf{h}_{t-1})) =$

$$\frac{q_{\boldsymbol{\phi}}(\boldsymbol{\psi}_t|\mathbf{x}_t, \mathbf{h}_{t-1}) + \sum_{k=1}^{K} q_{\boldsymbol{\phi}}(\boldsymbol{\psi}_t^{(k)}|\mathbf{x}_t, \mathbf{h}_{t-1})}{K+1}. \tag{12}$$

This leads to an asymptotically exact ELBO that satisfies $\underline{\mathcal{L}}_0 = \underline{\mathcal{L}}$ and $\lim_{K \to \infty} \underline{\mathcal{L}}_K = \mathcal{L}$.

4. EXPERIMENTS

Sequential MNIST. We first evaluate the performance of SIS-RNN in the task of sequentially generating pixels in MNIST digits, which is a common benchmarking test in evaluating sequence modeling methods. We consider the binarized MNIST dataset as in [13]. Following the previous works [1], we used 60,000 samples for training and 10,000 for testing. A Gated Recurrent Unit (GRU) with one layer of 64 hidden units was the backbone RNN in SIS-RNN. Two 64-dimensional fully-connected layers were adopted to model φ^{prior} in (8). We used a neural network with three 128-dimensional fully-connected layers as φ^{encoder} while injecting [150, 100, 50] dimensional Bernoulli noise. The model was trained for 2000 epochs using the Adam optimizer with a 0.001 learning rate at the mini-batch size of 128. K in (12) gradually increased from 1 to 100 during the first 500 epochs and remained constant after that. We used cyclic annealing [14] as the KL annealing in **ELBO** to gradually impose the prior regularization term and avoid posterior collapse. The performance of SIS-RNN and the comparison with other methods are provided in Table 1. We report exact negative log-likelihood (NLL), approximate NLL (with $\approx$ sign), or the variational lower bound (with $\leq$ sign) based on the competing methods. While 64 hidden units were chosen to have the same number of parameters as competing methods, we show increasing the number of hidden units to 128 significantly improves the performance of SIS-RNN without overfitting.

IAM-OnDB. This human handwriting dataset contains 13,040 handwriting lines written by 500 writers [22]. The writing trajectories are represented as a sequence of (x, y) coordinates together with binary indicators of pen-up/pen-down. We followed [11, 23] to preprocess and split the dataset. The experimental setup for IAM-OnDB is the same as that of the sequential MNIST experiment except that we used 256 hidden units for GRU to have the same number of parameters

Table 1. Comparison of the negative log-likelihood (NLL) between various algorithms for sequential MNIST.

Model	NLL
DBN 2hl	≈84.55
NADE	88.33
EoNADE-5 2hl	84.68
DLGM 8 [15]	≈85.51
DARN 1hl [16]	≈84.13
DRAW [16]	≤80.97
PixelVAE [17]	≈79.02
P-Forcing$_{(3\text{-layers})}$ [18]	79.58
PixelRNN$_{(1\text{-layer})}$ [19]	80.75
PixelRNN$_{(7\text{-layers})}$ [19]	79.20
MatNets [20]	78.50
Z-Forcing$_{(1\text{-layer})}$ [1]	≤ 80.60
Z-Forcing$_{(1\text{-layer})}$ + aux [1]	≤ 80.09
TwinNet$_{(3\text{-layers})}$ [21]	≤ 79.12
VRNN$_{(1\text{-layer})}$	≤ 74.15
SIS-RNN$_{(1\text{-layer})}$ 64	**71.90**
SIS-RNN$_{(1\text{-layer})}$ 128	**70.57**

Table 2. Comparison of the average NLL between various algorithms for IAM-OnDB.

Model	Average NLL
RNN [11]	-1358
VRNN [11]	≤-1384
WAVENET [23]	-1021
SWAVENET [23]	≤-1301
STCN [23]	≤-1338
STCN-DENSE [23]	≤-1796
SIS-RNN$_{(1\text{-LAYER})}$	**-1973**

with competing methods. We report the average negative log-likelihood of test examples in Table 2. For SIS-RNN, WaveNet, and RNN, we report the exact log-likelihood, while in the other cases, we report the variational lower bound (with $\leq$ sign). Our results show that SIS-RNN achieves higher log-likelihood, which supports our expectation that implicit latent random variables are helpful when modeling complex sequences.

Language modeling. Due to their powerful model capacity by distribution-based latent representations, VAEs have become the generative models of choice for dealing with many natural language processing (NLP) tasks including language modeling [24]. This flexible representation allows capturing holistic properties of sentences, such as text style, topic, and high-level linguistic and semantic features. Generated samples from the prior latent distribution can further produce diverse and well-formed sentences through simple deterministic decoding.

Despite its popularity, 1) the adopted auto-regressive decoder, which is often implemented with an RNN, tends to ignore the latent variables in decoding, yielding "posterior collapse" [24, 14]; 2) the Gaussian assumption imposed on the variational distribution restricts its variational inference capacity. While there exists a variety of methods to address the first problem by either changing the decoder [25] or applying KL annealing [24, 14], only a few works addressed the latter one including semi-amortized VAE (SA-VAE) [26].

Table 3. Comparison of language modeling on two datasets.

Dataset	Model	NLL	PPL	KL
YAHOO	VAE-LSTM [24]	337.3	68.31	0.0
	VAE-TRANSFORMER [27]	328.6	61.6	0.7
	SA-VAE [26]	327.2	60.1	5.2
	SIS-RNN	**326.7**	**59.8**	4.2
PTB	VAE-LSTM [24]	102.1	105.2	0.0
	VAE-TRANSFORMER [27]	101.5	102.4	0.2
	CYCLIC-VAE [14]	103.1	110.5	3.5
	SA-VAE [26]	102.6	107.1	1.2
	SIS-RNN	**101.2**	**101.8**	1.6

It has been shown that having one latent variable for each sentence is more effective than including one latent variable for each word [14]. Therefore, for this experiment, we customized our SIS-RNN to have only one stochastic latent variable for each sequence of data, i.e. sentence. More specifically, we only infer one variational latent variable from the last hidden state of RNN. Moreover, we used a self-attention transformer as the decoder, i.e. φ^{decoder}, similar to [27]. We used the same experimental setting as in the previous works [24, 25, 27]. The rest of the hyper-parameters of our model are the same as those of our IAM-OnDB experiment. We consider two public datasets, the Yahoo [25] and Penn Treebank (PTB) [24]. While PTB is a relatively small dataset with sentences of varying lengths, Yahoo contains more samples with longer sentences. Table 3 shows the perplexity (PPL), sentence-level NLL and KL divergence of test samples. Not only SIS-RNN outperforms other methods in terms of NLL and PPL, but also checking KL values indicates that SIS-RNN does not suffer from posterior collapse. *"how to stay in hot water when i get dizzy?"*, *"i just like this girl and we have been friends for 4 yrs."*, and *"i hate it personally."* are the generated examples from the model trained on Yahoo. *"he probably showed it this month as a fundamental policy which includes the best of <unk> and sales <EOS>"* and *"the market 's bullish trend is underway <EOS>"* are two generated examples from the model trained on PTB.

5. CONCLUSION

We have proposed SIS-RNN, the first stochastic recurrent latent variable model with more expressive variational posteriors. We argue that more flexible variational inference in SIS-RNN is a key to better modeling of the dependency in the sequential data. We have tested SIS-RNN on three different tasks with SIS-RNN outperforming competing methods substantially.

6. ACKNOWLEDGEMENT

The presented materials are based upon the work supported by the National Science Foundation under Grants ECCS-1839816, IIS-1848596, CCF-1553281, IIS-1812641, IIS-1812699, and CCF-1934904. We also thank Texas A&M High Performance Research Computing and Texas Advanced Computing Center for providing computational resources to perform experiments in this work.

7. REFERENCES

[1] Anirudh Goyal Alias Parth Goyal, Alessandro Sordoni, Marc-Alexandre Côté, Nan Rosemary Ke, and Yoshua Bengio, "Z-forcing: Training stochastic recurrent networks," in *Advances in neural information processing systems*, 2017.

[2] Ehsan Hajiramezanali, Arman Hasanzadeh, Nick Duffield, Krishna R Narayanan, Mingyuan Zhou, and Xiaoning Qian, "Variational graph recurrent neural networks," in *Advances in neural information processing systems*, 2019.

[3] Jen-Tzung Chien and Chun-Wei Wang, "Variational and hierarchical recurrent autoencoder," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3202–3206.

[4] Ehsan Hajiramezanali, Mahdi Imani, Ulisses Braga-Neto, Xiaoning Qian, and Edward R Dougherty, "Scalable optimal bayesian classification of single-cell trajectories under regulatory model uncertainty," *BMC genomics*, vol. 20, no. 6, pp. 435, 2019.

[5] Ehsan Hajiramezanali, Siamak Zamani Dadaneh, Paul de Figueiredo, Sing-Hoi Sze, Mingyuan Zhou, and Xiaoning Qian, "Differential expression analysis of dynamical sequencing count data with a gamma markov chain," *arXiv preprint arXiv:1803.02527*, 2018.

[6] Ehsan Hajiramezanali, Siamak Zamani Dadaneh, Alireza Karbalayghareh, Mingyuan Zhou, and Xiaoning Qian, "Bayesian multi-domain learning for cancer subtype discovery from next-generation sequencing count data," in *Advances in Neural Information Processing Systems*, 2018, pp. 9115–9124.

[7] David M Blei, Alp Kucukelbir, and Jon D McAuliffe, "Variational inference: A review for statisticians," *Journal of the American Statistical Association*, vol. 112, no. 518, 2017.

[8] Mingzhang Yin and Mingyuan Zhou, "Semi-implicit variational inference," *arXiv preprint arXiv:1805.11183*, 2018.

[9] Arman Hasanzadeh, Ehsan Hajiramezanali, Nick Duffield, Krishna R Narayanan, Mingyuan Zhou, and Xiaoning Qian, "Semi-implicit graph variational auto-encoders," in *Advances in neural information processing systems*, 2019.

[10] Shahin Boluki, Randy Ardywibowo, Siamak Zamani Dadaneh, Mingyuan Zhou, and Xiaoning Qian, "Learnable Bernoulli dropout for Bayesian deep learning," *arXiv preprint arXiv:2002.05155*, 2020.

[11] Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel, Aaron C Courville, and Yoshua Bengio, "A recurrent latent variable model for sequential data," in *Advances in neural information processing systems*, 2015, pp. 2980–2988.

[12] Siamak Zamani Dadaneh, Shahin Boluki, Mingyuan Zhou, and Xiaoning Qian, "Arsm gradient estimator for supervised learning to rank," *arXiv preprint arXiv:1911.00465*, 2019.

[13] Hugo Larochelle and Iain Murray, "The neural autoregressive distribution estimator," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*.

[14] Xiaodong Liu, Jianfeng Gao, Asli Celikyilmaz, Lawrence Carin, et al., "Cyclical annealing schedule: A simple approach to mitigating kl vanishing," *arXiv preprint arXiv:1903.10145*.

[15] Tim Salimans, Diederik Kingma, and Max Welling, "Markov chain monte carlo and variational inference: Bridging the gap," in *International Conference on Machine Learning*, 2015.

[16] Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Rezende, and Daan Wierstra, "DRAW: A recurrent neural network for image generation," in *International Conference on Machine Learning*, 2015, pp. 1462–1471.

[17] Ishaan Gulrajani, Kundan Kumar, Faruk Ahmed, Adrien Ali Taiga, Francesco Visin, David Vazquez, and Aaron Courville, "Pixelvae: A latent variable model for natural images," *arXiv preprint arXiv:1611.05013*, 2016.

[18] Alex M Lamb, Anirudh Goyal ALIAS PARTH GOYAL, Ying Zhang, Saizheng Zhang, Aaron C Courville, and Yoshua Bengio, "Professor forcing: A new algorithm for training recurrent networks," in *Advances In Neural Information Processing Systems*, 2016, pp. 4601–4609.

[19] Aaron Van Oord, Nal Kalchbrenner, and Koray Kavukcuoglu, "Pixel recurrent neural networks," in *ICML 2016*.

[20] Philip Bachman, "An architecture for deep, hierarchical generative models," in *NIPS 2016*.

[21] Dmitriy Serdyuk, Nan Rosemary Ke, Alessandro Sordoni, Adam Trischler, Chris Pal, and Yoshua Bengio, "Twin networks: Matching the future for sequence generation," 2018.

[22] Marcus Liwicki and Horst Bunke, "Iam-ondb-an on-line english sentence database acquired from handwritten text on a whiteboard," in *Document Analysis and Recognition, 2005. Proceedings. Eighth International Conference on*. IEEE, 2005.

[23] Guokun Lai, Bohan Li, Guoqing Zheng, and Yiming Yang, "Stochastic wavenet: A generative latent variable model for sequential data," *arXiv preprint arXiv:1806.06116*, 2018.

[24] Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio, "Generating sentences from a continuous space," *arXiv preprint arXiv:1511.06349*, 2015.

[25] Zichao Yang, Zhiting Hu, Ruslan Salakhutdinov, and Taylor Berg-Kirkpatrick, "Improved variational autoencoders for text modeling using dilated convolutions," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 3881–3890.

[26] Yoon Kim, Sam Wiseman, Andrew C Miller, David Sontag, and Alexander M Rush, "Semi-amortized variational autoencoders," *arXiv preprint arXiv:1802.02550*, 2018.

[27] Zhiting Hu, Zichao Yang, Tiancheng Zhao, Haoran Shi, Junxian He, Di Wang, Xuezhe Ma, Zhengzhong Liu, Xiaodan Liang, Lianhui Qin, et al., "Texar: A modularized, versatile, and extensible toolbox for text generation," in *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, 2018.