INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Self Attention in Variational Sequential Learning for Summarization

Jen-Tzung Chien, Chun-Wei Wang

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

Abstract

Attention mechanism plays a crucial role in sequential learning for many speech and language applications. However, it is challenging to develop a *stochastic attention* in a sequence-to-sequence model which consists of two recurrent neural networks (RNNs) as the encoder and decoder. The problem of *posterior collapse* happens in variational inference and results in the estimated latent variables close to a standard Gaussian prior so that the information from input sequence is disregarded in learning process. This paper presents a new recurrent autoencoder for sentence representation where a *self attention* scheme is incorporated to activate the interaction between inference and generation in training procedure. In particular, a stochastic RNN decoder is implemented to provide additional latent variable to fulfill self attention for sentence reconstruction. The posterior collapse is alleviated. The latent information is sufficiently attended in variational sequential learning. During test phase, the estimated prior distribution of decoder is sampled for stochastic attention and generation. Experiments on Penn Treebank and Yelp 2013 show the desirable generation performance in terms of perplexity. The visualization of attention weights also illustrates the usefulness of self attention. The evaluation on DUC 2007 demonstrates the merit of variational recurrent autoencoder for document summarization.

Index Terms: sequence generation, variational autoencoder, sequence-to-sequence learning, attention mechanism

1. Introduction

Attention mechanism has been practically developed in sequential learning [1] for spatial and temporal data and successfully applied for image caption [2], speech recognition [3, 4], machine translation [5, 6, 7], document summarization [8, 9], textual entailment [10], reading comprehension [11, 12] to name a few. Basically, system performance of an encoder-decoder network can be improved with attention scheme especially when the input sequence is long or contains rich information. Such an attention aims to simulate human visual and hearing systems so as to concentrate on a certain part of a spatial or temporal sequence to generate desirable output sequence at each time step. There have been enormous works carried out as attention approaches to sequential learning and employed in different applications. Recently, self attention, also known as the intra-attention, has been successfully proposed in different speech and language processing tasks. Basically, self attention is implemented to produce a sequence representation by mutually relating different tokens of an input sequence. This mechanism yields a powerful semantic representation for the corresponding word sequence. In the literature, the well-known *transformer* [7] was extensively built with the self attention without the usage of convolutional neural network and recurrent neural network (RNN). Transformer has achieved state-of-the-art performance in machine translation and other natural language processing tasks. A very recent work [13] proposed a generative adversarial network [14] based on self attention which has attained an excellent performance in image generation.

This paper presents a self attention model in variational autoencoder (VAE) [15, 16, 17] which is developed for stochastic sequential learning from a heterogeneous sequence database. Our goal is to present a robust solution to sequence generation and representation for summarization. However, the learning process is challenging because the encoder will be disregarded when generating an output sequence from latent space. This is caused due to the posterior collapse in variational inference. Self attention is merged to tackle this issue in construction of recurrent autoencoder. Typically, self attention is employed as an connector between RNN encoder and decoder which encourages the interaction between the inference model and the generative model in a sequence-to-sequence learning. Owing to the capability of stochastic RNN [18, 19], we build a decoder which provides an additional latent variable to carry out the stochastic attention to calculate the context vector at each time step as the weighted sum of hidden states of an encoder. Given this model, the decoder will sample the features to determine the context vector to fulfill self attention in test phase. A variational sequential learning is developed to build a latent variable model of encoder and decoder which is driven by self attention and applied for semantic representation and document summairzation.

2. Background Survey

2.1. Recurrent variational autoencoder

Sequential learning for semantic representation is crucial for speech and language processing. RNN-based variational autoencoder was proposed for sequential learning of music and text [16, 17] where two RNNs were used as the encoder and decoder to build the variational recurrent autoencoder (VRAE). The encoder compresses an input sequence $\mathbf{x} = \{\mathbf{x}_t\}_{t=1}^T$ into a latent representation $\mathbf{z}$ based on the variational distribution $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ while the decoder reconstructs the samples from latent space using the generative distribution $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$. Figure 1 illustrates how a sequence is reconstructed via VARE. RNN encoder and decoder with parameters $\{\boldsymbol{\phi}, \boldsymbol{\theta}\}$ are estimated by maximizing the variational lower bound of log likelihood

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

where the Kullback-Leibler (KL) divergence is minimized to regularize to a standard Gaussian prior $p(\mathbf{z})$. The reparameterization trick [15] is applied to compute the stochastic gradients. However, VRAE suffers from the posterior collapse in variational sequential learning where KL term in Eq. (1) tends to be vanished, i.e. $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) \approx p(\mathbf{z})$, due to the *autoregression* in RNN decoder. Latent variable $\mathbf{z}$ could not reflect the information from sequential data $\mathbf{x}$. Local optimum likely happens. Such a problem was handled by either weakening the decoder [20, 21] or strengthening the encoder [22]. This study presents a meaningful solution based on a decoder with self attention.

Copyright © 2019 ISCA

http://dx.doi.org/10.21437/Interspeech.2019-1548

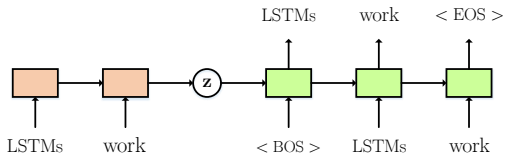

Figure 1: *Recurrent variational autoencoder.*

2.2. Attention mechanism

Attention scheme is beneficial for sequential learning which estimates a sequence-to-sequence model for a mapping function from an input sequence $\mathbf{x} = \{\mathbf{x}_i\}_{i=1}^{T_x}$ to a target sequence $\mathbf{y} = \{\mathbf{y}_t\}_{t=1}^{T_y}$ for speech recognition and machine translation. Their lengths T_x and T_y are different. Sequence-to-sequence model consists of two RNNs or two long short-term memories (LSTMs) as the encoder and decoder. In [5, 23, 24], the encoder was implemented by a bidirectional LSTM to calculate the hidden states $\{\mathbf{s}_i\}_{i=1}^{T_x}$ while LSTM decoder was performed to estimate the hidden state $\mathbf{h}_t$ at each time t by a recurrent function using previous hidden state $\mathbf{h}_{t-1}$ and target vector

$$\mathbf{h}_t = f(\mathbf{h}_{t-1}, \mathbf{y}_{t-1}, \mathbf{c}_t) \quad \text{where} \quad \mathbf{c}_t = \sum_{i=1}^{T_x} \alpha_{ti}\mathbf{s}_i. \tag{2}$$

$\mathbf{c}_t$ is a context vector which depends on all source hidden states encoded from input sequence. This vector is computed as a weighted sum of source hidden states $\{\mathbf{s}_i\}_{i=1}^{T_x}$ using the attention weights $\boldsymbol{\alpha}_t = \{\alpha_{ti}\}_{i=1}^{T_x}$ which are calculated by

$$\alpha_{ti} = \frac{\exp(\text{score}(\mathbf{h}_{t-1}, \mathbf{s}_i))}{\sum_{j=1}^{T_x} \exp(\text{score}(\mathbf{h}_{t-1}, \mathbf{s}_j))}. \tag{3}$$

The score function between hidden states $\mathbf{h}_t$ and $\mathbf{s}_j$ of decoder and encoder at times t and j, respectively, was defined by [5]

$$\text{score}(\mathbf{h}_t, \mathbf{s}_i) = \mathbf{w}_a^\top \tanh\left(\mathbf{W}_a\mathbf{h}_t + \mathbf{V}_a\mathbf{s}_i\right) \tag{4}$$

where $\{\mathbf{w}_a, \mathbf{W}_a, \mathbf{V}_a\}$ denote the attention parameters. Owing to this attention mechanism, LSTM decoder implements an attended representation for each input word and accordingly carries out a precise sequential prediction for target words.

3. Sequential Learning and Self Attention

This paper presents the self attention in a variational recurrent autoencoder for sentence generation where semantic representation is performed with the same input and output sequences $\mathbf{x} = \{\mathbf{x}_t\}_{t=1}^{T}$. Different from VRAE [16, 17], we sufficiently learn the latent space for stochastic recurrent decoder with the complementary information extracted by self attention. The variational sequence modeling with self attention in inference and generative processes is detailed in what follows.

3.1. Model construction

Figure 2(a) illustrates the graphical representation of the proposed method. This model consists two primary latent variables $\mathbf{z}_{\text{enc}}$ and $\{\mathbf{z}_t\}_{t=1}^{T}$ for representation of the whole sentence $\mathbf{x}$ and the individual words $\{\mathbf{x}_t\}_{t=1}^{T}$, respectively, based on the stochastic LSTM decoder. $\widetilde{\mathbf{c}}_t$ is an auxiliary latent variable for context vector driven by self attention using the same sentence $\mathbf{x}$. First, the hidden state of LSTM encoder is obtained by

$$\mathbf{s}_t = f_{\boldsymbol{\phi}}^{\text{enc}}(\mathbf{x}_t, \mathbf{s}_{t-1}) \tag{5}$$

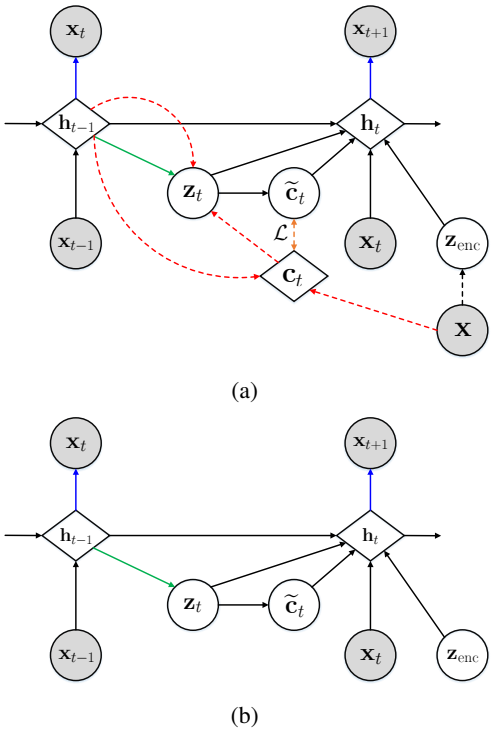

Figure 2: *(a) Self attention in variational autoencoder. Black and red dash lines denote the inference models of a sentence and a word, respectively. Blue and green solid lines denotes the generative model. Orange line denotes auxiliary prediction. Diamond variable is deterministic. Circle variable is stochastic. (b) Generative process where the inference process is removed.*

while the hidden state of LSTM decoder is calculated by

$$\mathbf{h}_t = f_{\boldsymbol{\theta}}^{\text{dec}}(\mathbf{x}_t, \mathbf{h}_{t-1}, \mathbf{z}_t, \widetilde{\mathbf{c}}_t, \mathbf{z}_{\text{enc}}). \tag{6}$$

Notably, three latent variables $\mathbf{z}_t$, $\widetilde{\mathbf{c}}_t$ and $\mathbf{z}_{\text{enc}}$ are merged in LSTM decoder. The decoder is strengthened to avoid mode collapse in variational sequential learning. The functions $f_{\boldsymbol{\phi}}^{\text{enc}}$ and $f_{\boldsymbol{\theta}}^{\text{dec}}$ are basically the standard LSTM cells. In the *inference* stage, the latent variable $\mathbf{z}_t$ depends on $\mathbf{h}_{t-1}$ and $\mathbf{c}_t$ where $\mathbf{c}_t$ is the context vector computed as a weighted sum of the hidden states of encoder $\{\mathbf{s}_t\}_{t=1}^{T}$ as computed in Eq. (2) using the attention weights shown in Eqs. (3)(4). Different from [5, 23, 24], *self attention* is performed in stochastic LSTM decoder at each time step t using the same sequence $\mathbf{x} = \{\mathbf{x}_t\}_{t=1}^{T}$. In the *generative* stage, the next output $\mathbf{x}_{t+1}$ is generated by using the hidden state $\mathbf{h}_t$ which is computed with $\{\mathbf{x}_t, \mathbf{h}_{t-1}, \mathbf{z}_t, \widetilde{\mathbf{c}}_t, \mathbf{z}_{\text{enc}}\}$ as shown in Figure 2(b). However, in test time, the sequence generation is performed according to the recursion in Eq. (6) using the samples of latent variables $\mathbf{z}_t$, $\widetilde{\mathbf{c}}_t$ and $\mathbf{z}_{\text{enc}}$ at each time t which are Gaussians with the parameters estimated by variational inference and optimization. Self attention is run even in test session. For sequential learning, the marginal likelihood of $\mathbf{x} = \{\mathbf{x}_t\}_{t=1}^{T}$ is integrated with respect to $\mathbf{z}_t$, $\widetilde{\mathbf{c}}_t$ and $\mathbf{z}_{\text{enc}}$ by using the parameters $\boldsymbol{\theta}$ in four decomposed distributions

$$p(\mathbf{x}) = \int_{\mathbf{z}_{\text{enc}}} p_{\boldsymbol{\theta}}(\mathbf{z}_{\text{enc}}) \prod_{t=0}^{T-1} \int_{\mathbf{z}_t} \int_{\widetilde{\mathbf{c}}_t} p_{\boldsymbol{\theta}}(\mathbf{x}_{t+1}|\mathbf{x}_{\le t}, \mathbf{z}_t, \widetilde{\mathbf{c}}_t, \mathbf{z}_{\text{enc}}) \times p_{\boldsymbol{\theta}}(\mathbf{z}_t|\mathbf{x}_{\le t}, \mathbf{z}_{\text{enc}}) p_{\boldsymbol{\theta}}(\widetilde{\mathbf{c}}_t|\mathbf{z}_t) d\mathbf{z}_t d\widetilde{\mathbf{c}}_t d\mathbf{z}_{\text{enc}}. \tag{7}$$

3.2. Optimization for stochastic recurrent decoder

The optimization over marginal likelihood in Eq. (7) is intractable due to the coupling of latent variables. We therefore maximize the variational evidence lower bound (ELBO) with latent variables $\mathbf{z}_t$, $\widetilde{\mathbf{c}}_t$ and $\mathbf{z}_{\text{enc}}$ in stochastic recurrent decoder

$$\begin{aligned}\mathcal{L}(\mathbf{x};\boldsymbol{\theta},\boldsymbol{\varphi},\boldsymbol{\phi}) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}_{\text{enc}}|\mathbf{x})}\Bigg[\sum_{t=0}^{T-1}\mathbb{E}_{q_{\boldsymbol{\varphi}}(\mathbf{z}_t|\mathbf{x}_{\le t},\mathbf{z}_{\text{enc}})}\Bigg[\mathbb{E}_{p_{\boldsymbol{\theta}}(\widetilde{\mathbf{c}}_t|\mathbf{z}_t)}\Big[\\ \log p_{\boldsymbol{\theta}}(\mathbf{x}_{t+1}|\mathbf{x}_{\le t},\mathbf{z}_t,\widetilde{\mathbf{c}}_t,\mathbf{z}_{\text{enc}}) + \log p_{\boldsymbol{\theta}}(\widetilde{\mathbf{c}}_t|\mathbf{z}_t)\Big]\Bigg] \\ - \mathcal{D}_{\text{KL}}(q_{\boldsymbol{\varphi}}(\mathbf{z}_t|\mathbf{x}_{\le t},\mathbf{z}_{\text{enc}})\|p_{\boldsymbol{\theta}}(\mathbf{z}_t|\mathbf{x}_{\le t},\mathbf{z}_{\text{enc}}))\Bigg] \\ - \mathcal{D}_{\text{KL}}\left(q_{\boldsymbol{\phi}}(\mathbf{z}_{\text{enc}}|\mathbf{x})\|p_{\boldsymbol{\theta}}(\mathbf{z}_{\text{enc}})\right).\end{aligned} \tag{8}$$

Eq. (8) is derived by incorporating the variational distributions for two primary latent variables $\mathbf{z}_t$ and $\mathbf{z}_{\text{enc}}$

$$q_{\boldsymbol{\varphi}}(\mathbf{z}_t|\mathbf{x}_{\le t},\mathbf{z}_{\text{enc}}) = \mathcal{N}\left(\boldsymbol{\mu}_{z,t},\text{diag}(\boldsymbol{\sigma}^2_{z,t})\right) \tag{9}$$

$$q_{\boldsymbol{\phi}}(\mathbf{z}_{\text{enc}}|\mathbf{x}) = \mathcal{N}\left(\boldsymbol{\mu}_{\mathbf{z}_{\text{enc}}},\text{diag}(\boldsymbol{\sigma}^2_{\mathbf{z}_{\text{enc}}})\right) \tag{10}$$

where the Gaussian mean and variance parameters are based on two fully connected neural networks (FC-NNs)

$$[\boldsymbol{\mu}_{z,t},\boldsymbol{\sigma}^2_{z,t}] = f^{(q)}_{\boldsymbol{\varphi}}(\mathbf{h}_{t-1},\mathbf{c}_t) \tag{11}$$

$$[\boldsymbol{\mu}_{\mathbf{z}_{\text{enc}}},\boldsymbol{\sigma}^2_{\mathbf{z}_{\text{enc}}}] = f^{(q)}_{\boldsymbol{\phi}}(\mathbf{s}_T). \tag{12}$$

Meaningfully, the latent code $\mathbf{z}_t$ is driven by previous hidden state $\mathbf{h}_{t-1}$ and current context vector $\mathbf{c}_t$ via self attention while the latent code $\mathbf{z}_{\text{enc}}$ is determined by hidden state in the last time $\mathbf{s}_T$ using the whole sequence $\mathbf{x}$. Notably, the auxiliary latent variable $\widetilde{\mathbf{c}}_t$ is reflected by $\mathbf{z}_t$ which is fully connected by self attention with $\mathbf{c}_t$. The ELBO in Eq. (8) can be rewritten as

$$\begin{aligned}\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}_{\text{enc}}|\mathbf{x})}\Bigg[\sum_{t=0}^{T-1}\mathbb{E}_{q_{\boldsymbol{\varphi}}(\mathbf{z}_t|\mathbf{h}_{t-1},\mathbf{c}_t)}\Bigg[\mathbb{E}_{p_{\boldsymbol{\theta}}(\widetilde{\mathbf{c}}_t|\mathbf{z}_t)}\Big[\log p_{\boldsymbol{\theta}}(\mathbf{x}_{t+1}|\mathbf{h}_t) \\ + \log p_{\boldsymbol{\theta}}(\widetilde{\mathbf{c}}_t|\mathbf{z}_t)\Big]\Bigg] - \mathcal{D}_{\text{KL}}(q_{\boldsymbol{\varphi}}(\mathbf{z}_t|\mathbf{h}_{t-1},\mathbf{c}_t)\|p_{\boldsymbol{\theta}}(\mathbf{z}_t|\mathbf{h}_{t-1}))\Bigg] \\ - \mathcal{D}_{\text{KL}}\left(q_{\boldsymbol{\phi}}(\mathbf{z}_{\text{enc}}|\mathbf{x})\|p_{\boldsymbol{\theta}}(\mathbf{z}_{\text{enc}})\right).\end{aligned} \tag{13}$$

Here, the information $\{\mathbf{x}_{\le t},\mathbf{z}_t,\widetilde{\mathbf{c}}_t,\mathbf{z}_{\text{enc}}\}$ has been encoded and revealed by $\mathbf{h}_t$. The auxiliary cost due to $p_{\boldsymbol{\theta}}(\widetilde{\mathbf{c}}_t|\mathbf{z}_t)$ is imposed in learning objective to carry out self attention which can be further adjusted by a hyperparameter. In maximization of ELBO, the prior distribution $p_{\boldsymbol{\theta}}(\mathbf{z}_{\text{enc}}) = \mathcal{N}(\mathbf{0},\mathbf{I})$ is assumed and the prior neural networks are specified as

$$p_{\boldsymbol{\theta}}(\mathbf{x}_{t+1}|\mathbf{x}_{\le t},\mathbf{z}_t,\widetilde{\mathbf{c}}_t,\mathbf{z}_{\text{enc}}) = \mathcal{N}\left(\boldsymbol{\mu}_{x,t},\text{diag}(\boldsymbol{\sigma}^2_{x,t})\right) \tag{14}$$

$$p_{\boldsymbol{\theta}}(\mathbf{z}_t|\mathbf{x}_{\le t},\mathbf{z}_{\text{enc}}) = \mathcal{N}\left(\boldsymbol{\mu}_{h,t},\text{diag}(\boldsymbol{\sigma}^2_{h,t})\right) \tag{15}$$

$$p_{\boldsymbol{\theta}}(\widetilde{\mathbf{c}}_t|\mathbf{z}_t) = \mathcal{N}\left(\boldsymbol{\mu}_{\widetilde{\mathbf{c}},t},\text{diag}(\boldsymbol{\sigma}^2_{\widetilde{\mathbf{c}},t})\right) \tag{16}$$

where the Gaussian means and variances are calculated by

$$[\boldsymbol{\mu}_{x,t},\boldsymbol{\sigma}^2_{x,t}] = f^{(o)}_{\boldsymbol{\theta}}(\mathbf{h}_t) \tag{17}$$

$$[\boldsymbol{\mu}_{h,t},\boldsymbol{\sigma}^2_{h,t}] = f^{(h)}_{\boldsymbol{\theta}}(\mathbf{h}_{t-1}) \tag{18}$$

$$[\boldsymbol{\mu}_{\widetilde{\mathbf{c}},t},\boldsymbol{\sigma}^2_{\widetilde{\mathbf{c}},t}] = f^{(c)}_{\boldsymbol{\theta}}(\mathbf{z}_t) \tag{19}$$

using the FC-NNs for distributions $p_{\boldsymbol{\theta}}(\mathbf{x}_{t+1}|\mathbf{h}_t)$, $p_{\boldsymbol{\theta}}(\mathbf{z}_t|\mathbf{h}_{t-1})$ and $p_{\boldsymbol{\theta}}(\widetilde{\mathbf{c}}_t|\mathbf{z}_t)$ with inputs $\mathbf{h}_t$, $\mathbf{h}_{t-1}$ and $\mathbf{z}_t$, respectively. The whole model parameters and variational parameters are formed by $\boldsymbol{\Theta} = \{\boldsymbol{\theta},\boldsymbol{\varphi},\boldsymbol{\phi}\}$ with five FC-NNs $\{f^{(o)}_{\boldsymbol{\theta}}(\mathbf{h}_t), f^{(h)}_{\boldsymbol{\theta}}(\mathbf{h}_{t-1}), f^{(c)}_{\boldsymbol{\theta}}(\mathbf{z}_t), f^{(q)}_{\boldsymbol{\varphi}}(\mathbf{h}_{t-1},\mathbf{c}_t), f^{(q)}_{\boldsymbol{\phi}}(\mathbf{s}_T)\}$.

Algorithm 1: Training (or inference) process for variational recurrent autoencoder with self attention

initialize parameters $\boldsymbol{\theta}$, $\boldsymbol{\varphi}$, $\boldsymbol{\phi}$, hyperparamter α, states $\mathbf{s}_0$, $\mathbf{h}_0$
for *number of iterations* **do**
  Forward pass: *recurrent encoder*
  for $t = 1,\cdots,T$ **do**
    $\mathbf{s}_t \leftarrow f^{\text{enc}}_{\boldsymbol{\phi}}(\mathbf{x}_t,\mathbf{s}_{t-1})$
  end
  $q_{\boldsymbol{\phi}}(\mathbf{z}_{\text{enc}}|\mathbf{x}) \leftarrow f^{(q)}_{\boldsymbol{\phi}}(\mathbf{s}_T)$
  $\mathbf{z}_{\text{enc}}$ is sampled from $q_{\boldsymbol{\phi}}(\mathbf{z}_{\text{enc}}|\mathbf{x})$
  Forward pass: *recurrent decoder*
  for $t = 0,\cdots,T-1$ **do**
    $p_{\boldsymbol{\theta}}(\mathbf{z}_t|\mathbf{x}_{\le t},\mathbf{z}_{\text{enc}}) \leftarrow f^{(h)}_{\boldsymbol{\theta}}(\mathbf{h}_{t-1})$
    $\mathbf{c}_t$ is calculated by attention
    $q_{\boldsymbol{\varphi}}(\mathbf{z}_t|\mathbf{x}_{\le t},\mathbf{z}_{\text{enc}}) \leftarrow f^{(q)}_{\boldsymbol{\varphi}}(\mathbf{h}_{t-1},\mathbf{c}_t)$
    $\mathbf{z}_t$ is sampled from $q_{\boldsymbol{\varphi}}(\mathbf{z}_t|\mathbf{x}_{\le t},\mathbf{z}_{\text{enc}})$
    $p_{\boldsymbol{\theta}}(\widetilde{\mathbf{c}}_t|\mathbf{z}_t) \leftarrow f^{(c)}_{\boldsymbol{\theta}}(\mathbf{z}_t)$
    $\widetilde{\mathbf{c}}_t$ is sampled from $p_{\boldsymbol{\theta}}(\widetilde{\mathbf{c}}_t|\mathbf{z}_t)$
    $\mathbf{h}_t \leftarrow f^{\text{dec}}_{\boldsymbol{\theta}}(\mathbf{x}_t,\mathbf{h}_{t-1},\mathbf{z}_t,\widetilde{\mathbf{c}}_t,\mathbf{z}_{\text{enc}})$
    $p_{\boldsymbol{\theta}}(\mathbf{x}_{t+1}|\mathbf{x}_{\le t},\mathbf{z}_t,\widetilde{\mathbf{c}}_t,\mathbf{z}_{\text{enc}}) \leftarrow f^{(o)}_{\boldsymbol{\theta}}(\mathbf{h}_t)$
    accumulate the variational objective $\mathcal{L}(\mathbf{x};\boldsymbol{\theta},\boldsymbol{\varphi},\boldsymbol{\phi})$
  end
  Backward pass
  compute gradients $\frac{\partial\mathcal{L}}{\partial\boldsymbol{\theta}}$, $\frac{\partial\mathcal{L}}{\partial\boldsymbol{\varphi}}$, $\frac{\partial\mathcal{L}}{\partial\boldsymbol{\phi}}$
  update feedforward and recurrent parameters $\boldsymbol{\theta}$, $\boldsymbol{\varphi}$, $\boldsymbol{\phi}$
end

3.3. Algorithms for inference and generation

Algorithm 1 illustrates the training procedure of sequence-to-sequence model. Recurrent and feedforward parameters $\{f^{\text{enc}}_{\boldsymbol{\phi}}, f^{\text{dec}}_{\boldsymbol{\theta}}\}$ and $\{f^{(o)}_{\boldsymbol{\theta}}, f^{(h)}_{\boldsymbol{\theta}}, f^{(c)}_{\boldsymbol{\theta}}, f^{(q)}_{\boldsymbol{\varphi}}, f^{(q)}_{\boldsymbol{\phi}}\}$ are jointly trained by the stochastic backpropagation algorithm based on the gradients of $\mathcal{L}(\mathbf{x};\boldsymbol{\theta},\boldsymbol{\varphi},\boldsymbol{\phi})$ with respect to individual parameters. In this conditional generative model, $\widetilde{\mathbf{c}}_t$ is seen as a predictor of $\mathbf{c}_t$ similar to [18, 19]. The auxiliary reconstruction of $\widetilde{\mathbf{c}}_t$ enforces the latent variable $\mathbf{z}_t$ contained with attention information. Algorithm 2 shows the generative procedure of a new sequence $\{\widehat{\mathbf{x}}_t\}_{t=1}^T$ from begin of sentence (bos) $\mathbf{x}_0$. The auxiliary variable or attention predictor $\widetilde{\mathbf{c}}_t$ is sampled for generation of $\widehat{\mathbf{x}}_t$ at each time t for self attention even without future words $\mathbf{x}_{>t}$.

Algorithm 2: Generative process with self attention

require $\boldsymbol{\theta}$, $\boldsymbol{\phi}$, $\mathbf{x}_0$
$\mathbf{z}_{\text{enc}}$ is sampled from $q_{\boldsymbol{\phi}}(\mathbf{z}_{\text{enc}}|\mathbf{x})$
$\mathbf{h}_0$ is initialized
for $t = 0,\cdots,T-1$ **do**
  $\mathbf{z}_t$ is sampled from $p_{\boldsymbol{\theta}}(\mathbf{z}_t|\mathbf{h}_{t-1})$
  $\widetilde{\mathbf{c}}_t$ is generated by $p_{\boldsymbol{\theta}}(\widetilde{\mathbf{c}}_t|\mathbf{z}_t)$
  compute $\mathbf{h}_t \leftarrow f^{\text{dec}}_{\boldsymbol{\theta}}(\mathbf{x}_t,\mathbf{h}_{t-1},\mathbf{z}_t,\widetilde{\mathbf{c}}_t,\mathbf{z}_{\text{enc}})$
  output $\widehat{\mathbf{x}}_{t+1} = \arg\max_{\mathbf{x}_{t+1}} p_{\boldsymbol{\theta}}(\mathbf{x}_{t+1}|\mathbf{h}_t)$
end

4. Experiments

4.1. Experimental setup

The proposed method was evaluated for semantic representation by using the training, validation and test sets from two datasets: Penn TreeBank (PTB) [25, 26, 27] and Yelp 2013 [28, 29]. PTB was a benchmark dataset for language model with 10K vocabulary words. Yelp 2013 was a review dataset with 15K vocabulary words from the Yelp Data Challenge of year 2013. In average, there were 21 and 48 words per sentence in PTB and Yelp,

respectively. The perplexity of test sentences was examined. In addition, Document Understanding Conference (DUC) 2007 (http://duc.nist.gov) was assessed for document summarization. This corpus provided the reference summary for individual document. The automatic summary was limited to 250 words at most. The NIST evaluation tool ROUGE (http://berouge.com) was adopted. ROUGE-1 was used to measure the matched unigrams between reference summary and automatic summary in terms of recall, precision and F-measure. The sentence or document representation was evaluated to select the representative sentences from multiple documents. The sentences with the smallest KL divergence between document and sentence models were selected. The sentence-based latent Dirichlet allocation (sLDA) [30] was included. Detailed setup was referred to [31].

LSTM and VRAE [16] were implemented for language model in comparison of the proposed VRAE with self attention (VRAE-SA). Different VRAEs use LSTMs for both encoder and decoder. All models used one-layer LSTM for both encoder and decoder with an embedding size 512 for $\mathbf{x}_t$ and the number of hidden units 256 for $\mathbf{s}_t$, $\mathbf{h}_t$ and $\mathbf{c}_t$. The dimension of latent variables was 32 for $\mathbf{z}_t$, $\mathbf{z}_{\text{enc}}$ and $\widetilde{\mathbf{c}}_t$. The one-layer FC-NNs were calculated for Gaussian means and variances. Mini-batch size was 32. All models were optimized with 20 epochs by using Adam optimizer with initial learning rate 0.001 which was decreased by a factor of 2 every 2 epochs after 10 epochs. There was a dropout layer with probability 0.5 over the input-to-hidden layer of the LSTM decoder. Gradient clipping was applied with maximum norm 5. The KL-cost annealing strategy [17] was utilized to partially alleviate posterior collapse.

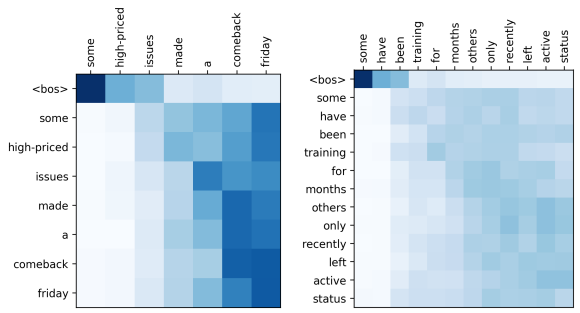

Figure 3: *Self-attention weight maps for two sentences in PTB.*

Model	NLL	KL	PPL
LSTM	102.27	-	132.89
VRAE	101.45	4.86	127.78
VRAE-SA w/o aux	99.82	5.80	118.22
VRAE-SA w aux	**99.19**	**6.16**	**114.68**

Table 1: *NLL and PPL using different methods in PTB.*

4.2. Experimental results

Figure 3 visualizes the self-attention weights for different sentences. Self attention helps to identify useful information for sentence reconstruction. Table 1 and Table 2 compare the negative log-likelihood (NLL) and the perplexity (PPL) of different models by using the PTB and Yelp 2013, respectively. Lower value implies better result. Yelp 2013 has larger variety than PTB. The KL divergence $\mathcal{D}_{\text{KL}}\left(q_{\boldsymbol{\phi}}(\mathbf{z}_{\text{enc}}|\mathbf{x})\|p_{\boldsymbol{\theta}}(\mathbf{z}_{\text{enc}})\right)$ is also listed. Higher KL means less likely posterior collapse and better learning representation. To investigate the *stochastic* self attention, VRAE-SA has the realizations without and with the auxiliary term $p_{\boldsymbol{\theta}}(\widetilde{\mathbf{c}}_t|\mathbf{z}_t)$ in Eq. (13). The best result of NLL and PPL among different models is marked in bold. VRAE-SA obtains the highest KL value so as to learn the most informative latent codes. The stochastic information learned with auxiliary cost does help self attention. Table 3 reports the results of recall, precision and F-measure of document summarization in DUC 2007 where different methods are compared. The semantic representation for sentences and documents is learned and used to extract the latent variables $\widetilde{\mathbf{c}}_T$ from mean vectors of FC-NN $f_{\boldsymbol{\theta}}^{(c)}(\mathbf{z}_T)$ for a document as well as each sentences in this document $\{\mathbf{x}_t\}_{t=1}^T$. Sentence ranking is then performed by using the corresponding context vectors $\widetilde{\mathbf{c}}_T$ of a document and its different sentences to find a summary. Different recurrent machines (LSTM, VRAE, VRAE-SA) work much better than sLDA. The results of F-measure are consistent with those of NLL and PPL. In this comparison, the highest F-measure 0.431 is obtained for document summarization by using the proposed VRAE with stochastic self attention. Source codes are accessible at https://github.com/NCTUMLlab/.

Model	NLL	KL	PPL
LSTM	196.69	-	62.91
VRAE	196.28	2.25	62.38
VRAE-SA w/o aux	193.52	4.07	58.86
VRAE-SA w aux	**191.98**	**6.98**	**56.98**

Table 2: *NLL and PPL using different methods in Yelp 2013.*

Model	Recall	Precision	F-measure
sLDA [30]	0.337	0.390	0.362
LSTM	0.431	0.392	0.411
VRAE	0.438	0.399	0.418
VRAE-SA w/o aux	0.448	0.408	0.427
VRAE-SA w aux	**0.451**	**0.413**	**0.431**

Table 3: *Recall, precision and F-measure for document summarization using different methods in DUC 2007.*

5. Conclusions

We proposed a self-attention model for sequence generation based on variational autoencoder which was employed in semantic representation and summarization. The stochastic recurrent decoder was constructed with self attention and formulated according to variational inference. The additional latent variable in stochastic decoder was incorporated to reflect self attention which allowed us to estimate the context vectors for sentences and documents. The missing of self-attention information was sufficiently compensated in generation procedure for test data. Experimental results showed that the proposed variational sequential model could mitigate the issue of the posterior collapse and improved the performance in terms of perplexity for sentence generation and F-measure for document summarization. Future investigation for generation of new sentences will be studied in other applications, e.g. natural language generation and composition for dialogue with different styles.

6. References

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

[2] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. of International Conference on Machine Learning*, 2015, pp. 2048–2057.

[3] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Advances in Neural Information Processing Systems*, 2015, pp. 577–585.

[4] 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.

[5] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. of International Conference on Learning Representations*, 2015.

[6] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder–decoder for statistical machine translation," in *Proc. of Conference on Empirical Methods in Natural Language Processing*, 2014, pp. 1724–1734.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[8] A. M. Rush, S. Chopra, and J. Weston, "A neural attention model for abstractive sentence summarization," in *Proc. of Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 379–389.

[9] R. Paulus, C. Xiong, and R. Socher, "A deep reinforced model for abstractive summarization," in *Proc. of International Conference on Learning Representations*, 2018.

[10] Z. Lin, M. Feng, C. N. d. Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio, "A structured self-attentive sentence embedding," in *Proc. of International Conference on Learning Representations*, 2017.

[11] J. Cheng, L. Dong, and M. Lapata, "Long short-term memory-networks for machine reading," in *Proc. of Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 551–561.

[12] J.-T. Chien and T.-A. Lin, "Supportive attention in end-to-end memory networks," in *Proc. of IEEE International Workshop on Machine Learning for Signal Processing*, 2018, pp. 1–6.

[13] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," *arXiv preprint arXiv:1805.08318*, 2018.

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

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

[17] 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.

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

[19] S. Shabanian, D. Arpit, A. Trischler, and Y. Bengio, "Variational Bi-LSTMs," *arXiv preprint arXiv:1711.05717*, 2017.

[20] 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.

[21] 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.

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

[23] T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proc. of Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 1412–1421.

[24] J. Su, S. Wu, D. Xiong, Y. Lu, X. Han, and B. Zhang, "Variational recurrent neural machine translation," in *Proc. of AAAI Conference on Artificial Intelligence*, 2018.

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

[26] 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.

[27] 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.

[28] J. Xu, D. Chen, X. Qiu, and X. Huang, "Cached long short-term memory neural networks for document-level sentiment classification," in *Proc. of Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 1660–1669.

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

[30] Y.-L. Chang and J.-T. Chien, "Latent Dirichlet learning for document summarization," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, pp. 1689–1692.

[31] J.-T. Chien, "Hierarchical theme and topic modeling," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 3, pp. 565–578, 2016.