VARIATIONAL AND HIERARCHICAL RECURRENT AUTOENCODER

Jen-Tzung Chien and Chun-Wei Wang, National Chiao Tung University, Hsinchu, Taiwan

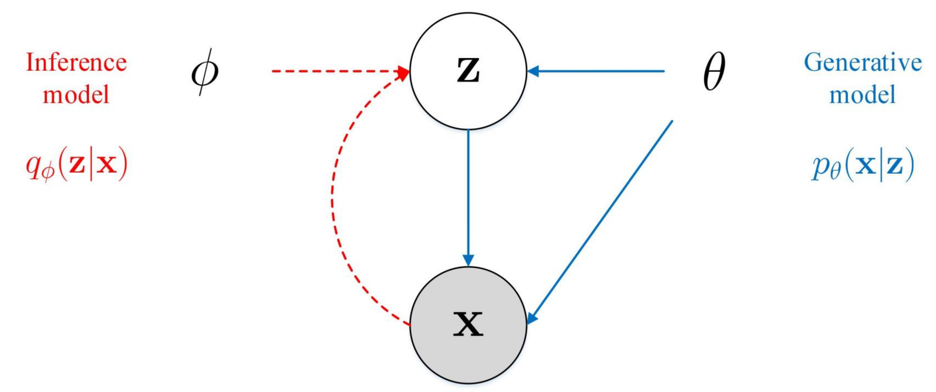
Introduction

- We propose a hierarchical latent variable model for stochastic sequential learning which copes with the issue of posterior collapse in sequence generation
- We strengthen the capability of the encoder by using two different networks
  - long short-term memory (LSTM)
  - pyramid bidirectional LSTM (pBLSTM)
- Global and local dependencies in latent structure are characterized by a sophisticated model and sufficiently learned in stochastic sequence autoencoder

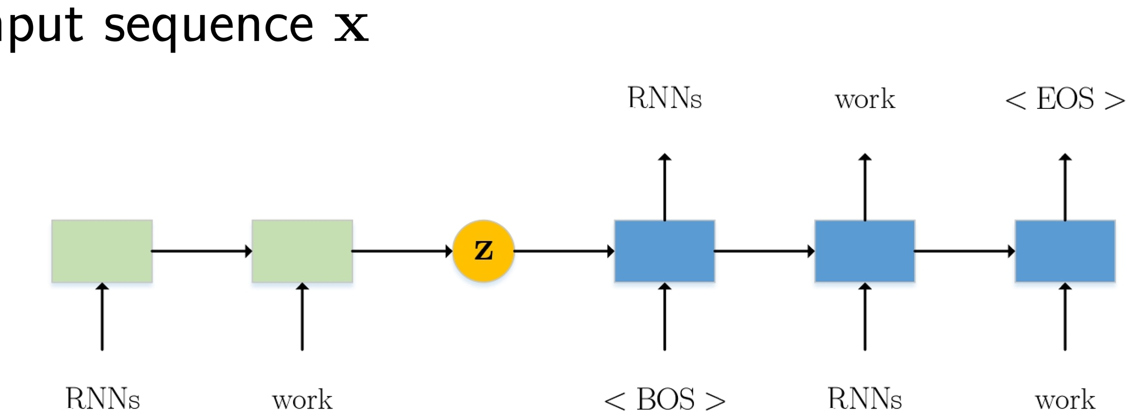
Background Survey

- Variational autoencoder (VAE) (Kingma and Welling, 2013)
  - estimates the distribution of latent variable $\mathbf{z}$ and uses latent information to reconstruct the input signal $\mathbf{x}$
  - learns the encoder $q_\phi(\mathbf{z}|\mathbf{x})$ with parameter ϕ and decoder $p_\theta(\mathbf{x}|\mathbf{z})$ with parameter θ by maximizing a variational lower bound of log likelihood

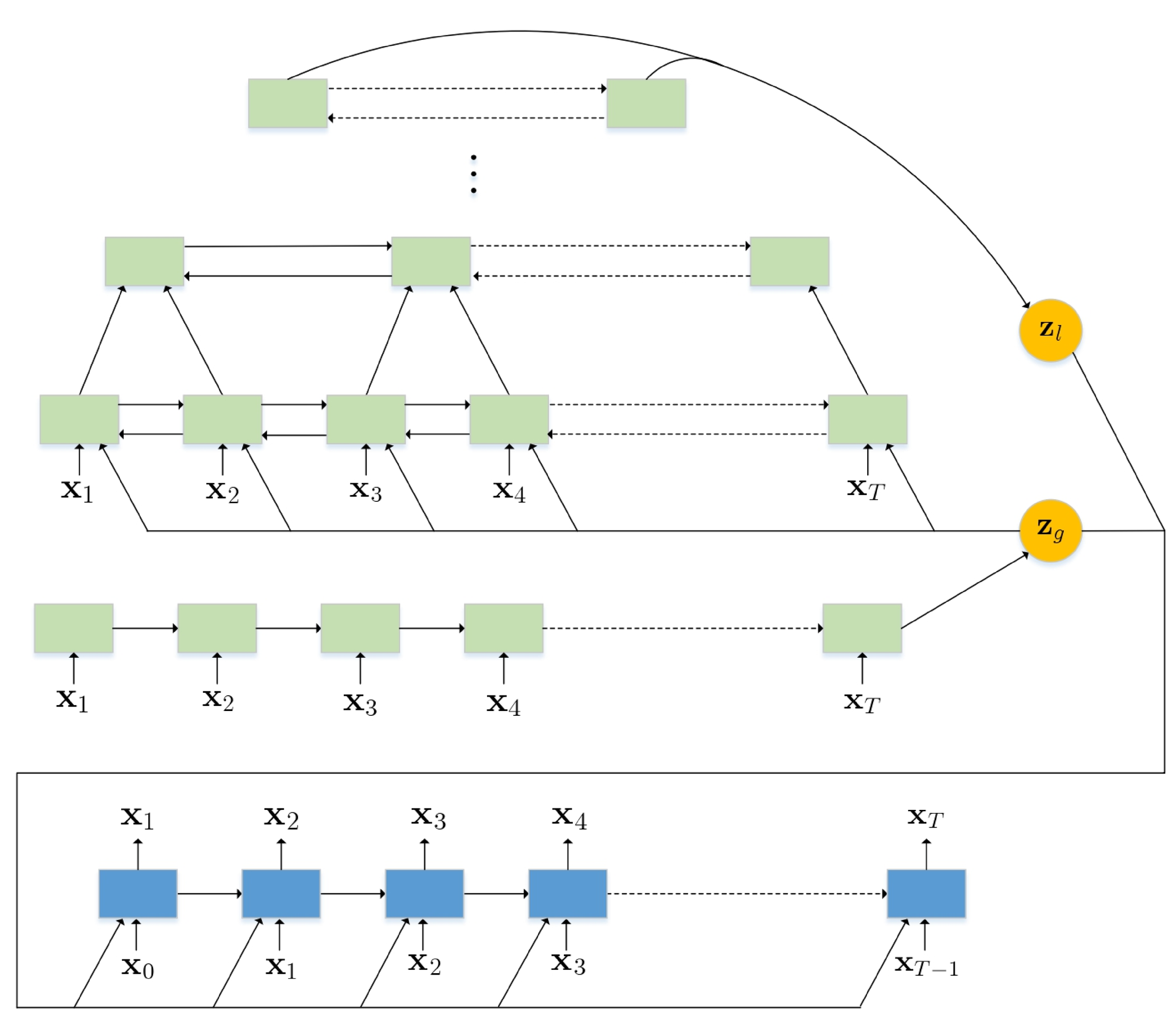
$$\mathcal{L}(\mathbf{x};\theta,\phi) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \mathcal{D}_{\text{KL}}(q_\phi(\mathbf{z}|\mathbf{x})\|p(\mathbf{z}))$$



- Variational recurrent autoencoder (VRAE) (Yang et al., 2017)
  - consists of two RNNs for both encoder and docoder for reconstruction of a sequence data
  - encoder 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 $\mathbf{h}_T$
  - decoder $p_\theta(\mathbf{x}|\mathbf{z})$ 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 $\widehat{\mathbf{x}} = \{\widehat{\mathbf{x}}_t\}_{t=1}^T$ for reconstruction of input sequence $\mathbf{x}$



Variational and Hierarchical Model



- Pyramid bidirectional long short-term memory
  - reduces the time resolution and captures the local features $\mathbf{h}$ in latent space by using the pyramid bidirectional LSTM

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

- Model optimization
  - marginal likelihood of sequence data $\mathbf{x}$ is yielded by

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

  - likelihood function for sequence generation

$$p_\theta(\mathbf{x}|\mathbf{z}_l,\mathbf{z}_g) = \mathcal{N}(\mu_x, \text{diag}(\sigma_x^2))$$

  - variational posterior for inference of latent variables

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

  - variational lower bound of the proposed hierarchical VRAE can be derived from marginal likelihood to find

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

    - ▷ first term reflects the reconstruction error
    - ▷ remaining two terms denote the KL divergence due to local variables $\mathbf{z}_l$ and global variable $\mathbf{z}_g$

Experiments

- Experimental setup
  - baseline system was built by using LSTM language model (denoted by RNNLM)
  - VRAE and hierarchical VRAE were carried out with encoder and docoder
    - ▷ one-layer LSTM as both encoder and decoder
    - ▷ embedding size 300 and hidden units of size 256
  - hierarchical VRAE additionally employed
    - ▷ three-layer pBLSTM with 256 hidden units
  - training settings and evaluation metrics
    - ▷ Penn TreeBank (PTB) ($|V| = 10K$) & Yelp 2013 ($15K$)
    - ▷ minibatch size 32, 20 epoches, latent dimension 16
    - ▷ Adam optimizer, drop prob. 0.5, KL-cost annealing
    - ▷ negative log-likelihood (NLL), KL divergence, perplexity

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.

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

Linear interpolation of two sentences

Conclusions

- We employed a LSTM and a pyramid bidirectional LSTM as encoders to characterize global and local variables, respectively
- The proposed method mitigated the posterior collapse and improved the prediction performance for sentence generation
- A stochastic and hierarchical latent representations was learned
- Document summarization is now under investigation