[Paper review 41]

Neural Variational Inference for Text Processing

(Miao, et al., 2016)

[Contents]

1. Abstract

Introduce **Variational Inference** framework for **generative & conditional models of text**

- traditional) analytic approximation for intractable distn
- this paper) construct **inference network** conditioned on the discrete text input

Validate using 2 applications

- **1) generative document modeling**
- **2) supervised question answering**

2. Introduction

Probabilistic generative models in NLP

- 1) ability use unlabelled data effectively
- 2) incorporate abundant linguistic features
- 3) learn interpretable dependencies among data

$\rightarrow$ but as the model becomes complex.... becomes intractable! (due to high dimensional integrals)

$\rightarrow$ use MCMC or VI

Problem of MCMC & VI

- MCMC : computational cost
- VI : confined due to underestimation of posterior variance

Introduces **NEURAL VARIATIONAL FRAMEWORK** for generative models of text, inspired by **VAE**

$\rightarrow$ main idea : building **inference network** with DNN

- analytic approximation (X)

  model the posterior probability (O) .... thus strong generalization ability
- due to DNN : capture of learning complex non-linear distn
- use reparam tricks

  (to train by back-prop of unbiased & low variance gradients w.r.t latent variables)

Propose

- NVDM (Neural Variational Document Model)
- NASM (Neural Answer Selection Model)

NVDM

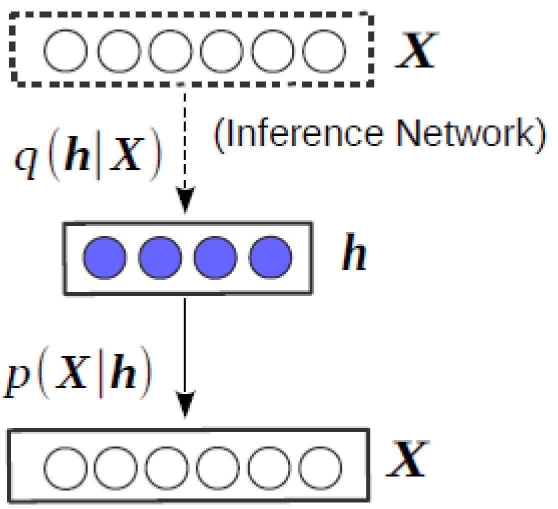

Figure 1. NVDM for document modelling.

- **unsupervised** generative model of text
- extract "continuous semantic latent variable" for each document
- like VAE
  - **MLP encoder** (X : bag-of-words documents $\rightarrow$ Z : continuous latent distn)
  - **Softmax decoder** (reconstructs document!)

    (each word is generated directly from dense continuous document representation)

NASM

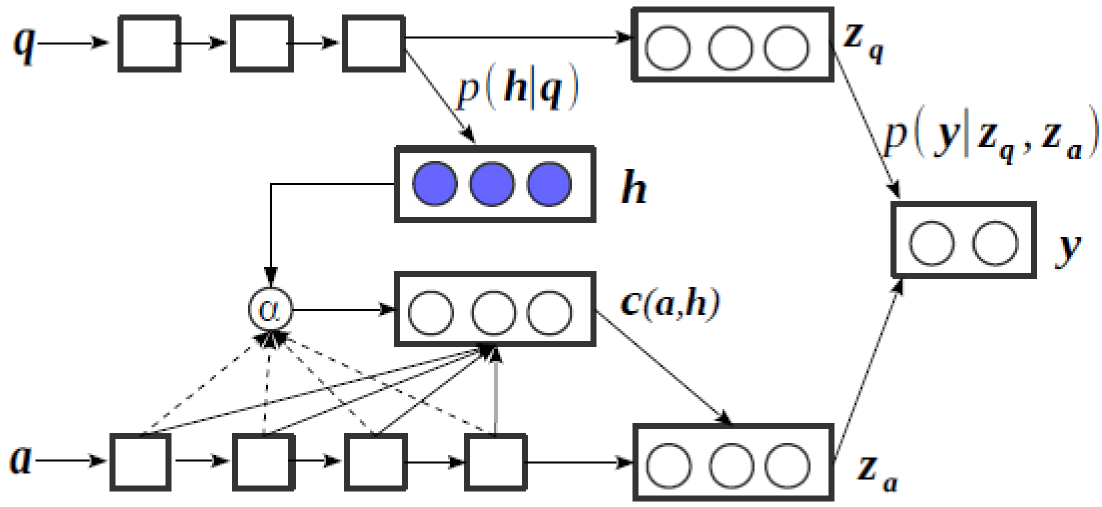

Figure 2. NASM for question answer selection.

- **supervised** conditional model which imbues **LSTMs**
- **attention model**

  (to focus on the phrases of an answer that are strongly connected to the question semantics)

Summary

demonstrate the effectiveness of **Neural Variational Inference for text processing** on 2 tasks

the models are simple, expressive, trained efficiently (with scalable stochastic gradient back-prop)

suitable for unsupervised & supervised task

3. Neural Variational Inference Framework

latent variable model : popular in NLP, but hard to be effective & efficient in complex structures

$\rightarrow$ thus propose generic **"neural variational inference framework"** that can be applied to

- unsupervised NVDM
- supervised NASM

Generative model

- latent variable h

  (= stochastic units in DNN)
- x, y : observed parent & child nodes of h
- joint pdf : $p_\theta(x, y) = \sum_h p_\theta(\boldsymbol{y} \mid \boldsymbol{h}) p_\theta(\boldsymbol{h} \mid \boldsymbol{x}) p(\boldsymbol{x})$

ELBO :

$$
\begin{aligned}
\mathcal{L} &= \mathbb{E}_{q(\boldsymbol{h})}\left[\log p_\theta(\boldsymbol{y} \mid \boldsymbol{h}) p_\theta(\boldsymbol{h} \mid \boldsymbol{x}) p(\boldsymbol{x}) - \log q(\boldsymbol{h})\right] \\
&\leqslant \log \int \frac{q(\boldsymbol{h})}{q(\boldsymbol{h})} p_\theta(\boldsymbol{y} \mid \boldsymbol{h}) p_\theta(\boldsymbol{h} \mid \boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{h} = \log p_\theta(\boldsymbol{x}, \boldsymbol{y})
\end{aligned}.
$$

- $q(\boldsymbol{h})$ should approach true posterior $p(\boldsymbol{h} \mid \boldsymbol{x}, \boldsymbol{y})$
- parameterized diagonal Gaussian

  $q_\phi(\boldsymbol{h} \mid \boldsymbol{x}, \boldsymbol{y}) = \mathcal{N}\left(\boldsymbol{h} \mid \boldsymbol{\mu}(\boldsymbol{x}, \boldsymbol{y}), \operatorname{diag}\left(\boldsymbol{\sigma}^2(\boldsymbol{x}, \boldsymbol{y})\right)\right).$

3 steps to construct **Inference network**

- step 1) Construct **vector representations of the observed variables** : $u = f_x(x), v = f_y(y)$.
  - $f_x(\cdot)$ and $f_y(\cdot)$: DNN
- step 2) Assemble a **joint representation**: $\pi = g(\boldsymbol{u}, \boldsymbol{v})$.
  - $g(\cdot)$: MLP that concatenates vector representations of the conditioning variables
- step 3) **Parameterise the variational distribution** over the latent variable: $\mu = l_1(\pi), \log \sigma = l_2(\pi)$
  - $l(\cdot)$: linear transformation & outputs the params of Gaussian

During training,

- model params θ
- inference network params ϕ

are updated using stochastic-backprop

(1) model params θ

$\nabla_\theta \mathcal{L} \simeq \frac{1}{L} \sum_{l=1}^L \nabla_\theta \log p_\theta\left(\boldsymbol{y} \mid \boldsymbol{h}^{(l)}\right) p_\theta\left(\boldsymbol{h}^{(l)} \mid \boldsymbol{x}\right)$

where $h \sim q_\phi(h \mid x, y)$

(2) inference network params ϕ

need reparam trick!

- $h = \mu + \sigma \cdot \epsilon$.

  where $\epsilon^{(l)} \sim \mathcal{N}(0, I)$
- to reduce variance!

update of ϕ can be carried out, using gradients w.r.t. μ and σ

$s(\boldsymbol{h}) = \log p_\theta(\boldsymbol{y} \mid \boldsymbol{h}) p_\theta(\boldsymbol{h} \mid \boldsymbol{x}) - \log q_\phi(\boldsymbol{h} \mid \boldsymbol{x}, \boldsymbol{y})$

$\nabla_\mu \mathcal{L} \simeq \frac{1}{L} \sum_{l=1}^L \nabla_{\boldsymbol{h}^{(l)}}\left[s\left(\boldsymbol{h}^{(l)}\right)\right].$

$\nabla_\sigma \mathcal{L} \simeq \frac{1}{2L} \sum_{l=1}^L \boldsymbol{\epsilon}^{(l)} \nabla_{\boldsymbol{h}^{(l)}}\left[s\left(\boldsymbol{h}^{(l)}\right)\right].$

4. NVDM (Neural Variational Document Model)

"Unsupervised learning"

$h \in R^K$: continuous hidden variable , used to represent its semantic content

interpret NVDM as VAE

- MLP encoder : $q(\boldsymbol{h} \mid \boldsymbol{X})$
- softmax decoder : $p(\boldsymbol{X} \mid \boldsymbol{h})=\prod_{i=1}^{N} p\left(\boldsymbol{x}_{i} \mid \boldsymbol{h}\right)$

To maximize log likelihood (= $\log \sum_{h} p(\boldsymbol{X} \mid \boldsymbol{h}) p(\boldsymbol{h})$),

$\rightarrow$ maximize ELBO $\mathcal{L}=\mathbb{E}_{q_{\phi}(\boldsymbol{h} \mid \boldsymbol{X})}\left[\sum_{i=1}^{N} \log p_{\theta}\left(\boldsymbol{x}_{i} \mid \boldsymbol{h}\right)\right]-D_{\mathrm{KL}}\left[q_{\phi}(\boldsymbol{h} \mid \boldsymbol{X}) \| p(\boldsymbol{h})\right]$

- N : number of words in the document
- $p(\mathbf{h})$: Gaussian prior for h

Encoder

- inference network
- modeled by Gaussian
- $q_{\phi}(\boldsymbol{h} \mid \boldsymbol{X})=\mathcal{N}\left(\boldsymbol{h} \mid \boldsymbol{\mu}(\boldsymbol{X}), \operatorname{diag}\left(\boldsymbol{\sigma}^{2}(\boldsymbol{X})\right)\right)$.

  where $\pi=g\left(f_{X}^{\mathrm{MLP}}(\boldsymbol{X})\right)$. and $\mu=l_{1}(\boldsymbol{\pi}), \log \sigma=l_{2}(\boldsymbol{\pi})$

Decoder

- modeled by softmax (= multinomial logistic regression)
- $p_{\theta}\left(\boldsymbol{x}_{i} \mid \boldsymbol{h}\right)=\frac{\exp \left\{-E\left(\boldsymbol{x}_{i} ; \boldsymbol{h}, \theta\right)\right)\}}{\sum_{j=1}^{|V|} \exp \left\{-E\left(\boldsymbol{x}_{j} ; \boldsymbol{h}, \theta\right)\right\}}$.

  where $E\left(\boldsymbol{x}_{i} ; \boldsymbol{h}, \theta\right)=-\boldsymbol{h}^{T} \boldsymbol{R} \boldsymbol{x}_{i}-\boldsymbol{b}_{x_{i}}$

ELBO (= $\mathcal{L}=\mathbb{E}_{q_{\phi}(\boldsymbol{h} \mid \boldsymbol{X})}\left[\sum_{i=1}^{N} \log p_{\theta}\left(\boldsymbol{x}_{i} \mid \boldsymbol{h}\right)\right]-D_{\mathrm{KL}}\left[q_{\phi}(\boldsymbol{h} \mid \boldsymbol{X}) \| p(\boldsymbol{h})\right]$) can be optimized,

by back-prop of stochastic gradients w.r.t θ and ϕ

5. NASM (Neural Answer Selection Model)

after reviewing Attention mechanism...