Advanced Topics on Sequence Generation

Lili Mou
doublepower.mou@gmail.com
http://sei.pku.edu.cn/~moull12
Outline

- The Basics
- BiRNN as Generators
- Reinforcement Learning
- Variational Autoencoder
The Basics

- Seq2seq
- Attention
Sequence to Sequence Generation

- Training phrase: X, Y, and Z are the ground truth (words in the corpus)
- Predicting phrase: X, Y, and Z are those generated by RNN
- Seq2seq model is essentially an LM (of XYZ) conditioned on another LM (of ABC)

The Attention Mechanism

- During sequence generation, the output sequence's hidden state $h_t$ is related to
  - That of the last time step $h_{t-1}$, and
  - A context vector $c$, which is a combination of the input sequence's states

$$h_t = \text{RNN}(h_{t-1}, c) = f(W[h_{t-1}; c])$$

Context Vector

The context vector $\mathbf{c}$ is a combination of the input sequence's states

$$\mathbf{c} = \sum_i \alpha_i \mathbf{c}_i$$

where the coefficient $\alpha_i$ is related to

- The local context $\mathbf{c}_i$, and
- The last output state $\mathbf{h}_{t-1}$
- $\alpha_i$ is normalized

$$\alpha_i = \frac{\exp\{\tilde{\alpha}_i\}}{\sum_j \exp\{\tilde{\alpha}_j\}}$$

$$\tilde{\alpha}_i = W[\mathbf{h}_{t-1}; \mathbf{c}_i]$$
But...

- Deep learning is far beyond CNNs, RNNs, etc.
Outline

- The Basics
- BiRNN as Generators
- Reinforcement Learning
- Variational Autoencoder
## Bidirectional Recurrent Neural Networks as Generative Models

Mathias Berglund  
Aalto University, Finland

Tapani Raiko  
Aalto University, Finland

Mikko Honkala  
Nokia Labs, Finland

Leo Kärkkäinen  
Nokia Labs, Finland

Akos Vetek  
Nokia Labs, Finland

Juha Karhunen  
Aalto University, Finland

- Motivation: Use LMs of both directions during language generation
Figure 1: Structure of the simple RNN (left) and the bidirectional RNN (right).

The output $y_t$ is traditionally determined by

$$P(y_t | \{x_d\}_{d \neq t}) = \phi \left( W_y^f h_t^f + W_y^b h_t^b + b_y \right),$$

but we propose the use of

$$P(y_t | \{x_d\}_{d \neq t}) = \phi \left( W_y^f h_{t-1}^f + W_y^b h_{t+1}^b + b_y \right),$$

where

$$h_t^f = \tanh \left( W_h^f h_{t-1}^f + W_x^f x_t + b_h^f \right)$$

$$h_t^b = \tanh \left( W_h^b h_{t+1}^b + W_x^b x_t + b_h^b \right).$$
Probabilistic Interpretation

- **Unidirectional distribution**
  \[ P_{\text{unidirectional}}(X) = \prod_{t=1}^{T} P(x_t \mid \{x_d\}_{d=1}^{t-1}) \]

- **Bidirectional distribution**
  - Interpretation I (Generative Stochastic Networks, GSN)
    Asymptotic distribution when sampling
    \[ P_{\text{BRNN}}(x_t \mid \{x_d\}_{d \neq t}) \]
    Essentially the distribution defined by Gibbs sampling
  - Interpretation II (Neural Autoregressive Distribution Estimator, NADE)
    \[ O_t: \text{a permutation of time indexes } 1..T \]
    \[ P_{\text{NADE}}(X \mid o_d) = \prod_{d=1}^{T} P(x_{od} \mid \{x_{oe}\}_{e=1}^{d-1}) \]
## Training and Inference

<table>
<thead>
<tr>
<th></th>
<th>Training</th>
<th>Inference</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSN</td>
<td>Assume the whole sentence is known</td>
<td>Gibbs sampling</td>
</tr>
<tr>
<td>NADE</td>
<td>Set some input to a missing value</td>
<td>Start from random initialization, and compute the prob. in a one-pass fashion</td>
</tr>
<tr>
<td>Bayesian MCMC</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[
P_{\text{RNN}}(x_t = a \mid \{x_d\}_{d \neq t})
\propto P_{\text{RNN}}(x_t = a \mid \{x_d\}_{d=1}^{t-1})\, P_{\text{RNN}}(\{x_e\}_{e=t+1}^T \mid x_t = a, \{x_d\}_{d=1}^{t-1})
= \prod_{\tau=t}^T P_{\text{RNN}}(x_\tau \mid \{x_d\}_{d=1}^{\tau-1}) \bigg|_{x_t = a}
\]
Pros and Cons

- **Pros**
  - 😊 The constraint of “sequence” is slacked

- **Cons**
  - 😞 Virtually impossible to be used in practice
C.f. our B/F LMs

- stochastic **gradient** - based algorithm for `<unk>` - based convex optimization
- **gradient** - **free** learning with a stochastic block model
- on global convergence of **sub** - **gradient** descent for sparse graphs
- a stochastic **gradient descent** algorithm for `<unk>` - `<unk>` systems
- a **stochastic gradient descent** algorithm for `<unk>` - `<unk>` systems

- deep **convolutional neural networks** for object detection
- **semi** - supervised learning with **convolutional neural networks**
- semantic image **segmentation** with **convolutional neural networks** for `<unk>`
- efficient object **tracking** with **convolutional neural networks**
- deep **convolutional neural networks** for language
**Step I**
Generate the middle part with the second constraint as additional input.

**Step II**
Backward generation.

**Step III**
Forward generation.
a stochastic gradient descent algorithm for unk - unk systems
a stochastic gradient descent algorithm for unk - unk systems
on global convergence of sub - gradient descent for sparse graphs
convergence of the gradient descent algorithm for unk problems in the plane
stochastic gradient descent for non - convex optimization
stochastic gradient descent for non - convex optimization
the complexity of the unk stochastic problem and its effect
deep convolutional neural networks for object detection
deep convolutional neural networks for object detection
deep learning for unk
deep convolutional neural networks with unk features
convolutional neural networks for unk detection
object detection using deep neural networks
community detection in social networks : a survey
Outline

- The Basics
- BiRNN as Generators
- Reinforcement Learning
- Variational Autoencoder
Sequence-Level Training

- Motivation: We don't have the ground truth

  In a dialogue system, for example, “The nature of open-domain conversations shows that a variety of replies are plausible, but some are more meaningful, and others are not.”

- Optimize the sequence generator as a whole in terms of external metrics
REINFORCE

- Define an external cost function on a generated sequence
- Generate words by sampling
- Take the derivative of generated samples

\[
L_\theta = - \sum_{w_1^g, \ldots, w_T^g} p_\theta(w_1^g, \ldots, w_T^g) r(w_1^g, \ldots, w_T^g) = -\mathbb{E}_{w_1^g, \ldots, w_T^g \sim p_\theta} r(w_1^g, \ldots, w_T^g)
\]

\[
\partial p(w)=p(w) \partial \log p(w) \text{ because } p(w)=\exp\{\log p(w)\}
\]

\[
\partial J = \sum_{w} \partial p(w|\ldots) r(w) = \sum_{w} p(w)[ \partial \log p(w) ] r(w)
\]

\[
(p_\theta(w_{t+1}|w_t^g, h_{t+1}, c_t) - 1(w_{t+1}^g))
\]

where \(o_t\) is the input to the softmax.
Tricks

- Redefine $r$ as $r - r_{\text{bar}}$
  - $r_{\text{bar}}$: estimated average reward
- Combination of cross-entropy loss and external metrics
  - Simulated annealing
    - First $n$ words: cross entropy loss
    - Last N-n words: external reward
    - Decrease $n$
Outline

- The Basics
- BiRNN as Generators
- Reinforcement Learning
- Variational Autoencoder
Variational Autoencoder

Penalize:
– Reconstruction error
– KL(posterior || prior)
A Very Brief (and maybe oversimplified) Introduction to VAE

$$x = \{x_1, x_2, \cdots\}$$ is observed variables

$$z = \{z_1, z_2, \cdots\}$$ is hidden/latent variables

$$z \rightarrow x$$

—Karl Marx
How God creates the world
Rain $\rightarrow$ Wet

- The marginal distribution of $x$ is defined as

$$p(x) = \int p(x, z) \, dz = \int p(z)p(x|z) \, dz$$

or

$$\log p(x) = KL(q_\phi(z|x)||p_\theta(z|x)) + \mathcal{L}(\theta, \phi; x)$$

where

$$KL(q(z|x)||p(z|x)) = \int q(z|x) \log \left\{ \frac{q(z)}{p(z|x)} \right\} \, dz \quad (\geq 0)$$

$$\mathcal{L}(\theta, \phi; x) = \int q(z|x) \log \left\{ \frac{p(x, z)}{q(z)} \right\} \, dz$$

How man recognizes the world
Wet $\rightarrow$ Rain
How God creates the world  
Rain $\rightarrow$ Wet

- The marginal distribution of $x$ is defined as

$$
\text{log } \int q(z) p(x | z) \, dz
$$

or

$$
\text{log } \int q(z) p(x | z) \, dz
$$

where

$$
KL(q(z) \| p(z | x)) = \int (x, z) \log \left\{ \frac{q(z)}{p(z | x)} \right\} \, dz \quad (\geq 0)
$$

How man recognizes the world  
Wet $\rightarrow$ Rain
The marginal distribution of $x$ is defined as

$$p(x) = \int p(x, z) \, dz = \int p(z)p(x|z) \, dz$$

or

$$\log p(x) = KL(q_\phi(z|x)\|p_\theta(z|x)) + \mathcal{L}(\theta, \phi; x)$$

where

$$KL(q(z|x)\|p(z|x)) = \int q(z|x) \log \left\{ \frac{q(z)}{p(z|x)} \right\} \, dz \quad (\geq 0)$$

$$\mathcal{L}(\theta, \phi; x) = \int q(z|x) \log \left\{ \frac{p(x, z)}{q(z)} \right\} \, dz$$
Factorized Posterior Assumption

Assumption

\[ q(Z) = \prod_{i=1}^{M} q_i(Z_i) \]

Optimize \( \mathcal{L}(q) \) w.r.t a group \( Z_j \) at a time

A case study of Gaussian-Gamma distribution
Variational Autoencoder

- Reparametrization

\[ z|x = g_\phi(x, \epsilon) \]

where \( g(\cdot) \) is a deterministic function and \( \epsilon \sim \mathcal{N} \)

- Use NNs to recognize Z and then reconstruct X

\[
\text{Input} \xrightarrow{\text{NN}} \phi \xrightarrow{\text{Sampling}} \text{Sample} \xrightarrow{\text{NN}} X
\]
Variational Autoencoder

\[ p(x) \geq \mathcal{L}(\theta, \phi; x) = \mathbb{E}_{q_{\phi}(z|x)} \left[ -\log q_{\phi}(z|x) + \log p_{\theta}(x, z) \right] \]
\[ = -KL(q_{\phi}(z|x) \| p(z)) + \mathbb{E}_{q_{\phi}(z|x)}[\log p_{\theta}(x|z)] \]

\[ \mathcal{L}(\theta, \phi; x^{(i)}) \approx \frac{1}{2} \sum_{j=1}^{J} \left( 1 + \log((\sigma_j^{(i)})^2) - (\mu_j^{(i)})^2 - (\sigma_j^{(i)})^2 \right) + \frac{1}{L} \sum_{l=1}^{L} \log p_{\theta}(x^{(i)}|z^{(i,l)}) \]
Variational Autoencoder

Penalize:
- Reconstruction error
- $\text{KL}(\text{posterior} \parallel \text{prior})$
A Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues

Iulian V. Serban*, Alessandro Sordoni*, Ryan Lowe○, Laurent Charlin○, Joelle Pineau○, Aaron Courville* and Yoshua Bengio†
<table>
<thead>
<tr>
<th>Context</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>jamaica blood!! &lt;url&gt; → lol. yes [name]!. [name]! → what’s poppin in toronto homie</td>
<td><strong>VHRED</strong>: nothing much. just chillin. what’s good with you?</td>
</tr>
<tr>
<td></td>
<td><strong>LSTM</strong>: something!!!!!!!!</td>
</tr>
<tr>
<td></td>
<td><strong>HRED</strong>: something!!!!!!</td>
</tr>
<tr>
<td>whhhhhyyyyy is it suddenly cold in tallahassee today?? this is florida! how can i dream of the beach when my fingers are icy cold?? → sittin on deck in sun looking at lake travis in austin! enjoying day b4 birthday</td>
<td><strong>VHRED</strong>: the sun is shining!!!</td>
</tr>
<tr>
<td></td>
<td><strong>LSTM</strong>: i know!!!!!!!!!!!</td>
</tr>
<tr>
<td></td>
<td><strong>HRED</strong>: something!!!!!!</td>
</tr>
<tr>
<td>reaches for [name]! → - approaches him warily - hey there buddy. - swings him into the air - how you doin? → tilts his head and your wariness and reaches for your face [name].</td>
<td><strong>VHRED</strong>: - looks up at you and smiles - you’re so cute. - kisses your check -</td>
</tr>
<tr>
<td></td>
<td><strong>LSTM</strong>: - smiles - i’m fine thank you. how are you?</td>
</tr>
<tr>
<td></td>
<td><strong>HRED</strong>: - grins, looking at you as i look at you -</td>
</tr>
<tr>
<td>i think i want an ipad2. but i already have an ibook. want vs. need. → if you have an ibook. then you deserve an ipad!</td>
<td><strong>VHRED</strong>: that’s what i’m thinking. i’m going to buy a new one.</td>
</tr>
<tr>
<td></td>
<td><strong>LSTM</strong>: i want one!!!!</td>
</tr>
<tr>
<td></td>
<td><strong>HRED</strong>: i don’t know what to do with it. i don’t know what to do.</td>
</tr>
<tr>
<td>ah echt? dan anrztwoord ik volgende keer via twitter als je me smst. → haha doe maar p. goed bezig hoor myrthe d → goed bezig? die snap ik even niet p</td>
<td><strong>VHRED</strong>: haha, ik heb er geen zin in de something, maar ik heb er geen something meer something</td>
</tr>
<tr>
<td></td>
<td><strong>LSTM</strong>: haha something something something</td>
</tr>
<tr>
<td></td>
<td><strong>HRED</strong>: something something something something something something something</td>
</tr>
<tr>
<td>i love katie waissel more than like, anything in life. → i wanna c the show again! i need &lt;number&gt; meet her again! n i want a b-day tweet nxt week doubt that will happen x</td>
<td><strong>VHRED</strong>: me too. i’m going to have to wait for her to come back. i’m so excited.</td>
</tr>
<tr>
<td></td>
<td><strong>LSTM</strong>: i’m sure she will. i’m sure she’ll be fine.</td>
</tr>
<tr>
<td></td>
<td><strong>HRED</strong>: i’m sure she’ll be fine, i’m sure she’ll be fine.</td>
</tr>
<tr>
<td>my math teacher over here talkn bout nickels nd i’m tryna make big money i need to be in calculus → lol</td>
<td><strong>VHRED</strong>: i need to go to the gym and get my hair done lol</td>
</tr>
<tr>
<td></td>
<td><strong>LSTM</strong>: lol</td>
</tr>
<tr>
<td></td>
<td><strong>HRED</strong>: how u been</td>
</tr>
</tbody>
</table>
Thank you for listening!

Q&A and Discussion