Variational Sequential Labelers for Semi-Supervised Learning

Mingda Chen, Qingming Tang, Karen Livescu, Kevin Gimpel
Sequence Labeling

Part-of-Speech (POS) Tagging

This item is a small one and easily missed.

Named Entity Recognition (NER)

EU rejects German call to boycott British lamb.
Overview

❖ Latent-variable generative models for sequence labeling
❖ 0.8 ~ 1% absolute improvements over 8 datasets without structured inference
❖ 0.1 ~ 0.3% absolute improvements from adding unlabeled data
Why latent-variable models?

- Natural way to incorporate unlabeled data
- Ability to disentangle representations via the configuration of latent variables
- Allow us to use neural variational methods
Variational Autoencoder (VAE)

[Kingma and Welling, ICLR’14; Rezende and Mohamed, ICML’15]
Variational Autoencoder (VAE)

[Kingma and Welling, ICLR’14; Rezende and Mohamed, ICML’15]

\[
\log p_\theta(x_{1:T}) \geq \mathbb{E}_{z \sim q_\phi(\cdot|x_{1:T})} \left[ \log p_\theta(x_{1:T}|z) \right] - KL(q_\phi(z|x_{1:T}) \| p_\theta(z))
\]

Reconstruction Loss

KL divergence

Evidence Lower Bound (ELBO)
Conditional Variational Autoencoder

Observation $x_{1:T}$

Latent variable $z$

Given context $C$
Conditional Variational Autoencoder

\[
\log p_\theta(x_{1:T}\mid c) \geq \mathbb{E}_{z \sim q_\phi(\cdot \mid x_{1:T}, c)} \left[ \log p_\theta(x_{1:T}\mid z) \right] - KL(q_\phi(z\mid x_{1:T}, c)\parallel p_\theta(z\mid c))
\]
$x - t$

The input words other than the word at position $t$
$x - t$

The input words other than the word at position $t$

This item is a small one and easily missed.

$x - 1$
$x_t$

The input words other than the word at position $t$

This item is a small one and easily missed.
Variational Sequential Labeler (VSL)
Variational Sequential Labeler (VSL)

\[
\log p_\theta(x_t|x_{-t}) \geq \mathbb{E}_{z_t \sim q_\phi(\cdot | x_{1:T}, t)} \left[ \log p_\theta(x_t | z_t) \right] - KL(q_\phi(z_t | x_{1:T}, t) \| p_\theta(z_t | x_{-t}))
\]

ELBO
Variational Sequential Labeler (VSL)

\[
\log p_\theta(x_t|x_{-t}) \geq \mathbb{E}_{z_t \sim q_\phi(\cdot|x_{1:T}, t)} [\log p_\theta(x_t|z_t)] - KL(q_\phi(z_t|x_{1:T}, t) \parallel p_\theta(z_t|x_{-t}))
\]
Variational Sequential Labeler (VSL)

\[
\log p_\theta(x_t|x_{-t}) \geq \mathbb{E}_{z_t \sim q_\phi(\cdot|x_{1:T}, t)} \left[ \log p_\theta(x_t|z_t) \right] - KL(q_\phi(z_t|x_{1:T}, t) || p_\theta(z_t|x_{-t}))
\]
$\log p_\theta(x_t|x_{-t}) \geq \mathbb{E}_{z_t \sim q_\phi(\cdot|x_{1:T}, t)}[\log p_\theta(x_t|z_t)] - KL(q_\phi(z_t|x_{1:T}, t)\| p_\theta(z_t|x_{-t}))$
Variational Sequential Labeler (VSL)

\[
\log p_\theta(x_t | x_{-t}) \geq \mathbb{E}_{z_t \sim q_\phi(\cdot | x_{1:T}, t)} \left[ \log p_\theta(x_t | z_t) \right] - KL(q_\phi(z_t | x_{1:T}, t) \| p_\theta(z_t | x_{-t}))
\]
VSL: Training and Testing

Training

❖ Maximize $\text{ELBO} - \alpha \cdot \text{CL}$ where $\alpha$ is a hyperparameter
❖ Use one sample from Gaussian distribution using reparameterization trick

Testing

❖ Use the mean of Gaussian distribution
Variants of VSL

VSL-G

$z_t$

$\mathbf{x}_t$

Position of classifier
Variants of VSL

\[ \mathbf{x}_t \]

\[ \mathbf{z}_t \]

\[ \mathbf{x}_{-t} \]

VSL-G

Stands for “Gaussian”
Variants of VSL

- VSL-G
  - Stands for “Gaussian”

- VSL-GG-Flat
  - Position of classifier

VSL-G

Stands for “Gaussian”
Variants of VSL

- **VSL-G**
  - Stands for “Gaussian”

- **VSL-GG-Flat**
  - Position of classifier
Variants of VSL

- VSL-G
  - Stands for “Gaussian”

- VSL-GG-Flat
  - Position of classifier
Variants of VSL

- VSL-G
  - Stands for “Gaussian”
- VSL-GG-Flat
- VSL-GG-Hier

Position of classifier
Variants of VSL

- VSL-G
  - Stands for "Gaussian"

- VSL-GG-Flat

- VSL-GG-Hier

Position of classifier
Experiments

❖ Twitter POS Dataset
  ➢ Subset of 56 million English tweets as unlabeled data
  ➢ 25 tags

❖ Universal Dependencies POS Datasets
  ➢ 20% of original training set as labeled data
  ➢ 50% of original training set as unlabeled data
  ➢ 6 languages
  ➢ 17 tags

❖ CoNLL 2003 English NER Dataset
  ➢ 10% of original training set as labeled data
  ➢ 50% of original training set as unlabeled data
  ➢ BIOES labeling scheme
Results

**Accuracy**
- Twitter: 90.8 (Effect of unlabeled data), 91.1 (VSL-G (supervised)), 91.4 (VSL-GG-Flat (supervised)), 91.6 (BiGRU)
- NER: 87.6 (Effect of unlabeled data), 87.8 (VSL-G (supervised)), 88.0 (VSL-GG-Flat (supervised)), 88.4 (BiGRU)

**F1 Score**
- Twitter: 91.2 (VSL-G (supervised)), 91.5 (VSL-GG-Flat (supervised)), 91.9 (BiGRU)
- NER: 87.9 (VSL-G (supervised)), 88.1 (VSL-GG-Flat (supervised)), 88.6 (BiGRU)
Universal Dependencies POS

<table>
<thead>
<tr>
<th>Language</th>
<th>BiGRU</th>
<th>VSL-G</th>
<th>VSL-GG-Flat</th>
<th>VSL-GG-Hier</th>
</tr>
</thead>
<tbody>
<tr>
<td>French</td>
<td>95.9</td>
<td>96.1</td>
<td>96.4</td>
<td></td>
</tr>
<tr>
<td>German</td>
<td>92.6</td>
<td>92.8</td>
<td>93.3</td>
<td></td>
</tr>
<tr>
<td>Indonesian</td>
<td>92.2</td>
<td>92.3</td>
<td>92.4</td>
<td>92.8</td>
</tr>
<tr>
<td>Spanish</td>
<td>94.7</td>
<td>94.8</td>
<td>95.0</td>
<td>95.3</td>
</tr>
</tbody>
</table>
t-SNE Visualization

- Each point represents a word token
- Color indicates gold standard POS tag in Twitter dev set

BiGRU baseline
t-SNE Visualization

y (label) variable

VSL-GG-Hier

z variable

VSL-GG-Flat
Effect of Position of Classification Loss

VSL-GG-Hier

Position of classifier
Effect of Position of Classification Loss

VSL-GG-Hier

VSL-GG-Hier with classifier on $z_t$

Position of classifier
Effect of Position of Classification Loss

VSL-GG-Hier

VSL-GG-Hier with classifier on $z_t$

VSL-GG-Hier-z

Position of classifier
Effect of Position of Classification Loss

- Twitter: VSL-GG-Hier-z (91.1) vs. VSL-GG-Hier (91.6)
- NER: VSL-GG-Hier-z (87.8) vs. VSL-GG-Hier (88.4)
- UD average: VSL-GG-Hier-z (94.4) vs. VSL-GG-Hier (95.0)
Effect of Position of Classification Loss

Hierarchical structure is only helpful when classification loss and reconstruction loss are attached to different latent variables.
Effect of Variational Regularization (VR)

VR

\[ \text{KL divergence between approximated posterior and prior} \]

\[ + \]

Randomness in the latent space
Effect of VR

![Graph showing the effect of VR on accuracy and F1 score for Twitter and NER tasks.](image-url)
Effect of Unlabeled data

❖ Evaluate VSL-GG-Hier on Twitter dataset
❖ Subsample unlabeled data from 56 million tweets
❖ Vary the number of unlabeled data
Effect of Unlabeled data
Summary

❖ We introduced VSLs for semi-supervised learning
❖ Best VSL uses multiple latent variable and arranged in hierarchical structure
❖ Hierarchical structure is only helpful when classification loss and reconstruction loss are attached to different latent variables
❖ VSLs show consistent improvements across 8 datasets over a strong baseline
Thank you!