Structured Inference Networks for Nonlinear State Space Models

Rahul G. Krishnan, Uri Shalit, David Sontag
New York University
30 Sep 2016

Chris Cremer
CSC2541
Nov 4 2016
Overview

• VAE
• Gaussian State Space Models
• Inference Network
• Results
Recap - VAE

Generative Model
\[ p_\theta(x|z) = \mathcal{N}(\mu_\theta(z), \Sigma_\theta(z)) \]
\[ p_\theta(z) = \mathcal{N}(0, I) \]

Recognition Network
\[ q_\phi(z|x) = \mathcal{N}(\mu_\phi(x), \Sigma_\phi(x)) \]

Use MLP to model the mean and covariance

Learning and Inference → Maximize Lower Bound
\[ \log p_\theta(x) \geq \mathbb{E}_{q_\phi(z|x)} \left[ \log p_\theta(x|z) \right] - \text{KL}(q_\phi(z|x) || p_\theta(z)) \]

- **Reconstruction Loss**
  - Calculated by sampling \( q_\phi(z|x) \) with reparameterization trick

- **Divergence from Prior**
  - Analytic equation
Gaussian State Space Models

Generative Model

\[ z_t \sim \mathcal{N}(G_\alpha(z_{t-1}, \Delta_t), S_\beta(z_{t-1}, \Delta_t)) \text{ (Transition)} \quad x_t \sim \mathcal{N}(F_\kappa(z_t)) \text{ (Emission)} \]

- HMM with continuous hidden state
- If transition and emission are linear Gaussian, then we can do inference analytically (Kalman Filter)
- Deep Markov Model:
  - Transition and emission distributions are parametrized by MLPs
  - Inference: VAE
Inference – Factorized Lower Bound

\[
\log p_\theta(x) \geq \mathbb{E}_{q_\phi(z|x)} \left[ \log p_\theta(x|z) \right] - \text{KL}(q_\phi(z|x)||p_\theta(z))
\]

- **Reconstruction Loss**: Calculated by sampling \(q_\phi(z|x)\) with reparameterization trick.
- **Divergence from Prior**: Analytic equation.

\[
\log p_\theta(\bar{x}) \geq \mathbb{E}_{q_\phi(\bar{z}|\bar{x})} \left[ \log p_\theta(\bar{x}|\bar{z}) \right] - \text{KL}(q_\phi(\bar{z}|\bar{x})||p_\theta(\bar{z}))
\]

- **Reconstruction Loss**: Calculated by sampling \(q_\phi(\bar{z}|\bar{x})\) with reparameterization trick.
- **Divergence from Prior**: Analytic equation.

\[
= \sum_{t=1}^{T} \mathbb{E}_{q_\phi(z_t|\bar{x})} \left[ \log p_\theta(x_t|z_t) \right] - \text{KL}(q_\phi(z_t|\bar{x})||p_\theta(z_t)) - \sum_{t=2}^{T} \mathbb{E}_{q_\phi(z_{t-1},z_t|\bar{x})} \left[ \text{KL}(q_\phi(z_t|z_{t-1},\bar{x})||p_\theta(z_t|z_{t-1})) \right]
\]

- **Reconstruction Loss**: Calculated by sampling \(q_\phi(z_t|\bar{x})\) with reparameterization trick.
- **Divergence from Prior**: Analytic equation.
- **Divergence from Prior**: Analytic equation.
Inference Networks

- Evaluate possibilities for the inference networks
  - Mean-Field Model (MF) vs Structured Model (ST)
  - Observations from past (L), future (R), or both (LR)
- Combiner Function: MLP that combines the previous state with the RNN output

\[
\log p_\theta(\mathbf{x}) \geq \sum_{t=1}^{T} \mathbb{E}_{q_\phi(z_t|\mathbf{x})} \left[ \log p_\theta(x_t|z_t) \right] - \text{KL}(q_\phi(z_1|\mathbf{x}) \parallel p_\theta(z_1)) - \sum_{t=2}^{T} \mathbb{E}_{q_\phi(z_{t-1}|\mathbf{x})} \left[ \text{KL}(q_\phi(z_t|z_{t-1}, \mathbf{x}) \parallel p_\theta(z_t|z_{t-1})) \right]
\]

<table>
<thead>
<tr>
<th>Inf. Network</th>
<th>Variational Approximation</th>
<th>Implementation</th>
</tr>
</thead>
<tbody>
<tr>
<td>MF-LR</td>
<td>(q(z_t</td>
<td>x_1, \ldots x_T))</td>
</tr>
<tr>
<td>MF-L</td>
<td>(q(z_t</td>
<td>x_1, \ldots x_t))</td>
</tr>
<tr>
<td>ST-L</td>
<td>(q(z_t</td>
<td>z_{t-1}, x_1, \ldots x_t))</td>
</tr>
<tr>
<td>DKS</td>
<td>(q(z_t</td>
<td>z_{t-1}, x_t, \ldots x_T))</td>
</tr>
<tr>
<td>ST-LR</td>
<td>(q(z_t</td>
<td>z_{t-1}, x_1, \ldots x_T))</td>
</tr>
</tbody>
</table>

Deep Kalman Smoothing (ST-R)
Inference Networks Results

Polyphonic music data (Boulanger-Lewandowski et al., 2012)
- Sequence of 88-dimensional binary vectors corresponding to the notes of a piano
- Report held-out negative log-likelihood (NLL)

Results:
- **ST-LR** and **DKS** substantially outperform **MF-LR** and **ST-L**
  - Due to previous state \((z_{t-1})\) and future observations \((x_t, \ldots, x_T)\)
- \(z_{t-1}\) summarizes past observations \((x_1, \ldots, x_t)\)
- **DKS** network has half the parameters of the **ST-LR**
Model Comparison

Results:
• Increasing the complexity of the generative model improves the likelihood (DMM vs DMM-Aug)
• DMM-Aug (DKS) obtains better results on all datasets (except LV-RNN on JSB)
• Demonstrates the inference network’s ability to learn powerful generative models

<table>
<thead>
<tr>
<th>Methods</th>
<th>JSB</th>
<th>Nottingham</th>
<th>Piano</th>
<th>Musedata</th>
</tr>
</thead>
<tbody>
<tr>
<td>DMM</td>
<td>6.388</td>
<td>2.770</td>
<td>7.835</td>
<td>6.831</td>
</tr>
<tr>
<td></td>
<td>(6.926)</td>
<td>(2.964)</td>
<td>(7.980)</td>
<td>(6.989)</td>
</tr>
<tr>
<td></td>
<td>{6.856}</td>
<td>{2.954}</td>
<td>{8.246}</td>
<td>{6.203}</td>
</tr>
<tr>
<td></td>
<td>(6.773)</td>
<td>(2.856)</td>
<td>(7.721)</td>
<td>(6.476)</td>
</tr>
<tr>
<td></td>
<td>{6.692}</td>
<td>{2.872}</td>
<td>{8.025}</td>
<td>{5.766}</td>
</tr>
<tr>
<td>HMSBN</td>
<td>(8.0473)</td>
<td>(5.2354)</td>
<td>(9.563)</td>
<td>(9.741)</td>
</tr>
<tr>
<td></td>
<td>{7.9970}</td>
<td>{5.1231}</td>
<td>{9.786}</td>
<td>{8.9012}</td>
</tr>
<tr>
<td>STORN</td>
<td>6.91</td>
<td>2.85</td>
<td>7.13</td>
<td>6.16</td>
</tr>
<tr>
<td>RNN</td>
<td>8.71</td>
<td>4.46</td>
<td>8.37</td>
<td>8.13</td>
</tr>
<tr>
<td>TSBN</td>
<td>{7.48}</td>
<td>{3.67}</td>
<td>{7.98}</td>
<td>{6.81}</td>
</tr>
<tr>
<td>LV-RNN</td>
<td>3.99</td>
<td>2.72</td>
<td>7.61</td>
<td>6.89</td>
</tr>
</tbody>
</table>
EHR Patient Data

• What would happen if the patient received diabetic medication or not?
Conclusion

• Structured Inference Networks for Nonlinear State Space Models

$$\mathcal{L}(\mathcal{X}; (\theta, \phi)) = \sum_{t=1}^{T} \mathbb{E}_{q_{\phi}(z_{t} | \mathcal{X})} \left[ \log p_{\theta}(x_{t} | z_{t}) \right] - \text{KL}(q_{\phi}(z_{1} | \mathcal{X}) || p_{\theta}(z_{1}))$$

$$- \sum_{t=2}^{T} \mathbb{E}_{q_{\phi}(z_{t-1} | \mathcal{X})} \left[ \text{KL}(q_{\phi}(z_{t} | z_{t-1}, \mathcal{X}) || p_{\theta}(z_{t} | z_{t-1})) \right].$$

VAE for sequential data
Questions?