Correlated Variational Auto-Encoders

Da Tang\textsuperscript{1}  Dawen Liang\textsuperscript{2}  Tony Jebara\textsuperscript{1,2}  Nicholas Ruozzi\textsuperscript{3}

\textsuperscript{1}Columbia University
\textsuperscript{2}Netflix Inc.
\textsuperscript{3}The University of Texas at Dallas

June 11, 2019
Variational Auto-Encoders (VAEs)

▶ Learn stochastic low dimensional latent representations for high dimensional data:

\[ q_\lambda(z \mid x) \quad p_\theta(x \mid z) \]

Data \( x \) \hspace{1cm} Latent representation \( z \) \hspace{1cm} Reconstruction \( \hat{x} \)

\[
L(\lambda, \theta) = \sum_{i=1}^{n} \left[ \mathbb{E}_{q_\lambda(z \mid x_i)} \left[ \log p_\theta(x_i \mid z_i) \right] - \text{KL}(q_\lambda(z \mid x_i) \mid \mid p_0(z_i)) \right].
\]
Variational Auto-Encoders (VAEs)

- Learn stochastic low dimensional latent representations for high dimensional data:

- Model the likelihood and the inference distribution independent among data points in the objective (the ELBO):

\[
\mathcal{L}(\lambda, \theta) = \sum_{i=1}^{n} (\mathbb{E}_{q_{\lambda}(z_i| x_i)} [\log p_{\theta}(x_i | z_i)] - KL(q_{\lambda}(z_i| x_i) || p_0(z_i))).
\]
Motivation

- VAEs assume the prior is i.i.d. among data points.
Motivation

- VAEs assume the prior is i.i.d. among data points.

- If we know information about correlations between data points (e.g., networked data), we can incorporate it into the generative process of VAEs.
Given an undirected correlation graph $G = (V, E)$ for data $x_1, \ldots, x_n$, where $V = \{v_1, \ldots, v_n\}$ and $E = \{(v_i, v_j) : x_i$ and $x_j$ are correlated$\}$. 

Directly applying a correlated prior of $z = (z_1, \ldots, z_n)$ on general undirected graphs is hard.
Given an undirected correlation graph $G = (V, E)$ for data $x_1, \ldots, x_n$, where $V = \{v_1, \ldots, v_n\}$ and $E = \{(v_i, v_j) : x_i$ and $x_j$ are correlated$\}$.

Directly applying a correlated prior of $z = (z_1, \ldots, z_n)$ on general undirected graphs is hard.
Define the prior of $z$ as a uniform mixture over all Maximal Acyclic Subgraphs of $G$:

$$p_{0}^{corr}(z) = \frac{1}{|\mathcal{A}_G|} \sum_{G'=(V,E') \in \mathcal{A}_G} p_{0}^{G'}(z).$$
We apply a uniform mixture over acyclic subgraphs since we have closed-form correlated distributions for acyclic graphs:

\[
p_0^{G'}(z) = \prod_{i=1}^{n} p_0(z_i) \prod_{(v_i, v_j) \in E'} \frac{p_0(z_i, z_j)}{p_0(z_i)p_0(z_j)}.
\]
Correlated Priors

We apply a uniform mixture over acyclic subgraphs since we have closed-form correlated distributions for acyclic graphs:

\[
p_{0}^{G'}(z) = \prod_{i=1}^{n} p_{0}(z_i) \prod_{(v_i, v_j) \in E'} \frac{p_{0}(z_i, z_j)}{p_{0}(z_i)p_{0}(z_j)}.
\]
We apply a uniform mixture over acyclic subgraphs since we have closed-form correlated distributions for acyclic graphs:

\[ p^G_G(z) = \prod_{i=1}^{n} p_0(z_i) \prod_{(v_i, v_j) \in E'} \frac{p_0(z_i, z_j)}{p_0(z_i)p_0(z_j)}. \]
Inference with a Weighted Objective

Define a new ELBO for general graphs:

$$\log p_\theta(x) = \log \mathbb{E}_{p_0^{\text{corr}}(z)}[p_\theta(x|z)]$$

$$\geq \frac{1}{|A_G|} \sum_{G' \in A_G} \left( \mathbb{E}_{q_\lambda^{G'}(z|x)}[\log p_\theta(x|z)] - \text{KL}(q_\lambda^{G'}(z|x) || p_0^{G'}(z)) \right)$$

$$:= \mathcal{L}(\lambda, \theta)$$

where $q_\lambda^{G'}$ is defined in the same way as for the priors:

$$q_\lambda^{G'}(z) = \prod_{i=1}^n q_\lambda(z_i|x_i) \prod_{(v_i, v_j) \in E'} \frac{q_\lambda(z_i, z_j|x_i, x_j)}{q_\lambda(z_i|x_i)q_\lambda(z_j|x_j)}.$$
The loss function is intractable due to the potentially exponential many subgraphs.

The weighted loss is tractable. The weights can be computed from the pseudo-inverse of the Laplacian matrix of $G$. 
The loss function is intractable due to the potentially exponential many subgraphs.

Represent the average loss on acyclic subgraphs as a weighted average loss on edges.

The weighted loss is tractable. The weights can be computed from the pseudo-inverse of the Laplacian matrix of $G$. 

Inference with a Weighted Objective
The loss function is intractable due to the potentially exponential many subgraphs.

Represent the average loss on acyclic subgraphs as a weighted average loss on edges.

The weighted loss is tractable. The weights can be computed from the pseudo-inverse of the Laplacian matrix of $G$. 
## Empirical Results

**Table: Link prediction test NCRR**

<table>
<thead>
<tr>
<th>Method</th>
<th>Test NCRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>VAE</td>
<td>0.0052 ± 0.0007</td>
</tr>
<tr>
<td>GraphSAGE</td>
<td>0.0115 ± 0.0025</td>
</tr>
<tr>
<td>CVAE</td>
<td><strong>0.0171 ± 0.0009</strong></td>
</tr>
</tbody>
</table>

**Table: Spectral clustering scores**

<table>
<thead>
<tr>
<th>Method</th>
<th>NMI scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>VAE</td>
<td>0.0031 ± 0.0059</td>
</tr>
<tr>
<td>GraphSAGE</td>
<td>0.0945 ± 0.0607</td>
</tr>
<tr>
<td>CVAE</td>
<td><strong>0.2748 ± 0.0462</strong></td>
</tr>
</tbody>
</table>

**Table: User matching test RR**

<table>
<thead>
<tr>
<th>Method</th>
<th>Test RR</th>
</tr>
</thead>
<tbody>
<tr>
<td>VAE</td>
<td>0.3498 ± 0.0167</td>
</tr>
<tr>
<td>CVAE</td>
<td><strong>0.7129 ± 0.0096</strong></td>
</tr>
</tbody>
</table>
CVAE accounts for correlations between data points that are known *a priori*. It can adopt a correlated variational density function to achieve a better variational approximation.

Future work includes extending to correlated VAEs with higher-order correlations.
CVAE accounts for correlations between data points that are known \textit{a priori}. It can adopt a correlated variational density function to achieve a better variational approximation.

Future work includes extending to correlated VAEs with higher-order correlations.
Thanks!

Poster #219