Outline

Reviewed paper

- Ladder architecture
  - *Deconstructing the Ladder Network Architecture.* (ICML 2016)

- Semi-supervised learning
  - *Semi-Supervised Learning with Ladder Networks.* (NIPS 2015)

- Generative learning
  - *Ladder Variational Autoencoders.* (NIPS 2016)
  - *Learning Hierarchical Features from Generative Models.* (ICML 2017)

(From a greeting card found in Target)
Vanilla Ladder Network

Key Ideas

- Denoising representation at every level of the model
- With skip connections the decoder can recover details discarded by the encoder
Semi-supervised learning

Notations

- $l \in \{1, \ldots, L\}$: layer index
- $z^{(l)}$: clean latent code
- $\tilde{z}^{(l)}$: noisy latent code
- $\hat{z}^{(l)}$: de-noised latent code

- $\tilde{x}, \{\tilde{z}^{(l)}\}, \tilde{y} = \text{Enc}(x, \text{Noisy})$
- $x, \{z^{(l)}\}, y = \text{Enc}(x, \text{Clean})$
- $\{\hat{z}^{(l)}\} = \text{Dec}(\{\tilde{z}^{(l)}\})$
Algorithm::Encoder

- $\tilde{h}(0) = \tilde{z}(0) = F(x, \xi^{(0)}), h(0) = z(0) = F_0(x, \emptyset)$
- For $l = 1, \ldots, L$
  - $[\tilde{z}^{(l)}, z^{(l)}] = \text{BN}(F_l([\tilde{h}^{(l-1)}, h^{(l-1)}], [\xi^{(l)}, \emptyset]))$
  - $[\tilde{h}^{(l)}, h^{(l)}] = \text{Act}(\gamma^{(l)} \odot ([\tilde{z}^{(l)}, z^{(l)}] + \beta^{(l)}))$
- $P(\tilde{y}|x) = H(\tilde{h}^{(L)})$
Algorithm::Decoder

- \( u^L = \text{BN}(\tilde{h}^{(L)}) \)
- For \( l = L - 1, \ldots, 0 \):
  - \( u^{(l)} := \text{BN}(G(\hat{z}^{(l+1)})) \), \( \hat{v}^{(l)} = T(\tilde{z}^{(l)}, u^{(l)}) \)
  - \( \tilde{z}^{(l)} = \text{BN}(\hat{v}^{(l)}, \mu_{\text{clean}}, \sigma_{\text{clean}}) \)
- Likelihood loss: \( \ell_{\text{lik}}(x, y) = -\log P(\tilde{y} = y|x) \)
- Reconstruction loss: \( \ell_{\text{recon}}(x) = \sum_{l=0}^{L} \lambda_l \| z^{(l)} - \hat{z}^{(l)} \|_2^2 \)

Semi-supervised setting

- \( \mathcal{D} \): full dataset, \( \mathcal{D}_l \): labeled subset
- \( \text{Loss} = \mathbb{E}_{(X,Y) \sim \mathcal{D}_l}[\ell_{\text{lik}}(X, Y)] + \lambda \mathbb{E}_{X \sim \mathcal{D}}[\ell_{\text{recon}}(X)] \)
## Basic Concepts of Variational Auto-encoder

### Inference model
- \( q_\phi(z|x) = \prod_{i=1}^{L} q_\phi(z_i|z_{i-1}) \)
- \( q_\phi(z_i|z_{i-1}) \sim \mathcal{N}(\mu_{q,i}(z_{i-1}), \sigma_{q,i}^2(z_{i-1})) \)
- \( z_0 = x \)

### Generative model
- \( p_\theta(x, z) = \prod_{i=0}^{L} p_\phi(z_i|z_{i+1}) \)
- \( p_\phi(z_i|z_{i+1}) = \mathcal{N}(\mu_{p,i}(z_{i+1}), \sigma_{p,i}^2(z_{i+1})) \)
- \( z_0 = x, z_{L+1} = \emptyset \)

### Evidence lower bound:
- \( \log p(x) \geq \text{ELBO} := \mathbb{E}_{q_\phi(z|x)} \left[ \frac{p(x,z)}{q_\phi(z|x)} \right] \)
Ladder Network Semi-supervision Generative-ladder Models

Ladder Variational Auto-encoder (LVAE)

LVAE Inference Model

- **Forward pass**
  - \( h_l = f_l(h_{l-1}) \)
  - \((\hat{\mu}_{q,l}, \hat{\sigma}_{q,l}^2) = g_l(h_l)\)

- **Backward sampling**
  - \( q_{\phi}(z|x) = \prod_{l=1}^{L} q_{\phi}(z_l|z_{l+1}, x) \)
  - \( \mu_l = \frac{\hat{\mu}_{q,l} \hat{\sigma}_{q,l}^{-2} + \mu_{p,l} \sigma_{p,l}^{-2}}{\hat{\sigma}_{q,l}^{-2} + \sigma_{p,l}^{-2}}, \sigma_{l}^2 = \frac{1}{\hat{\sigma}_{q,l}^{-2} + \sigma_{p,l}^{-2}} \)
  - Precision weighted Gaussian
  - \( q_{\phi}(z_l|z_{l+1}, x) \sim \mathcal{N}(\mu_l, \sigma_{l}^2) \)
Hacks

Importance Weighting

- \( L_K = \mathbb{E}_{q_\phi(z^{(1)}|x)} \cdots \mathbb{E}_{q_\phi(z^{(K)}|x)} \left[ \log \left( K^{-1} \sum_{k=1}^{K} \frac{p(x,z^{(k)})}{q_\phi(z^{(k)}|x)} \right) \right] \)
- \( \log p(x) \leftarrow \cdots \geq L_K \geq \cdots \geq L_1 = \text{ELBO} \)

Warm-up

- Problem: many units collapsed to prior early in training
- Solution: anneal the KL loss term on the prior
- Limitation: not directly compatible with the IW scheme
Variational Ladder Auto-encoder (VLAЕ)

Basic Argument

- Given sufficient representation power, distribution can be recovered without hierarchical latent structure.
  - Gibbs chain \( \{Z_1 \sim q(z_1|x), X \sim p(x|z_1)\} \) converges to \( p(x) \)

Key Assumption

- Latent codes are independent of each other
  - \( p(x, z) = p(x|z) \prod_{l=1}^{L} p(z_l) \)
Variational Ladder Auto-encoder (VLAE)

VLAE Solution

- Replace the hierarchical dependency with transformation complexity
Variational Ladder Auto-encoder (VLAE)

Generative Model

- $h_L = f_L(z_L)$
- $h_l = f_l(h_{l+1}, z_l)$, $l = 1, \ldots, L - 1$
- $x \sim G(\cdot; f_0(h_1))$
- $f_l(h_{l-1}, z_l) = u_l([h_{l+1}, v_l(z_l)])$, $u_l(\cdot)$, $v_l(\cdot)$ neural nets

Inference Model

- $h_l = g_l(h_{l-1})$
- $z_l = \mu_l(h_l) + \sigma(h_l) \odot \epsilon_l$
## Table 1: A collection of previously reported MNIST test errors in the permutation invariant setting followed by the results with the Ladder network. * = SVM. Standard deviation in parenthesis.

<table>
<thead>
<tr>
<th>Test error % with # of used labels</th>
<th>100</th>
<th>1000</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Semi-sup. Embedding [15]</td>
<td>16.86</td>
<td>5.73</td>
<td>1.5</td>
</tr>
<tr>
<td>Transductive SVM [from 15]</td>
<td>16.81</td>
<td>5.38</td>
<td>1.40*</td>
</tr>
<tr>
<td>MTC [16]</td>
<td>12.03</td>
<td>3.64</td>
<td>0.81</td>
</tr>
<tr>
<td>Pseudo-label [17]</td>
<td>10.49</td>
<td>3.46</td>
<td></td>
</tr>
<tr>
<td>AtlasRBF [18]</td>
<td>8.10 (± 0.95)</td>
<td>3.68 (± 0.12)</td>
<td>1.31</td>
</tr>
<tr>
<td>DGN [19]</td>
<td>3.33 (± 0.14)</td>
<td>2.40 (± 0.02)</td>
<td>0.96</td>
</tr>
<tr>
<td>DBM, Dropout [20]</td>
<td></td>
<td></td>
<td>0.79</td>
</tr>
<tr>
<td>Adversarial [21]</td>
<td></td>
<td></td>
<td>0.78</td>
</tr>
<tr>
<td>Virtual Adversarial [22]</td>
<td>2.12</td>
<td>1.32</td>
<td>0.64 (± 0.03)</td>
</tr>
<tr>
<td>Baseline: MLP, BN, Gaussian noise</td>
<td>21.74 (± 1.77)</td>
<td>5.70 (± 0.20)</td>
<td>0.80 (± 0.03)</td>
</tr>
<tr>
<td>Γ-model (Ladder with only top-level cost)</td>
<td>3.06 (± 1.44)</td>
<td>1.53 (± 0.10)</td>
<td>0.78 (± 0.03)</td>
</tr>
<tr>
<td>Ladder, only bottom-level cost</td>
<td>1.09 (± 0.32)</td>
<td>0.90 (± 0.05)</td>
<td>0.59 (± 0.03)</td>
</tr>
<tr>
<td>Ladder, full</td>
<td><strong>1.06 (± 0.37)</strong></td>
<td><strong>0.84 (± 0.08)</strong></td>
<td><strong>0.57 (± 0.02)</strong></td>
</tr>
</tbody>
</table>

**Classification error**
Reported Results :: LVAE :: MNIST

<table>
<thead>
<tr>
<th>VAE</th>
<th>VAE +BN</th>
<th>VAE +BN+WU</th>
<th>LVAE +BN+WU</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Image" /></td>
<td><img src="image2.png" alt="Image" /></td>
<td><img src="image3.png" alt="Image" /></td>
<td><img src="image4.png" alt="Image" /></td>
</tr>
<tr>
<td><img src="image5.png" alt="Image" /></td>
<td><img src="image6.png" alt="Image" /></td>
<td><img src="image7.png" alt="Image" /></td>
<td><img src="image8.png" alt="Image" /></td>
</tr>
<tr>
<td><img src="image9.png" alt="Image" /></td>
<td><img src="image10.png" alt="Image" /></td>
<td><img src="image11.png" alt="Image" /></td>
<td><img src="image12.png" alt="Image" /></td>
</tr>
<tr>
<td><img src="image13.png" alt="Image" /></td>
<td><img src="image14.png" alt="Image" /></td>
<td><img src="image15.png" alt="Image" /></td>
<td><img src="image16.png" alt="Image" /></td>
</tr>
</tbody>
</table>

Latent representation
Unit informativeness
(measured in KL, darker is better)
Reported Results :: LVAE :: MNIST

<table>
<thead>
<tr>
<th>Model Configuration</th>
<th>Log-likelihood ≤ \log p(x)</th>
</tr>
</thead>
<tbody>
<tr>
<td>VAE 1-layer + NF [18]</td>
<td>-85.10</td>
</tr>
<tr>
<td>IWAE, 2-layer + IW=1 [3]</td>
<td>-85.33</td>
</tr>
<tr>
<td>IWAE, 2-layer + IW=50 [3]</td>
<td>-82.90</td>
</tr>
<tr>
<td>VAE, 2-layer + VGP [21]</td>
<td>-81.90</td>
</tr>
<tr>
<td>LVAE, 5-layer</td>
<td>-82.12</td>
</tr>
<tr>
<td>LVAE, 5-layer + finetuning</td>
<td>-81.84</td>
</tr>
<tr>
<td>LVAE, 5-layer + finetuning + IW=10</td>
<td>-81.74</td>
</tr>
</tbody>
</table>

Log-likelihood vs. training epochs graph.

Chenyang Tao

Ladder networks
Reported Results :: VLAE :: MNIST

Latent representations
Reported Results :: VLAE :: CelebA

Latent representations
Be Cautious

**LVAE**
- Did not report generated samples.
- Heuristic implementation choices without any arguments.

**VLAE**
- Did not report test likelihoods.
- Some arguments are nebulous.

**More General Principled Argument**