10807
Topics in Deep Learning
Russ Salakhutdinov
Machine Learning Department
rsalakhu@cs.cmu.edu
http://www.cs.cmu.edu/~rsalakhu/10807_2016/

Variational Autoencoders
Gaussian Policy: Continuous Actions

- Remember stochastic policy
  \[ \pi_\theta(s, a) = \mathbb{P}[a \mid s, \theta] \]
- In continuous action spaces, a Gaussian policy is natural
- Mean is a linear combination of state features
  \[ \mu(s) = \phi(s)^\top \theta \]
  Nonlinear extension: replace \( \phi(s) \) with a deep neural network with trainable weights \( w \)
- Variance may be fixed \( \sigma^2 \), or can also parameterized
- Policy is Gaussian
  \[ a \sim \mathcal{N}(\mu(s), \sigma^2) \]
Multimodal Outputs

- Remember stochastic policy

\[ \pi_\theta(s, a) = \mathbb{P}[a \mid s, \theta] \]

- But what if stochastic policy has multiple modes?

- Model-based RL: Dynamics of the environment can be multimodal.

\[ P(s_{t+1} \mid a_t, s_t) \]
Helmholtz Machines


Diagram:

- Approximate Inference
- Generative Process
- Input data
- \( P(h^3) \)
- \( P(h^2|h^3) \)
- \( P(h^1|h^2) \)
- \( P(x|h^1) \)
Variational Autoencoders (VAEs)

- The VAE defines a generative process in terms of ancestral sampling through a cascade of hidden stochastic layers:

\[
p(x|\theta) = \sum_{h^1, \ldots, h^L} p(h^L|\theta)p(h^{L-1}|h^L, \theta) \cdots p(x|h^1, \theta)
\]

Each term may denote a complicated nonlinear relationship

- \(\theta\) denotes parameters of VAE.
- \(L\) is the number of **stochastic** layers.
- Sampling and probability evaluation is tractable for each \(p(h^\ell|h^{\ell+1})\).
Variational Autoencoders (VAEs)

• The VAE defines a generative process in terms of ancestral sampling through a cascade of hidden stochastic layers:

\[ p(x | \theta) = \sum_{h^1, \ldots, h^L} p(h^L | \theta)p(h^{L-1} | h^L, \theta) \cdots p(x | h^1, \theta) \]

› Given state, we can generate a distribution over actions:

\[ \pi_\theta(s, a) = \mathbb{P}[a | s, \theta] \]

› Conditional VAE: neural networks with stochastic and deterministic layers
VAE: Example

• The VAE defines a generative process in terms of ancestral sampling through a cascade of hidden stochastic layers:

\[
p(x|\theta) = \sum_{h^1, h^2} p(h^2|\theta)p(h^1|h^2, \theta)p(x|h^1, \theta)
\]

This term denotes a one-layer neural net.

• \(\theta\) denotes parameters of VAE.

• \(L\) is the number of stochastic layers.

• Sampling and probability evaluation is tractable for each \(p(h^\ell|h^{\ell+1})\).
Recognition Network

- The recognition model is defined in terms of an analogous factorization:

\[ q(h|x, \theta) = q(h^1|x, \theta)q(h^2|h^1, \theta) \cdots q(h^L|h^{L-1}, \theta) \]

- We assume that

\[ h^L \sim \mathcal{N}(0, I) \]

- The conditionals:

\[ p(h^\ell|h^{\ell+1}) \]

\[ q(h^\ell|h^{\ell-1}) \]

are Gaussians with diagonal covariances.
Variational Bound

• The VAE is trained to maximize the variational lower bound:

\[
\log p(x) = \log \mathbb{E}_{q(h|x)} \left[ \frac{p(x, h)}{q(h|x)} \right] \geq \mathbb{E}_{q(h|x)} \left[ \log \frac{p(x, h)}{q(h|x)} \right] = \mathcal{L}(x)
\]

\[
\mathcal{L}(x) = \log p(x) - D_{KL} (q(h|x))\|p(h|x))
\]

• Trading off the data log-likelihood and the KL divergence from the true posterior.

• Hard to optimize the variational bound with respect to the recognition network (high-variance).

• Key idea of Kingma and Welling is to use reparameterization trick.
Reparameterization Trick

• Assume that the recognition distribution is Gaussian:

\[ q(h^\ell | h^{\ell-1}, \theta) = \mathcal{N}(\mu(h^{\ell-1}, \theta), \Sigma(h^{\ell-1}, \theta)) \]

with mean and covariance computed from the state of the hidden units at the previous layer.

• Alternatively, we can express this in term of auxiliary variable:

\[ \epsilon^\ell \sim \mathcal{N}(0, I) \]

\[ h^\ell (\epsilon^\ell, h^{\ell-1}, \theta) = \Sigma(h^{\ell-1}, \theta)^{1/2} \epsilon^\ell + \mu(h^{\ell-1}, \theta) \]
Reparameterization Trick

• Assume that the recognition distribution is Gaussian:
\[ q(h^l|h^{l-1}, \theta) = \mathcal{N}(\mu(h^{l-1}, \theta), \Sigma(h^{l-1}, \theta)) \]

• Or
\[ \epsilon^l \sim \mathcal{N}(0, I) \]
\[ h^l(\epsilon^l, h^{l-1}, \theta) = \Sigma(h^{l-1}, \theta)^{1/2} \epsilon^l + \mu(h^{l-1}, \theta) \]

• The recognition distribution \( q(h^l|h^{l-1}, \theta) \) can be expressed in terms of a deterministic mapping:
\[ h(\epsilon, x, \theta), \quad \text{with} \quad \epsilon = (\epsilon^1, \ldots, \epsilon^L) \]

Deterministic Encoder
Distribution of \( \epsilon \) does not depend on \( \theta \)
Computing the Gradients

• The gradient w.r.t the parameters: both recognition and generative:

\[ \nabla_\theta \mathbb{E}_{h \sim q(h|x,\theta)} \left[ \log \frac{p(x, h|\theta)}{q(h|x, \theta)} \right] \]

\[ = \nabla_\theta \mathbb{E}_{\epsilon_1, \ldots, \epsilon^L \sim \mathcal{N}(0, I)} \left[ \log \frac{p(x, h(\epsilon, x, \theta)|\theta)}{q(h(\epsilon, x, \theta)|x, \theta)} \right] \]

\[ = \mathbb{E}_{\epsilon_1, \ldots, \epsilon^L \sim \mathcal{N}(0, I)} \left[ \nabla_\theta \log \frac{p(x, h(\epsilon, x, \theta)|\theta)}{q(h(\epsilon, x, \theta)|x, \theta)} \right] \]

Gradients can be computed by backprop. The mapping \( h \) is a deterministic neural net for fixed \( \epsilon \).
Computing the Gradients

• The gradient w.r.t the parameters: recognition and generative:

\[ \nabla_\theta \mathbb{E}_{h \sim q(h|x,\theta)} \left[ \log \frac{p(x, h|\theta)}{q(h|x, \theta)} \right] = \mathbb{E}_{\epsilon^1, \ldots, \epsilon^L \sim \mathcal{N}(0, I)} \left[ \nabla_\theta \log \frac{p(x, h(\epsilon, x, \theta)|\theta)}{q(h(\epsilon, x, \theta)|x, \theta)} \right] \]

• Approximate expectation by generating k samples from \( \epsilon \):

\[ \frac{1}{k} \sum_{i=1}^{k} \nabla_\theta \log w(x, h(\epsilon_i, x, \theta), \theta) \]

where we defined unnormalized importance weights:

\[ w(x, h, \theta) = \frac{p(x, h|\theta)}{q(h|x, \theta)} \]

• VAE update: Low variance as it uses the log-likelihood gradients with respect to the latent variables.
VAE: Assumptions

• Remember the variational bound:

\[ \mathcal{L}(x) = \log p(x) - D_{KL} (q(h|x)) \| p(h|x)) \]

• The variational assumptions must be approximately satisfied.

• The posterior distribution must be approximately factorial (common practice) and predictable with a feed-forward net.

• We show that we can relax these assumptions using a tighter lower bound on marginal log-likelihood.
Importance Weighted Autoencoders

• Consider the following k-sample importance weighting of the log-likelihood:

\[
\mathcal{L}_K(x) = \mathbb{E}_{h_1, \ldots, h_k \sim q(h|x)} \left[ \log \frac{1}{k} \sum_{i=1}^{k} \frac{p(x, h_i)}{q(h_i|x)} \right]
\]

\[
= \mathbb{E}_{h_1, \ldots, h_k \sim q(h|x)} \left[ \log \frac{1}{k} \sum_{i=1}^{k} w_i \right]
\]

where \( h_1, \ldots, h_k \) are sampled from the recognition network.
Importance Weighted Autoencoders

• Consider the following k-sample importance weighting of the log-likelihood:

\[
\mathcal{L}_k(x) = \mathbb{E}_{h_1, \ldots, h_k \sim q(h|x)} \left[ \log \frac{1}{k} \sum_{i=1}^{k} \frac{p(x, h_i)}{q(h_i|x)} \right]
\]

• This is a lower bound on the marginal log-likelihood:

\[
\mathcal{L}_k(x) = \mathbb{E} \left[ \log \frac{1}{k} \sum_{i=1}^{k} w_i \right] \leq \log \mathbb{E} \left[ \frac{1}{k} \sum_{i=1}^{k} w_i \right] = \log p(x)
\]

• Special Case of k=1: Same as standard VAE objective.

• Using more samples \(\rightarrow\) Improves the tightness of the bound.
IWAES vs. VAEs

• Draw $k$-samples form the recognition network $q(h|x)$
  - or $k$-sets of auxiliary variables $\epsilon$.

• Obtain the following Monte Carlo estimate of the gradient:

$$\nabla_\theta \mathcal{L}_k(x) \approx \sum_{i=1}^{k} \tilde{w}_i \nabla_\theta \log w(x, h(\epsilon_i, x, \theta), \theta)$$

• Compare this to the VAE’s estimate of the gradient:

$$\nabla_\theta \mathcal{L}(x) \approx \frac{1}{k} \sum_{i=1}^{k} \nabla_\theta \log w(x, h(\epsilon_i, x, \theta), \theta)$$
VAE: Intuition

- The gradient of the log weights decomposes:

\[
\nabla_\theta \log w(x, h(\epsilon_i, x, \theta), \theta) = \nabla_\theta \log p(x, h(\epsilon_i, x, \theta) | \theta) - \log q(h(\epsilon_i, x, \theta) | x, \theta)
\]

First term:

- **Decoder**: encourages the generative model to assign high probability to each \( h^l | h^{l+1} \).

- **Encoder**: encourages the recognition net to adjust its latent states \( h \) so that the generative network makes better predictions.
VAE: Intuition

- The gradient of the log weights decomposes:

\[ \nabla_{\theta} \log w(x, h(\epsilon_i, x, \theta), \theta) \]

\[ = \nabla_{\theta} \log p(x, h(\epsilon_i, x, \theta)|\theta) - \log q(h(\epsilon_i, x, \theta)|x, \theta) \]

**Deterministic**
- Encoder
- Decoder

Second term:
- **Encoder**: encourages the recognition network to have a spread-out distribution over predictions.
Two Architectures

• For the MNIST experiments, we considered two architectures:

1-stochastic layer

\[ x \rightarrow 784 \rightarrow 200 \rightarrow 200 \rightarrow 50 \rightarrow h^1 \]

Deterministic Layers

Stochastic Layers

2-stochastic layers

\[ h^1 \rightarrow 100 \rightarrow 100 \rightarrow 50 \rightarrow h^2 \]

Deterministic Layers

\[ x \rightarrow 784 \rightarrow 200 \rightarrow 200 \rightarrow 100 \rightarrow 100 \rightarrow h^1 \]
## MNIST Results

<table>
<thead>
<tr>
<th># stoch. layers</th>
<th>$k$</th>
<th>VAE</th>
<th>IWAE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>NLL</td>
<td>NLL</td>
</tr>
<tr>
<td></td>
<td></td>
<td>active units</td>
<td>active units</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>86.76</td>
<td>19</td>
</tr>
<tr>
<td>5</td>
<td>86.47</td>
<td>20</td>
<td>85.54</td>
</tr>
<tr>
<td>50</td>
<td>86.35</td>
<td>20</td>
<td>84.78</td>
</tr>
</tbody>
</table>
## MNIST Results

<table>
<thead>
<tr>
<th># stoch. layers</th>
<th>$k$</th>
<th>VAE NLL</th>
<th>active units</th>
<th>IWAE NLL</th>
<th>active units</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>86.76</td>
<td>19</td>
<td>86.76</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>86.47</td>
<td>20</td>
<td>85.54</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>86.35</td>
<td>20</td>
<td>84.78</td>
<td>25</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>85.33</td>
<td>16+5</td>
<td>85.33</td>
<td>16+5</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>85.01</td>
<td>17+5</td>
<td>83.89</td>
<td>21+5</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>84.78</td>
<td>17+5</td>
<td>82.90</td>
<td>26+7</td>
</tr>
</tbody>
</table>
Good Generative Model?

Handwritten Characters
Good Generative Model?

Handwritten Characters
Good Generative Model?

Handwritten Characters

Simulated          Real Data
<table>
<thead>
<tr>
<th>Handwritten Characters</th>
<th>Real Data</th>
<th>Simulated</th>
</tr>
</thead>
</table>

Good Generative Model?
Good Generative Model?

Handwritten Characters
Motivating Example
(Mansimov, Parisotto, Ba, Salakhutdinov, 2015)

- Can we generate images from natural language descriptions?

A **stop sign** is flying in blue skies

A **pale yellow school bus** is flying in blue skies

A **herd of elephants** is flying in blue skies

A **large commercial airplane** is flying in blue skies
Overall Model

Variational Autoencoder
Overall Model
(Mansimov, Parisotto, Ba, Salakhutdinov, 2015)

- **Generative Model**: Stochastic Recurrent Network, chained sequence of Variational Autoencoders, with a single stochastic layer.

Gregor et. al. 2015
Overall Model
(Mansimov, Parisotto, Ba, Salakhutdinov, 2015)

- **Generative Model**: Stochastic Recurrent Network, chained sequence of Variational Autoencoders, with a single stochastic layer.

- **Recognition Model**: Deterministic Recurrent Network.
• Maximize the variational lower bound on the marginal log-likelihood of the correct image $x$ given the caption $y$:

$$
\mathcal{L} = \sum_Z Q(Z|x, y) \log P(x|Z, y) - D_{KL}(Q(Z|x, y)\|P(Z|y)) \\
\leq \log P(x|y)
$$
MS COCO Dataset

• Contains 83K images.

• Each image contains 5 captions.

• Standard benchmark dataset for many of the recent image captioning systems.
Flipping Colors

A yellow school bus parked in the parking lot

A red school bus parked in the parking lot

A green school bus parked in the parking lot

A blue school bus parked in the parking lot
Flipping Backgrounds

A very large commercial plane flying in clear skies.

A very large commercial plane flying in rainy skies.

A herd of elephants walking across a dry grass field.

A herd of elephants walking across a green grass field.
Flipping Objects

The decadent chocolate **desert** is on the table.

A bowl of bananas is on the table.

A vintage photo of a **cat**.

A vintage photo of a **dog**.
Qualitative Comparison

A group of people walk on a beach with surf boards

Our Model

LAPGAN (Denton et. al. 2015)

Conv-Deconv VAE

Fully Connected VAE
Novel Scene Compositions

A toilet seat sits open in the bathroom

A toilet seat sits open in the grass field

Ask Google?