Variational Autoencoders

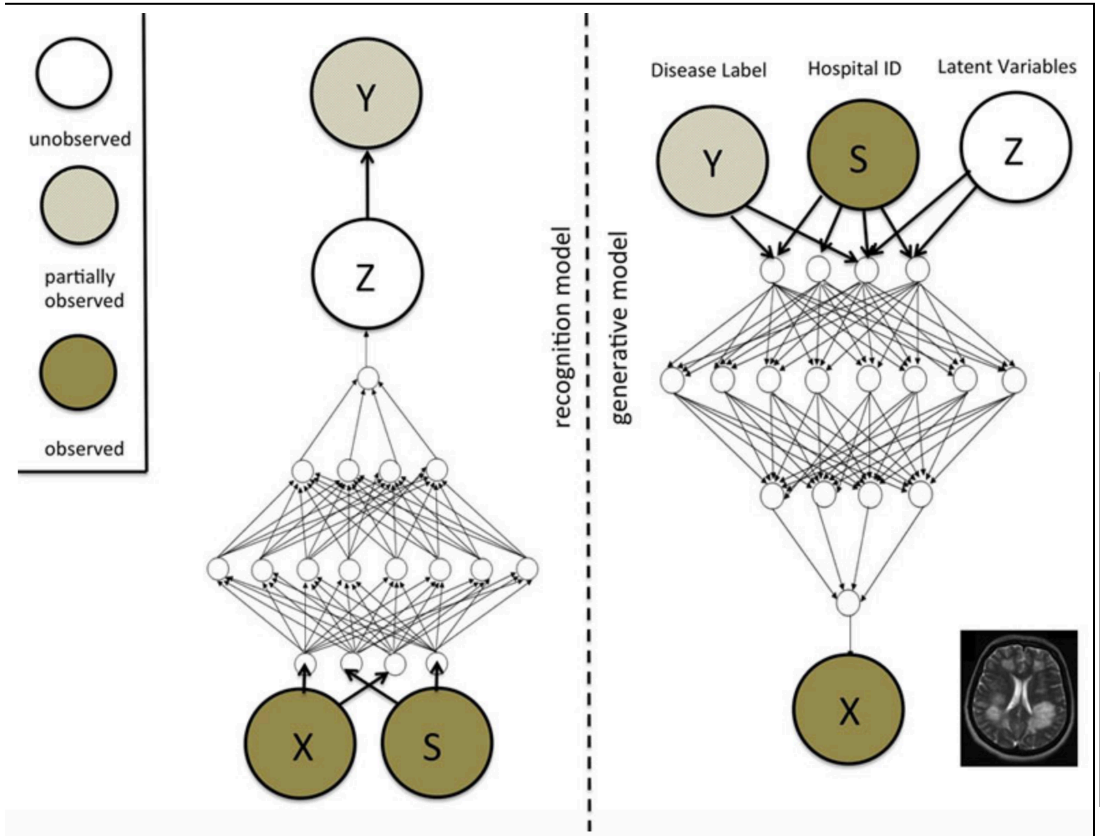
Sargur N. Srihari
srihari@cedar.buffalo.edu

Topics

1. Variational autoencoders (VAE)

Variational Autoencoder (VAE)

- Combines two types of models: *discriminative* and *generative* models into a single framework



Right: generative PGM with inputs:
1. class label y (diseases)
2. nuisance variables s (hospital identifiers)
3. latent variables z (size, shape, other brain properties)

Provides causality of observation

Left: Discriminative deep nn model
Input: observed variables
Generates posterior distributions over latent variables and possibly (if unobserved) class labels.

Performs Inference of latent variables necessary to perform variational updates

The models are trained jointly using the variational EM framework

Variational Autoencoder (VAE)

- VAE is a directed model that uses
  - Learned approximate inference
  - Trained purely with gradient-based methods
- VAE generates a sample from the model,
  - First draw sample z from code distribution $p_{\text{model}}(z)$.
  - Sample is then run through a differentiable generator network $g(z)$
  - $\boldsymbol{x}$ is sampled from distribution $p_{\text{model}}(\boldsymbol{x}; g(z)) = p_{\text{model}}(\boldsymbol{x} | g(z))$
  - However during training the approximate inference network (or encoder) $q(z|\boldsymbol{x})$ is used to obtain z and $p_{\text{model}}(\boldsymbol{x}|z)$ is viewed as a decoder network

The VAE model

- Method for modeling a data distribution using a collection of independent latent variables
  - Generative model: $p(x,\boldsymbol{z})=p(x|\boldsymbol{z})p(\boldsymbol{z})$
    - x is a r.v. representing observed data
    - $\boldsymbol{z}$ is a collection of latent variables
  - $p(x|\boldsymbol{z})$ is parameterized by a deep neural network (decoder)
  - Components of $\boldsymbol{z}$ are independent Bernoulli or Gaussian
  - Learned approx inference trained using gradient descent
    - $q(\boldsymbol{z}|x)=N(\mu,\sigma^2\mathrm{I})$ whose parameters are given by another deep network (encoder)
  - Thus we have $\boldsymbol{z} \sim \mathrm{Enc}(x)=q(\boldsymbol{z}|x)$ and $y{\sim}\mathrm{Dec}(\boldsymbol{z})=p(x|\boldsymbol{z})$

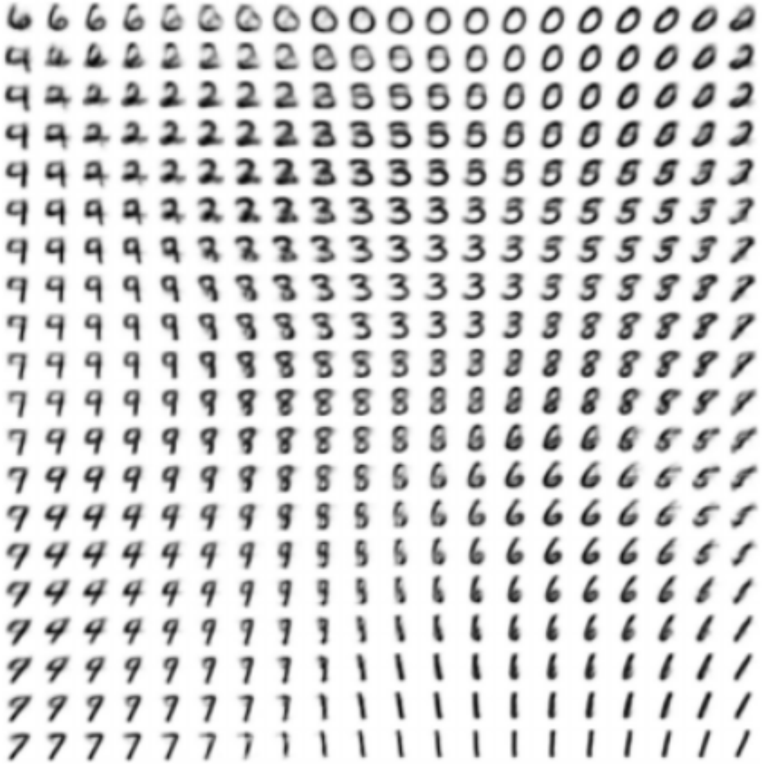
Key insight of VAE

- They can be trained by maximizing variational lower bound $\mathcal{L}(q)$ associated with data point $\boldsymbol{x}$

$$\begin{aligned}\mathcal{L}(q) &= \mathbb{E}_{\boldsymbol{z}\sim q(\boldsymbol{z}|\boldsymbol{x})} \log p_{\text{model}}(\boldsymbol{z},\boldsymbol{x}) + \mathcal{H}(q(\mathbf{z}\mid\boldsymbol{x}))\\ &= \mathbb{E}_{\boldsymbol{z}\sim q(\boldsymbol{z}|\boldsymbol{x})} \log p_{\text{model}}(\boldsymbol{x}\mid\boldsymbol{z}) - D_{\text{KL}}(q(\mathbf{z}\mid\boldsymbol{x})\|p_{\text{model}}(\mathbf{z}))\\ &\le \log p_{\text{model}}(\boldsymbol{x}).\end{aligned}$$

  - where $\mathrm{E}_{z\sim q(z|\boldsymbol{x})}\log\, p_{\text{model}}(z,\, \boldsymbol{x})$ is the joint log-likelihood of the visible and hidden variables under the approximate posterior over the latent variables
  - $\mathcal{H}(q(z|\boldsymbol{x})$ is the entropy of the approximate posterior
  - When q is chosen to be Gaussian with noise added to a predicted mean, maximizing this entropy term encourages increasing σ

VAE : 2-D coordinate systems learned for high-dimensional manifolds



Disentangling FoVs

- During training, only supervision is class labels
- Specified FoVs
  - Images captured from different viewpoints
  - Strong supervision: pairs of images with two different objects at same viewing angle
- Unspecified FoVs
  - Labels unavailable
- A disentaglement method
  - Combine *variational autoencoder* with *adversarial training*