Lecture12 Generative Models

sgc

March 24, 2022

1 Supervised vs Unsupervised Learning

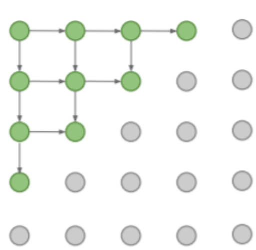
Supervised Learning : Sample (data + labels) Examples: Classification, regression, object,detection, semantic segmentation, image captioning, etc.
Unsupervised Learning : Sample (data, no labels) Learn Hidden Structure Examples: Clustering, dimensionality reduction, feature learning, density estimation, etc.

2 Generative Modeling

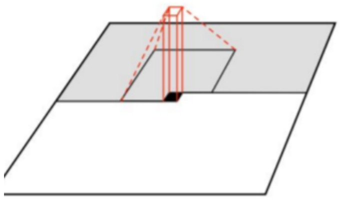
Given training data. generate new samples from same distribution: Learn p_model that approximates p_data.
Can learn important features to reconstruct data set.

3 PixelRNN and PixelCNN

FVBN: maximize likelihood of training data :$p(x) = \prod_{i=1}^{n} p(x_i|x_1, ..., x_{i-1})$ x_i is pixel No.i.
Explicit Model: Explicitly compute likelihood.
PixelRNN generate image start from the corner. Depend on previous pixels.



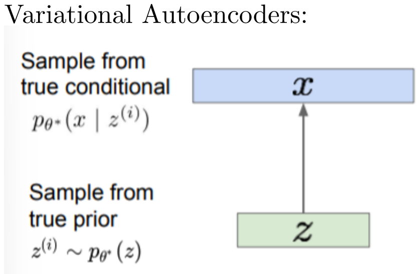
But it is slow.
PixelCNN: Still start from corner, but now suing a CNN over context region.



Relatively faster but still slow

4 Variational Autoencoders VAE

Autoencoders reconstruct data through latent z that represents the most meaningful features.
Autoencoders can be used to make a supervised model more efficient.
Variational Autoencoders:



x is an image, and z is latent factors to generate x: Choose p(z) to be simple, p(x—z) is complex (neural network).
Maximize likelihood: $p_\theta(x) = \int p_\theta(z) p_\theta(x|z)$,but it is intractable.
Use distribution $q_\phi(x|z)$ to approximate $p_\theta(x|z)$

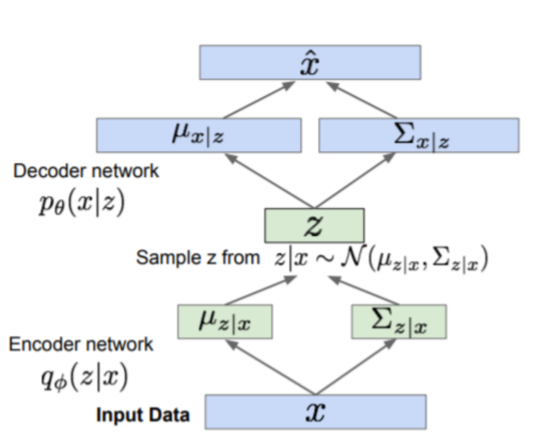
$$
\begin{aligned}
\log p_\theta(x^{(i)}) &= \mathbf{E}_{z \sim q_\phi(z|x^{(i)})}\left[\log p_\theta(x^{(i)})\right] \quad (p_\theta(x^{(i)}) \text{ Does not depend on } z) \\
&= \mathbf{E}_z\left[\log \frac{p_\theta(x^{(i)} \mid z) p_\theta(z)}{p_\theta(z \mid x^{(i)})}\right] \quad \text{(Bayes' Rule)} \\
&= \mathbf{E}_z\left[\log \frac{p_\theta(x^{(i)} \mid z) p_\theta(z)}{p_\theta(z \mid x^{(i)})} \frac{q_\phi(z \mid x^{(i)})}{q_\phi(z \mid x^{(i)})}\right] \quad \text{(Multiply by constant)} \\
&= \mathbf{E}_z\left[\log p_\theta(x^{(i)} \mid z)\right] - \mathbf{E}_z\left[\log \frac{q_\phi(z \mid x^{(i)})}{p_\theta(z)}\right] + \mathbf{E}_z\left[\log \frac{q_\phi(z \mid x^{(i)})}{p_\theta(z \mid x^{(i)})}\right] \quad \text{(Logarithms)} \\
&= \boxed{\underbrace{\mathbf{E}_z\left[\log p_\theta(x^{(i)} \mid z)\right] - D_{KL}(q_\phi(z \mid x^{(i)}) \,||\, p_\theta(z))}_{\mathcal{L}(x^{(i)}, \theta, \phi)}} + \underbrace{D_{KL}(q_\phi(z \mid x^{(i)}) \,||\, p_\theta(z \mid x^{(i)}))}_{\boxed{\geq 0}}
\end{aligned}
$$

Decoder: reconstruct the input data

Encoder: make approximate posterior distribution close to prior

Tractable lower bound which we can take gradient of and optimize! ($p_\theta(x|z)$ differentiable, KL term differentiable)

Optimizing TLB.



Encoder network is used to lessen KL divergence, and Decoder network is used to maximize Expectation.For every minibatch of input data: compute this forward pass, and then backprop.
After training, just sample z from prior and use decoder network.

5 Generative Adversarial Networks (GANs)

GANs is a Two-player game consists of the Discriminator network that tries to distinguish between real and fake images and the Generator network that tries to fool the discriminator by generating real-looking images.
Objective function:

Discriminator outputs likelihood in (0,1) of real image

Minimax objective function:

$$\min_{\theta_g} \max_{\theta_d} \left[\mathbb{E}_{x\sim p_{data}} \log \underbrace{D_{\theta_d}(x)}_{\text{Discriminator output for real data x}} + \mathbb{E}_{z\sim p(z)} \log(1 - \underbrace{D_{\theta_d}(G_{\theta_g}(z))}_{\text{Discriminator output for generated fake data G(z)}}) \right]$$

- Discriminator (θ_d) wants to **maximize objective** such that D(x) is close to 1 (real) and D(G(z)) is close to 0 (fake)
- Generator (θ_g) wants to **minimize objective** such that D(G(z)) is close to 1 (discriminator is fooled into thinking generated G(z) is real)

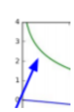
Alternate between:

1. **Gradient ascent** on discriminator

$$\max_{\theta_d} \left[\mathbb{E}_{x\sim p_{data}} \log D_{\theta_d}(x) + \mathbb{E}_{z\sim p(z)} \log(1 - D_{\theta_d}(G_{\theta_g}(z))) \right]$$

2. **Instead: Gradient ascent** on generator, different objective

$$\max_{\theta_g} \mathbb{E}_{z\sim p(z)} \log(D_{\theta_d}(G_{\theta_g}(z)))$$



Use Gradient ascent get higher gradient for bad samples and works better.

for number of training iterations **do**

for k steps **do**

- Sample minibatch of m noise samples $\{\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(m)}\}$ from noise prior $p_g(\boldsymbol{z})$.
- Sample minibatch of m examples $\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}\}$ from data generating distribution $p_{\text{data}}(\boldsymbol{x})$.
- Update the discriminator by ascending its stochastic gradient:

$$\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} \left[\log D_{\theta_d}(x^{(i)}) + \log(1 - D_{\theta_d}(G_{\theta_g}(z^{(i)}))) \right]$$

end for

- Sample minibatch of m noise samples $\{\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(m)}\}$ from noise prior $p_g(\boldsymbol{z})$.
- Update the generator by ascending its stochastic gradient (improved objective):

$$\nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^{m} \log(D_{\theta_d}(G_{\theta_g}(z^{(i)})))$$

end for

Some find k==1 more stable.