M2177.0043 Introduction to Deep Learning

Lecture 13: Deep Generative Models[1]

Hyun Oh Song[1]

[1]Dept. of Computer Science and Engineering, Seoul National University

April 28, 2020

[1]Many slides and figures adapted Justin Johnson

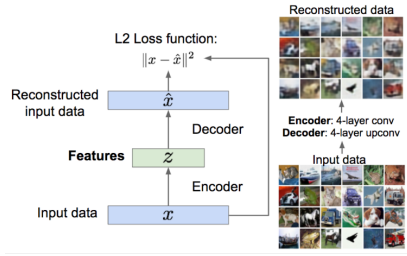
Last time

- Architectures
- Recurrent networks

Outline

Unsupervised learning

- **Data:** x, Just data, no labels.
- **Goal:** Learn some underlying hidden structure of the data
- **Examples:** Clustering, dimensionality reduction, feature learning, density estimation, etc



Autoencoders
(Feature learning)

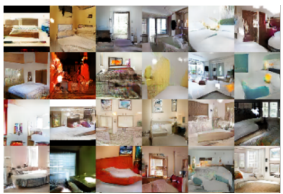
Supervised vs unsupervised

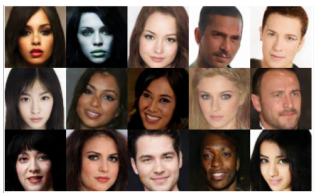
- **Data:** (x,y), x is data, y is label
- **Goal:** Learn a function to map x to y.
- **Examples:** Classification, regression, object detection, semantic segmentation, image captioning, etc.

- **Data:** x, Just data, no labels.
- **Goal:** Learn some underlying hidden structure of the data
- **Examples:** Clustering, dimensionality reduction, feature learning, density estimation, etc

Why Generative models?

- Realistic samples for artwork, super-resolution, colorization, etc.
- Data augmentation
- Feature learning





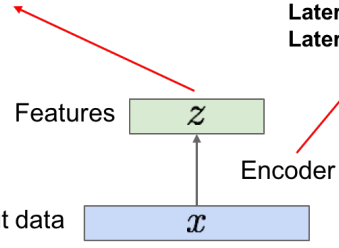
Outline

Autoencoders

z usually smaller than **x** (dimensionality reduction)

Q: Why dimensionality reduction?

A: Want features to capture meaningful factors of variation in data

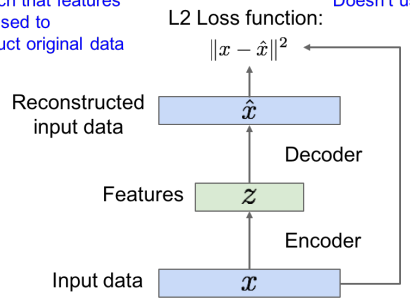


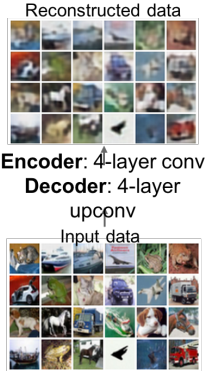
Originally: Linear + nonlinearity (sigmoid)
Later: Deep, fully-connected
Later: ReLU CNN

Autoencoders

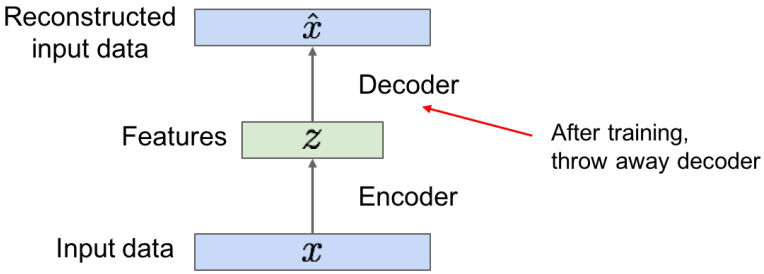
Train such that features can be used to reconstruct original data

Doesn't use labels!

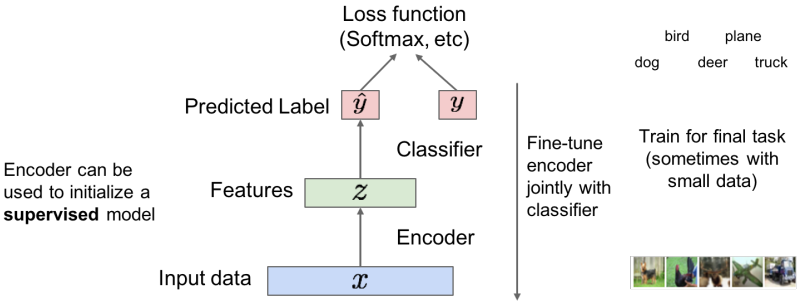




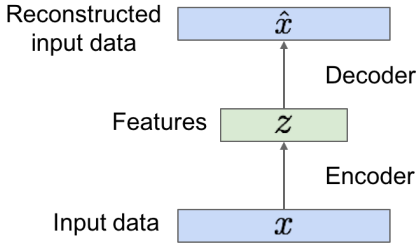
Autoencoders



Autoencoders



Autoencoders



Autoencoders can reconstruct data, and can learn features to initialize a supervised model

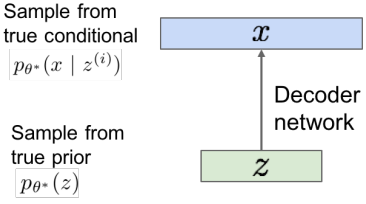
Features capture factors of variation in training data. Can we generate new images from an autoencoder?

Outline

Variational autoencoders

Probabilistic spin on autoencoders - will let us sample from the model to generate data.

Variational autoencoders



We want to estimate the true parameters θ^* of this generative model.

How to train the model?

Learn model parameters to maximize likelihood of training data

$$p_\theta(x) = \int p_\theta(z) p_\theta(x|z) dz$$

Q: What is the problem with this?

Intractable!

Kingma and Welling, "Auto-Encoding Variational Bayes", ICLR 2014

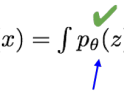
Variational autoencoders

Data likelihood: $p_\theta(x) = \int p_\theta(z) p_\theta(x|z) dz$

Kingma and Welling, “Auto-Encoding Variational Bayes”, ICLR 2014

Variational autoencoders

Data likelihood: $p_\theta(x) = \int p_\theta(z) p_\theta(x|z) dz$



Kingma and Welling, “Auto-Encoding Variational Bayes”, ICLR 2014

Variational autoencoders

Data likelihood: $p_\theta(x) = \int p_\theta(z) p_\theta(x|z) dz$

Decoder neural network

Kingma and Welling, "Auto-Encoding Variational Bayes", ICLR 2014

Variational autoencoders

Data likelihood: $p_\theta(x) = \int p_\theta(z) p_\theta(x|z) dz$

😢 ✔ ✔

Intractible to compute p(x|z) for every z!

Kingma and Welling, “Auto-Encoding Variational Bayes”, ICLR 2014

Variational autoencoders

Data likelihood: $p_\theta(x) = \int p_\theta(z) p_\theta(x|z) dz$

Posterior density also intractable: $p_\theta(z|x) = p_\theta(x|z) p_\theta(z) / p_\theta(x)$

Intractable data likelihood

Kingma and Welling, “Auto-Encoding Variational Bayes”, ICLR 2014

Variational autoencoders

Data likelihood: $p_\theta(x) = \int p_\theta(z) p_\theta(x|z) dz$

Posterior density also intractable: $p_\theta(z|x) = p_\theta(x|z) p_\theta(z) / p_\theta(x)$

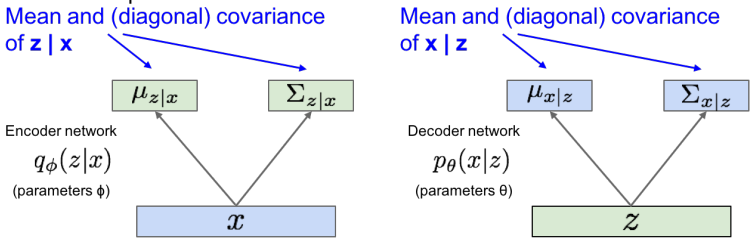
Solution: In addition to decoder network modeling $p_\theta(x|z)$, define additional encoder network $q_\phi(z|x)$ that approximates $p_\theta(z|x)$

Will see that this allows us to derive a lower bound on the data likelihood that is tractable, which we can optimize

Kingma and Welling, “Auto-Encoding Variational Bayes”, ICLR 2014

Variational autoencoders

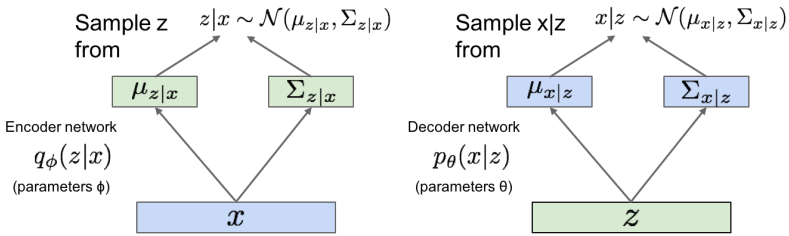
Since we're modeling probabilistic generation of data, encoder and decoder networks are probabilistic



Kingma and Welling, "Auto-Encoding Variational Bayes", ICLR 2014

Variational autoencoders

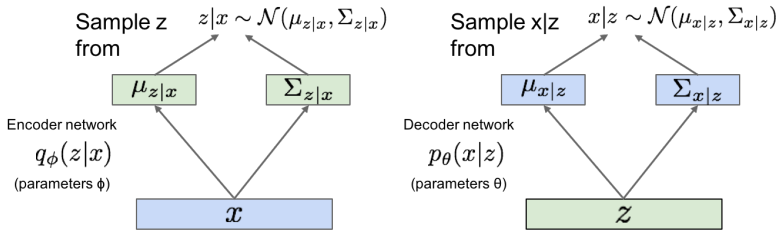
Since we're modeling probabilistic generation of data, encoder and decoder networks are probabilistic



Kingma and Welling, "Auto-Encoding Variational Bayes", ICLR 2014

Variational autoencoders

Since we're modeling probabilistic generation of data, encoder and decoder networks are probabilistic



Encoder and decoder networks also called "recognition"/"inference" and "generation" networks

Kingma and Welling, "Auto-Encoding Variational Bayes", ICLR 2014

KL divergence

Definition

For distributions P and Q of continuous random variable, the Kullback-Leibler divergence is defined to be the integral:

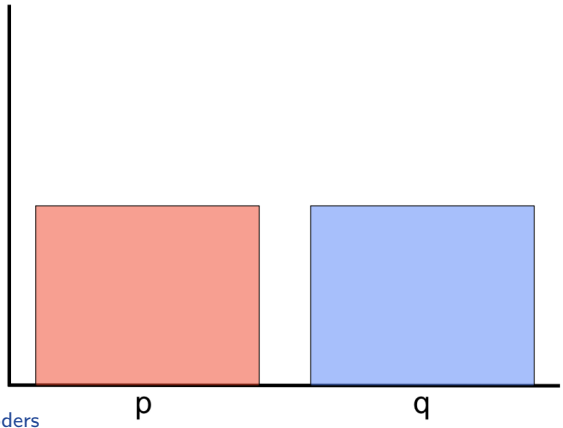
$$D_{KL}(P \parallel Q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} dx = \mathbb{E}_p \left[\log \frac{p(x)}{q(x)} \right],$$

where p and q denote the densities of P and Q.

Note that despite the intuition of divergence as a distance measure, $D(p \parallel q) \neq D(q \parallel p)$. It is the case however, that if $p = q$, the $D(p \parallel q) = 0$.

KL divergence - support[2]

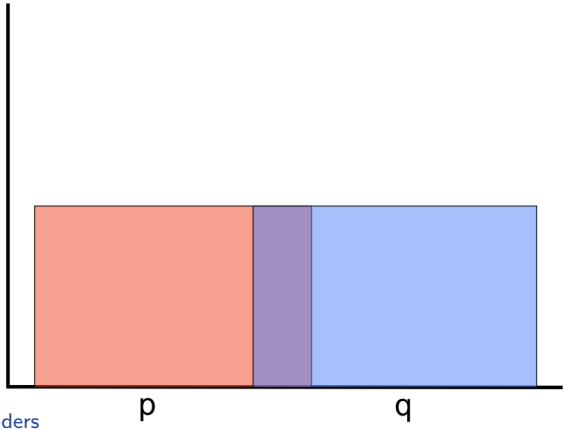
- KL divergence is defined iff $Q(x) = 0 \implies P(x) = 0$.
- Disjoint support, infinite KL



[2] Pictures taken from CMU15-859

KL divergence - support[3]

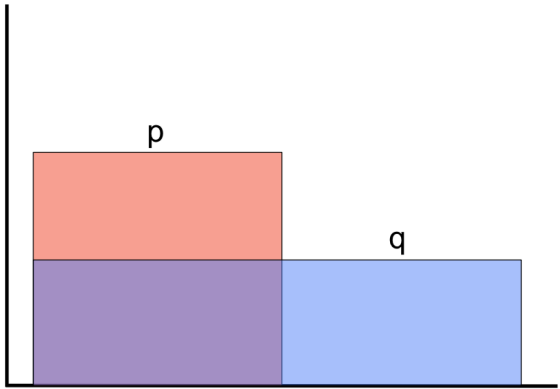
- KL divergence is defined iff $Q(x) = 0 \implies P(x) = 0$.
- Overlapping support but neither is a subset of the other. Infinite KL



[3] Pictures taken from CMU15-859

KL divergence - support[4]

- KL divergence is defined iff $Q(x) = 0 \implies P(x) = 0$.
- Supp(p) $\subseteq$ Supp(q). Finite KL.



[4] Variational Autoencoders CMU15-859

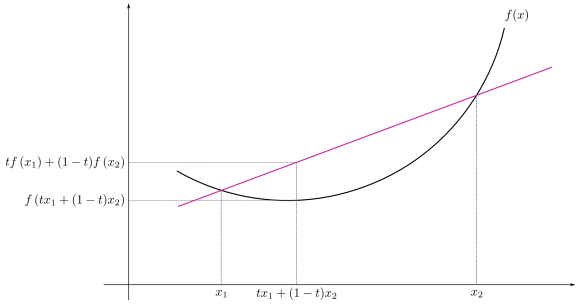
KL divergence - summary

- $D_{KL}(p \parallel q)$ is finite only if the support of p is contained within the support of q
- Definition of KL divergence uses the following conventions (Cover and Thomas, Elements of Information Theory)

$$0 \log \frac{0}{0} = 0, \quad 0 \log \frac{0}{q(x)} = 0, \quad p(x) \log \frac{p(x)}{0} = \infty$$

- Note, KL divergence can be infinite even if $p(x)$ and $q(x)$ are nonzero for all x (*i.e.* KL between Cauchy and Normal density).

Jensen's inequality[5]



Definition

Let f be a convex function and X be a random variable. Then

$$\mathbb{E}[f(X)] \geqslant f(\mathbb{E}X).$$

[5] From wikipedia

KL divergence properties

Theorem 1

Non-negativity $D_{KL}(p \parallel q) \geqslant 0$

Proof.

$D_{KL}(p \parallel q) = \mathbb{E}_p \left[\log \frac{p(x)}{q(x)}\right]$. By convexity of $f(x) = -\log(x)$ and Jensen's inequality, we have that

$$D_{KL}(p \parallel q) = \mathbb{E}_p \left[-\log \frac{q(x)}{p(x)}\right] \geqslant -\log \mathbb{E}_p \frac{q(x)}{p(x)} = 0,$$

where in the last step we used that $\int q(x)dx = 1$. Thus we conclude that KL-divergence is always non-negative. ■

Outline

ELBO Derivation method I

- Examine the KL divergence between the intractable posterior $p_\theta(z \mid x^{(i)})$ and the variational distribution $q_\phi(z \mid x^{(i)})$.

$$D_{KL}(q_\phi(z \mid x^{(i)}) \parallel p_\theta(z \mid x^{(i)})) = \mathbb{E}_{z \sim q_\phi(\cdot \mid x^{(i)})}\left[\log \frac{q_\phi(z \mid x^{(i)})}{p_\theta(z \mid x^{(i)})}\right]$$

$$= \mathbb{E}_{z \sim q_\phi(\cdot \mid x^{(i)})}\left[\log q_\phi(z \mid x^{(i)}) - \log p_\theta(x^{(i)} \mid z) - \log p_\theta(z) + \log p_\theta(x^{(i)})\right]$$

Above is from applying Bayes rule. Log prior doesn't depend on z. Omit distribution on z from below for brevity.

$$\begin{aligned}\log p_\theta(x^{(i)}) &= \mathbb{E}_z \log p_\theta(x^{(i)} \mid z) + \mathbb{E}_z\left[\log p_\theta(z) - \log q_\phi(z \mid x^{(i)})\right] \\ &\quad + D_{KL}(q_\phi(z \mid x^{(i)}) \parallel p_\theta(z \mid x^{(i)})) \\ &= \mathbb{E}_z \log p_\theta(x^{(i)} \mid z) - D_{KL}(q_\phi(z \mid x^{(i)}) \parallel p_\theta(z)) \\ &\quad + D_{KL}(q_\phi(z \mid x^{(i)}) \parallel p_\theta(z \mid x^{(i)}))\end{aligned}$$

ELBO Derivation method I

- Mathematical meaning of individual terms

$$\log p_\theta(x^{(i)}) = \underbrace{\underbrace{\mathbb{E}_z \log p_\theta(x^{(i)} \mid z)}_{\text{Decoder gives } p_\theta(x|z)} - \underbrace{D_{KL}(q_\phi(z \mid x^{(i)}) \parallel p_\theta(z))}_{\text{KL between two Gaussians has closed form sol}}}_{:=\mathcal{L}(x^{(i)},\theta,\phi),\ \text{Differentiable lower bound}} + \underbrace{D_{KL}(q_\phi(z \mid x^{(i)}) \parallel p_\theta(z \mid x^{(i)}))}_{p_\theta(z|x^{(i)})\ \text{intractable but KLD is always greater than zero}}$$

- $\log p_\theta(x^{(i)}) \geqslant \mathcal{L}(x^{(i)}, \theta, \phi)$ called variational lower bound (ELBO).
- The optimization problem is to maximize the lower bound (instead of maximizing the data log likelihood)
  $\theta^*, \phi^* = \text{argmax}_{\theta,\phi} \sum_{i=1}^{N} \mathcal{L}(x^{(i)}, \theta, \phi)$

ELBO Derivation method II

- Unlike the previous derivation method, start from the log likelihood

$$\log p_\theta(x^{(i)}) = \mathbb{E}_{z\sim q_\phi(\cdot|x^{(i)})} \log p_\theta(x^{(i)}) \quad (p_\theta(x^{(i)}) \text{ does not depend on } z)$$

$$= \mathbb{E}_z \log \frac{p_\theta(x^{(i)} \mid z) p_\theta(z)}{p_\theta(z \mid x^{(i)})}$$

(Bayes' Rule, note $p_\theta(z \mid x^{(i)})$ intractable)

$$= \mathbb{E}_z \log \frac{p_\theta(x^{(i)} \mid z) p_\theta(z)}{p_\theta(z \mid x^{(i)})} \frac{q_\phi(z \mid x^{(i)})}{q_\phi(z \mid x^{(i)})} \qquad \text{(Mult. by 1)}$$

$$= \mathbb{E}_z \log p_\theta(x^{(i)} \mid z) - \mathbb{E}_z \log \frac{q_\phi(z \mid x^{(i)})}{p_\theta(z)} + \mathbb{E}_z \log \frac{q_\phi(z \mid x^{(i)})}{p_\theta(z \mid x^{(i)})}$$

$$= \underbrace{\mathbb{E}_z \log p_\theta(x^{(i)} \mid z) - D_{KL}(q_\phi(z \mid x^{(i)}) \parallel p_\theta(z))}_{:=\mathcal{L}(x^{(i)},\theta,\phi),\ \text{Differentiable lower bound}}$$

$$+ \underbrace{D_{KL}(q_\phi(z \mid x^{(i)}) \parallel p_\theta(z \mid x^{(i)}))}_{\geqslant 0,\ \text{Intractable}}$$

ELBO Derivation method III

- Start from the joint distribution and apply Jensen's inequality

$$
\begin{aligned}
\log p_\theta(x^{(i)}) &= \log \int p_\theta(x^{(i)}, z) dz \\
&= \log \int \frac{p_\theta(x^{(i)}, z)}{q_\phi(z \mid x^{(i)})} q_\phi(z \mid x^{(i)}) dz && \text{(Mult. by 1)} \\
&\geqslant \int q_\phi(z \mid x^{(i)}) \log \frac{p_\theta(x^{(i)}, z)}{q_\phi(z \mid x^{(i)})} dz && \text{(Jensen's inequality)} \\
&= \mathbb{E}_{z \sim q_\phi(\cdot \mid x^{(i)})} \log \frac{p_\theta(x^{(i)} \mid z) p_\theta(z)}{q_\phi(z \mid x^{(i)})} \\
&= \mathbb{E}_z \log p_\theta(x^{(i)} \mid z) - \mathbb{E}_z \log \frac{q_\phi(z \mid x^{(i)})}{p_\theta(z)} \\
&= \underbrace{\mathbb{E}_z \log p_\theta(x^{(i)} \mid z) - D_{KL}(q_\phi(z \mid x^{(i)}) \parallel p_\theta(z))}_{:= \mathcal{L}(x^{(i)}, \theta, \phi),\ \text{Differentiable lower bound}}
\end{aligned}
$$

Outline

ELBO intuition

$$\mathcal{L}(x^{(i)}, \theta, \phi) = \underbrace{\mathbb{E}_{z \sim q_\phi(\cdot|x^{(i)})} \log p_\theta(x^{(i)} \mid z)}_{\text{Decoder gives } p_\theta(x|z)} - \underbrace{D_{KL}(q_\phi(z \mid x^{(i)}) \parallel p_\theta(z))}_{\text{KL between two Gaussians has closed form sol}}$$

- $\mathbb{E}_z \log p_\theta(x^{(i)} \mid z)$: Expectation of all samples of z sampled from passing $x^{(i)}$ through the encoder network ($q_\phi(\cdot \mid x^{(i)})$), sampling z, computing the log likelihood of decoder reconstruction from the sample z, ($p_\theta(\cdot \mid z)$). Encourage good reconstruction.
- $D_{KL}(q_\phi(z \mid x^{(i)}) \parallel p_\theta(z))$: Make the variational approximate posterior distribution close to prior.
- Typically, we use standard normal or uniform distribution for the prior.

KL between Gaussians

- $D_{KL}(\mathcal{N}(\mu_0, \Sigma_0) \parallel \mathcal{N}(\mu_1, \Sigma_1)) =$
  $\frac{1}{2}\left(\mathbf{Tr}\left(\Sigma_1^{-1}\Sigma_0\right) + (\mu_1 - \mu_0)^\intercal \Sigma_1^{-1}(\mu_1 - \mu_0) - k + \log\left(\frac{\det \Sigma_1}{\det \Sigma_0}\right)\right)$
- In our case, this simplifies to:
  $D_{KL}(\mathcal{N}(\mu(x^{(i)}), \Sigma(x^{(i)})) \parallel \mathcal{N}(0, I)) =$
  $\frac{1}{2}\left(\mathbf{Tr}\,\Sigma(x^{(i)}) + \mu(x^{(i)})^\intercal \mu(x^{(i)}) - k - \log\det\Sigma(x^{(i)})\right)$
- In case of diagonal covariance matrix:
  $D_{KL}(\mathcal{N}(\mu(x^{(i)}), \boldsymbol{\sigma}(x^{(i)})) \parallel \mathcal{N}(0, I)) =$
  $\frac{1}{2}\sum_j^d 1 + \log\sigma_j^2(x^{(i)}) - \mu_j^2(x^{(i)}) - \sigma_j^2(x^{(i)})$
- Derive these for fun at home!
- Can you eyeball the simple case from the general case?

VAE - training

Putting it all together: maximizing the likelihood lower bound

$$\underbrace{\mathbf{E}_z\left[\log p_\theta(x^{(i)} \mid z)\right] - D_{KL}(q_\phi(z \mid x^{(i)}) \,||\, p_\theta(z))}_{\mathcal{L}(x^{(i)}, \theta, \phi)}$$

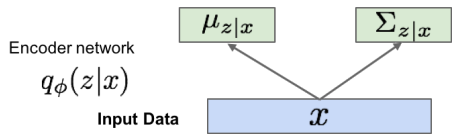
Let's look at computing the bound (forward pass) for a given minibatch of input data

Input Data x

VAE - training

Putting it all together: maximizing the likelihood lower bound

$$\underbrace{\mathbf{E}_z\left[\log p_\theta(x^{(i)} \mid z)\right] - D_{KL}(q_\phi(z \mid x^{(i)}) \,||\, p_\theta(z))}_{\mathcal{L}(x^{(i)}, \theta, \phi)}$$

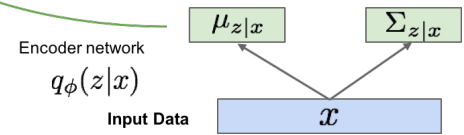


VAE - training

Putting it all together: maximizing the likelihood lower bound

$$\underbrace{\mathbf{E}_z\left[\log p_\theta(x^{(i)} \mid z)\right] - D_{KL}(q_\phi(z \mid x^{(i)}) \,||\, p_\theta(z))}_{\mathcal{L}(x^{(i)}, \theta, \phi)}$$

Make approximate posterior distribution close to prior

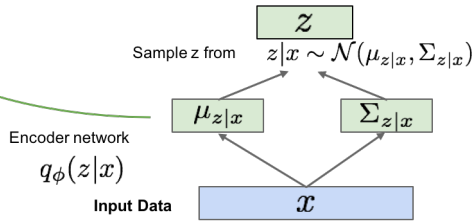


VAE - training

Putting it all together: maximizing the likelihood lower bound

$$\underbrace{\mathbf{E}_z\left[\log p_\theta(x^{(i)} \mid z)\right] - D_{KL}(q_\phi(z \mid x^{(i)}) \,||\, p_\theta(z))}_{\mathcal{L}(x^{(i)}, \theta, \phi)}$$

Make approximate posterior distribution close to prior

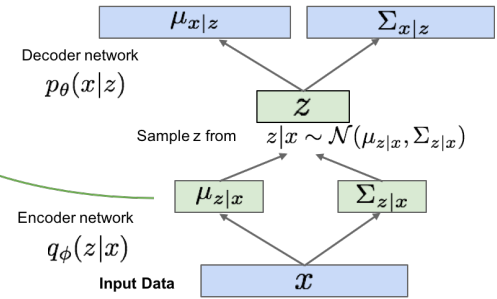


VAE - training

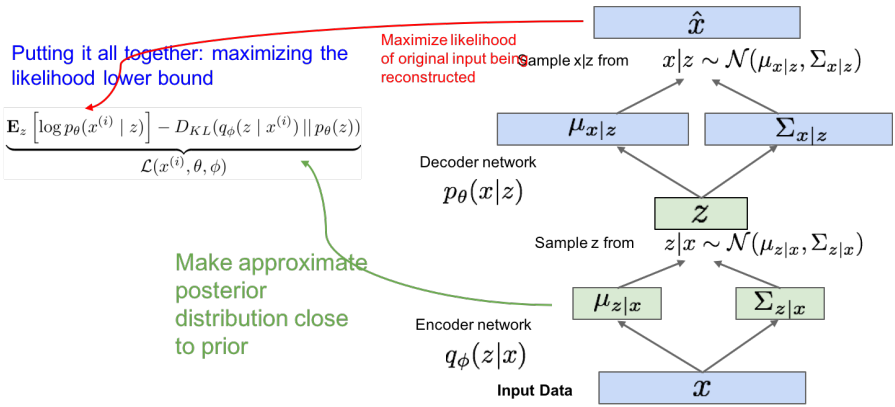
Putting it all together: maximizing the likelihood lower bound

$$\underbrace{\mathbf{E}_z\left[\log p_\theta(x^{(i)} \mid z)\right] - D_{KL}(q_\phi(z \mid x^{(i)}) \,||\, p_\theta(z))}_{\mathcal{L}(x^{(i)}, \theta, \phi)}$$

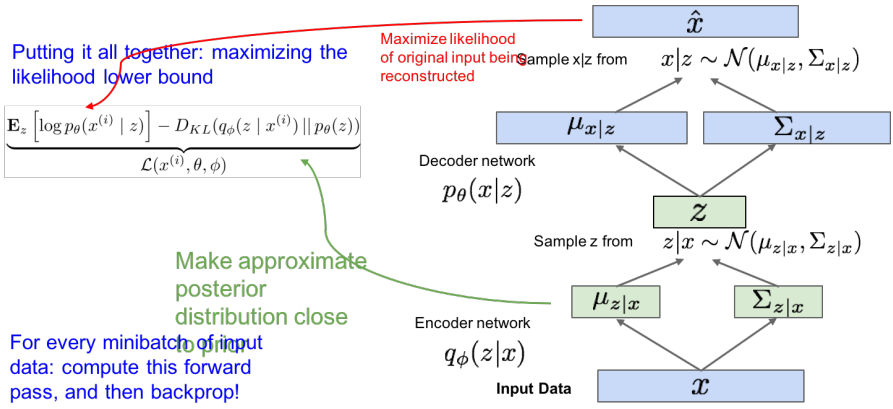
Make approximate posterior distribution close to prior



VAE - training



VAE - training

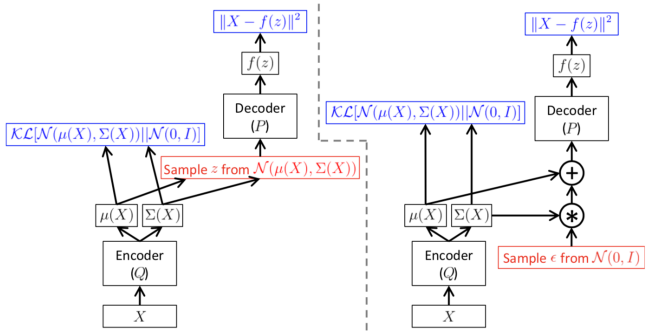


Reparameterization trick

- Reconstruction term $\mathbb{E}_z \log p_\theta(x^{(i)} \mid z)$ has sampling procedure which is not differentiable.
- Move sampling to an input layer. Given $\mu(x^{(i)}), \Sigma(x^{(i)})$ from the encoder $q_\phi(z \mid x^{(i)})$, we can sample from $\mathcal{N}(\mu(x^{(i)}), \Sigma(x^{(i)}))$ by first sampling $\epsilon \sim \mathcal{N}(0, I)$, then computing $z = \mu(x^{(i)}) + A(x^{(i)})\epsilon$. Remember the homework problem?
- A could be a Cholesky factor or $QD^{1/2}$ factor of $\Sigma(x^{(i)})$
- Reconstruction is now reparameterized as $\mathbb{E}_{\epsilon \sim \mathcal{N}(0,I)} \log p_\theta(x^{(i)} \mid z) = \mu(x^i) + A(x^{(i)})\epsilon$

Reparameterization trick[6]

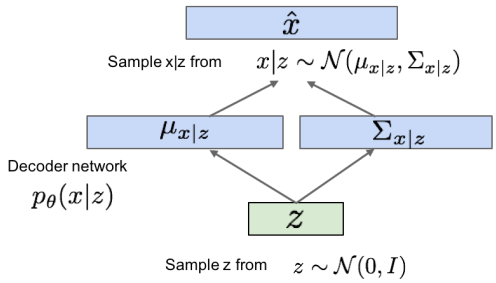
- Left is without the reparameterization trick and right is with it. Red shows sampling operations that are non-differentiable. Blue shows loss layers (with Gaussian log likelihood). The feedforward behavior of these networks is identical, but backpropagation can be applied only to the right network.



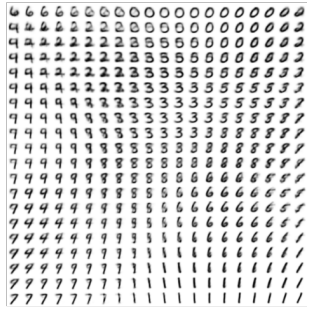
from Carl Doersch arXiv 1606.05908

VAE - inference

Use decoder network. Now sample z from prior!

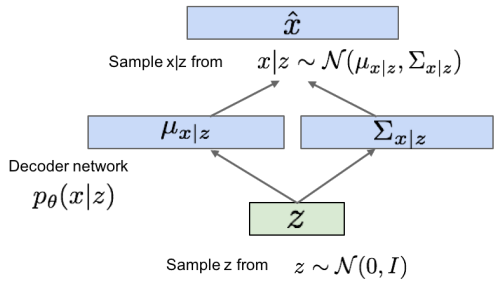


Kingma and Welling, “Auto-Encoding Variational Bayes”, ICLR 2014



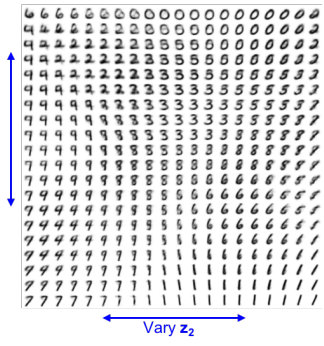
VAE - inference

Use decoder network. Now sample z from prior!



Kingma and Welling, “Auto-Encoding Variational Bayes”, ICLR 2014

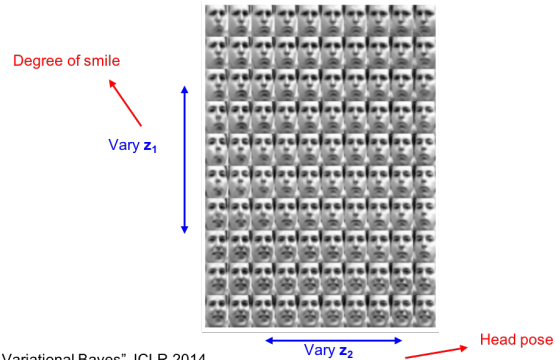
Data manifold for 2-d **z**



VAE - inference

Diagonal prior on **z** => independent latent variables

Different dimensions of **z** encode interpretable factors of variation



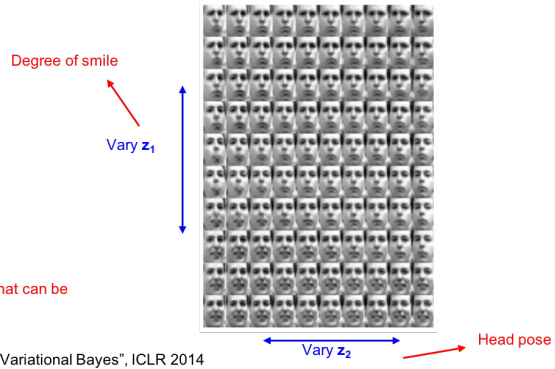
Kingma and Welling, “Auto-Encoding Variational Bayes”, ICLR 2014

VAE - inference

Diagonal prior on **z** => independent latent variables

Different dimensions of **z** encode interpretable factors of variation

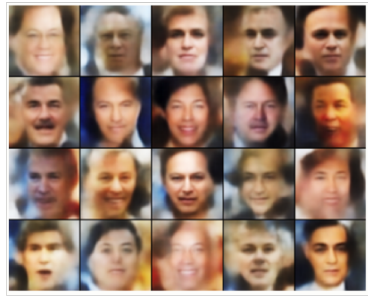
Also good feature representation that can be computed using $q_\phi(z|x)$!



Kingma and Welling, “Auto-Encoding Variational Bayes”, ICLR 2014

VAE - inference

32x32 CIFAR-10



Labeled Faces in the Wild

Summary

- Principled approach to generative models
- Allows inference of $q(z|x)$ and can be useful feature representation for other tasks
- Active areas of research: More flexible approximations, e.g. richer approximate posterior instead of diagonal Gaussian. Incorporating structure in latent variables (as opposed to i.i.d z).
- How about discrete latent variables? (See Jeong & Song ICML19)