DSC291: Machine Learning with Few Labels

Unsupervised Learning

Zhiting Hu

Lecture 19, May 17, 2024

UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE

This Lecture

- Variational Autoencoders (30mins)

- Presentation #1 (10mins):
  - Zehan Li, Dense Passage Retrieval for Open-Domain Question Answering

- Presentation #2 (10mins):
  - Zhaoyang Li, BLINK : Multimodal Large Language Models Can See but Not Perceive

Google form for presentation questions and feedback:

Recap: Computing Gradients of Expectations

- Loss: $\mathcal{L} = \mathbb{E}_{q_\lambda(\boldsymbol{z})}[f_\lambda(\boldsymbol{z})]$
- Score gradient

$$\nabla_\lambda \mathcal{L} = \boxed{\mathbb{E}_{q_\lambda(\boldsymbol{z})}}[f_\lambda(\boldsymbol{z})\nabla_\lambda \log q_\lambda(\boldsymbol{z}) + \nabla_\lambda f_\lambda(\boldsymbol{z})]$$

  - Pros: generally applicable to any distribution $q(z|\lambda)$
  - Cons: empirically has high variance → slow convergence

- Reparameterization gradient

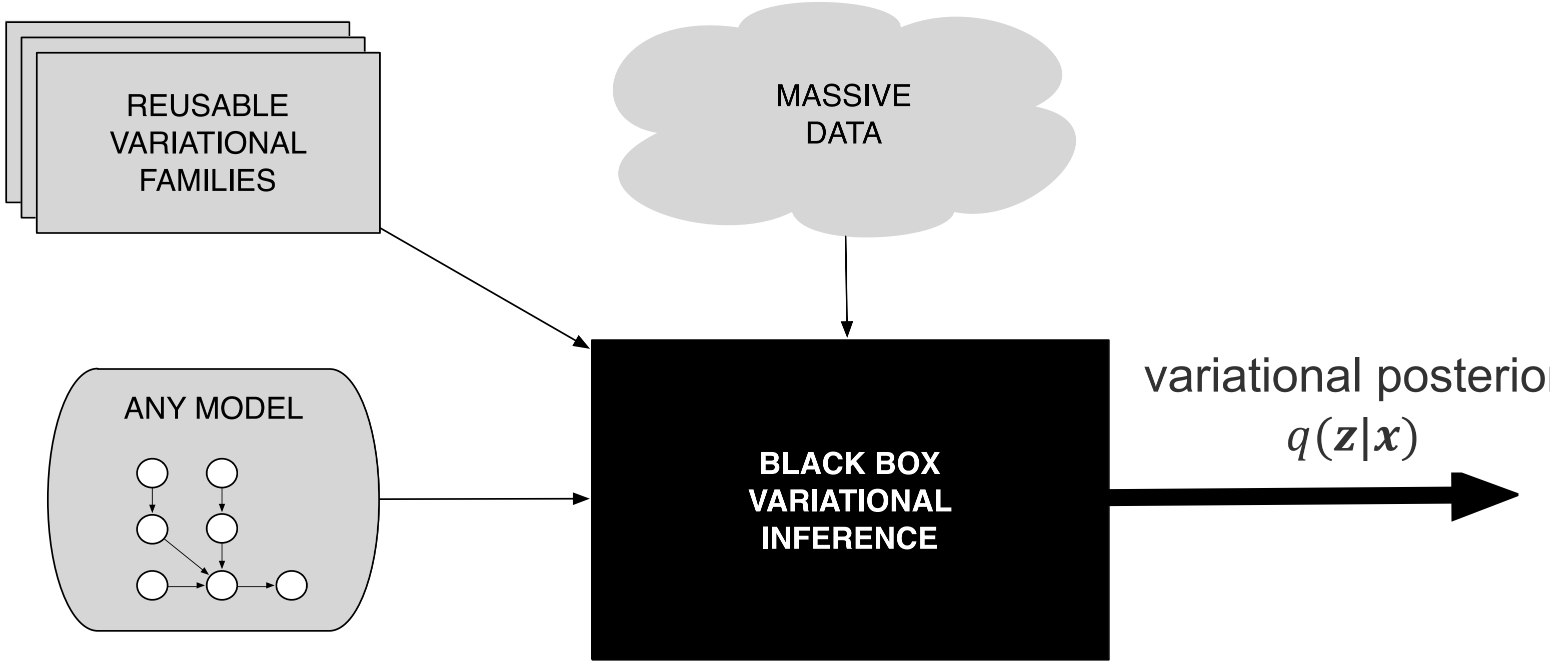
$$\nabla_\lambda \mathcal{L} = \boxed{\mathbb{E}_{\epsilon \sim s(\epsilon)}}[\nabla_{\boldsymbol{z}} f_\lambda(\boldsymbol{z})\, \nabla_\lambda t(\epsilon, \lambda)]$$

$\varepsilon \sim s(\varepsilon)$

$z = t(\varepsilon, \lambda)$

  - Pros: empirically, lower variance of the gradient estimate
  - Cons: Not all distributions can be reparameterized

Recap: Black-box Variational Inference (BBVI)



- Easily use variational inference with **any model**
- **No mathematical work** beyond specifying the model
- Perform inference with **massive data**

(Courtesy: Blei et al., 2018)

$$\mathcal{L} = \mathbb{E}_{q_\lambda(z)}[f_\lambda(z)]$$

$$\nabla_\lambda \mathcal{L} = \mathbb{E}_{q_\lambda(z)}[f_\lambda(z)\nabla_\lambda \log q_\lambda(z) + \nabla_\lambda f_\lambda(z)]$$

BBVI with the score gradient

- ELBO:

$$\mathcal{L}(\lambda) = \mathbb{E}_{q(z|\lambda)}[\log p(x, z)] - \mathbb{E}_{q(z|\lambda)}[\log q(z|\lambda)]$$

$$= \mathbb{E}_{q_\lambda}[\log P(x,z) - \log q(z|\lambda)]$$

- **Question:** what's the score gradient w.r.t. λ ?

$$\nabla_\lambda \mathcal{L} = \mathrm{E}_q[\nabla_\lambda \log q(z|\lambda)(\log p(x,z) - \log q(z|\lambda))]$$

$$\nabla_\lambda L = \mathbb{E}_{q_\lambda}\left[(\log P(x,z) - \log q_\lambda)\nabla_\lambda \log q_\lambda(z) + \nabla_\lambda \log q_\lambda\right]$$

$$\mathbb{E}_{q_\lambda}[\nabla_\lambda \log q_\lambda] = \mathbb{E}_{q_\lambda}\left[\frac{1}{q_\lambda}\nabla_\lambda q_\lambda\right]$$

$$= \int q_\lambda \cdot \frac{1}{q_\lambda}\nabla_\lambda q_\lambda = \int \nabla_\lambda q_\lambda = \nabla_\lambda \int q_\lambda = \nabla_\lambda 1 = 0$$

[Ranganath et al.,14]

$$\boxed{\begin{aligned}\mathcal{L} &= \mathbb{E}_{q_\lambda(\mathbf{z})}[f_\lambda(\mathbf{z})] \\ \nabla_\lambda \mathcal{L} &= \mathbb{E}_{q_\lambda(\mathbf{z})}[f_\lambda(\mathbf{z})\nabla_\lambda \log q_\lambda(\mathbf{z}) + \nabla_\lambda f_\lambda(\mathbf{z})]\end{aligned}}$$

BBVI with the score gradient

- ELBO:

$$\mathcal{L}(\lambda) = \mathbb{E}_{q(\mathbf{z}|\boldsymbol{\lambda})}[\log p(\boldsymbol{x}, \mathbf{z})] - \mathbb{E}_{q(\mathbf{z}|\boldsymbol{\lambda})}[\log q(\mathbf{z}|\lambda)]$$

- **Question:** what's the score gradient w.r.t. λ ?

$$-\mathbb{E}_{q_\lambda}[\log P(x,z) - \log q(z|\lambda)]$$

$$\nabla_\lambda \mathcal{L} = \mathrm{E}_q[\nabla_\lambda \log q(z|\lambda)(\log p(x,z) - \log q(z|\lambda))]$$

- Compute noisy unbiased gradients of the ELBO with Monte Carlo samples from the variational distribution

$$\nabla_\lambda \mathcal{L} \approx \frac{1}{S}\sum_{s=1}^{S} \nabla_\lambda \log q(z_s|\lambda)(\log p(x, z_s) - \log q(z_s|\lambda)),$$

$$\text{where } z_s \sim q(z|\lambda).$$

[Ranganath et al.,14]

BBVI with the reparameterization gradient

$$\mathcal{L} = \mathbb{E}_{q_\lambda(\boldsymbol{z})}[f_\lambda(\boldsymbol{z})]$$

$$\nabla_\lambda \mathcal{L} = \mathbb{E}_{\boldsymbol{\epsilon}\sim \boldsymbol{s}(\boldsymbol{\epsilon})}[\nabla_{\boldsymbol{z}} f_\lambda(\boldsymbol{z})\, \nabla_\lambda t(\epsilon, \lambda)]$$

- ELBO:

$$\mathcal{L}(\lambda) = \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{\lambda})}[\log p(\boldsymbol{x}, \boldsymbol{z})] - \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{\lambda})}[\log q(\boldsymbol{z}|\lambda)]$$

- **Question:** what's the reparamerization gradient w.r.t. λ ?

$$\begin{array}{c} \epsilon \sim s(\epsilon) \\ z = t(\epsilon, \lambda) \end{array} \iff z \sim q(z|\lambda)$$

$$\nabla_\lambda \mathcal{L} = \mathrm{E}_{\epsilon\sim s(\epsilon)}[\nabla_z[\log p(x, z) - \log q(z)]\, \nabla_\lambda t(\epsilon, \lambda)]$$

Variational Autoencoders (VAEs)

Variational Auto-Encoders (VAEs)

VAEs are a combination of the following ideas:

- Variational Inference
  - ELBO
- Variational distribution parametrized as neural networks
- Reparameterization trick

[Courtesy: Dhruv, CS 4803]

Variational Auto-Encoders (VAEs)

- Model $p_\theta(\boldsymbol{x}, \boldsymbol{z}) = p_\theta(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})$
  - $p_\theta(\boldsymbol{x}|\boldsymbol{z})$: a.k.a., generative model, generator, (probabilistic) decoder, …
  - $p(\boldsymbol{z})$: prior, e.g., Gaussian
- Assume variational distribution $q_\phi(\boldsymbol{z}|\boldsymbol{x})$
  - E.g., a Gaussian distribution parameterized as **deep neural networks**
  - a.k.a, recognition model, inference network, (probabilistic) encoder, …
- ELBO:

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{x}) = \mathrm{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}, \boldsymbol{z})] + \mathrm{H}(q_\phi(\boldsymbol{z}|\boldsymbol{x}))$$
$$= \mathrm{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] - \mathrm{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x}) \,||\, p(\boldsymbol{z}))$$

Reconstruction

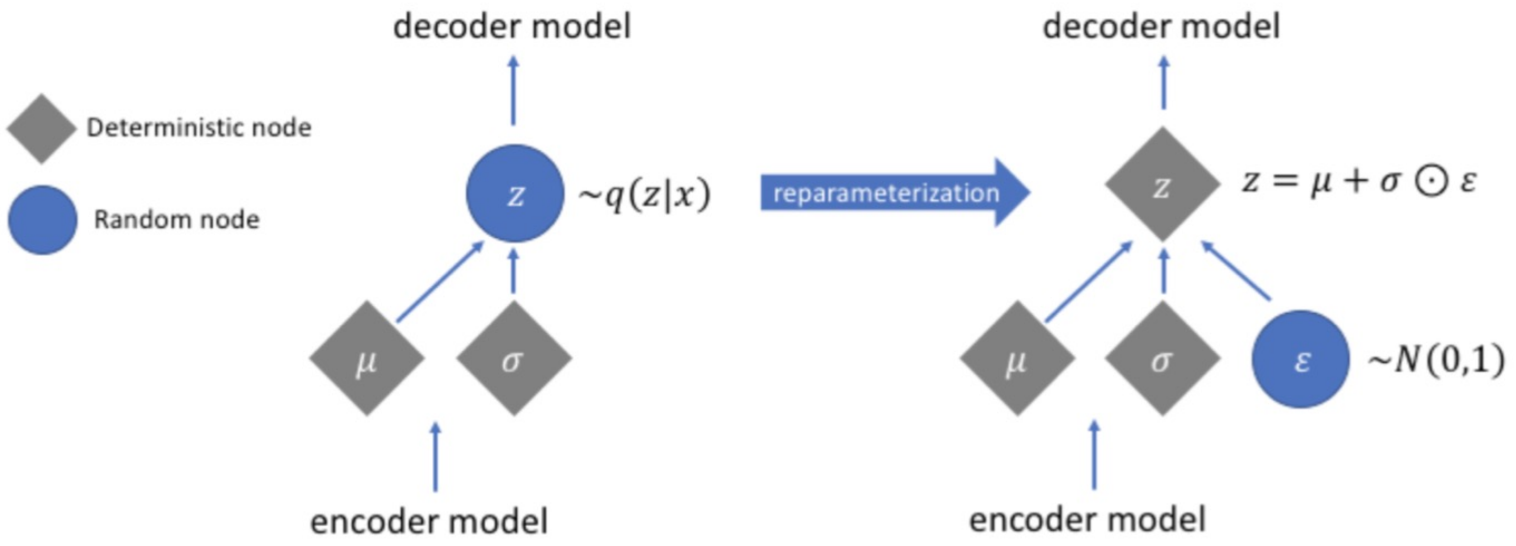
Divergence from prior
(KL divergence between two Guassians
has an analytic form)

Variational Auto-Encoders (VAEs)

- ELBO:

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{x}) = \mathrm{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}, \boldsymbol{z})] + \mathrm{H}(q_\phi(\boldsymbol{z}|\boldsymbol{x}))$$
$$= \mathrm{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] - \mathrm{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x}) \,||\, p(\boldsymbol{z}))$$

- Reparameterization:
  - $[\boldsymbol{\mu}; \boldsymbol{\sigma}] = f_\phi(\boldsymbol{x})$ (a neural network)
  - $\boldsymbol{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \boldsymbol{N}(\mathbf{0}, \mathbf{1})$



Variational Auto-Encoders (VAEs)

- ELBO:

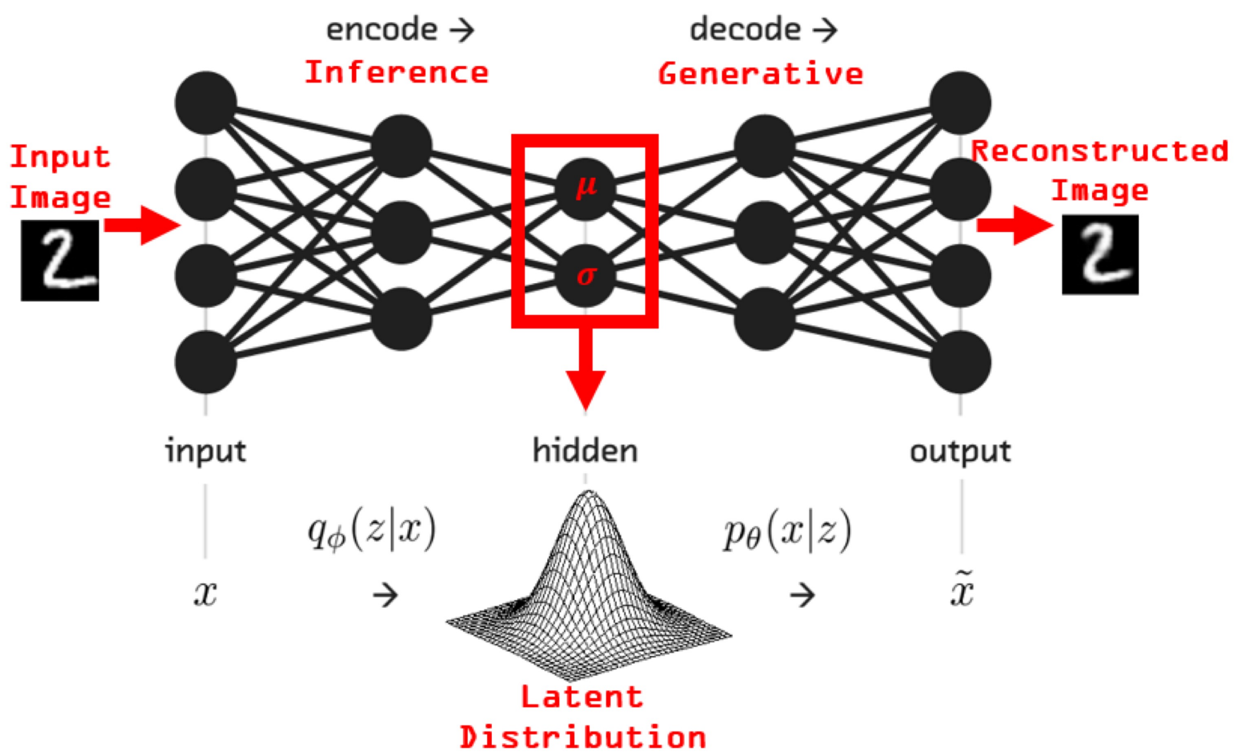
$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{x}) = \mathrm{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}, \boldsymbol{z})] + \mathrm{H}(q_\phi(\boldsymbol{z}|\boldsymbol{x}))$$
$$= \mathrm{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] - \mathrm{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x}) \,||\, p(\boldsymbol{z}))$$

- Reparameterization:
  - $[\boldsymbol{\mu}; \boldsymbol{\sigma}] = f_\phi(\boldsymbol{x})$ (a neural network)
  - $\boldsymbol{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \boldsymbol{N}(\mathbf{0}, \mathbf{1})$

$$\nabla_{\boldsymbol{\phi}} \mathcal{L} = \mathrm{E}_{\epsilon \sim N(\mathbf{0},\mathbf{1})}\left[\nabla_{\boldsymbol{z}} \left[\log p_\theta(\boldsymbol{x}, \boldsymbol{z}) - \log q_\phi(\boldsymbol{z}|\boldsymbol{x})\right] \nabla_\phi z(\epsilon, \boldsymbol{\phi})\right]$$

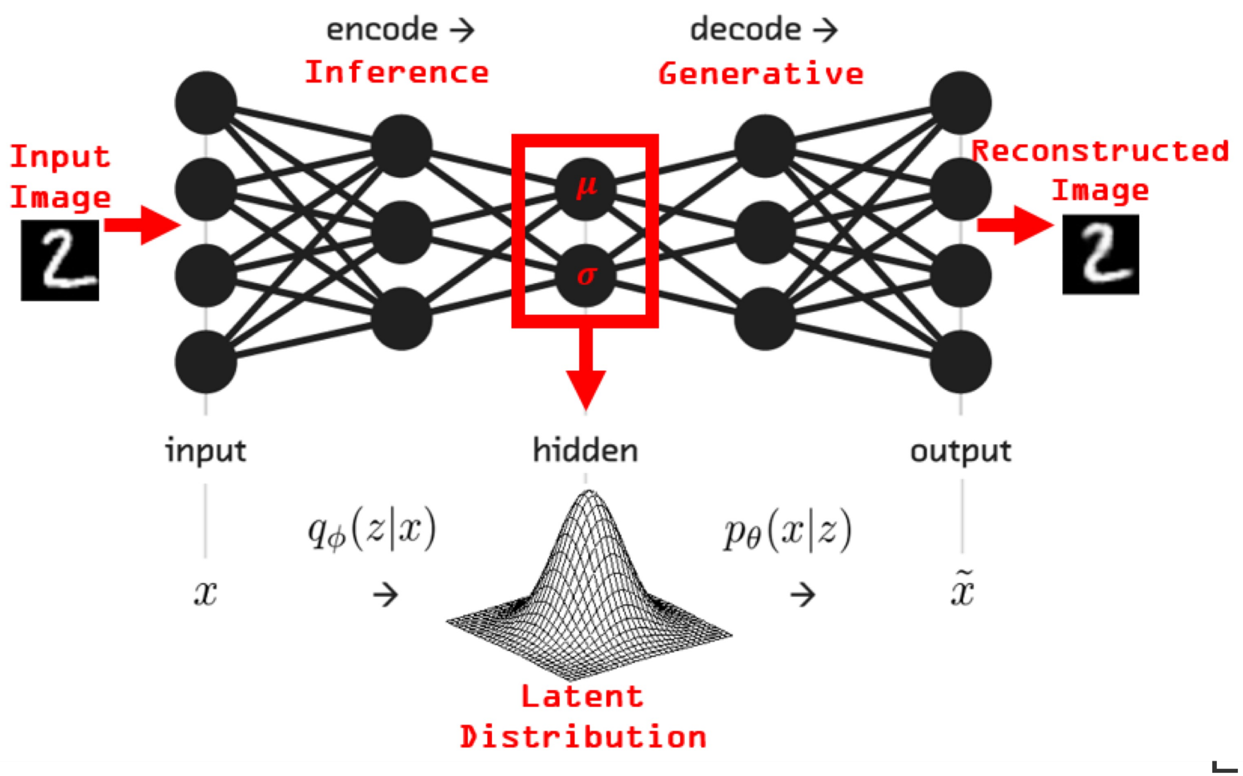
$$\nabla_\theta \mathcal{L} = \mathrm{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\nabla_\theta \log p_\theta(\boldsymbol{x}, \boldsymbol{z})]$$

Example: VAEs for images



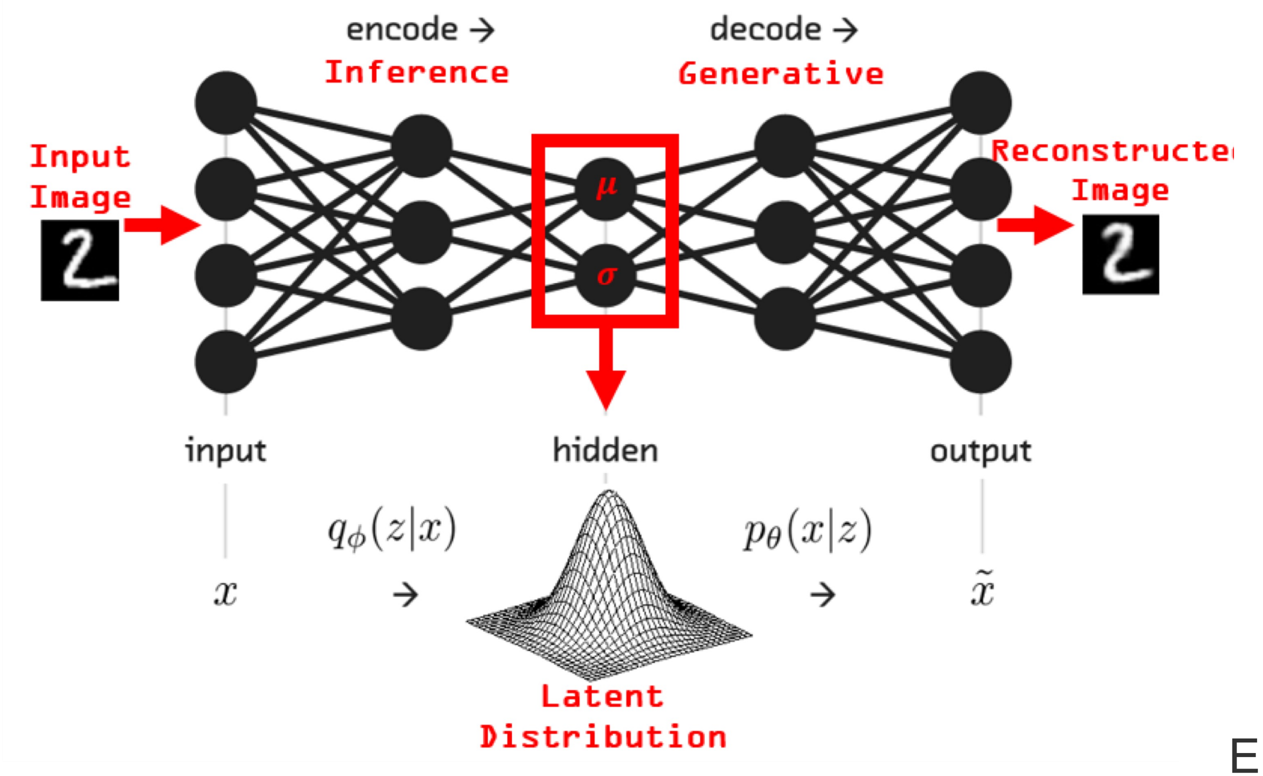
[https://www.kaggle.com/rvislaywade/visualizing-mnist-using-a-variational-autoencoder]

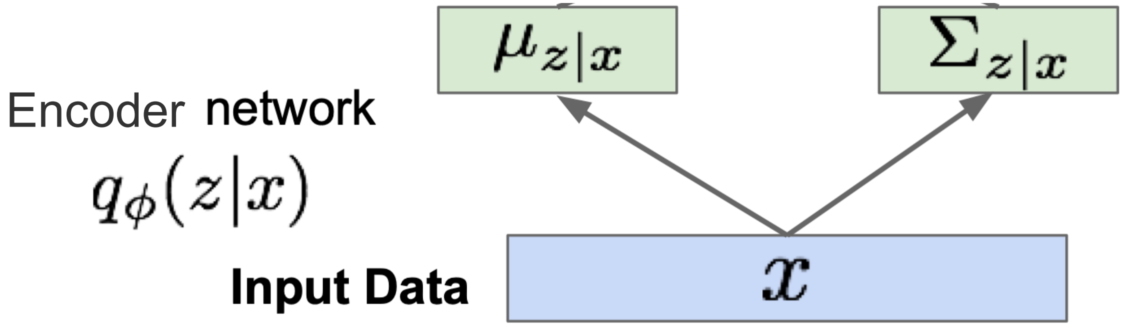
Example: VAEs for images



[Courtesy: Fei-Fei Li, Justin Johnson, Serena Yeung, CS 231n]

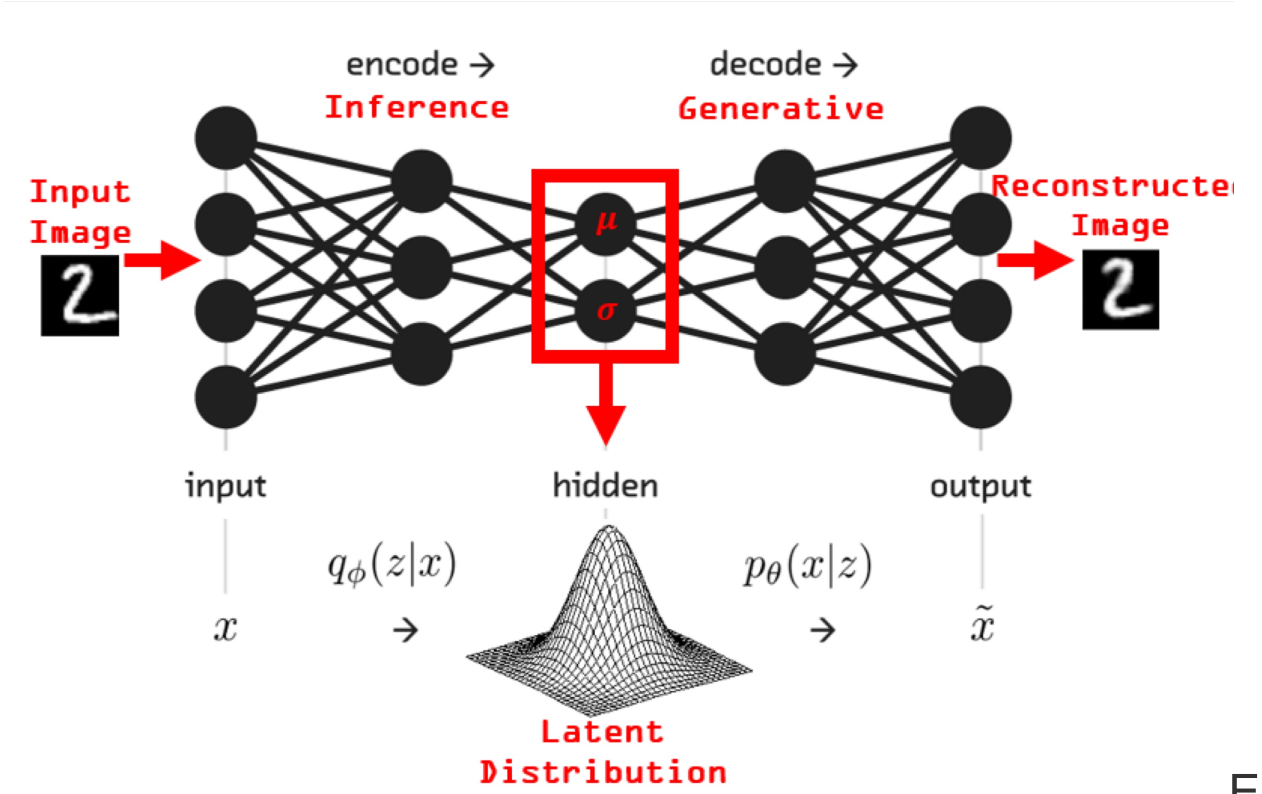
Example: VAEs for images

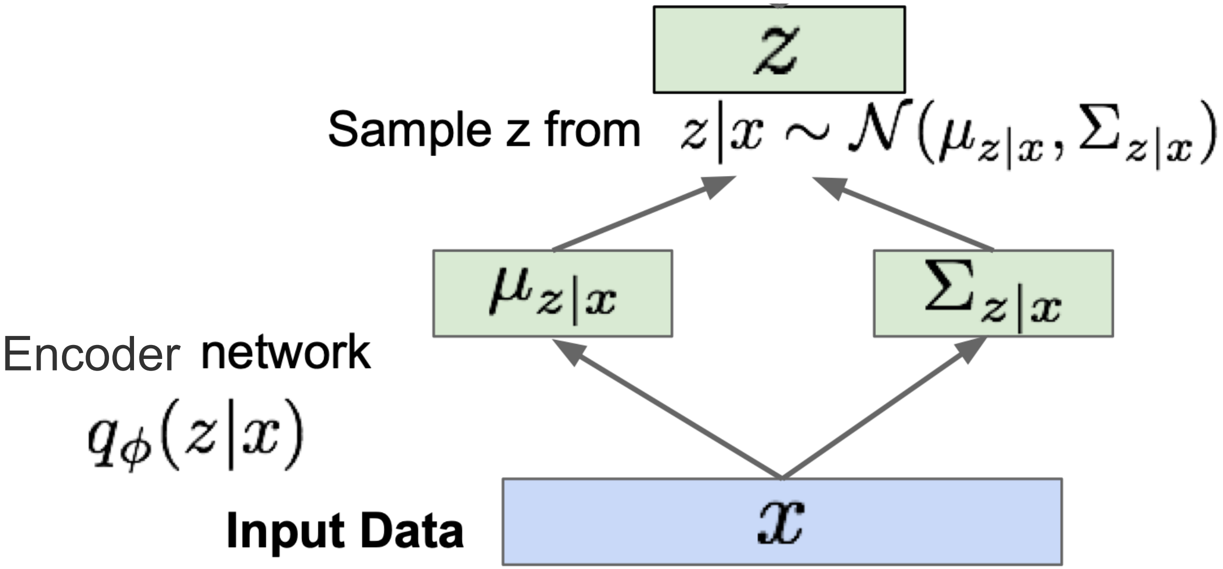




[Courtesy: Fei-Fei Li, Justin Johnson, Serena Yeung, CS 231n]

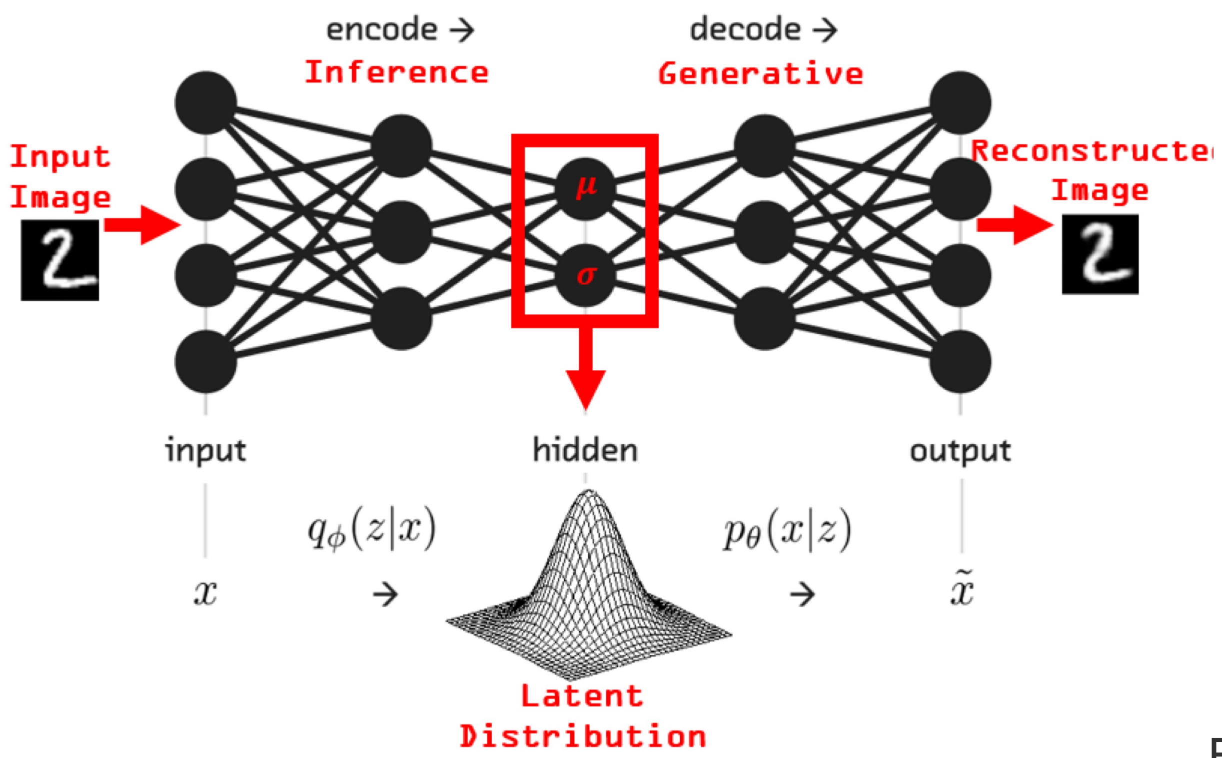
Example: VAEs for images

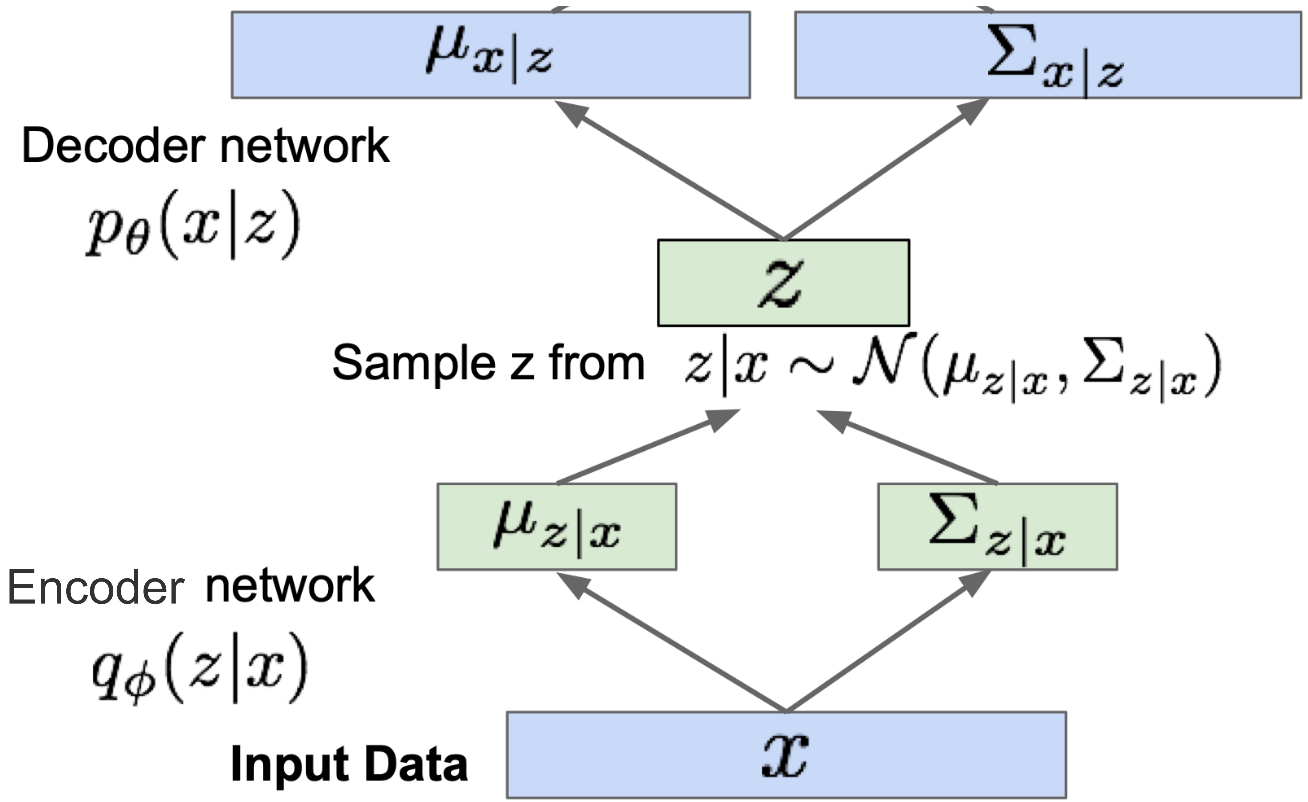




[Courtesy: Fei-Fei Li, Justin Johnson, Serena Yeung, CS 231n]

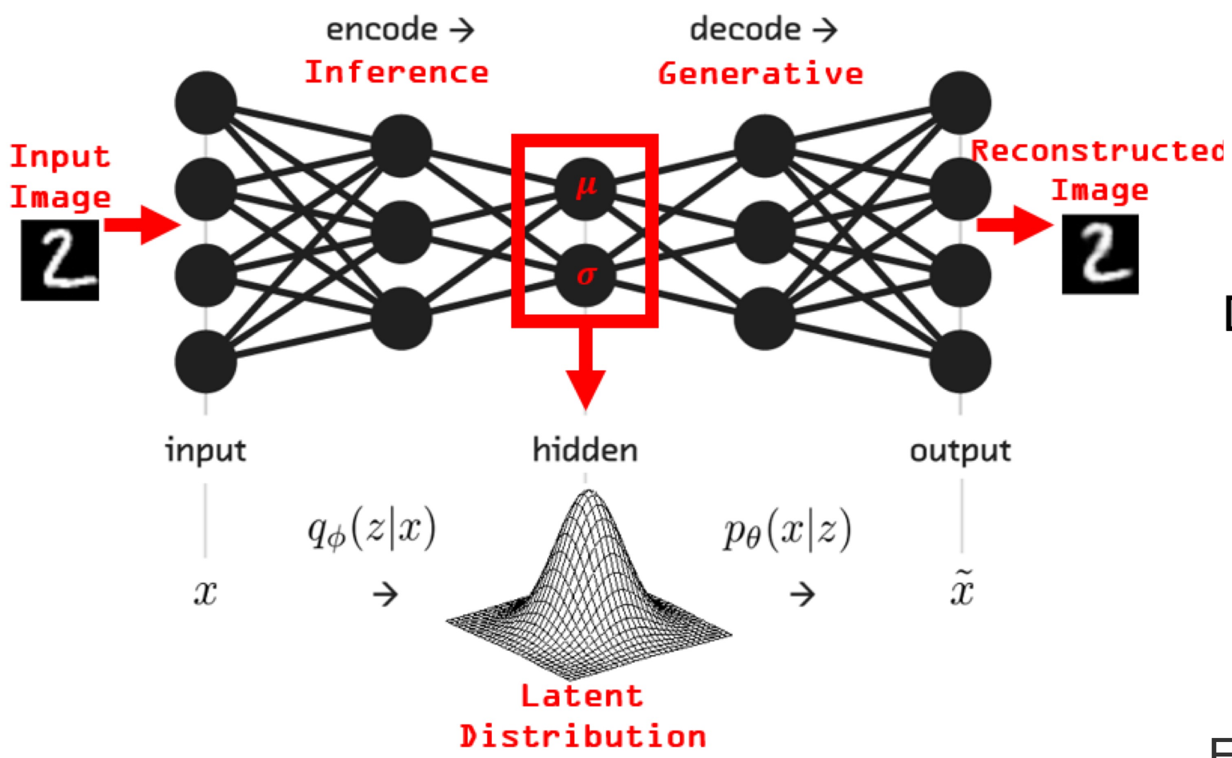
Example: VAEs for images

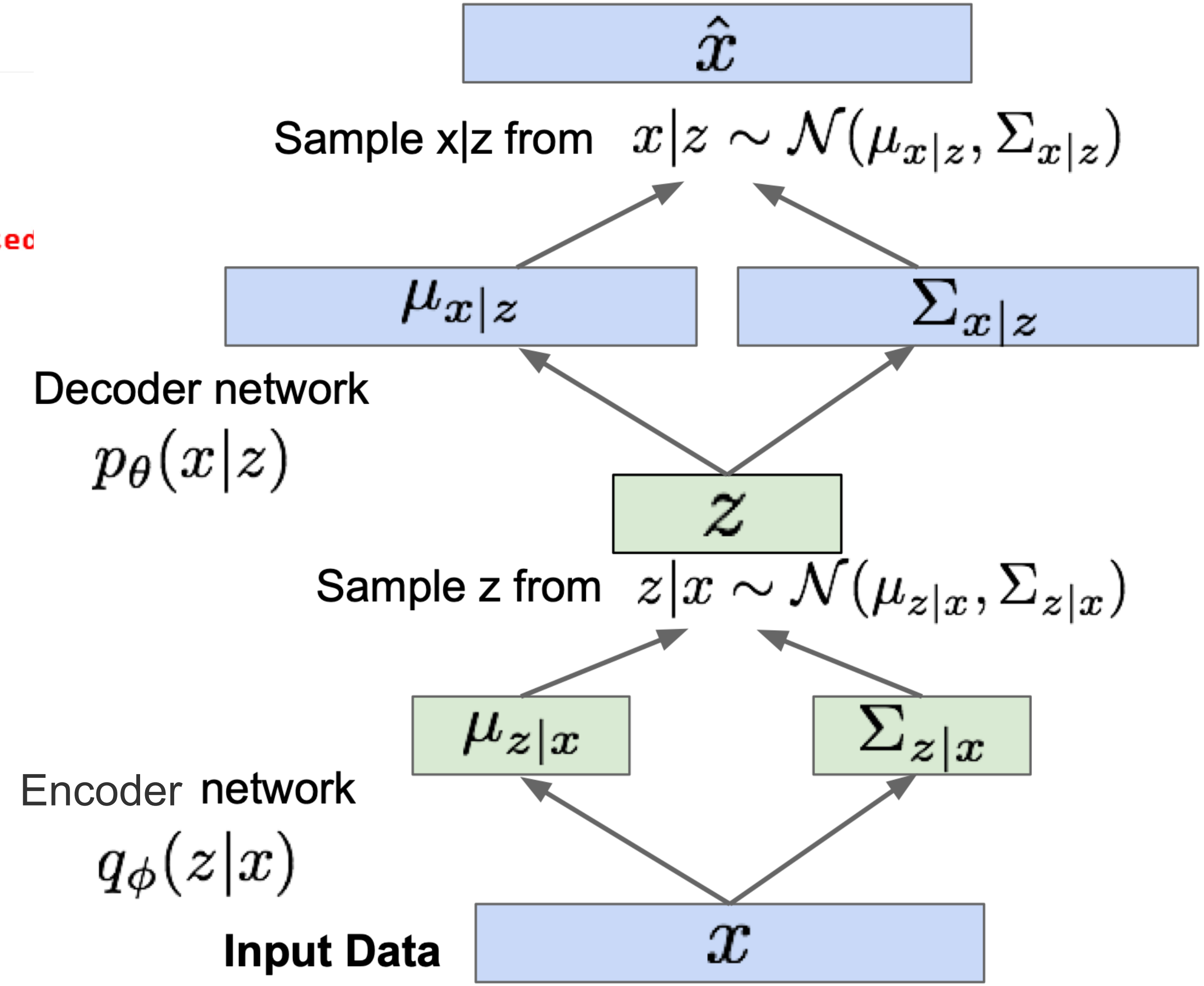




[Courtesy: Fei-Fei Li, Justin Johnson, Serena Yeung, CS 231n]

Example: VAEs for images



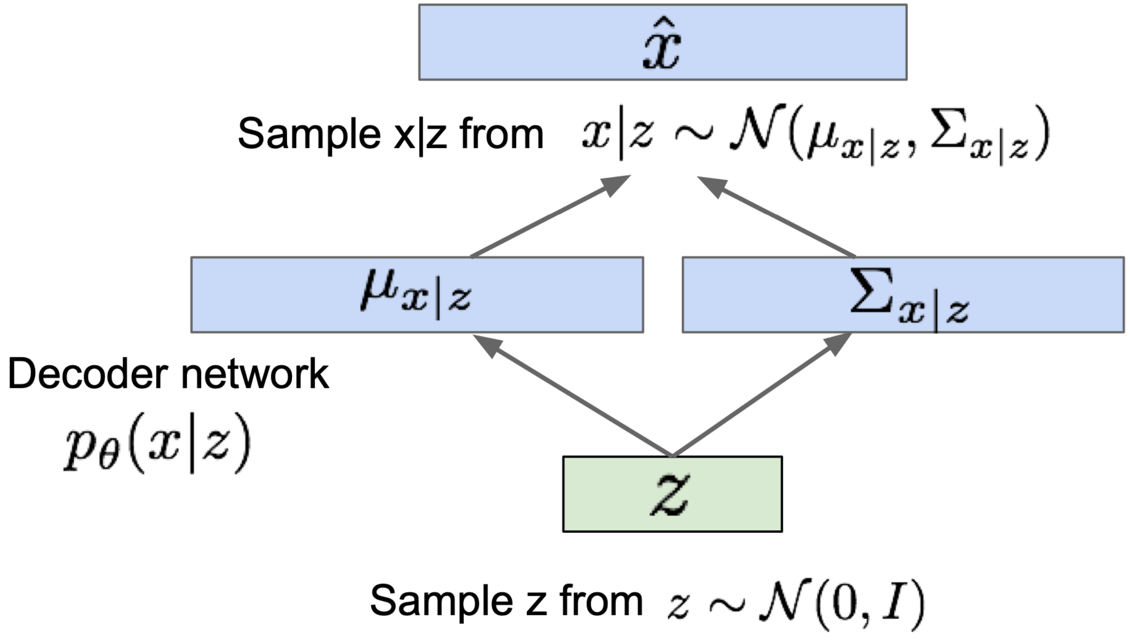


[Courtesy: Fei-Fei Li, Justin Johnson, Serena Yeung, CS 231n]

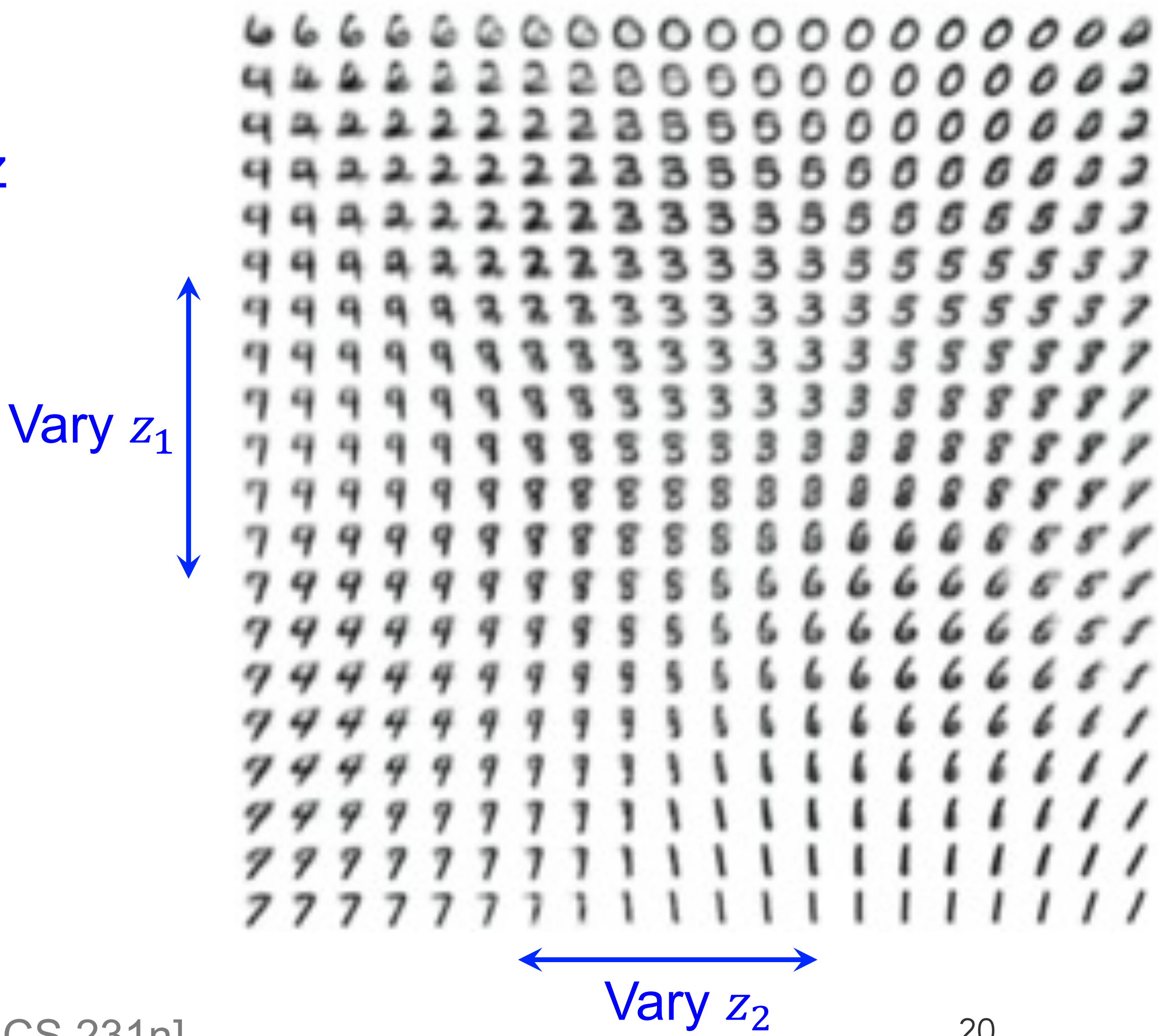
Example: VAEs for images

Generating samples:

- Use decoder network. Now sample z from prior!



Data manifold for 2-d **z**

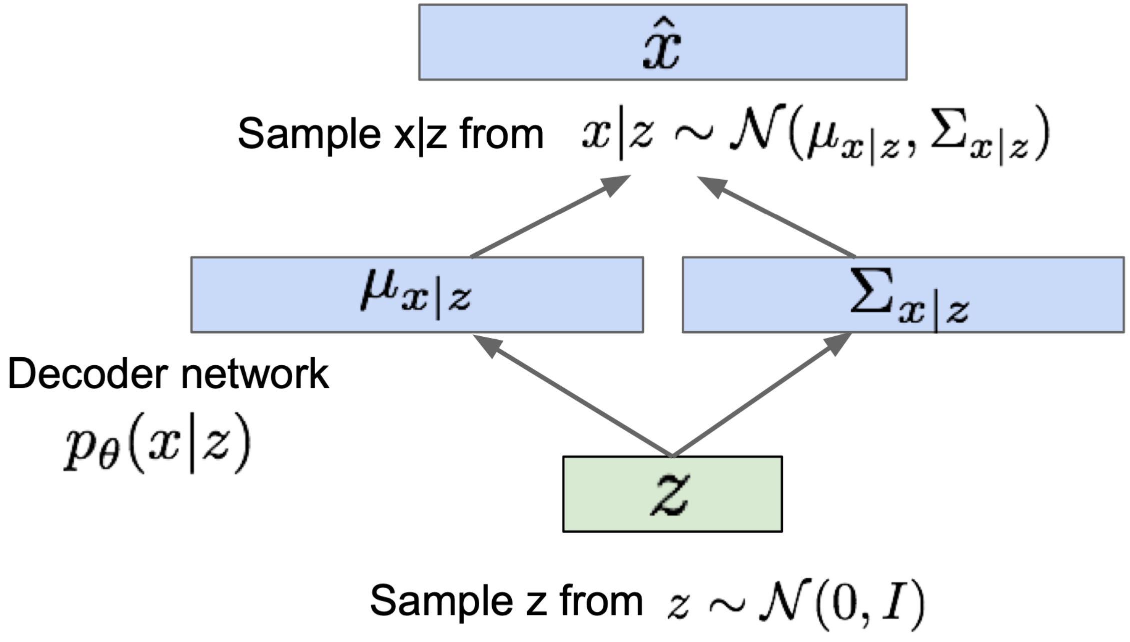


[Courtesy: Fei-Fei Li, Justin Johnson, Serena Yeung, CS 231n]

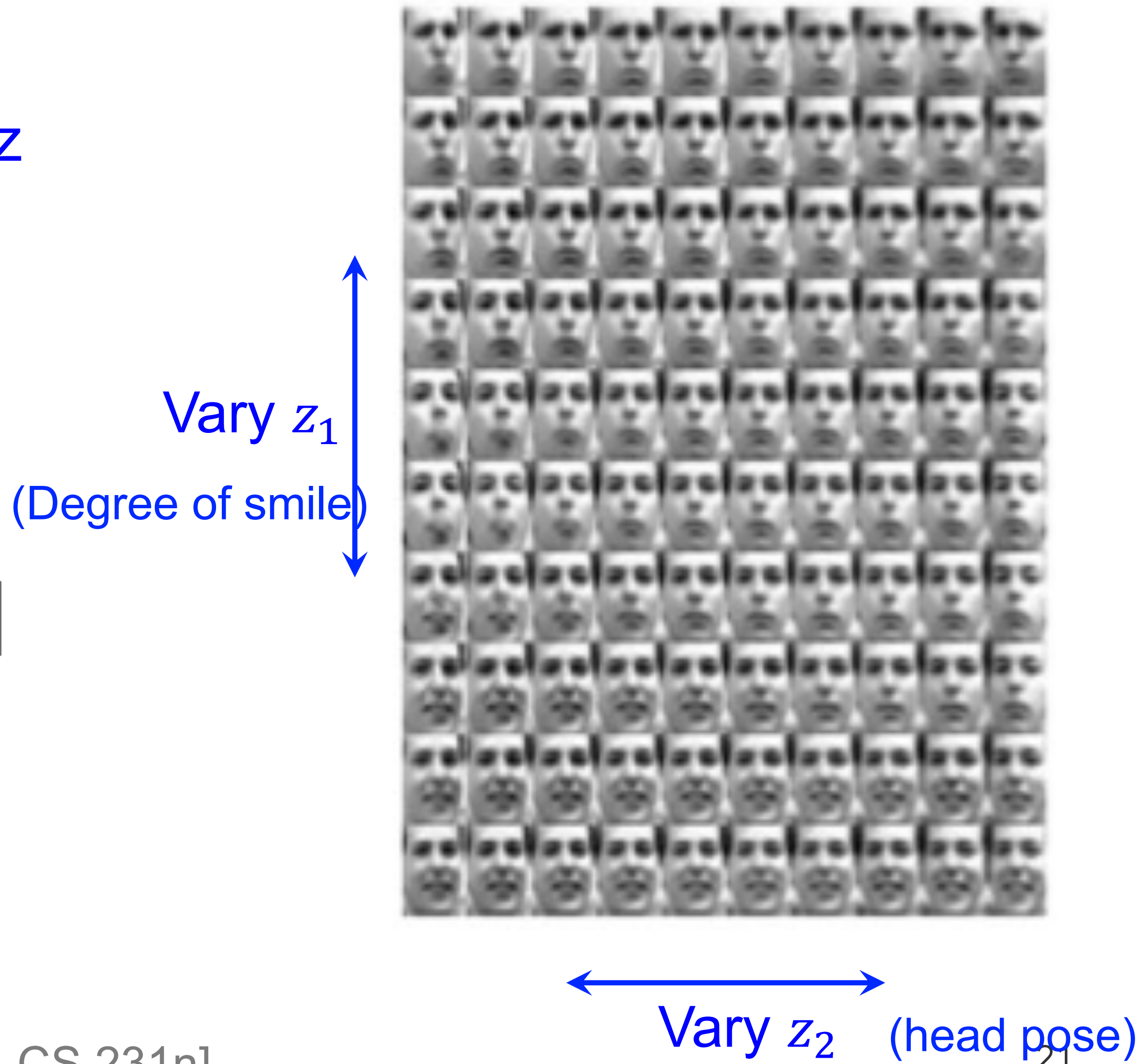
Example: VAEs for images

Generating samples:

- Use decoder network. Now sample z from prior!



Data manifold for 2-d **z**



[Courtesy: Fei-Fei Li, Justin Johnson, Serena Yeung, CS 231n]

Example: VAEs for text

- Latent code interpolation and sentences generation from VAEs [Bowman et al., 2015].

“ i want to talk to you . ”
“i want to be with you . ”
“i do n’t want to be with you . ”
i do n’t want to be with you .
she did n’t want to be with him .

Note: Amortized Variational Inference

- Variational distribution as an inference model $q_{\phi}(\boldsymbol{z}|\boldsymbol{x})$ with parameters $\boldsymbol{\phi}$ (which was traditionally factored over samples)
- Amortize the cost of inference by learning a **single** data-dependent inference model
- The trained inference model can be used for quick inference on new data

Variational Auto-encoders: Summary

- A combination of the following ideas:
  - Variational Inference: ELBO
  - Variational distribution parametrized as neural networks
  - Reparameterization trick

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{x}) = \underbrace{[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})]}_{\text{Reconstruction}} - \underbrace{\mathrm{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x}) \,||\, p(\boldsymbol{z}))}_{\text{Divergence from prior}}$$

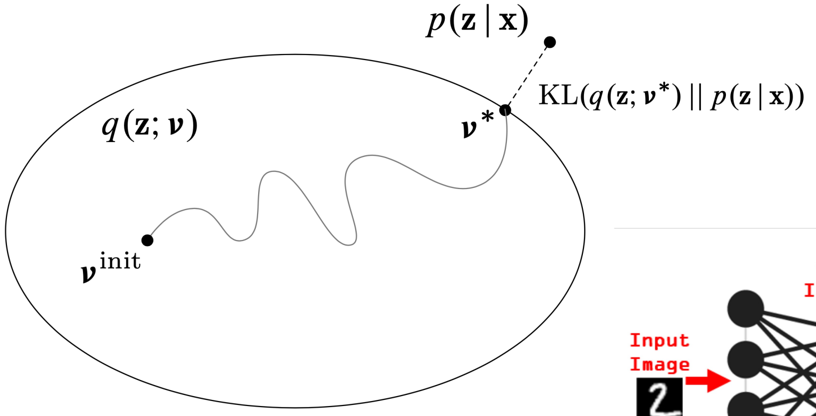
(Razavi et al., 2019)

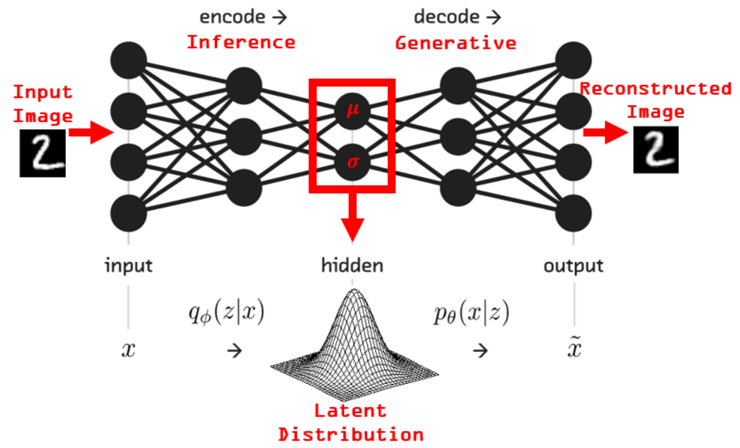
- Pros:
  - Principled approach to generative models
  - Allows inference of $q(z|x)$, can be useful feature representation for other tasks
- Cons:
  - Samples blurrier and lower quality compared to GANs
  - Tend to collapse on text data

Summary: Supervised / Unsupervised Learning

- Supervised Learning
  - Maximum likelihood estimation (MLE)

- Unsupervised learning
  - Maximum likelihood estimation (MLE) with latent variables
    - Marginal log-likelihood
  - EM algorithm for MLE
    - ELBO / Variational free energy
  - Variational Inference
    - ELBO / Variational free energy
    - Variational distributions
      - Factorized (mean-field VI)
      - Mixture of Gaussians (Black-box VI)
      - Neural-based (VAEs)





Questions?