# DSC190: Machine Learning with Few Labels

## Unsupervised Learning

**Zhiting Hu**

Lecture 18, November 8, 2024

UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE

# Outline

Unsupervised learning: Variational Auto-Encoders

Presentations

- **Bobby Zhu**: Visualizing Data using t-SNE
- **Vivian Zhao**: Exploiting Audio-Visual Features with Pretrained AV-HuBERT for Multi-Modal Dysarthric Speech Reconstruction
- **Abhinav Sanisetty**: VideoPoet: A Large Language Model for Zero-Shot Video Generation
- **Zhiqing Wang**: Divide and Conquer: Leveraging Intermediate Feature Representations for Quantized Training of Neural Networks
- **Feiyang Jiang**: LoRA: Low-Rank Adaptation of Large Language Models

# Recall: Black-box Variational Inference (BBVI)

- Probabilistic model: $\boldsymbol{x}$ -- observed variables, $\boldsymbol{z}$ -- latent variables
- Variational distribution $q_\lambda(\boldsymbol{z}|\boldsymbol{x})$ with parameters $\lambda$, e.g.,
  - Gaussian mixture distribution:
    - "A Gaussian mixture model is a universal approximator of densities, in the sense that any smooth density can be approximated with any specific nonzero amount of error by a Gaussian mixture model with enough components." (Deep Learning book, pp.65)
  - Deep neural networks

$$\mathcal{L}(\lambda) \triangleq \mathrm{E}_{q_\lambda(z)}[\log p(x, z) - \log q(z)]$$

- ELBO to be maximized:

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

- Want to compute the gradient w.r.t variational parameters $\lambda$

[Ranganath et al.,14]

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

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

# BBVI with the score gradient

- 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 score gradient w.r.t. $\lambda$ ?

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

## BBVI with the score gradient

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

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

- 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 score gradient w.r.t. $\lambda$ ?

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

- 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

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

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

- 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. $\lambda$ ?

$$\begin{array}{l}\epsilon \sim s(\epsilon) \\ z = t(\epsilon, \lambda)\end{array} \quad \Leftrightarrow \quad z \sim q(z|\lambda)$$

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

# 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

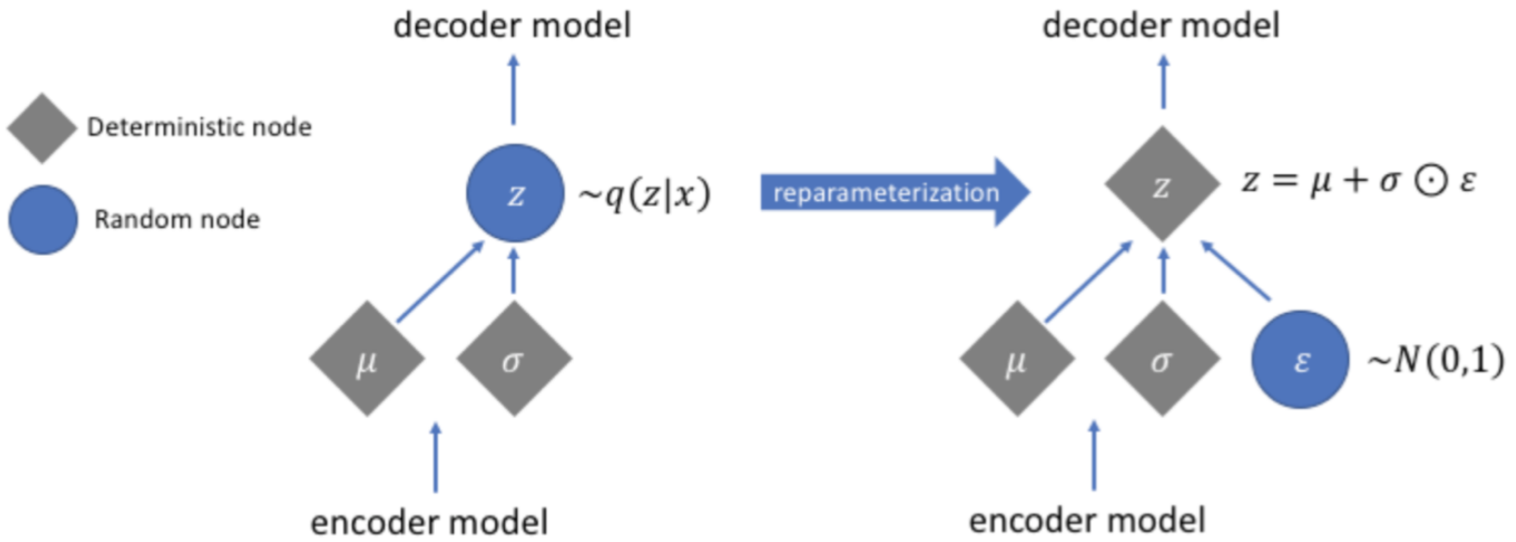
Divergence from prior
(KL divergence between two Guassians has an analytic form)

# 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:

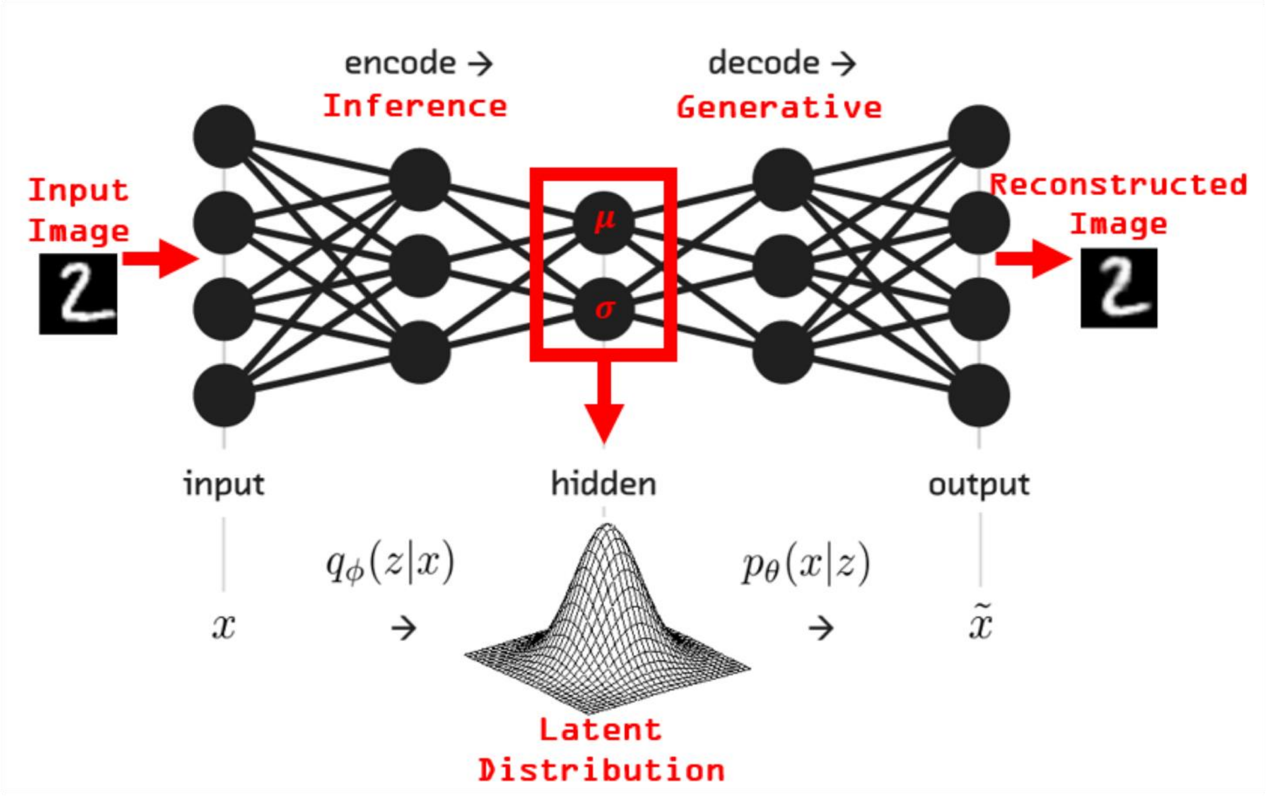
$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{x}) = \mathrm{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}, \boldsymbol{z})] + \mathrm{H}(q_\phi(\boldsymbol{z}|\boldsymbol{x}))$$
$$= \mathrm{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})]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

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

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

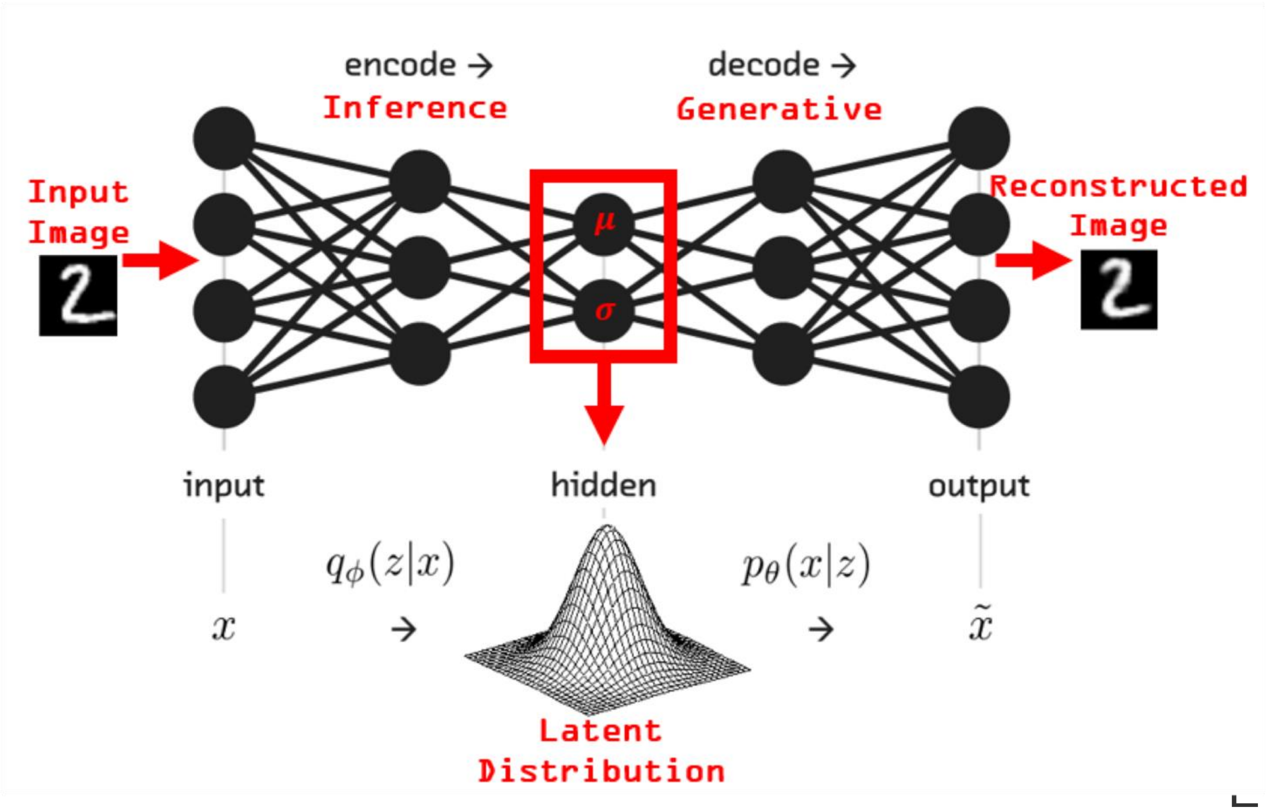
$$\nabla_\theta \mathcal{L} = \mathrm{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\nabla_\theta \log p_\theta(\boldsymbol{x}, \boldsymbol{z})]$$

# Example: VAEs for images




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

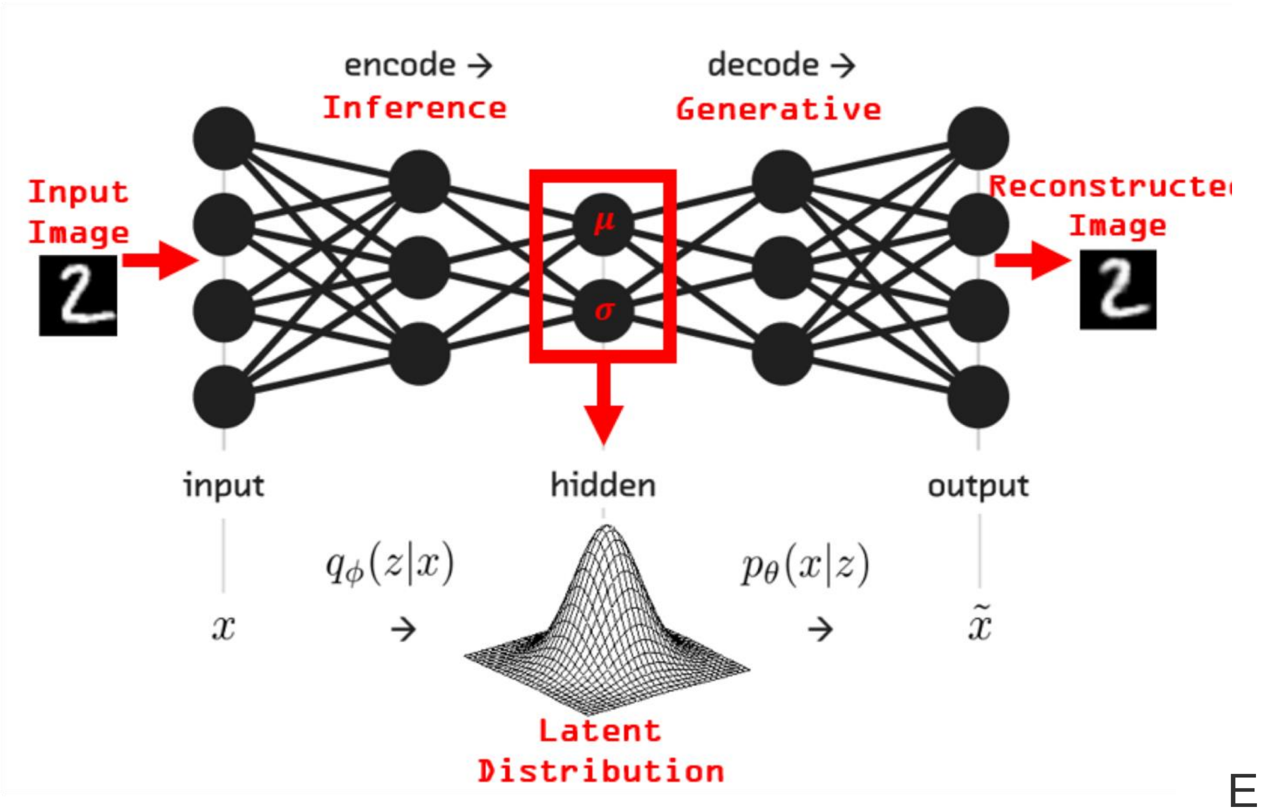
# Example: VAEs for images

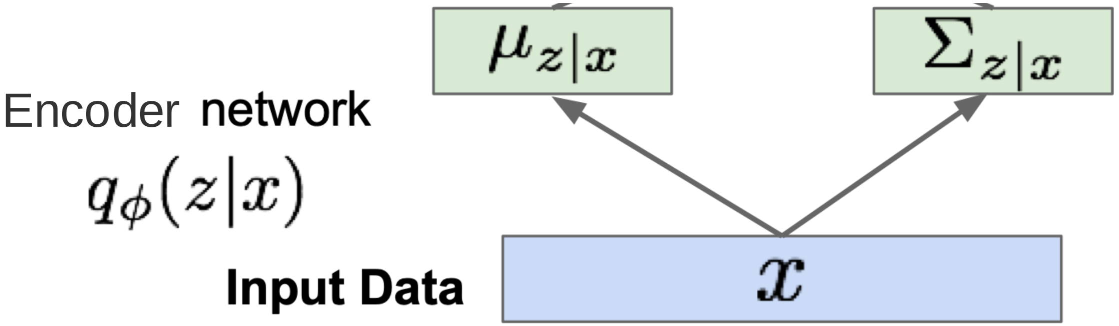



**Input Data** $x$

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

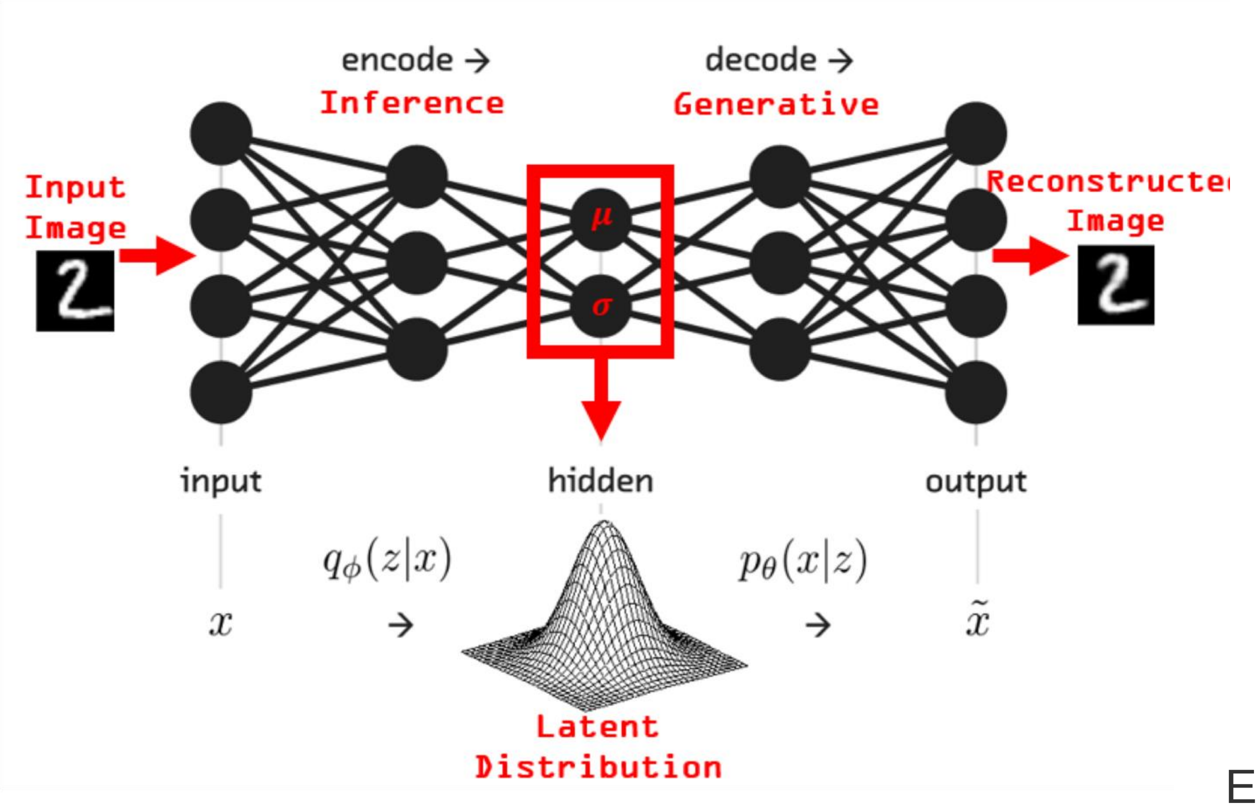
# Example: VAEs for images

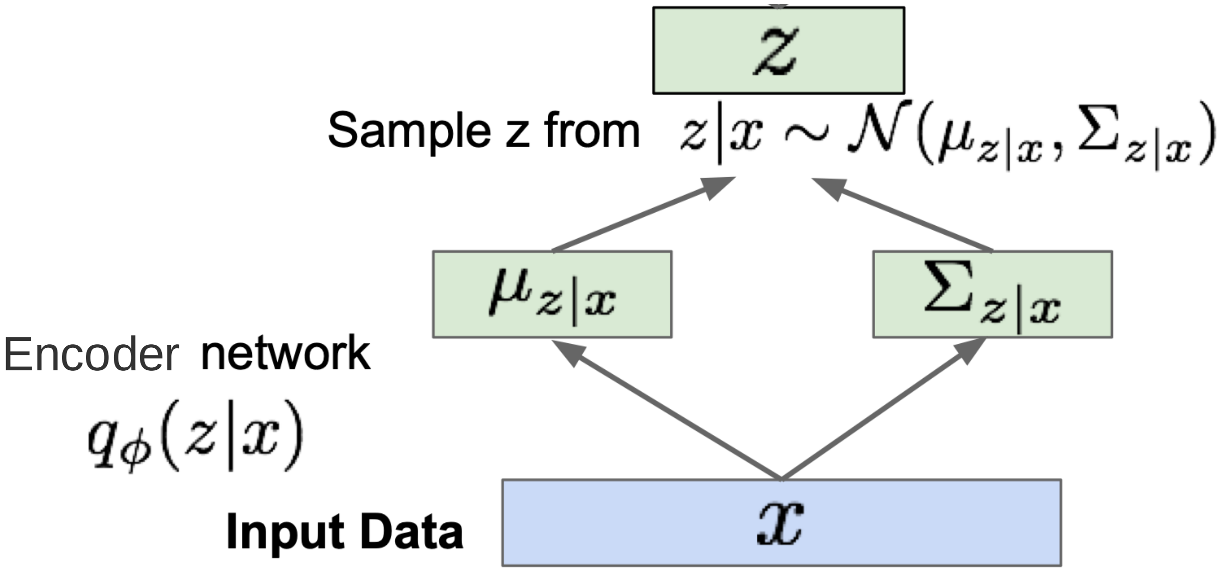



[Courtesy: Fei-Fei Li, Justin Johnson, Serena Yeung, CS 231n]

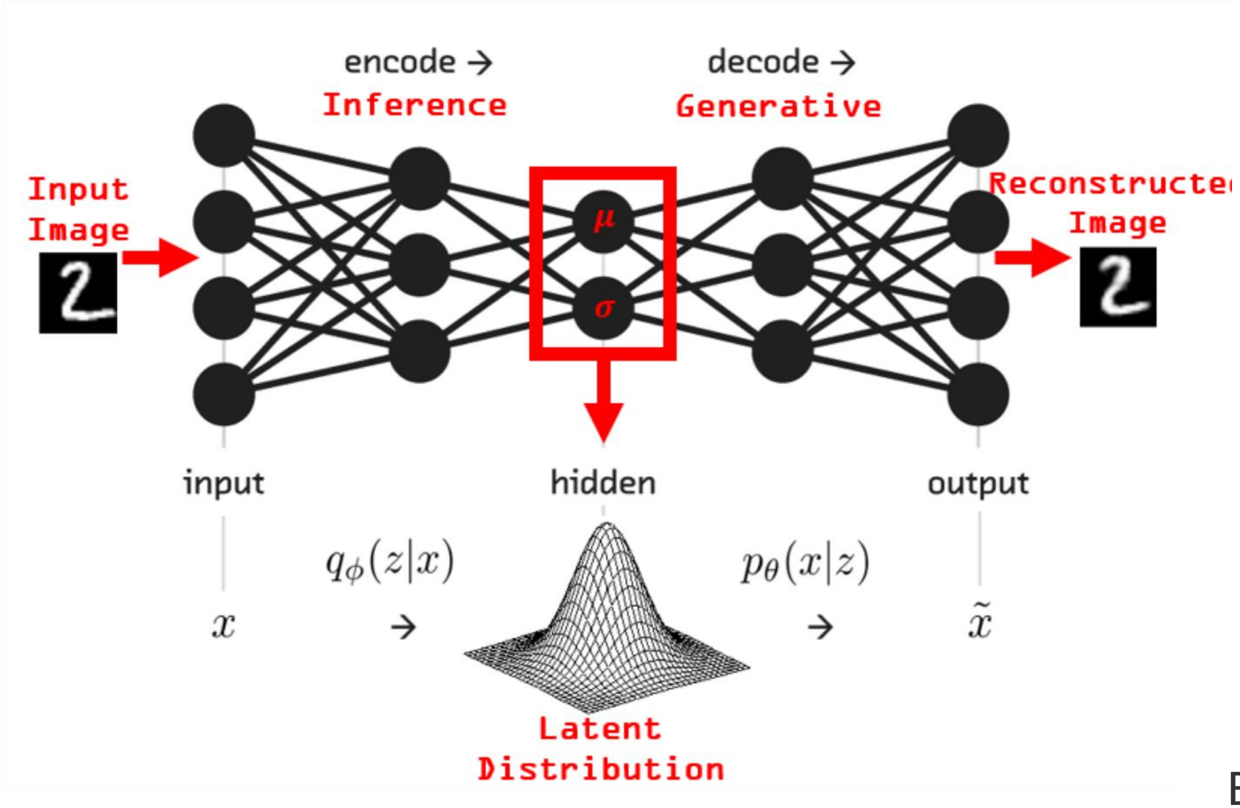
# Example: VAEs for images

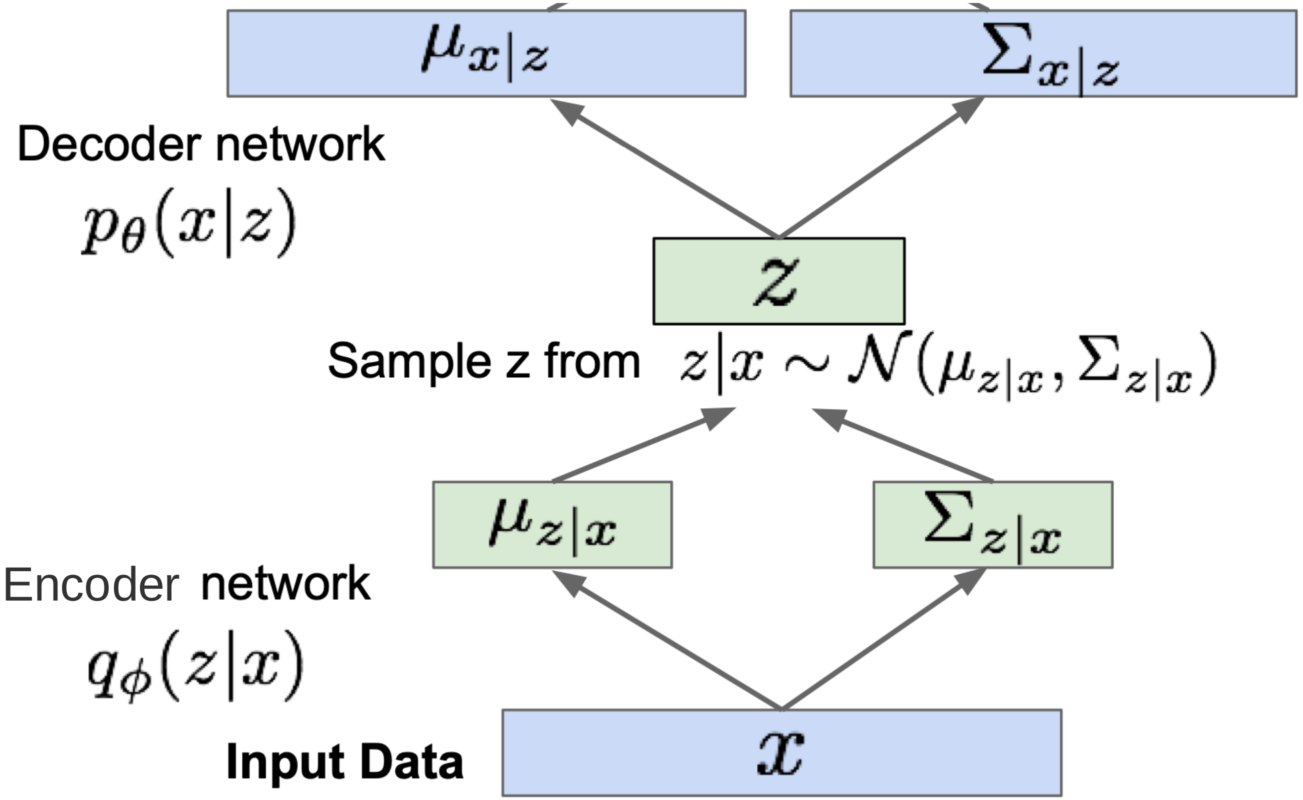



[Courtesy: Fei-Fei Li, Justin Johnson, Serena Yeung, CS 231n]

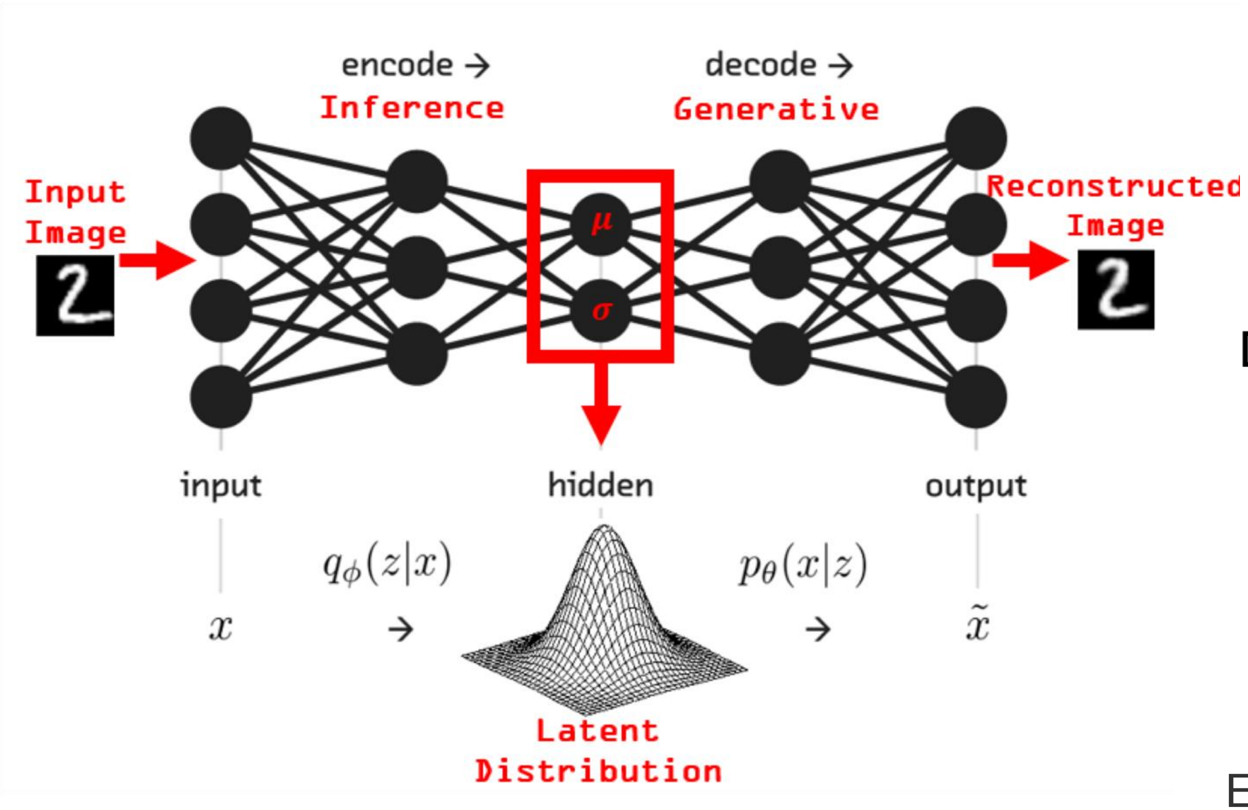
# Example: VAEs for images

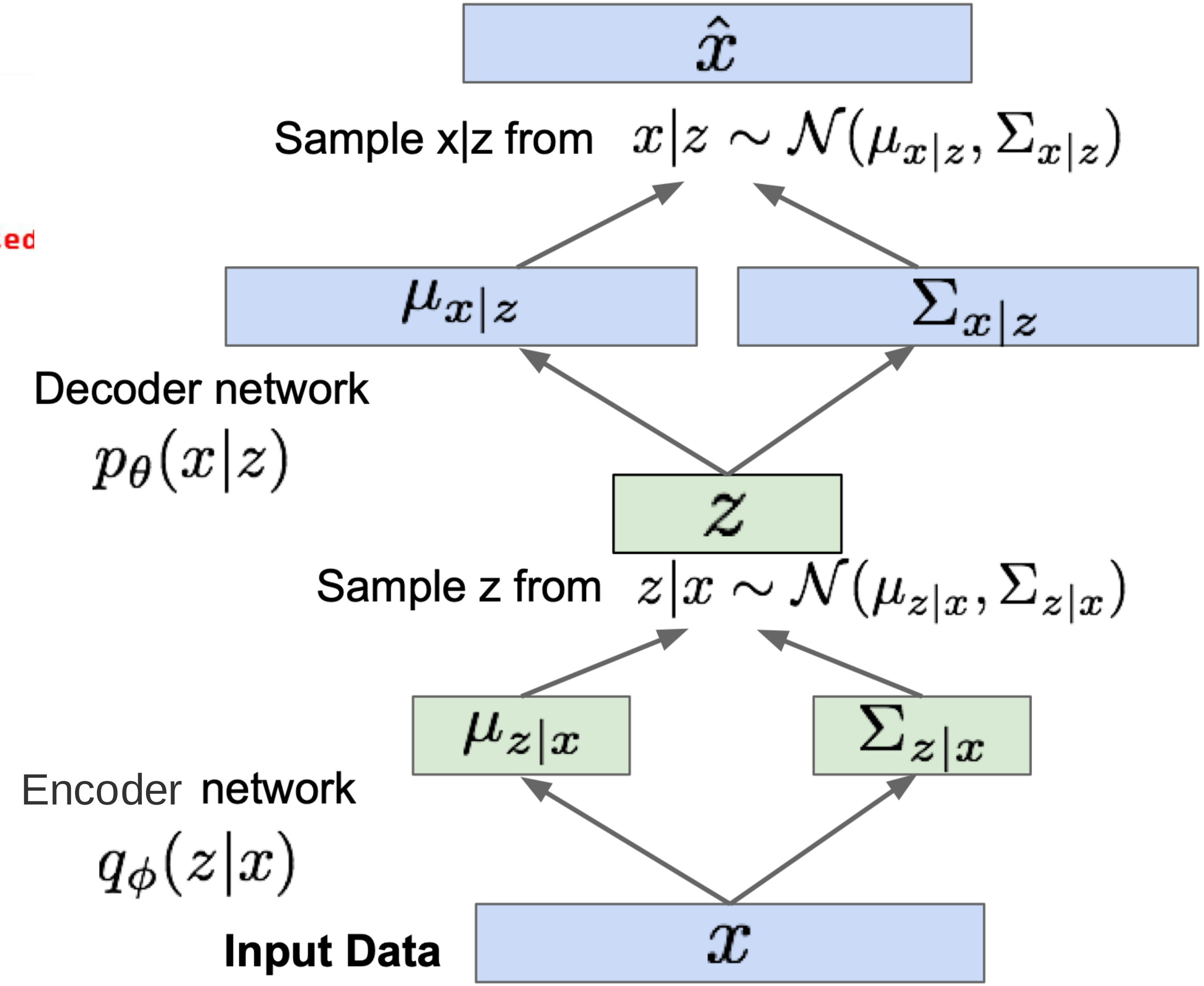



[Courtesy: Fei-Fei Li, Justin Johnson, Serena Yeung, CS 231n]

# Example: VAEs for images



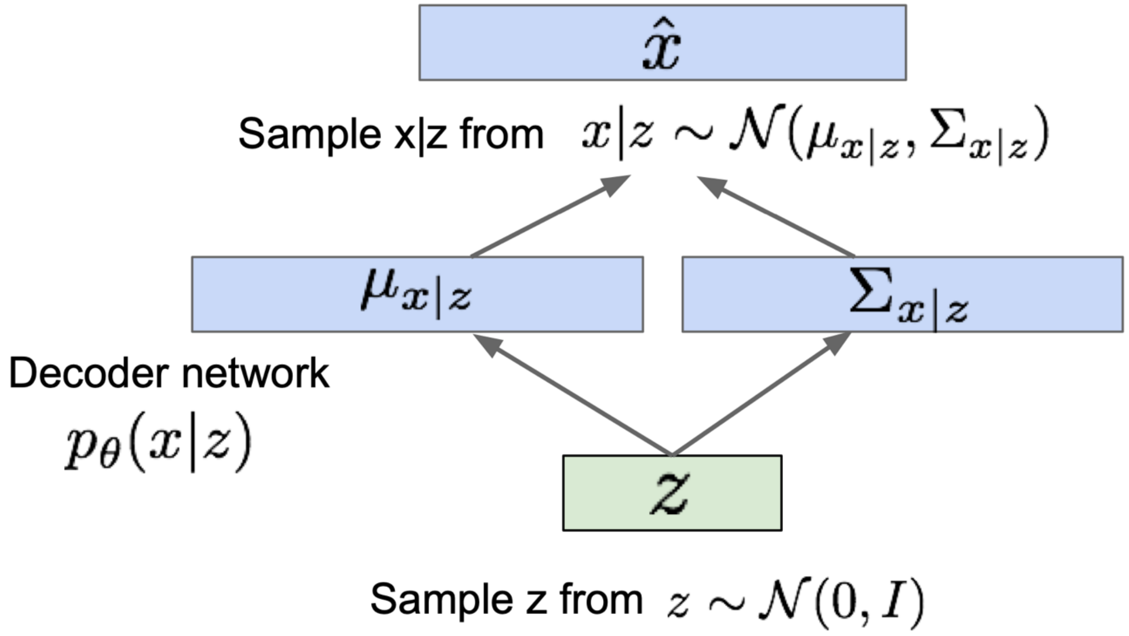




[Courtesy: Fei-Fei Li, Justin Johnson, Serena Yeung, CS 231n]

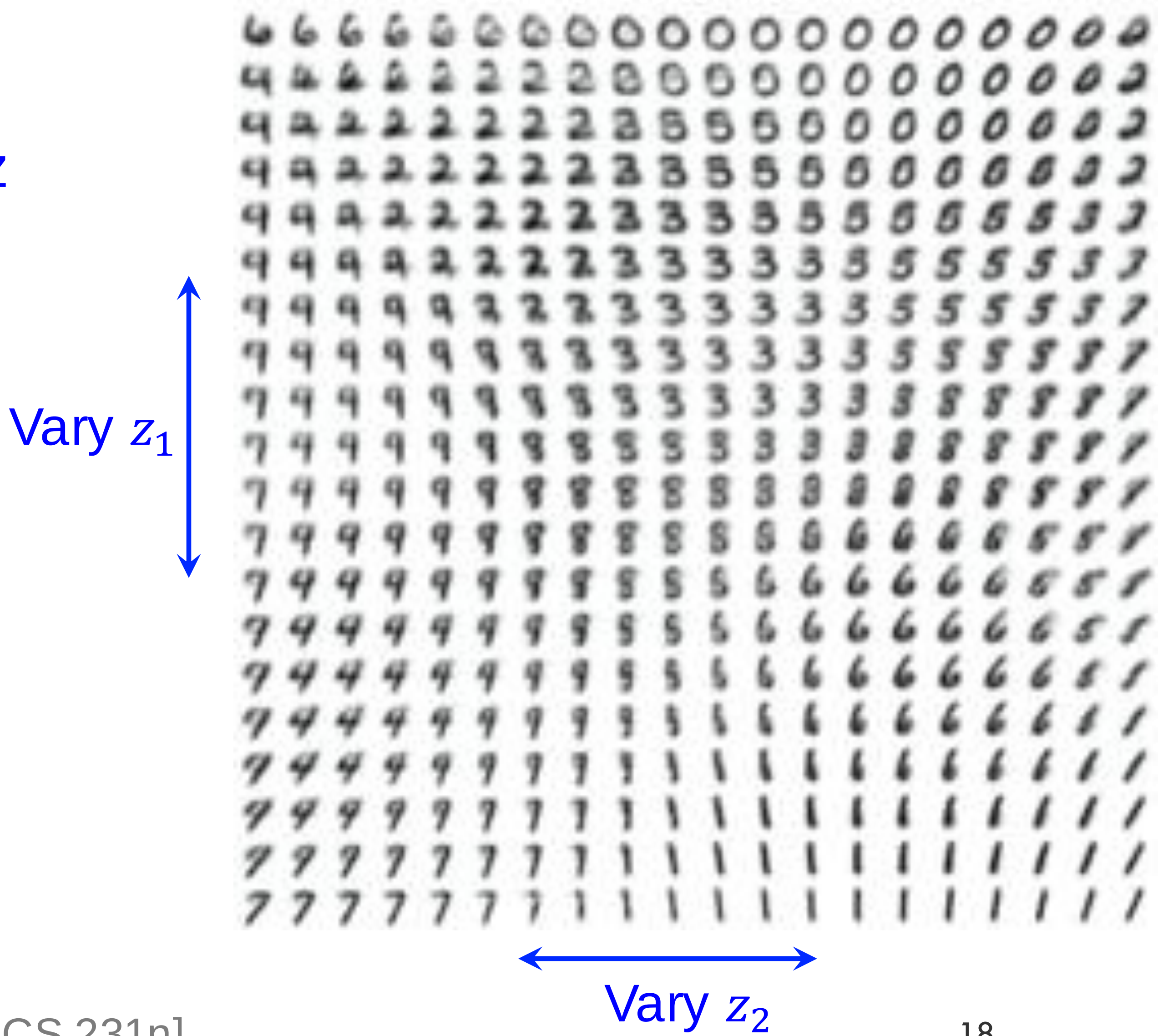
# Example: VAEs for images

Generating samples:

- Use decoder network. Now sample z from prior!




Data manifold for 2-d **z**

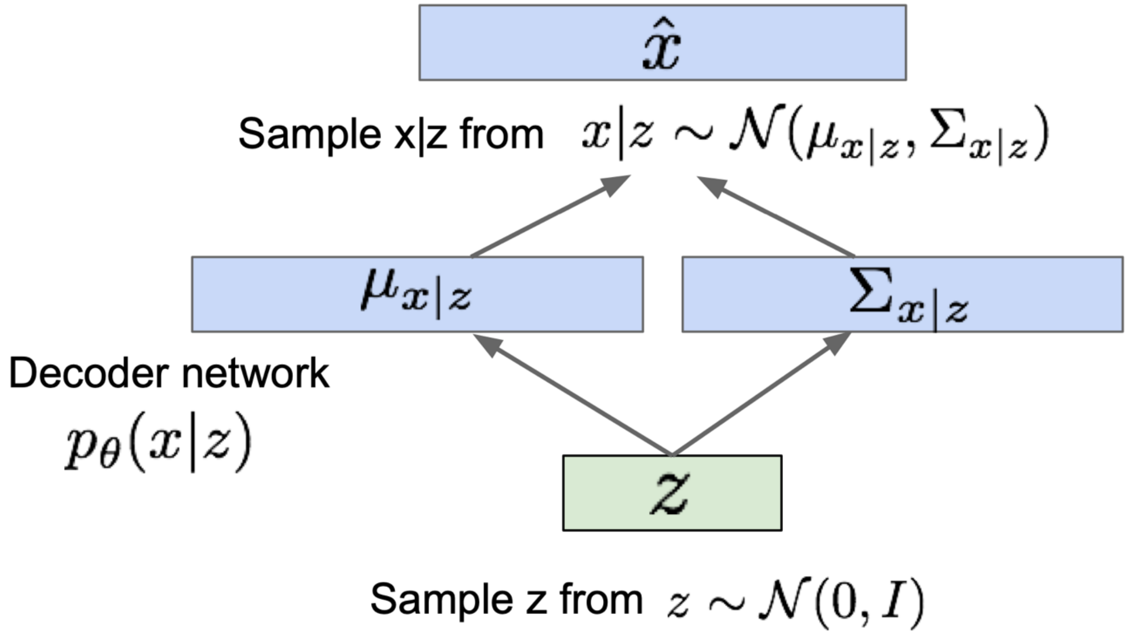



[Courtesy: Fei-Fei Li, Justin Johnson, Serena Yeung, CS 231n]

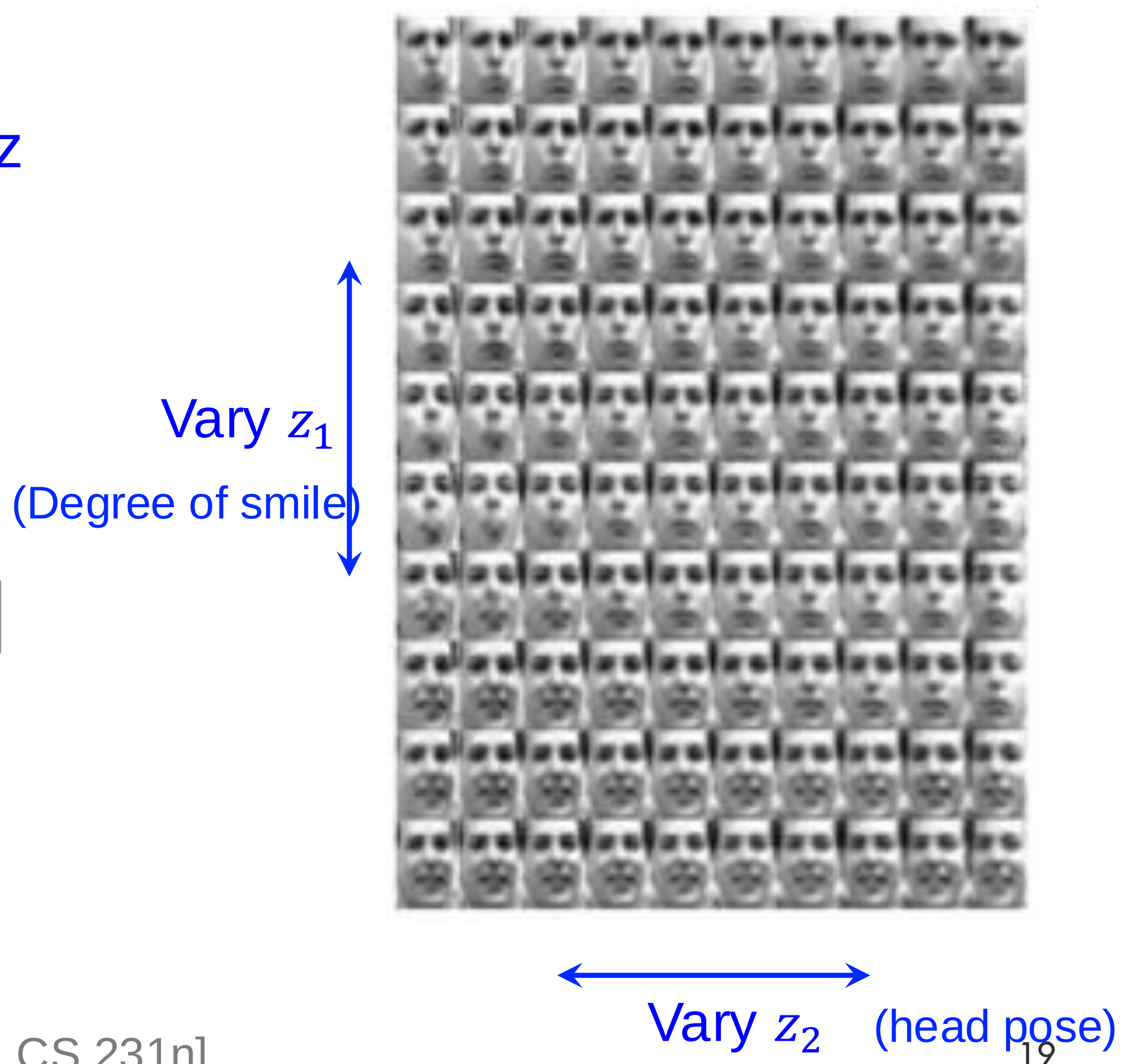
# Example: VAEs for images

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

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

Reconstruction

Divergence from prior

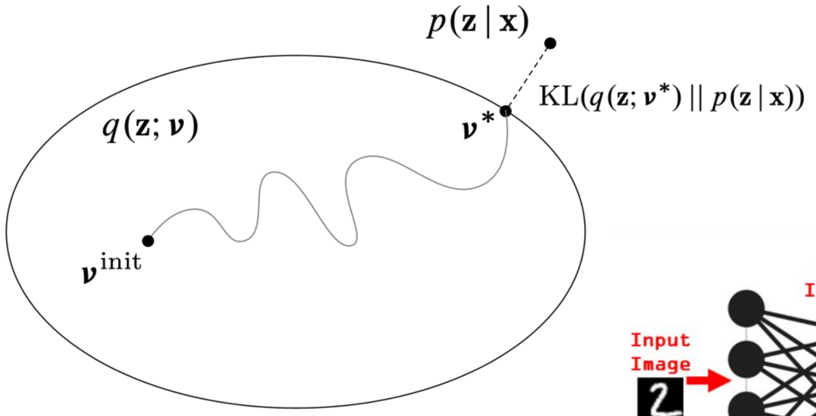
(Razavi et al., 2019)

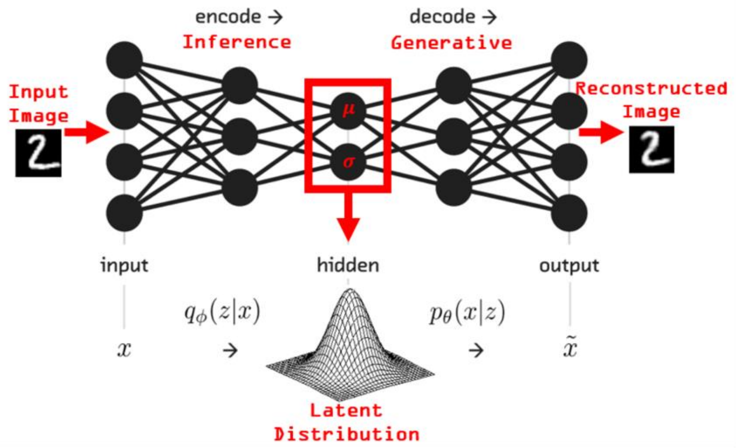
- Pros:
  - Principled approach to generative models
  - Allows inference of $q(z|x)$, can be useful feature representation for other tasks
- Cons:
  - Samples blurrier and lower quality compared to GANs
  - Tend to collapse on text data

# 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)



# Presentations

# Questions?