Lecture 17: Deep Generative Models III: VAE & GAN

Lan Xu
SIST, ShanghaiTech
Fall, 2021

Outline

- VAEs
  - Inverse graphics network
  - Attribute2Image
- Generative Adversarial Networks
  - Implicit generative models
  - Adversarial learning

Acknowledgement: Feifei Li et al's cs231n notes

The Variational Autencoder: overview

- Learning:
  - Given a large dataset of observations $\mathbf{X} = \{\mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(N)}\}$
  - Estimating the parameters in Deep LVM

$$p(x) = \int p(x, z)\ dz \quad \text{where} \quad p(\boldsymbol{x}, \boldsymbol{z}) = \boxed{p(\boldsymbol{x} \mid \boldsymbol{z})} p(\boldsymbol{z})$$

  - Based on Maximum Likelihood

$$\max \sum_{i=1}^{N} \log p(x^{(i)})$$

  - Direct optimization is challenging: use EM learning strategy
  - Jointly learning inference model with the deep latent variable model

VAE objective

- Recall lower bound of the data log likelihood

$$
\begin{aligned}
\log p_\theta(x) &= \log \int_z p_\theta(x,z)dz \\
&= \log \int_z q_\phi(z|x)\frac{p_\theta(x,z)}{q_\phi(z|x)}dz \\
&\geq \int_z q_\phi(z|x)\log \frac{p_\theta(x,z)}{q_\phi(z|x)}dz \quad \text{(Jensen's Inequality)} \\
&= \mathbf{E}_{q_\phi(z|x)}\left[\log p_\theta(x,z) - \log q_\phi(z|x)\right] = \mathcal{L}(x;\theta,\phi)
\end{aligned}
$$

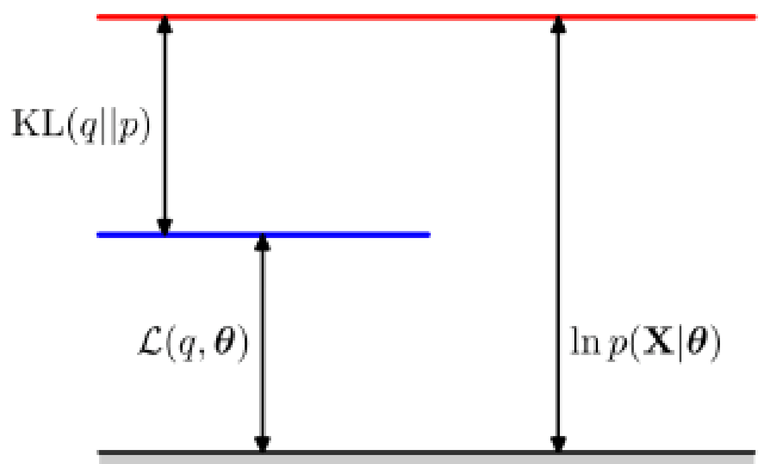
$$
\log p_\theta(x) = \boxed{\mathcal{L}(x;\theta,\phi)} + D_{KL}(q_\phi(z|x)||p_\theta(z|x))
$$

  - Learning: maximize the lower bound of data likelihood
  - The evidence lower bound (ELBO)

Review: VAEs

- Main ideas:
  - Introduce a parametric model $q_\phi(z \mid x)$ to approximate the true posterior $p_\theta(z \mid x)$
  - Jointly learn approximate posterior with the deep latent variable model
  - Variational EM: lower-bound of the Maximum Likelihood

$$\log p_\theta(x) = \boxed{\mathcal{L}(x;\theta,\phi)} + D_{KL}(q_\phi(z|x)||p_\theta(z|x))$$



Bishop – Pattern Recognition and Machine Learning

Review: VAEs

- Main ideas:
  - Introduce a parametric model $q_\phi(z \mid x)$ to approximate the true posterior $p_\theta(z \mid x)$
  - Jointly learn approximate posterior with the deep latent variable model
  - Variational EM: lower-bound of the Maximum Likelihood

$$
\begin{aligned}
\mathcal{L}(\theta, \phi, x) &= \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x, z) - \log q_\phi(z \mid x)\right] \\
&= \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x \mid z) + \log p_\theta(z) - \log q_\phi(z \mid x)\right] \\
&= -D_{\mathrm{KL}}\left(q_\phi(z \mid x) \| \, p_\theta(z)\right) + \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x \mid z)\right]
\end{aligned}
$$

***regularization* term** ***reconstruction* term**

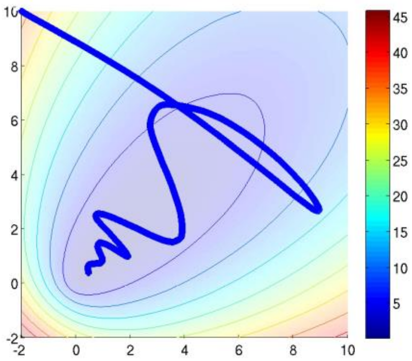
VAE learning

- EM perspective
  - Expectation Maximization alternately optimizes the ELBO, $\mathcal{L}(q,\theta)$, with respect to q (the E step) and θ (the M step)
  - Initialize $\theta^{(0)}$
  - At each iteration $t = 1, \ldots$
    - **E step:** Hold $\theta^{(t-1)}$ fixed, find $q^{(t)}$ which maximizes $\mathcal{L}\left(q, \theta^{(t-1)}\right)$
    - **M step:** Hold $q^{(t)}$ fixed, find $\theta^{(t)}$ which maximizes $\mathcal{L}\left(q^{(t)}, \theta\right)$

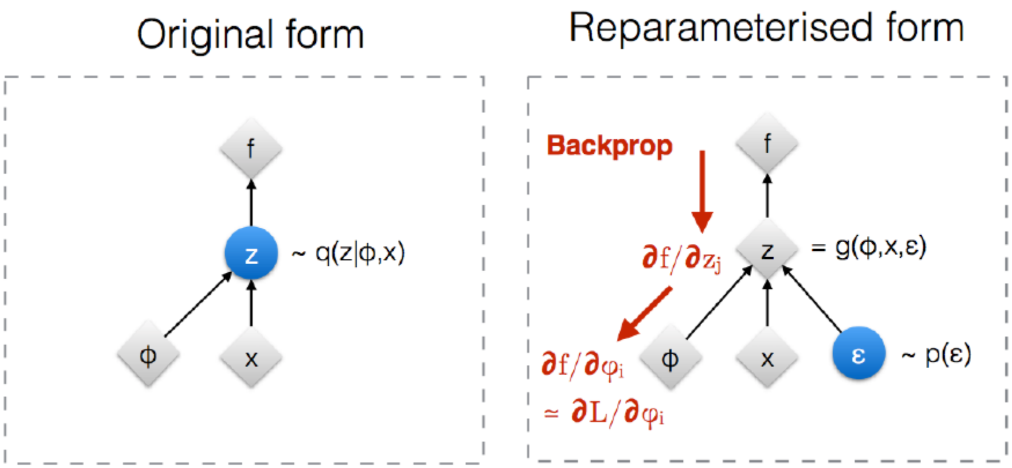
$$\mathcal{L}(x,\phi,\theta) = E_{q_\phi(z|x)}[\log p_\theta(x,z) - \log q_\phi(z|x)]$$

Two views of Learning VAE

- Optimization interpretation
  - Stochastic gradient-based



- Network interpretation
  - Backpropagation



Optimization interpretation

- Recall VAE objective

$$\mathcal{L}(x, \phi, \theta) = E_{q_\phi(z|x)}[\log p_\theta(x, z) - \log q_\phi(z|x)]$$

  - Or rewrite as $\mathcal{L}(x, \phi, \theta) = E_{q_\phi(z|x)}[f_{\phi,\theta}(x, z)]$

- Often no analytic solution to exact gradient

$$\nabla_{\phi,\theta}\mathcal{L}(x, \phi, \theta)$$

  - Solution: stochastic gradient ascent
  - Requires unbiased estimates of gradient
  - Can use small minibatches or single point of data

$$\nabla_\phi \mathcal{L}(x, \phi, \theta) \approx \nabla_\phi f_{\phi,\theta}(x, z^{(i)}), \quad \boxed{z^{(i)} \sim q_\phi(z|x)}$$

High variance for gradient estimation

Reparameterization trick

- Reparameterize $\mathbf{z}^{(i)} \sim q_\phi(\mathbf{z}|\mathbf{x})$ using a differentiable transformation of an auxiliary noise variable $\boldsymbol{\epsilon}$

$$\mathbf{z} = g_\phi(\epsilon, \mathbf{x}) \quad \text{with} \quad \epsilon \sim q(\epsilon)$$

  - Then we can write the ELBO as

$$\mathcal{L}(x, \phi, \theta) = E_{q_\phi(z|x)}[f_{\phi,\theta}(x, z)] = E_{q(\epsilon)}[f_{\phi,\theta}(x, g_\phi(\epsilon, \mathbf{x}))]$$

  - And its gradient estimation with L samples

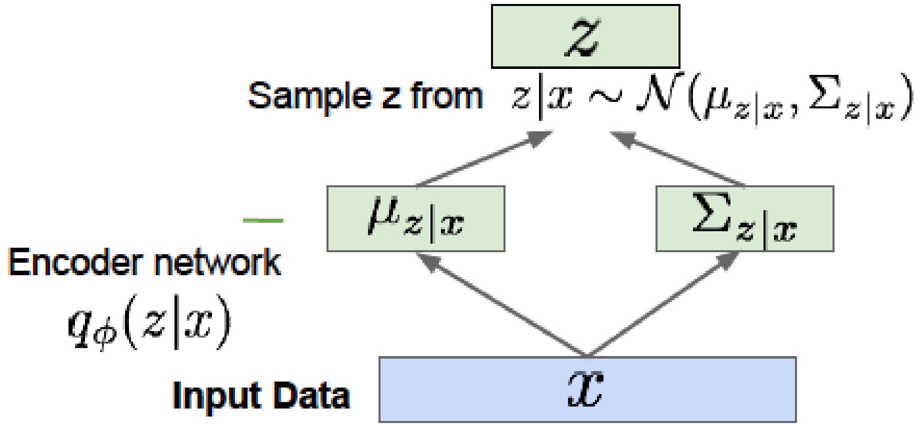
$$\nabla_\phi \mathcal{L}(x, \phi, \theta) = E_{q(\epsilon)}[\nabla_\phi f_{\phi,\theta}(x, z)] \approx \frac{1}{L} \sum_{i=1}^{L} \nabla_\phi f_{\phi,\theta}(x, g_\phi(\epsilon^{(i)}, x)), \quad \boxed{\epsilon^{(i)} \sim q(\epsilon)}$$

VAE Example

- Univariate Gaussian $z \sim p(z|x) = \mathcal{N}(\mu, \sigma^2)$

$$z = \mu + \sigma\epsilon \qquad \epsilon \sim \mathcal{N}(0,1)$$

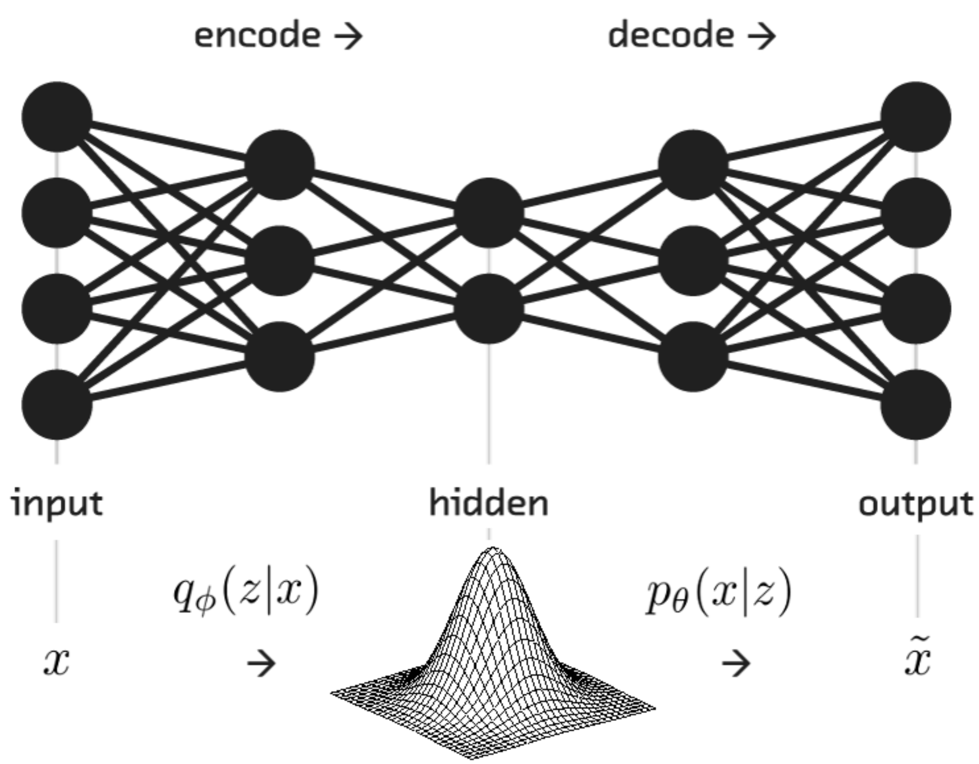
$$\mathbb{E}_{\mathcal{N}(z;\mu,\sigma^2)}\left[f(z)\right] = \mathbb{E}_{\mathcal{N}(\epsilon;0,1)}\left[f(\mu+\sigma\epsilon)\right] \simeq \frac{1}{L}\sum_{l=1}^{L} f(\mu+\sigma\epsilon^{(l)})$$



Autoencoder Interpretation

- Objective $\mathcal{L}(x, \phi, \theta) = -D_{KL}(q_\phi(z|x)||p_\theta(z)) + E_{q_\phi(z|x)}[\log p_\theta(x|z)]$

Regularization term Reconstruction term



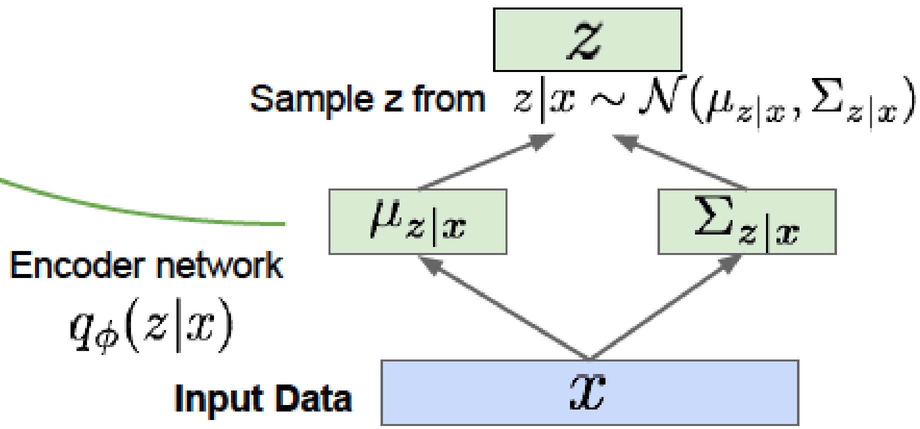
VAE Example

- Learning objective

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 Example

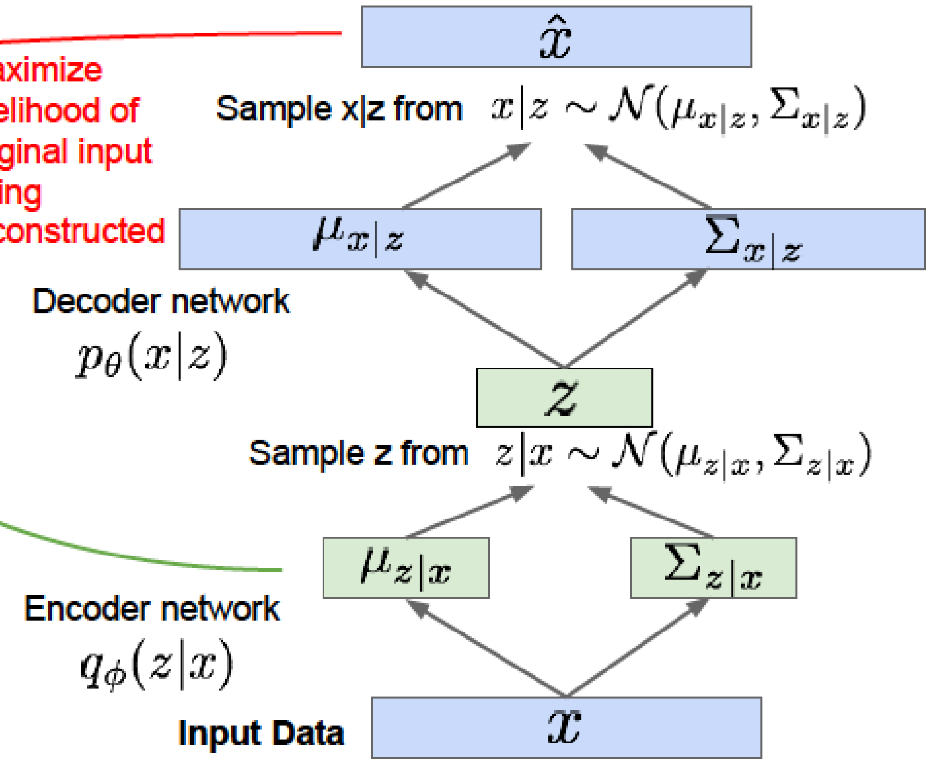
- Learning objective

Putting it all together: maximizing the likelihood lower bound

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

Maximize likelihood of original input being reconstructed

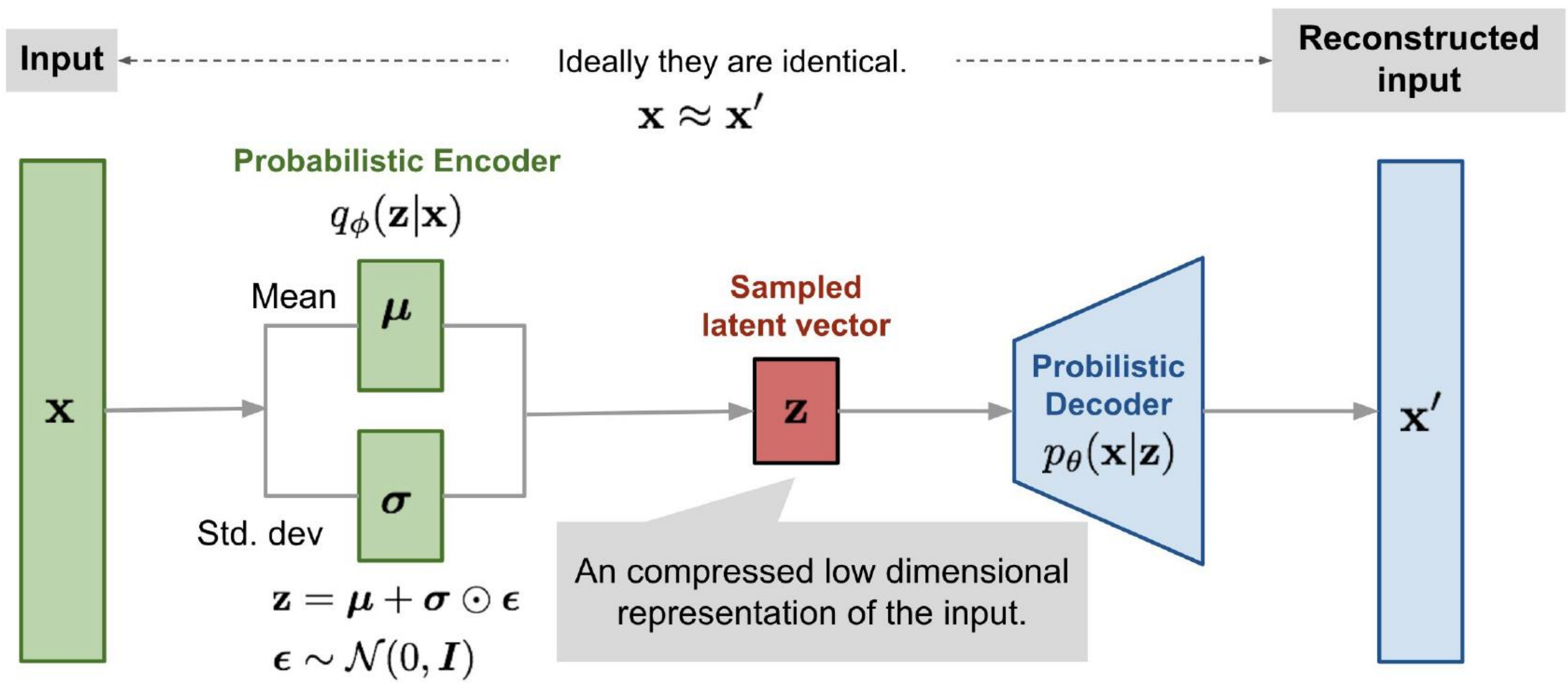
Make approximate posterior distribution close to prior



Autoencoder Interpretation

- Objective $\mathcal{L}(x, \phi, \theta) = -D_{KL}(q_\phi(z|x)||p_\theta(z)) + E_{q_\phi(z|x)}[\log p_\theta(x|z)]$

Regularization term Reconstruction term



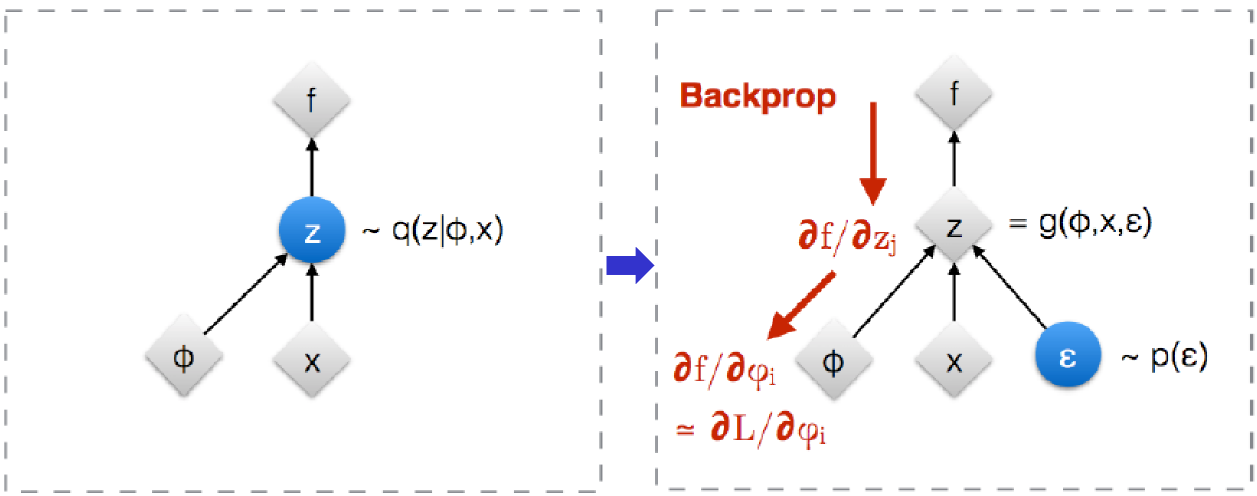
- The objective function can be represented as an Autoencoder-like computation graph.

Network interpretation

$$\mathcal{L}(x,\phi,\theta) = E_{q_\phi(z|x)}[f_{\phi,\theta}(x,z)]$$

$$\mathcal{L}(x,\phi,\theta) = E_{q(\epsilon)}[f_{\phi,\theta}(x,z)] \approx \frac{1}{L}\sum_{i=1}^{L} f_{\phi,\theta}(x, g_\phi(\epsilon^{(i)},x)), \quad \epsilon^{(i)} \sim q(\epsilon)$$

Original form | Reparameterised form



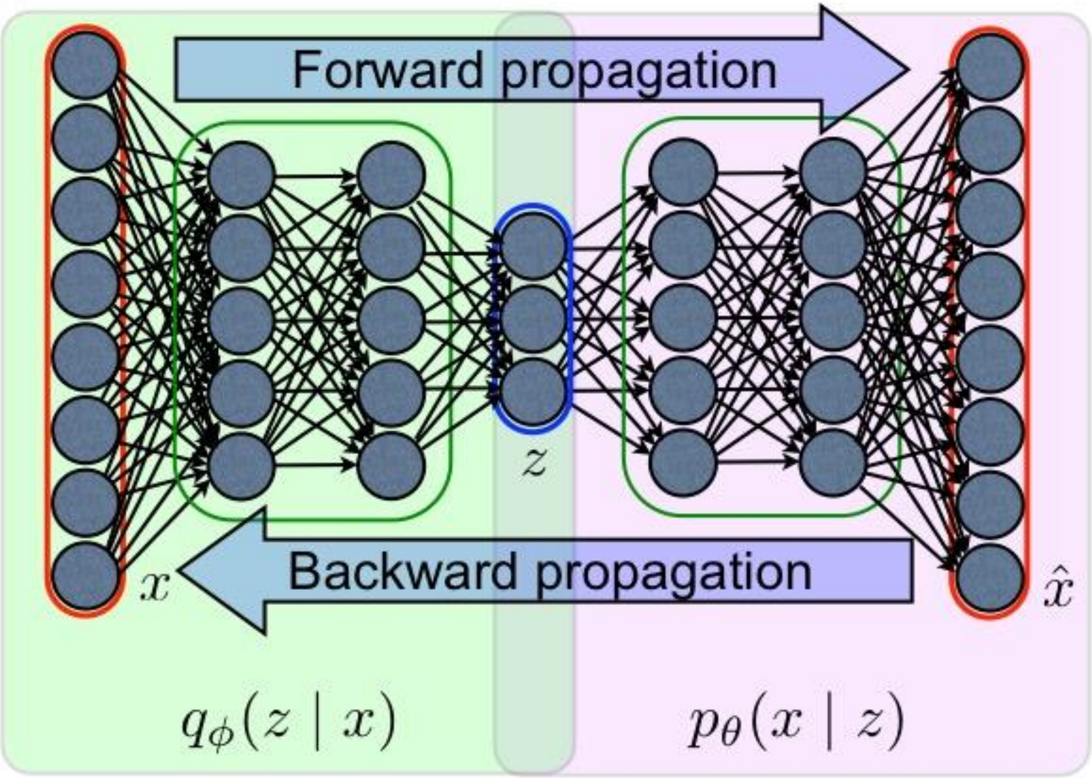
: Deterministic node

: Random node

[Kingma, 2013]
[Bengio, 2013]
[Kingma and Welling 2014]
[Rezende et al 2014]

Training with Backpropagation

- Due to reparametrization trick, we can simultaneously train both the generative model and the inference model by optimizing the variational bound using the gradient backpropagation.



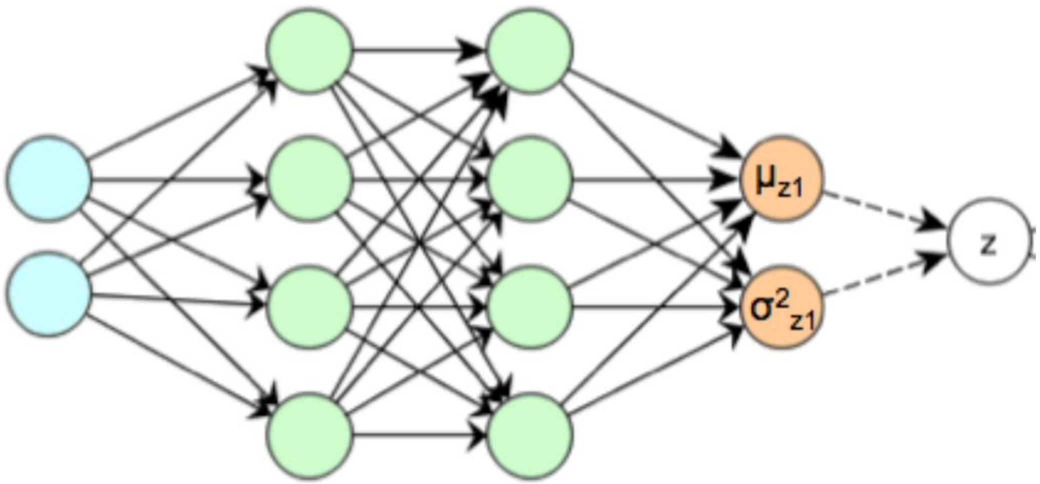
1D Gaussian Case

- We can compute the KL regularization in close form

Use N(0,1) as prior for p(z)

$q(z|x^{(i)})$ is Gaussian with parameters $(\mu^{(i)},\sigma^{(i)})$ determined by NN

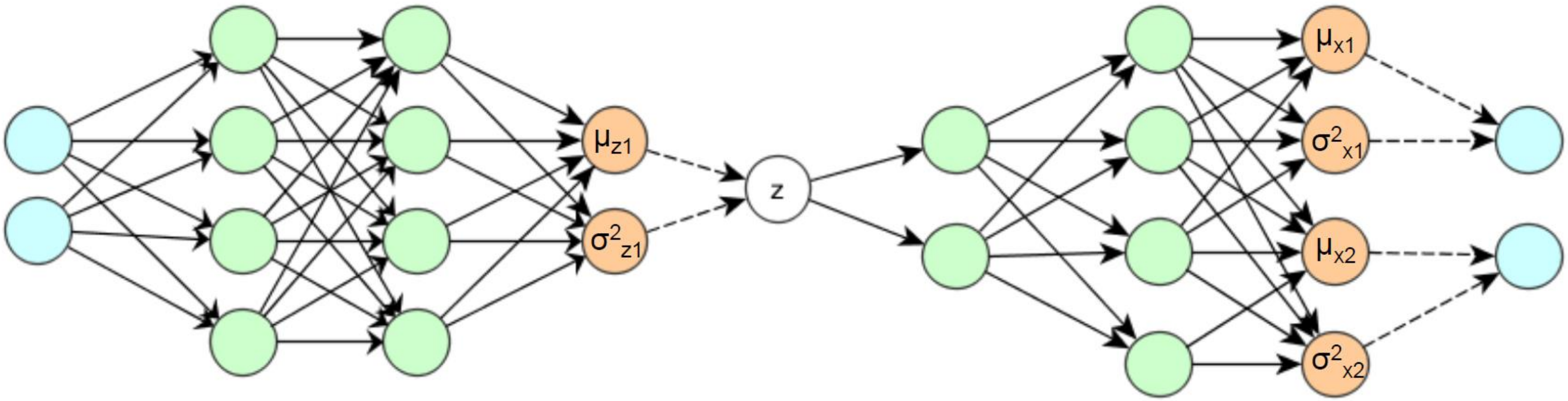
$$-D_{KL}\left(q(z|x^{(i)})||p(z)\right) = \frac{1}{2}\sum_{j=1}^{J}\left(1 + \log(\sigma_{z_j}^{(i)^2}) - \mu_{z_j}^{(i)^2} - \sigma_{z_j}^{(i)^2}\right)$$



1D Gaussian Case

- Overall loss function for BP

Prior p(z) ~ N(0,1) and p, q Gaussian, extension to dim(z) > 1 trivial



Cost: Regularisation

$$-D_{KL}\left(q(z|x^{(i)})||p(z)\right) = \frac{1}{2}\sum_{j=1}^{J}\left(1 + \log(\sigma_{z_j}^{(i)^2}) - \mu_{z_j}^{(i)^2} - \sigma_{z_j}^{(i)^2}\right)$$

We use mini batch gradient decent to optimize the cost function over all $x^{(i)}$ in the mini batch

Cost: Reproduction

$$-\log\left(p(x^{(i)}|z^{(i)})\right) = \sum_{j=1}^{D}\frac{1}{2}\log(\sigma_{x_j}^2) + \frac{(x_j^{(i)} - \mu_{x_j})^2}{2\sigma_{x_j}^2}$$

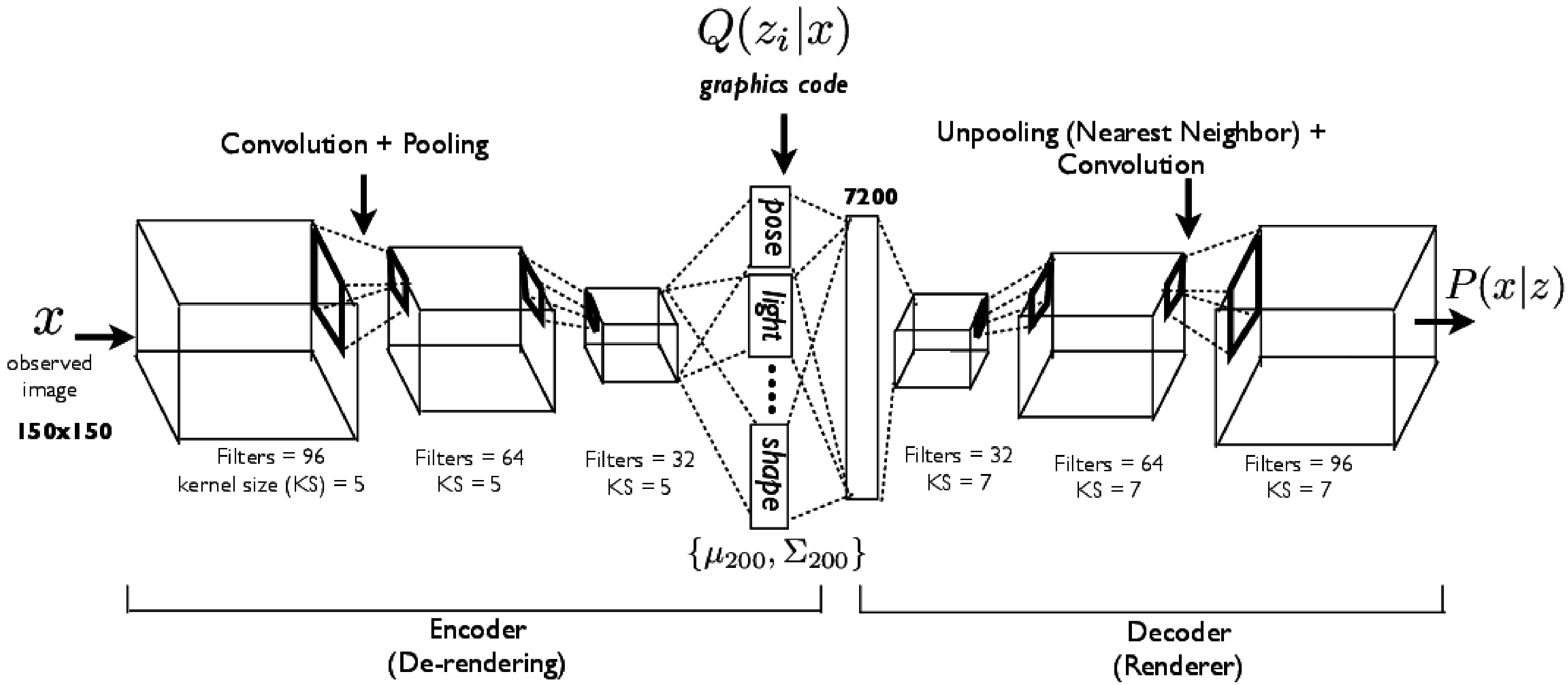
Least Square for constant variance

Interpreting the latent space

https://arxiv.org/pdf/1610.00291.pdf

Vision task I – Inverse graphics network

- Main ideas



	z_1	z_2	z_3	$z_{[4,n]}$
$Z =$	z_1	z_2	z_3	$z_{[4,n]}$
corresponds to	ϕ	α	ϕ_L	intrinsic properties (shape, texture, etc)

Vision task I – Inverse graphics network

- Mini-batch training

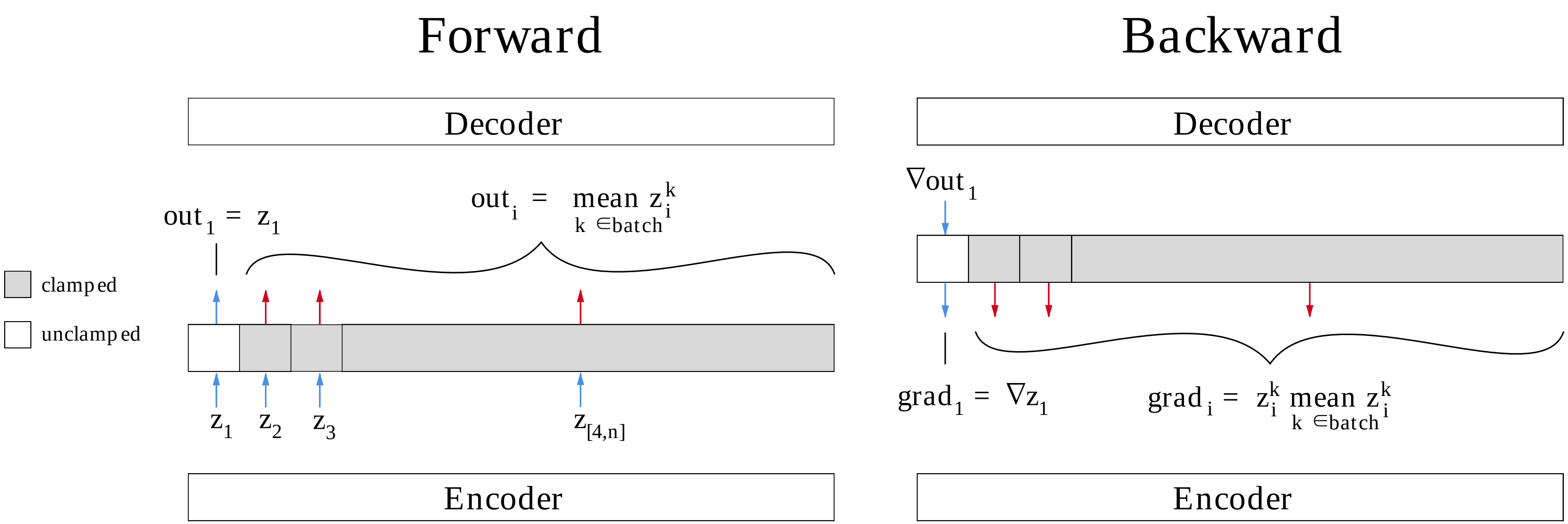
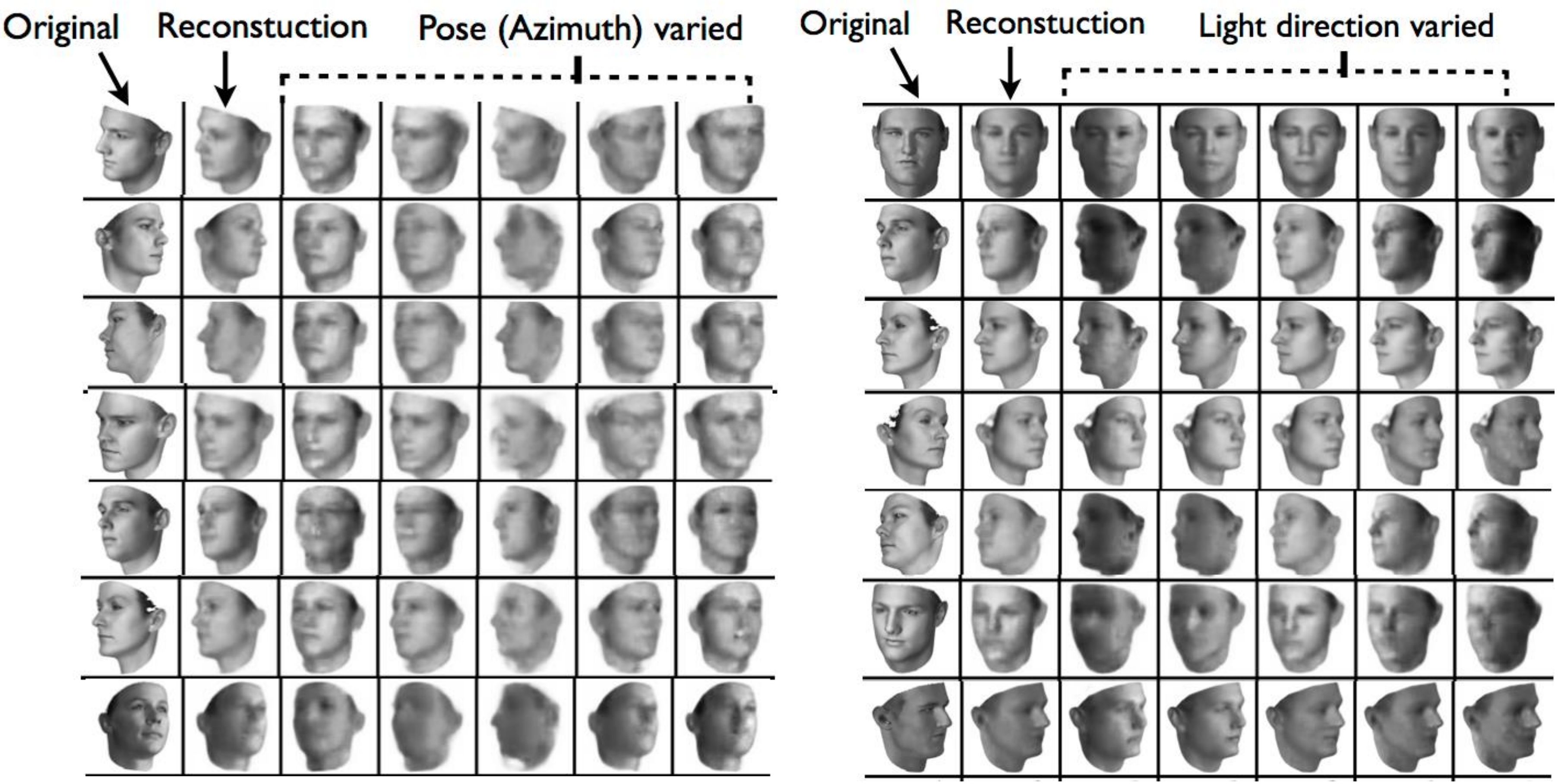


Figure 3: **Training on a minibatch in which only ϕ, the azimuth angle of the face, changes.** During the forward step, the output from each component $z_i \neq z_1$ of the encoder is altered to be the same for each sample in the batch. This reflects the fact that the generating variables of the image (e.g. the identity of the face) which correspond to the desired values of these latents are unchanged throughout the batch. By holding these outputs constant throughout the batch, the single neuron z_1 is forced to explain all the variance within the batch, i.e. the full range of changes to the image caused by changing ϕ. During the backward step z_1 is the only neuron which receives a gradient signal from the attempted reconstruction, and all $z_i \neq z_1$ receive a signal which nudges them to be closer to their respective averages over the batch. During the complete training process, after this batch, another batch is selected at random; it likewise contains variations of only one of $\phi, \alpha, \phi_L, intrinsic$; all neurons which do not correspond to the selected latent are clamped; and the training proceeds.

Vision task I – Inverse graphics network

- Results



Vision task I – Inverse graphics network

- MoFA: Model-based Face Autoencoder (ICCV2017)

Vision task I – Inverse graphics network

- FML: Face model learning from videos (CVPR2019)

Vision task I – Inverse graphics network

- DVP: Deep Video Portraits (SIGGRAPH 2018)

Deep Video Portraits

Hyeongwoo Kim[1] Pablo Garrido[2] Ayush Tewari[1] Weipeng Xu[1] Justus Thies[3]
Matthias Nießner[3] Patrick Pérez[2] Christian Richardt[4] Michael Zollhöfer[5] Christian Theobalt[1]

[1] *MPI Informatics* [2] *Technicolor* [3] *Technical University of Munich* [4] *University of Bath* [5] *Stanford University*

GENERATIONS
SIGGRAPH2018

Vision task I – Inverse graphics network

- Live Speech Portraits (SIGGRAPHAISA 2021)

SIGGRAPH ASIA 2021 TOKYO

LIVE

Live Speech Portraits: Real-Time Photorealistic Talking-Head Animation

Yuanxun Lu[1] Jinxiang Chai[2] Xun Cao[1]

[1] Nanjing University [2] Xmov

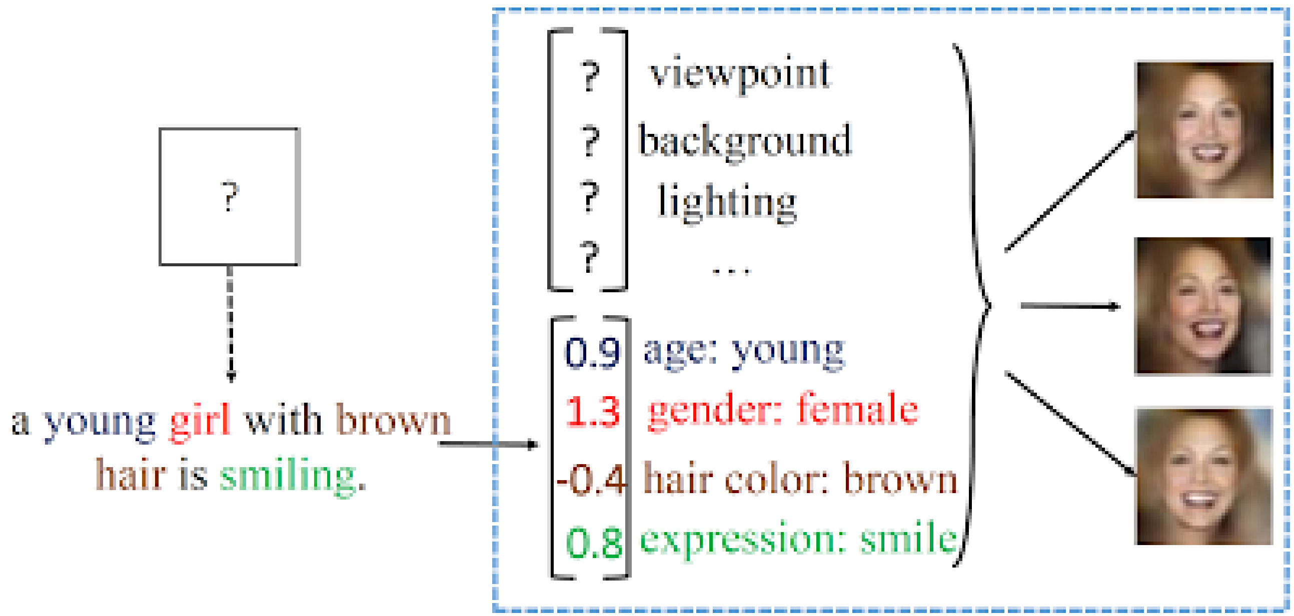
南京大學 NANJING UNIVERSITY

XMOV

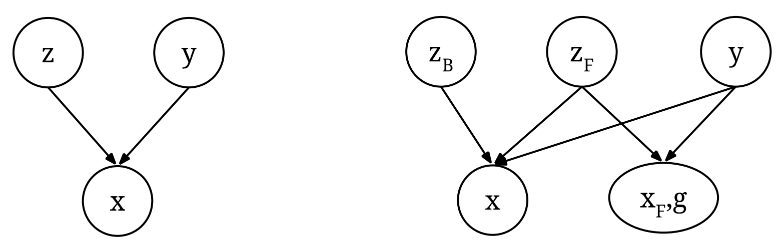
(with Audio)

Vision task II – Attribute2Image

- Main ideas



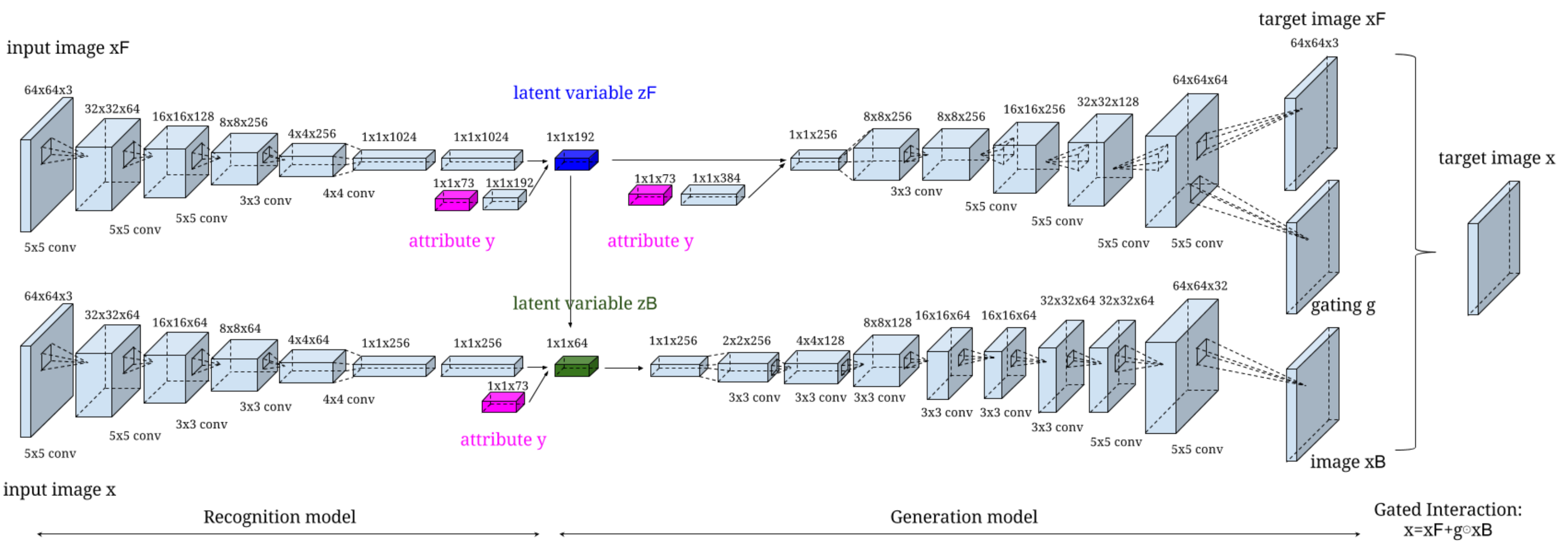
Attribute-conditioned Image Generation



(a) CVAE: $p_\theta(x|y,z)$ (b) disCVAE: $p_\theta(x, x_F, g|y, z_F, z_B)$

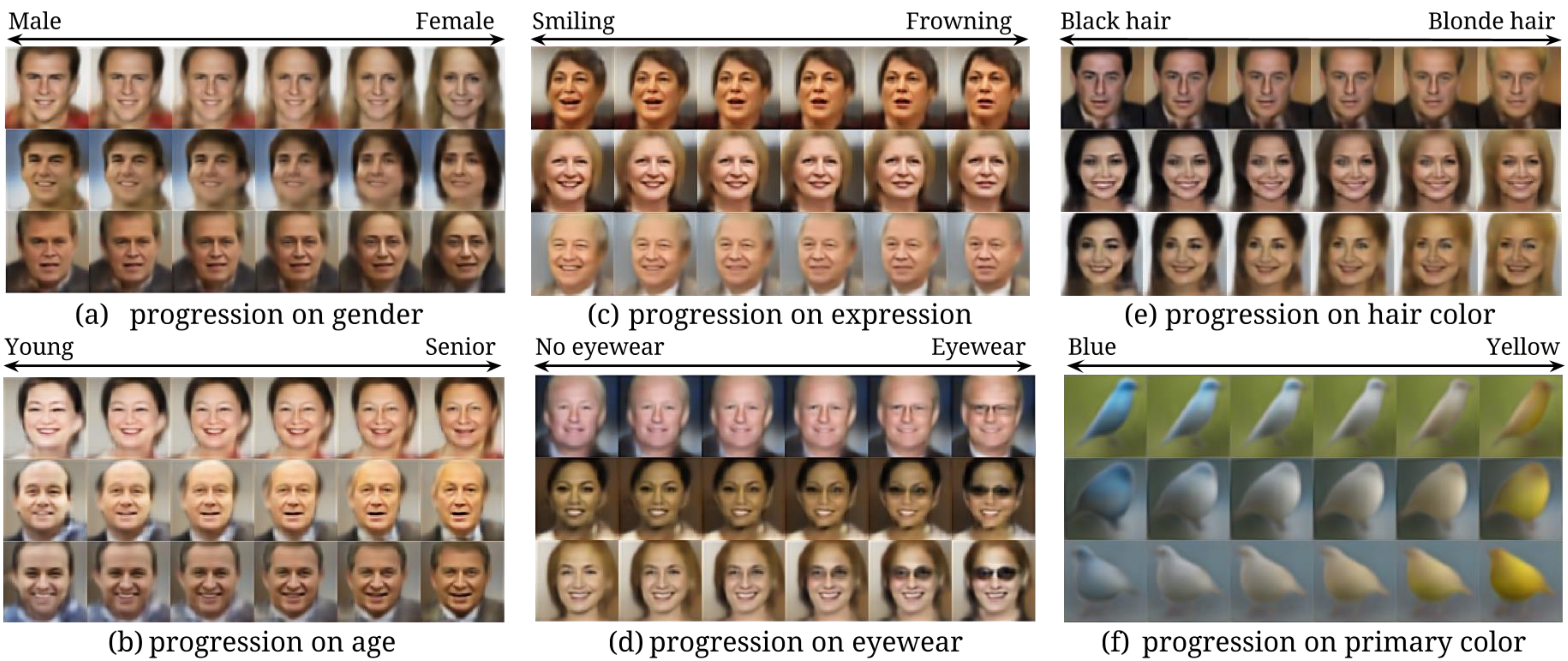
Vision task II – Attribute2Image

- Network structure



Vision task II – Attribute2Image

- Results



Problems of VAE

- Model capacity

- Note that the VAE requires 2 tractable distributions to be used:
  - The prior distribution $p(z)$ must be easy to sample from
  - The conditional likelihood $p(x|z, \theta)$ must be computable
- In practice this means that the 2 distributions of interest are often simple, for example uniform, Gaussian, or even isotropic Gaussian

Problems of VAE

- Blurry images

https://blog.openai.com/generative-models/

- The samples from the VAE look blurry
- Three plausible explanations for this
  - Maximizing the likelihood
  - Restrictions on the family of distributions
  - The lower bound approximation

Problems of VAE

- Blurry images

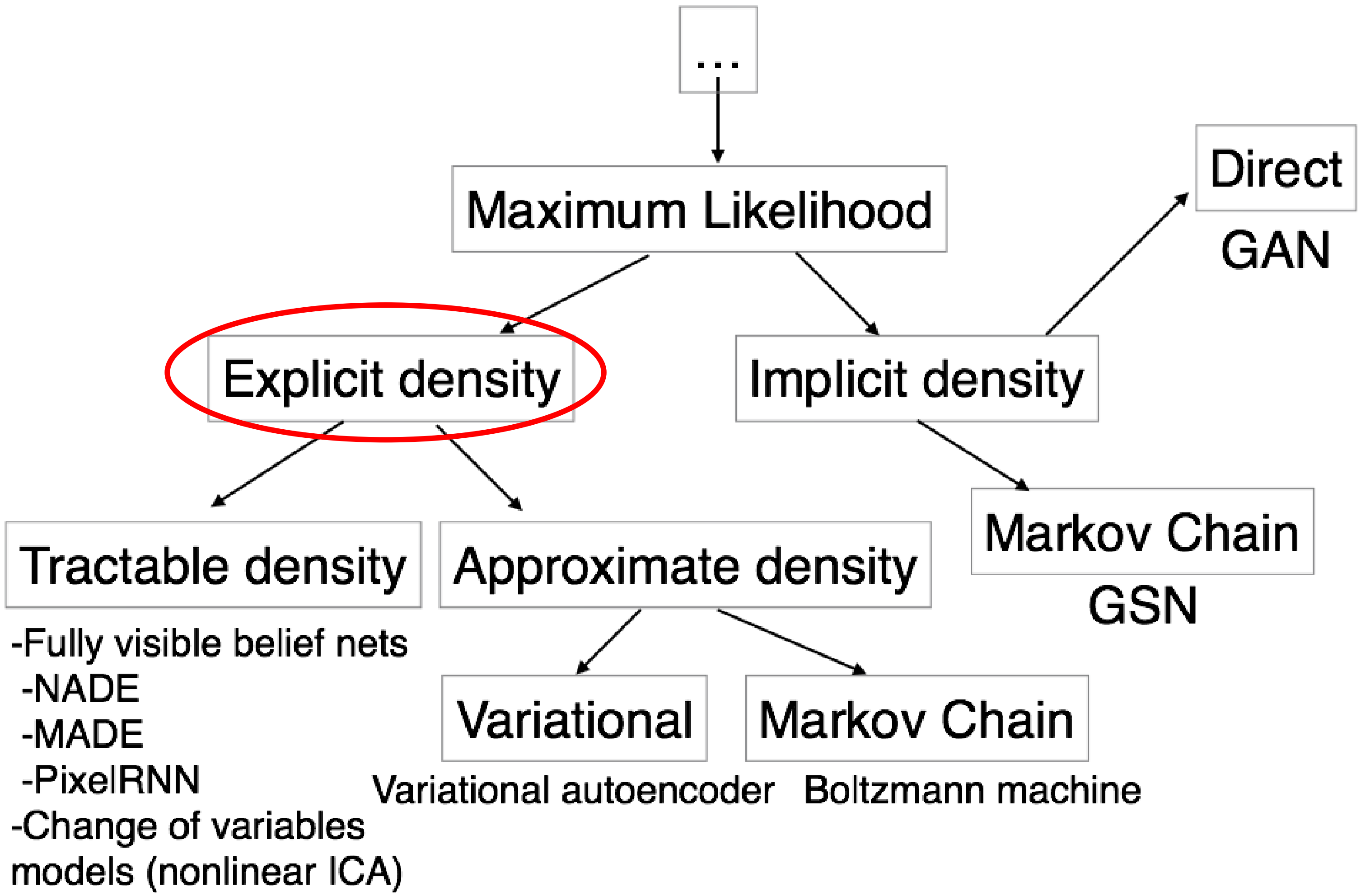
  - Recent investigations suggest that both the simple probability distributions and the variational approximation lead to blurry images
  - Kingma & colleages: Improving Variational Inference with Inverse Autoregressive Flow
  - Zhao & colleagues: Towards a Deeper Understanding of Variational Autoencoding Models
  - Nowozin & colleagues: f-gan: Training generative neural samplers using variational divergence minimization

Outline

- Vision applications of VAEs
  - Inverse graphics network
  - Attribute2Image
- Generative Adversarial Networks
  - Implicit generative models
  - Adversarial learning

Acknowledgement: Feifei Li et al's cs231n notes

Taxonomy of Generative Models



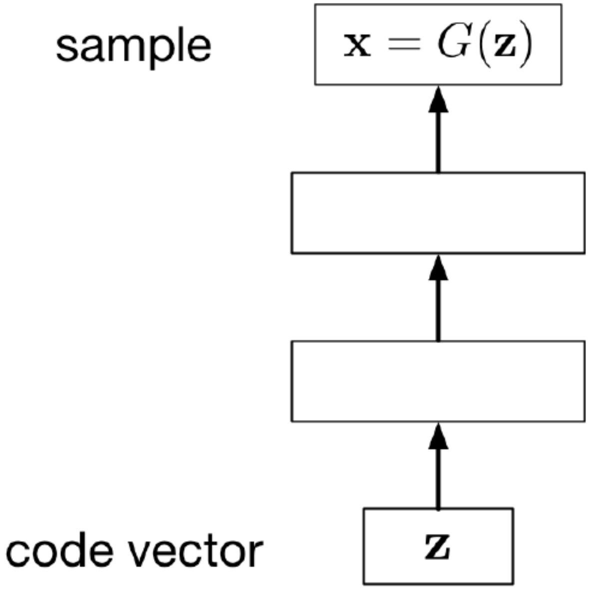
Implicit Generative Models

- Working with explicit model p(x) could be expensive
  - Variational Autoencoder (variational inference)
  - Boltzmann Machines (MCMC)

- Representation learning may not require p(x)
  - Sometimes we are more interested in taking samples from p(x) instead of p itself

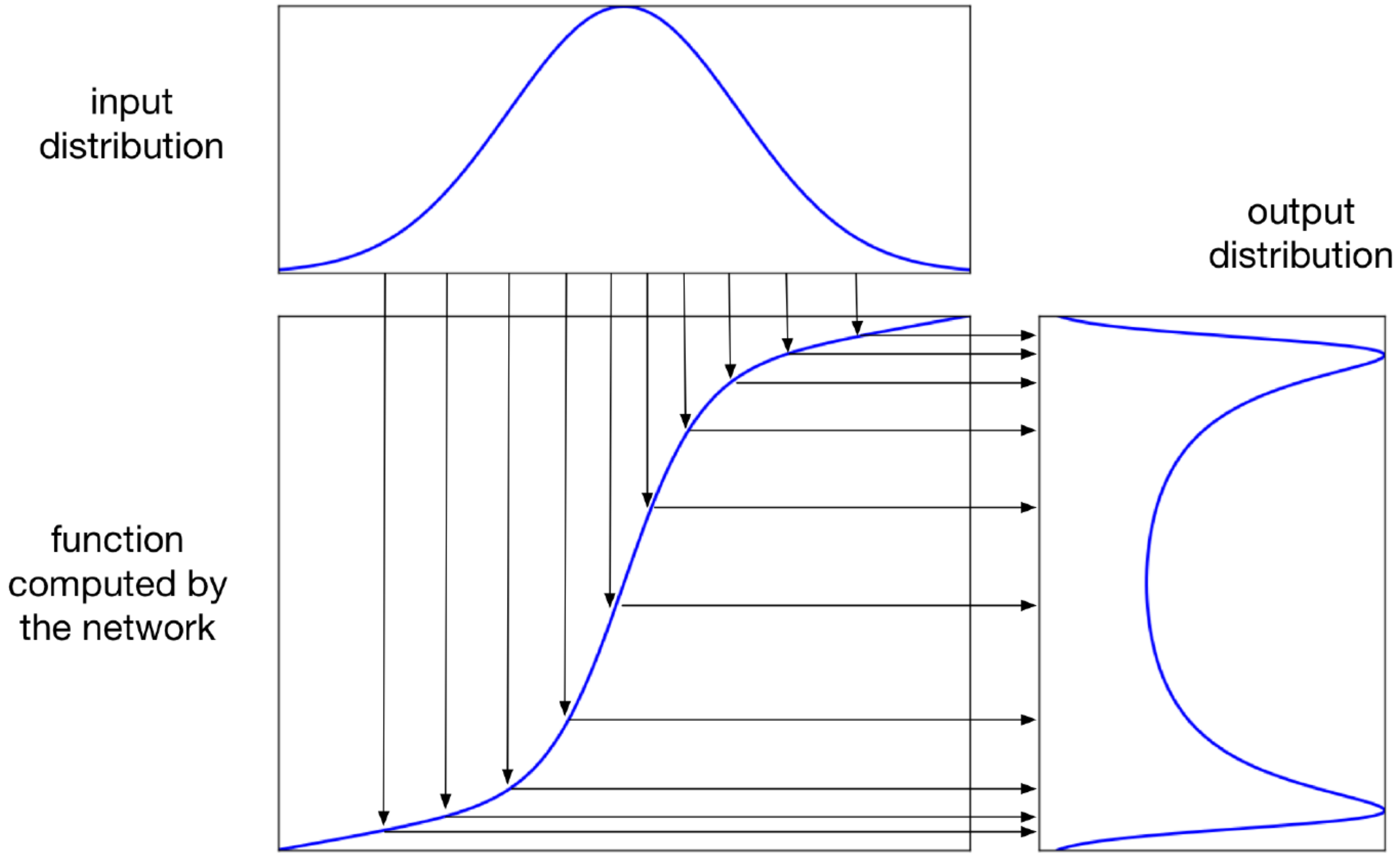
Implicit Generative Models

- Implicitly define a probability distribution
- Start by sampling the code vector z from a fixed, simple distribution
- A generator network computes a differentiable function G mapping z to an x in data space



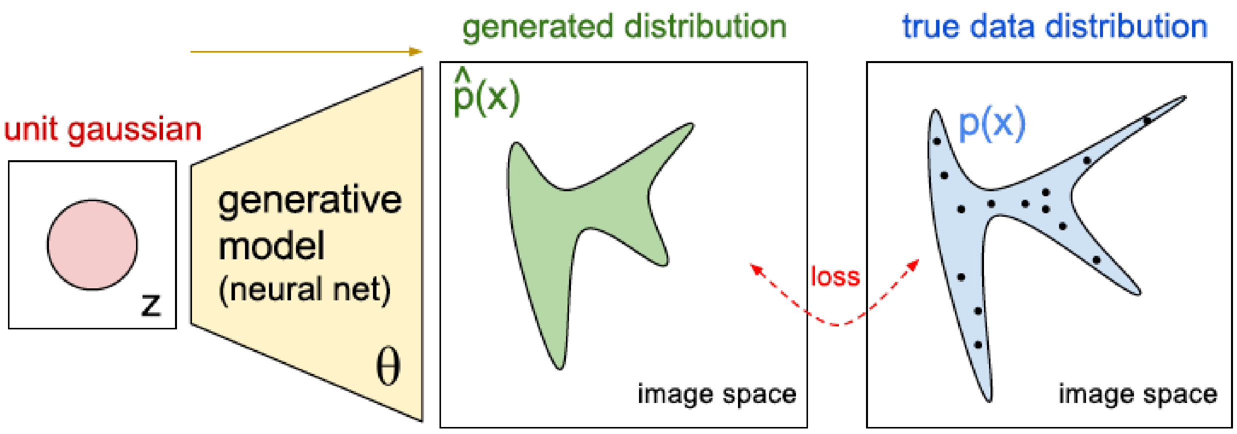
Implicit Generative Models

- Intuition: 1D example



Implicit Generative Models

- Intuition

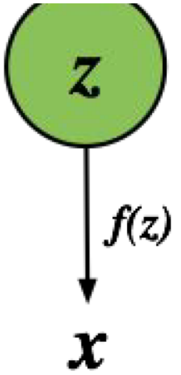


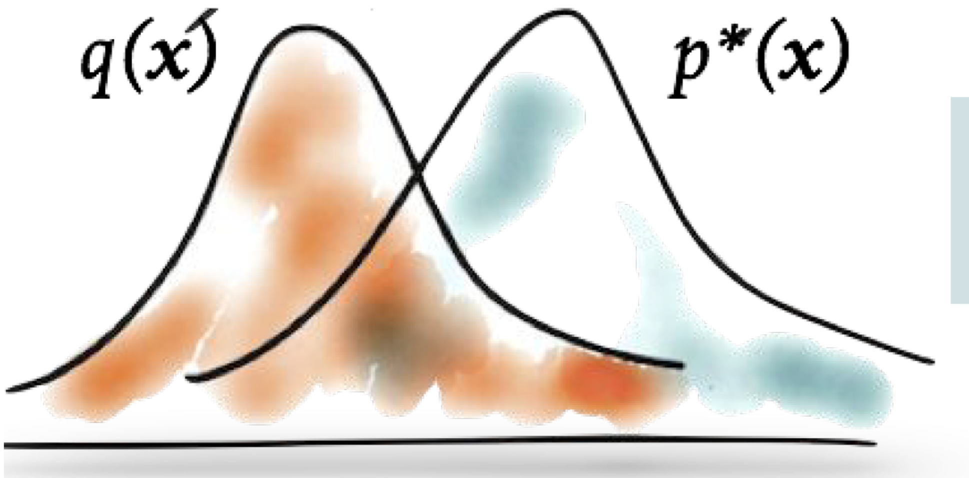
advocate/penalize samples within the blue/white region.

Learning by comparison

- Basic idea

For some models, we only have access to an unnormalised probability, partial knowledge of the distribution, or a simulator of data.

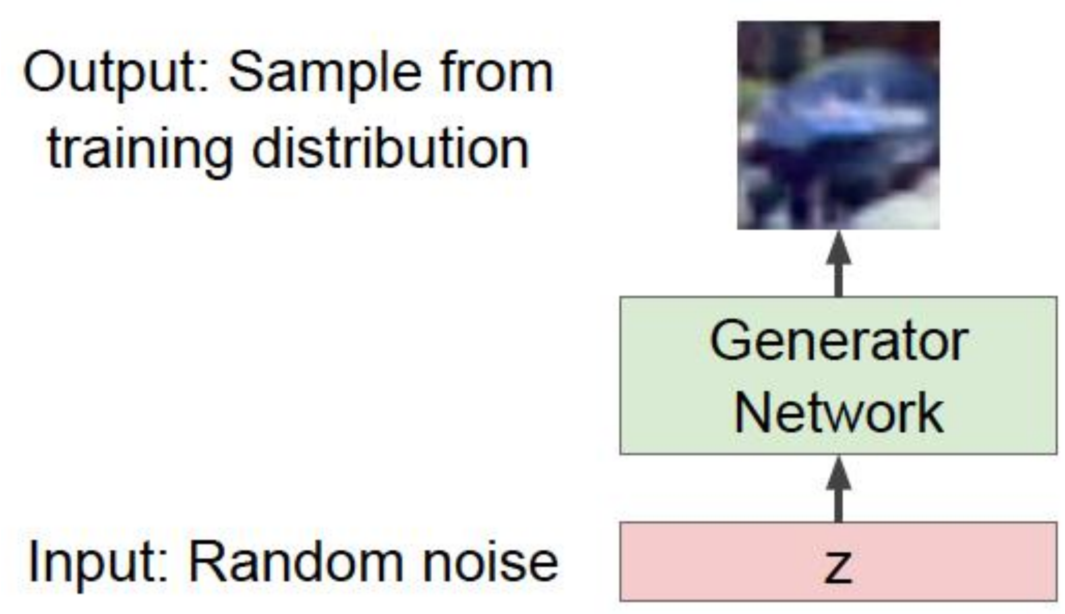




We compare the estimated distribution q(x) to the true distribution $p^*(x)$ using samples.

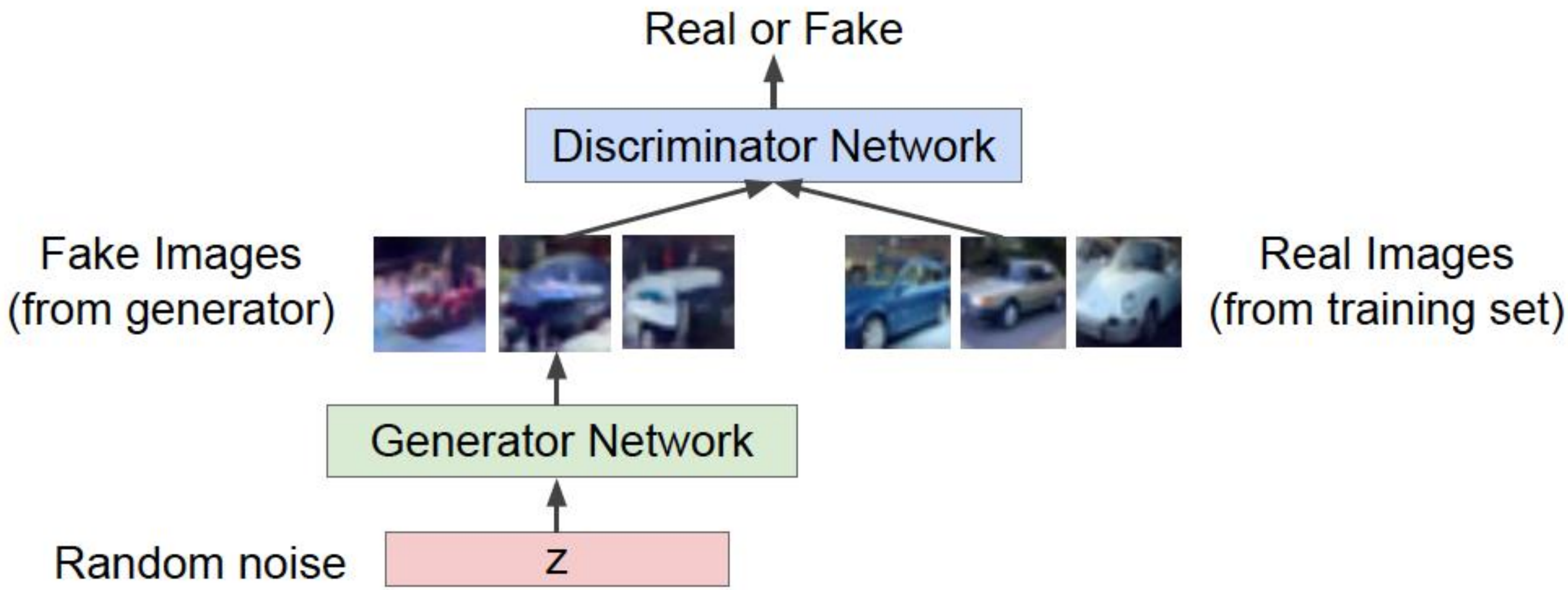
Generative Adversarial Networks

- Using a neural network to generate data



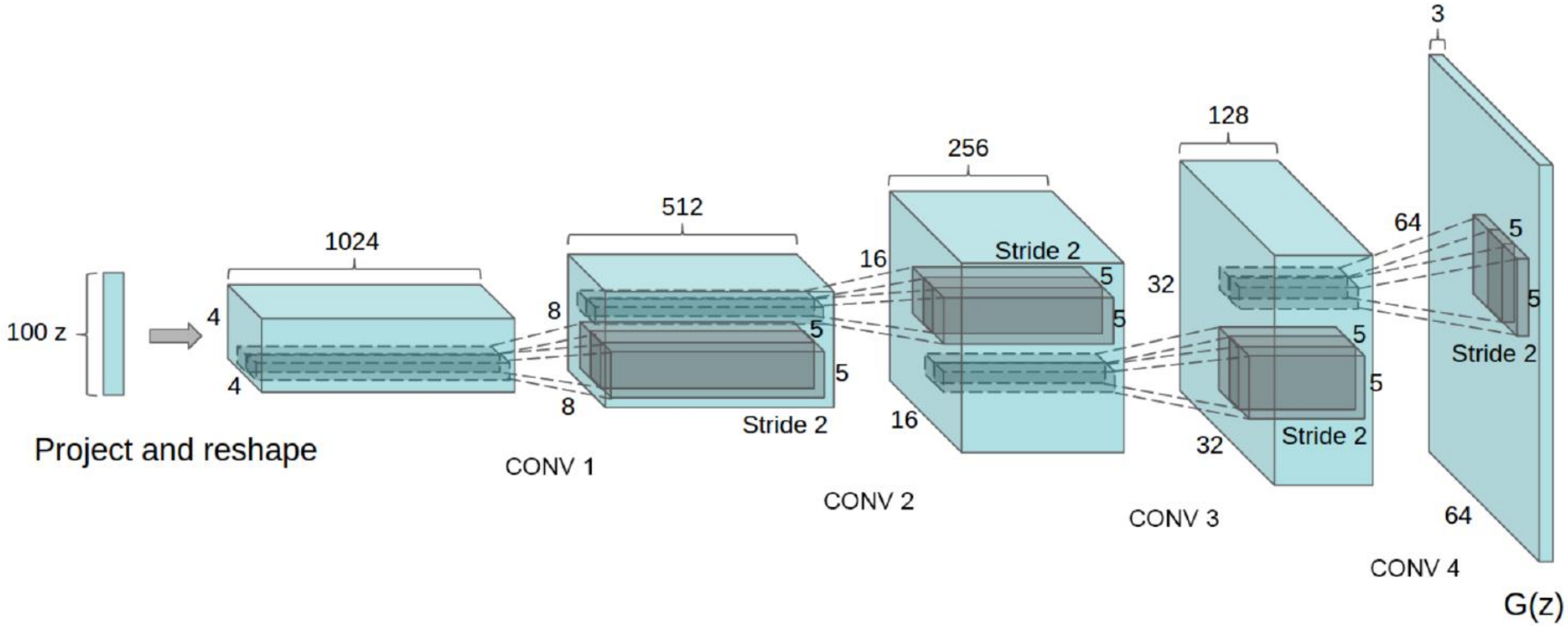
Generative Adversarial Networks

- Using another neural network to determine if the data is real or not



Typical generator architecture

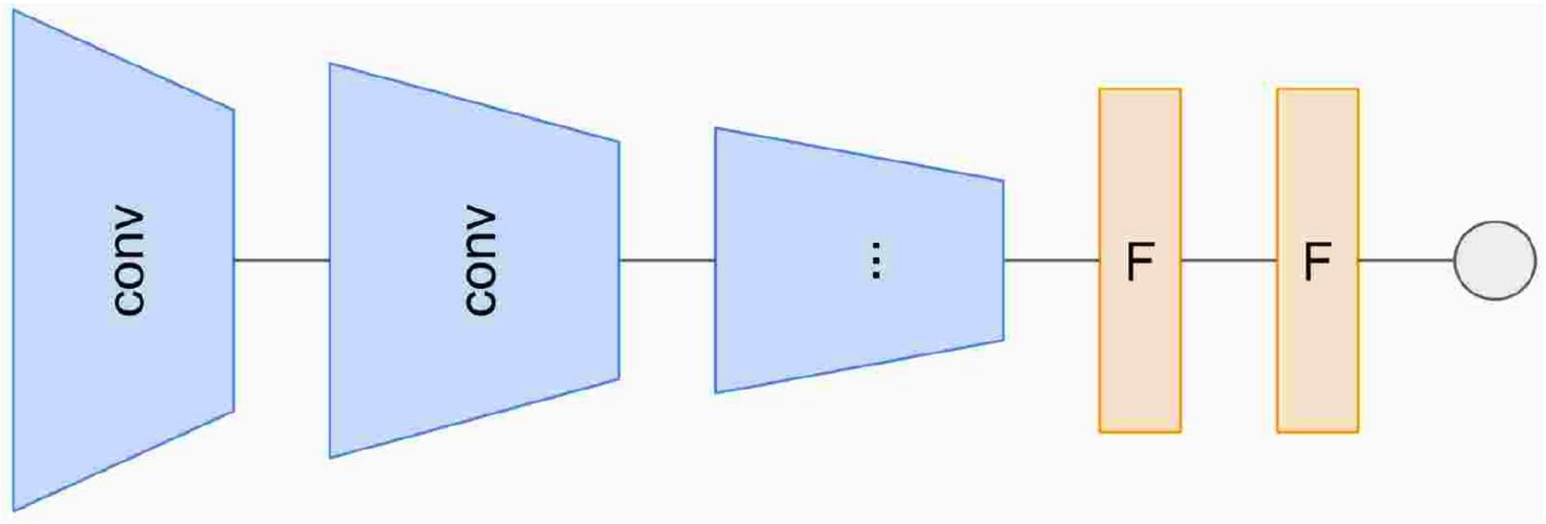
- For images



- Unit Gaussian distribution on z, typically 10-100 dim.
- Up-convolutional deep network (reverse recognition CNN)

Typical discriminator architecture

- For images



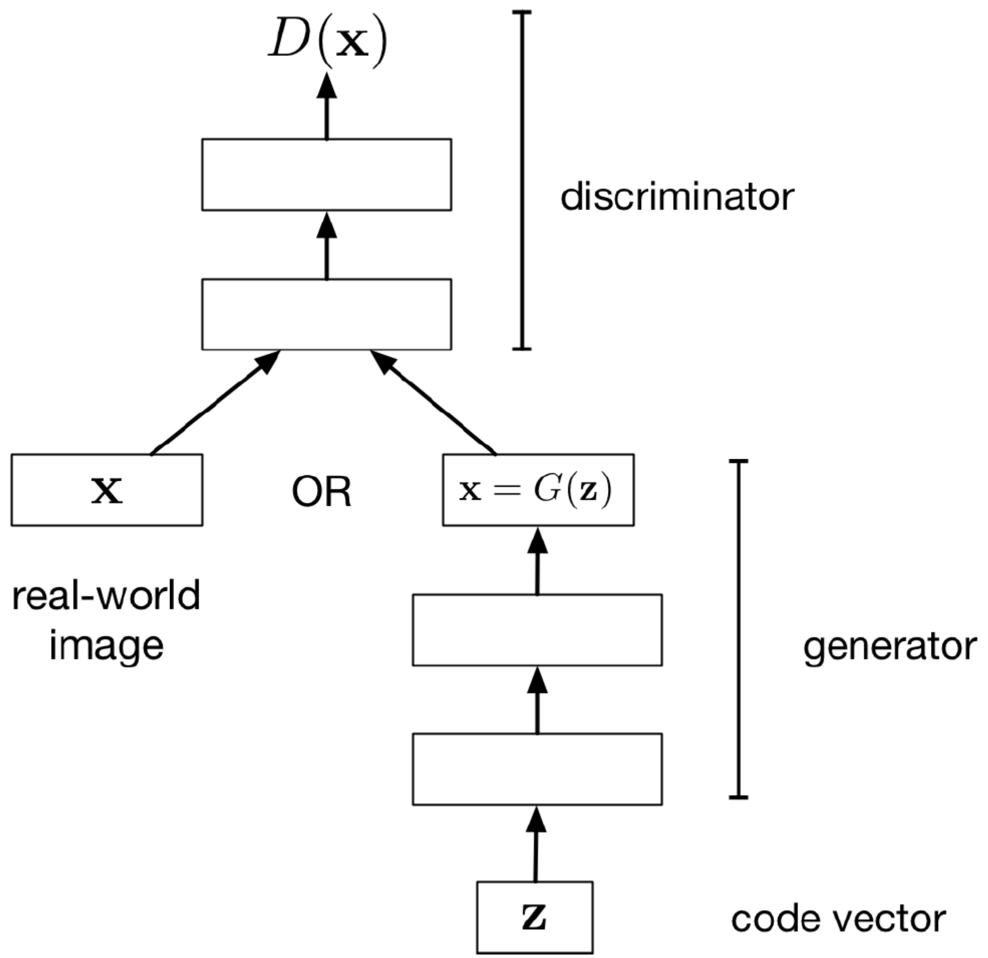
  - Recognition CNN model
  - Binary classification output: real / synthetic

Adversarial learning

- GAN objective for the generator is some complicated objective function defined by a neural network.
  - This means a new way of thinking about "distance".
  - We are training networks to minimize the "distance" or "divergence" between generated images and real images.
  - Instead of some hand-crafted distance metric like L1 or L2, we can make something completely new.
  - A neural network, with the right architecture, is arguably the definition of perceptual similarity (assuming our visual system is some sort of neural network).

Adversarial Learning

- Adversarial loss



Adversarial Learning

- Let D denote the discriminator's predicted probability of being real data
- Discriminator's cost function: cross-entropy loss for task of classifying real vs. fake images

$$\mathcal{J}_D = \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[-\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{z}}[-\log(1 - D(G(\mathbf{z})))]$$

- One possible cost function for the generator: the opposite of the discriminator's

$$\begin{aligned}\mathcal{J}_G &= -\mathcal{J}_D \\ &= \mathrm{const} + \mathbb{E}_{\mathbf{z}}[\log(1 - D(G(\mathbf{z})))]\end{aligned}$$

Two-player game

- Minimax formulation
  - The generator and discriminator are playing a zero-sum game against each other

$$\min_G \max_D \mathcal{J}_D$$

  - Using parametric models

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

Minimax objective function:

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

Summary

- Variational Autoencoders (VAEs)
  - VAE objective
  - VAE: Vision applications
- Generative Adversarial Networks (GAN)
  - Adversarial learning
- Next time:
  - GAN

- Reading material
  - https://arxiv.org/pdf/1503.03167.pdf (Deep Convolutional Inverse Graphics Network)
  - https://arxiv.org/pdf/2011.10063.pdf (Dual Contradistinctive Generative Autoencoder)
  - https://taesung.me/SwappingAutoencoder/ (Swapping Autoencoder for Deep Image Manipulation)

Acknowledgement: Feifei Li et al's cs231n notes