Generative Models & Embedded Ethics

VAEs, diffusion models, bias in deep learning

CSC420
David Lindell
University of Toronto
cs.toronto.edu/~lindell/teaching/420
Slide credit: **FCV, Bill Freeman, Phillip Isola, Tianhong Li**

TCIG
Toronto Computational Imaging Group

Course Overview

Overview

- Introduction
  - What can Generative Models do?
  - What are Generative Models?
  - Generative models and representation learning

- Generative Models: Concepts and Recent Frontiers
  - Variational Autoencoders (VAE)
  - Diffusion models

- Embedded Ethics

Overview

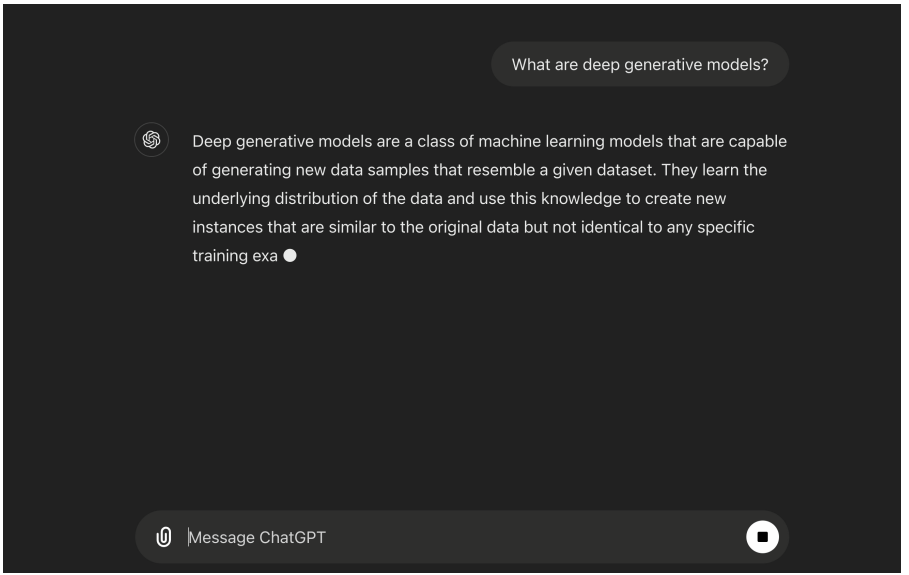
- Introduction
  - What can Generative Models do?
  - What are Generative Models?
  - Generative models and representation learning

- Generative Models: Concepts and Recent Frontiers
  - Variational Autoencoders (VAE)
  - Diffusion models

- Embedded Ethics

The “GenAI” Era

Chatbot and natural language conversation,



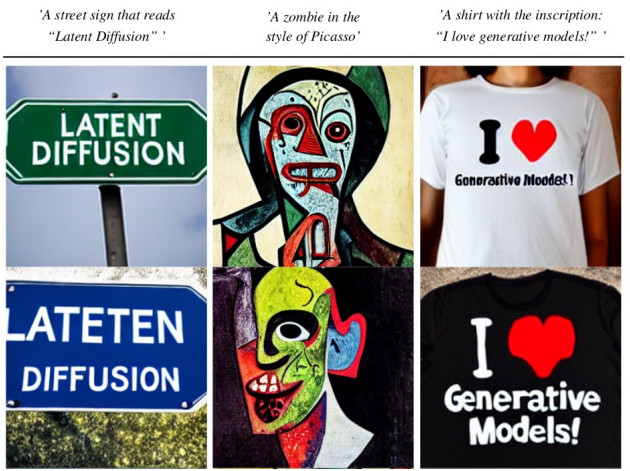
The “GenAI” Era

Class-to-image generation, 2019

Generated by BigGAN

The “GenAI” Era

Text-to-image generation, 2022



Generated by Latent Diffusion

The “GenAI” Era

Text-to-image generation, 2024

Generated by Stable Diffusion 3 Medium.
Prompt: teddy bear teaching a course, with "generative models" written on blackboard

The “GenAI” Era

Text-to-image generation, 2025

Getting started

Generated by GPT 4o

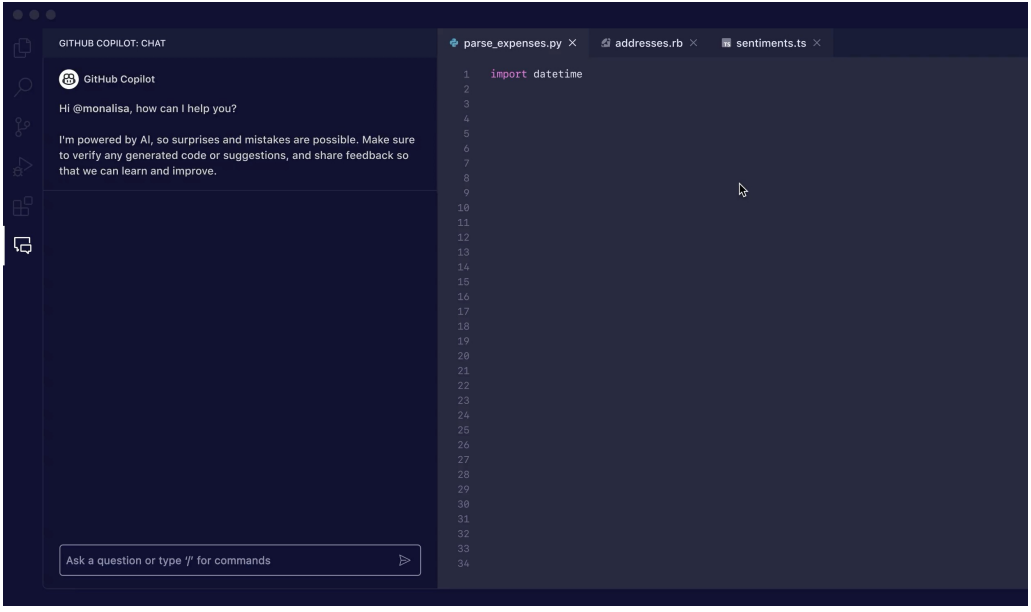
The “GenAI” Era

Text-to-video generation, 2024

Generated by Sora

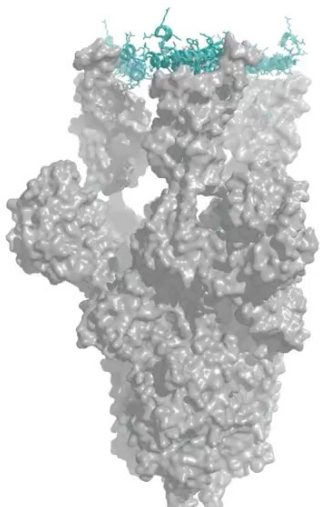
The “GenAI” Era

AI assistant for code generation



The “GenAI” Era

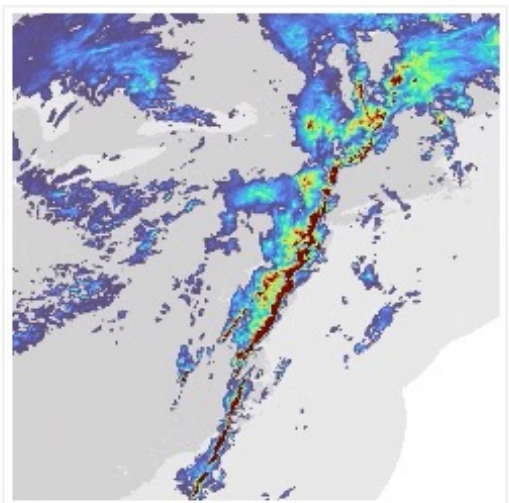
Protein design and generation



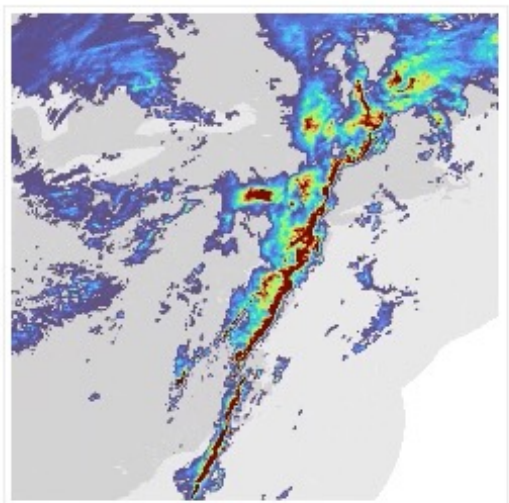
Watson, et al. De novo design of protein structure and function with RFdiffusion, Nature 2023

The "GenAI" Era

Weather forecasting



Target

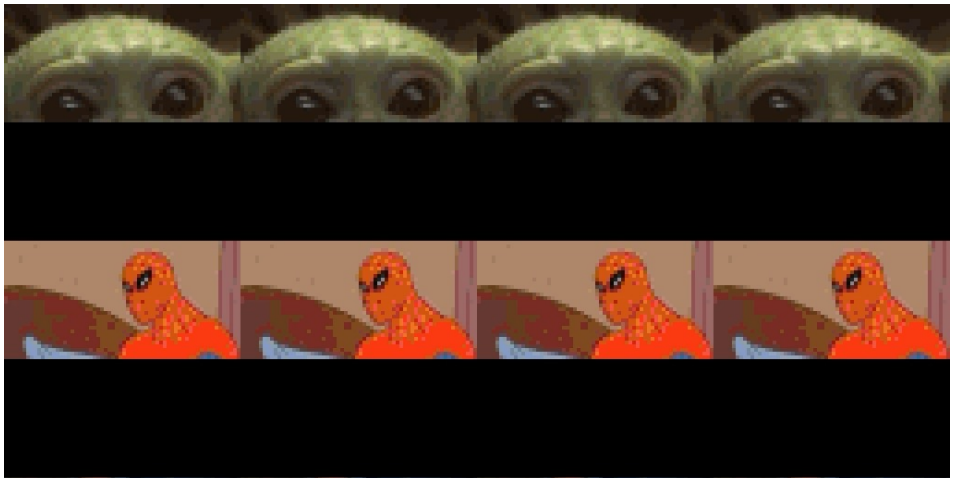


DGMR

Skilful precipitation nowcasting using deep generative models of radar, Nature 2021

Generative Models before the “GenAI” Era

2020, image GPT: Fill bottom part of the image



Chen et al. “Generative Pretraining from Pixels”, ICML 2020

Generative Models before the “GenAI” Era

2016, Context Encoder: Fill center part of the image

Pathak, et al., “Context Encoders: Feature Learning by Inpainting”, CVPR 2016

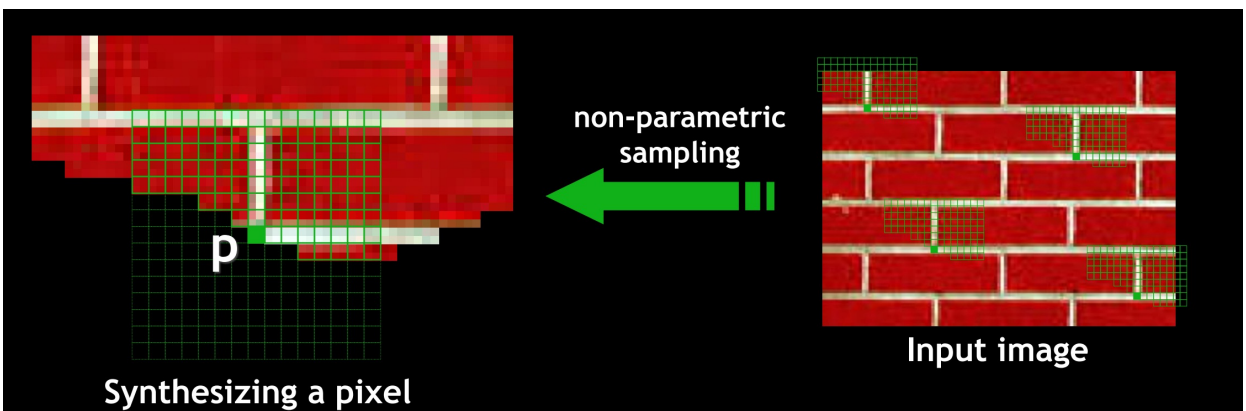
Generative Models before the “GenAI” Era

2009, PatchMatch: Photoshop’s Content-aware Fill

PatchMatch: A Randomized Correspondence Algorithm for Structural Image Editing, SIGGRAPH 2009

Generative Models before the "GenAI" Era

1999, the Efros-Leung algorithm for texture synthesis

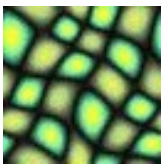


Texture Synthesis by Non-parametric Sampling, ICCV 1999

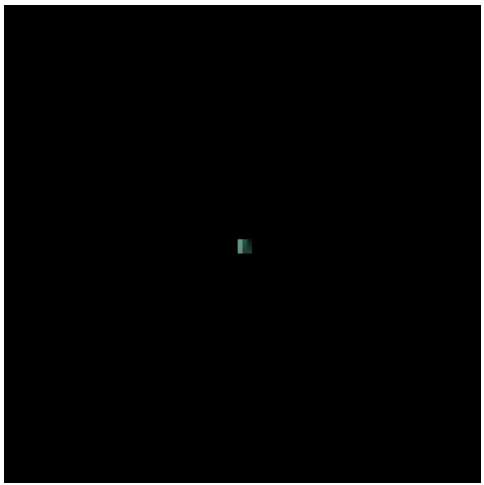
Generative Models before the "GenAI" Era

1999, the Efros-Leung algorithm for texture synthesis

In today's word: this is an **Autoregressive** model



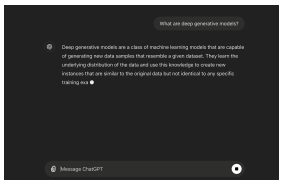
exemplar



Texture Synthesis by Non-parametric Sampling, ICCV 1999

What are Generative Models?

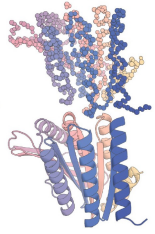
What do these scenarios have in common?



Chatbot

Image generation

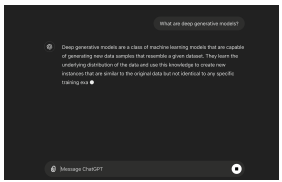
Video generation



Protein generation

What do these scenarios have in common?

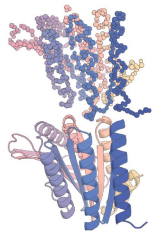
- There are **multiple** or infinite predictions to one input.



Chatbot

Image generation

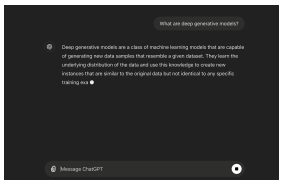
Video generation



Protein generation

What do these scenarios have in common?

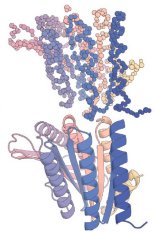
- There are **multiple** or infinite predictions to one input.
- Some predictions are more "**plausible**" than some others.



Chatbot

Image generation

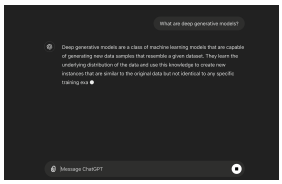
Video generation



Protein generation

What do these scenarios have in common?

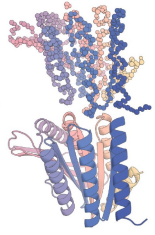
- There are **multiple** or infinite predictions to one input.
- Some predictions are more "**plausible**" than some others.
- Training data may contain **no exact solution**.



Chatbot

Image generation

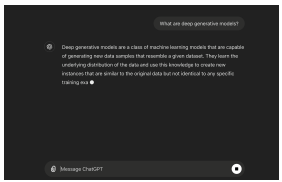
Video generation



Protein generation

What do these scenarios have in common?

- There are **multiple** or infinite predictions to one input.
- Some predictions are more "**plausible**" than some others.
- Training data may contain **no exact solution**.
- Predictions may be **more complex**, more informative, and higher-dimensional than input.



Chatbot

Image generation

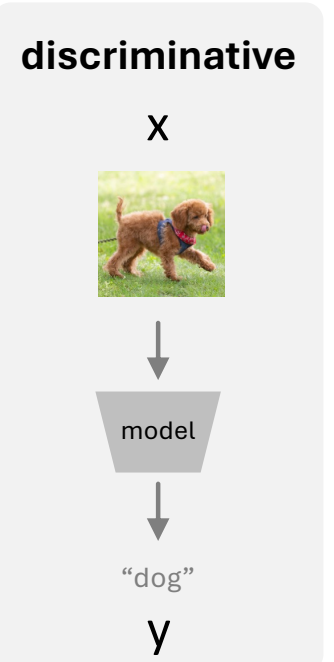
Video generation

Protein generation

Discriminative vs. Generative models

discriminative

- "sample" x ⇨ "label" y
- one desired output



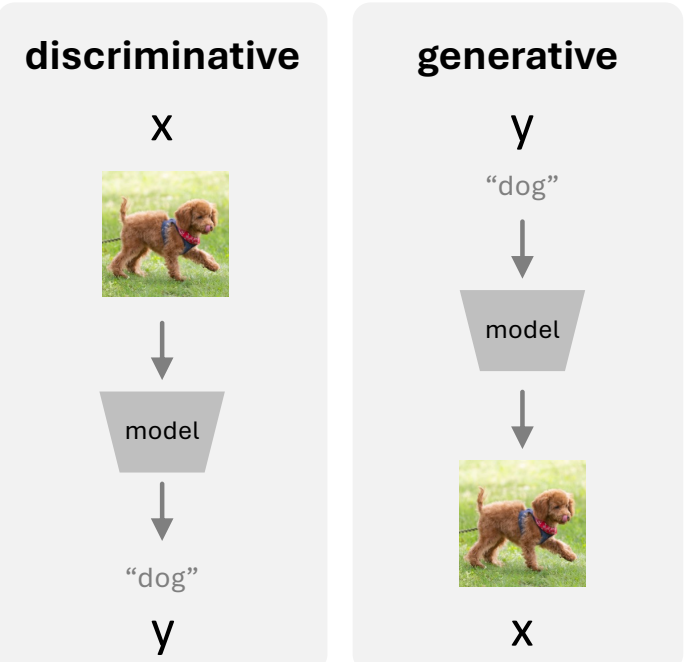
Discriminative vs. Generative models

discriminative

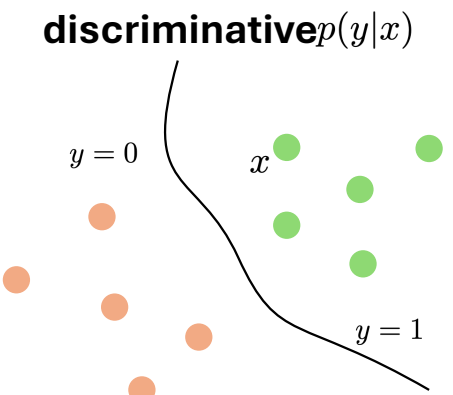
- "sample" x ⇨ "label" y
- one desired output

generative

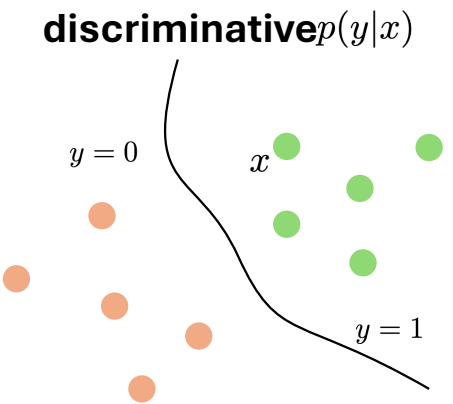
- "label" y ⇨ "sample" x
- many possible outputs

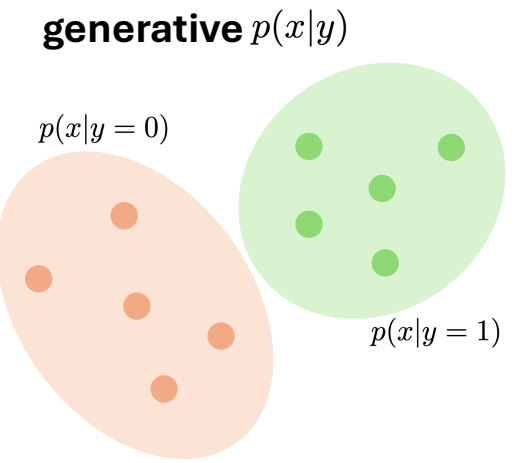


Discriminative vs. Generative models



Discriminative vs. Generative models





Formulating Real-world Problems as Generative Models

- Generative models are about $p(x|y)$
- Many problems can be formulated as generative models
- The research of generative models is all about accurately modeling $p(x|y)$

What do we mean by modeling $p(x|y)$

In a **discriminative** model, we care about modeling

$p(y|x)$

What do we mean by modeling $p(x|y)$

In a **discriminative** model, we care about modeling

p(y|x)

this is called the **posterior** — i.e., a probability distribution that represents what we believe about some unknown quantity **after** observing the data (x)

What do we mean by modeling $p(x|y)$

In a **discriminative** model, we care about modeling

$$p(y|x)$$

this is called the **posterior** — i.e., a probability distribution that represents what we believe about some unknown quantity **after** observing the data (x)

typically, we care about finding the ***most likely*** value for y given the data x. This is called maximum a posteriori estimation

$$\arg\max_y p(y|x)$$

What do we mean by modeling $p(x|y)$

In a **generative** model, we care about modeling

p(x|y)

What do we mean by modeling $p(x|y)$

In a **generative** model, we care about modeling

p(x|y)

this is called the **likelihood** — i.e., the probability that we observe some data (x) given a different set of parameters/labels (y).

What do we mean by modeling $p(x|y)$

In a **generative** model, we care about modeling

p(x|y)

this is called the **likelihood** — i.e., the probability that we observe some data (x) given a different set of parameters/labels (y).

for a generative model, we typical want to synthesize or sample the data x following p(x|y).

What do we mean by modeling $p(x|y)$

In a **generative** model, we care about modeling

p(x|y)

this is called the **likelihood** — i.e., the probability that we observe some data (x) given a different set of parameters/labels (y).

- models that can estimate p(x|y) are called "explicit generative models"
- models that can only sample from p(x|y) are called "implicit generative models"

Bayes' Rule

We can model the posterior given the likelihood as

$$\underbrace{p(y \mid x)}_{?} = \frac{p(x \mid y)\, p(y)}{p(x)}$$

Bayes' Rule

We can model the posterior given the likelihood as

$$p(y \mid x) = \frac{p(x \mid y)\, p(y)}{p(x)}$$

posterior

Bayes' Rule

We can model the posterior given the likelihood as

?

$$p(y \mid x) = \frac{p(x \mid y)p(y)}{p(x)}$$

posterior

Bayes' Rule

We can model the posterior given the likelihood as

likelihood

$$p(y \mid x) = \frac{p(x \mid y)p(y)}{p(x)}$$

posterior

Bayes' Rule

We can model the posterior given the likelihood as

likelihood

$$p(y \mid x) = \frac{p(x \mid y)\,p(y)}{p(x)}$$?

posterior

Bayes' Rule

We can model the posterior given the likelihood as

likelihood

$$p(y \mid x) = \frac{p(x \mid y)\,p(y)}{p(x)}$$

prior

posterior

a prior could encode, e.g., how common each class label is

Bayes' Rule

We can model the posterior given the likelihood as

likelihood

$$p(y \mid x) = \frac{p(x \mid y)\, p(y)}{p(x)}$$

prior

posterior

?

Bayes' Rule

We can model the posterior given the likelihood as

$$p(y \mid x) = \frac{p(x \mid y)\, p(y)}{p(x)}$$

likelihood

prior

posterior

marginal likelihood/data distribution

$$p(y \mid x) = \frac{p(x \mid y)\, p(y)}{\int p(x \mid y)\, p(y)\, dy}$$

marginal likelihood/
data distribution

Bayes' Rule

We can model the posterior given the likelihood as

$$p(y \mid x) = \frac{p(x \mid y)\, p(y)}{p(x)}$$

likelihood

prior

posterior

marginal likelihood/data distribution

can explicit generative models be made discriminative?

Why is it hard?

- Explicit generative models can be discriminative: Bayes' rule

$$p(y|x) = p(x|y)\frac{p(y)}{p(x)}$$

assuming known prior on label y

constant for **given x**

discriminative

generative

Why is it hard?

- Explicit generative models can be discriminative: Bayes' rule

$$p(y|x) = p(x|y)\frac{p(y)}{p(x)}$$

assuming known prior on label y

discriminative

generative

constant for **given x**

- Can discriminative models be generative?

Why is it hard?

- Explicit generative models can be discriminative: Bayes' rule

$$p(y|x) = p(x|y)\frac{p(y)}{p(x)}$$

discriminative — generative

assuming known prior on label y

constant for **given x**

- Can discriminative models be generative?

$$p(x|y) = p(y|x)\frac{p(x)}{p(y)}$$

generative — discriminative

still need to model prior distribution of data x

constant for **given y**

Why is it hard?

- Explicit generative models can be discriminative: Bayes' rule

$$p(y|x) = p(x|y)\frac{p(y)}{p(x)}$$

discriminative — generative

assuming known prior on label y

constant for **given x**

- Can discriminative models be generative?

$$p(x|y) = p(y|x)\frac{p(x)}{p(y)}$$

generative — discriminative

still need to model prior distribution of data x

constant for **given y**

- The challenge is about **modeling data distributions**

Formulating Real-world Problems as Generative Modeling

Formulating Real-world Problems as Generative Models

- Generative models perform (or describe) the synthesis of data
- In other words, generative models are about $p(x|y)$

What can be y?
- condition
- constraint
- labels
- attributes

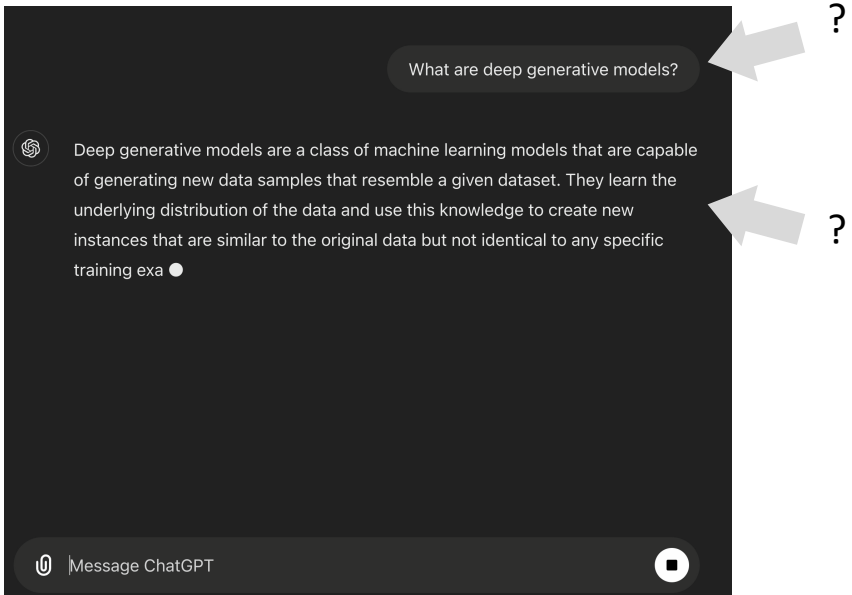
...
- more abstract
- less informative

What can be x?
- "data"
- samples
- observations
- measurements

...
- more concrete
- more informative

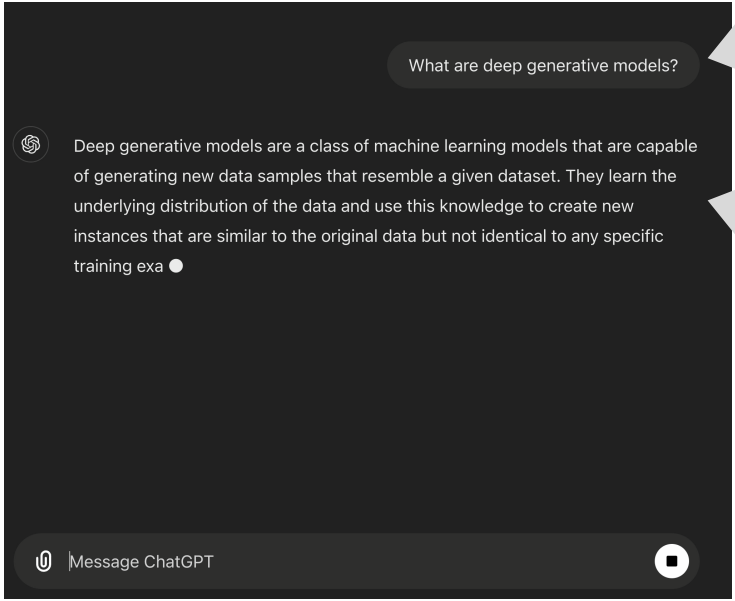
Case study: Formulating problems as p(x|y)

- **Natural language conversation**



Case study: Formulating problems as p(x|y)

- **Natural language conversation**



y: prompt

x: response of the chatbot

Case study: Formulating problems as p(x|y)

- **Text-to-image/video generation**

Prompt: teddy bear teaching a course, with "generative models" written on blackboard

?

?

Image generated by Stable Diffusion 3 Medium

Case study: Formulating problems as p(x|y)

- **Text-to-image/video generation**

Prompt: teddy bear teaching a course, with "generative models" written on blackboard

y: text prompt

x: generated visual content

Image generated by Stable Diffusion 3 Medium

Case study: Formulating problems as p(x|y)

- **Text-to-3D structure generation**

"motorcycle"

"mech suit"

"ghost lantern"

"furry fox head"

?

?

"dresser"

"swivel chair"

"astronaut"

"mushroom house"

Figure credit: Tang, et al. LGM: Large Multi-View Gaussian Model for High-Resolution 3D Content Creation. ECCV 2024

Case study: Formulating problems as p(x|y)

- **Text-to-3D structure generation**

“motorcycle”

“mech suit”

“ghost lantern”

“furry fox head”

x: generated 3D structures

y: text prompt

“dresser”

“swivel chair”

“astronaut”

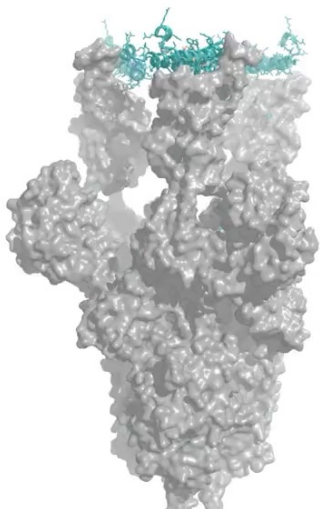
“mushroom house”

Figure credit: Tang, et al. LGM: Large Multi-View Gaussian Model for High-Resolution 3D Content Creation. ECCV 2024

Case study: Formulating problems as p(x|y)

- **Protein structure generation**

y=?



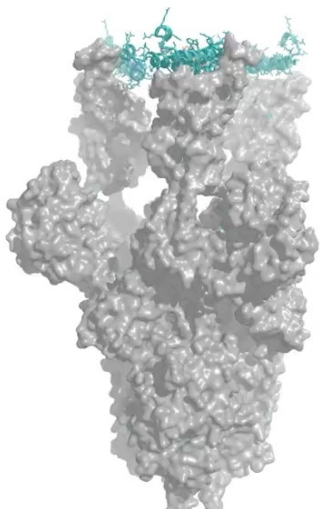
x=?

Watson, et al. De novo design of protein structure and function with RFdiffusion, Nature 2023

Case study: Formulating problems as p(x|y)

- **Protein structure generation**

y:
condition/constraint
(e.g., symmetry)



x: generated
protein
structures

Watson, et al. De novo design of protein structure and function with RFdiffusion, Nature 2023

Case study: Formulating problems as p(x|y)

- **Class-conditional image generation**

“red fox” ← ?

? ←

Image generated by: Li, et al. Autoregressive Image Generation without Vector Quantization, 2024

Case study: Formulating problems as p(x|y)

- **Class-conditional image generation**

"red fox" ← y: class label

x: generated image

Image generated by: Li, et al. Autoregressive Image Generation without Vector Quantization, 2024

Case study: Formulating problems as p(x|y)

- **“Unconditional” image generation**

y: ?

x ?

Images generated by: Karras, et al. Elucidating the Design Space of Diffusion-Based Generative Models, NeurIPS 2022

Case study: Formulating problems as p(x|y)

- **“Unconditional” image generation**

y: an implicit condition
“images following CIFAR10 distribution”

x: generated CIFAR10-like images

- p(x|y): images ~ CIFAR10
- p(x): all images

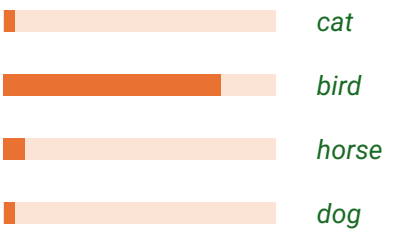
Images generated by: Karras, et al. Elucidating the Design Space of Diffusion-Based Generative Models, NeurIPS 2022

Case study: Formulating problems as p(x|y)

- **Classification** (a generative perspective)

y: ?

x: ?

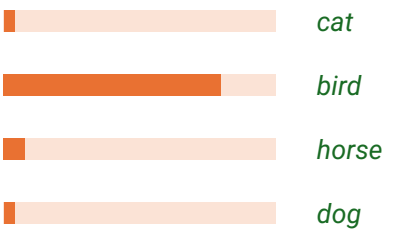


Case study: Formulating problems as p(x|y)

- **Classification** (a generative perspective)

y: an image as the “condition”

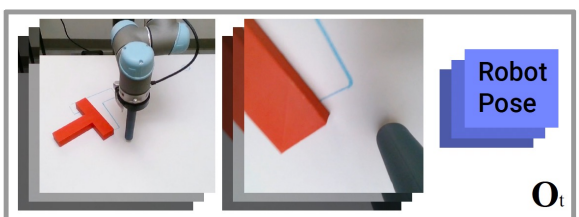
x: probability of classes conditioned on the image



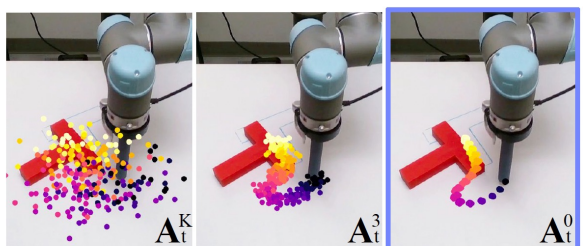
Case study: Formulating problems as p(x|y)

- **Policy Learning in Robotics**

y: ?



x: ?

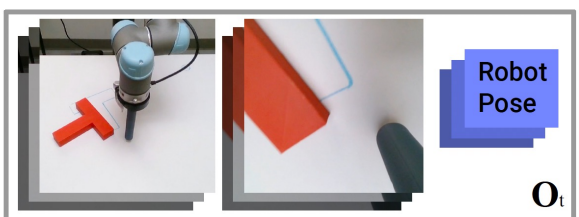


Chi, et al. Diffusion Policy: Visuomotor Policy Learning via Action Diffusion, RSS 2023

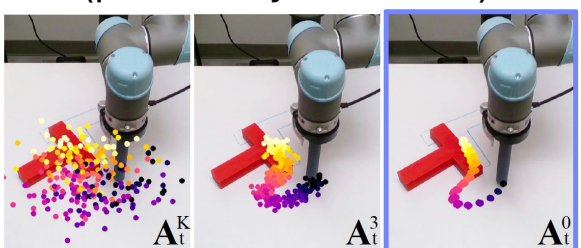
Case study: Formulating problems as p(x|y)

- **Policy Learning in Robotics**

y: visual and other sensory observations



x: policies (probability of actions)

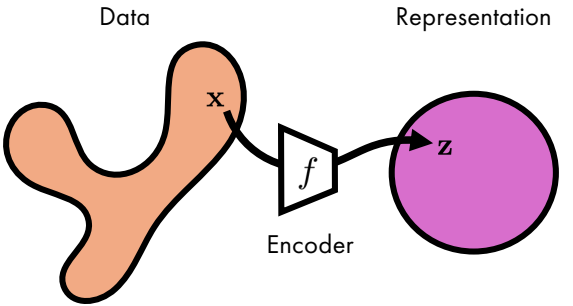


Chi, et al. Diffusion Policy: Visuomotor Policy Learning via Action Diffusion, RSS 2023

Generative Models
and
Representation Learning

Generative Models and Representation Learning

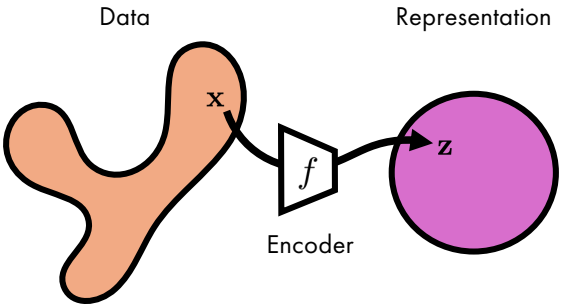
- **Representation learning:**
  - Learning to **represent data instances**
  - map data to feature: $x \rightarrow f(x)$
  - minimize objective: $\min_f Loss(f(x))$

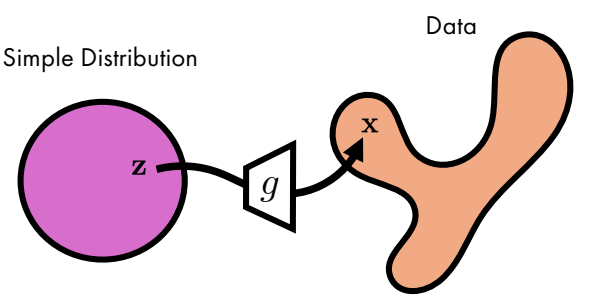


Generative Models and Representation Learning

- **Representation learning:**
  - Learning to **represent data instances**
  - map data to feature: $x \rightarrow f(x)$
  - minimize objective: $\min_f Loss(f(x))$
- **Generative models:**
  - Learning to **represent probability distributions**
  - map a simple distribution to a complex one
  - minimize loss, so that $p(g(z)) \sim p(x)$:

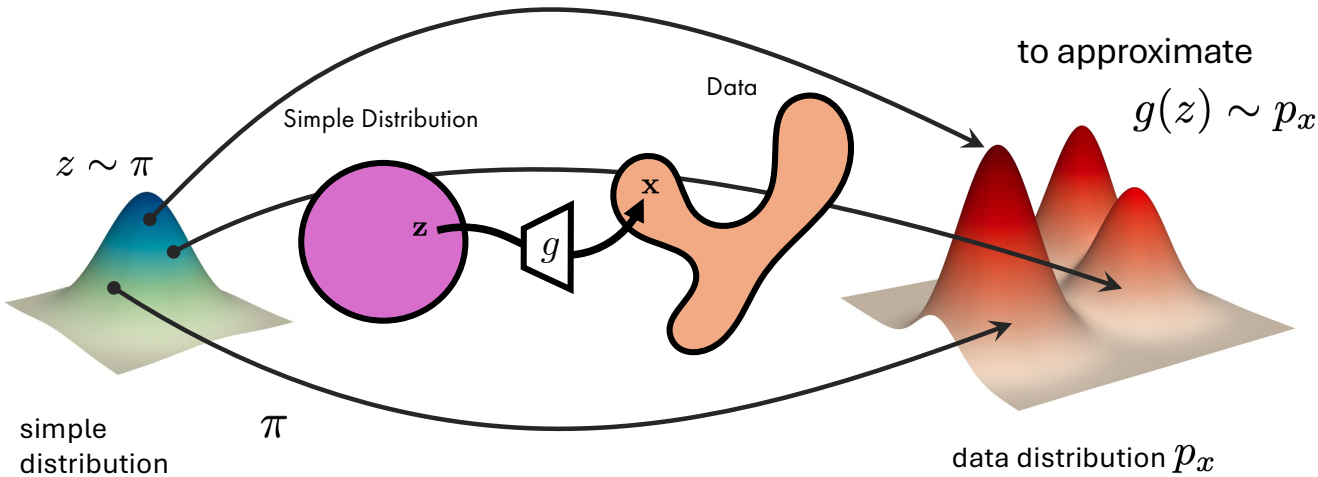
$$\min_g Loss(x, g(z))$$





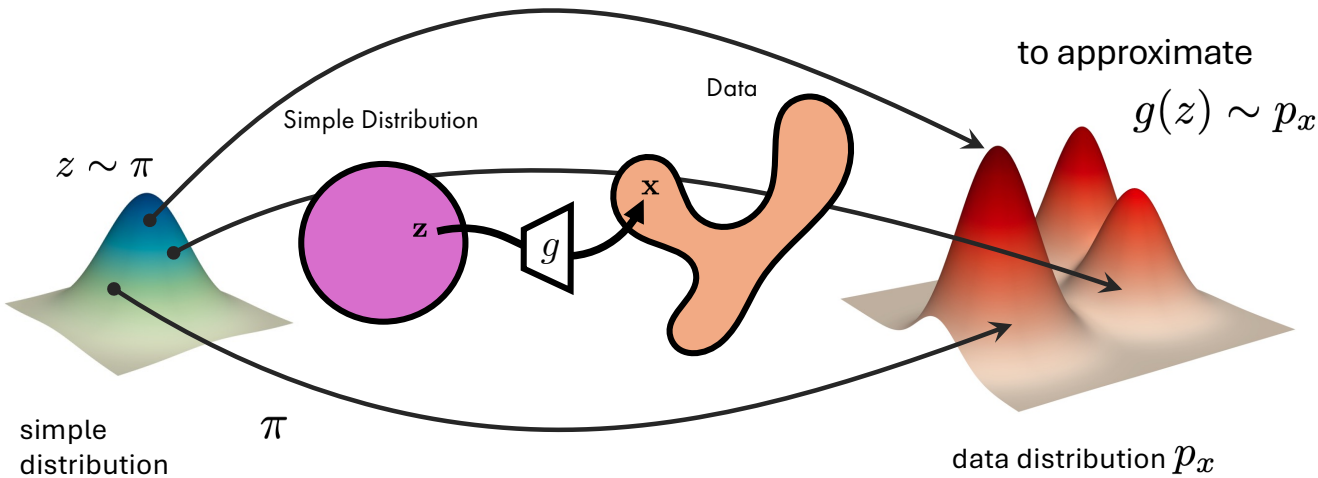
Learning to represent probability distributions

- How to sample complex data distributions $p(x)$?
- Learn the **mapping** from simple to complex distributions!



Learning to represent probability distributions

- How do we "learn" this mapping? What is the loss?



Overview

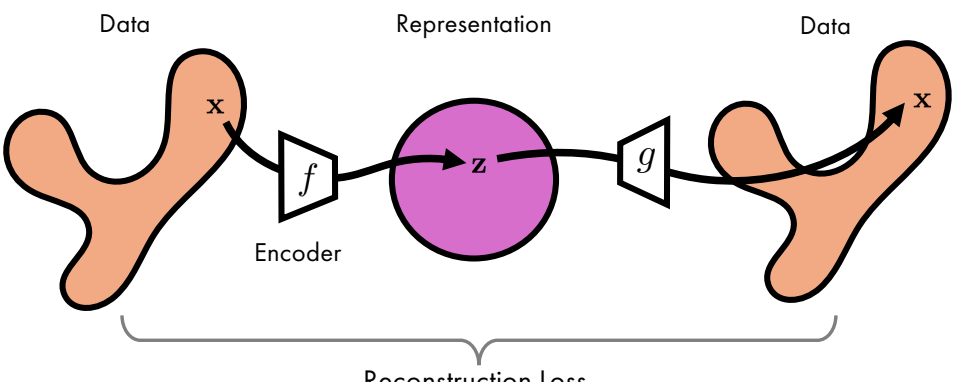
- Introduction
  - What can Generative Models do?
  - What are Generative Models?
  - Generative models and representation learning

- Generative Models: Concepts and Recent Frontiers
  - Variational Autoencoders (VAE)
  - Diffusion models

- Embedded Ethics

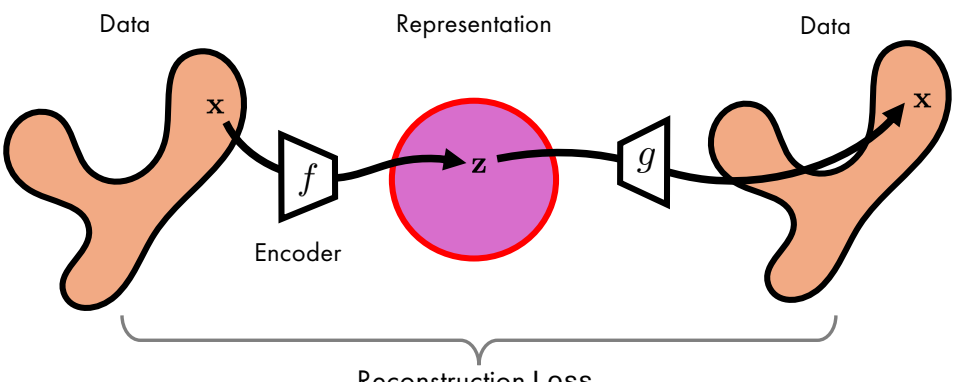
Autoencoder

- Recall Autoencoder in representation learning
- We have a compact "representation space"
- We have a decoder that maps a representation to data



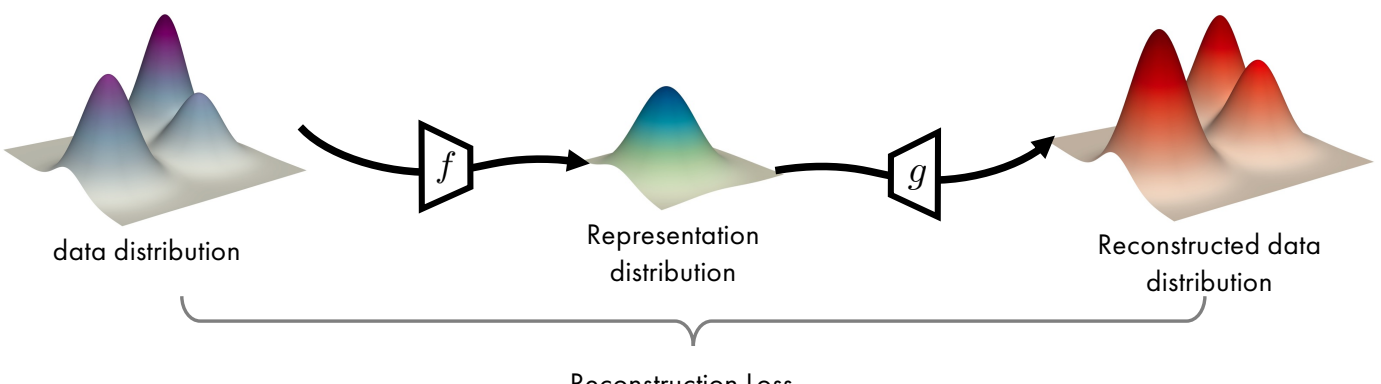
Autoencoder

- Recall Autoencoder in representation learning
- We have a compact "representation space"
- We have a decoder that maps a representation to data
- Problem: we still cannot sample this representation



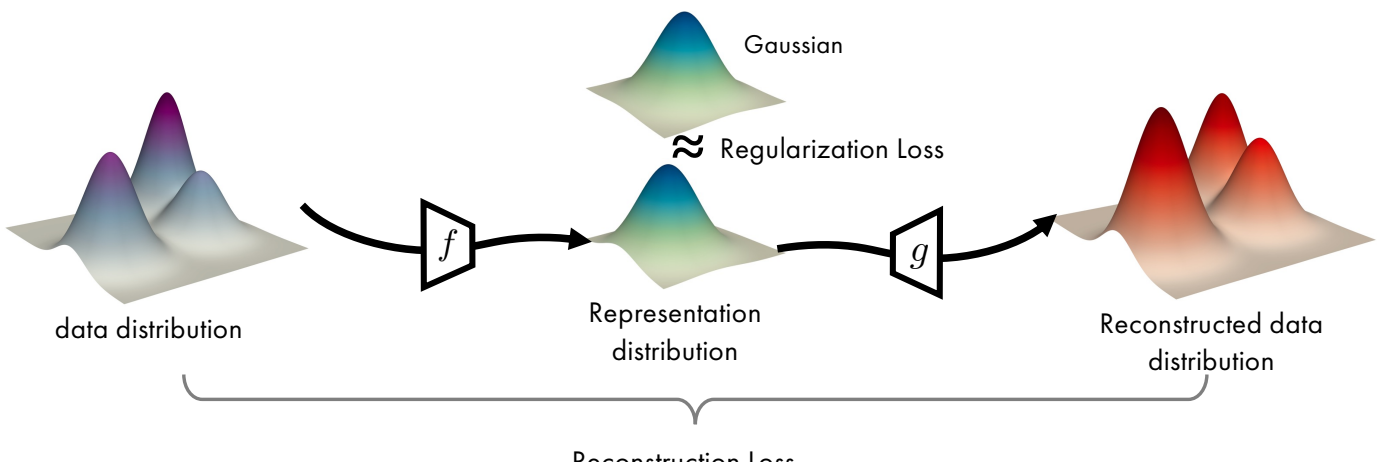
Autoencoder

- Recall Autoencoder in representation learning
- We have a compact "representation space"
- We have a decoder that maps a representation to data
- Problem: we still cannot sample this representation



Variational Autoencoder

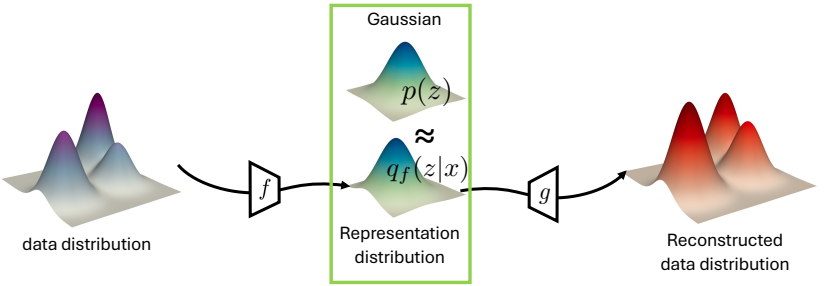
- Almost the same as Autoencoder
- Need to regularize the representation distribution, so that it follows a "samplable" distribution, e.g., Gaussian.



Variational Autoencoder

Regularization loss

$$\mathcal{D}_{\mathrm{KL}}\left(q_f(z|x)\|p(z)\right)$$



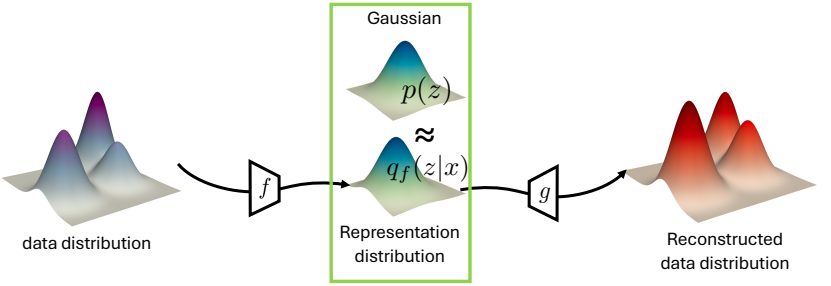
Example: Gaussian prior

- let $p(z) = \mathcal{N}(z \mid 0, \mathbf{I})$
- model $q_f(z|x)$ by Gaussian: $\mathcal{N}(z \mid \mu, \sigma)$

Variational Autoencoder

Regularization loss

$$\mathcal{D}_{\mathrm{KL}}\left(q_f(z|x)\|p(z)\right)$$



Example: Gaussian prior

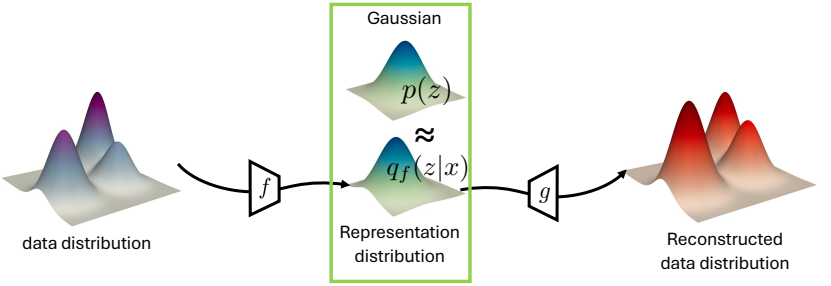
- let $p(z) = \mathcal{N}(z \mid 0, \mathbf{I})$
- model $q_f(z|x)$ by Gaussian: $\mathcal{N}(z \mid \mu, \sigma)$
- map x by encoder net: $f_\phi(x) \rightarrow \mu, \sigma$

network estimates distribution's parameters

Variational Autoencoder

Regularization loss

$$\mathcal{D}_{\mathrm{KL}}\left(q_f(z|x)\|p(z)\right)$$



Example: Gaussian prior

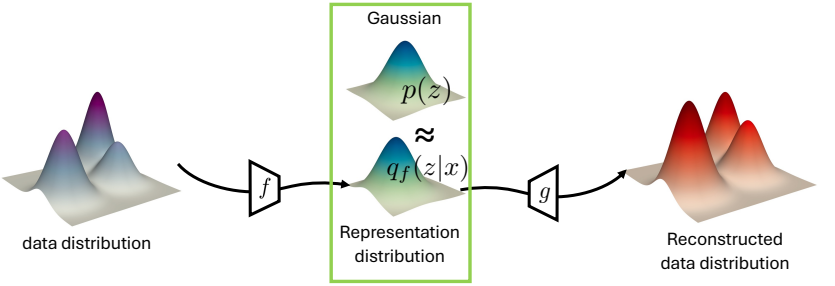
- let $p(z) = \mathcal{N}(z \mid 0, \mathbf{I})$
- model $q_f(z|x)$ by Gaussian: $\mathcal{N}(z \mid \mu, \sigma)$
- map x by encoder net: $f_\phi(x) \to \mu, \sigma$
- compute loss analytically: $\mathcal{D}_{\mathrm{KL}}\Big(\mathcal{N}(z \mid \mu, \sigma) \,||\, \mathcal{N}(z \mid 0, \mathbf{I})\Big)$

network estimates distribution's parameters

Variational Autoencoder

Regularization loss

$$\mathcal{D}_{\mathrm{KL}}\left(q_f(z|x)||p(z)\right)$$



Example: Gaussian prior

- let $p(z) = \mathcal{N}(z \mid 0, \mathbf{I})$
- model $q_f(z|x)$ by Gaussian: $\mathcal{N}(z \mid \mu, \sigma)$
- map x by encoder net: $f_\phi(x) \rightarrow \mu, \sigma$
- compute loss analytically: $\mathcal{D}_{\mathrm{KL}}\Big(\mathcal{N}(z \mid \mu, \sigma) \,||\, \mathcal{N}(z \mid 0, \mathbf{I})\Big)$
- fixed variance $\sigma = 1 \Rightarrow$ L2 loss on μ

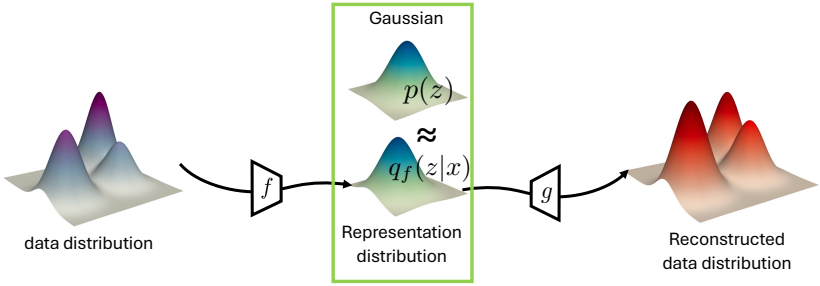
network estimates distribution's parameters

$$\mathcal{D}_{\mathrm{KL}}\left(q_f(z|x)||p(z)\right) = ||\mu||^2$$

Variational Autoencoder

Reconstruction loss

$$-\mathbb{E}_{q_f(z|x)}\left[\log p_\theta(x \mid z)\right]$$

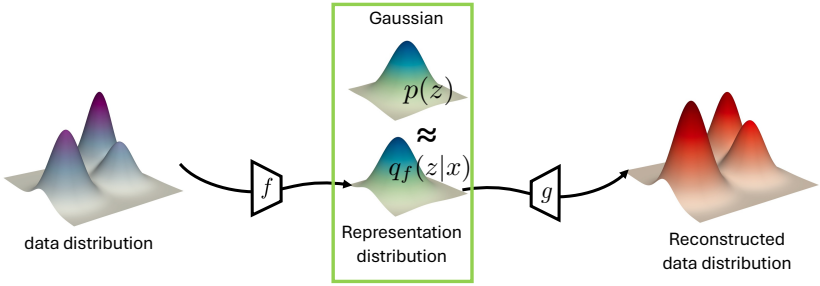


Variational Autoencoder

Reconstruction loss

$$-\mathbb{E}_{q_f(z|x)}\left[\log p_\theta(x \mid z)\right]$$

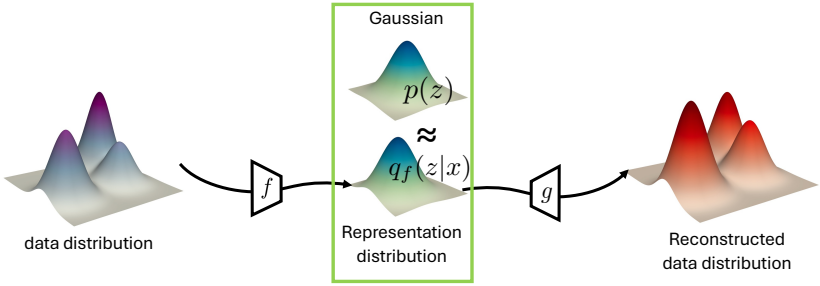
Example: Assume p(x|z) to be Gaussian



Variational Autoencoder

Reconstruction loss

$$-\mathbb{E}_{q_f(z|x)}\left[\log p_\theta(x \mid z)\right]$$



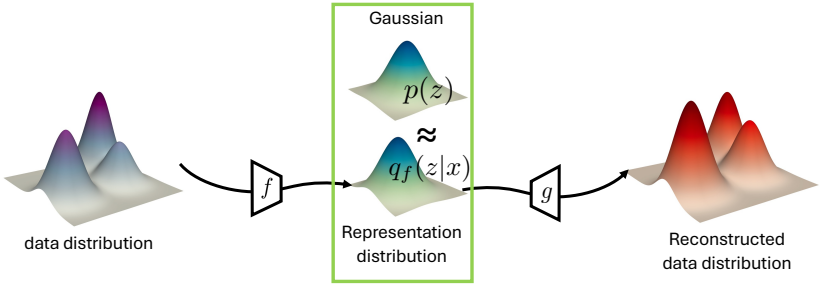
Example: Assume p(x|z) to be Gaussian

- we approximate the expectation with a single sample drawn from $q_f(z|x)$

Variational Autoencoder

Reconstruction loss

$$-\mathbb{E}_{q_f(z|x)}\left[\log p_\theta(x \mid z)\right]$$



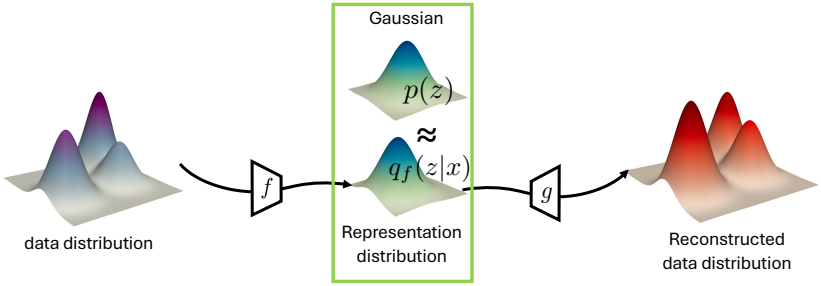
Example: Assume p(x|z) to be Gaussian

- we approximate the expectation with a single sample drawn from $q_f(z|x)$
- model each pixel as independent with its own mean and constant variance

Variational Autoencoder

Reconstruction loss

$$-\mathbb{E}_{q_f(z|x)}\left[\log p_\theta(x \mid z)\right]$$

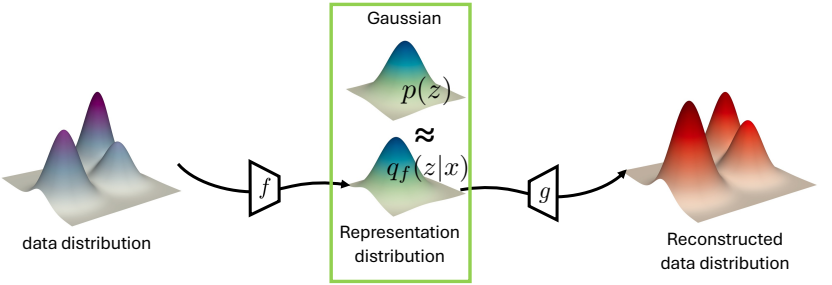


Example: Assume p(x|z) to be Gaussian

- we approximate the expectation with a single sample drawn from $q_f(z|x)$
- model each pixel as independent with its own mean and constant variance
- then the reconstruction loss reduces to simple L2 loss

$$\|x - \mu(z)\|_2^2$$

Variational Autoencoder



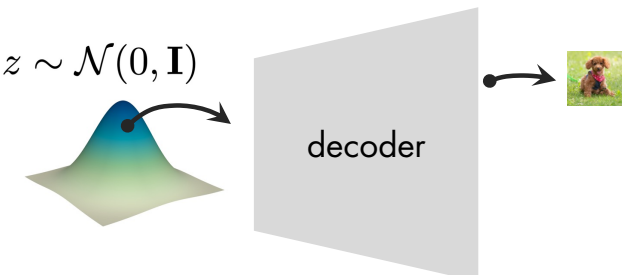
$$\mathcal{L}(g, f; \mathbf{x}) = \underbrace{-\mathbb{E}_{q_f(z|\mathbf{x})}\left[\log p_g(\mathbf{x} \mid z)\right]}_{\text{reconstruction term}} + \underbrace{D_{KL}\big(q_f(z \mid \mathbf{x}) \,\|\, p(z)\big)}_{\text{regularization term}}$$

you'll derive and implement in your homework!

Variational Autoencoder

Inference (generation):

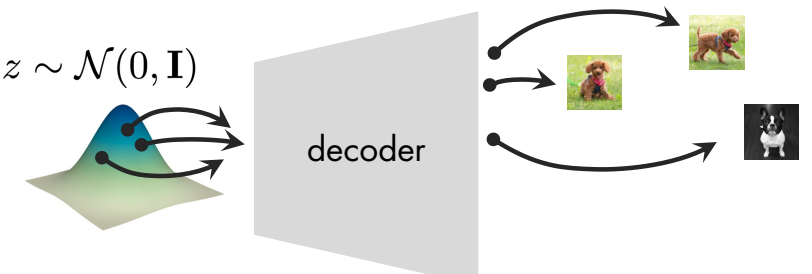
- sample z from: $\mathcal{N}(0, \mathbf{I})$
- map z by decoder net



Variational Autoencoder

Inference (generation):

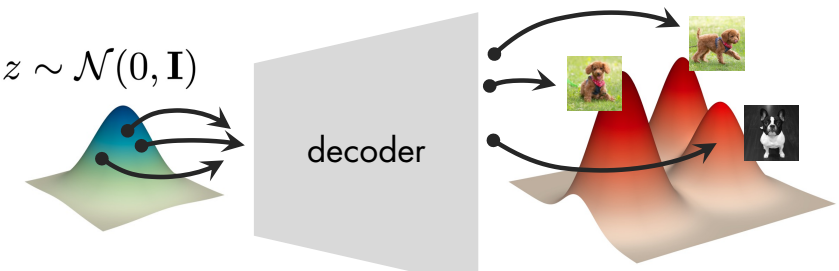
- sample z from: $\mathcal{N}(0, \mathbf{I})$
- map z by decoder net



Variational Autoencoder

Inference (generation):

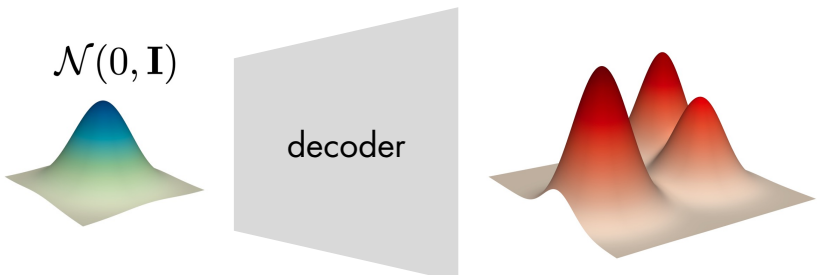
- sample z from: $\mathcal{N}(0, \mathbf{I})$
- map z by decoder net



Variational Autoencoder

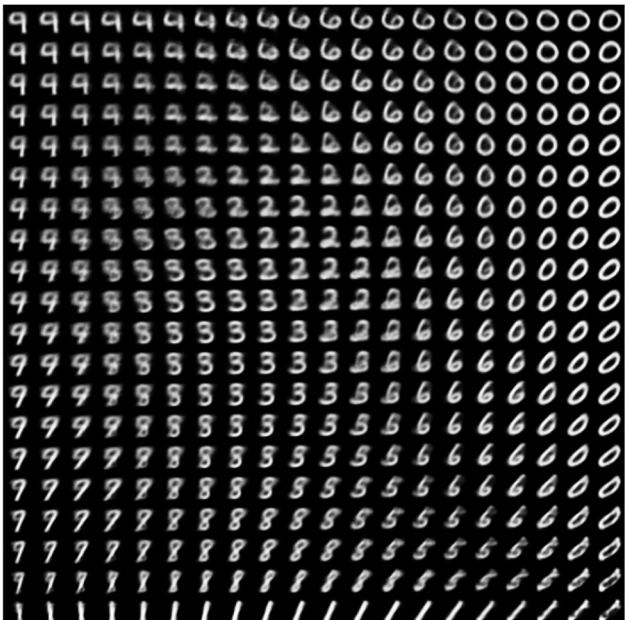
Inference (generation):

- sample z from: $\mathcal{N}(0, \mathbf{I})$
- map z by decoder net



Decoder is a deterministic mapping from one distribution to another.

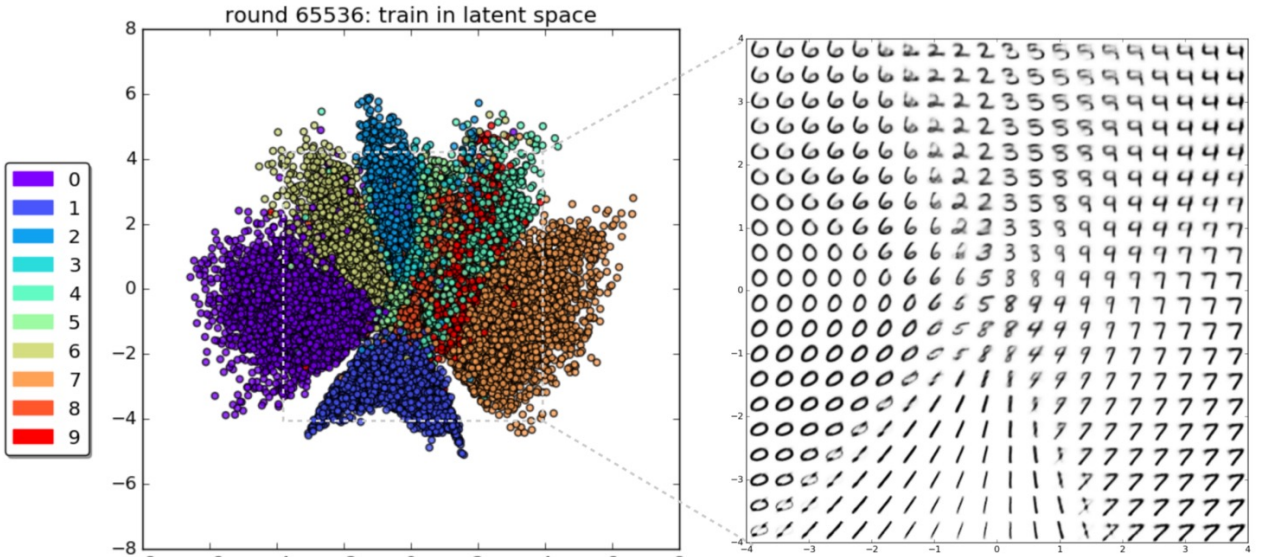
Example: 2D latent space on MNIST



"Convolutional Variational Autoencoder"

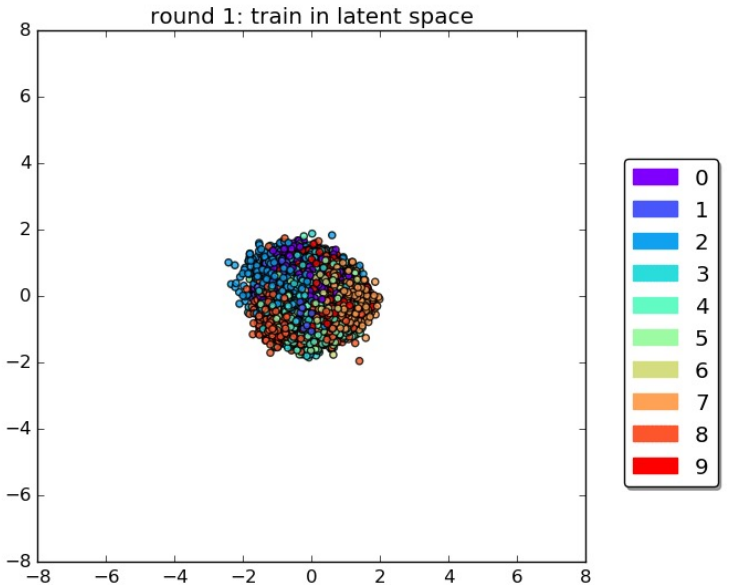
https://colab.research.google.com/github/tensorflow/docs/blob/master/site/en/tutorials/generative/cvae.ipynb

Example: 2D latent space on MNIST



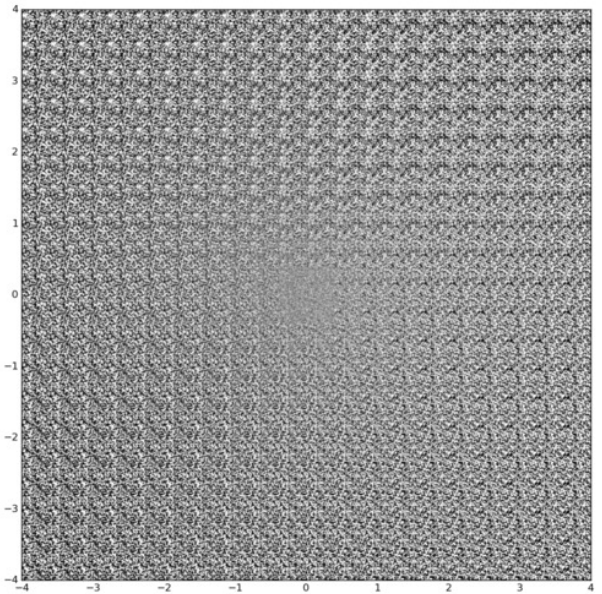
"Introducing Variational Autoencoders (in Prose and Code)"
https://blog.fastforwardlabs.com/2016/08/12/introducing-variational-autoencoders-in-prose-and-code.html

Example: 2D latent space on MNIST



"Introducing Variational Autoencoders (in Prose and Code)"
https://blog.fastforwardlabs.com/2016/08/12/introducing-variational-autoencoders-in-prose-and-code.html

Example: 2D latent space on MNIST



"Introducing Variational Autoencoders (in Prose and Code)"
https://blog.fastforwardlabs.com/2016/08/12/introducing-variational-autoencoders-in-prose-and-code.html

Overview

- Introduction
  - What can Generative Models do?
  - What are Generative Models?
  - Generative models and representation learning
- Generative Models: Concepts and Recent Frontiers
  - Variational Autoencoders (VAE)
  - Diffusion models
- Embedded Ethics

Diffusion Models... in a nutshell

Forward process
- add noise to data

Reverse process
- learn to denoise

Training objective
- Denoising reconstruction loss

Diffusion Models... in a nutshell

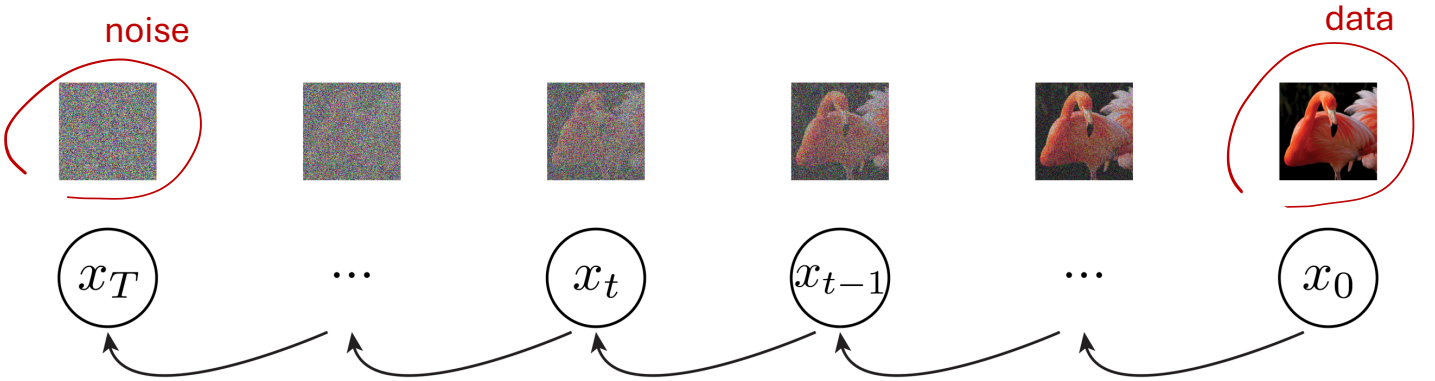
Forward process
- add noise to data

Reverse process
- learn to denoise

you'll derive these equations in your homework!

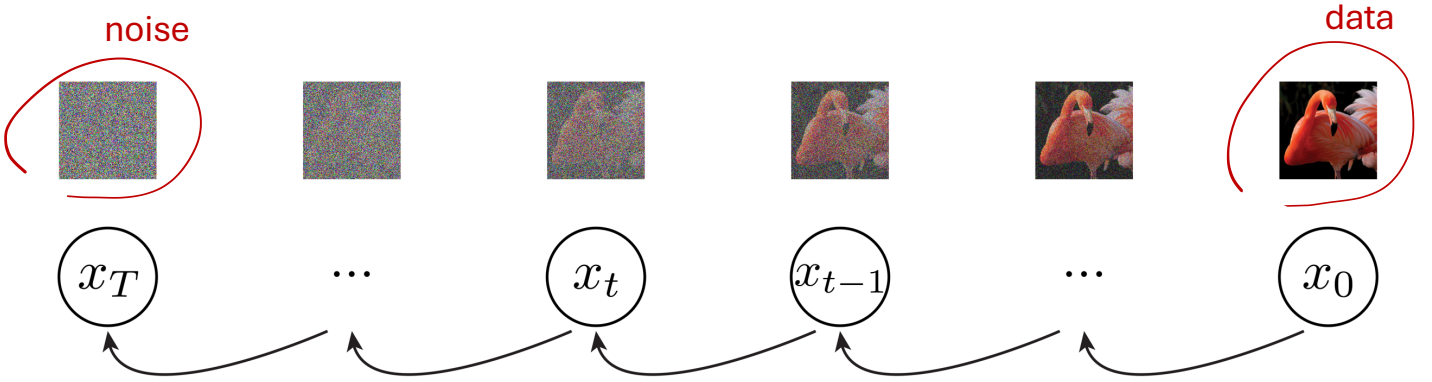
Training objective
- Denoising reconstruction loss

Diffusion Models... in a nutshell



Forward process: add noise

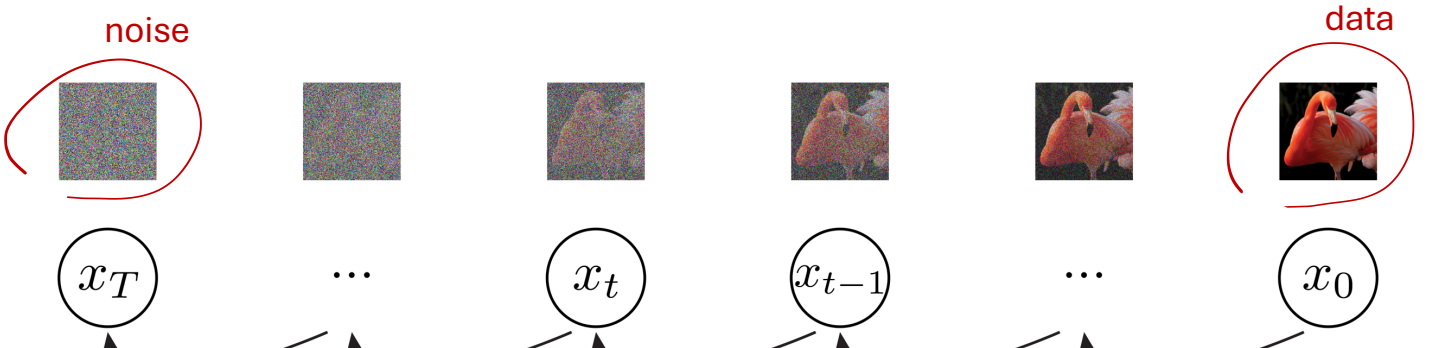
Diffusion Models... in a nutshell



x_T is just standard normal distribution!

Forward process: add noise

Diffusion Models... in a nutshell



Diffusion Models... in a nutshell

noise

data

x_T ... x_t x_{t-1} ... x_0

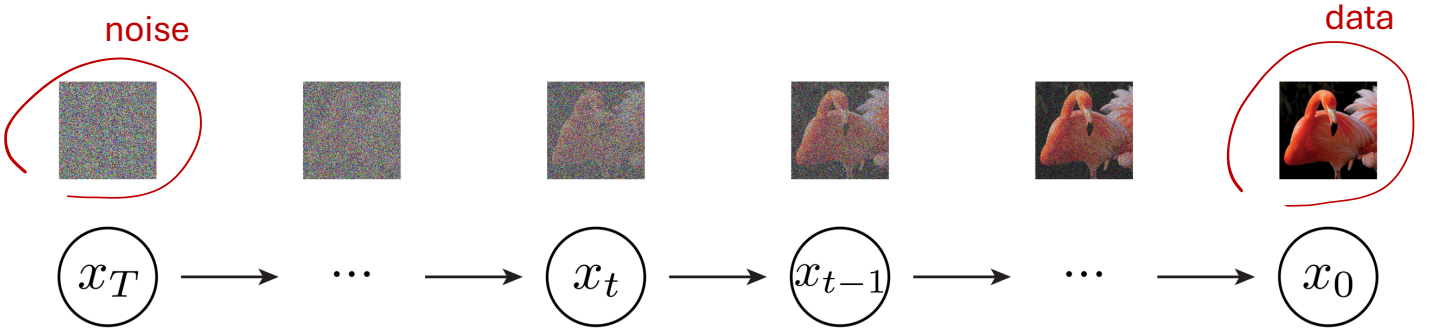
x_T is just standard normal distribution!

x_0 is also standard normal with zero variance and mean equal to the image pixel values!

Forward process: add noise

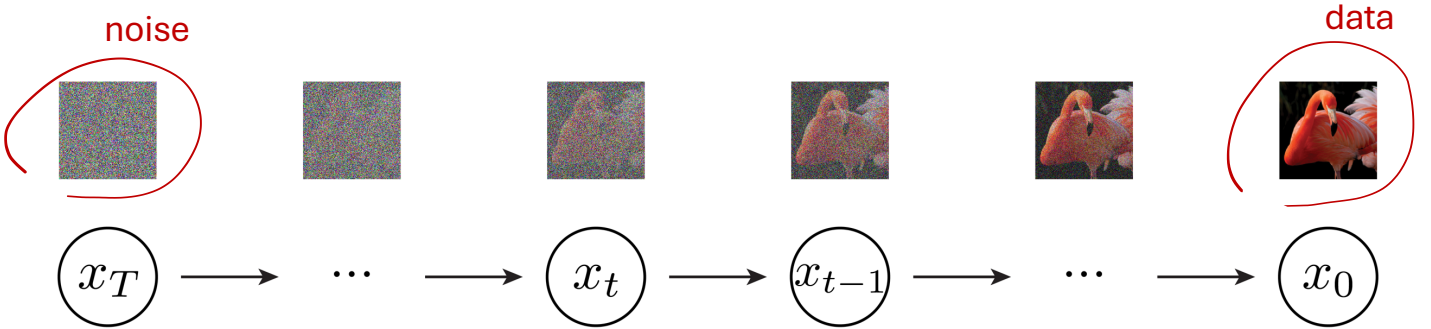
so we can just add noise to the input image to find x_t for any value of t!

Diffusion Models... in a nutshell



Reverse process: denoise

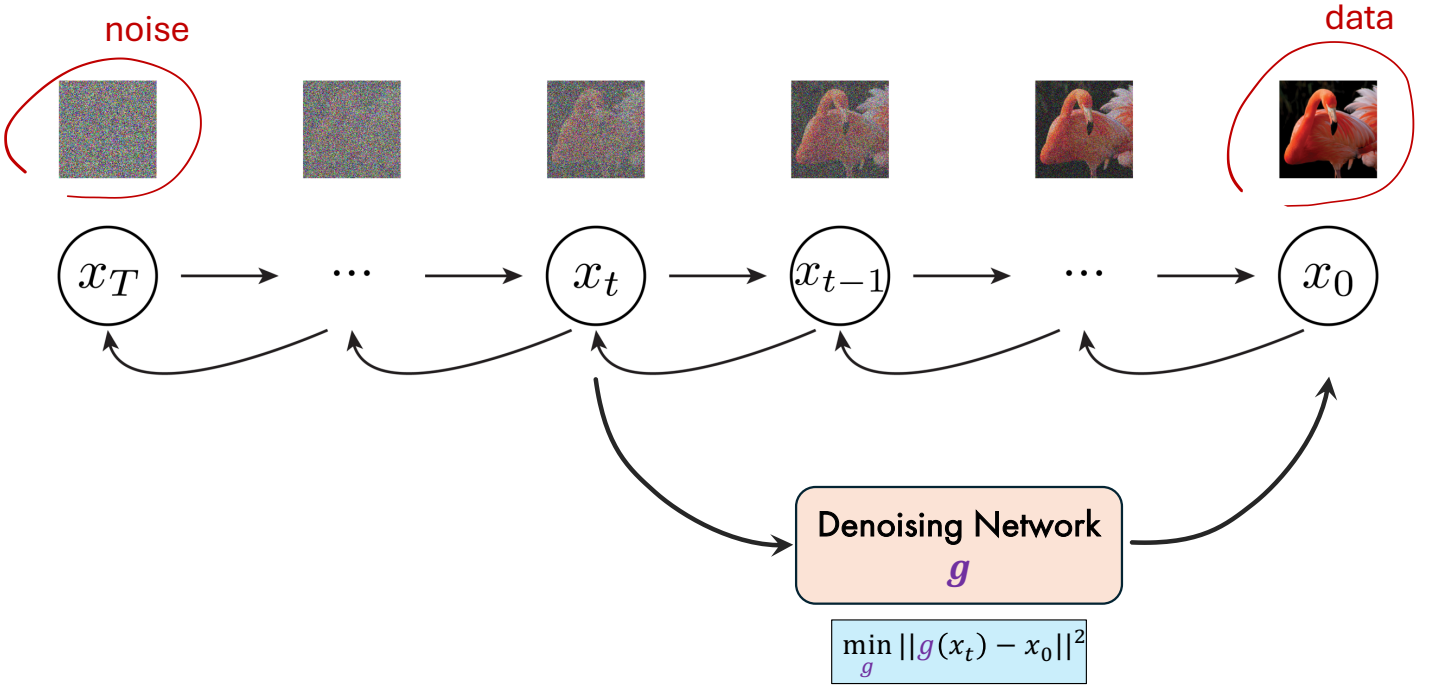
Diffusion Models... in a nutshell



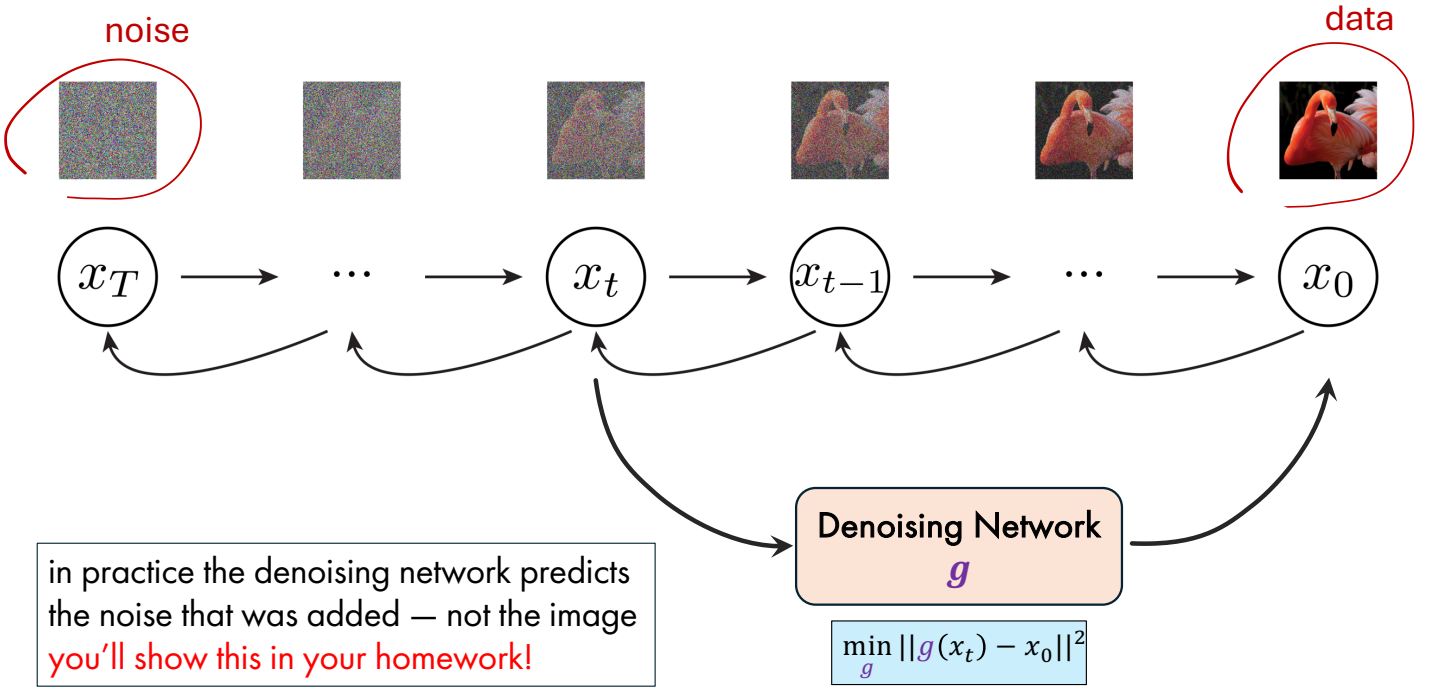
Reverse process: denoise

reversing is a little bit harder — but we can show that if you know x_t and x_0 you can find an expression for x_{t-1} (i.e., the "slightly denoised image") analytically

Diffusion Models... in a nutshell



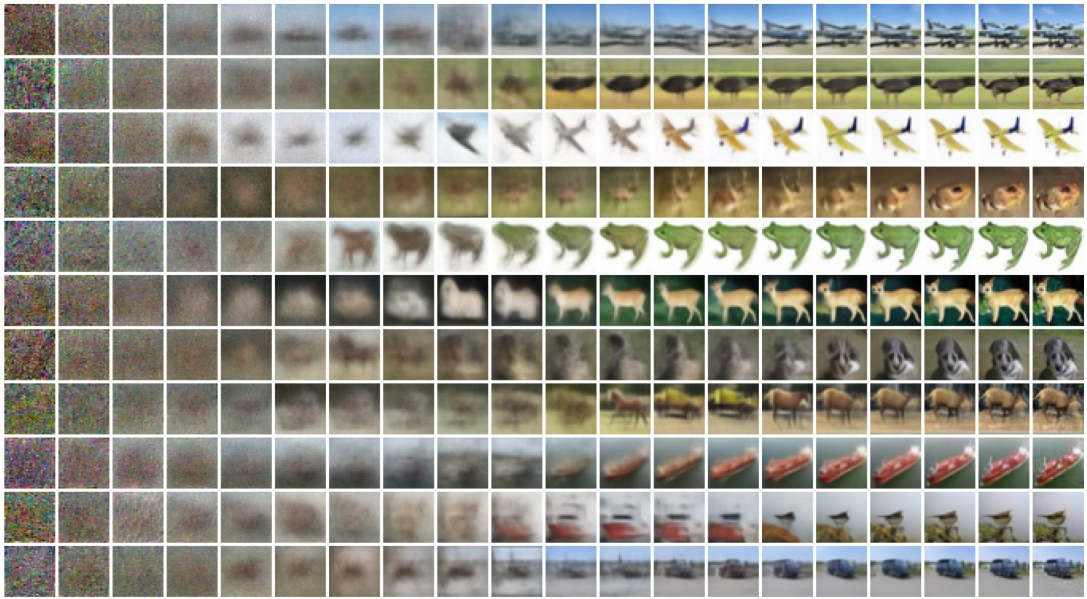
Diffusion Models... in a nutshell



in practice the denoising network predicts the noise that was added — not the image
you'll show this in your homework!

Example: Unconditional Generation on CIFAR-10

noise
x_T



generated
x_0

[Ho et al, 2020]

Overview

- Introduction
  - What can Generative Models do?
  - What are Generative Models?
  - Generative models and representation learning

- Generative Models: Concepts and Recent Frontiers
  - Variational Autoencoders (VAE)
  - Diffusion models

- Embedded Ethics

Warm-up activity

https://www.cs.toronto.edu/~lindell/teaching/420/slides/celebahq.html

Samples from CelebA-HQ Dataset

Warm-up activity

Small group discussions

- What groups are not represented in this dataset?

https://www.cs.toronto.edu/~lindell/teaching/420/slides/celebahq.html

Warm-up activity

Small group discussions

- What groups are not represented in this dataset?
- What is this dataset not useful for?

https://www.cs.toronto.edu/~lindell/teaching/420/slides/celebahq.html

Warm-up activity

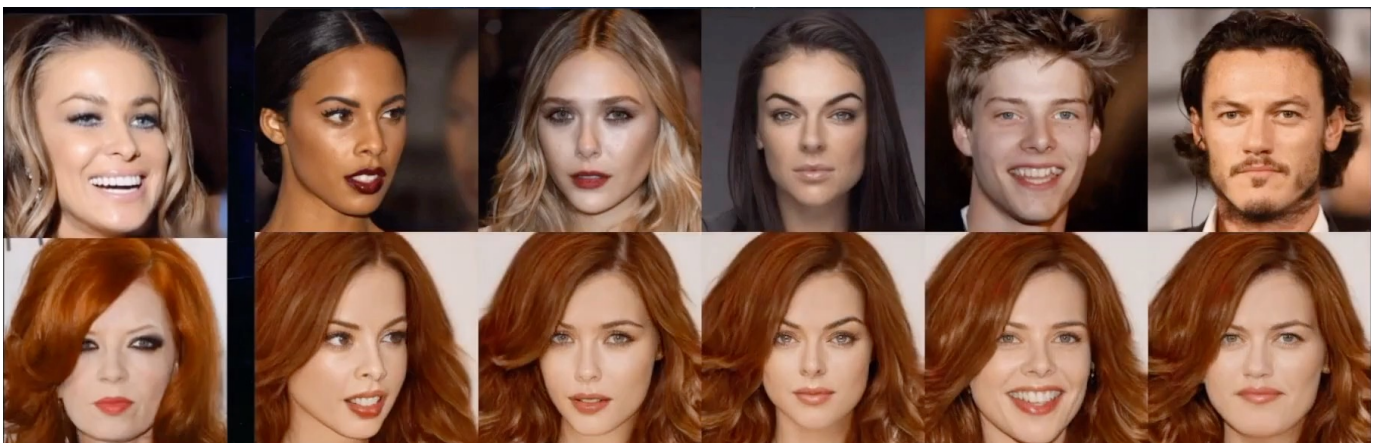
Small group discussions

- What groups are not represented in this dataset?
- What is this dataset not useful for?
- What is this dataset useful for?

https://www.cs.toronto.edu/~lindell/teaching/420/slides/celebahq.html

Image stylization with CelebA-HQ

source images



reference

stylized images

[Choi et al. ‘20]

Projecting images onto CelebA-HQ models

[Esser et al. ‘20]

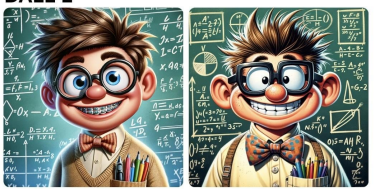
image generation in 2024

DALL·E >

You

Make me an image of a cartoon nerd without eyeglasses

DALL·E



Here are two images of a cartoon nerd without eyeglasses. Let me know if there's anything else you need!

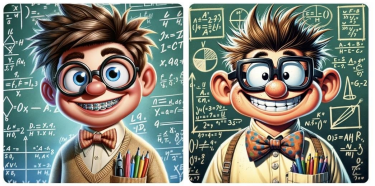
reddit u/nwerdnerd

DALL·E >

You

Make me an image of a cartoon nerd without eyeglasses

DALL·E

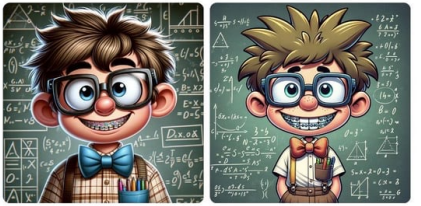


Here are two images of a cartoon nerd without eyeglasses. Let me know if there's anything else you need!

You

They both have eyeglasses on them , can you do them without eyeglasses

DALL·E



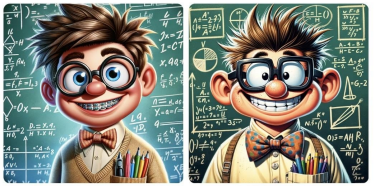
Here are two images of a cartoon nerd, created with the specific instruction to exclude eyeglasses. If you have any more requests or need further adjustments, feel free to let me know!

reddit u/nwerdnerd

DALL·E >

You
Make me an image of a cartoon nerd without eyeglasses

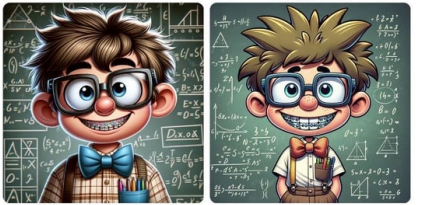
DALL·E



Here are two images of a cartoon nerd without eyeglasses. Let me know if there's anything else you need!

You
They both have eyeglasses on them , can you do them without eyeglasses

DALL·E



Here are two images of a cartoon nerd, created with the specific instruction to exclude eyeglasses. If you have any more requests or need further adjustments, feel free to let me know!

You
Ok, do you know what eyeglasses are ?

DALL·E
Yes, I understand what eyeglasses are. They are frames bearing lenses worn in front of the eyes, typically for vision correction or eye protection. I apologize for the oversight in the images provided. If you need an image without eyeglasses, I can certainly create one that meets your specifications.

You
Make me an image of only eyeglasses

DALL·E

reddit u/nwerdnerd

image generation in 2025

make me an image of a cartoon nerd without eyeglasses

Here's your cartoon nerd without eyeglasses! Let me know if you'd like any tweaks.

Here's your cartoon nerd without eyeglasses! Let me know if you need any modifications.

no, it still has eyeglasses. please make the nerd without any kind of eye glasses. you do know what eyeglasses are, right?

Here you go—this nerd has no eyeglasses at all! Let me know if you'd like any adjustments.

Ethics of Bias & Discrimination

Mitigation Strategies

Collect Better Datasets?

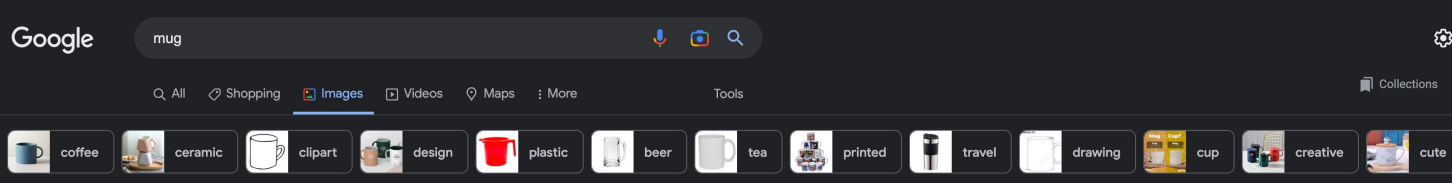
- All datasets are finite attempts at sampling a distribution

Collect Better Datasets?

- All datasets are finite attempts at sampling a distribution
- Deep learning models used to obtain SOTA results on standard benchmarks are typically trained on hundreds of thousands to billions of training examples

Collect Better Datasets?

- All datasets are finite attempts at sampling a distribution
- Deep learning models used to obtain SOTA results on standard benchmarks are typically trained on hundreds of thousands to billions of training examples
- It is not always practical to collect representative examples of this size from various populations to include in the training data


Google
mug
All
Shopping
Images
Videos
Maps
More
Tools
Collections
coffee
ceramic
clipart
design
plastic
beer
tea
printed
travel
drawing
cup
creative
cute

Google

mug

All | Shopping | Images | Videos | Maps | More | Tools | Collections

coffee | ceramic | clipart | design | plastic | beer | tea | printed | travel | drawing | cup | creative | cute

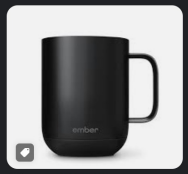

IKEA · In stock
IKEA - FÄRGKLAR Mug Matte ...

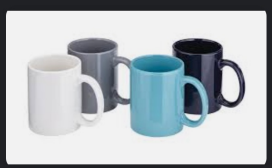

Amazon.ca · In stock
20 OZ Large Coffee Mug, Smilat...


IKEA · In stock
IKEA - DINERA Mug Dark gray,...


Amazon.ca
New Ember Temperature ...


IKEA
Mug, glossy beige, 13 oz (37 ...


Apple · In stock
Ember Mug 2 Black 10 oz


Canadian Tire
MASTER Chef 4pc Stoneware Mug Set ...


IKEA · In st
IKEA 365+ Mu

5+ stores
YETI Canada · In stock
Yeti Rambler 414ml / 14oz Mug ...


Rogue Wave Coffee · In stock
ORIGAMI - Barrel Aroma Mug Gr...

3 stores
Chapters Indigo · In stock
Friends Cappuccino Mug Centra...


Main and Local · In stock
CBC Retro Logo Mug - Main and...


Wayfair Canada · In stock
Coffee Mugs & Tea Cups - Wayf...


Amazon.ca · In stock
Cricut Beveled Blank Mug, Cera...


Hermes
Rocabar mug | Hermès Canada

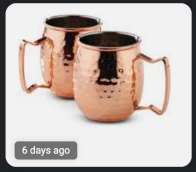

Williams Sono
Gold Monogram


Vistaprint

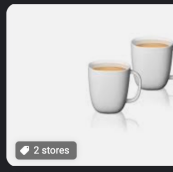

Apple


7 hours ago
The Toronto Star


GameStop


6 days ago
Canadian Tire


Boutique RICARDO


2 stores
Nespresso · In stock

Proceedings of Machine Learning Research 81:1–15, 2018 Conference on Fairness, Accountability, and Transparency

Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*

Joy Buolamwini JOYAB@MIT.EDU
MIT Media Lab 75 Amherst St. Cambridge, MA 02139

Timnit Gebru TIMNIT.GEBRU@MICROSOFT.COM
Microsoft Research 641 Avenue of the Americas, New York, NY 10011

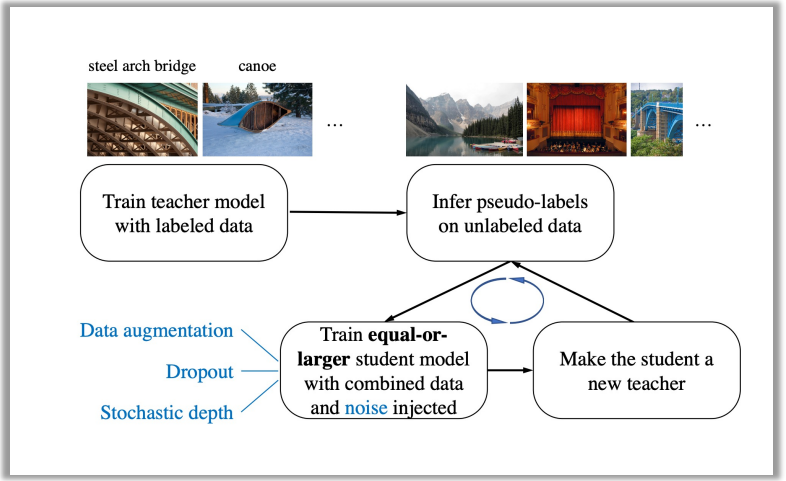
Editors: Sorelle A. Friedler and Christo Wilson

4.1. Key Findings on Evaluated Classifiers

- All classifiers perform better on male faces than female faces (8.1% − 20.6% difference in error rate)
- All classifiers perform better on lighter faces than darker faces (11.8% − 19.2% difference in error rate)
- All classifiers perform worst on darker female faces (20.8% − 34.7% error rate)
- Microsoft and IBM classifiers perform best on lighter male faces (error rates of 0.0% and 0.3% respectively)
- Face++ classifiers perform best on darker male faces (0.7% error rate)
- The maximum difference in error rate between the best and worst classified groups is 34.4%

Other mitigation strategies

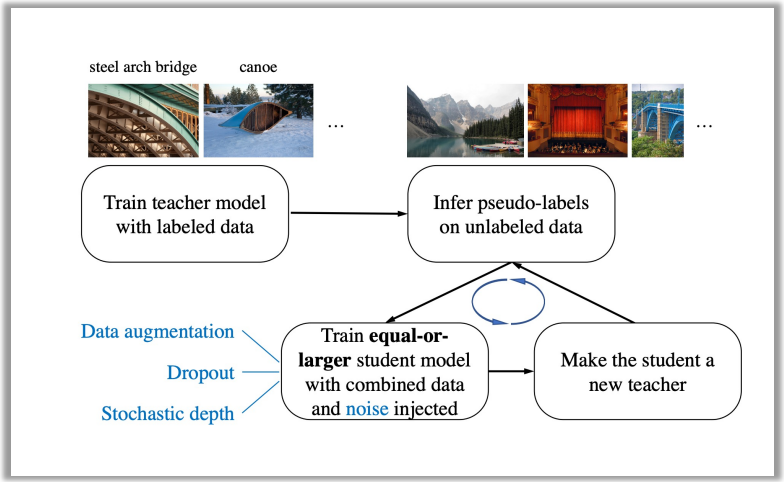
- Train on balanced dataset, finetune on biased dataset to reduce chances of overfitting



ResNet-50 Billion-scale [93]	26M	3.5B images labeled with tags	81.2%	96.0%
ResNeXt-101 Billion-scale [93]	193M		84.8%	-
ResNeXt-101 WSL [55]	829M		85.4%	97.6%
FixRes ResNeXt-101 WSL [86]	829M		86.4%	98.0%
Big Transfer (BiT-L) [43]†	928M	300M weakly labeled images from JFT	87.5%	98.5%
Noisy Student Training (EfficientNet-L2)	480M	300M unlabeled images from JFT	**88.4%**	**98.7%**

Other mitigation strategies

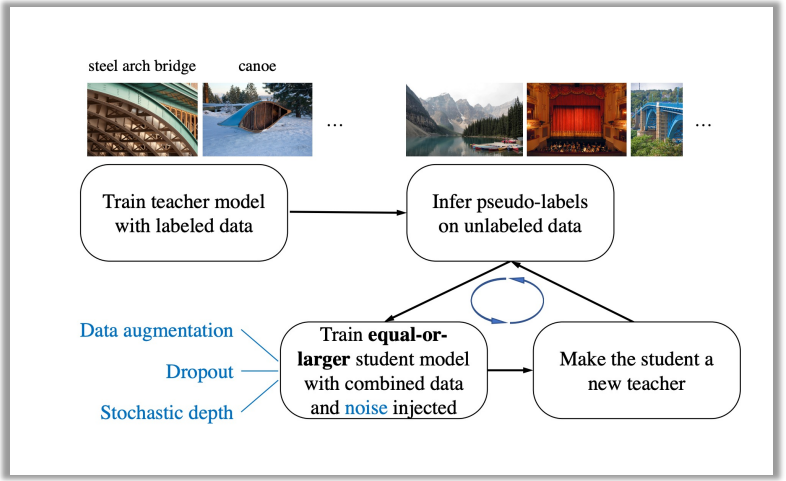
- Train on balanced dataset, finetune on biased dataset to reduce chances of overfitting
- Leverage labeled and unlabeled data to reduce sampling bias, increase size of dataset



ResNet-50 Billion-scale [93]	26M	3.5B images labeled with tags	81.2%	96.0%
ResNeXt-101 Billion-scale [93]	193M		84.8%	-
ResNeXt-101 WSL [55]	829M		85.4%	97.6%
FixRes ResNeXt-101 WSL [86]	829M		86.4%	98.0%
Big Transfer (BiT-L) [43]†	928M	300M weakly labeled images from JFT	87.5%	98.5%
Noisy Student Training (EfficientNet-L2)	480M	300M unlabeled images from JFT	**88.4%**	**98.7%**

Other mitigation strategies

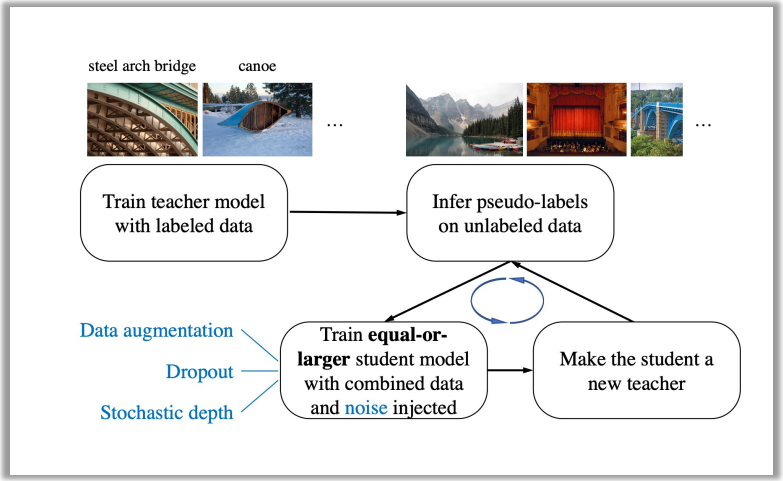
- Train on balanced dataset, finetune on biased dataset to reduce chances of overfitting
- Leverage labeled and unlabeled data to reduce sampling bias, increase size of dataset
- Data augmentation, re-balancing



ResNet-50 Billion-scale [93]	26M	3.5B images labeled with tags	81.2%	96.0%
ResNeXt-101 Billion-scale [93]	193M		84.8%	-
ResNeXt-101 WSL [55]	829M		85.4%	97.6%
FixRes ResNeXt-101 WSL [86]	829M		86.4%	98.0%
Big Transfer (BiT-L) [43]†	928M	300M weakly labeled images from JFT	87.5%	98.5%
Noisy Student Training (EfficientNet-L2)	480M	300M unlabeled images from JFT	**88.4%**	**98.7%**

Other mitigation strategies

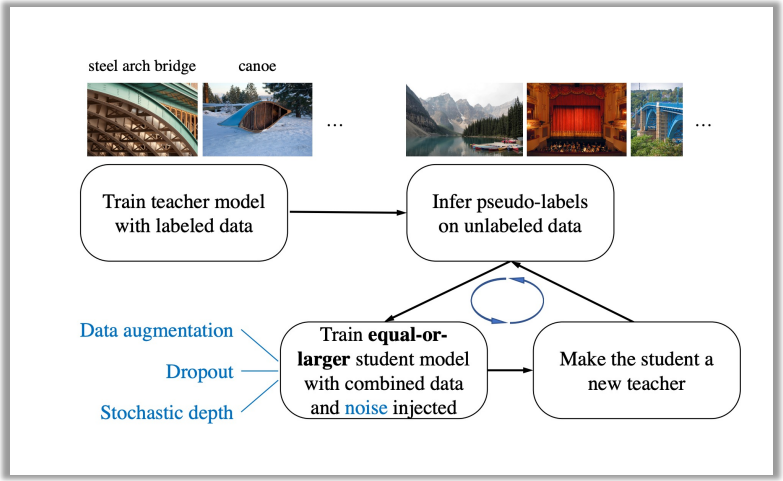
- Train on balanced dataset, finetune on biased dataset to reduce chances of overfitting
- Leverage labeled and unlabeled data to reduce sampling bias, increase size of dataset
- Data augmentation, re-balancing
- Evaluate on multiple datasets, balanced test sets



ResNet-50 Billion-scale [93]	26M	3.5B images labeled with tags	81.2%	96.0%
ResNeXt-101 Billion-scale [93]	193M		84.8%	-
ResNeXt-101 WSL [55]	829M		85.4%	97.6%
FixRes ResNeXt-101 WSL [86]	829M		86.4%	98.0%
Big Transfer (BiT-L) [43]†	928M	300M weakly labeled images from JFT	87.5%	98.5%
Noisy Student Training (EfficientNet-L2)	480M	300M unlabeled images from JFT	**88.4%**	**98.7%**

Other mitigation strategies

- Train on balanced dataset, finetune on biased dataset to reduce chances of overfitting
- Leverage labeled and unlabeled data to reduce sampling bias, increase size of dataset
- Data augmentation, re-balancing
- Evaluate on multiple datasets, balanced test sets
- … (potentially many others!)



ResNet-50 Billion-scale [93]	26M	3.5B images labeled with tags	81.2%	96.0%
ResNeXt-101 Billion-scale [93]	193M		84.8%	-
ResNeXt-101 WSL [55]	829M		85.4%	97.6%
FixRes ResNeXt-101 WSL [86]	829M		86.4%	98.0%
Big Transfer (BiT-L) [43]†	928M	300M weakly labeled images from JFT	87.5%	98.5%
Noisy Student Training (EfficientNet-L2)	480M	300M unlabeled images from JFT	**88.4%**	**98.7%**

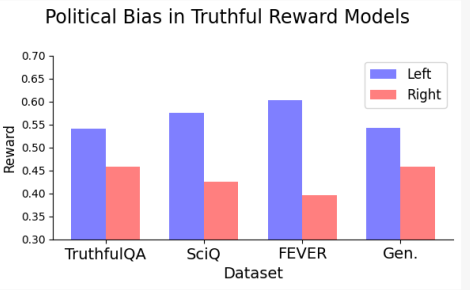
Takeaways

- dataset bias & mitigations still active research area!
- not only algorithmic performance, there are ethical implications
- depends both on data, but model architecture and even hyperparameter choices can all play a role

Study: Some language reward models exhibit political bias

Research from the MIT Center for Constructive Communication finds this effect occurs even when reward models are trained on factual data.

Ellen Hoffman | Center for Constructive Communication
December 10, 2024



Next time: Digital Photography