CSC413/2516 Lecture 9: Large Language Models and GANs

Jimmy Ba and Bo Wang

Overview

Four modern approaches to generative modeling:

- Autoregressive models (Lectures 3, 7, 8 and this lecture)
- Generative adversarial networks (this lecture)
- Diffusion models (next lecture)
- Variational autoencoders (next lecture)

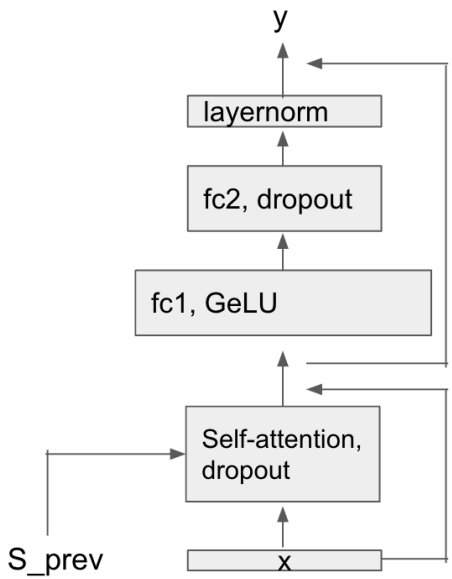
All four approaches have different pros and cons.

Transformers is all you need

- We have seen transformers in the previous lectures.
- In this lecture, we will take a look at how to scale up transformers to billions of parameters in models like ChatGPT.
  - Model architectures
  - Data and model parallelism

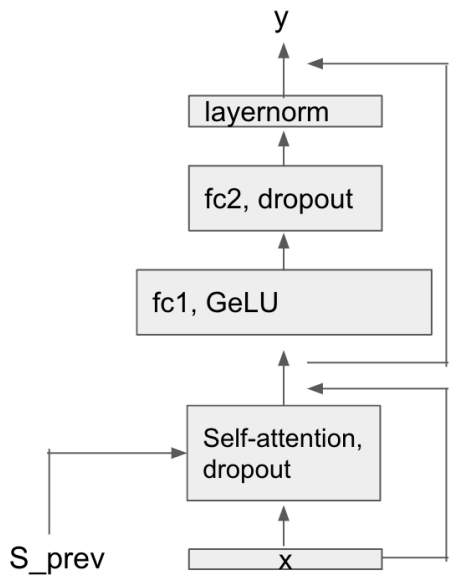
Model architectures

- The basic building block of generative pretrained transformer (GPT):



Model architectures

- Counting the number of parameters in a building block:



LN_#parameter = emb_size x 2

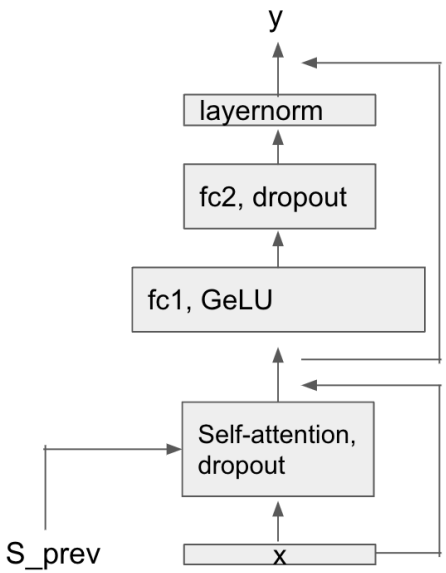
fc2_#parameter = emb_size x 4 x emb_size

fc1_#parameter = emb_size x emb_size x 4

self_attn_#parameter = emb_size x emb_size x 4

Model architectures

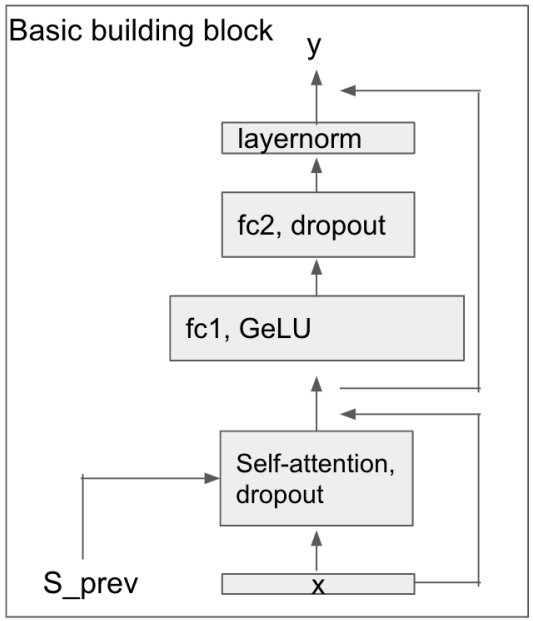
- The total number of parameters in a building block $\approx 12emb^2$



FFN width	8192, 16384, …, 49152 (4x emb)
#attention heads	32, 32, …, 96
embedding size	2048, 4096, …, 12288

Model architectures

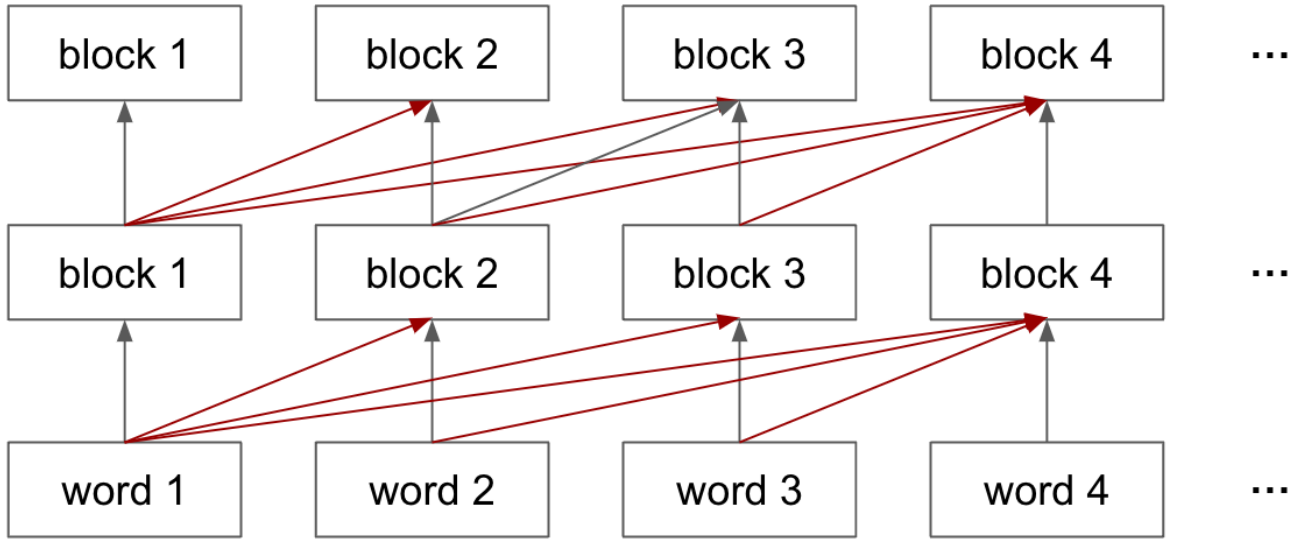
- The basic building block of generative pretrained transformer (GPT):



#training tokens	4B*, 20B*, …, 500B (~3x #P)
#parameters	1.3B, 6.6B, …, 175B
#layers	24, 32, …, 96
FFN width	8192, 16384, …, 49152 (4x emb)
#attention heads	32, 32, …, 96
embedding size	2048, 4096, …, 12288

Model architectures

- How many parameters in a GPT model with L layers and H emb size?



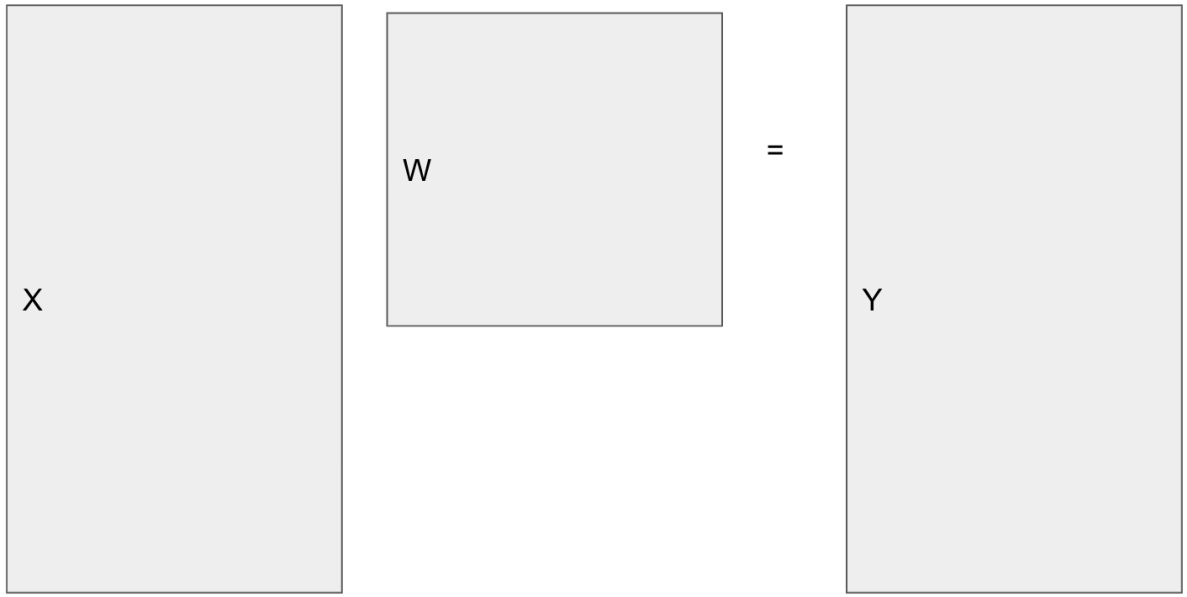
Model architectures

- Training data size v.s. model size in the recent GPT-like models:
- Are the models under trained or over trained?

	GPT-2	GPT-3	OPT	BLOOM	PaLM	Chinchilla
#params	1.6B	175B	175B	175B	540B	70B
#training tokens	9B	500B (92% eng)	180B	350B* (30% eng~105B)	780B (78% eng)	1400B

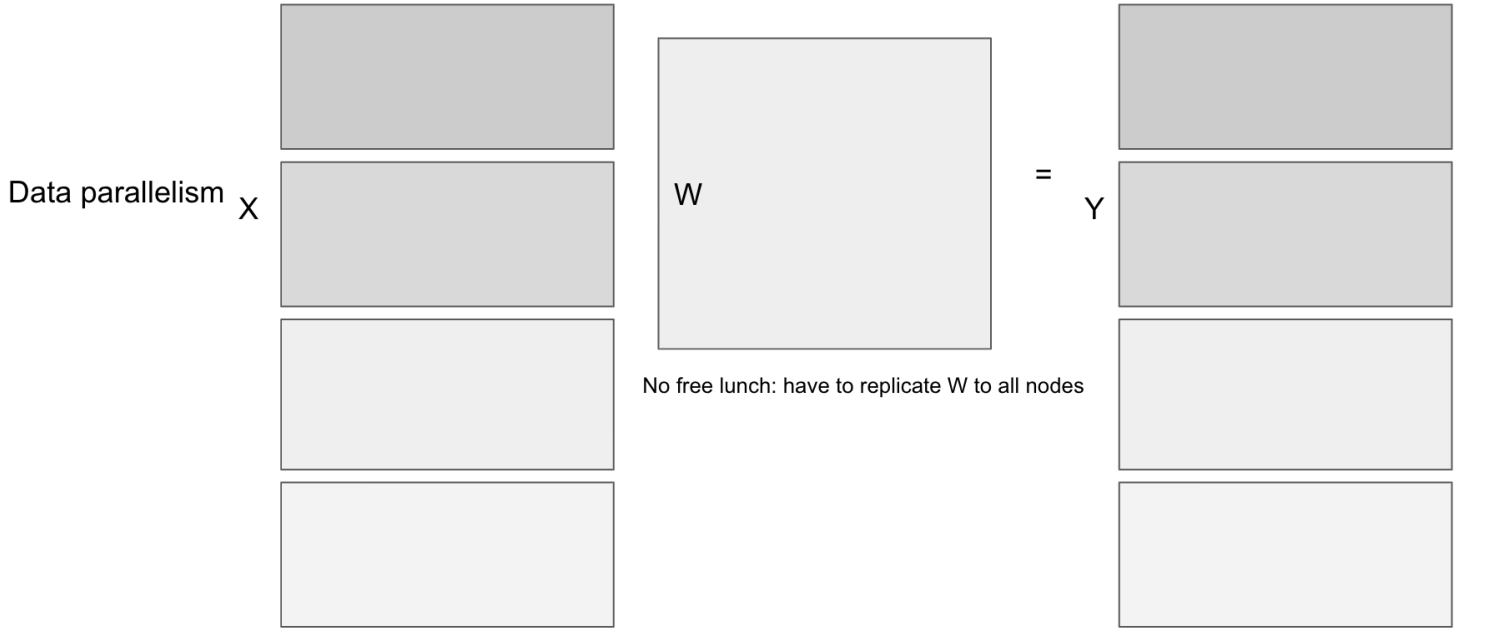
Data and Model Parallelism

- Matrix multiplication is the most basic operation in deep learning:



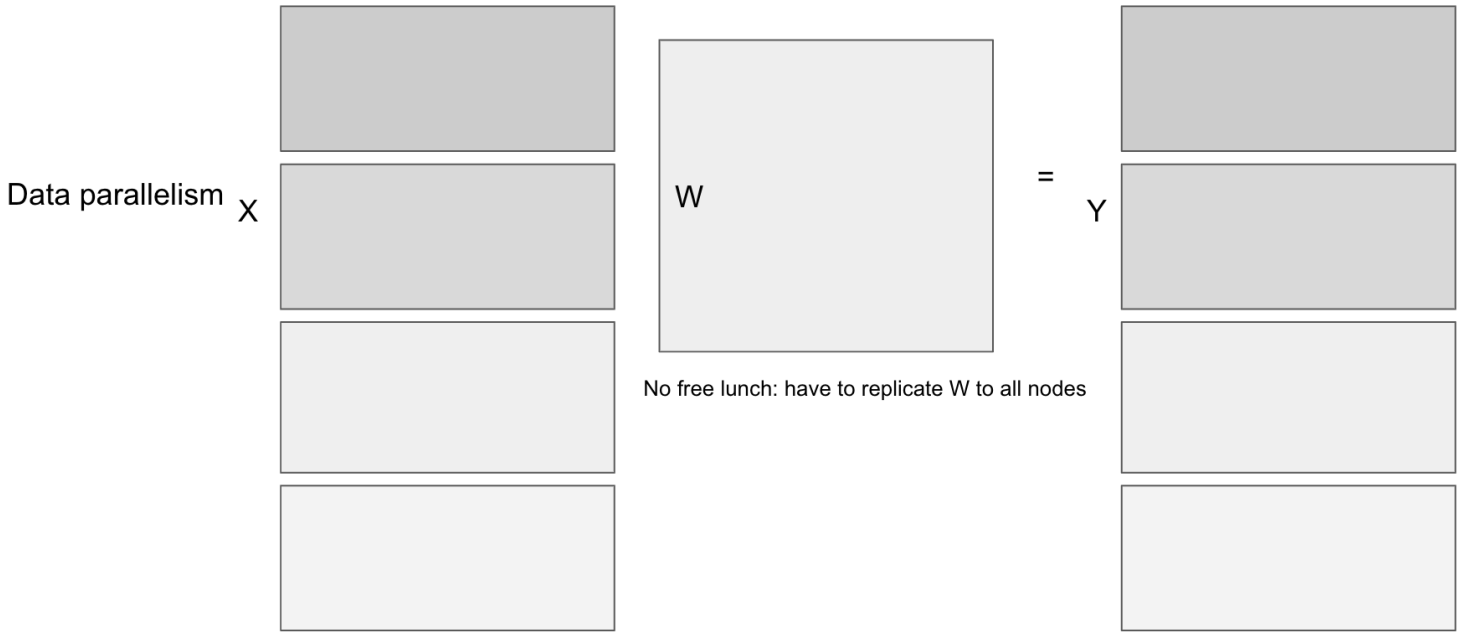
Data and Model Parallelism

- Naive data parallelism is the most common approach to utilize many parallel compute:



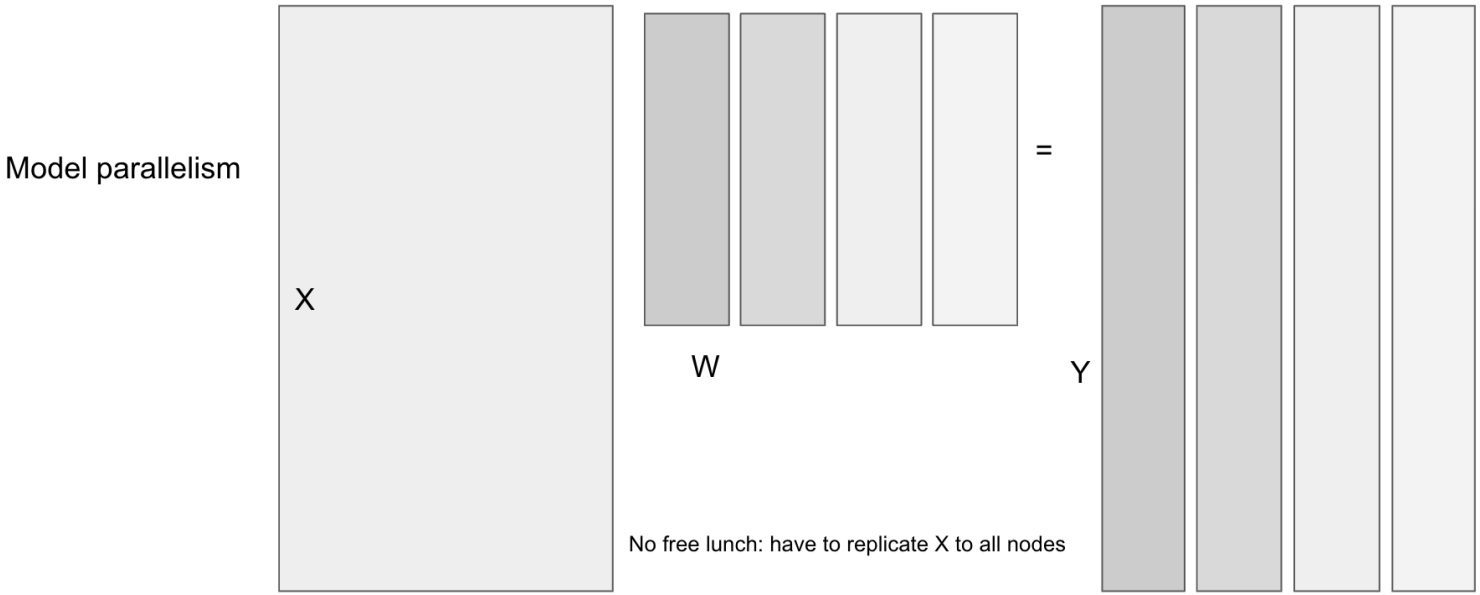
Data and Model Parallelism

- The communication cost in the naive data parallelism $\approx 2W$, why?



Data and Model Parallelism

- Model parallelism is similar to data parallelism just on another axis:



Data and Model Parallelism

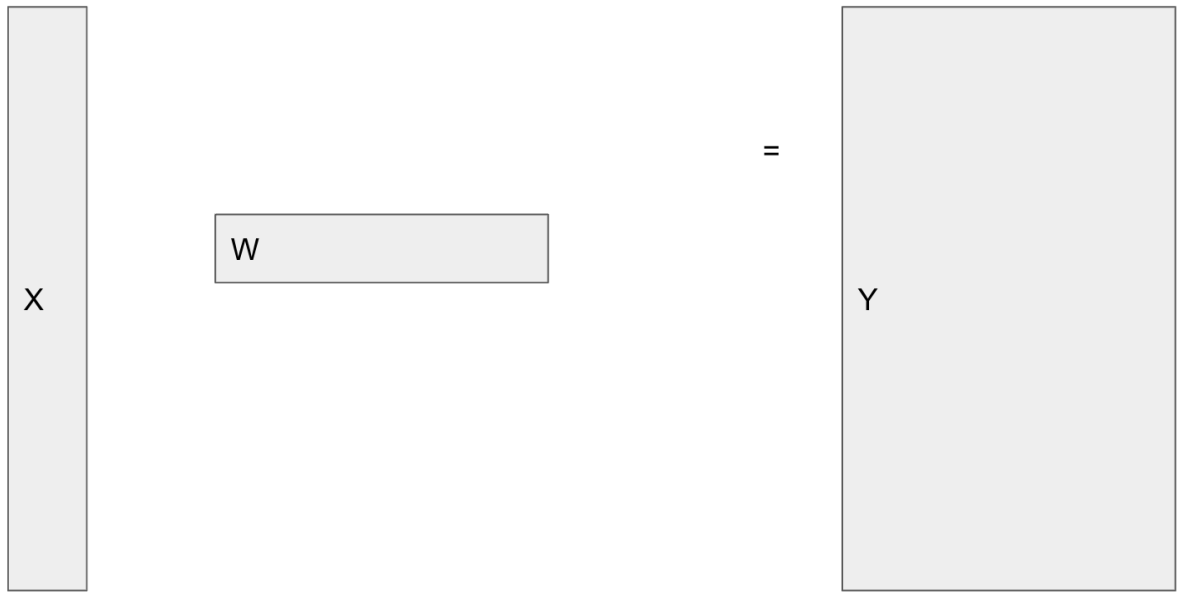
Example:

Data and Model Parallelism

Example:

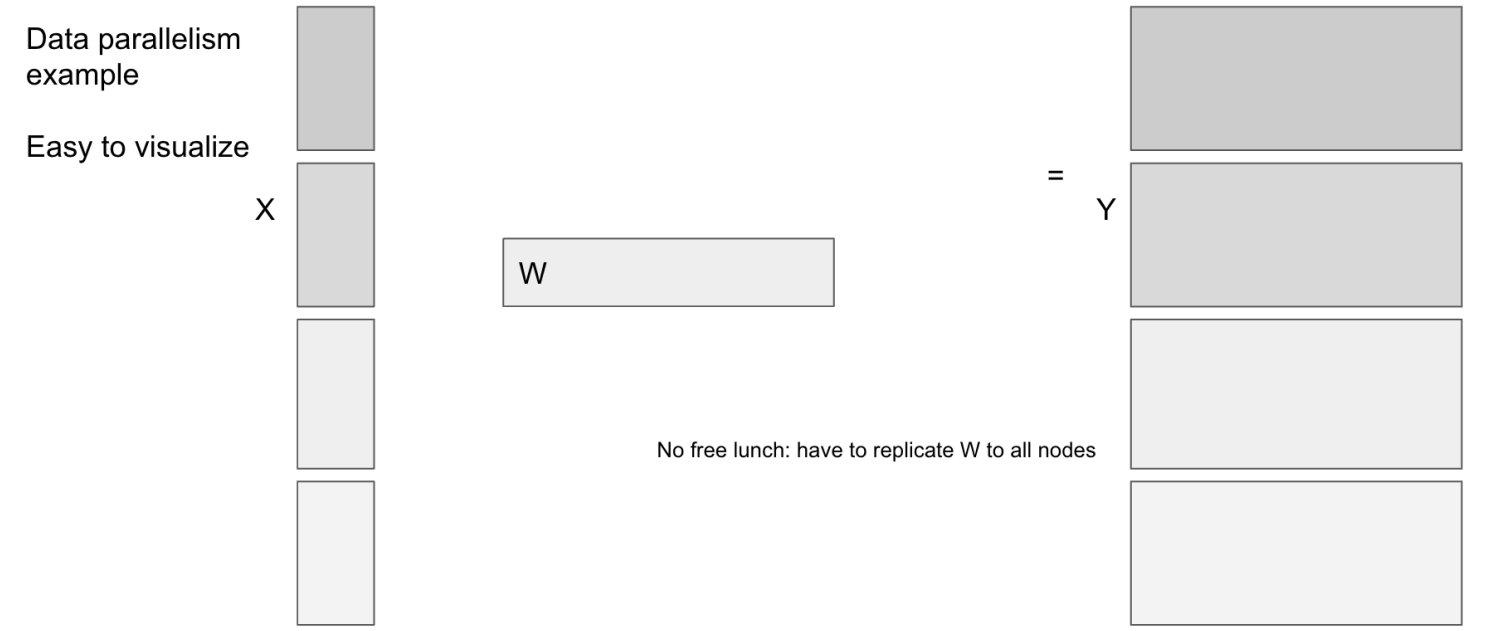
Data parallelism example

Easy to visualize



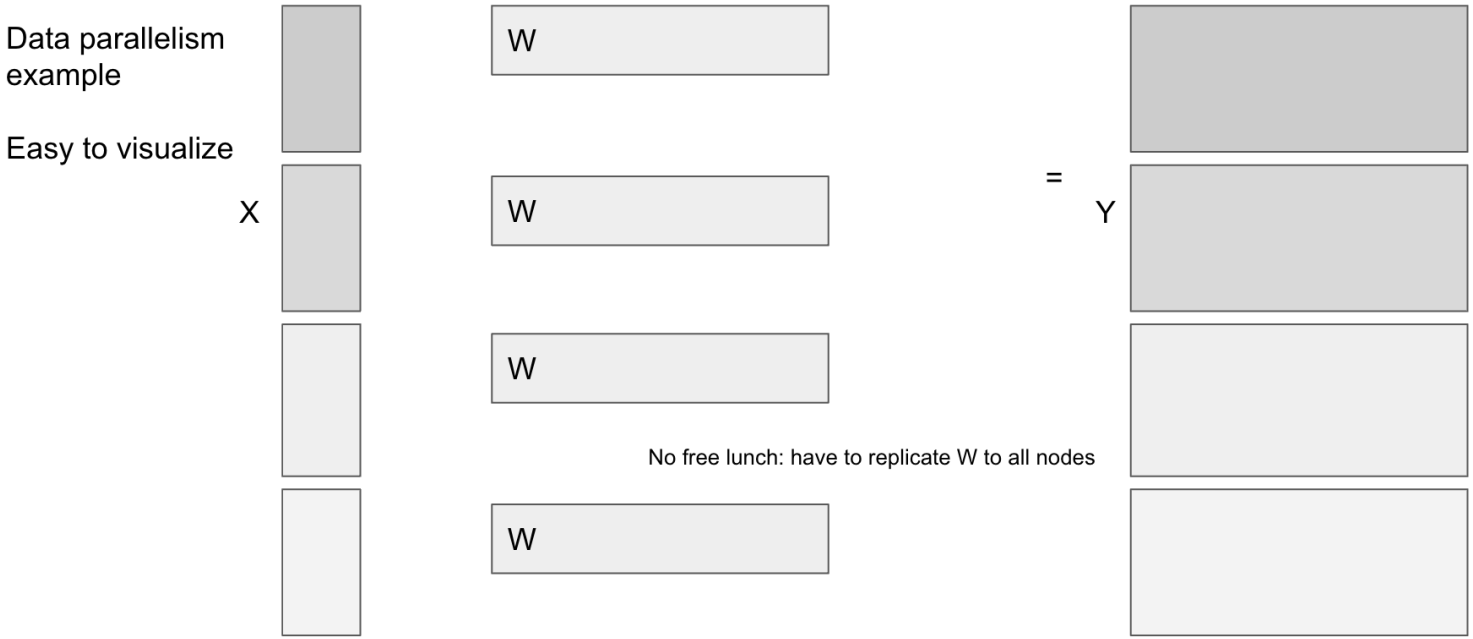
Data and Model Parallelism

Example:



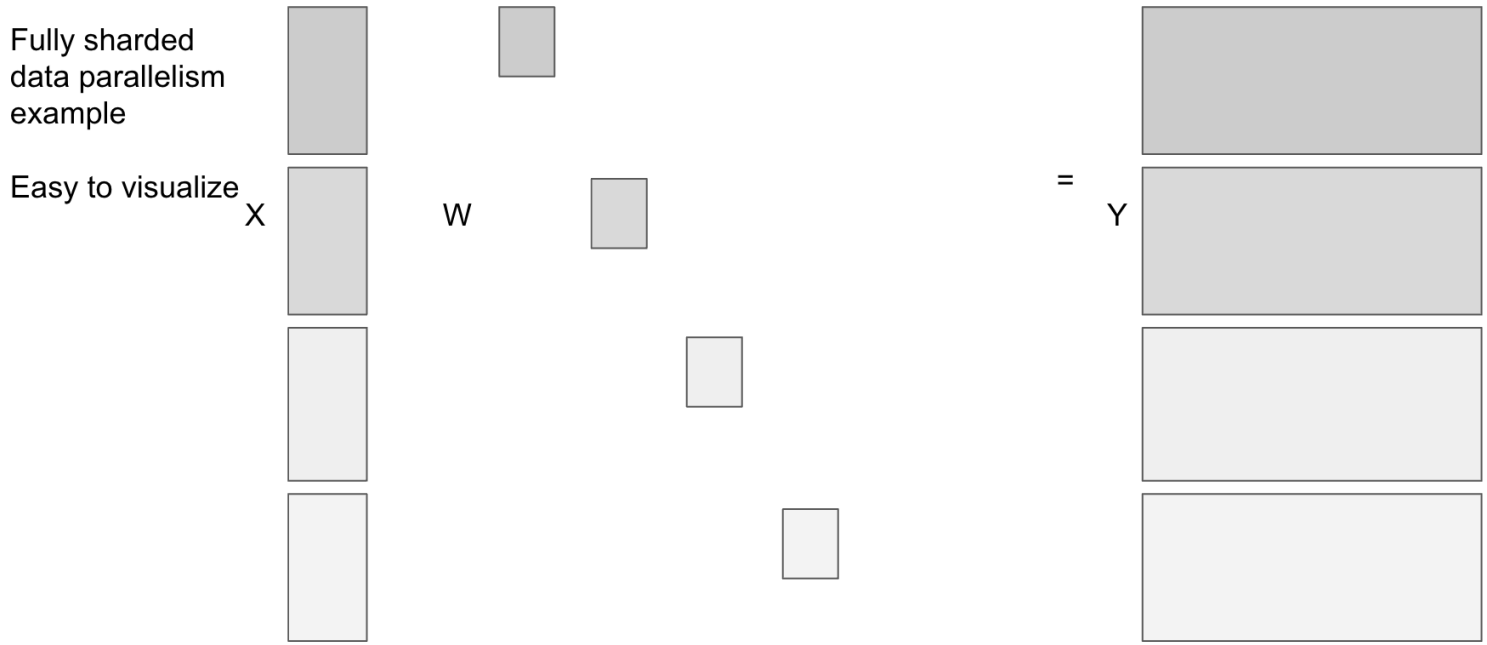
Data and Model Parallelism

Example:



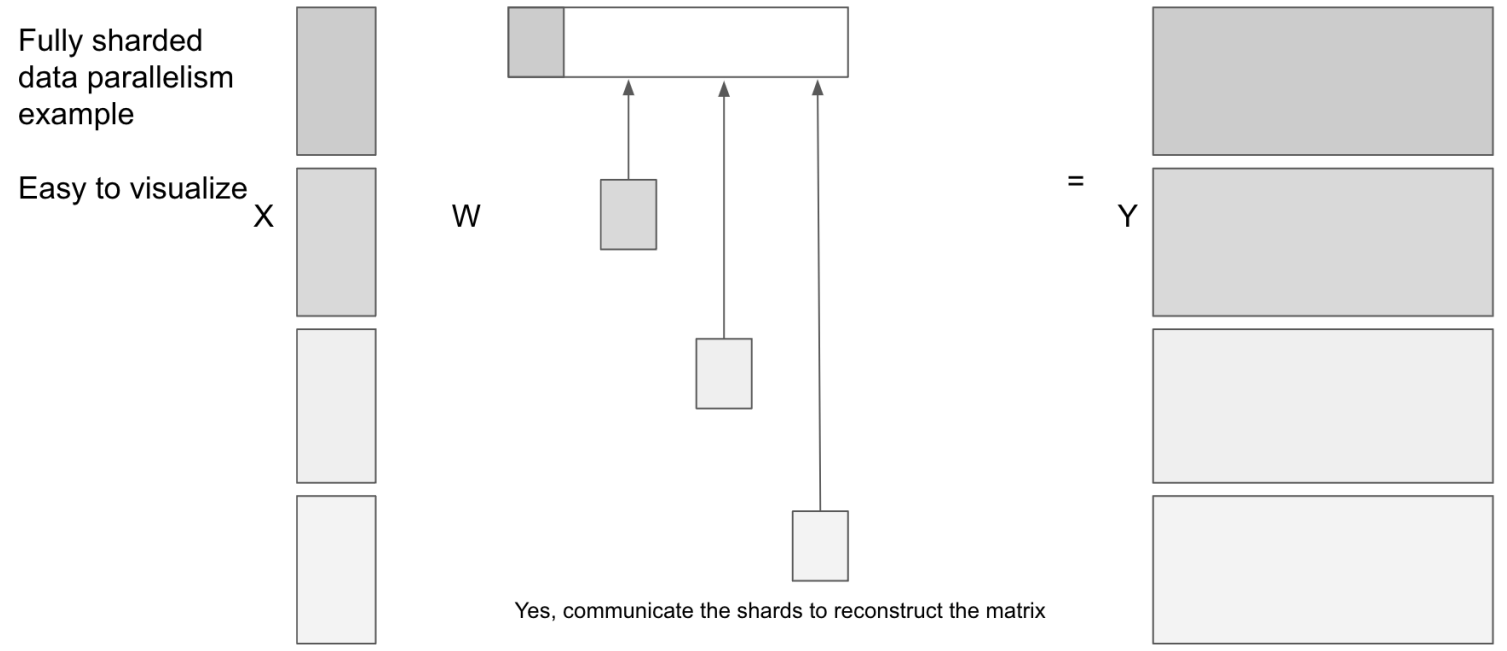
Data and Model Parallelism

Is it possible to reduce redundancy?



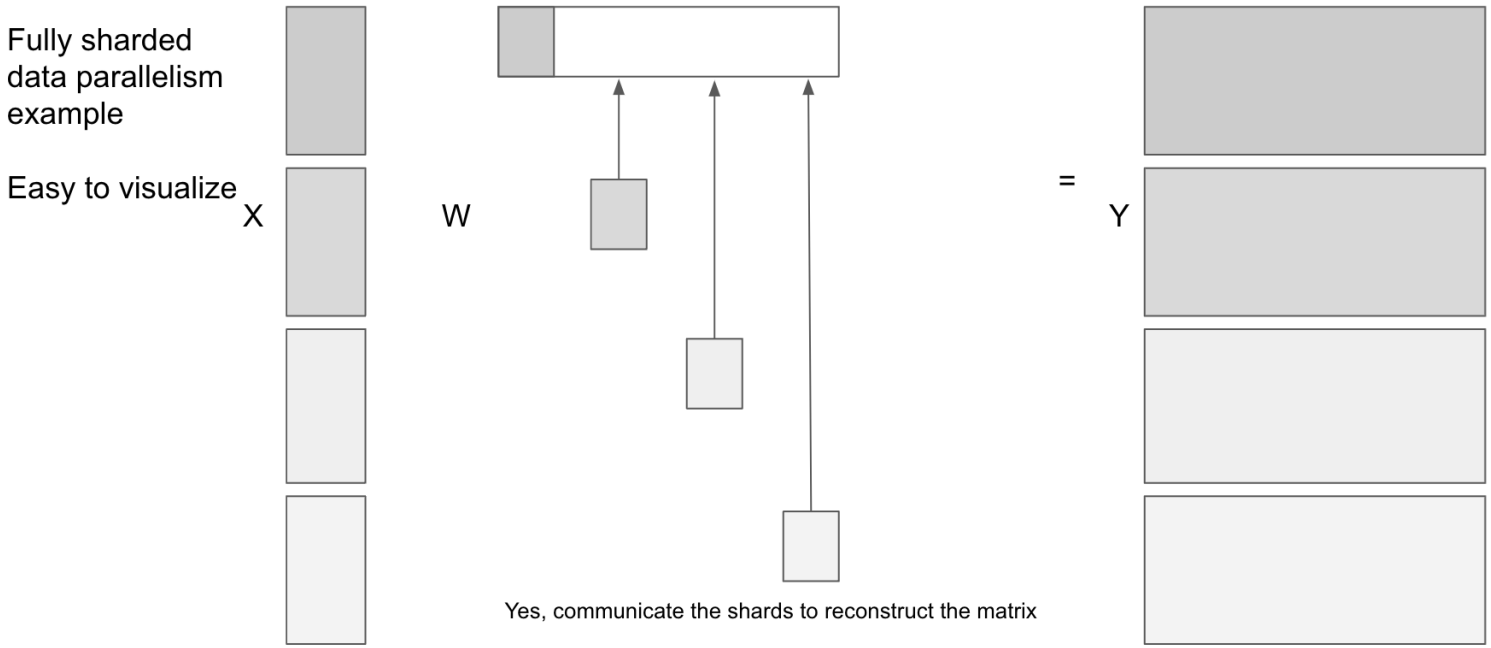
Data and Model Parallelism

We will now try to reconstruct our weight matrix before computing activations:



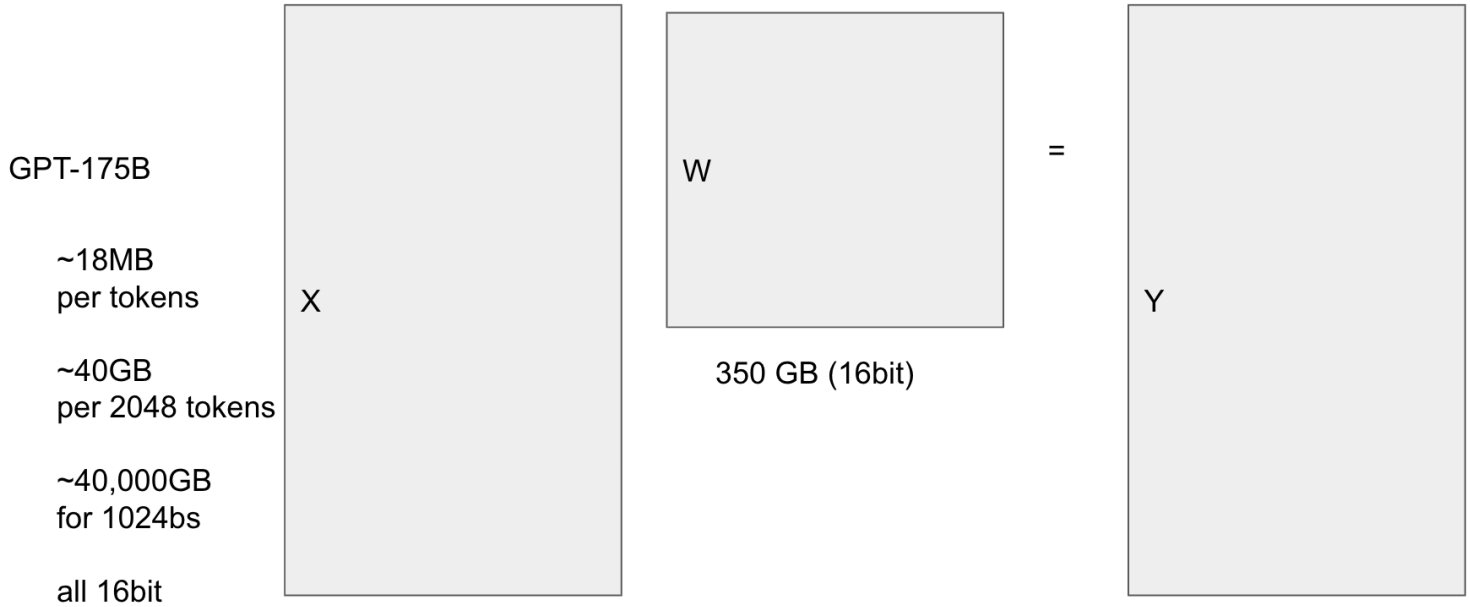
Data and Model Parallelism

Communication cost = parameters x 3. A trade-off between peak memory and communication



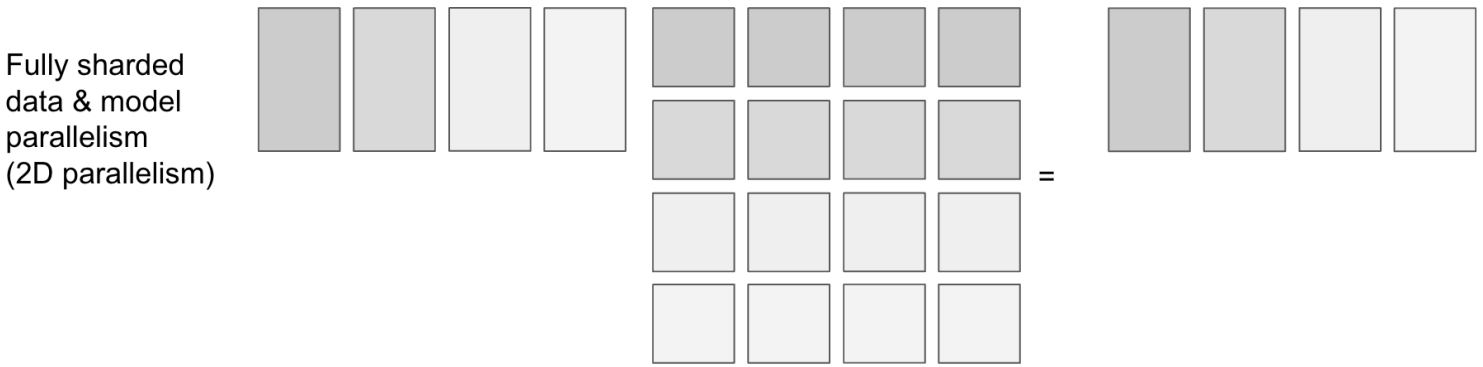
Data and Model Parallelism

Which strategy shall we use?



Data and Model Parallelism

Example:



After the break

After the break: **Other generative models**

Overview

- In generative modeling, we'd like to train a network that models a distribution, such as a distribution over images.
- One way to judge the quality of the model is to sample from it.
- This field has seen rapid progress:

2009

2015

2018

Overview

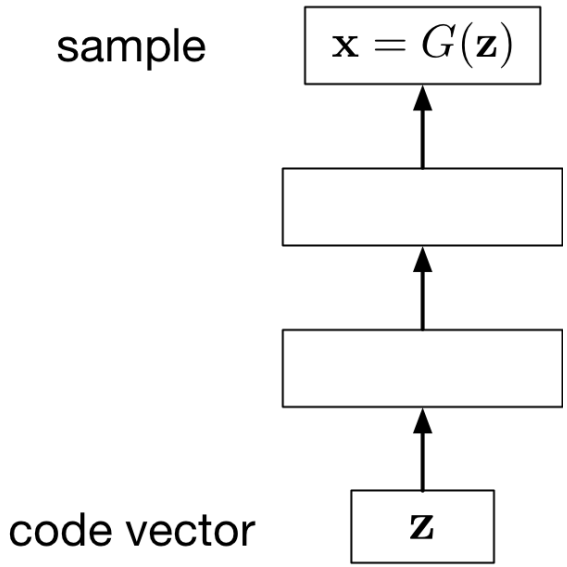
Four modern approaches to generative modeling:

- Autoregressive models (Lectures 3, 7, 8 and this lecture)
- Generative adversarial networks (this lecture)
- Diffusion models (next lecture)
- Variational autoencoders (next lecture)

All four approaches have different pros and cons.

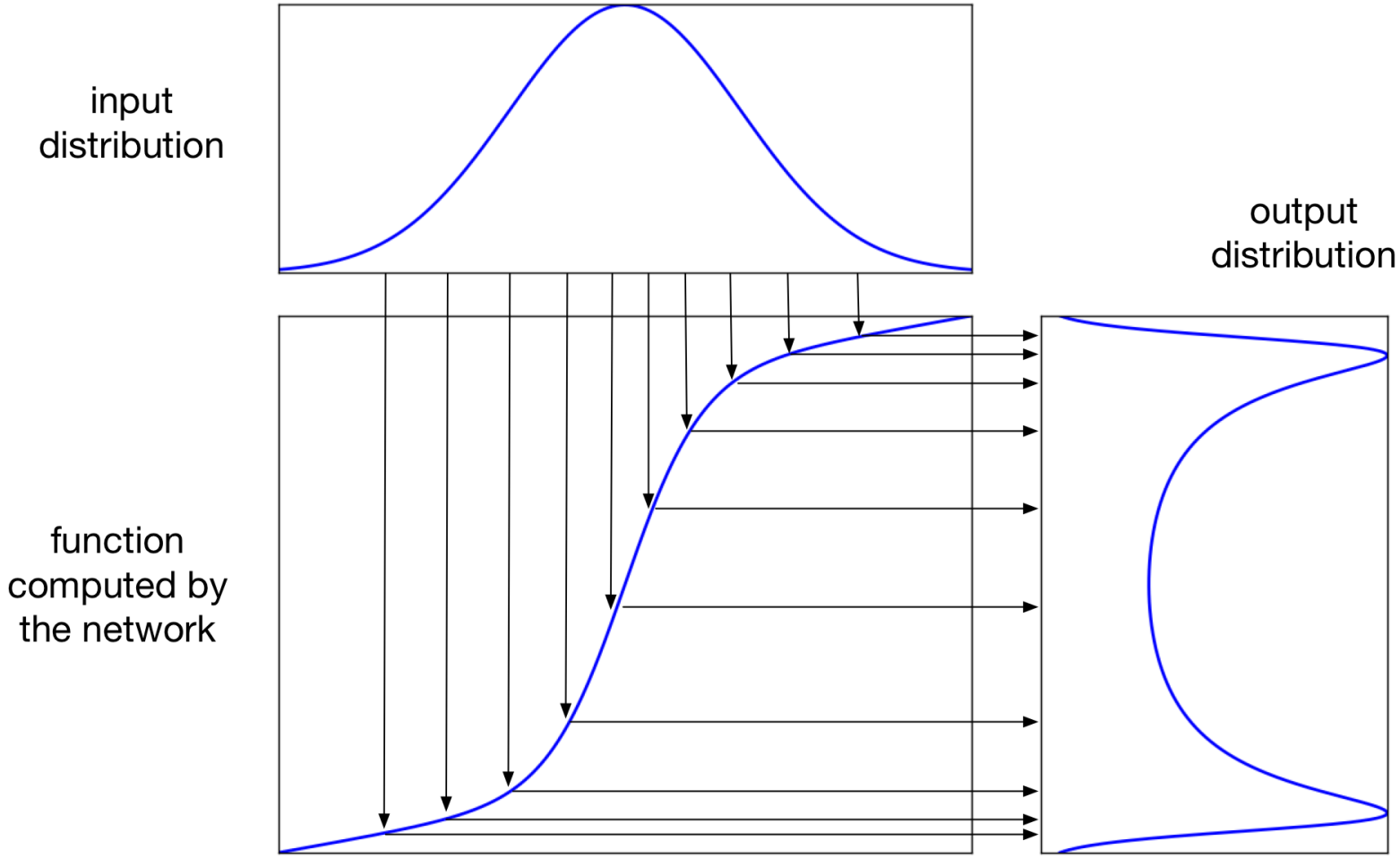
Generator Networks

- Autoregressive models explicitly predict a distribution at each step.
- Another approach to generative modeling is to train a neural net to produce approximate samples from the distribution.
- Start by sampling the code vector z from a fixed, simple distribution (e.g. spherical Gaussian)
- The generator network computes a differentiable function G mapping z to an x in data space

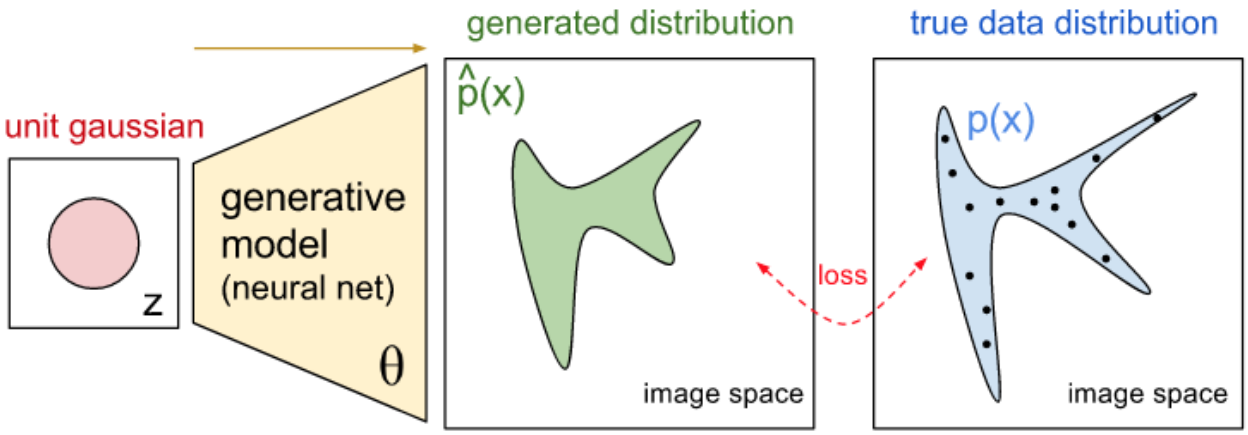


Generator Networks

A 1-dimensional example:

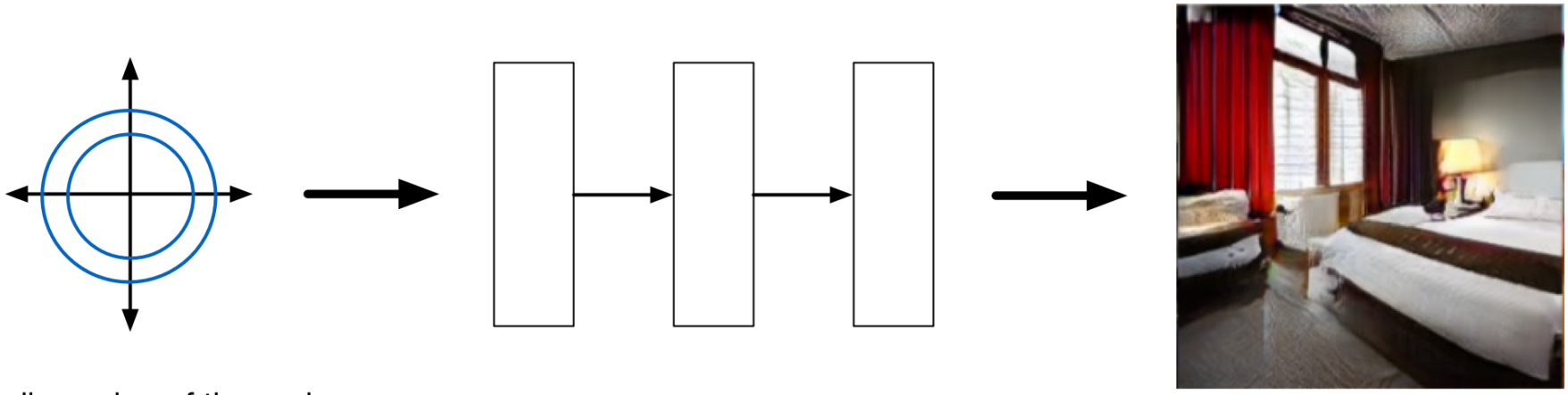


Generator Networks



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

Generator Networks



Each dimension of the code vector is sampled independently from a simple distribution, e.g. Gaussian or uniform.

This is fed to a (deterministic) generator network.

The network outputs an image.

This sort of architecture sounded preposterous to many of us, but amazingly, it works.

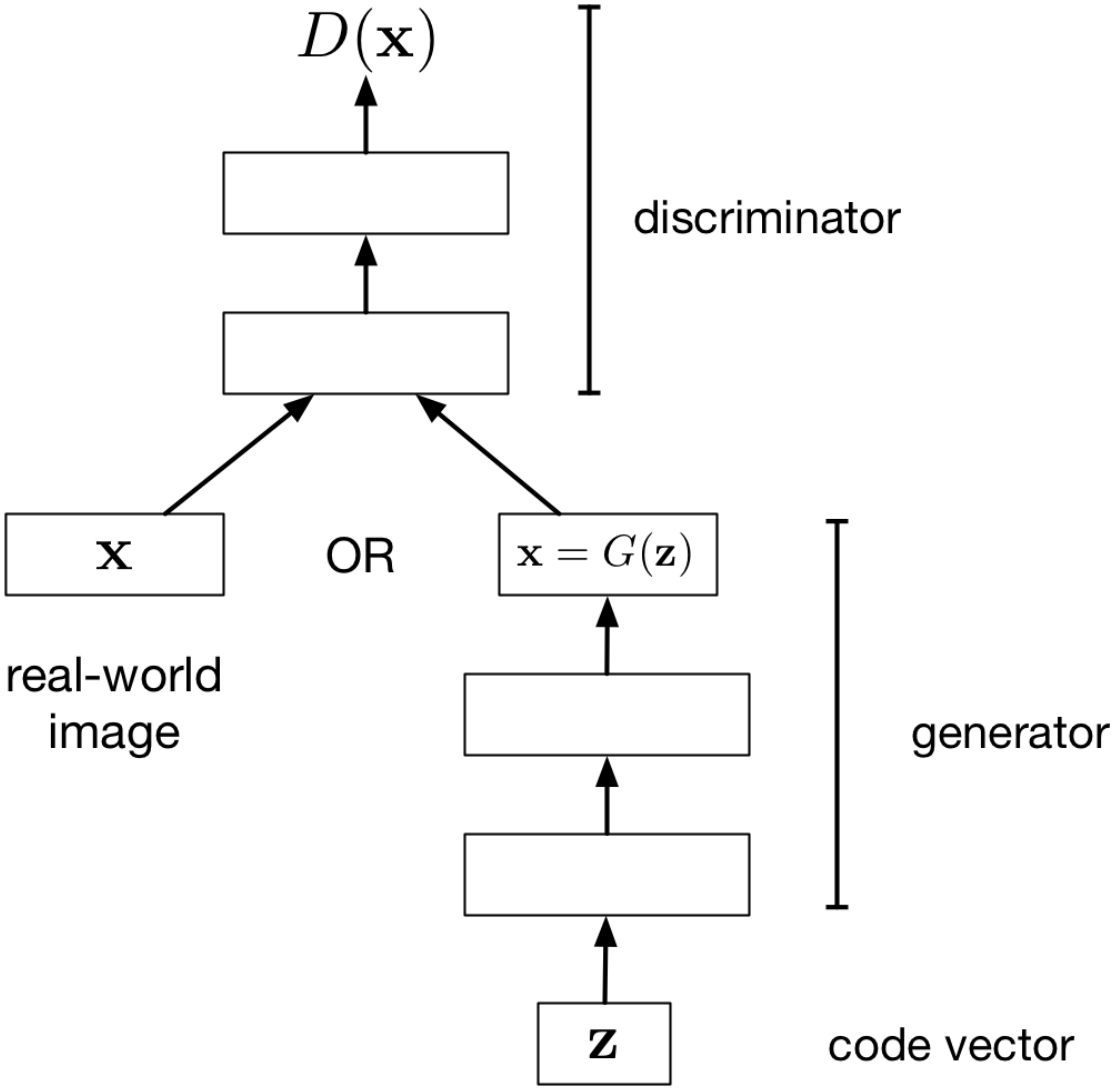
Generative Adversarial Networks

- Implicit generative models learn a mapping from random noise vectors to things that look like, e.g., images
- The advantage of implicit generative models: if you have some criterion for evaluating the quality of samples, then you can compute its gradient with respect to the network parameters, and update the network's parameters to make the sample a little better

Generative Adversarial Networks

- Implicit generative models learn a mapping from random noise vectors to things that look like, e.g., images
- The advantage of implicit generative models: if you have some criterion for evaluating the quality of samples, then you can compute its gradient with respect to the network parameters, and update the network's parameters to make the sample a little better
- The idea behind Generative Adversarial Networks (GANs): train two different networks
  - The generator network tries to produce realistic-looking samples
  - The discriminator network tries to figure out whether an image came from the training set or the generator network
- The generator network tries to fool the discriminator network

Generative Adversarial Networks



Generative Adversarial Networks

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

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

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

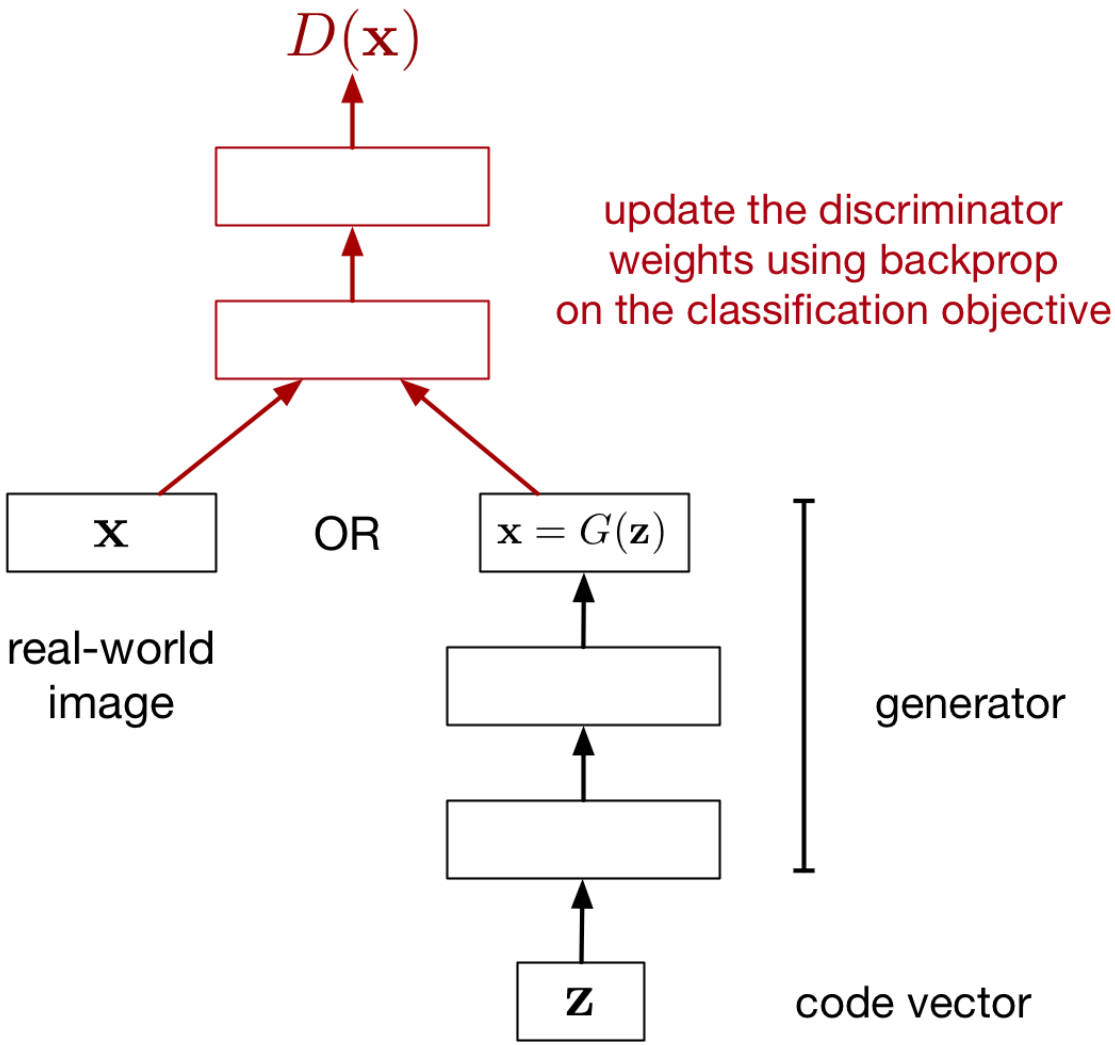
$$\begin{aligned}\mathcal{J}_G &= -\mathcal{J}_D \\ &= \text{const} + \mathbb{E}_{\mathsf{z}}[\log(1 - D(G(\mathsf{z})))]\end{aligned}$$

- This is called the minimax formulation, since the generator and discriminator are playing a zero-sum game against each other:

$$\max_G \min_D \mathcal{J}_D$$

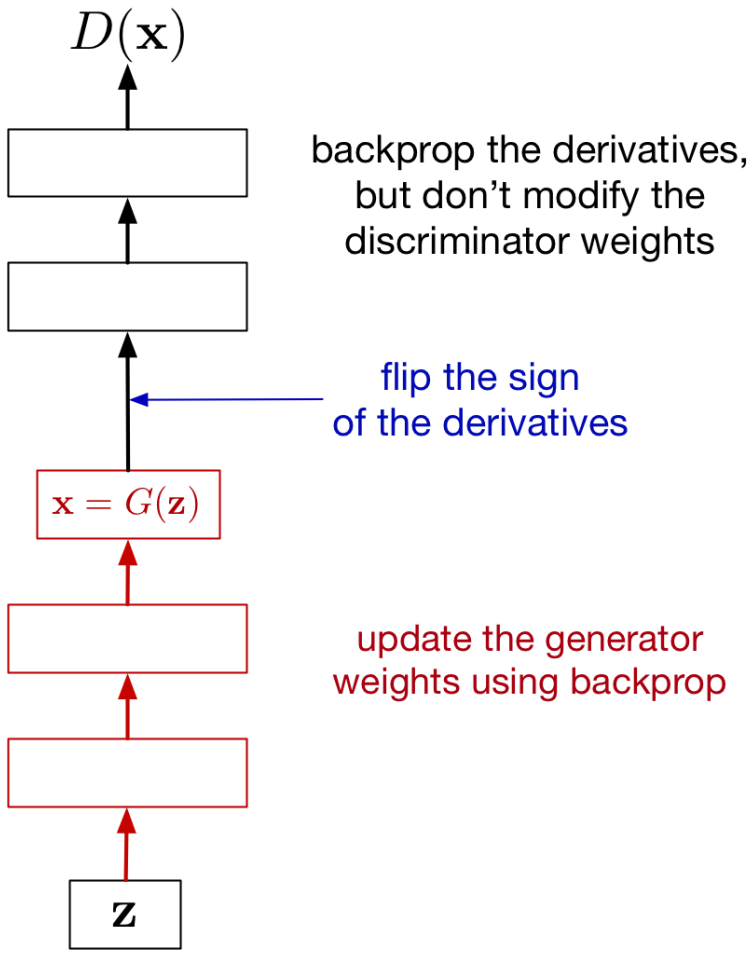
Generative Adversarial Networks

Updating the discriminator:



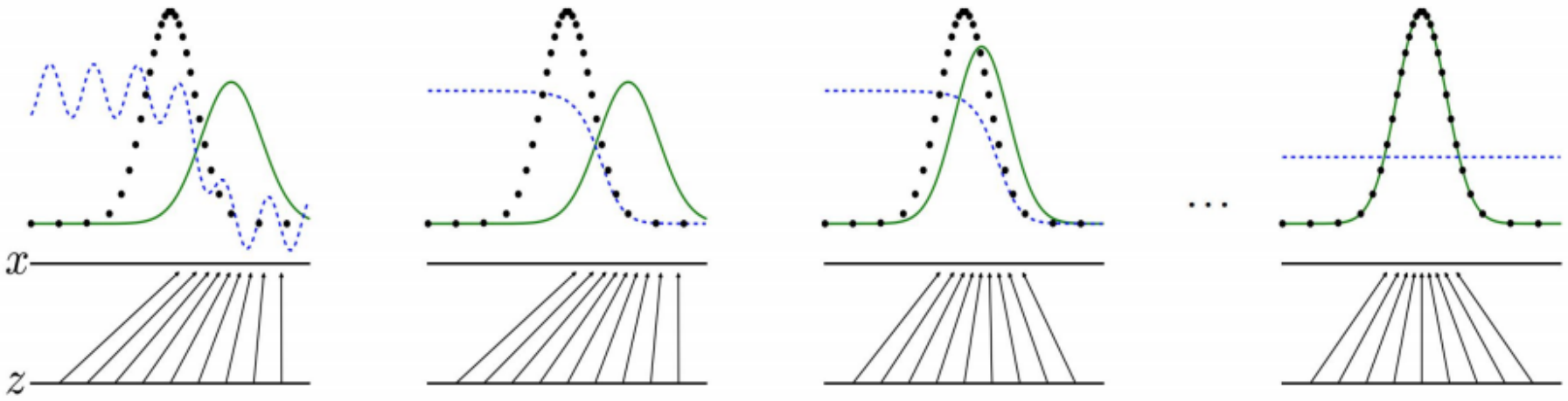
Generative Adversarial Networks

Updating the generator:



Generative Adversarial Networks

Alternating training of the generator and discriminator:



A Better Cost Function

- We introduced the minimax cost function for the generator:

$$\mathcal{J}_G = \mathbb{E}_z[\log(1 - D(G(z)))]$$

- One problem with this is saturation.
- Recall from our lecture on classification: when the prediction is really wrong,
  - "Logistic + squared error" gets a weak gradient signal
  - "Logistic + cross-entropy" gets a strong gradient signal
- Here, if the generated sample is really bad, the discriminator's prediction is close to 0, and the generator's cost is flat.

A Better Cost Function

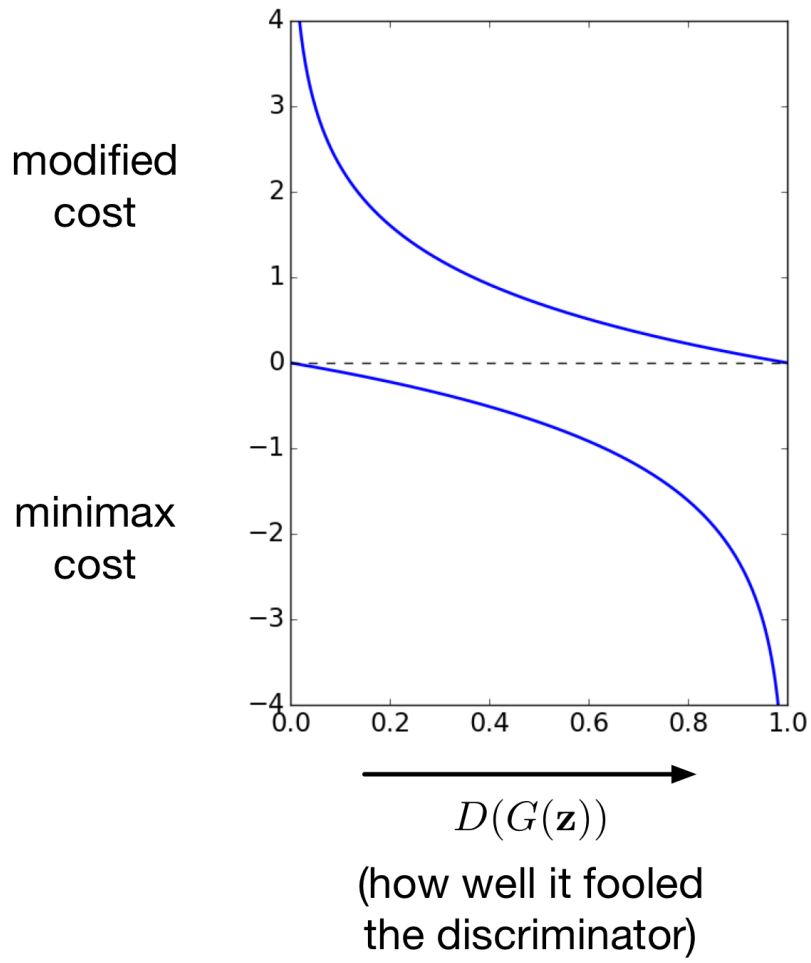
- Original minimax cost:

$$\mathcal{J}_G = \mathbb{E}_z[\log(1 - D(G(z)))]$$

- Modified generator cost:

$$\mathcal{J}_G = \mathbb{E}_z[-\log D(G(z))]$$

- This fixes the saturation problem.



Generative Adversarial Networks

- Since GANs were introduced in 2014, there have been hundreds of papers introducing various architectures and training methods.
- Most modern architectures are based on the Deep Convolutional GAN (DC-GAN), where the generator and discriminator are both conv nets.
- GAN Zoo: `https://github.com/hindupuravinash/the-gan-zoo`
  - Good source of horrible puns (VEEGAN, Checkhov GAN, etc.)

GAN Samples

Celebrities:

Karras et al., 2017. Progressive growing of GANs for improved quality, stability, and variation

GAN Samples

Bedrooms:

Karras et al., 2017. Progressive growing of GANs for improved quality, stability, and variation

GAN Samples

ImageNet object categories (by BigGAN, a much larger model with a bunch more engineering tricks):

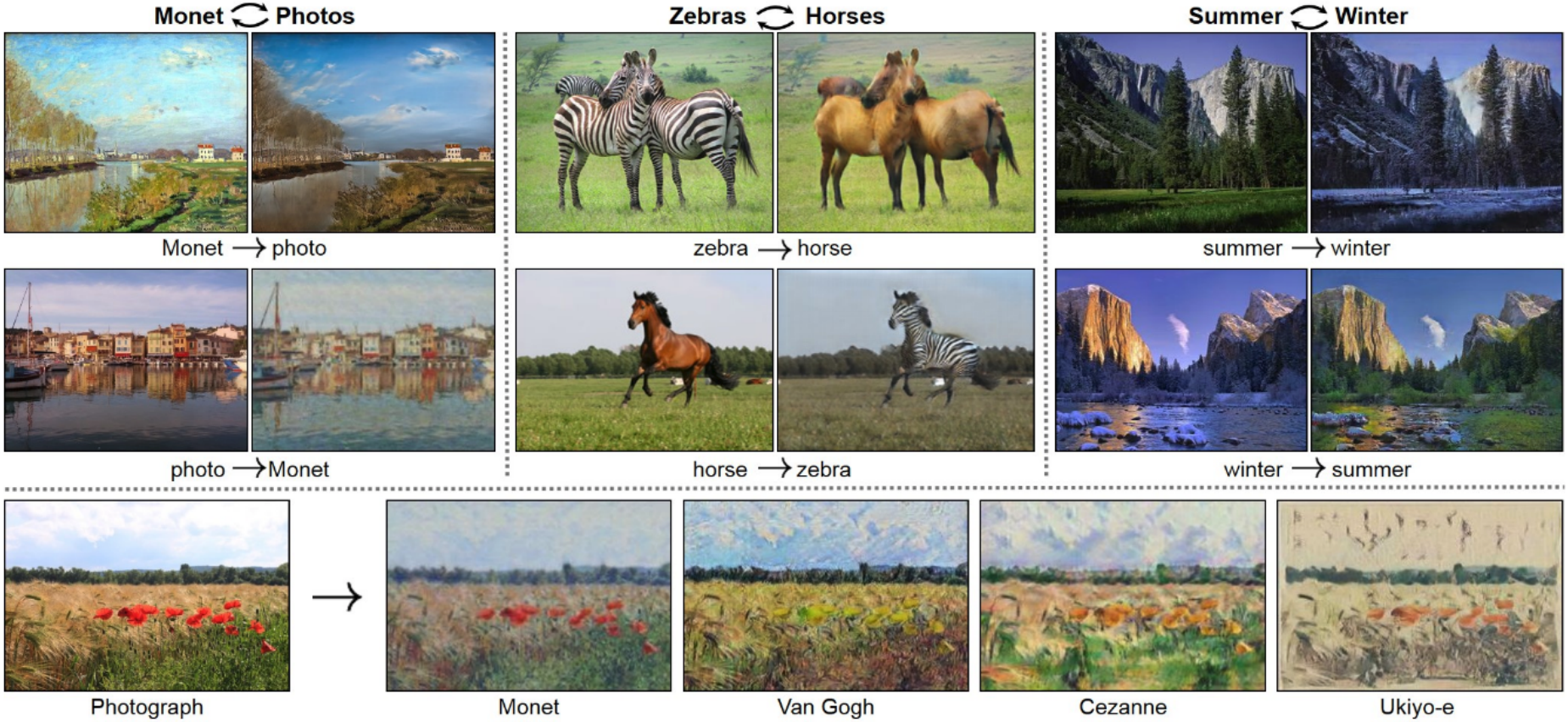
Brock et al., 2019. Large scale GAN training for high fidelity natural image synthesis.

GAN Samples

- GANs revolutionized generative modeling by producing crisp, high-resolution images.
- The catch: we don't know how well they're modeling the distribution.
  - Can't measure the log-likelihood they assign to held-out data.
  - Could they be memorizing training examples? (E.g., maybe they sometimes produce photos of real celebrities?)
  - We have no way to tell if they are dropping important modes from the distribution.
  - See Wu et al., "On the quantitative analysis of decoder-based generative models" for partial answers to these questions.

CycleGAN

Style transfer problem: change the style of an image while preserving the content.



Data: Two unrelated collections of images, one for each style

CycleGAN

- If we had paired data (same content in both styles), this would be a supervised learning problem. But this is hard to find.
- The CycleGAN architecture learns to do it from unpaired data.
  - Train two different generator nets to go from style 1 to style 2, and vice versa.
  - Make sure the generated samples of style 2 are indistinguishable from real images by a discriminator net.
  - Make sure the generators are cycle-consistent: mapping from style 1 to style 2 and back again should give you almost the original image.

CycleGAN

Real zebra image

The discriminator tries to distinguish generated zebra images from real ones

D

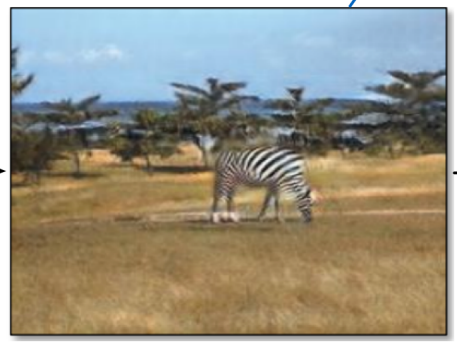
Discriminator loss: GAN generator objective, i.e. negative log probability D assigns to the sample being real

Reconstruction loss: squared error between the original image and the reconstruction

Input image (real horse image)

G_1

Generator 1 learns to map from horse images to zebra images while preserving the structure



Generated sample

G_2

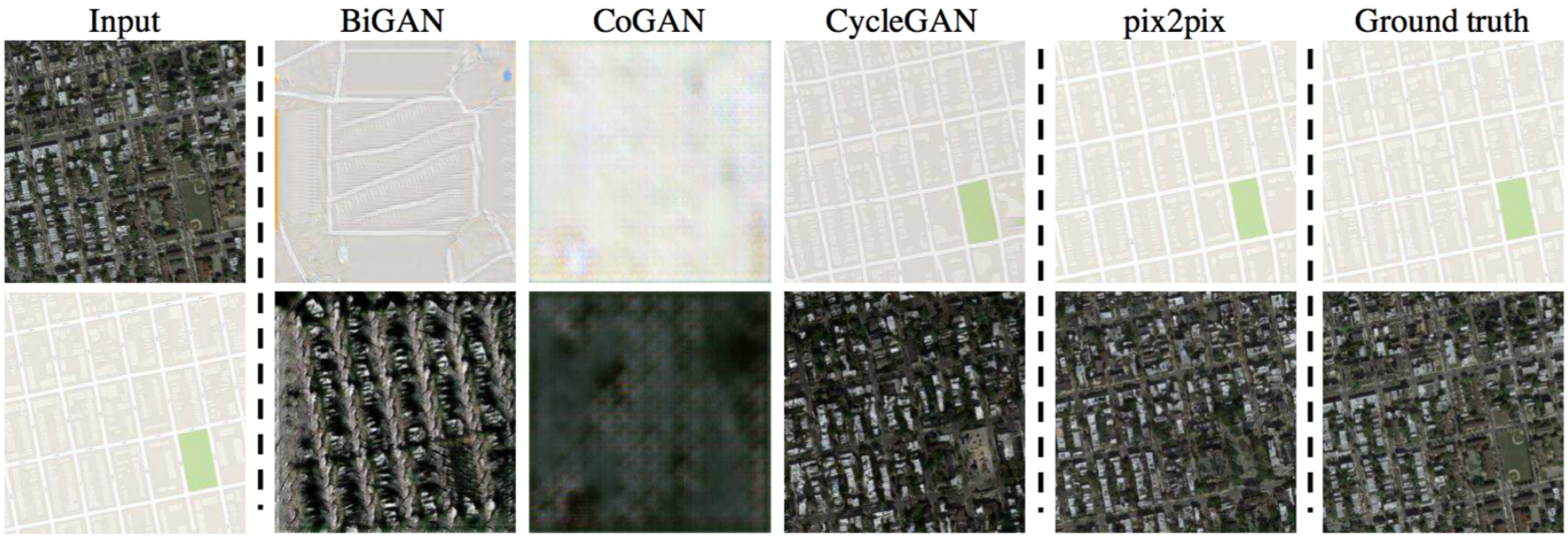
Generator 2 learns to map from zebra images to horse images while preserving the structure

Reconstruction

Total loss = discriminator loss + reconstruction loss

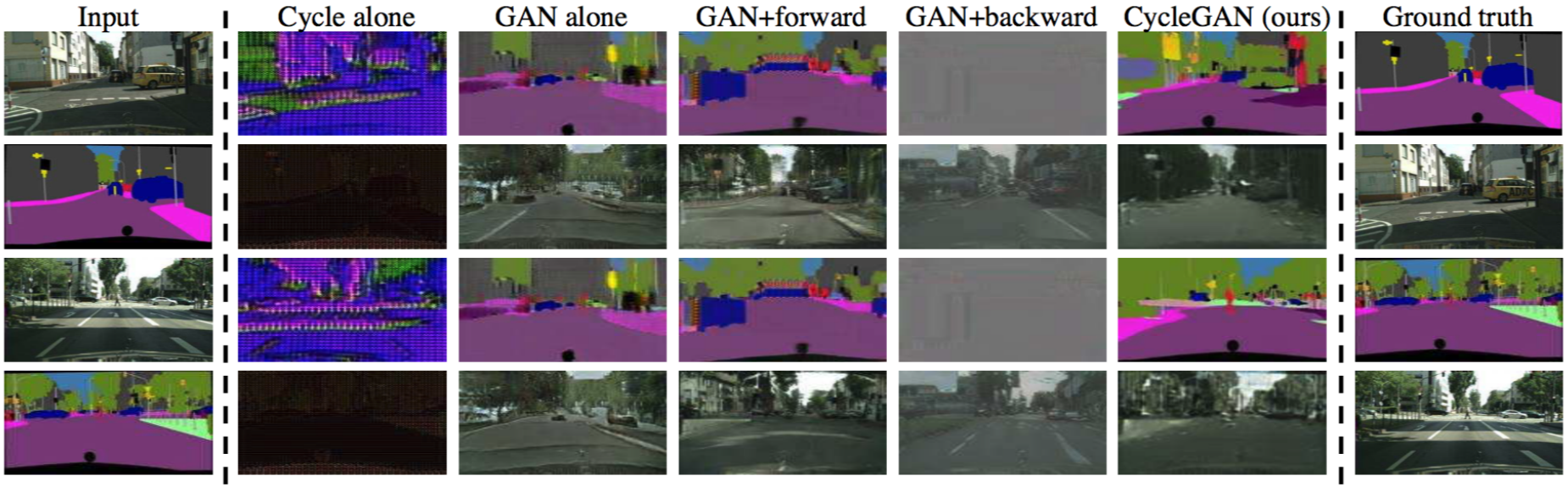
CycleGAN

Style transfer between aerial photos and maps:



CycleGAN

Style transfer between road scenes and semantic segmentations (labels of every pixel in an image by object category):



Overview

Four modern approaches to generative modeling:

- Autoregressive models (Lectures 3, 7, and 8)
- Generative adversarial networks (this lecture)
- Diffusion models (next lecture)
- Variational autoencoders (next lecture)

All four approaches have different pros and cons.

Trade-offs of Generative Approaches

- So far, we have seen four different approaches:
  - Autoregressive models (Lectures 3, 7, and 8)
  - Generative adversarial networks (this lecture)
  - Diffusion models (next lecture)
  - Variational autoencoders (next lecture)
- They all have their own pro and con. We often pick a method based on our application needs.
- Some considerations for computer vision applications:
  - Do we need to evaluate log likelihood of new data?
  - Do we prefer good samples over evaluation metric?
  - How imporant is representation learning, i.e. meaningful code vectors?
  - How much computational resource can we spent?

Trade-offs of Generative Approaches

- In summary:

	Log-likelihood	Sample	Representation	Computation
Autoregressive				
GANs				
Diffusion				
VAEs				

- To be continued...