

Introduction

January 2nd, 2025

Deep Learning (CS60010)

**Slides adapted from <http://udlbook.com>*

Logistics

Course Webpage: <http://cse.iitkgp.ac.in/~pawang/courses/DL25.html>

Lectures:

Wednesday - 11:00 - 12:00 (NR112)

Thursday - 12:00 - 13:00 (NR112)

Friday - 8:00 - 9:00 (NR112)

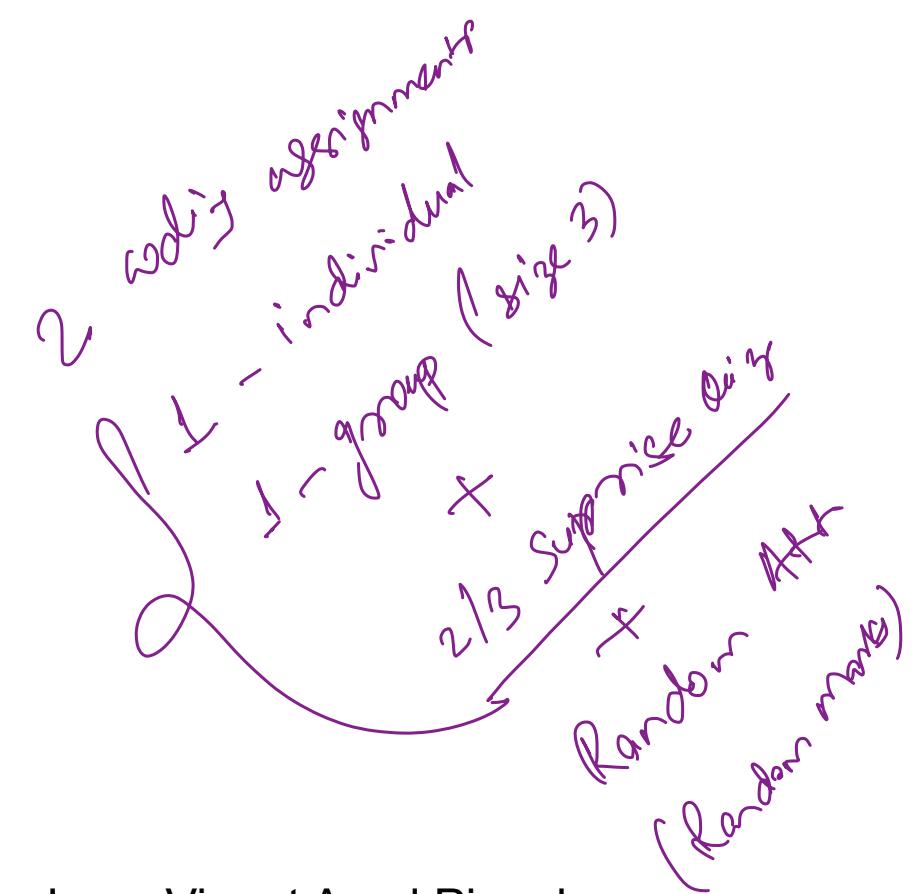
Teaching Assistants: Souryadip Ray, Pretam Ray, Rounak Saha, Shreyas Jena, Vineet Amol Pippal

Reference Books:

✓ Simon J.D. Prince. Understanding Deep Learning. MIT Press. (2023).

• Ian J. Goodfellow, Yoshua Bengio, and Aaron Courville. Deep learning. MIT Press. (2016). ✓

Evaluation Plan (Tentative): Mid-Sem (25%), End-Sem (40%), Coding Assignments and other class activities (35%)



Other Logistics

Please ask questions in the lectures



You can ask questions online (Microsoft Teams)

Get familiar with Pytorch (There will be some coding tutorials, but lot of resources online)

Use Google Colab / Jupyter Notebook

<https://udlbook.github.io/udlbook/>

A very nice resource

The Course Book

Harvard CS197

AI Research Experiences

GPT-3 · Large Language Models · VSCode · Git · Conda · Debugging · Linting ·
Reading AI papers · Literature Search · Hugging Face · Lightning · Vision
Transformer · PyTorch · Autograd · Experiment Organization · Weights and
Biases · Hyperparameter Search · Sweeps · Hydra · Research Ideas · Paper
Writing · AWS · GPU Training · Stable Diffusion · Colab · Accelerate · Gradio ·
Project Organization · Team Communication · Research Progress · Assertion-
Evidence · Slide Design · Statistical Testing ·

Pranav Rajpurkar PhD

Assistant Professor, Harvard University

Take your AI skills to the next level with this course.

https://docs.google.com/document/d/1uvAbEhbgs_M-uDMTzmOWRIYxqCkogKRXdKYYT98ooc/mobilebasic#h.2z3yllpny6or

< All teams

Other Logistics

join using the team code:

5uev23l

DL

Deep Learning Spring 2025

...

Home page

Class Notebook

Classwork

Assignments

Grades

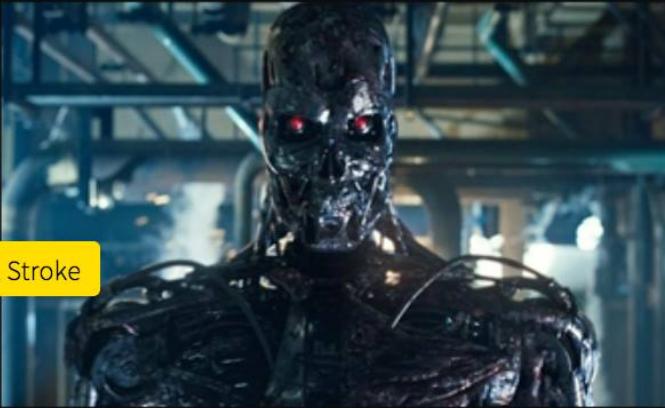
Reflect

Insights

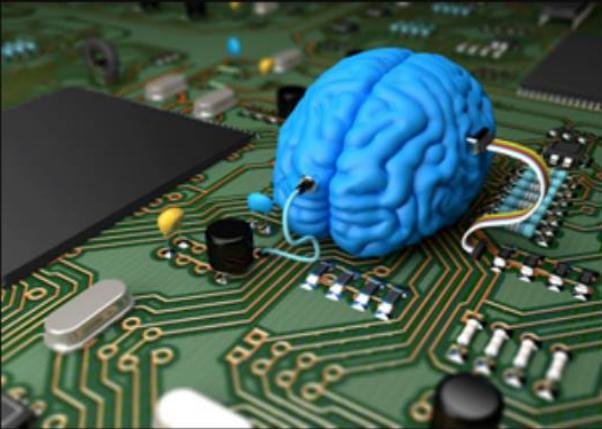
▼ Main Channels

General

Deep Learning



What society thinks I do



What my friends think I do



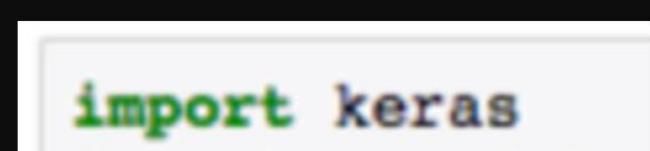
What other computer
scientists think I do



What mathematicians think I do



What I think I do



What I actually do

Artificial intelligence

- Building systems that simulate intelligent behavior
- Wide range of approaches: logic, search, probabilistic reasoning, ...

Artificial intelligence

- Building systems that simulate intelligent behavior
- Wide range of approaches: logic, search, probabilistic reasoning, ...

Machine learning

- A subset of AI that learns to make decisions by fitting mathematical models to observed data
- (incorrectly) almost synonymous with AI

Artificial intelligence

Image class
(10 animals)

GPT
Semi-supervised
learning

Riding a bicycle

Machine learning

Supervised learning ✓

Define a mapping from input data to an output prediction

Unsupervised learning ✓

Constructing a model from input data without corresponding output labels

Reinforcement learning ✓

Introduces the idea of an agent that lives in the world and learns to choose actions leading to high reward

Artificial intelligence

Machine learning

Supervised
learning

Unsupervised
learning

Reinforcement
learning

Deep learning

Deep Neural Networks contribute to each of the areas

Supervised learning

- Define a mapping from input to output
- Learn this mapping from paired input/output data examples

Regression

Real world input

6000 square feet,
4 bedrooms,
previously sold for
\$235K in 2005,
1 parking spot.

Model
input

$$\begin{bmatrix} 6000 \\ 4 \\ 235 \\ 2005 \\ 1 \end{bmatrix}$$

Model



Model
output

$$\begin{bmatrix} 340 \end{bmatrix}$$

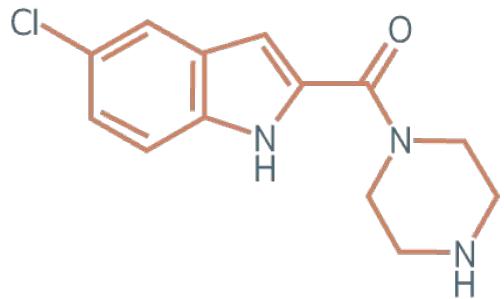
Real world output

Predicted price
is \$340k

- Univariate regression problem (one output, real value)
- Fully connected network

Graph regression

Real world input



Model
input

$$\begin{bmatrix} 1 \\ 0 \\ 1 \\ \vdots \\ 17 \\ 1 \\ 1 \\ \vdots \end{bmatrix}$$

Model



Model
output

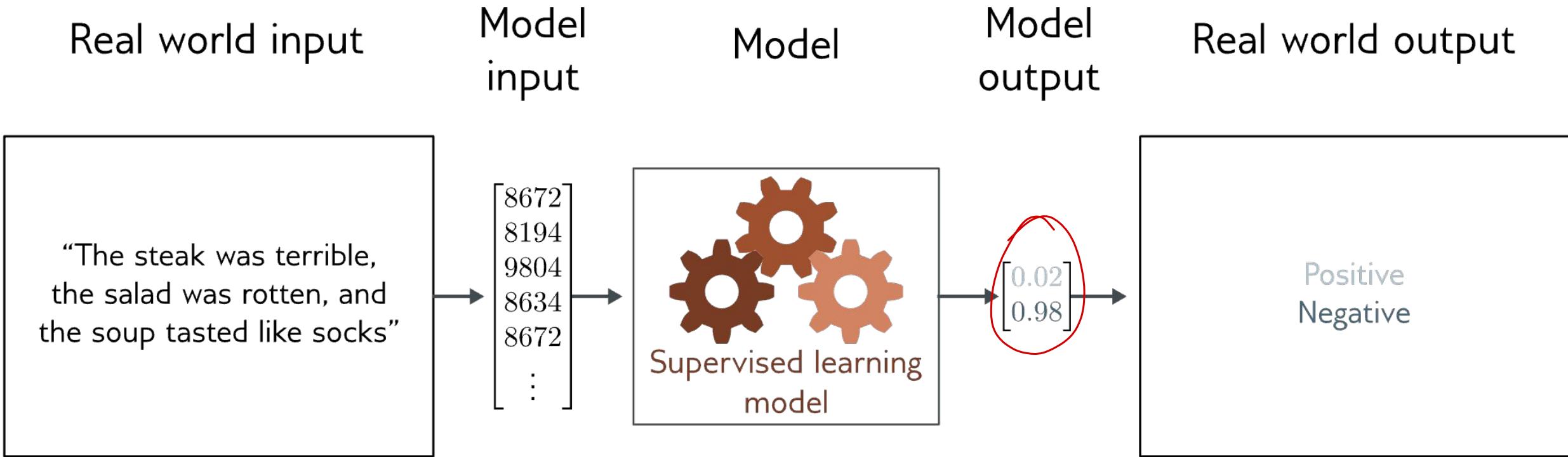
$$\begin{bmatrix} -12.9 \\ 56.4 \end{bmatrix}$$

Real world output

Freezing point
is -12.9°C
Boiling point
is 56.4°C

- Multivariate regression problem (>1 output, real value)
- Graph neural network

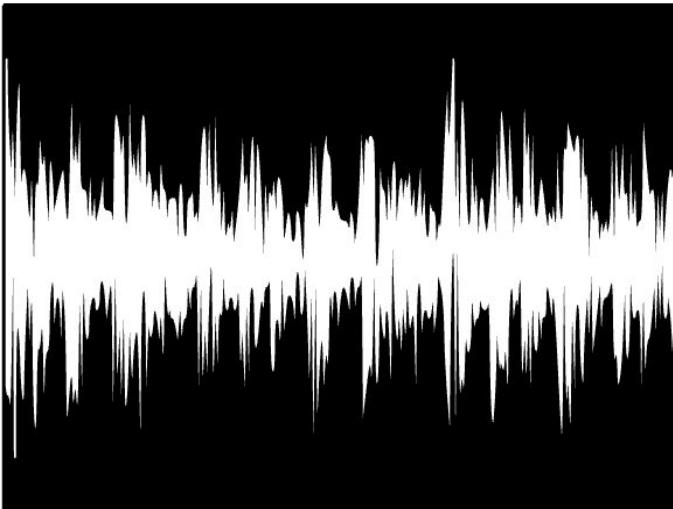
Text classification



- Binary classification problem (two discrete classes)
- Transformer network

Music genre classification

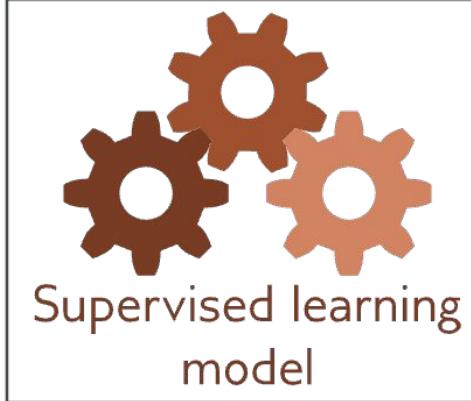
Real world input



Model
input

$$\begin{bmatrix} 125 \\ 12054 \\ 1253 \\ 6178 \\ 24 \\ 4447 \\ \vdots \end{bmatrix}$$

Model



Model
output

$$\begin{bmatrix} 0.03 \\ 0.52 \\ 0.18 \\ 0.07 \\ 0.12 \\ 0.08 \\ \vdots \\ 0.01 \end{bmatrix}$$

Real world output

Classical
Electronica
Hip Hop
Jazz
Pop
Metal
Punk

- Multiclass classification problem (discrete classes, >2 possible values)
- Recurrent neural network (RNN)

Image classification

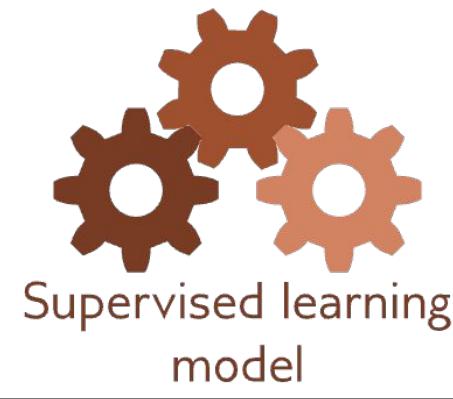
Real world input



Model
input

$$\begin{bmatrix} 124 \\ 140 \\ 156 \\ 128 \\ 142 \\ 157 \\ \vdots \end{bmatrix}$$

Model



Model
output

$$\begin{bmatrix} 0.00 \\ 0.00 \\ 0.01 \\ 0.89 \\ 0.05 \\ 0.00 \\ \vdots \\ 0.01 \end{bmatrix}$$

Real world output

Aardvark
Apple
Bee
Bicycle
Bridge
Clown
⋮

- Multiclass classification problem (discrete classes, >2 possible classes)
- Convolutional network

Terms

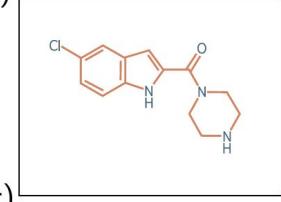
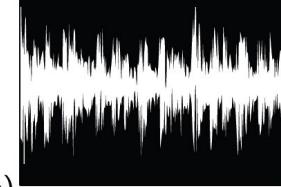
- Regression = continuous numbers as output
- Classification = discrete classes as output
- Two class and multiclass classification treated differently
- Univariate = one output
- Multivariate = more than one output

What do we see in these examples?

- There is a ***meaningful real-world input*** (Sentence, sound file, image, etc.)
- The input is encoded as a vector of numbers, forming the model input
- The model maps the input to an output vector
- The output vector is translated back to a ***meaningful real-word prediction.***

Structure in input

- a) Real world input

- b)

- c)
"The steak was terrible,
the salad was rotten, and
the soup tasted like socks"
- d)

- e)


Fixed-length vector containing values that describe the property. Example of a tabular data (Key-values)

May contain varying numbers of atoms that can be connected in different ways. The structure as well as the constituent atoms are needed

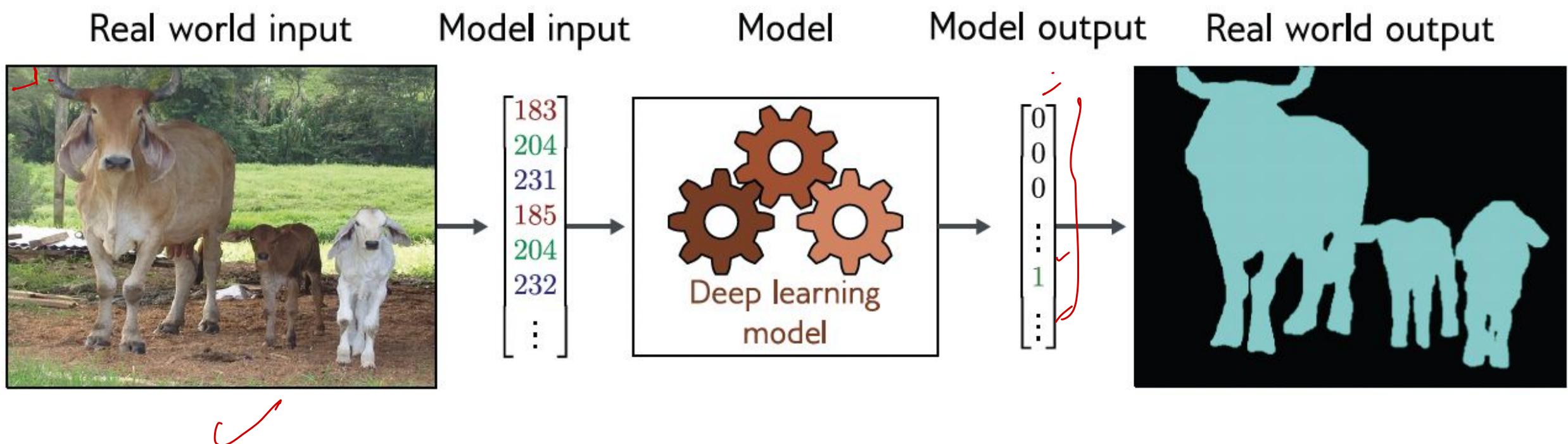
Body of text (variable length depending on number of words). One can use a fixed vocabulary and concatenate the word indices

Input may be of fixed size, but is very high-dimensional. Sampling rate: 44.1 KHz → 10 sec clip would be 441,000 integers

Two-dimensional structure. Input may be concatenated RGB values at each pixel

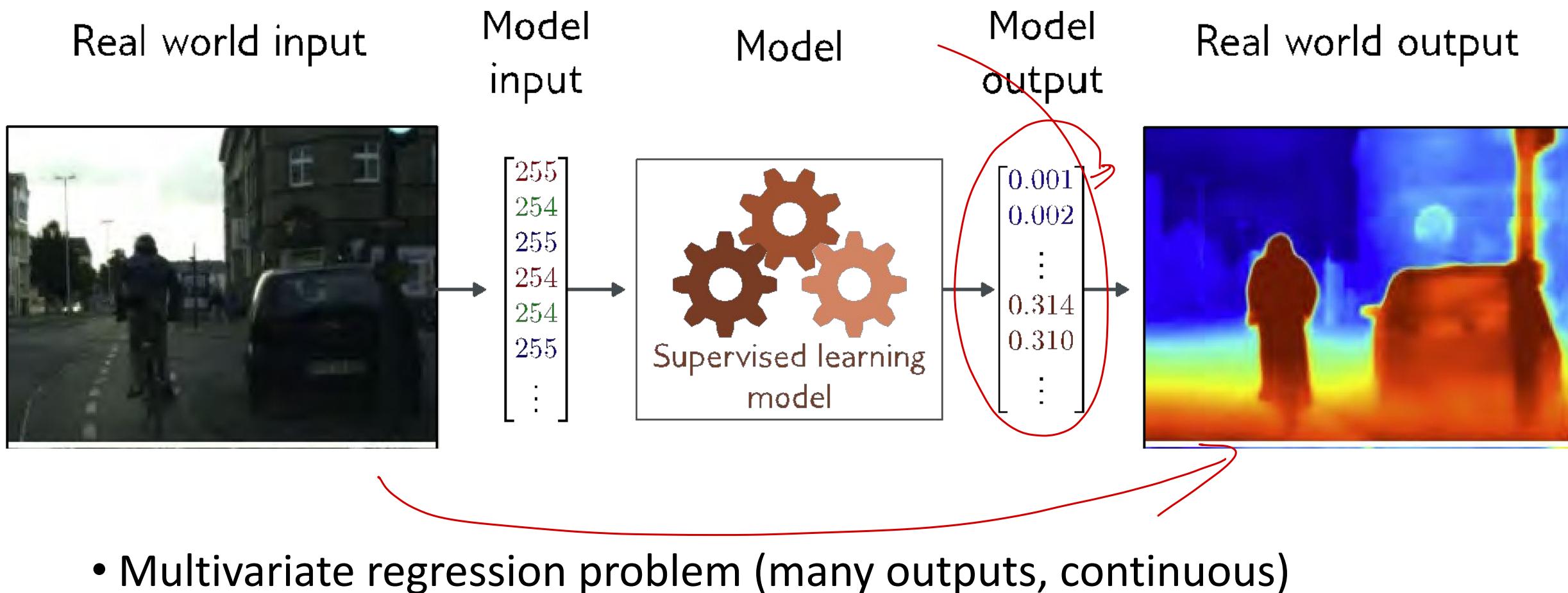
What about structure in output?

Image segmentation



- Multivariate binary classification problem (many outputs, two discrete classes)

Depth estimation



Translation

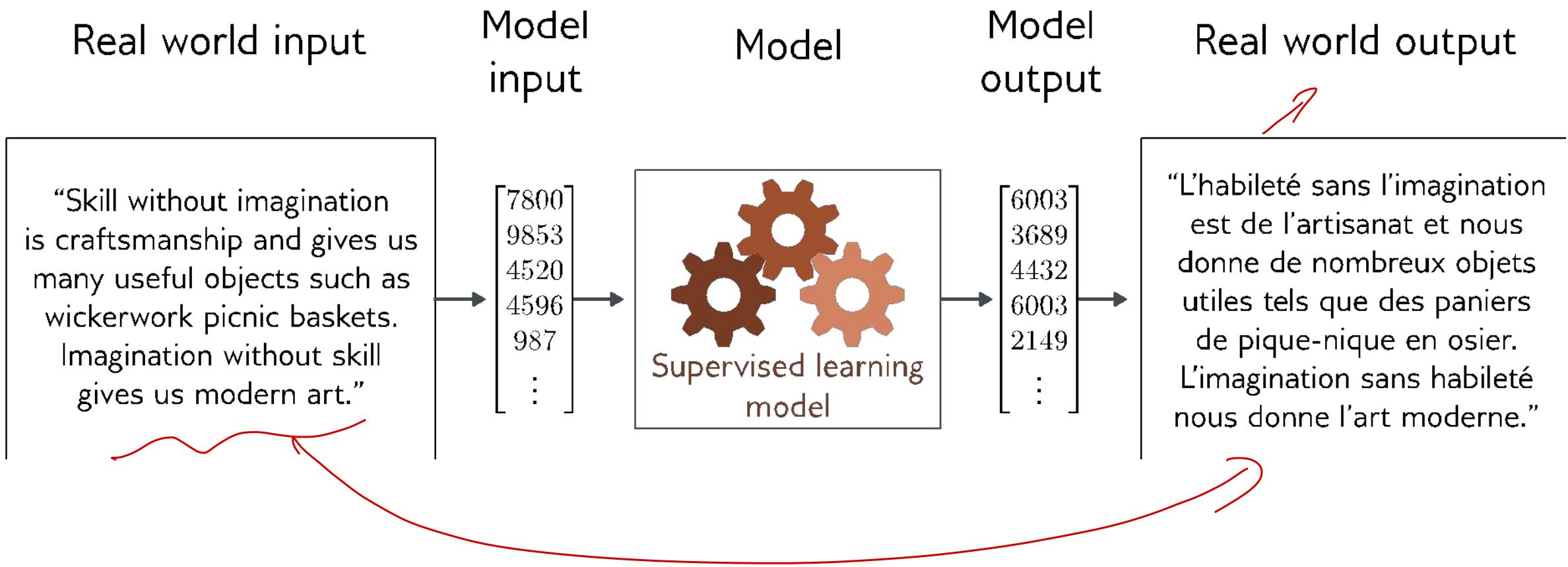


Image captioning

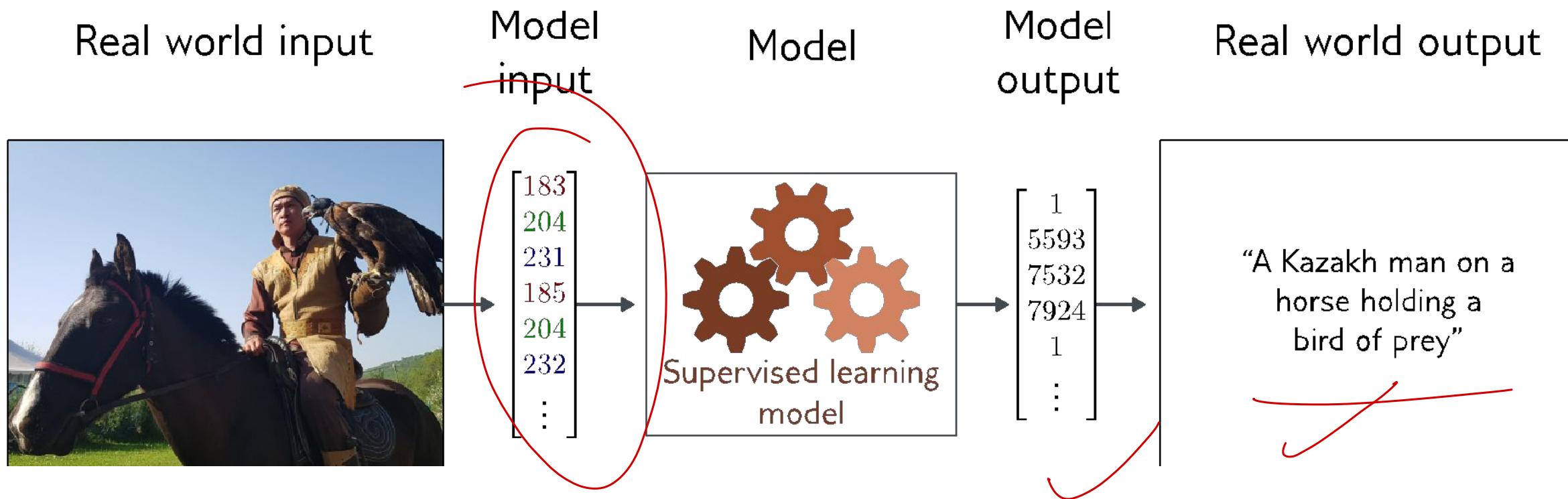
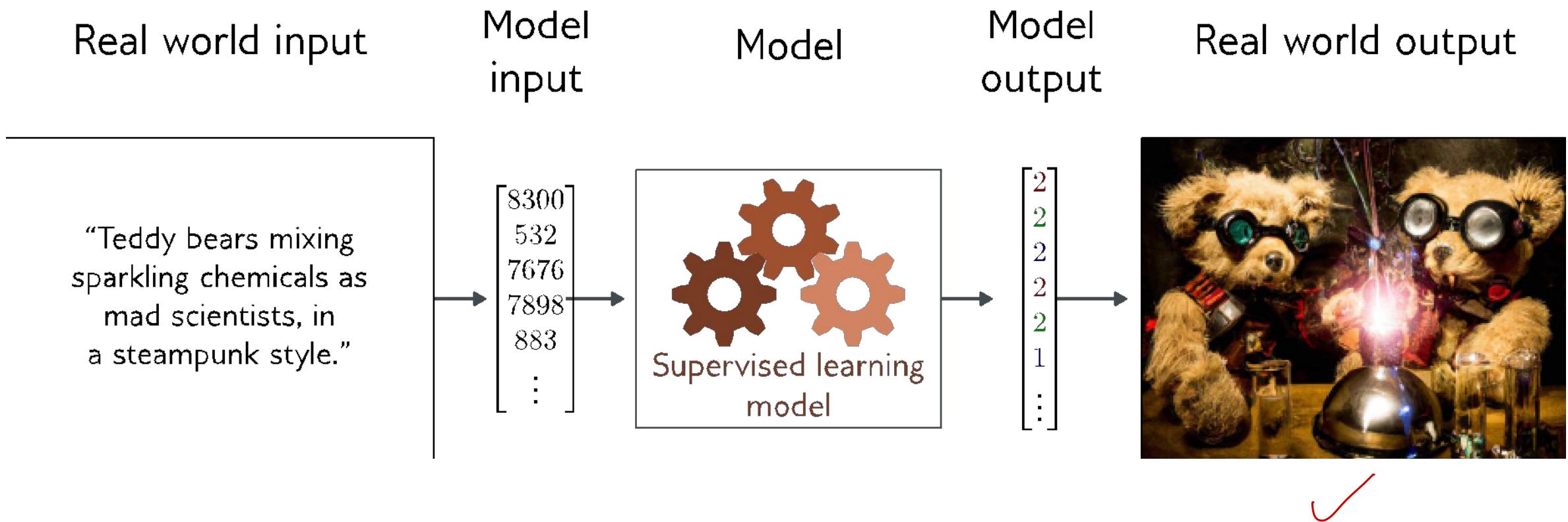


Image generation from text



Multimodal Input and Output

Unified-IO 2: Scaling Autoregressive Multimodal Models with Vision, Language, Audio, and Action

Jiasen Lu^{1*} , Christopher Clark^{1*} , Sangho Lee^{1*} , Zichen "Charles" Zhang^{1*} ,
Savya Khosla², Ryan Marten², Derek Hoiem², Aniruddha Kembhavi^{1,3} 

¹Allen Institute for AI; ²University of Illinois Urbana-Champaign; ³University of Washington

*Leading Authors, equal contribution.



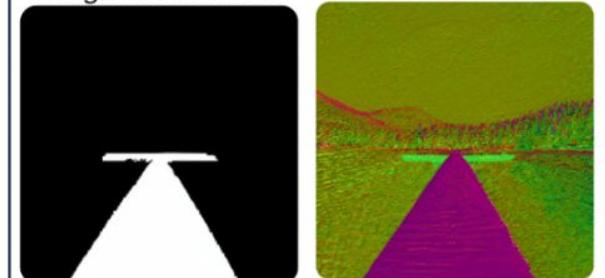
Multimodal Input and Output

Generate music about this scene. (original) 🔊

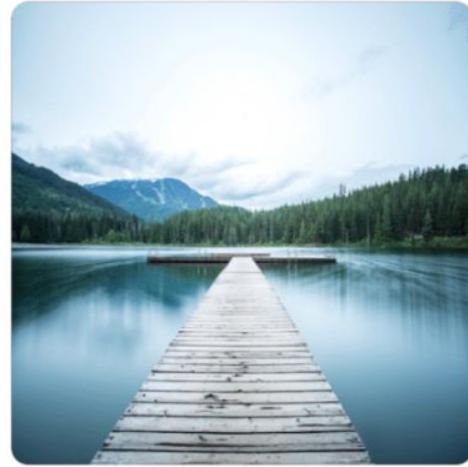
Detect the mountain. Detect the dock.



Segment the dock. Surface normals



Generate music about this scene. (haunted) 🔊



Please provide a short description of the image.
An image of dock on the water with trees in the background.

Please provide a detailed description of the image and share your personal impressions of the scene.
The image features a wooden jetty over a beautiful lake with numerous green trees surrounding the area. The scene gives off a peaceful and serene atmosphere. On the far side of the scene, a mountain range can be seen, stretching in the distance. In the middle of the scene, ...

What are the things I should be cautious about when I visit here?

When visiting this dock on the water surrounded by a forested area, there are several factors to consider:

1. Visibility: The visibility around the dock may be limited, so it is essential to take care while walking on the dock to avoid any sudden shifts in lighting conditions or obstructions on the shore.
2. Water levels: There might be flooding or excessive waves that could damage the wooden dock or affect the safety of the area. It is crucial to be mindful of the water levels to ...

Make the image haunted. Make the bridge red. Make it in midnight. Make it in sunset. Imagine you are da Vinci. Remove the bridge.



Unified-IO 2 can perform a multitude of multimodal tasks: captioning the image, following free-form instructions, image editing, object detection, semantic segmentation, surface normal, and image-based audio generation, etc. Here, we show the outputs of our model for a variety of prompts.

Free Form VQA



What food could you make with these ingredients ? Give me the recipe.

One delicious recipe using these ingredients is chocolate pudding! Here's the recipe:

Ingredients:

- 1 cup all-purpose flour, - ½ cup sugar

Instructions:

1. In a large bowl whisk together the flour, sugar..
2. In a separate bowl, mix together the eggs ...

Visual Parsing & Segmentation



Which fruits are in this image ? List them in json format with the name of the fruit as the key and the color of the fruit as the value.

```
{  
  "banana" : "Yellow",  
  "apple" : "Red",  
  "grapes" : "Green"  
}
```

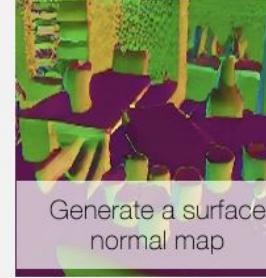
Segment grapes and one apple



Depth & Surface Normal



Generate a depth image



Generate a surface normal map



Generate a depth & surface normal map

Keypoint Estimation



Find the visible keypoints corresponding to the person located in the highlighted region.



Visual based Audio Generation



Generate an audio track for this band.



Generate an audio track for this band.

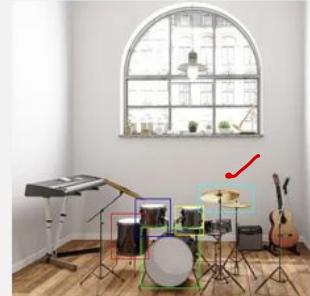


Visual Audio Localization



(drum sounds)

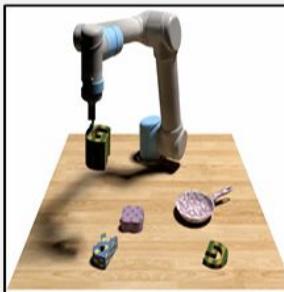
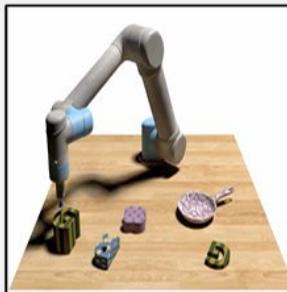
Identify the locations of the instruments producing the given sound.



Robotic Manipulation

= blicket
 = dax
 is kobar than

Put the less kobar blicket into the dax.



Future Frame Prediction

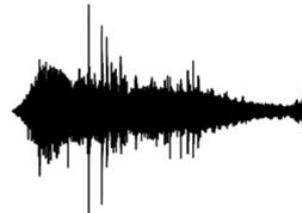


Given the initial image and a sequence of actions, predict the next frames

Generated Images



"Oil painting, cosmic horror painting,
elegant intricate ArtStation concept
art by Craig Mullins detailed"



(Pirate ship sailing ambience)



"Forward moving
camera view"



(Video of waterfall with camera moving forward)



(A video with sound with people watching panda and laughing)

→ "Panda eating bamboo,
people laughing."

So, how do we encode and decode multiple modalities?

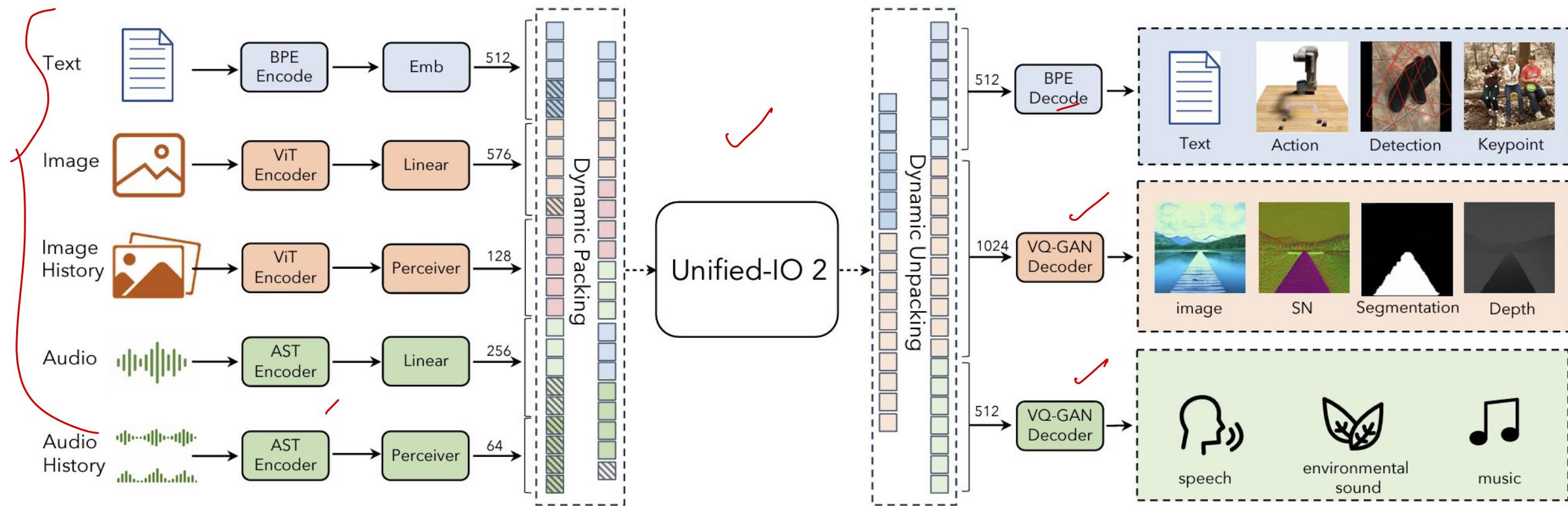
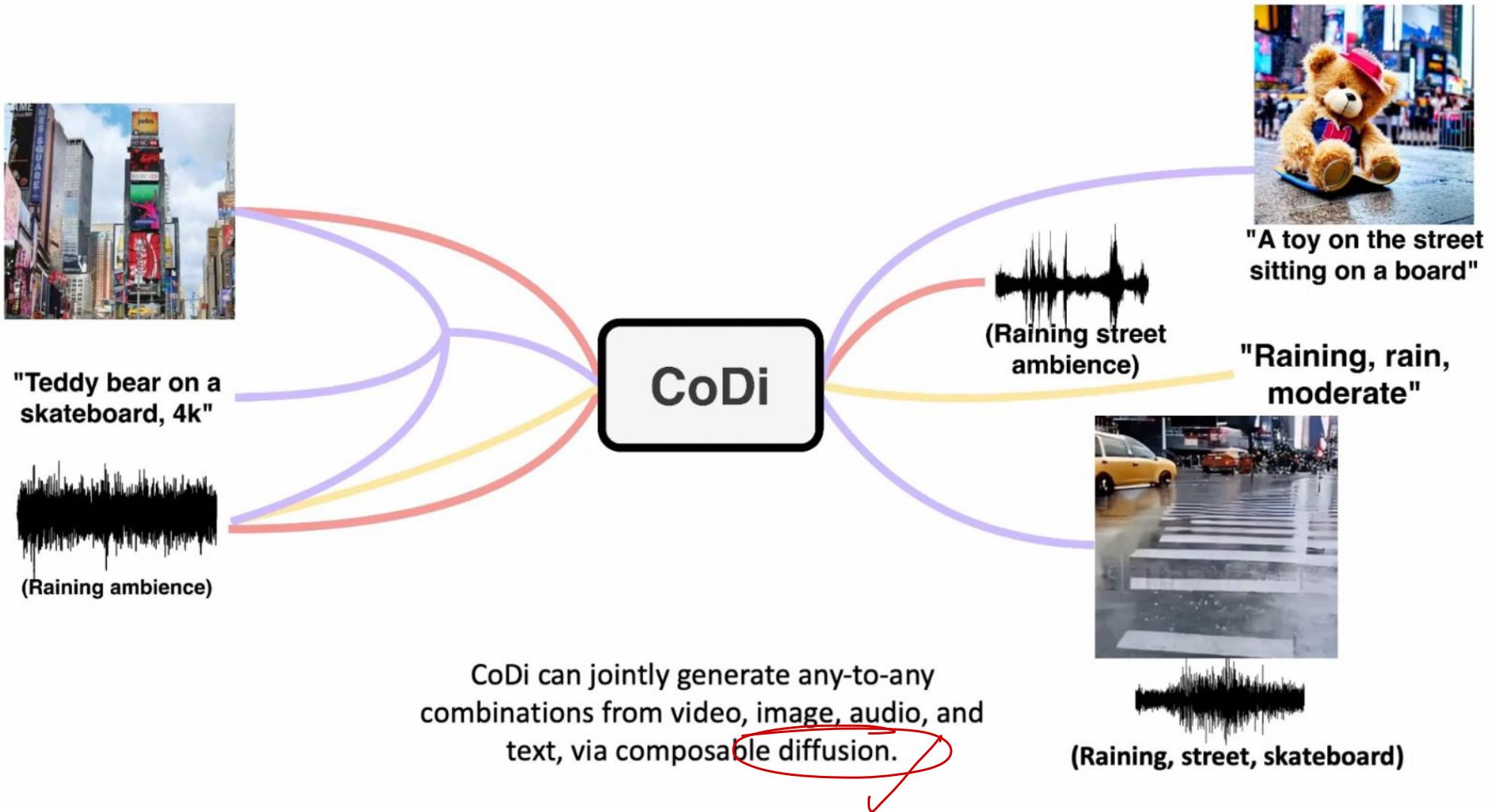


Figure 2. UNIFIED-IO 2 architecture. Input text, images, audio, or image/audio history are encoded into sequences of embeddings which are concatenated and used as input to an encoder-decoder transformer model. The **transformer** outputs discrete tokens that can be decoded into text, an image, or an audio clip.

CoDi: Any-to-Any Multimodal Generation



Artificial intelligence

Machine learning

Supervised
learning

Unsupervised
learning

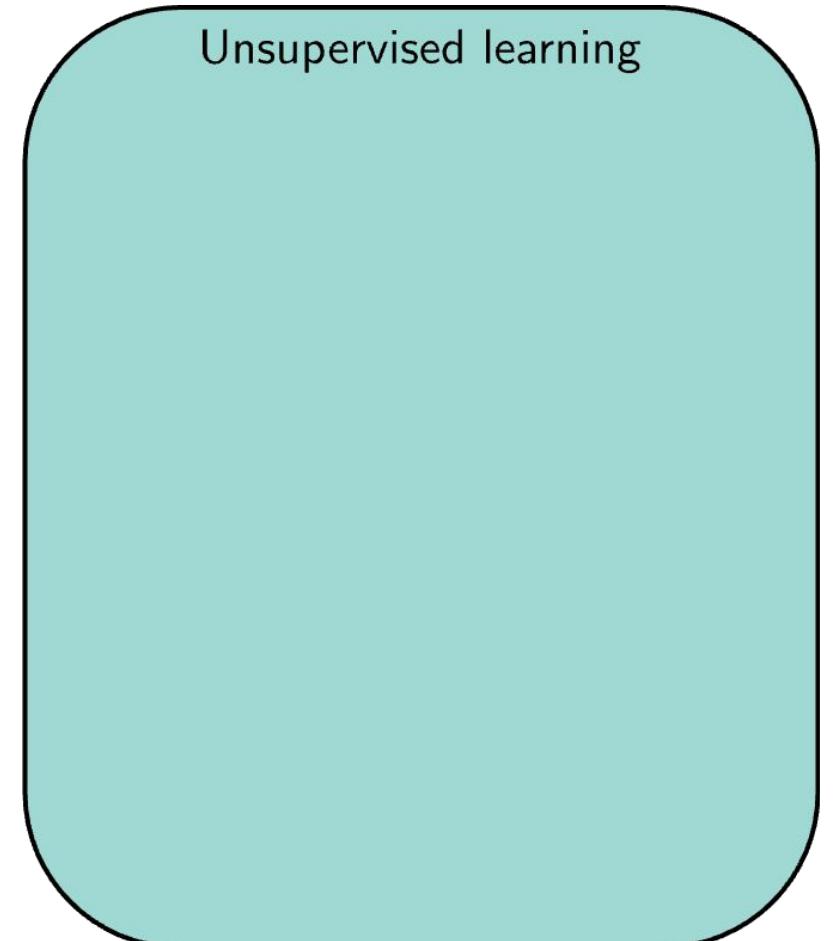
Reinforcement
learning

Deep learning



Unsupervised Learning

- Learning about a dataset without labels
 - Clustering
 - Generating new examples
 - Filling in missing data

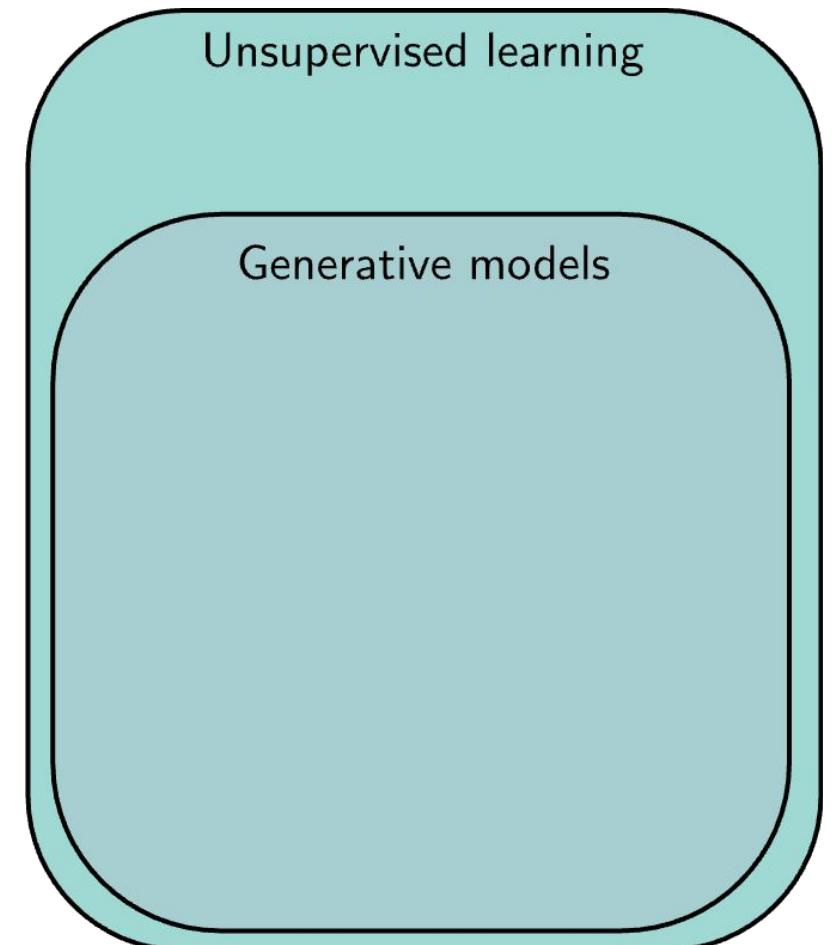


Unsupervised learning

Unsupervised Learning

- Learning about a dataset without labels
 - e.g., clustering
- Generative models can create examples
 - e.g., generative adversarial networks

GAN



Unsupervised Learning

- Learning about a dataset without labels
 - e.g., clustering
- Generative models can create examples
 - e.g., generative adversarial networks
- PGMs learn distribution over data
 - e.g., variational autoencoders,
 - e.g., normalizing flows,
 - e.g., diffusion models

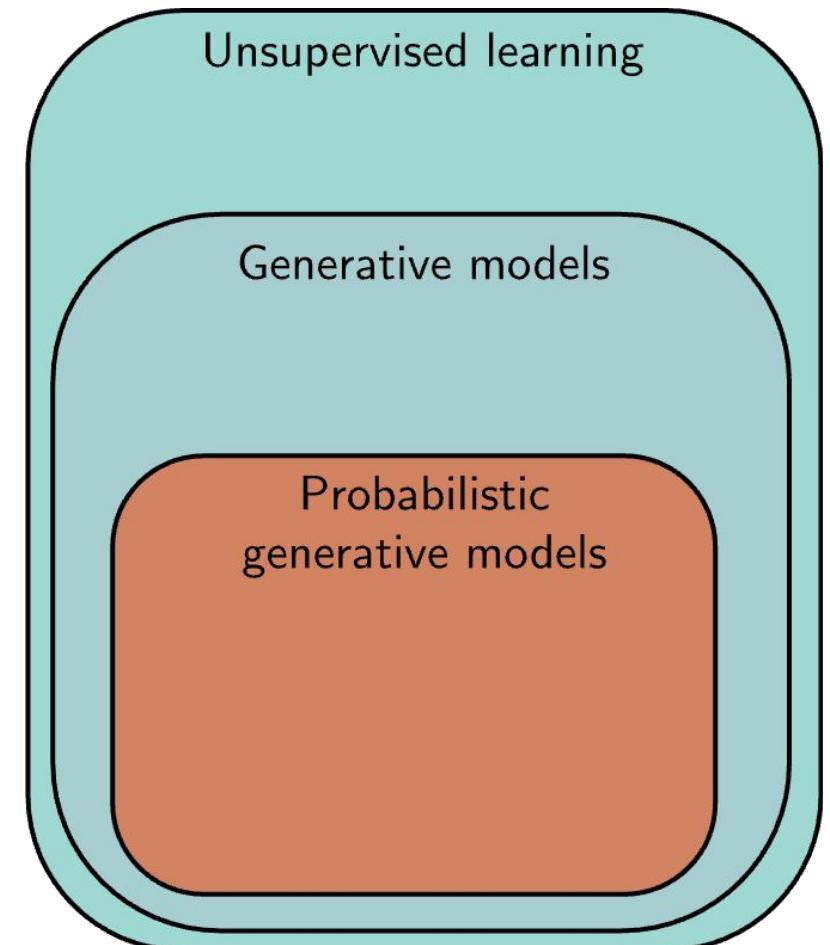
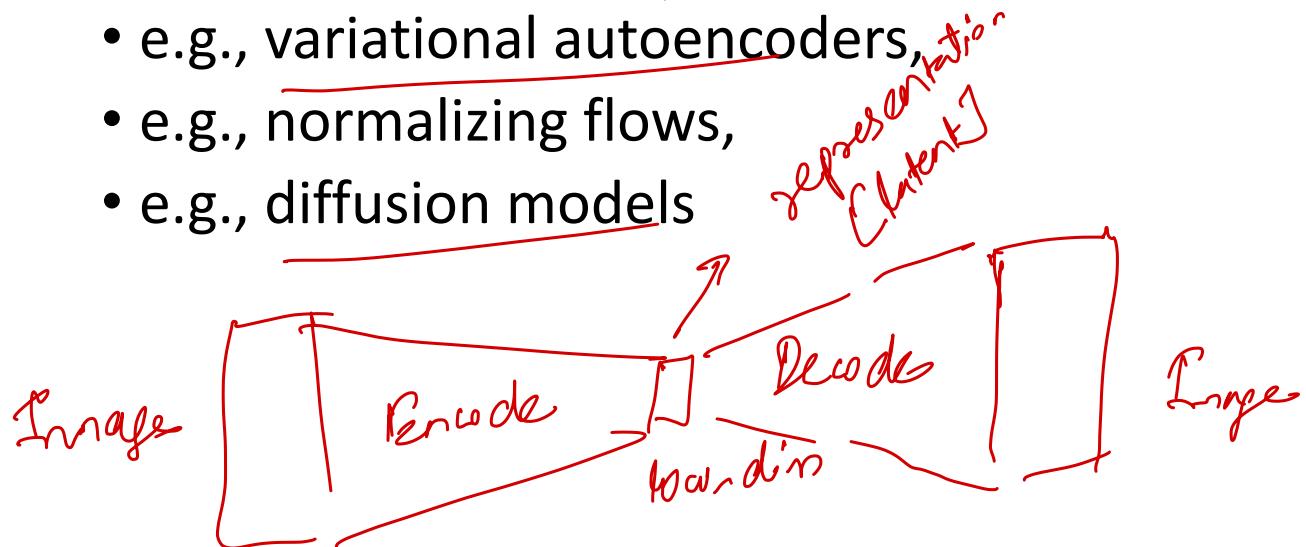




Figure 1.5 Generative models for images. Left: two images were generated from a model trained on pictures of cats. These are not real cats, but samples from a probability model. Right: two images generated from a model trained on images of buildings. Adapted from Karras et al. (2020b).

The moon had risen by the time I reached the edge of the forest, and the light that filtered through the trees was silver and cold. I shivered, though I was not cold, and quickened my pace. I had never been so far from the village before, and I was not sure what to expect. I had been walking for hours, and I was tired and hungry. I had left in such a hurry that I had not thought to pack any food, and I had not thought to bring a weapon. I was unarmed and alone in a strange place, and I did not know what I was doing.

I had been walking for so long that I had lost all sense of time, and I had no idea how far I had come. I only knew that I had to keep going. I had to find her. I was getting close. I could feel it. She was nearby, and she was in trouble. I had to find her and help her, before it was too late.

Figure 1.6 Short story synthesized from a generative model of text data. The model describes a probability distribution that assigns a probability to every output string. Sampling from the model creates strings that follow the statistics of the training data (here, short stories) but have never been seen before.

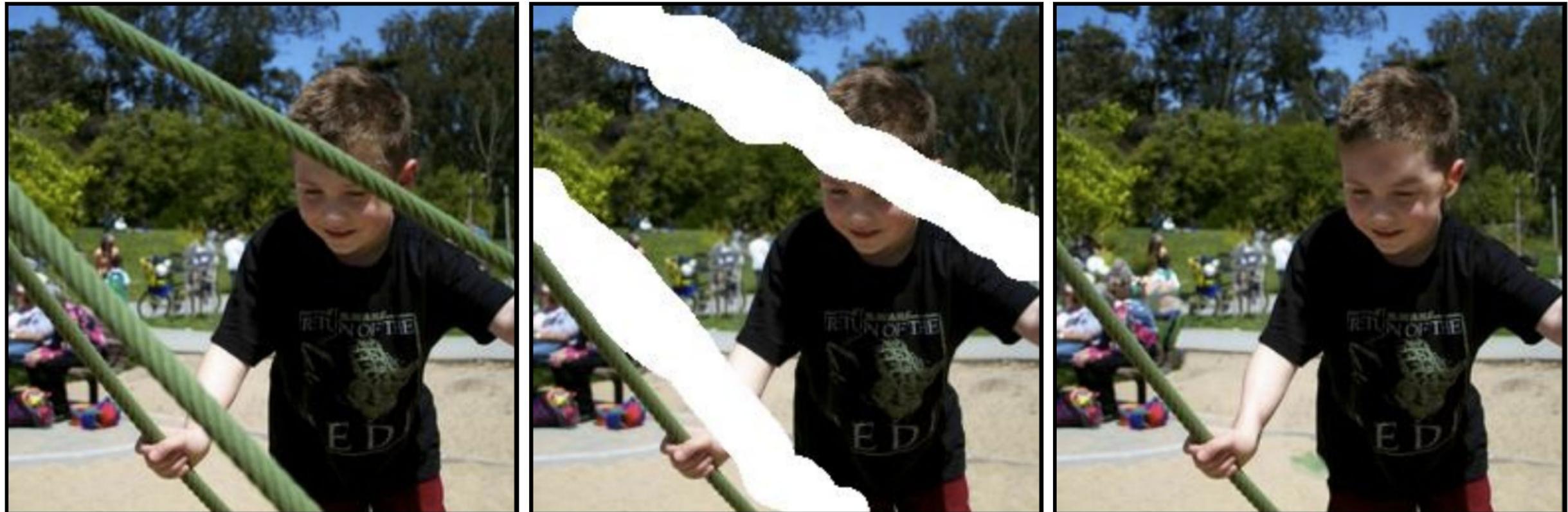


Figure 1.7 Inpainting. In the original image (left), the boy is obscured by metal cables. These undesirable regions (center) are removed and the generative model synthesizes a new image (right) under the constraint that the remaining pixels must stay the same. Adapted from Saharia et al. (2022a).

Denoising diffusion models

- Forward / noising process

- Sample data $p(\mathbf{x}_0) \rightarrow$ turn to noise



- Reverse / denoising process

- Sample noise $p_T(\mathbf{x}_T) \rightarrow$ turn into data



Reparameterization of the target

$$L[\phi_{1\dots T}] = \quad \quad \quad (18.36)$$

$$\sum_{i=1}^I -\log \left[\text{Norm}_{\mathbf{x}_i} [\mathbf{f}_1[\mathbf{z}_{i1}, \phi_1], \sigma_1^2 \mathbf{I}] \right] + \sum_{t=2}^T \frac{\beta_t^2}{(1-\alpha_t)(1-\beta_t)2\sigma_t^2} \left\| \mathbf{g}_t[\mathbf{z}_{it}, \phi_t] - \epsilon_{it} \right\|^2.$$

The log normal can be written as a least squares loss plus a constant C_i (section 5.3.1):

$$L[\phi_{1\dots T}] = \sum_{i=1}^I \frac{1}{2\sigma_1^2} \left\| \mathbf{x}_i - \mathbf{f}_1[\mathbf{z}_{i1}, \phi_1] \right\|^2 + \sum_{t=2}^T \frac{\beta_t^2}{(1-\alpha_t)(1-\beta_t)2\sigma_t^2} \left\| \mathbf{g}_t[\mathbf{z}_{it}, \phi_t] - \epsilon_{it} \right\|^2 + C_i.$$

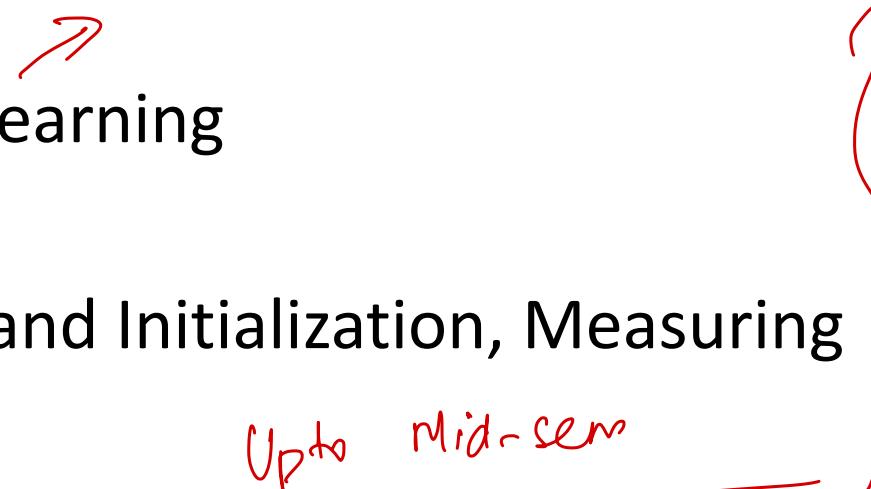
$$\frac{1}{2\sigma_1^2} \left\| \mathbf{x}_i - \mathbf{f}_1[\mathbf{z}_{i1}, \phi_1] \right\|^2 = \frac{1}{2\sigma_1^2} \left\| \frac{\beta_1}{\sqrt{1-\alpha_1}\sqrt{1-\beta_1}} \mathbf{g}_1[\mathbf{z}_{i1}, \phi_1] - \frac{\beta_1}{\sqrt{1-\alpha_1}\sqrt{1-\beta_1}} \epsilon_{i1} \right\|^2. \quad (18.37)$$

$$L[\phi_{1\dots T}] = \sum_{i=1}^I \sum_{t=1}^T \frac{\beta_t^2}{(1-\alpha_t)(1-\beta_t)2\sigma_t^2} \left\| \mathbf{g}_t[\mathbf{z}_{it}, \phi_t] - \epsilon_{it} \right\|^2$$



Tentative Plan

- Introduction to Deep Learning, Supervised learning
- Shallow and Deep Neural Networks
- Loss Functions, Training Models, Gradients and Initialization, Measuring Performance, Regularization
- Convolutional Neural Networks
- Transformers → GPT → Text, Image, Audio
- Multi-modal Foundational Models
- Generative Adversarial Networks, Variational Autoencoders
- Diffusion Models



Up to Mid-sem

Text, Image, Audio