COMP 6611C: Advanced Topics in Embedded AI Systems

Lecture 1: Machine Learning Basics

Xiaomin Ouyang

Assistant Professor

Department of Computer Science and Engineering, HKUST

香港科技大學
THE HONG KONG UNIVERSITY OF
SCIENCE AND TECHNOLOGY

Recap

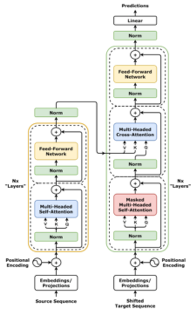
- ➢ Auditing students:
  - ➢ please email me your name/email/ID for joining canvas
- ➢ Late submission policy:
  - 20% reduction per day, request approval if having emergency
  - Only for project report and paper reviews, pre slides are more casual
- ➢ Q&A recording spreadsheet

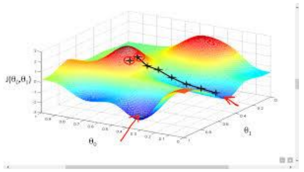
Outline

- **Overview of Machine Learning**
- Machine Learning Paradigms
- Model Architectures
- Machine Learning Systems
- Applications

What is Machine Learning

(3)





Data

Learning Algorithm

Computing Power

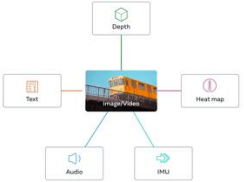
Model
(knowledge)

y=f(**x**)
X is the data
Y is the output

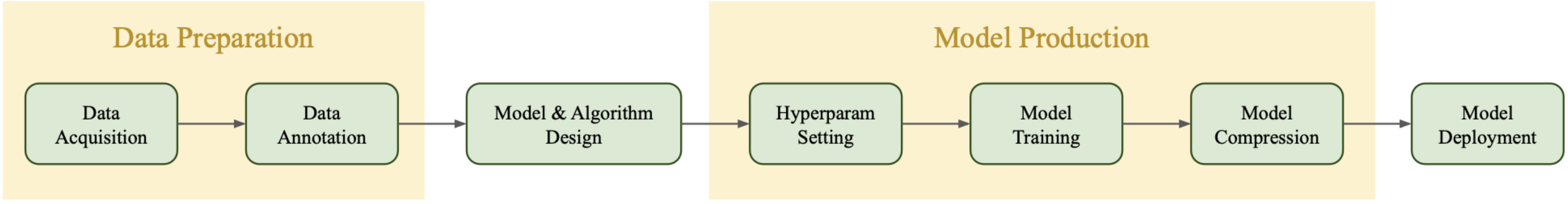
(1)

(4)

(2)



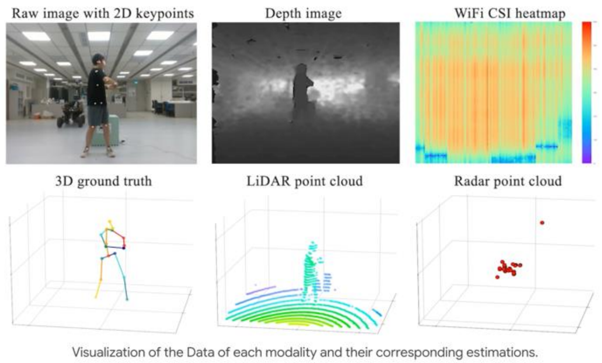
Workflow of Machine Learning



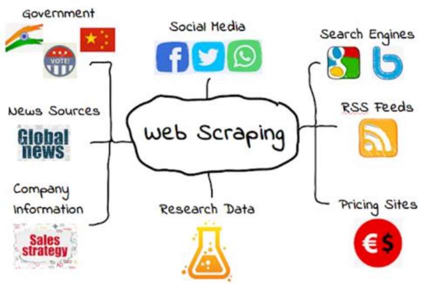
- Data Preparation
- Model & Algorithm Design
- Model Training
- Model Deployment and Inference

Data Preparation

➢ Data Acquisition



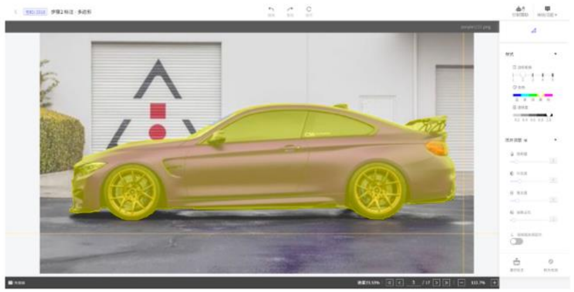
Captured by sensors



Crawled from web

Synthetic data

➢ Data Annotation

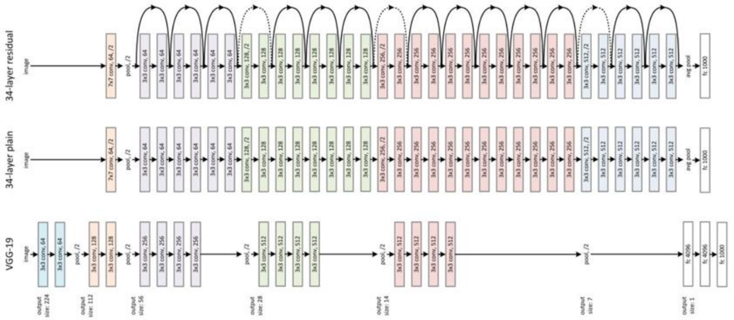


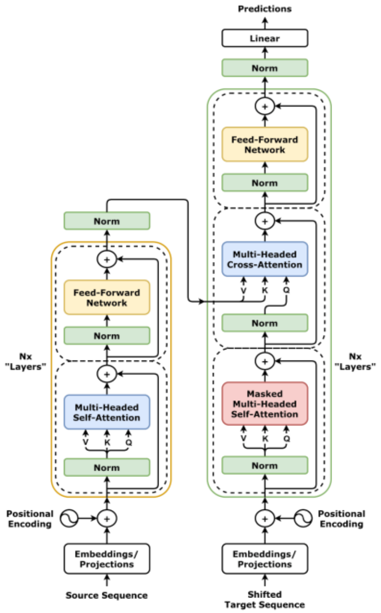
➢ Data Preprocess

- Normalization
- Feature selection
- Crop, resize
- Augmentation
- …

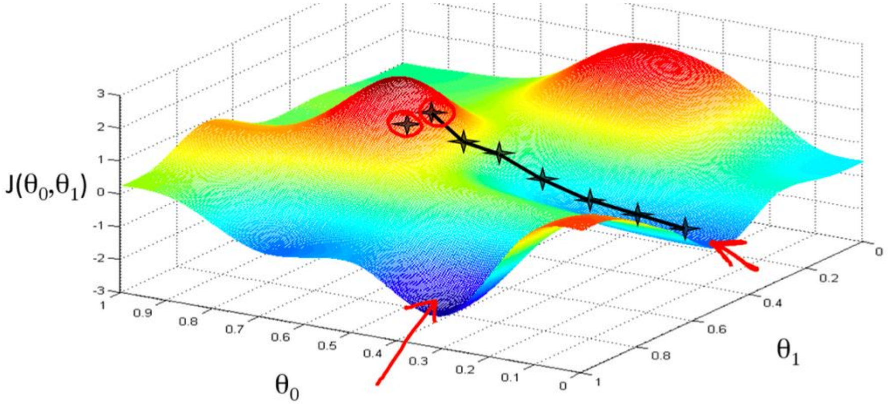
Model & Algorithm Design

- Model Architecture
  - MLP
  - CNN
  - RNN
  - ResNet
  - Transformers



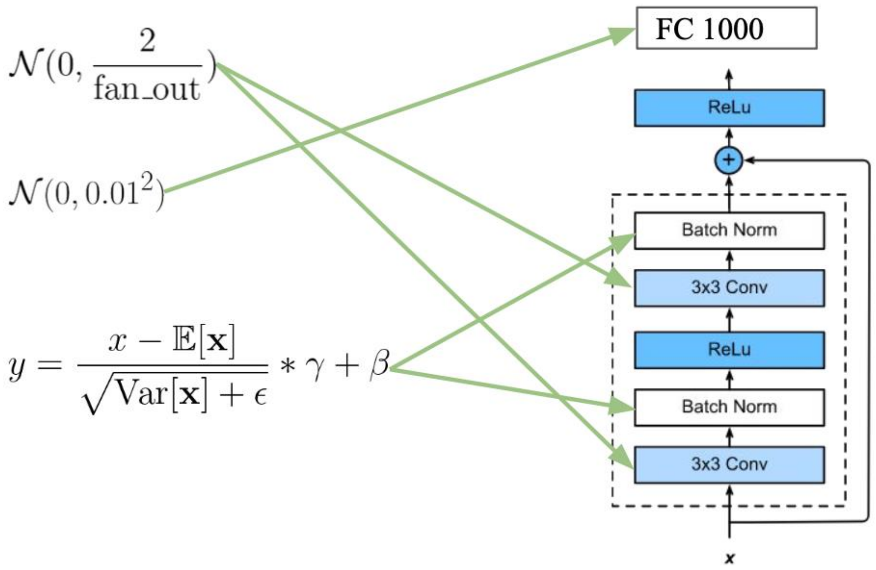


- Learning Algorithms
  - Gradient Descent
  - Stochastic Gradient Descent (SGD)
  - Adaptive Moment Estimation (Adam)



Model Training

➢ Weight Initialization



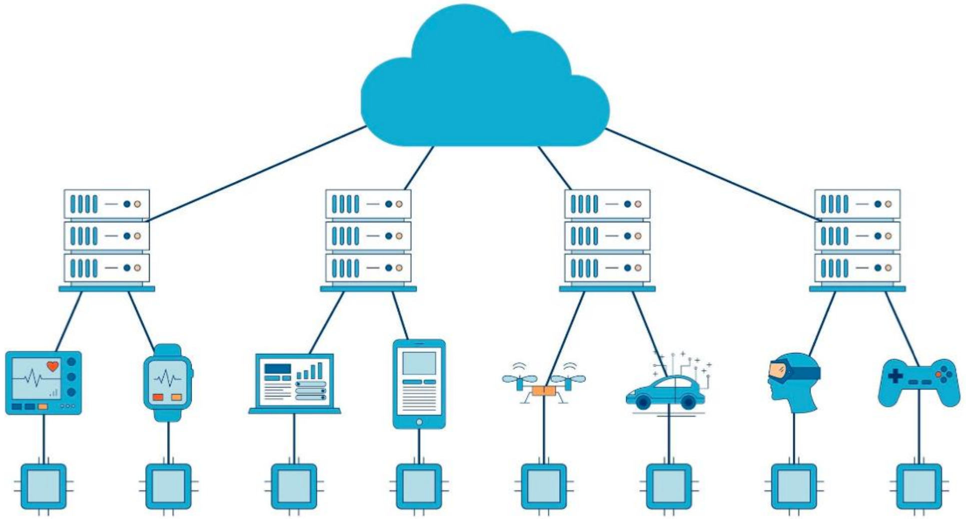
- Zero Initialization
- Random Initialization
- Kaiming initialization
- Gaussian distribution
- …

➢ Hyperparameters

- Model design
  - the number of layers
  - the size of each layer
  - other layer parameters
  - activation functions
- Learning algorithm
  - choice of optimizers
  - learning rate
  - batch size
  - dropout ratios

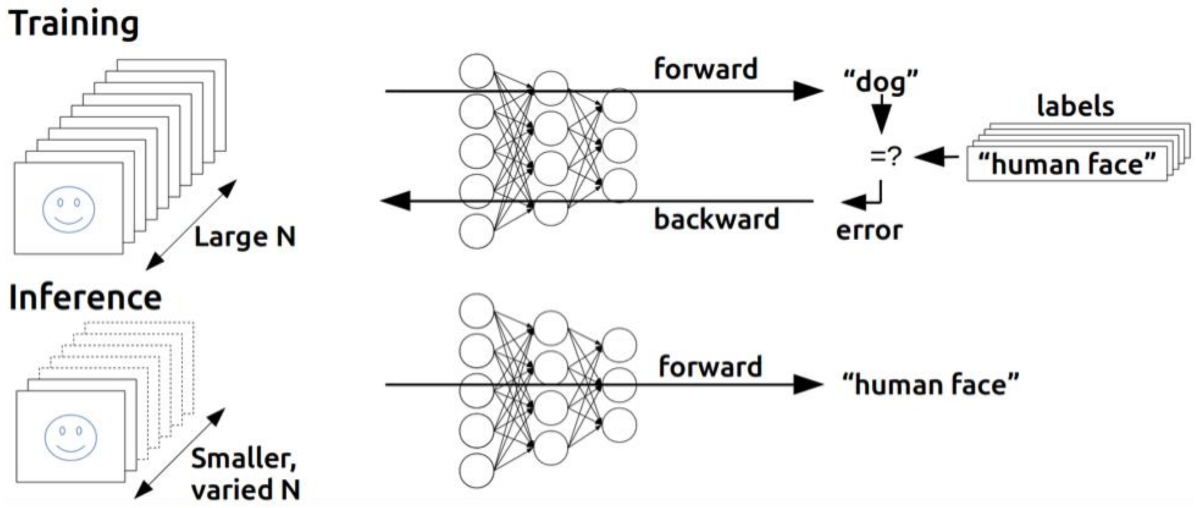
Model Deployment and Inference

➢ Model Deployment



- Cloud-Edge-Devices
- Model compression/partition
- Model finetuning and adaptation

➢ Model Inference

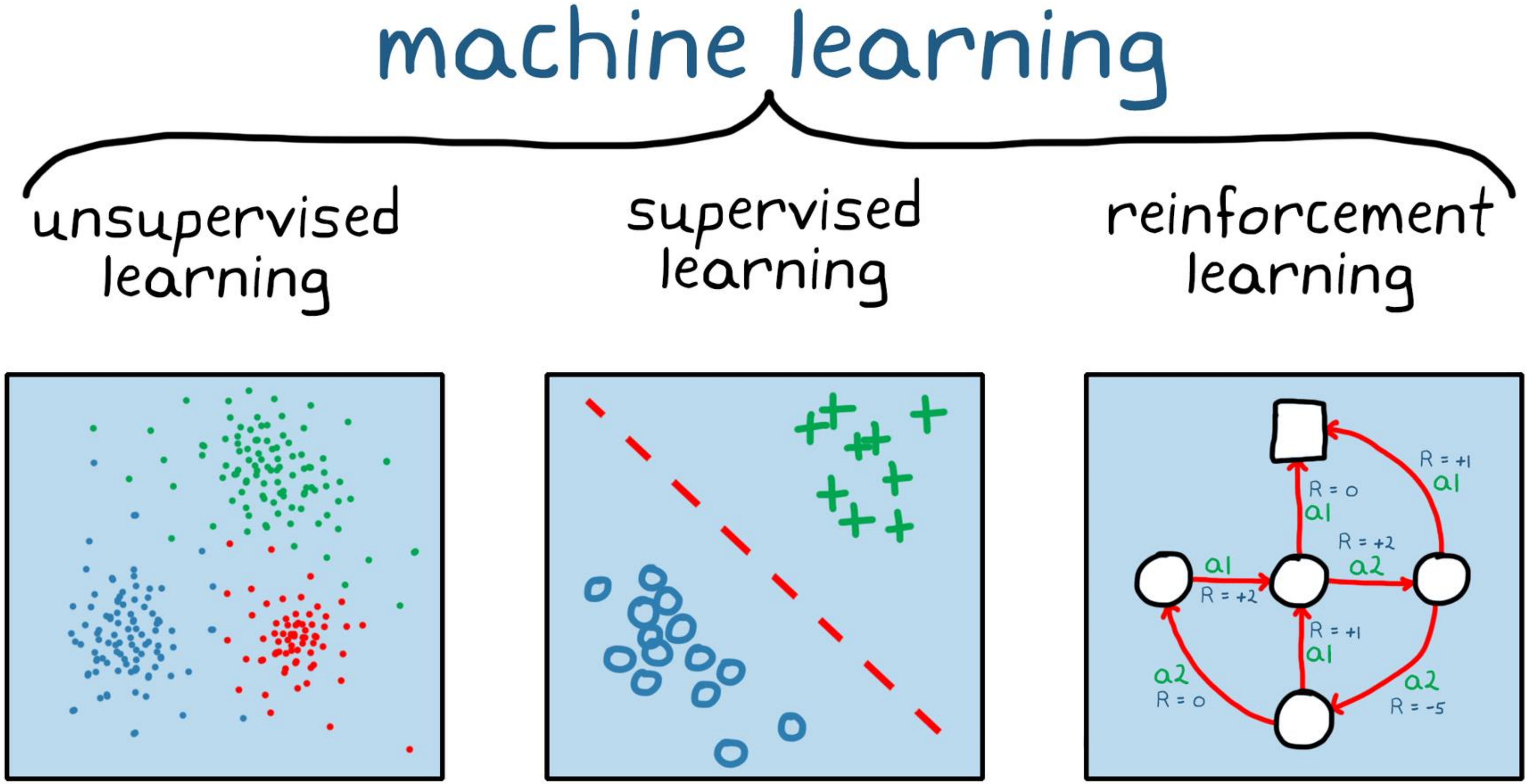


- Accuracy, latency, energy, memory

Outline

- Overview of Machine Learning
- **Machine Learning Paradigms**
- Model Architectures
- Machine Learning Systems
- Applications

Types of Machine Learning



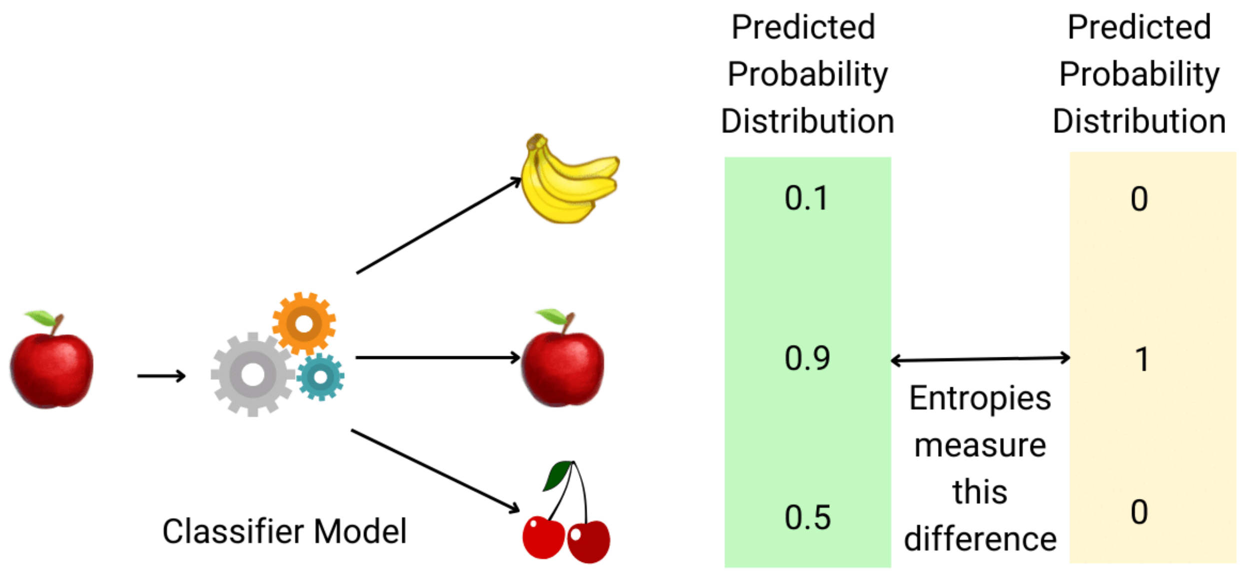
Data-driven

Task-driven

Learn from mistakes

Supervised Learning

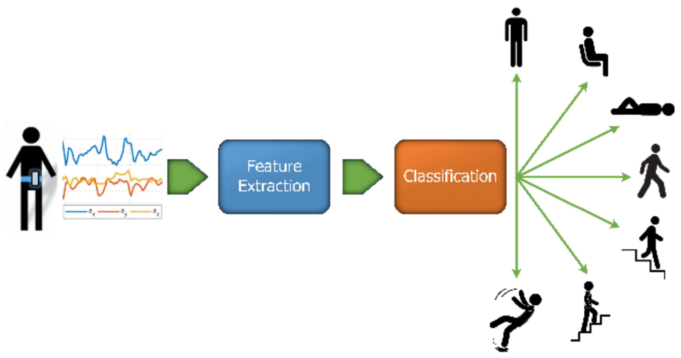
- Given a dataset with data and labels (**x**, y), find a function that **maps x** -> y

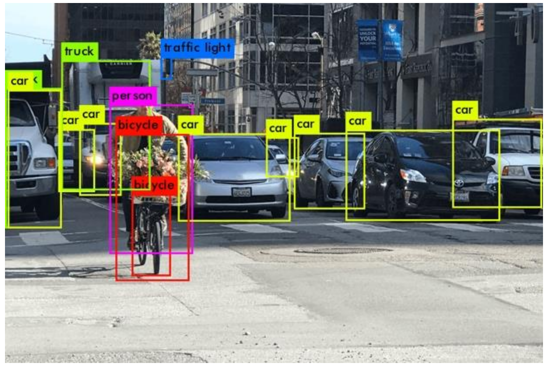


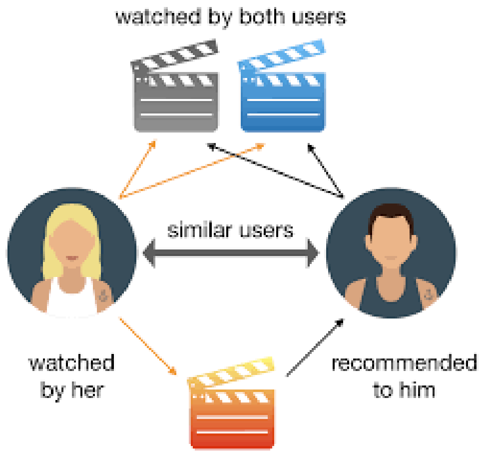
Training Loss is calculated by comparing predictions with y, e.g., cross entropy loss

Supervised Learning

➢ Applications







Classification

Regression

Detection

Recommendation

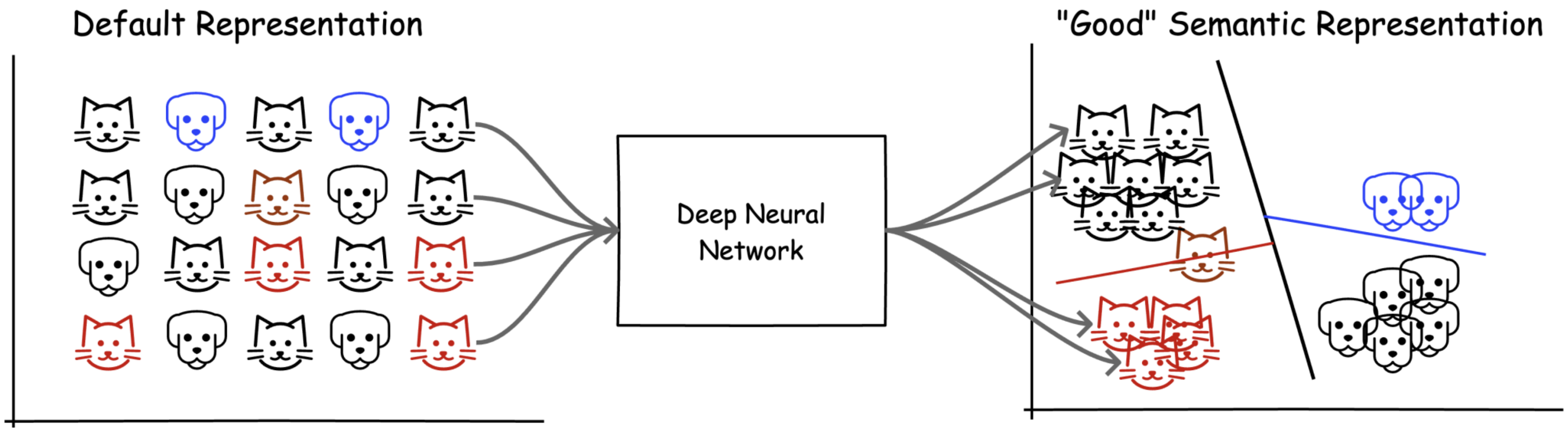
Supervised Learning

- Approaches

  - Linear Regression,
  - Logistic Regression
  - Decision Tree
  - Random Forests
  - Support Vector Machines (SVM)
  - K-Nearest Neighbors (KNN)
  - Naive Bayes
  - Neural Networks

Unsupervised Learning

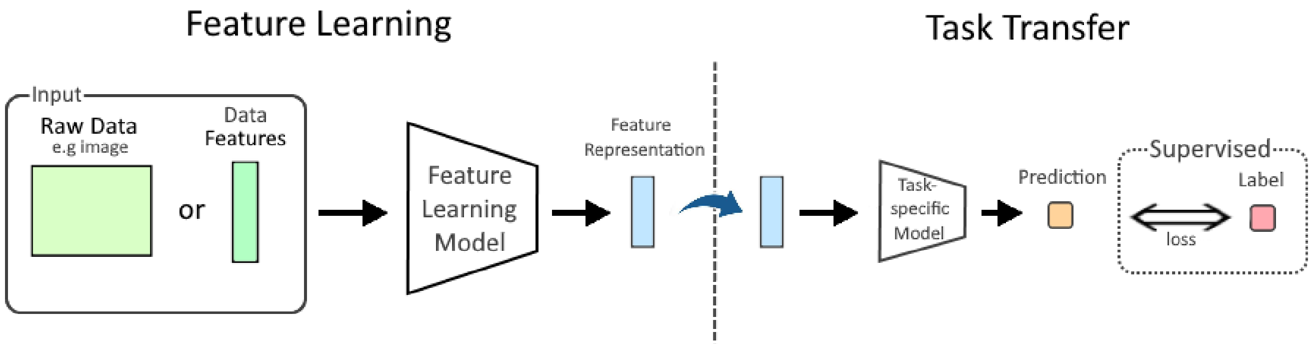
- Given a dataset with only data **x,** learn an **effective representation** of **x**



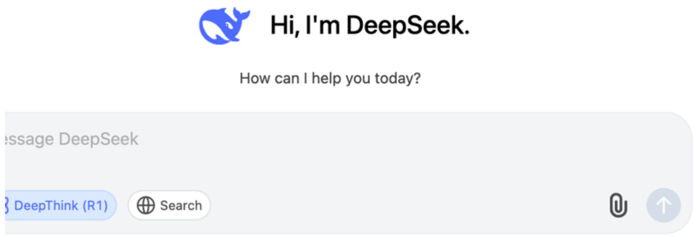
Unsupervised Learning

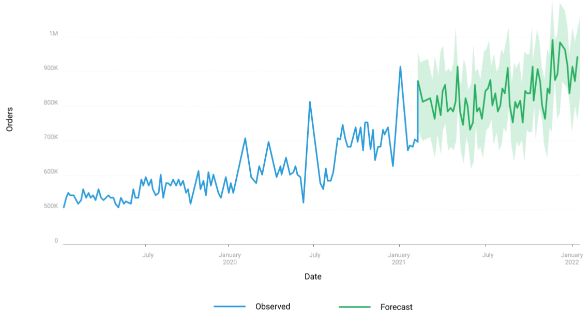
➢ Applications

Finetuned for Downstream Tasks



Generation

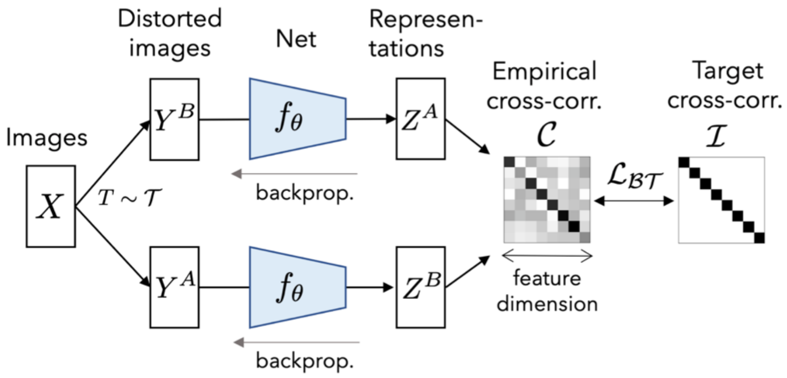


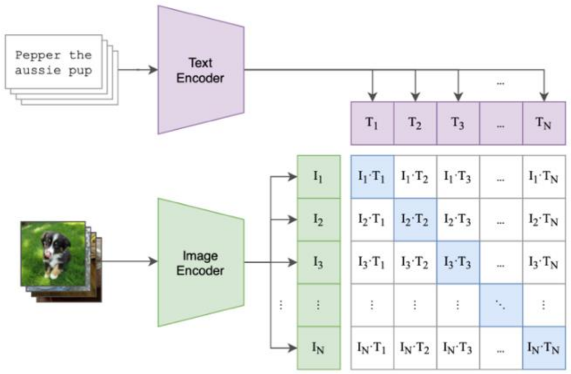


Sure, here is an image of a futuristic car driving through an old mountain road surrounded by nature:

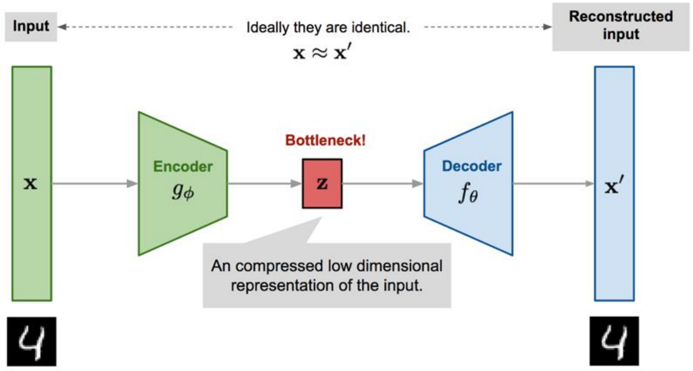
Unsupervised Learning

➢ Approaches

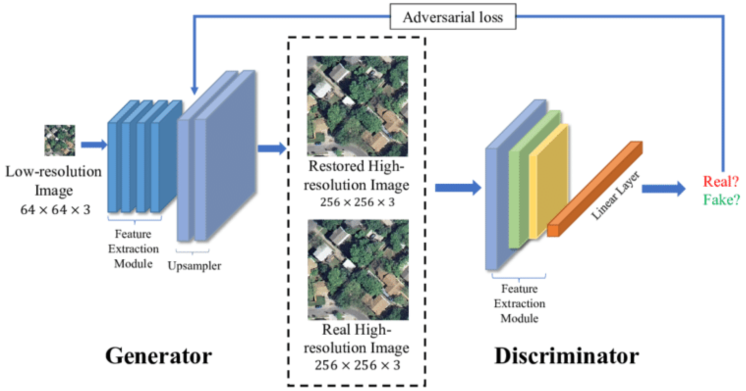




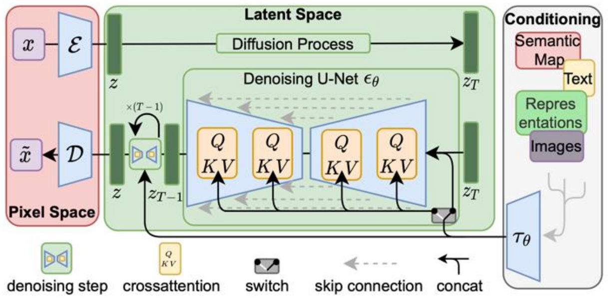
Contrastive Learning



VAE



GAN

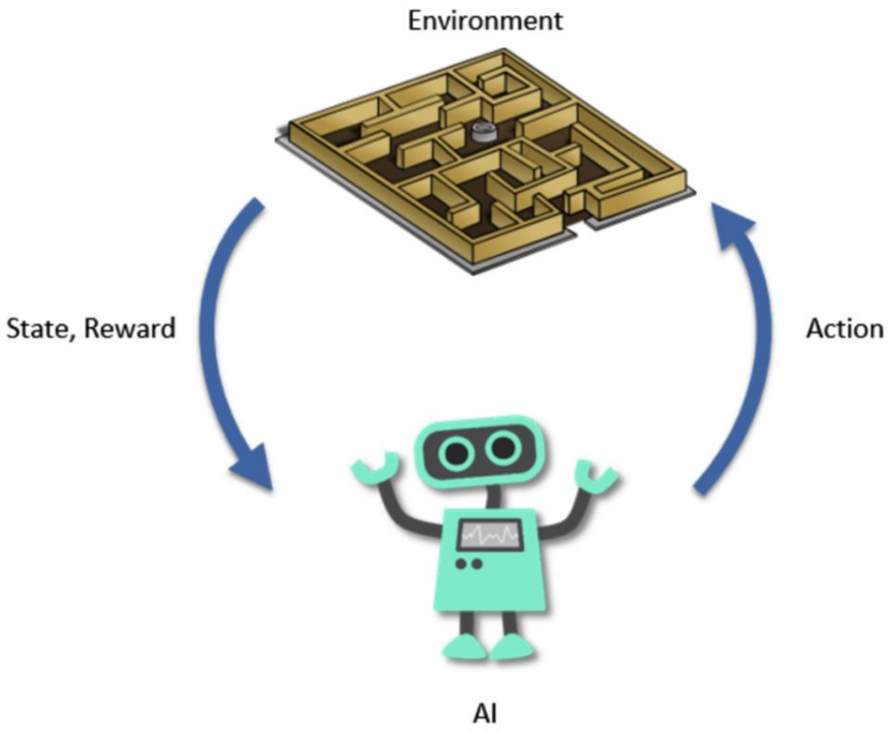


Diffusion

Reconstruction

Reinforcement Learning

- Given a dataset with state, action and reward (**s**, **a**, r), find a function to **maximize the reward** r

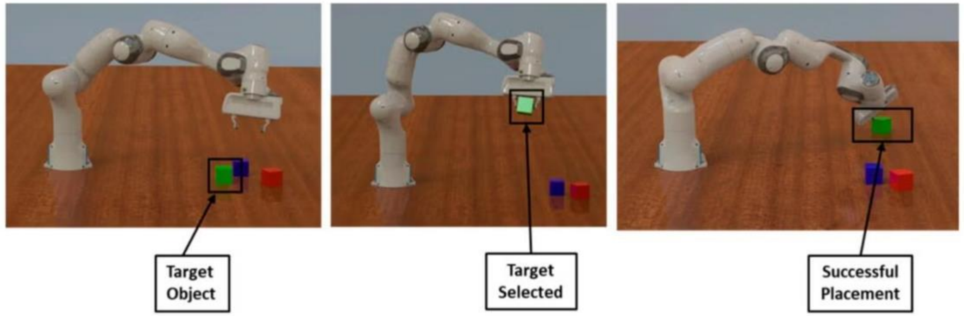


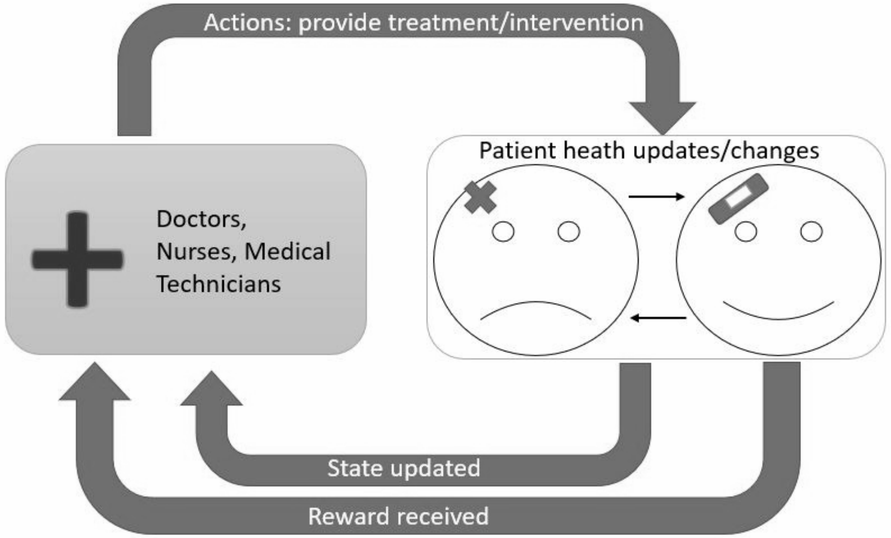
Reinforcement Learning

- Applications

Alpha Go

Robotic Control





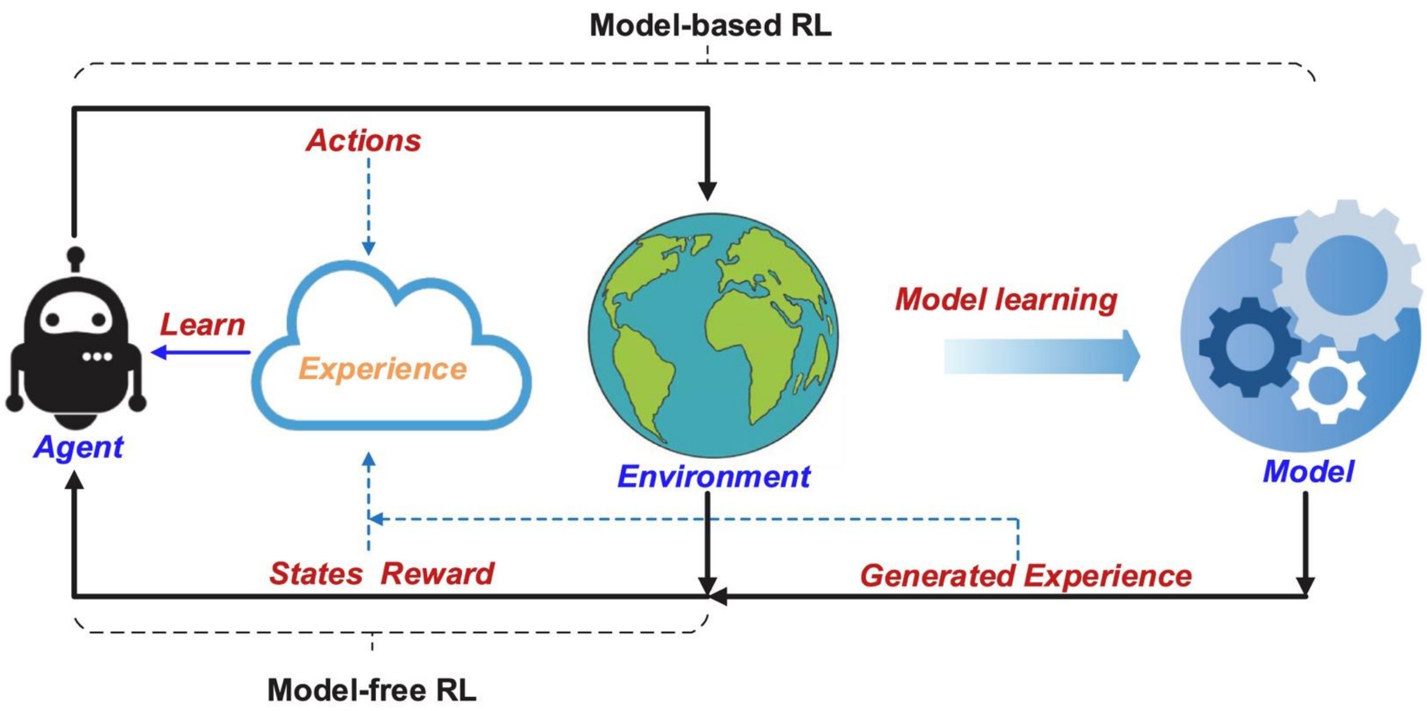
Health Intervention

Reinforcement Learning

- Approaches

  Model-based RL: build a model for the environment, sample-efficient

  Model-free RL: learn directly from environment, simpler to implement



How to choose different paradigms

- Lots of **labelled data**: supervised learning

- Lots of **unlabelled data**: unsupervised learning

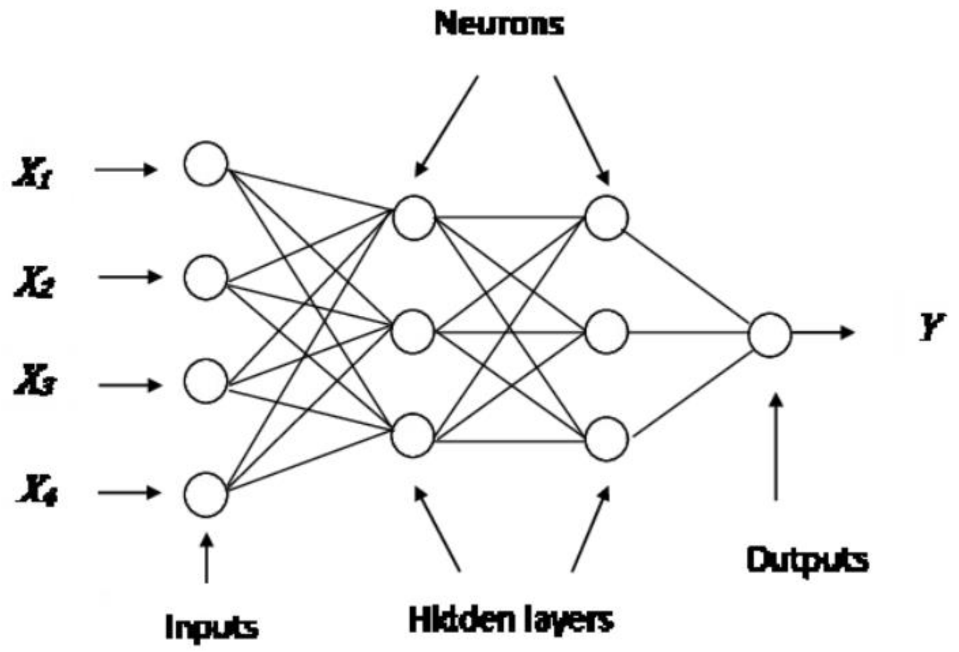
- No data labels, only **feedback signals**: reinforcement learning

Outline

- Overview of Machine Learning
- Machine Learning Paradigms
- **Model Architectures**
- Machine Learning Systems
- Applications

Multi-Layer Perceptron (MLP)

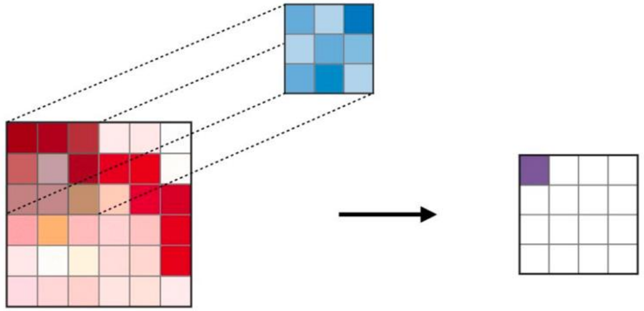
- Consists of multiple layers of neurons (fully connected layers), each taking the output of previous as input and generating outputs for the next layer.

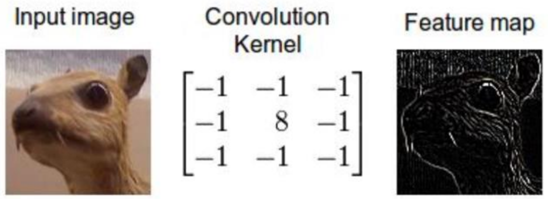


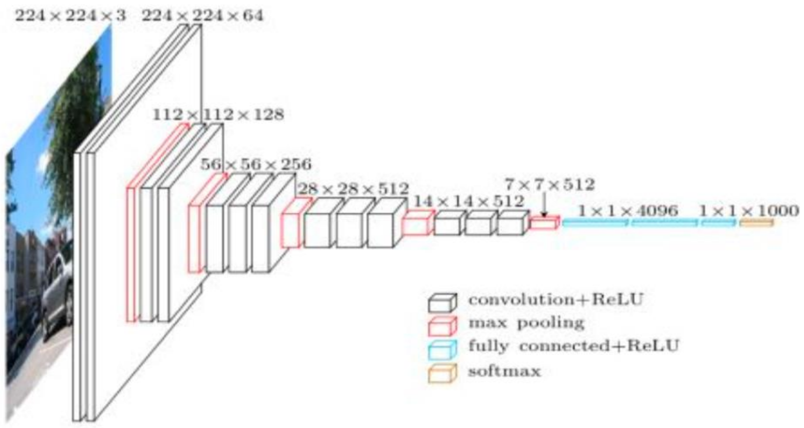
$$
\begin{aligned}
& & h_0 &= x \\
z_1 &= W_1 h_0 + b_1 & h_1 &= \sigma(z_1) \\
&\cdots & &\cdots \\
z_L &= W_L h_{L-1} + b_L & h_L &= \sigma(z_L) \\
& & y &= h_L
\end{aligned}
$$

Convolutional Neural Network (CNN)

- Extracts feature on small local receptive fields with shared kernel weights.



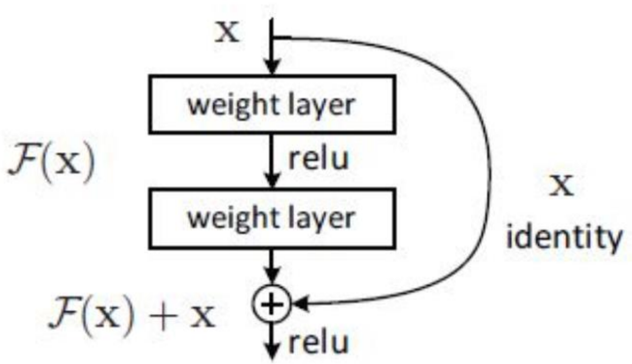


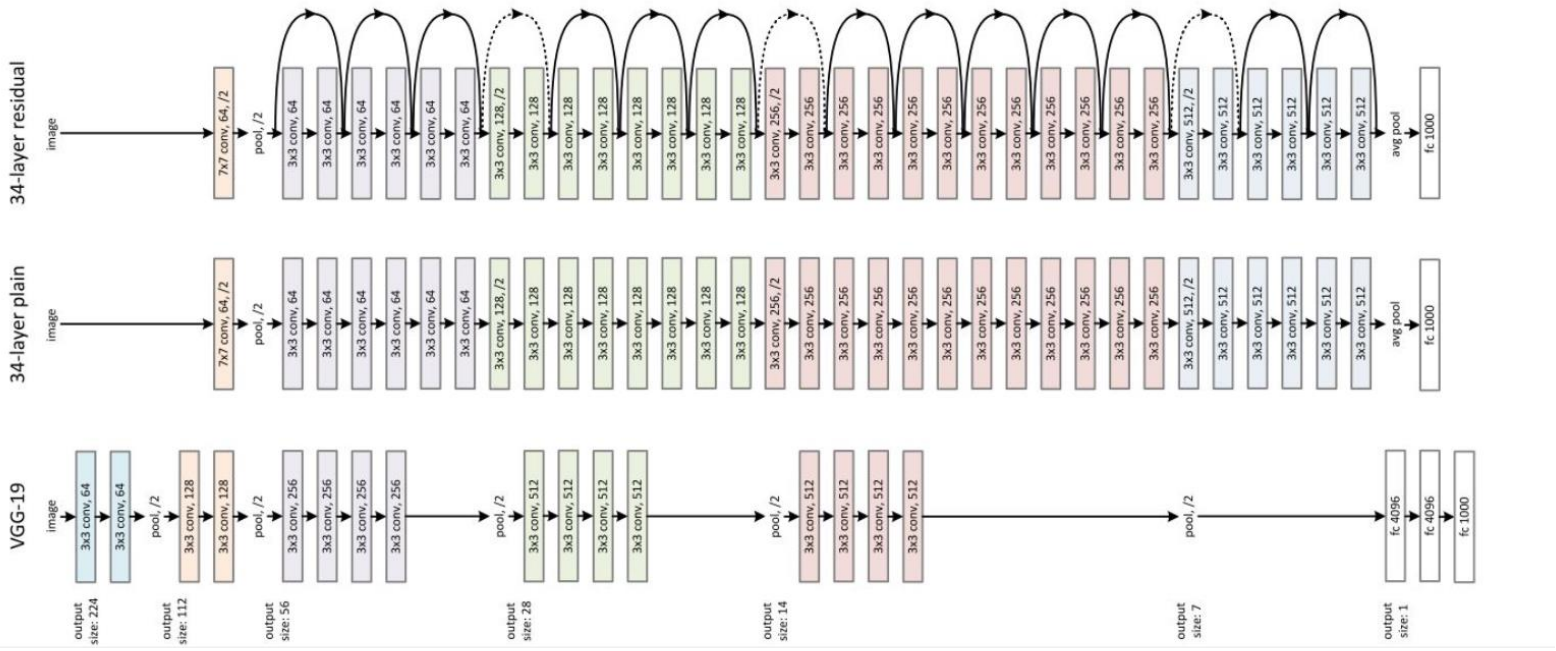


VGGNet (2014)

Residual Networks (ResNet)

➢ Introduces shortcut connections based on its residual learning paradigm and dramatically increases network depth above 1000.

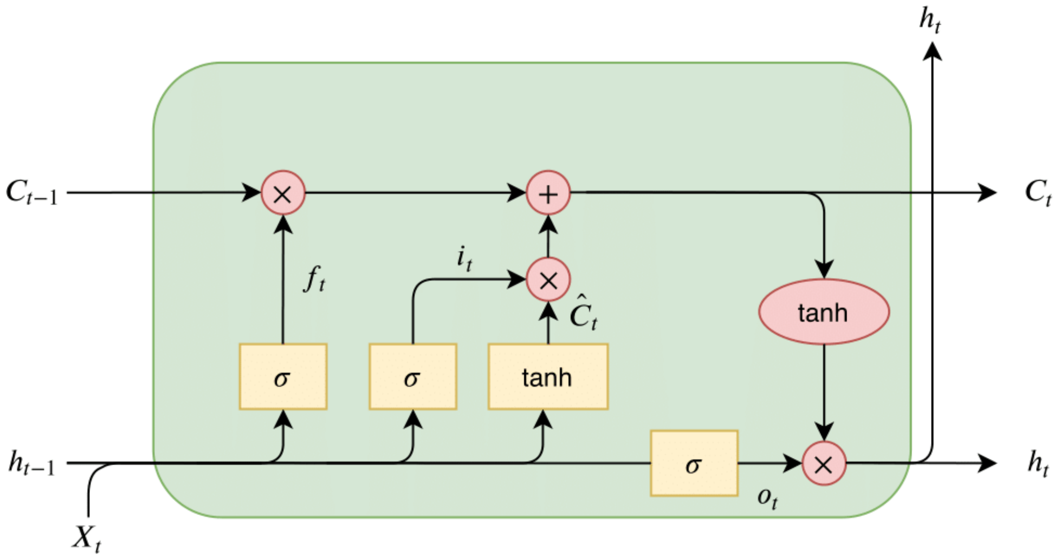




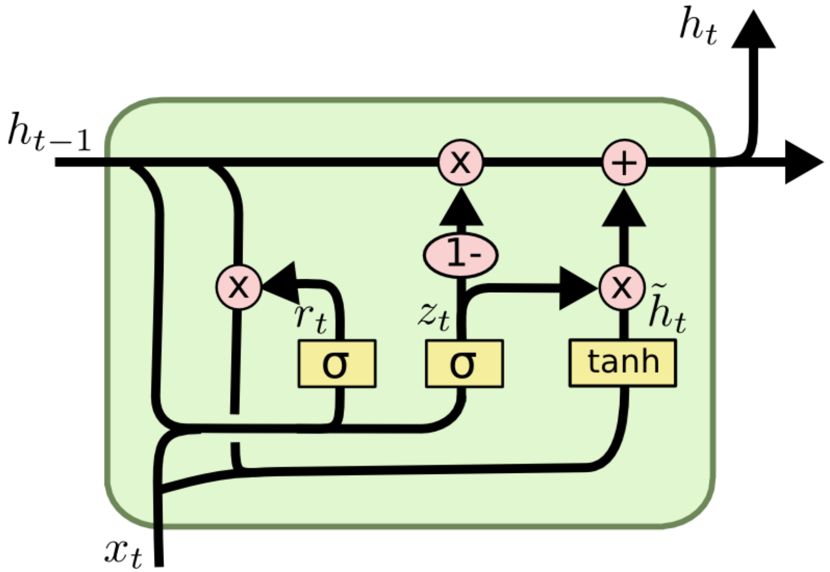
Given a target mapping H(x) and a network F(x). Fitting the full mapping F(x) = H(x) is harder than just fitting the residual F(x) = H(x)-x.

Recurrent Neural Network (RNN)

- Processing sequential data: connections between nodes form a sequence



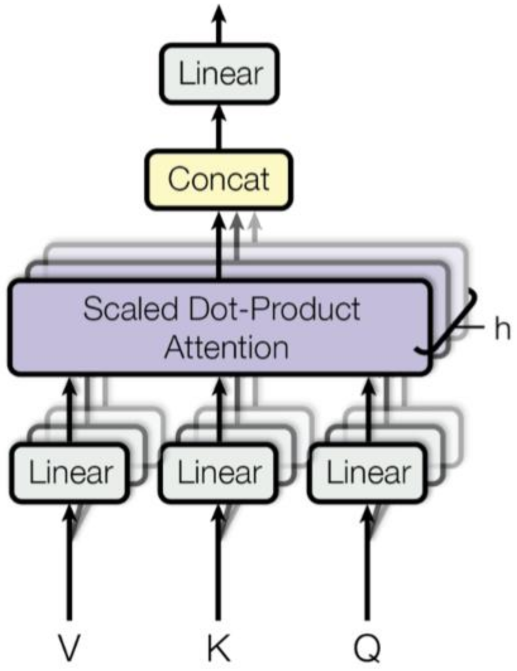
Long short-term memory (LSTM)



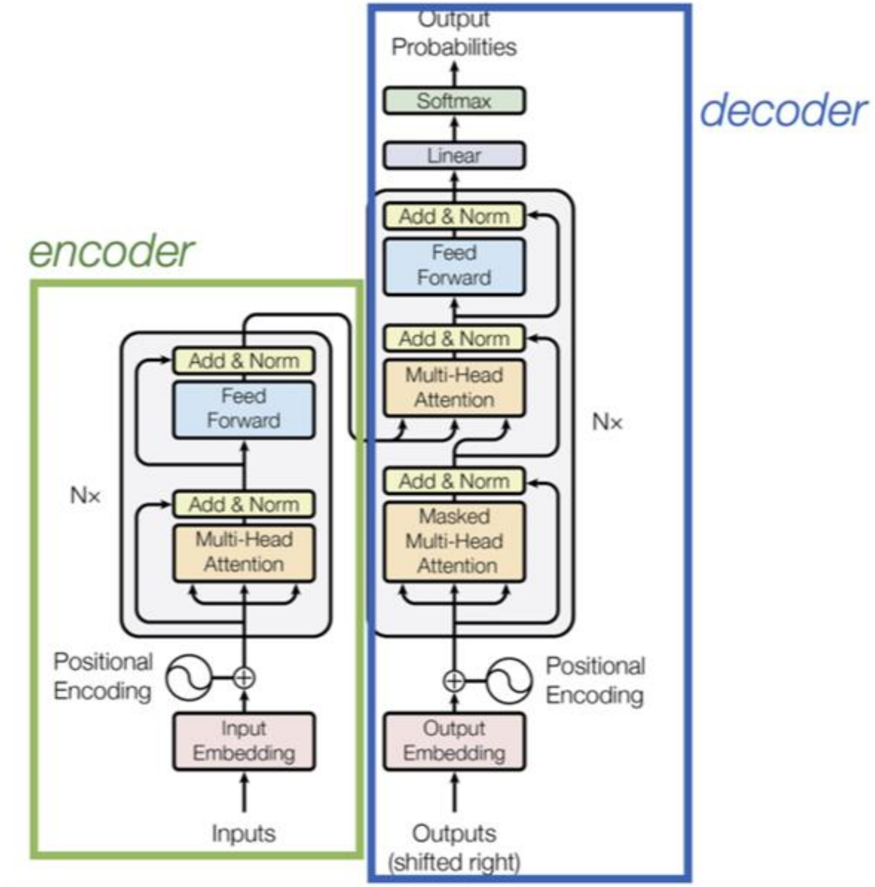
Gated Recurrent Unit (GRU)

Transformer

- ➢ Encoder-decoder architecture based on the multi-head Self-Attention



Multi-Head Self-Attention



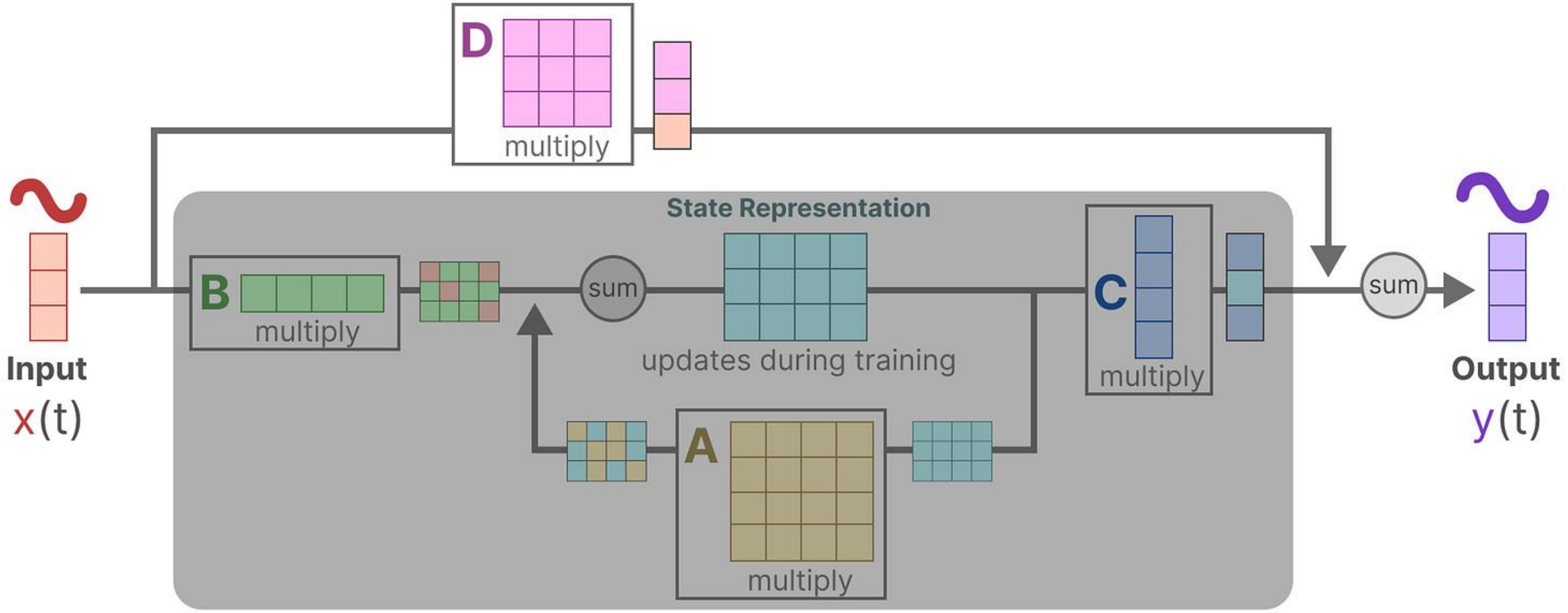
State Space Model (Mamba)

➢ From control theory: model a dynamic system via state representations

State equation $$\mathbf{h'}(t) = \mathbf{A}\mathbf{h}(t) + \mathbf{B}\mathbf{x}(t)$$

Output equation $$\mathbf{y}(t) = \mathbf{C}\mathbf{h}(t) + \mathbf{D}\mathbf{x}(t)$$

Handle very long sequences, generally with a lower number of parameters



State Space Model

Efficiently Modeling Long Sequences with Structured State Spaces.
Mamba: Linear-Time Sequence Modeling with Selective State Spaces.

How to choose different models

- General data: MLP, CNN, ResNet, Transformer

- Sequential data: RNN, Transformer, State Space Models

Outline

- Overview of Machine Learning
- Machine Learning Paradigms
- Model Architectures
- **Machine Learning Systems**
- Applications

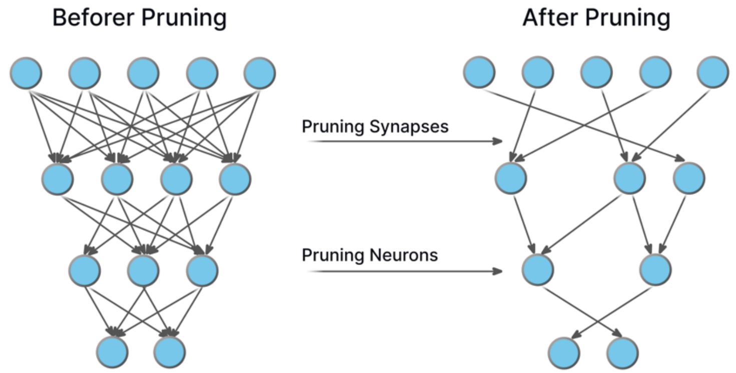
Machine Learning Systems

- Optimizing system performance of ML models

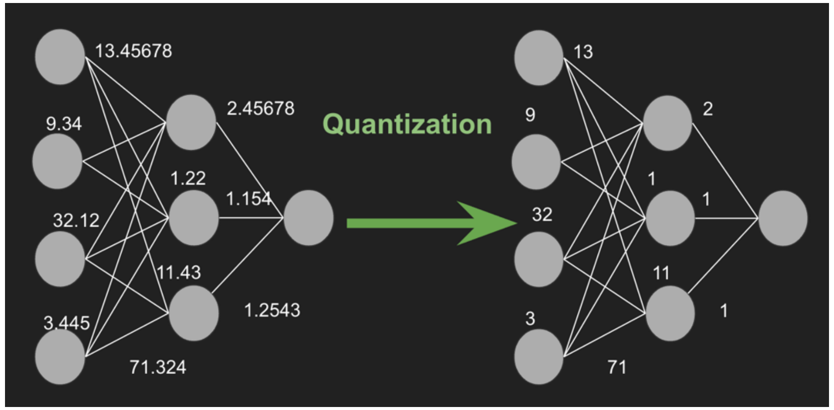
  - Model Compression
  - Parallel and Distributed Computing
  - Hardware Acceleration
  - Inference Optimization

Model Compression

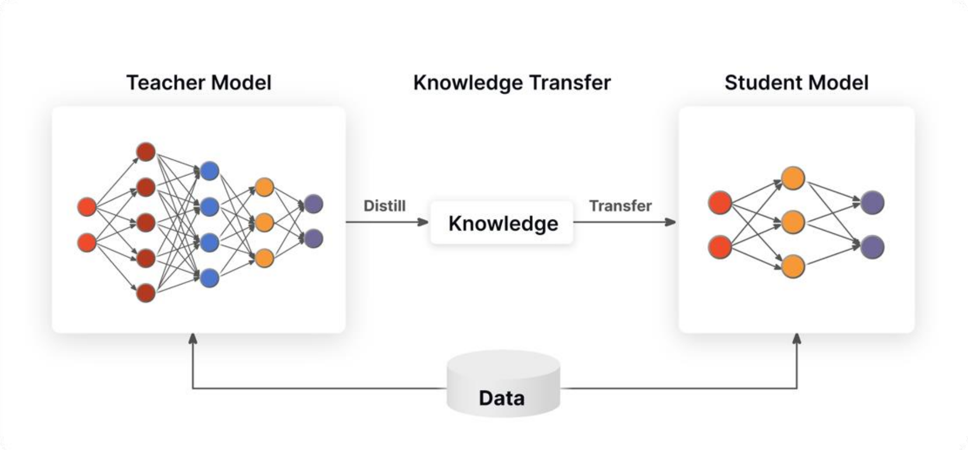
- Pruning



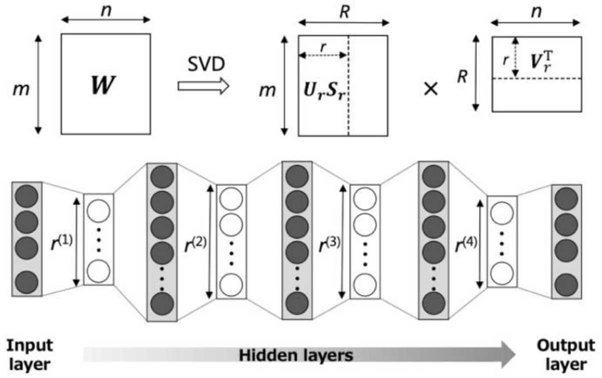
- Quantization



- Knowledge Distillation



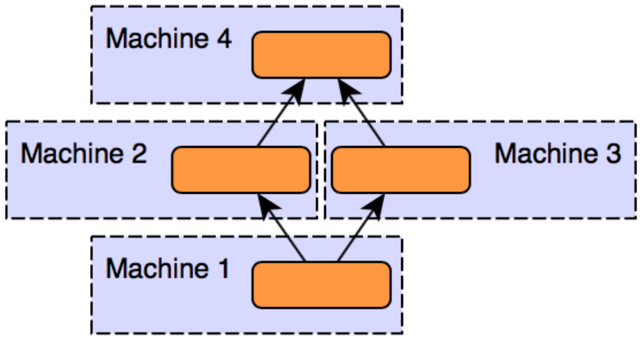
- Low-rank factorization



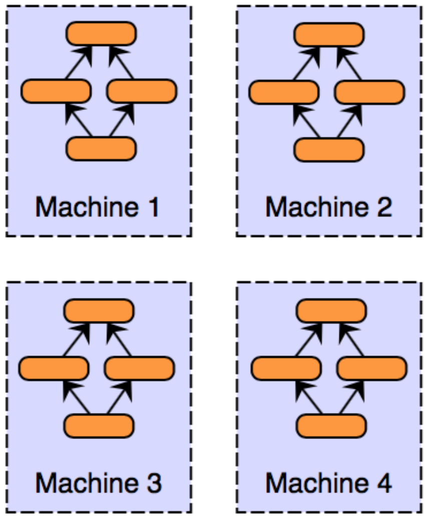
Parallel and Distributed Computing

- ➢ Distributed Training

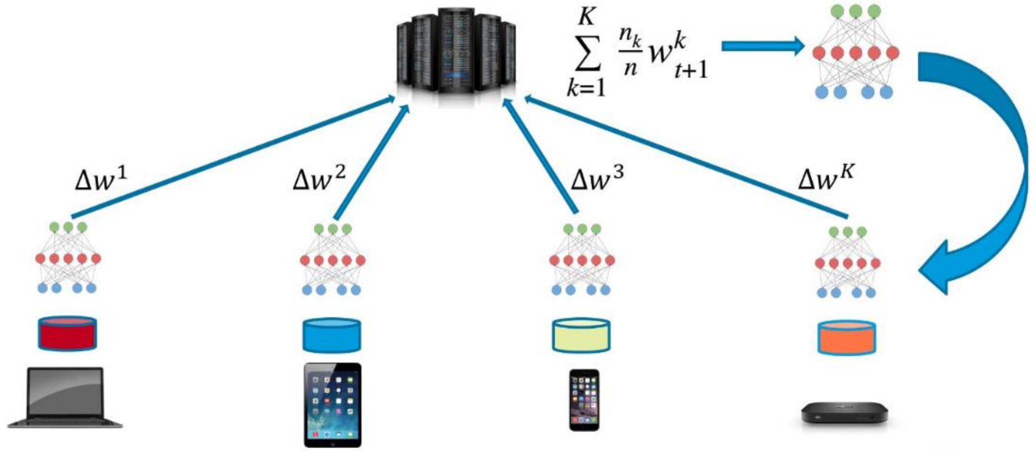
Model Parallelism



Data Parallelism



- ➢ Federated Learning



Hardware Acceleration

- GPUs, TPUs, and FPGAs

CPU
- Small models
- Small datasets
- Useful for design space exploration

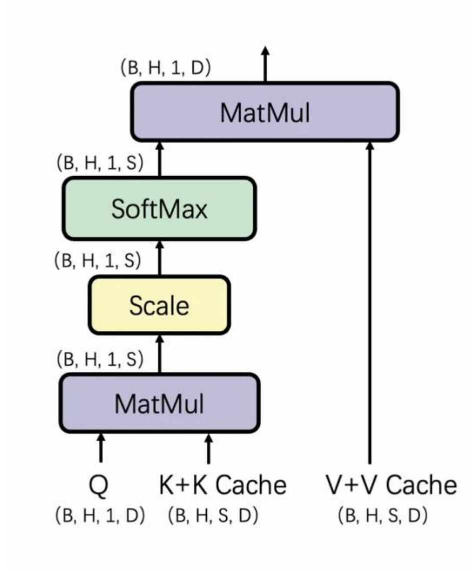
GPU
- Medium-to-large models, datasets
- Image, video processing
- Application on CUDA or OpenCL

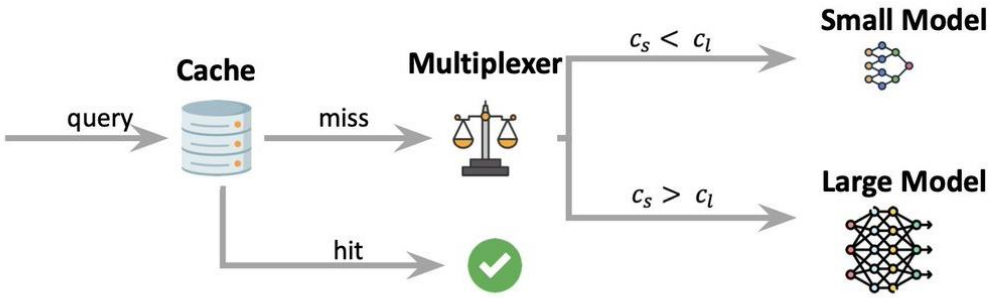
TPU
- Matrix computations
- Dense vector processing
- No custom TensorFlow operations

FPGA
- Large datasets, models
- Compute intensive applications
- High performance, high perf./cost ratio

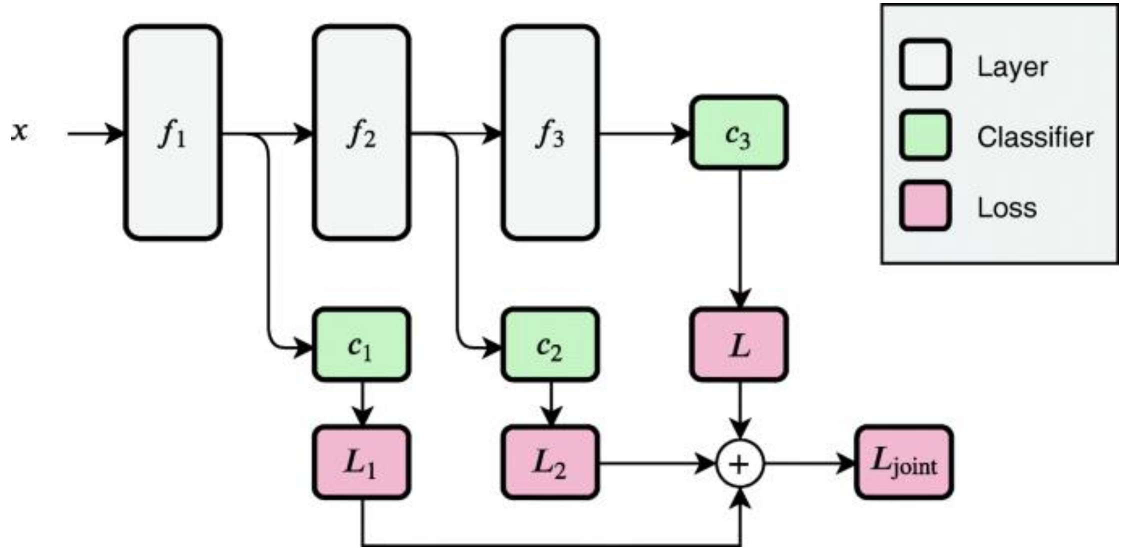
Inference Optimization

- Caching





- Progressive Inference



How to optimize ML systems

- Task requirements: Accuracy, Latency
- Resource constraints: Memory, Energy

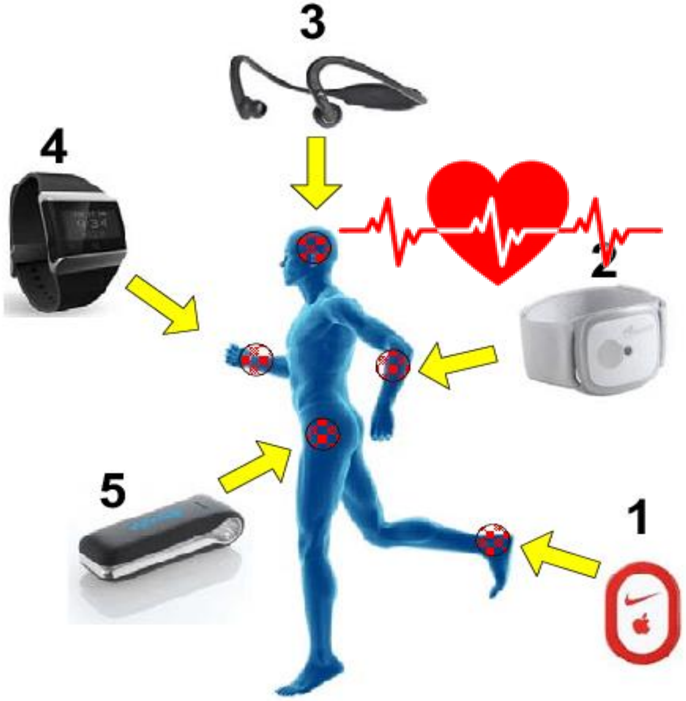
    - Model Compression
    - Parallel and Distributed Computing
    - Hardware Acceleration
    - Inference Optimization

Outline

- Overview of Machine Learning
- Machine Learning Paradigms
- Model Architectures
- Machine Learning Systems
- **Applications**

Smart Health

- Behavior monitoring, early disease diagnosis, personalized intervention



Fitness Tracking

Sleep Monitoring

Cognition Impairment Detection

Smart Home & Building

- Occupant detection, environment monitoring, localization, adaptive control

Autonomous Driving

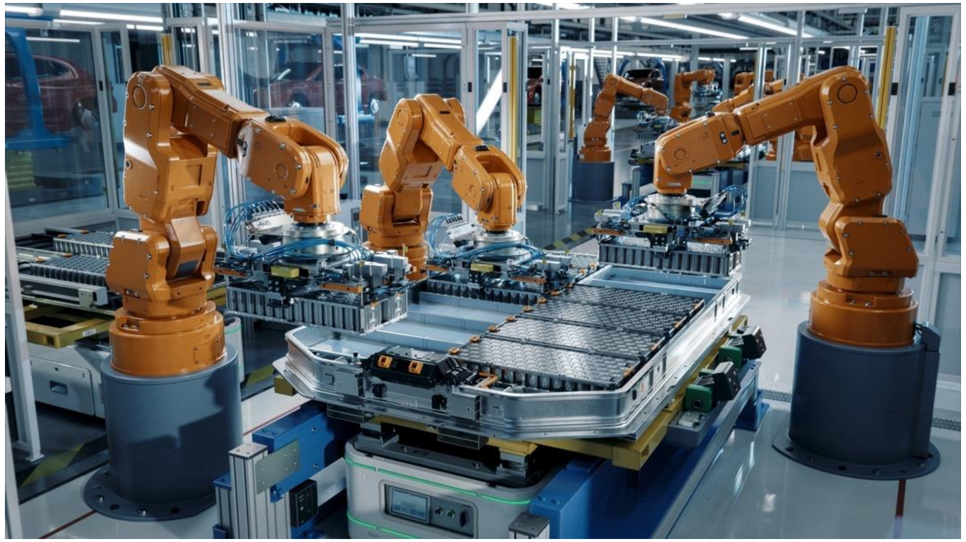
➤ Object detection, control

Smart City

➢ Traffic management, sustainability, public security

Other Applications

- Smart Manufacturing



- Smart Agriculture

Break

- **Next lecture: Challenges in Embedded AI Systems**
- **Website:**
  - **A shared spreadsheet for paper pre to be released on Weekends**
  - **Course APP and dataset to be released next Tuesday**
- **Any questions?**