



UNIVERSIDADE
CATÓLICA
PORTUGUESA

BRAGA

Deep Learning

Session 26

Multi Topic Overview

Applied Data Science

2024/2025



Multimodal Neural Networks

What is Multimodal Learning?

- In general, learning that involves multiple modalities
- This can manifest itself in different ways:
 - Input is one modality, output is another
 - Multiple modalities are learned jointly
 - One modality assists in the learning of another
 - ...

Multimodal Data

- Data is usually a collection of modalities



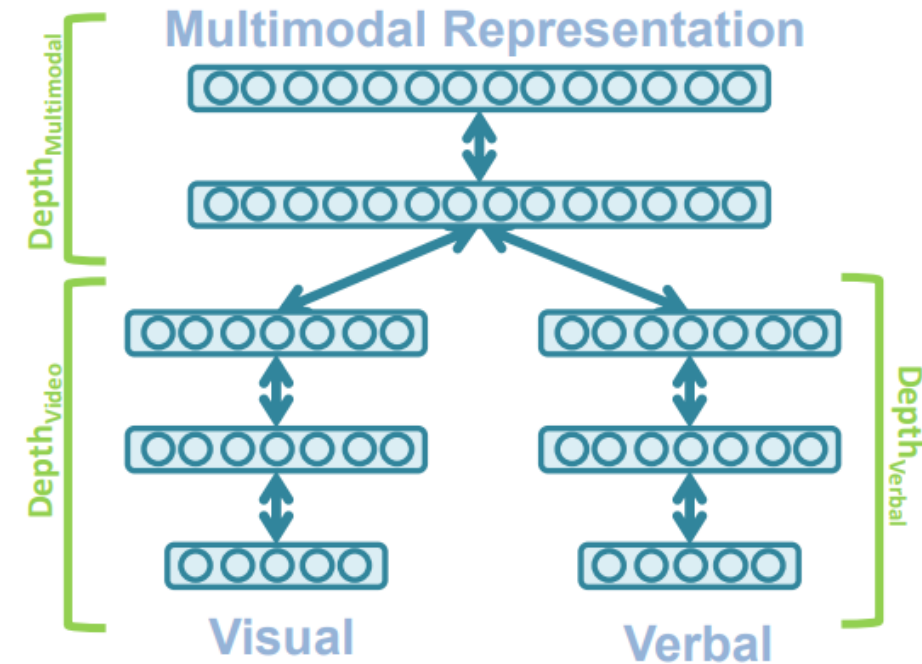
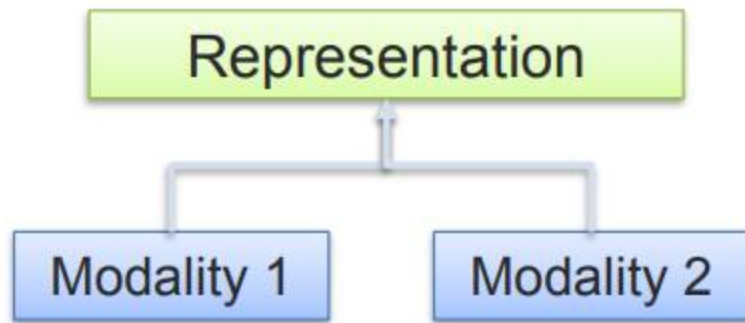
Sunset Pacific Ocean
Nikon D40 Baker Beach
San Francisco
Top20SunsetsOfOurHearts
California seashore
ocean



Why is Multimodal Learning Hard?

- Different representations
 - Learning how to represent and summarize multimodal data in away that exploits the complementarity and redundancy.

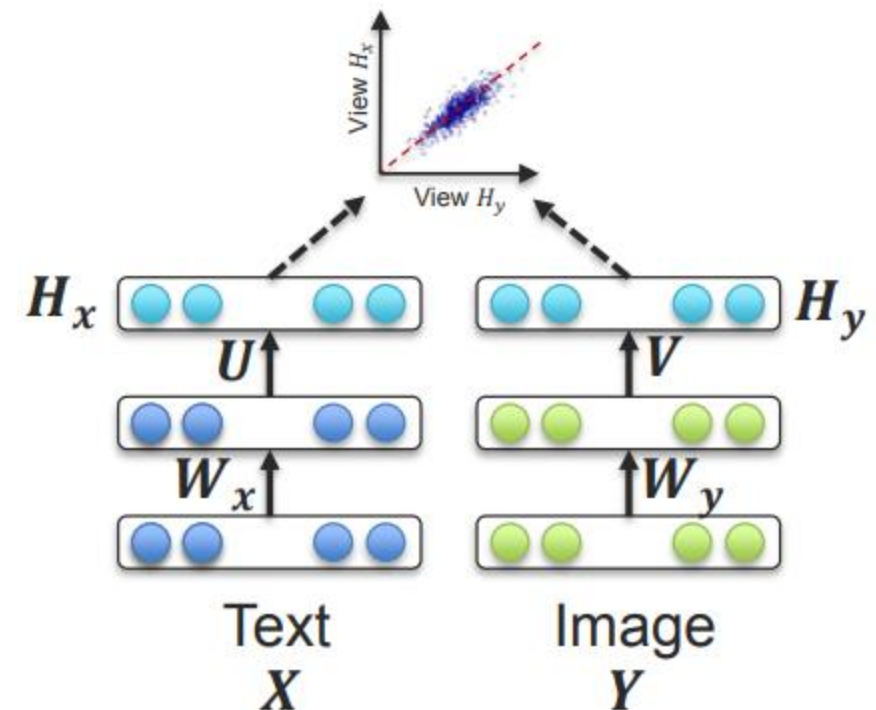
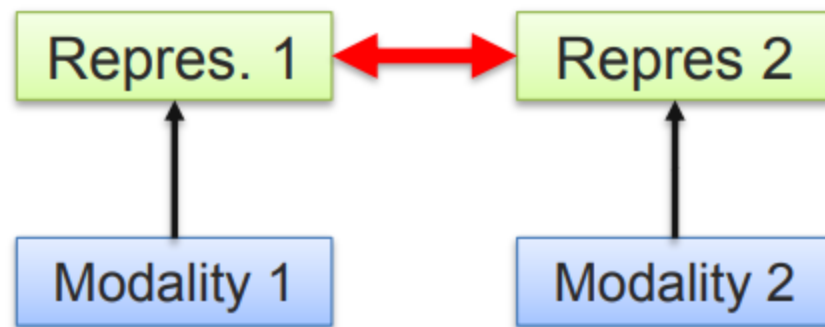
Joint representations:



Why is Multimodal Learning Hard?

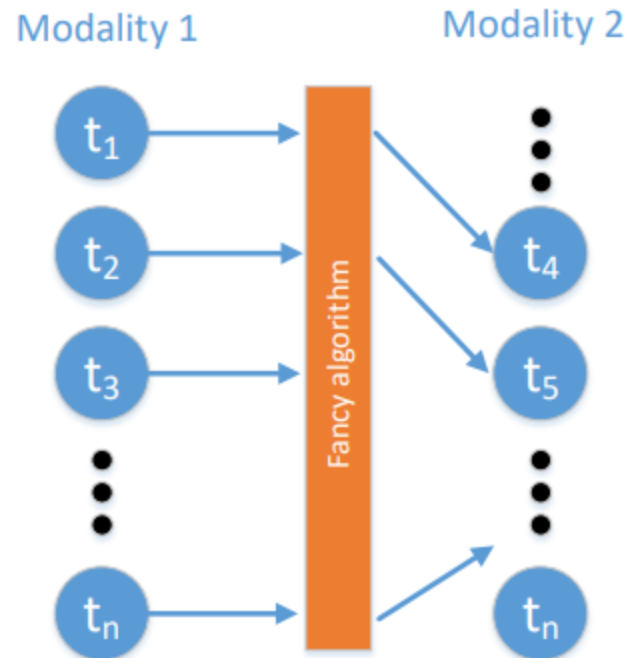
- Different representations
 - Learning how to represent and summarize multimodal data in away that exploits the complementarity and redundancy.

Coordinated representations:



Why is Multimodal Learning Hard?

- Alignment
 - Identify the direct relations between (sub)elements from two or more different modalities.



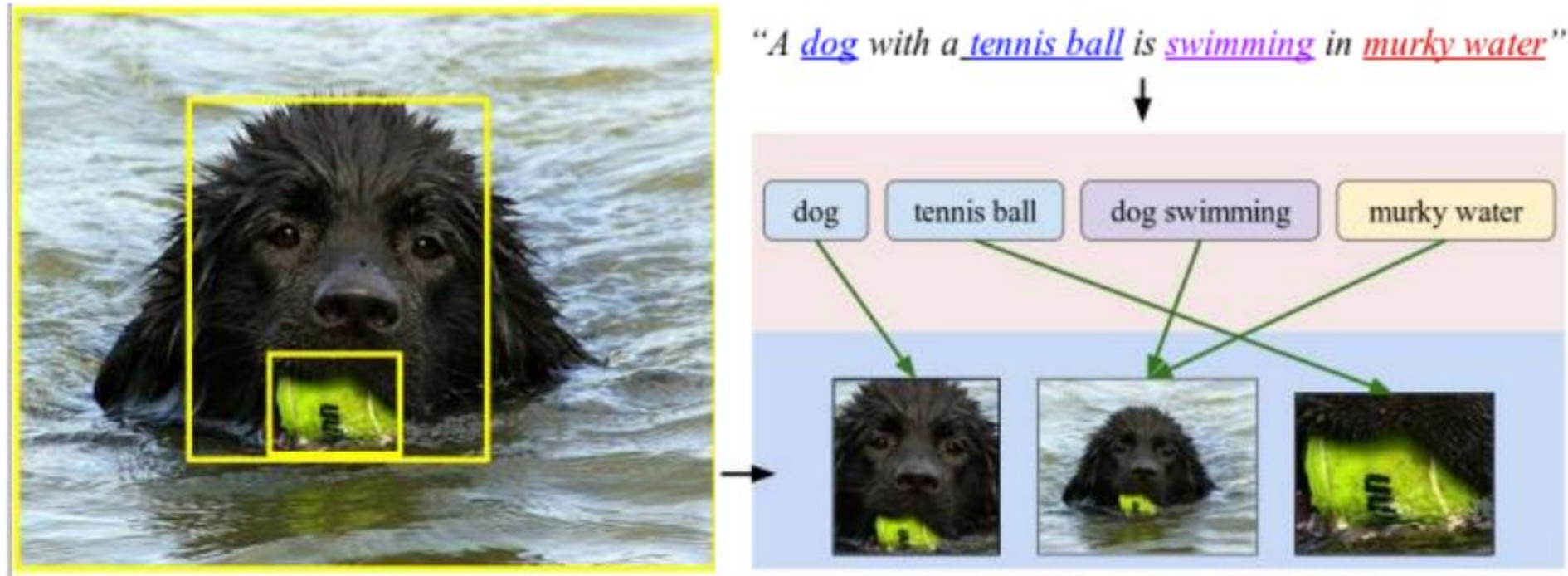
Explicit Alignment

The goal is to directly find correspondences between elements of different modalities

Implicit Alignment

Uses internally latent alignment of modalities in order to better solve a different problem

Implicit Alignment

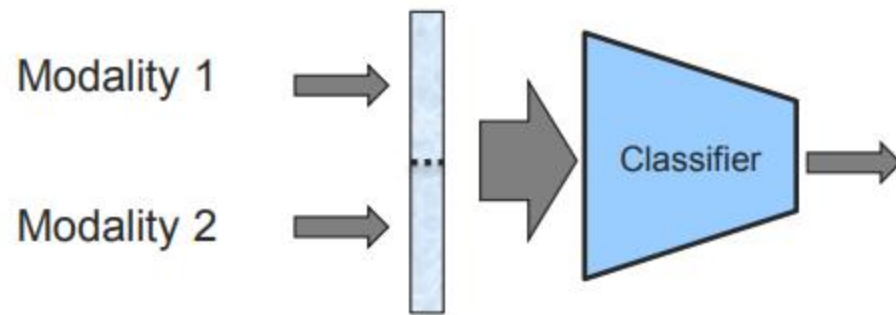


Why is Multimodal Learning Hard?

- Fusion
 - To join information from two or more modalities to perform a prediction task.

Model-Agnostic Approaches

1) Early Fusion



2) Late Fusion

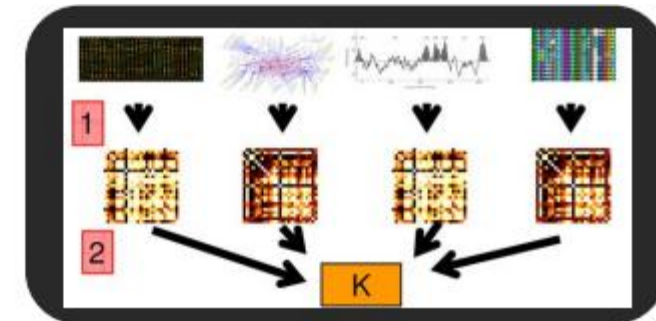


Why is Multimodal Learning Hard?

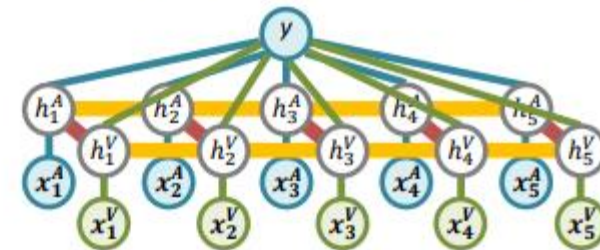
- Fusion
 - To join information from two or more modalities to perform a prediction task.

Model-Based (Intermediate) Approaches

- 1) Deep neural networks
- 2) Kernel-based methods
- 3) Graphical models



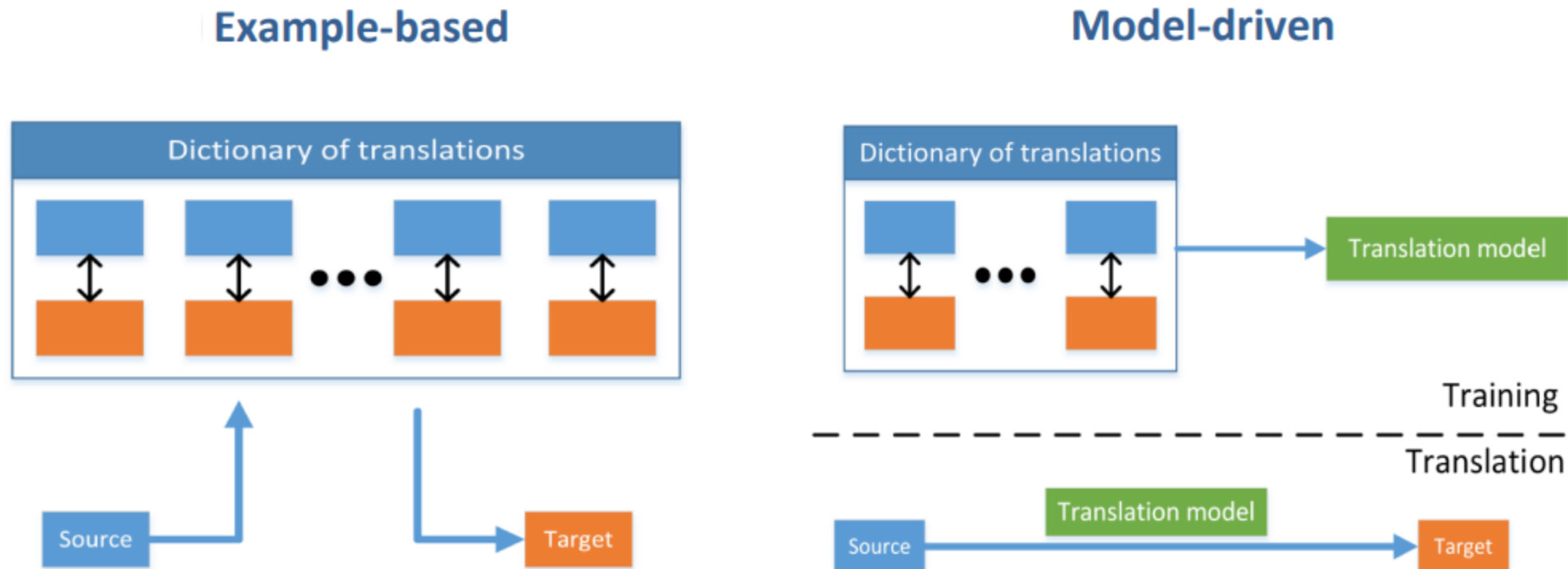
Multiple kernel learning



Multi-View Hidden CRF

Why is Multimodal Learning Hard?

- Translation
 - Process of changing data from one modality to another, where the translation relationship can often be open-ended or subjective.



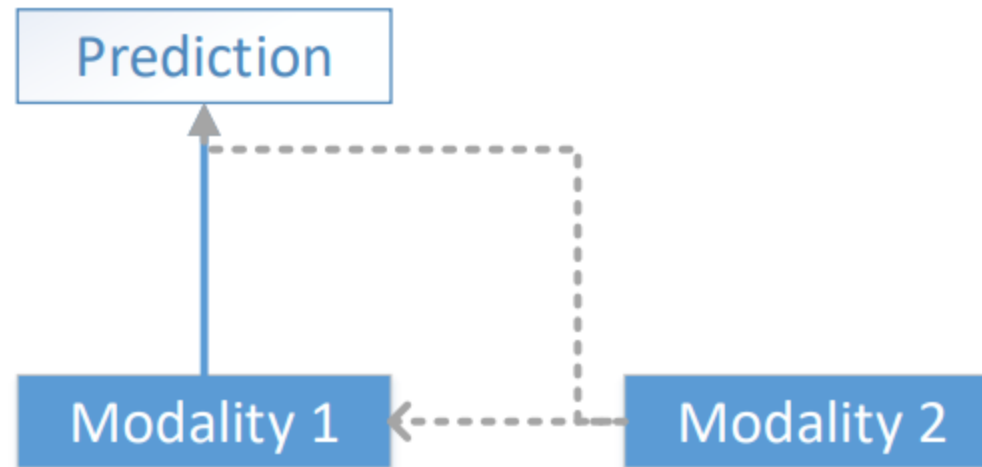
Why is Multimodal Learning Hard?

- Translation
 - Process of changing data from one modality to another, where the translation relationship can often be open-ended or subjective.



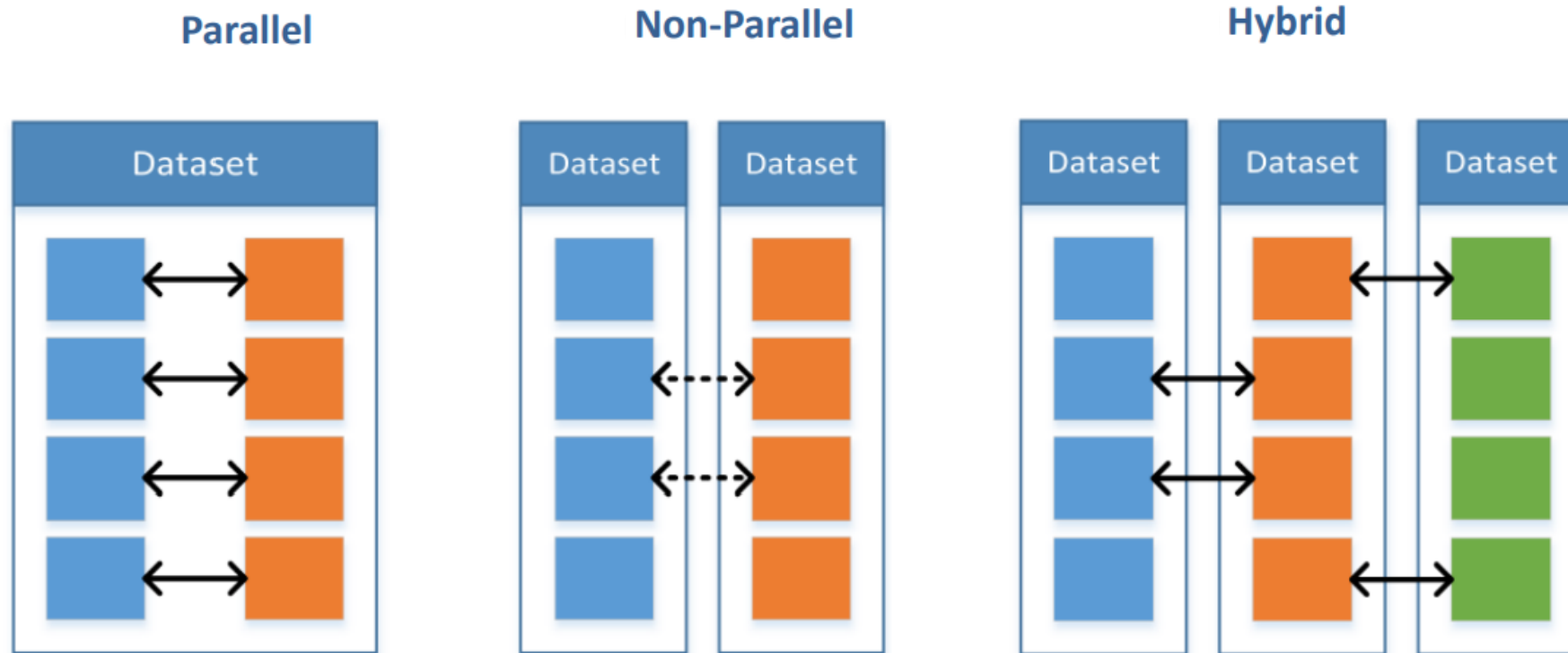
Why is Multimodal Learning Hard?

- Co-Learning:
 - Transfer knowledge between modalities, including their representations and predictive models.



Why is Multimodal Learning Hard?

- Co-Learning:
 - Transfer knowledge between modalities, including their representations and predictive models.



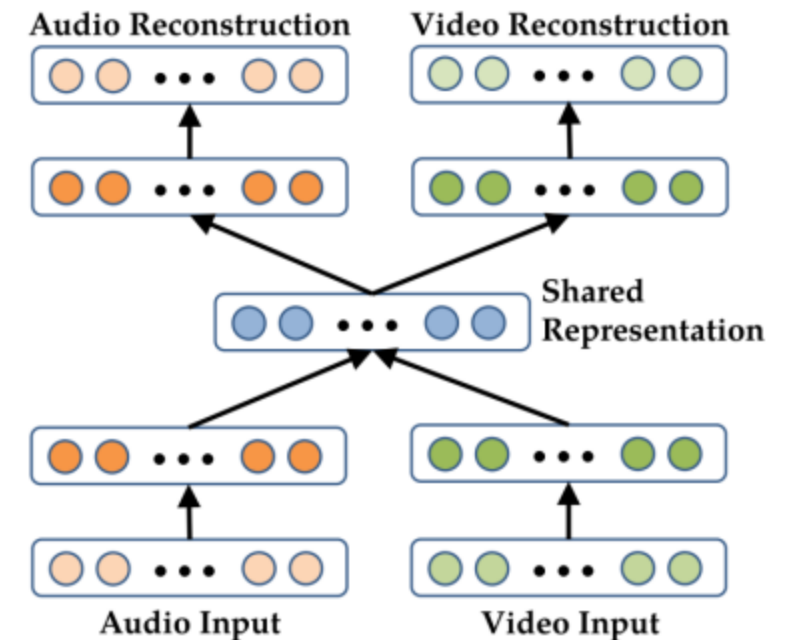
Multimodal Applications



	CHALLENGES				
APPLICATIONS	REPRESENTATION	TRANSLATION	FUSION	ALIGNMENT	CO-LEARNING
Speech Recognition and Synthesis					
Audio-visual Speech Recognition	✓		✓	✓	✓
(Visual) Speech Synthesis	✓	✓			
Event Detection					
Action Classification	✓		✓		✓
Multimedia Event Detection	✓		✓		✓
Emotion and Affect					
Recognition	✓		✓	✓	✓
Synthesis	✓	✓			
Media Description					
Image Description	✓	✓		✓	✓
Video Description	✓	✓	✓	✓	✓
Visual Question-Answering	✓		✓	✓	✓
Media Summarization	✓	✓	✓		
Multimedia Retrieval					
Cross Modal retrieval	✓	✓		✓	✓
Cross Modal hashing	✓				✓

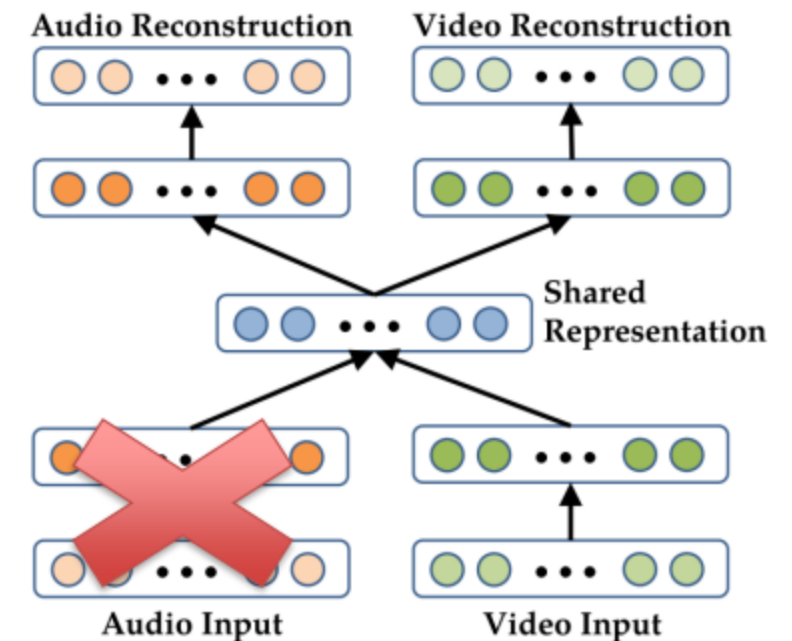
Deep Multimodal Autoencoders

- A deep representation learning approach
- A bimodal auto-encoder
 - Used for Audio-visual speech recognition



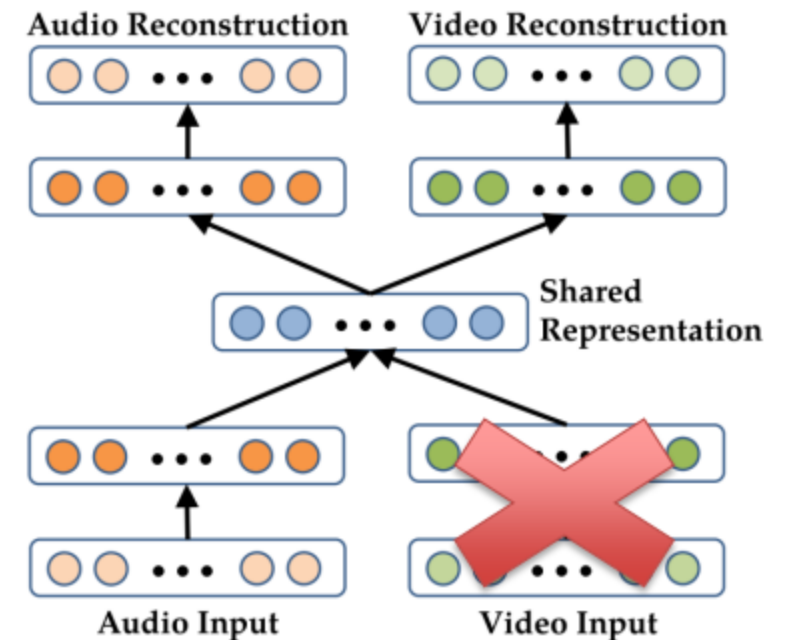
Deep Multimodal Autoencoders

- A deep representation learning approach
- A bimodal auto-encoder
 - Used for Audio-visual speech recognition



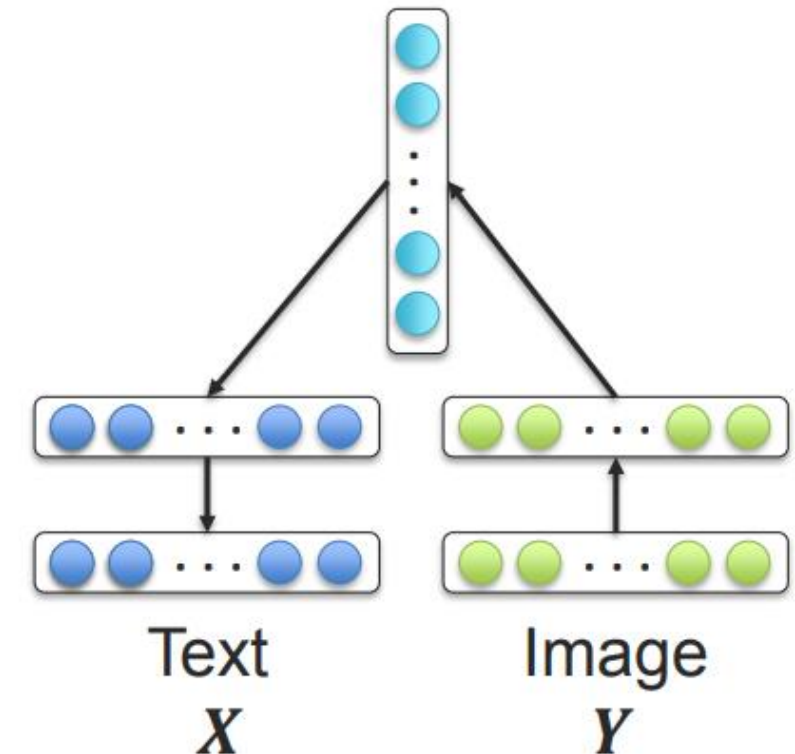
Deep Multimodal Autoencoders

- A deep representation learning approach
- A bimodal auto-encoder
 - Used for Audio-visual speech recognition



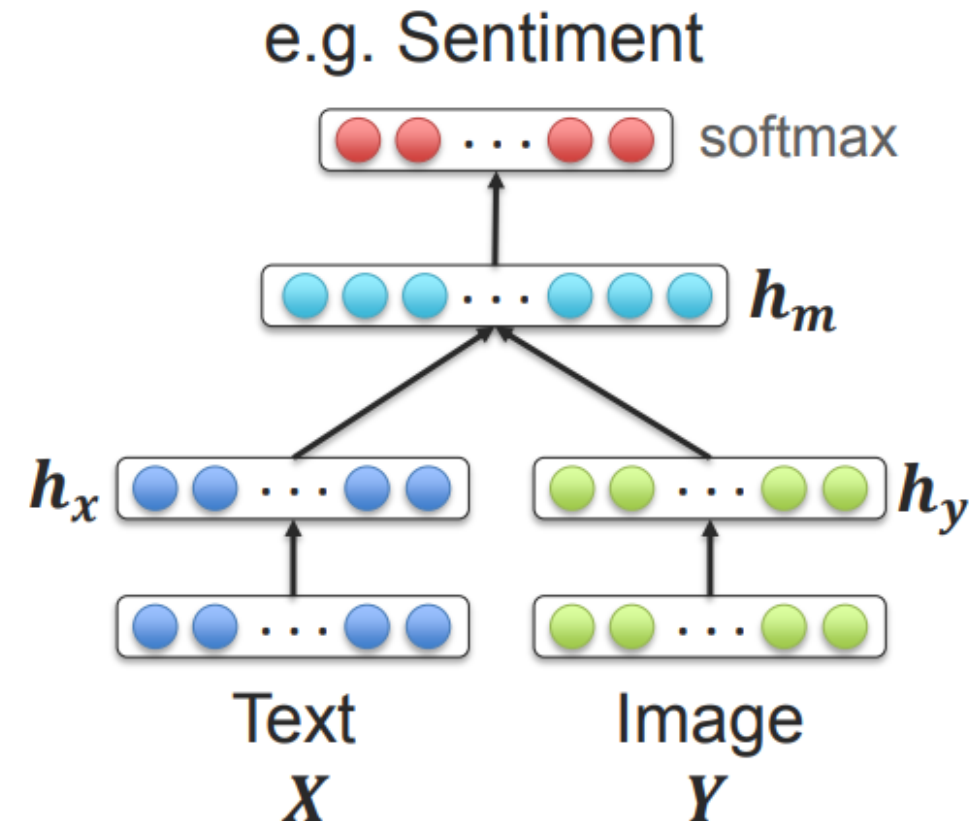
Multimodal Encoder-Decoder

- Visual modality often encoded using CNN
- Language modality will be decoded using LSTM
 - A simple multilayer perceptron will be used to translate from visual (CNN) to language (LSTM)



Multimodal Joint Representation

- For supervised learning tasks
- Joining the unimodal representations
 - Simple concatenation
 - Element-wise multiplication or summation
 - Multilayer perceptron



Multimodal Sentiment Analysis

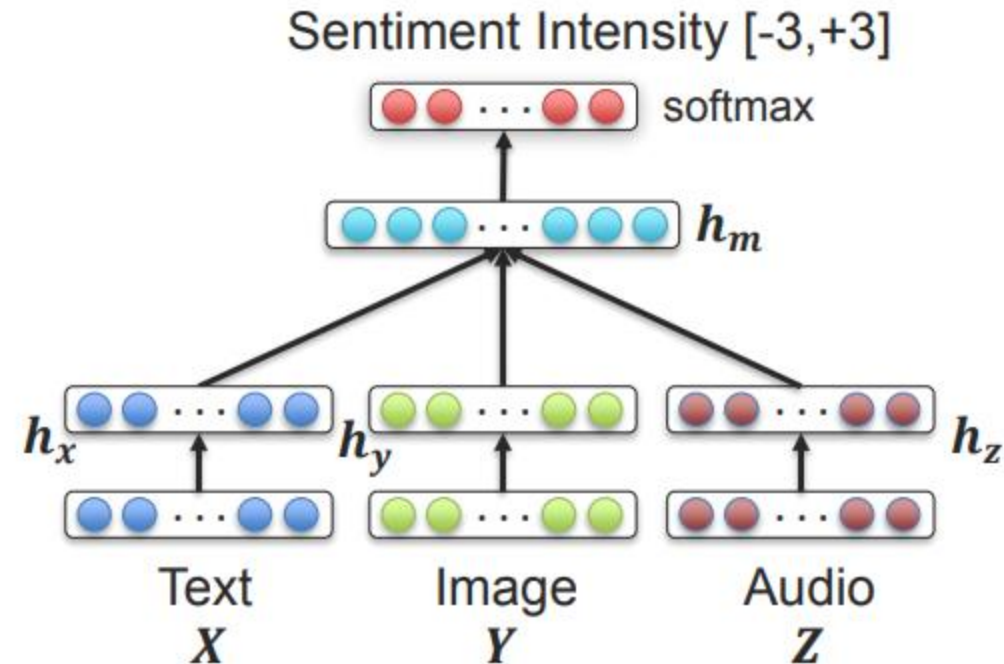
MOSI dataset (Zadeh et al, 2016)



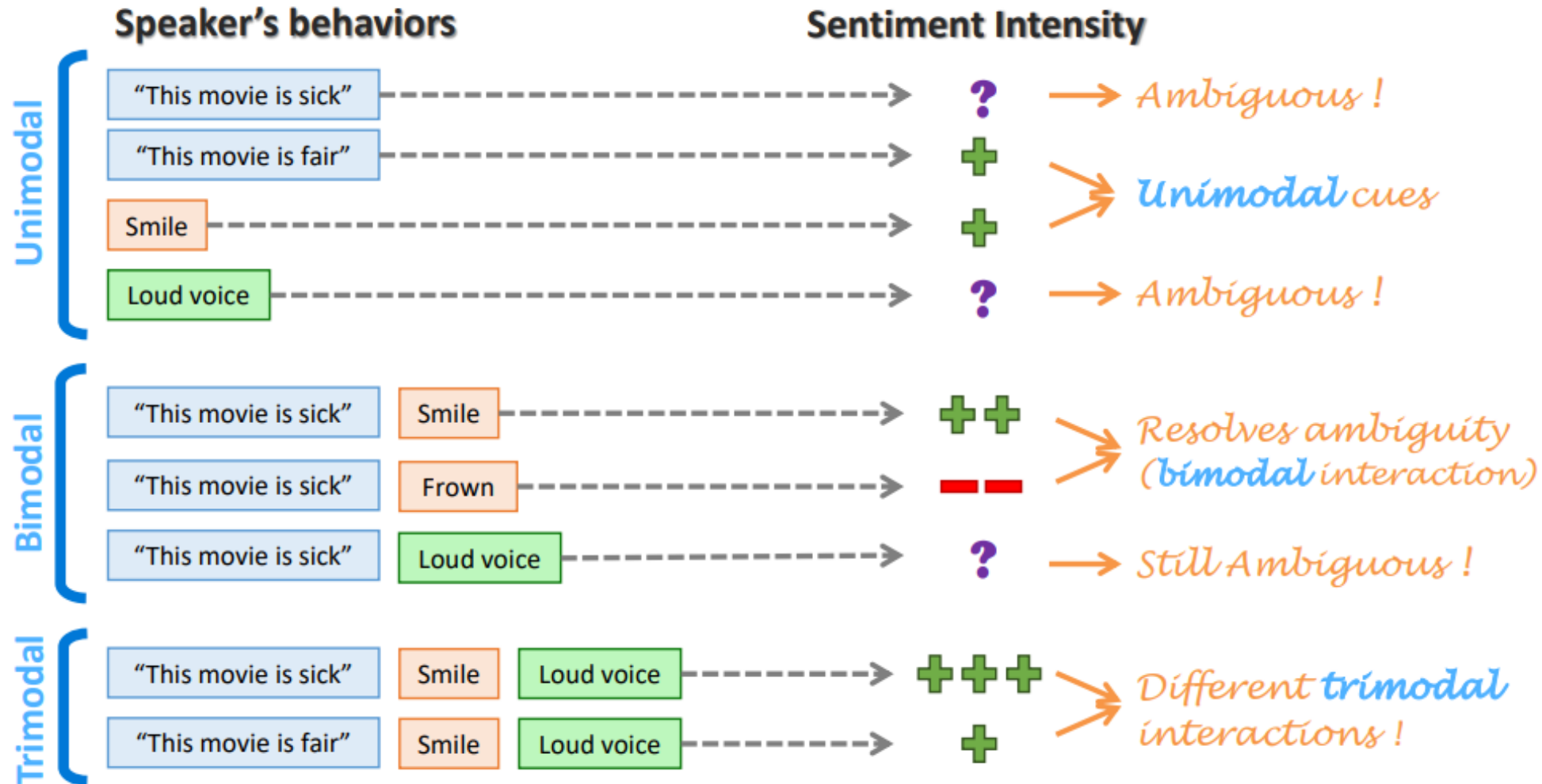
- 2199 subjective video segments
- Sentiment intensity annotations
- 3 modalities: text, video, audio

Multimodal joint representation:

$$h_m = f(W \cdot [h_x, h_y, h_z])$$



Unimodal, Bimodal and Trimodal Interactions







Transfer Learning

Transfer Learning


- Goal: : Avoid Always Relying on Large Labeled Datasets



- Expensive
- Relatively Slow to Build Dataset



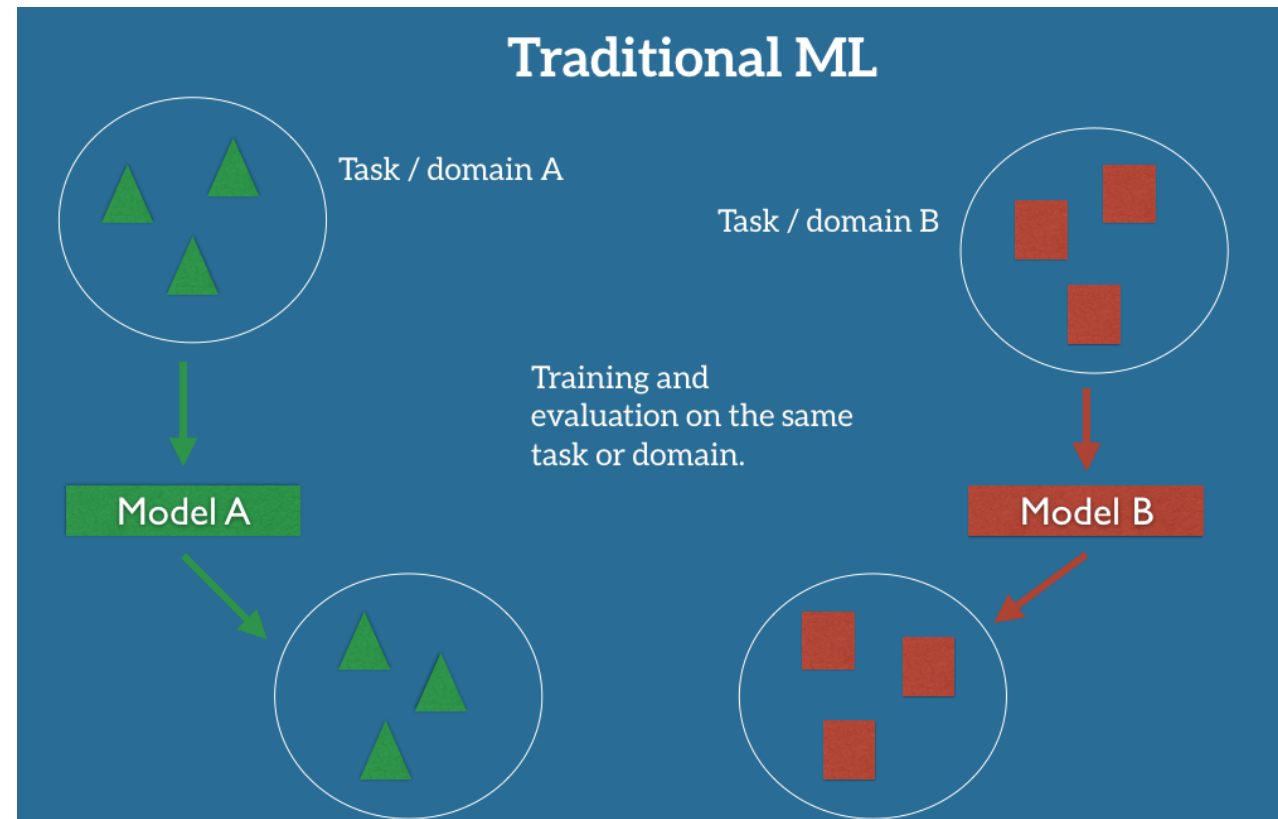
Places (2014) MS COCO (2014) Visual Genome (2016)



Places (2014) MS COCO (2014) Visual Genome (2016)

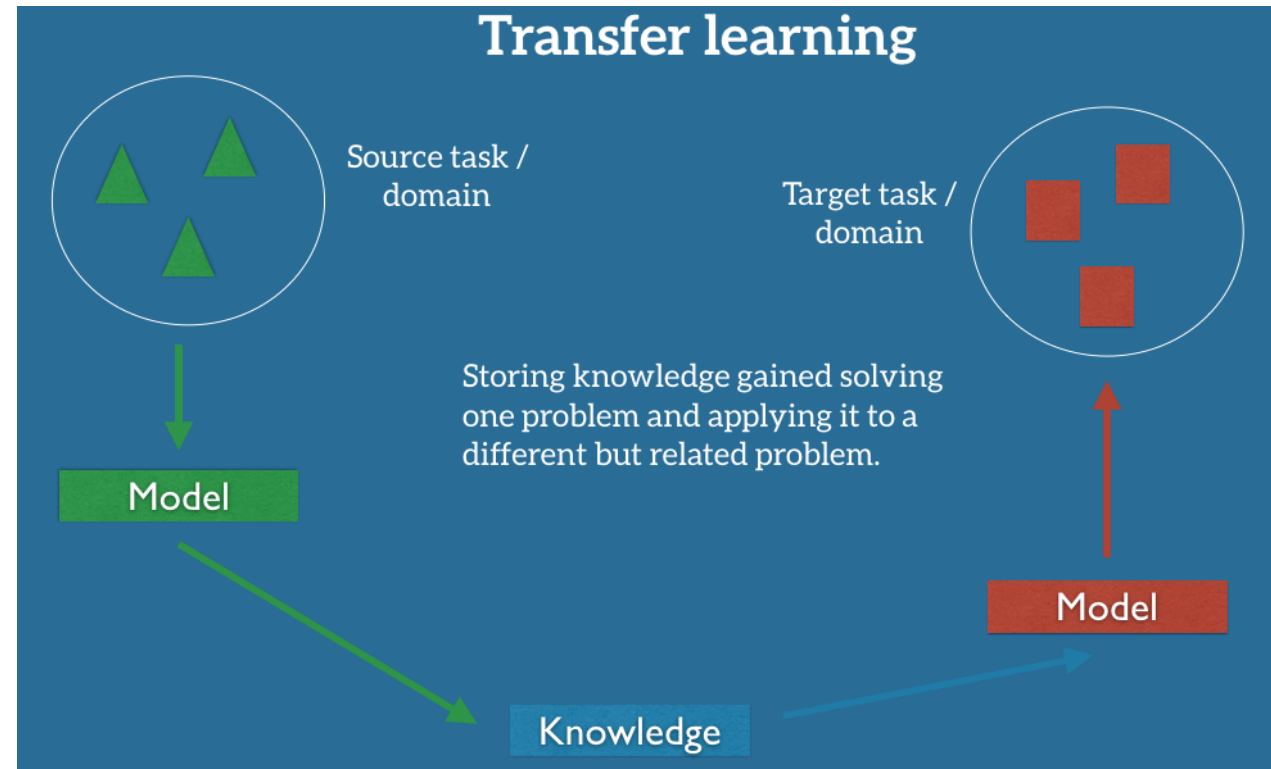
Transfer Learning

- Rather than Learn Solution from Scratch For Each Task/Domain Pair ...



Transfer Learning

- ... Improve the Learning for Conditions Not Observed During Training.



Transfer Learning

- Transfer Learning When Data Sampling Changes (e.g., Sentiment Classification)



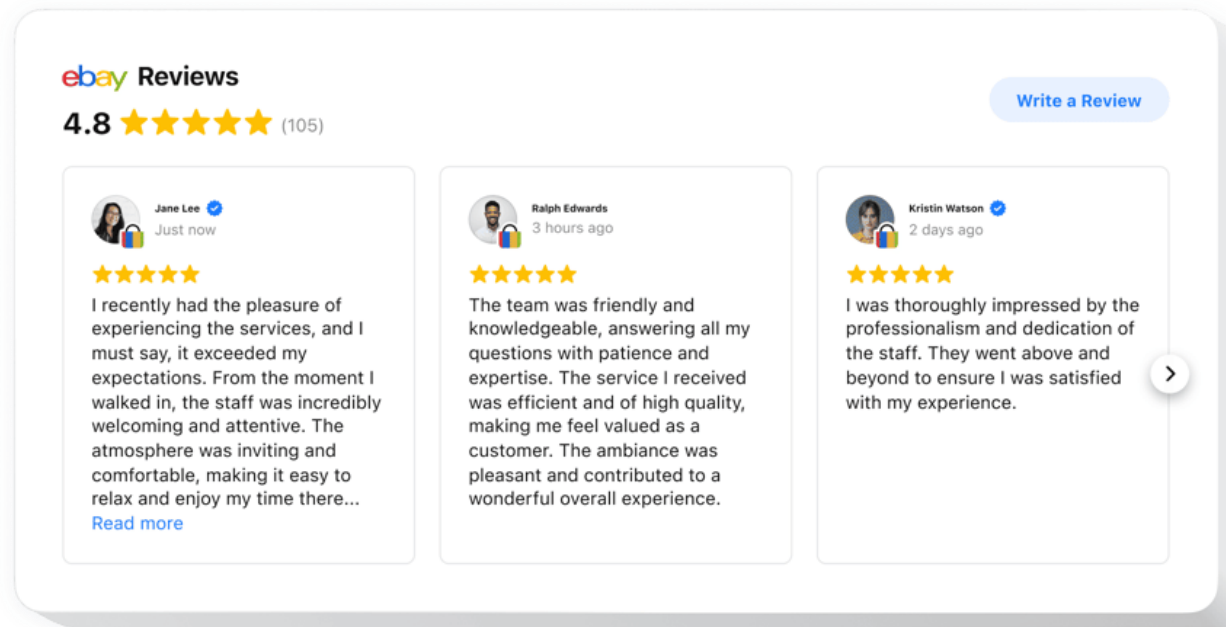
News (formal and lengthy)



Tweets (informal and brief)

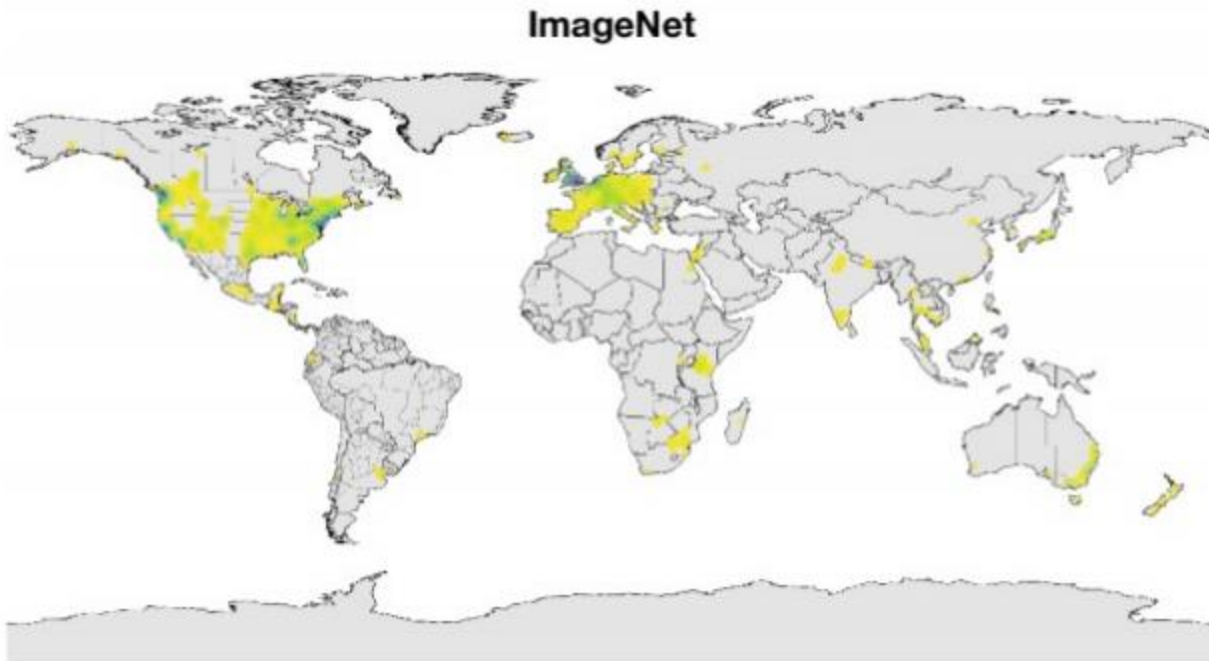
Transfer Learning

- Transfer Learning When Feature Space Changes (e.g., Sentiment Classification in Different Language)



Transfer Learning

- Transfer Learning When Target Categories Change (e.g., Items in Low Income Household vs ImageNet)



Ground truth: Soap

Nepal, 288 \$/month

Azure: food, cheese, bread, cake, sandwich

Clarifai: food, wood, cooking, delicious, healthy

Google: food, dish, cuisine, comfort food, spam

Amazon: food, confectionary, sweets, burger

Watson: food, food product, turmeric, seasoning

Tencent: food, dish, matter, fast food, nutriment



Ground truth: Soap

UK, 1890 \$/month

Azure: toilet, design, art, sink

Clarifai: people, faucet, healthcare, lavatory, wash closet

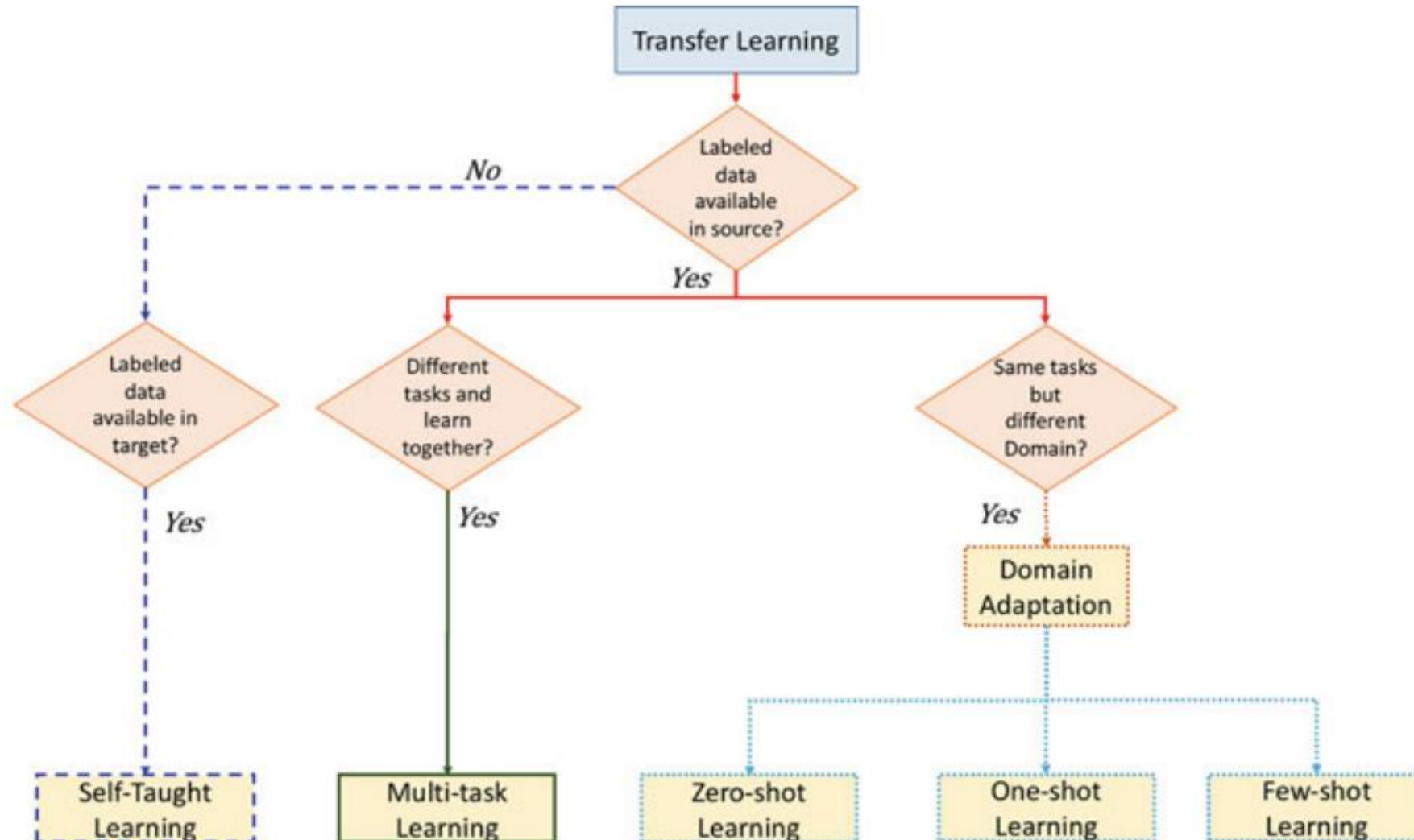
Google: product, liquid, water, fluid, bathroom accessory

Amazon: sink, indoors, bottle, sink faucet

Watson: gas tank, storage tank, toiletry, dispenser, soap dispenser

Tencent: lotion, toiletry, soap dispenser, dispenser, after shave

Transfer Learning Approaches



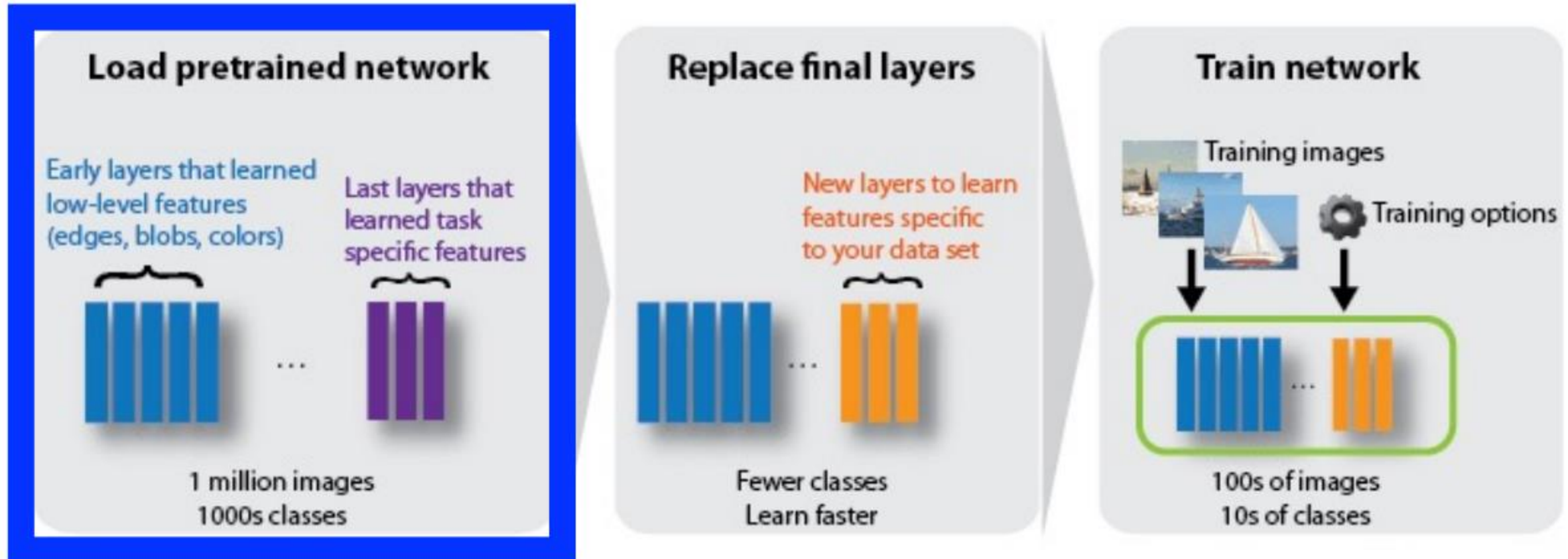
Transfer Learning: Key Challenges

- **What to transfer?** i.e., what knowledge generalizes
- **How to transfer?**
- **When to transfer?** i.e., transferring knowledge can harm performance

Transfer Learning: Self-Supervised

- Goal: create generalizable features

Key observation: features from a pretrained network can be useful for other datasets/tasks



Transfer Learning: Self-Supervised

- How Do Humans Learn?

With Supervision

Learn from instruction



Unsupervised

Learn from experience



Self-Supervised Learning

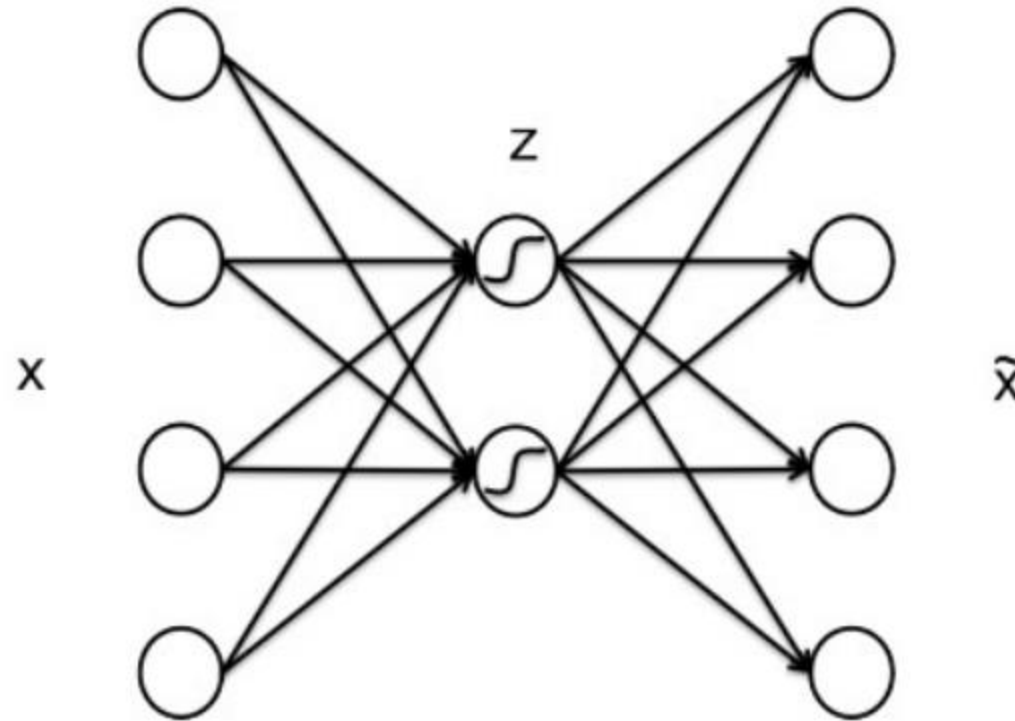
- Data Gives Supervision

- Relatively Cheap
- Can Collect Data Fast



Autoencoders

- Learn to copy the input to the output



Autoencoders

- Consists of two parts:
 - Encoder: compresses the input to a internal representation
 - Decoder: tries to reconstruct the input from the internal representation

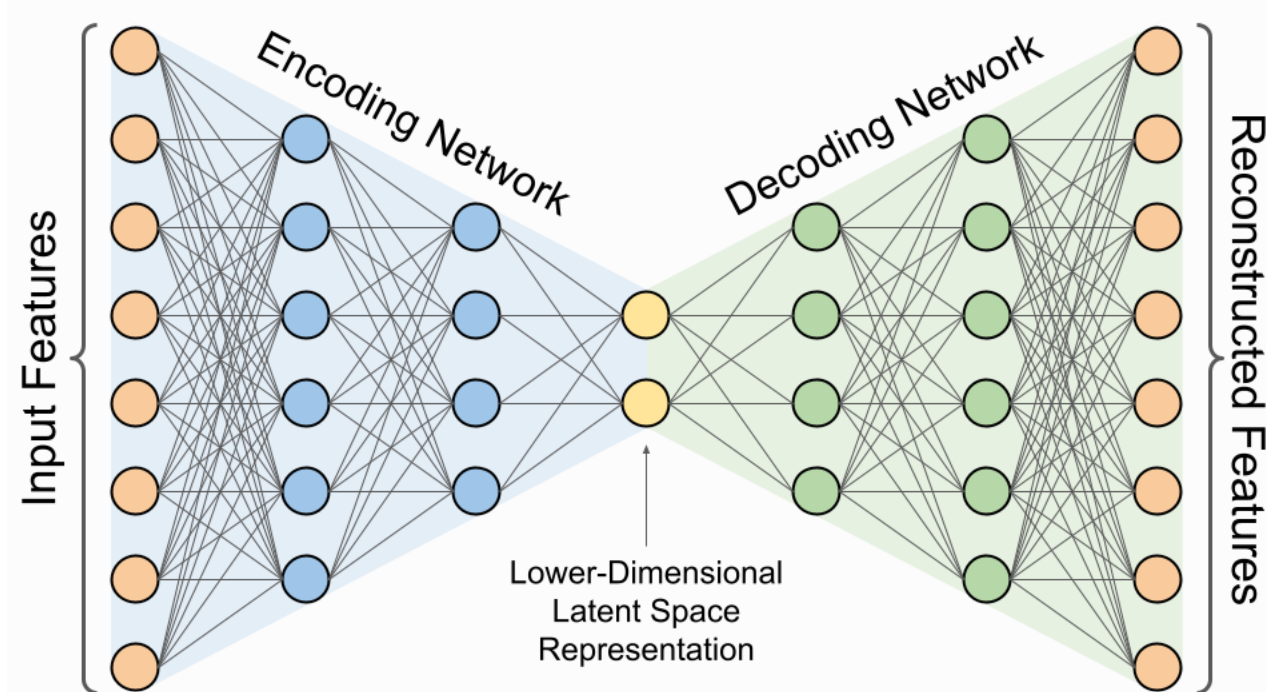


Image Autoencoder Architecture

- Given this input 620 x 426 image (264,120 pixels):



- What would a perfect autoencoder predict?

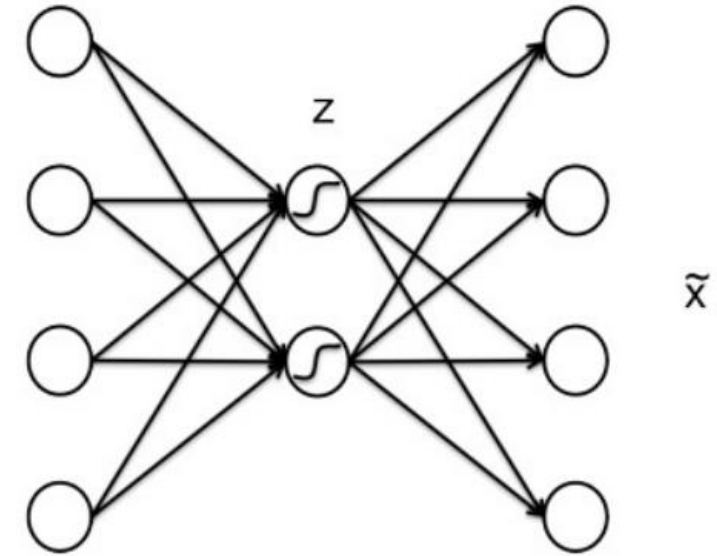


Image Autoencoder Architecture

- Given this input 620 x 426 image (264,120 pixels):



- What would a perfect autoencoder predict?
 - itself

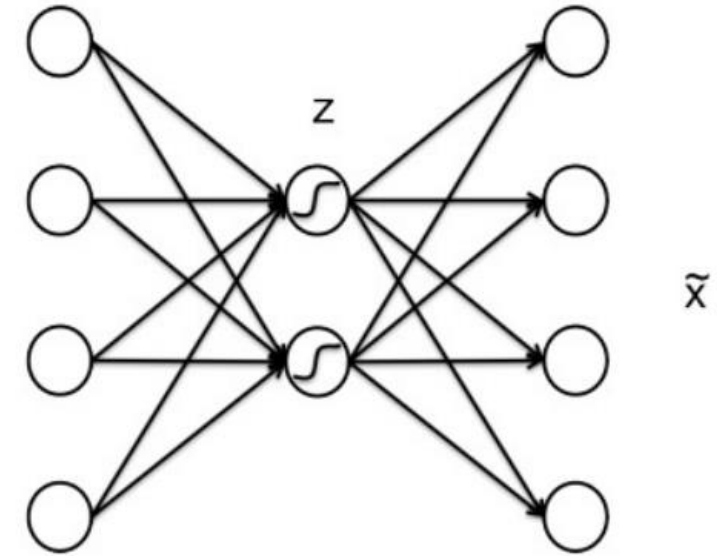


Image Autoencoder Architecture

- Given this input 620 x 426 image (264,120 pixels):



- What number of nodes are in the final layer?

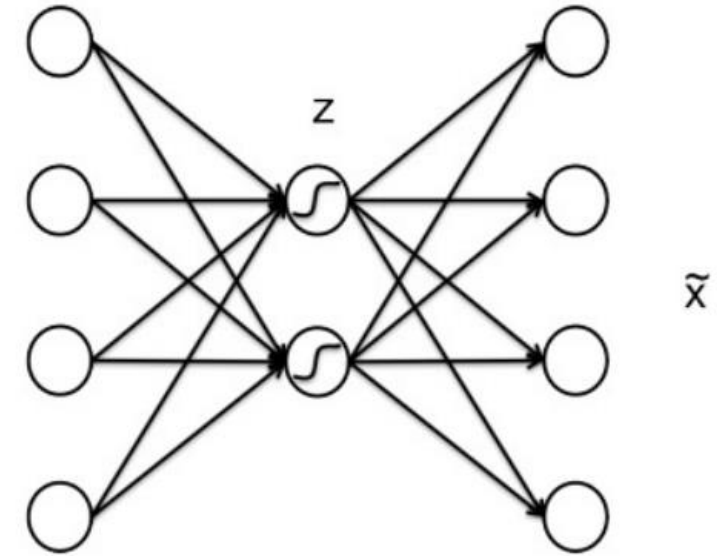
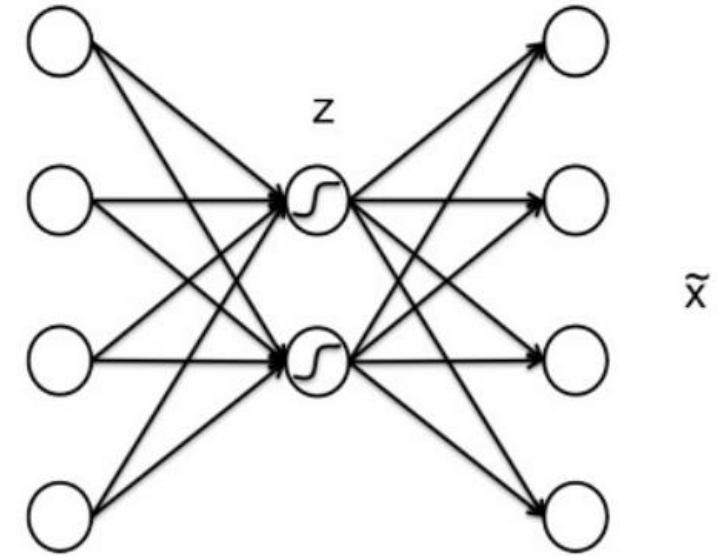


Image Autoencoder Architecture

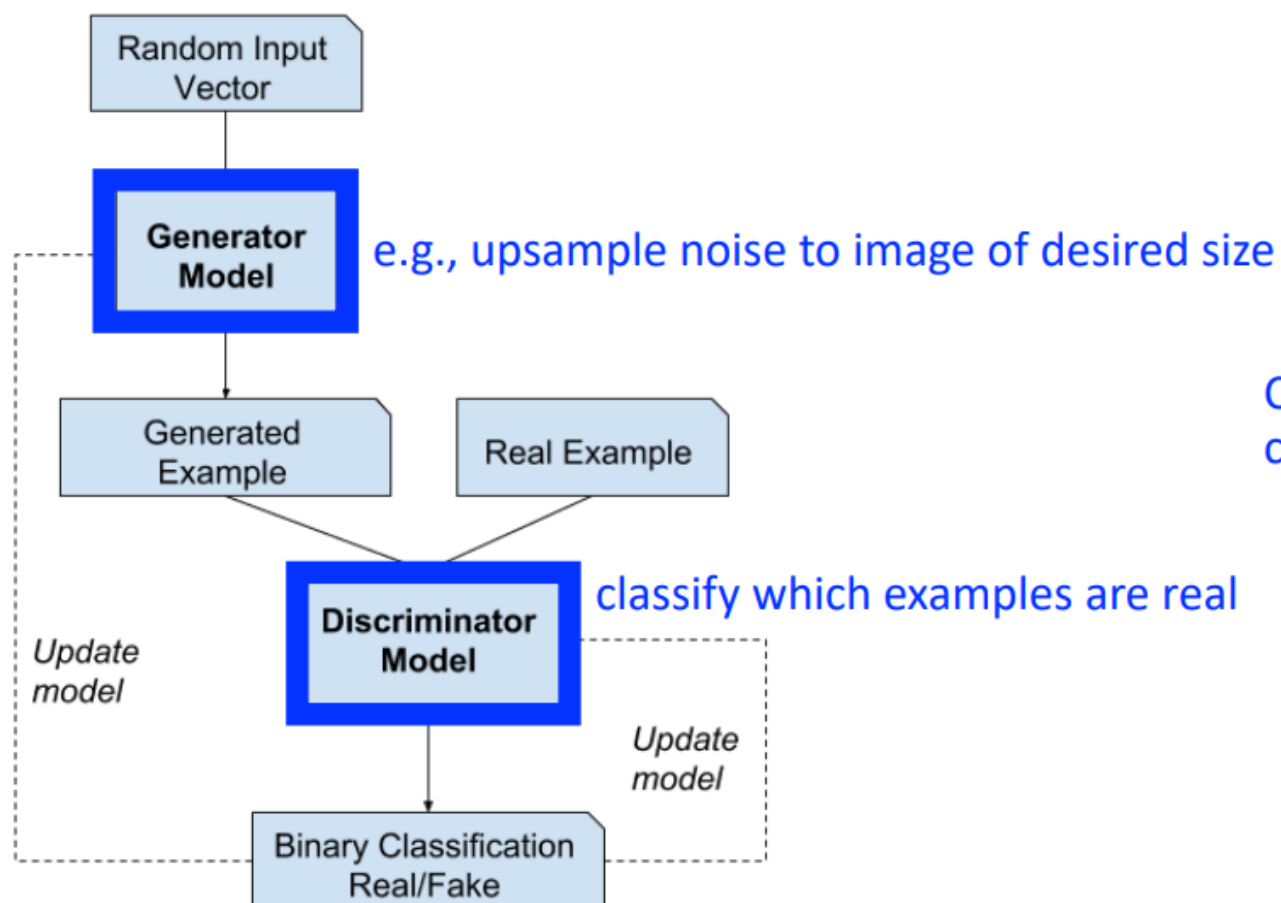
- Given this input 620 x 426 image (264,120 pixels):



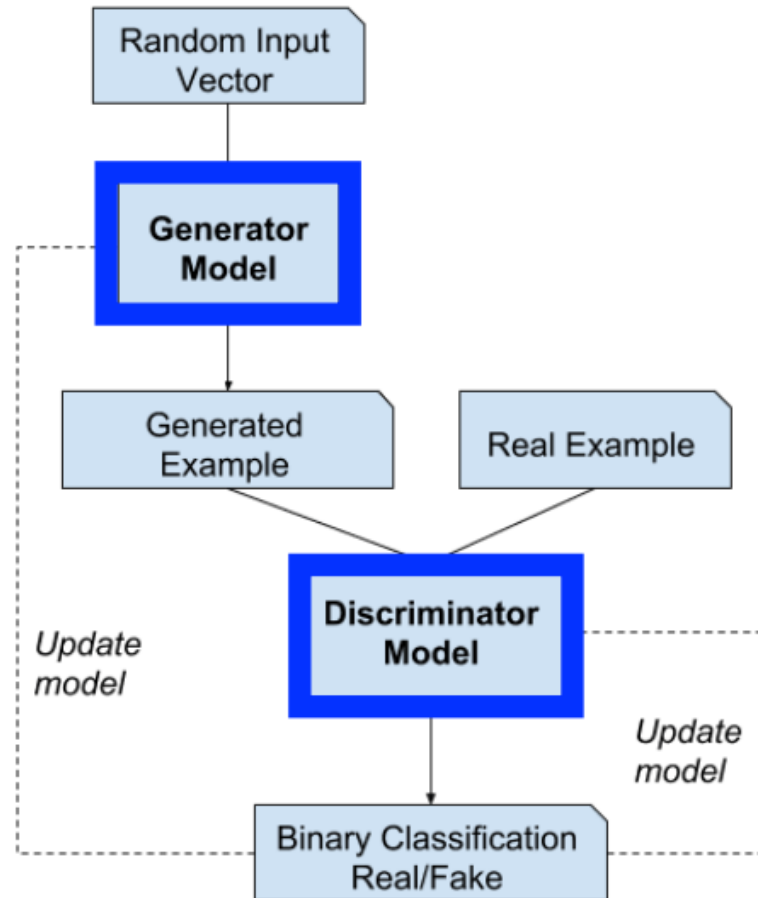
- What number of nodes are in the final layer?
 - 264,120



Generative Adversarial Networks (GANs)



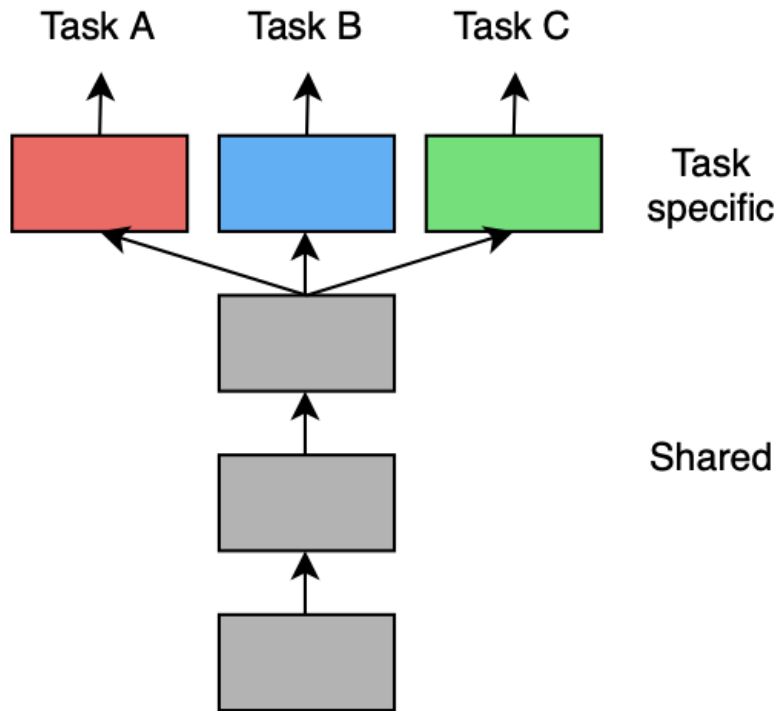
GAN Training



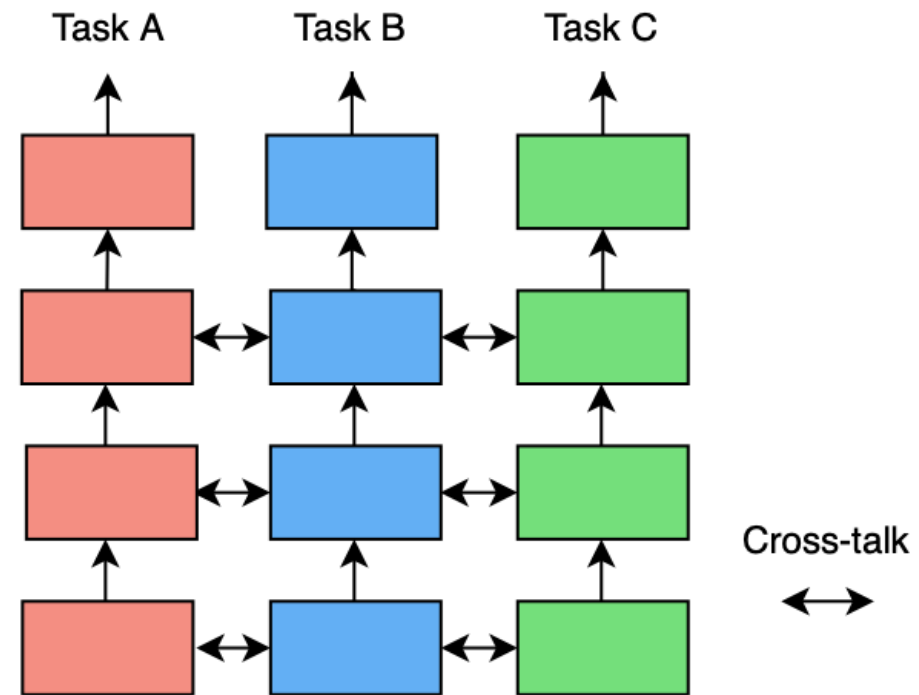
The two models are iteratively trained separately

- Train discriminator using fake and real images
- Train generator using just fake images and penalize it when the discriminator recognizes images are fake

Multi-Task Learning



(a) Hard parameter sharing



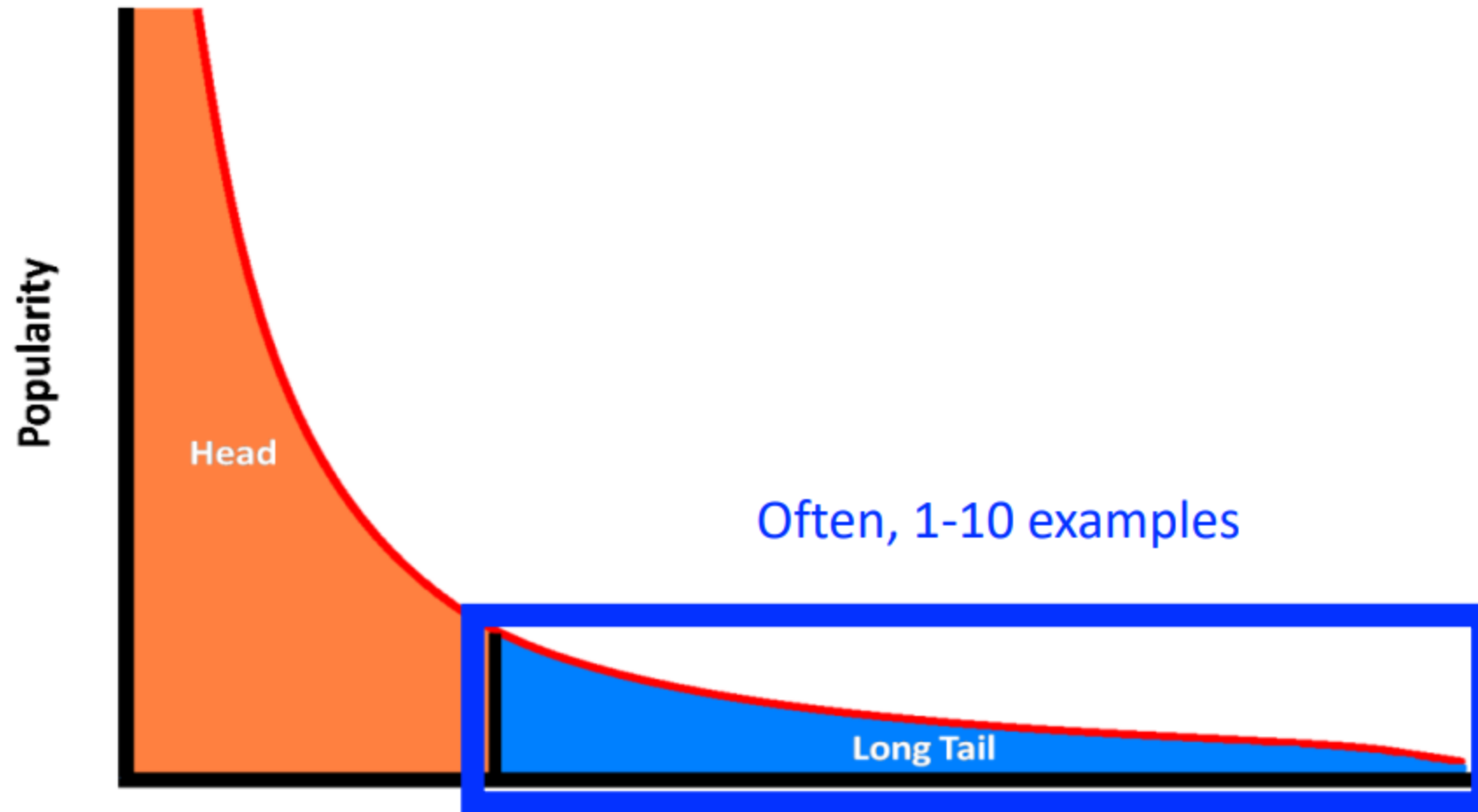
(b) Soft parameter sharing

Multi-Task Learning

- General Benefits from Parameter Sharing?
 - Data augmentation
 - Enables features found for one task to be available for another
 - Emphasizes generalizable features that are common across tasks

Zero/Few-Shot Learning

- Problem Set-up: Learn from Few Examples



Zero/Few-Shot Learning

- Intuition: Generalize Current Knowledge to Quickly Generalize to New Categories



What is this?

How many examples do you think you would need to see to recognize another one of these?

Zero/Few-Shot Learning

- Intuition: Generalize Current Knowledge to Quickly Generalize to New Categories



Could see 0 examples if you knew the object fuses a person on top with a horse on the bottom

Zero/Few-Shot Learning

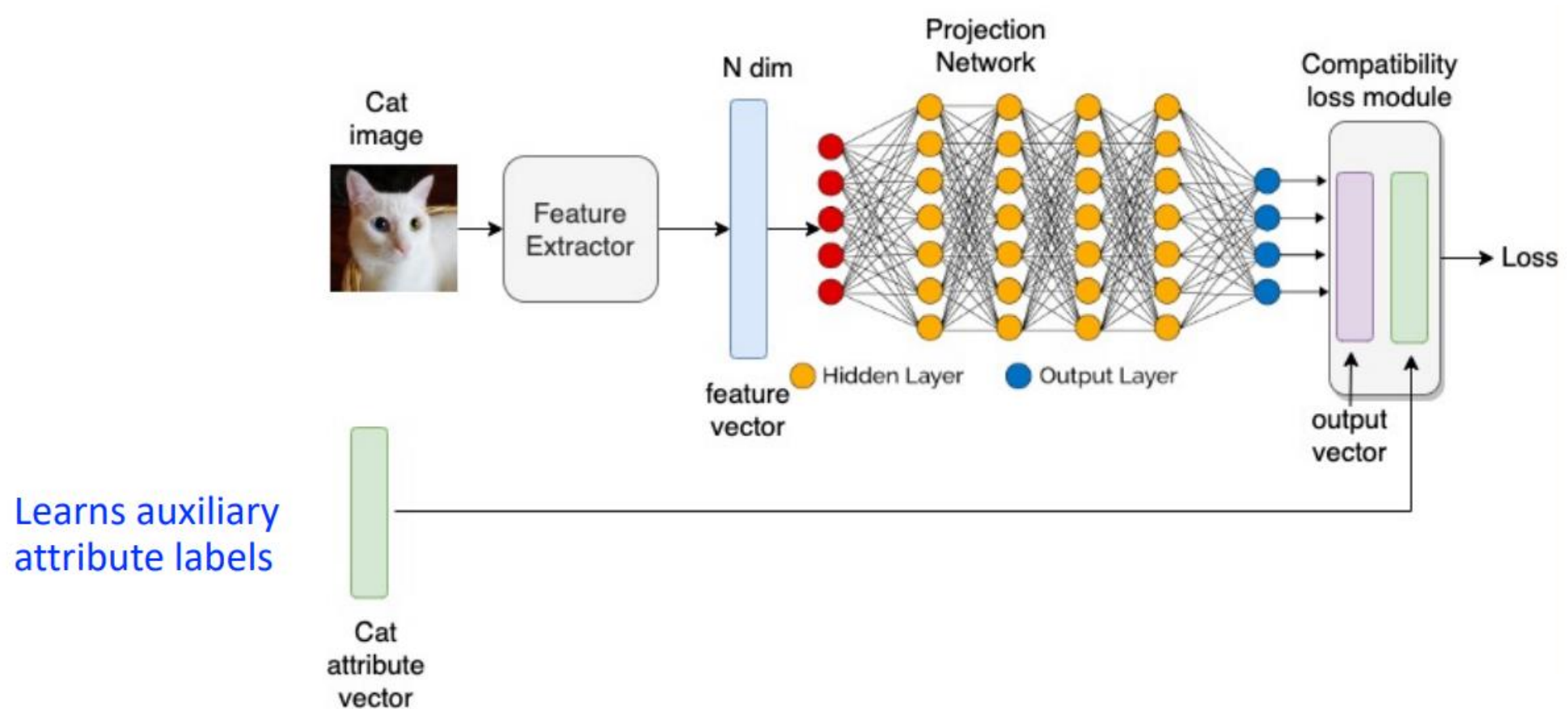
- Intuition: Generalize Current Knowledge to Quickly Generalize to New Categories



Could see 0 examples of a zebra if you knew it looks like a horse with black and white stripes

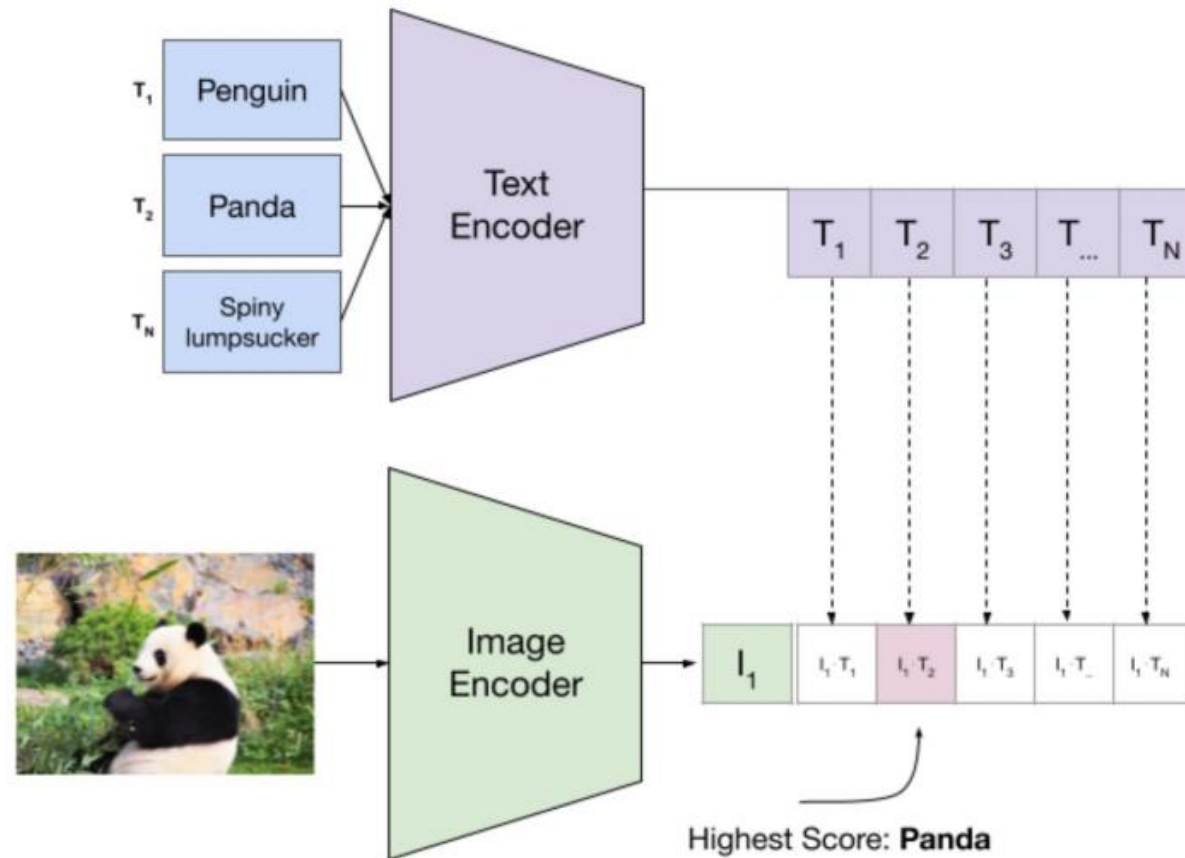
Zero/Few-Shot Learning

- Key Idea: Learn from Auxiliary Labels How to Perform a Different Task with Zero/Few Training Examples



Zero/Few-Shot Learning

- Contrastive Language-Image Pretraining (CLIP)



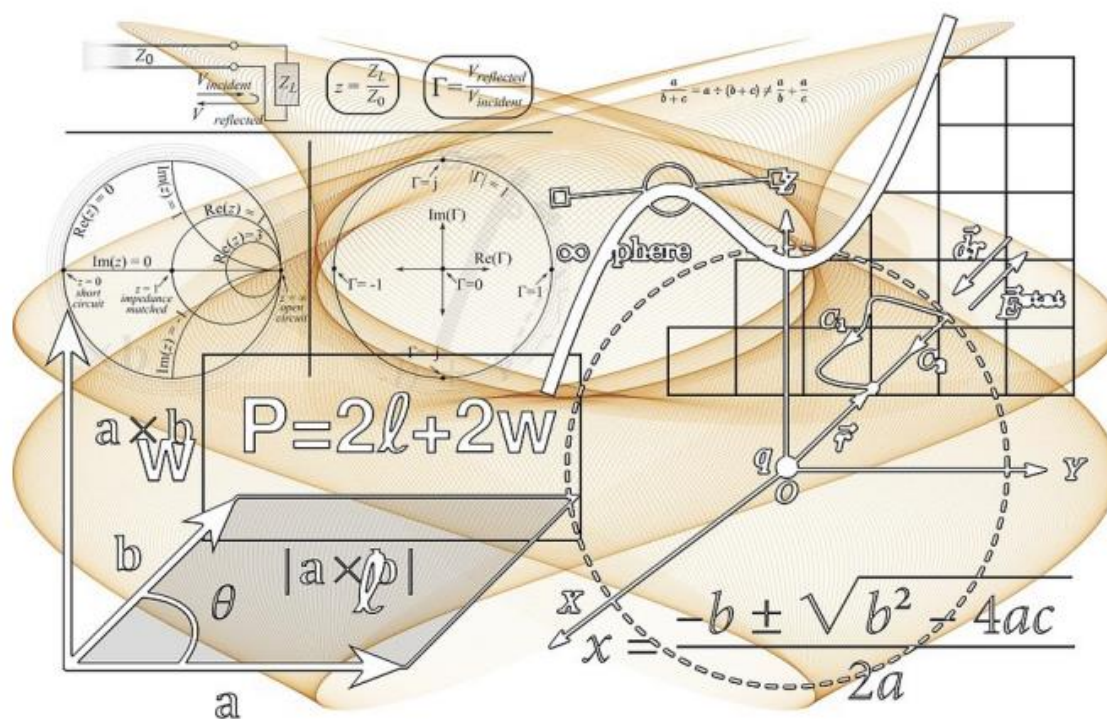


Efficient Learning

Efficient Learning

- How to teach machines so they learn faster?

Random Order of Examples



Meaningful Order of Examples

Table of Contents

<p>Letter from Dwight Kirby vi</p> <p>Introduction to Foldables</p> <p>Why Use Foldables as Manipulatives? vi</p> <p>Foldable Basics 1</p> <p>Choosing the Appropriate Foldable 2</p> <p>Folding Techniques</p> <p>Basic Foldable Types 3</p> <p>1-Piece Fold</p> <p>Book Fold 4</p> <p>Envelope Fold 5</p> <p>Bookend Fold 6</p> <p>Two-Side Fold 6</p> <p>2-Piece Fold</p> <p>Accordion 10</p> <p>Pocket Book 15</p> <p>Slotted Fold 12</p> <p>3-Piece Fold</p> <p>Tri-fold Book 13</p> <p>Three-Side Book 14</p> <p>Three-Side Book Variations 15</p> <p>Personal Fold in Minutes 16</p> <p>4-Piece Fold</p> <p>Expanded Look Book 17</p> <p>Four-Fold Book 18</p> <p>Envelope Fold 19</p> <p>Shocking Fold 24</p> <p>Four-Side Book 20</p> <p>Top-Tail Book 22</p> <p>Accordion Book 24</p> <p>Any Number of Pages</p> <p>Top-Up Book 25</p> <p>Building into Pages 26</p> <p>Expanded Table of Contents 27</p> <p>Folding in Certain Side Panels 28</p> <p>Circle Graph 29</p> <p>Orange-Side Book 30</p> <p>Handbook Book 31</p> <p>Project-Using Units</p> <p>Millimeter Project 32</p>	<p>Science Study Guides 33</p> <p>Science Maps 34</p> <p>Math Activities using Foldables</p> <p>Number Systems</p> <p>Whole Numbers 35</p> <p>Fractions 36</p> <p>Fractions: Adding and Subtracting 37</p> <p>Fractions: Multiplying and Dividing 38</p> <p>Rational Numbers 39</p> <p>Rational Numbers: Fractions 40</p> <p>Rational Numbers: Decimals 41</p> <p>Percent 42</p> <p>Polys 43</p> <p>Polynomials 44</p> <p>Arithmetic Patterns 45</p> <p>Real Number System 46</p> <p>Algebraic Patterns and Functions</p> <p>Geo and Algebraic 47</p> <p>Exponents 48</p> <p>Properties 49</p> <p>Equations 50</p> <p>Inequalities 51</p> <p>Relations and Functions 52</p> <p>Factors 53</p> <p>Multiples 54</p> <p>Monomials and Polynomials 55</p> <p>Formulas and Equations 56</p> <p>Geometry 57</p> <p>Statistics 58</p> <p>Geometry</p> <p>Points 59</p> <p>Lines and Line Segments 60</p> <p>Angles 61</p> <p>Angle Relationships 62</p> <p>Planes 63</p> <p>Triangles 64</p> <p>Triangles 65</p> <p>Right Triangles 66</p> <p>Right Triangle Relationships 67</p> <p>Quadrilaterals 68</p> <p>Squares, Rectangles, and Rhombi 69</p>
----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Efficient Learning

- Intuition: How to Teach a Child To Read?

Random Order of Examples



Meaningful Order of Examples



Efficient Learning

- Curriculum Learning

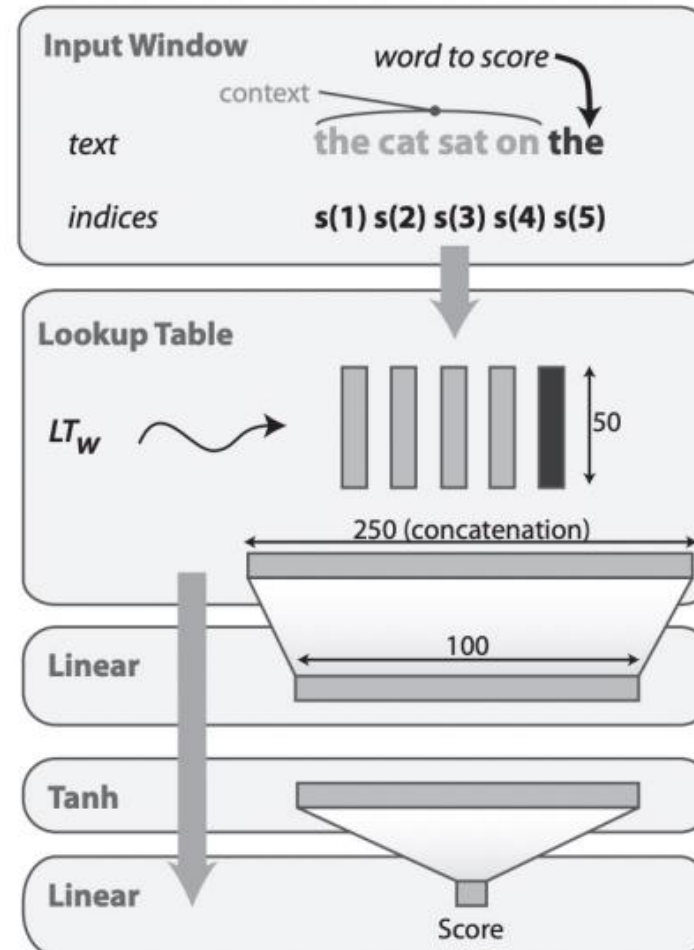
Architecture:

context size
set to 5

Easy: 5,000 most
frequent words

Hard: additional 5,000
words at each epoch
until 20,000 words

Examples with words
not in the vocab were
discarded from training

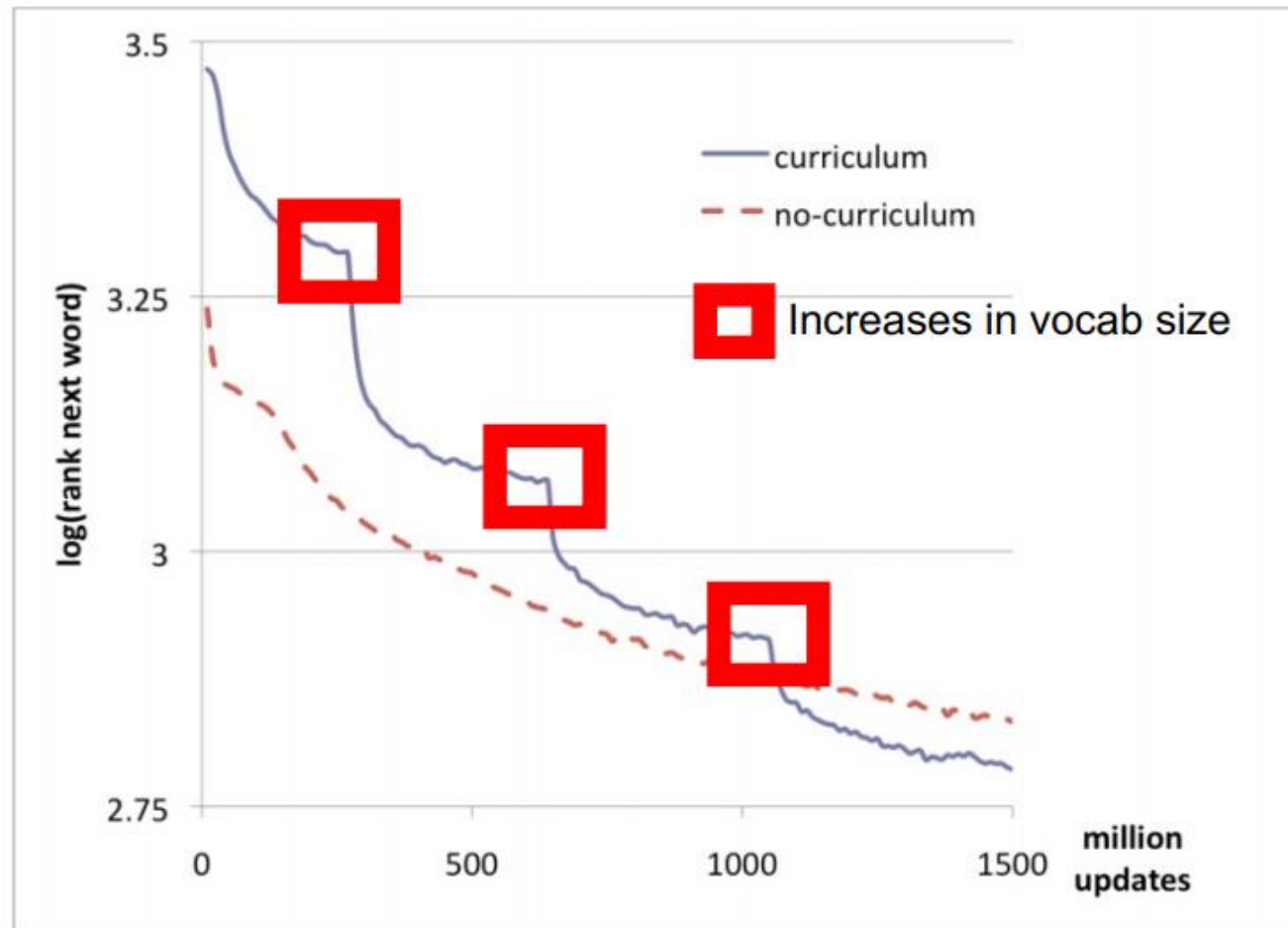


2. Predict the next word

Background music from a _____

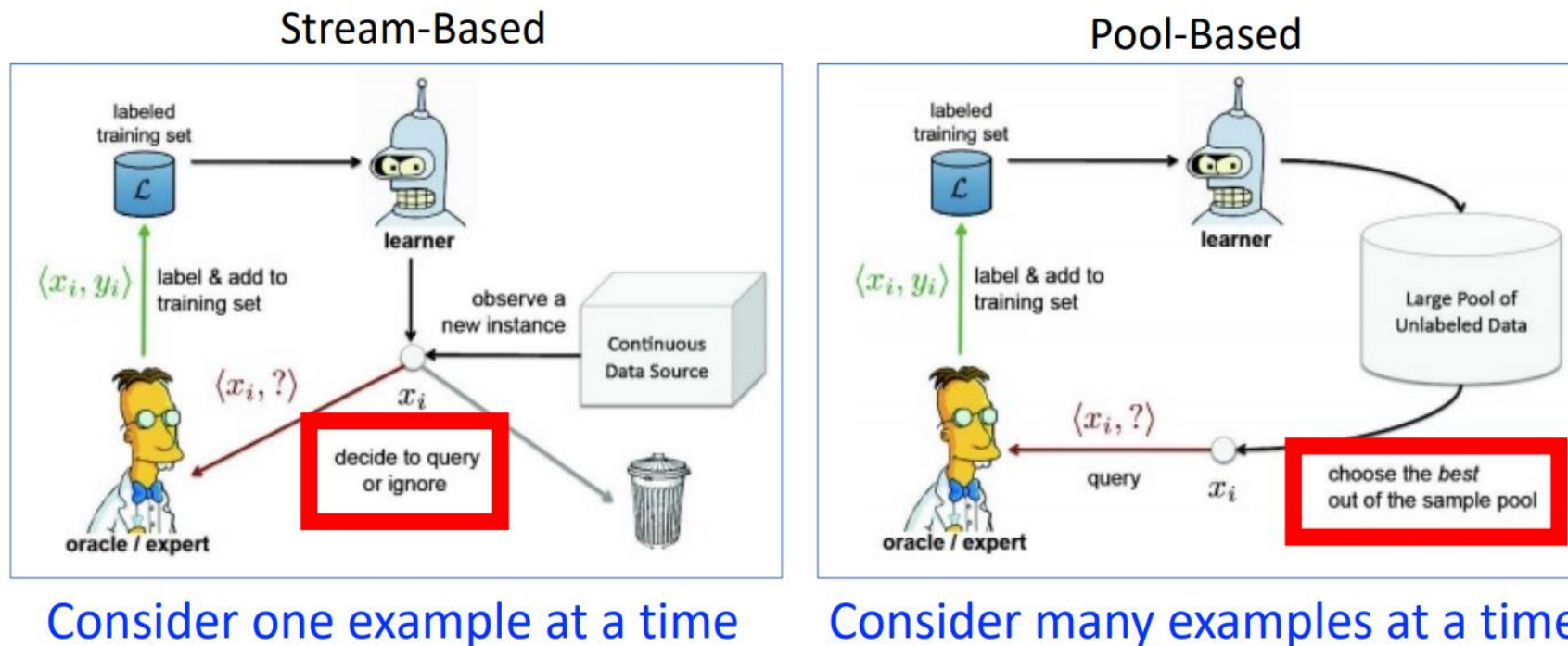
Efficient Learning

- Curriculum Learning: Next Word Prediction



Efficient Learning

- Active Learning:
 - Actively select the examples to label that would be most effective for learning rather than labelling all available data.
- Types of Active Learning:

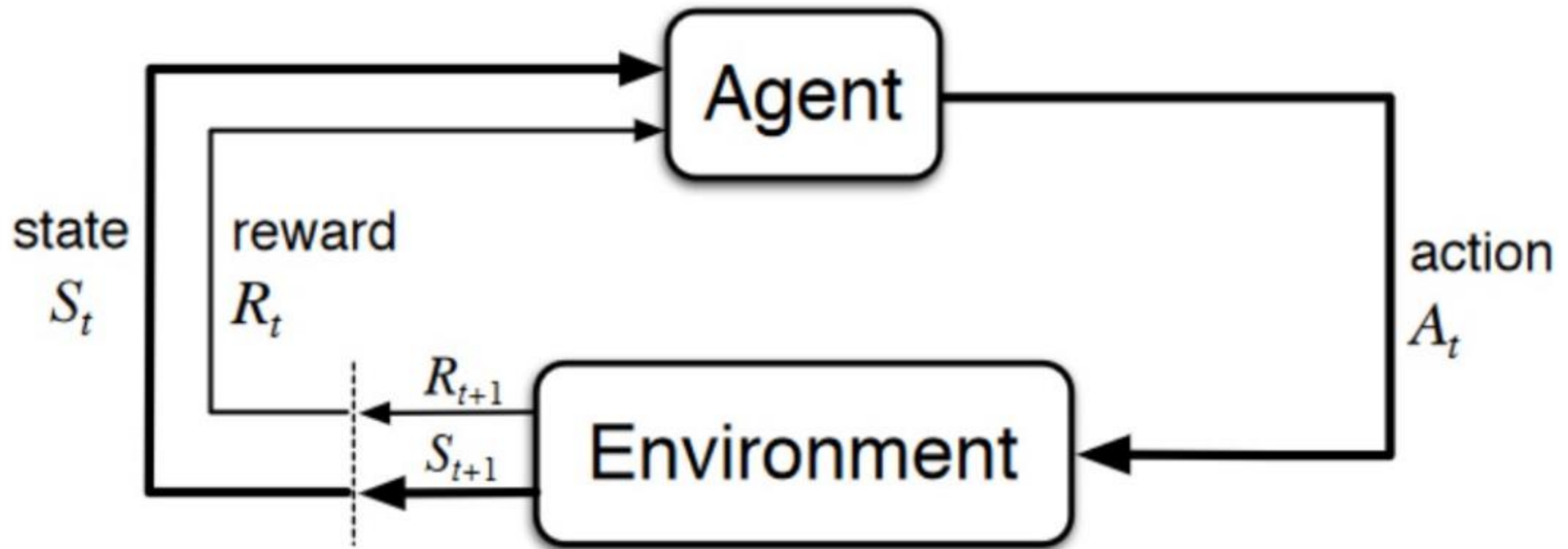




Reinforcement Learning

Reinforcement Learning

- Agent takes actions in an environment to maximize the total reward



Reinforcement Learning

- Intuition: Learning to Walk by Trial-and Error



Reinforcement Learning Applications

Learning to Walk in 20 Minutes

Russ Tedrake
Brain & Cognitive Sciences
Center for Bits and Atoms
Massachusetts Inst. of Technology
Cambridge, MA 02139
russt@csail.mit.edu

Teresa Weirui Zhang
Mechanical Engineering
Department
University of California, Berkeley
Berkeley, CA 94270
resa@berkeley.edu

H. Sebastian Seung
Howard Hughes Medical Institute
Brain & Cognitive Sciences
Massachusetts Inst. of Technology
Cambridge, MA 02139
seung@mit.edu



Reinforcement Learning Applications

Autonomous reinforcement learning on raw visual input data in a real world application

Sascha Lange, Martin Riedmiller
Department of Computer Science
Albert-Ludwigs-Universität Freiburg
D-79110, Freiburg, Germany
Email: [slange,riedmiller]@informatik.uni-freiburg.de

Arne Voigtländer
Shoogee GmbH & Co. KG
Krögerweg 16a
D-48155 Münster, Germany
Email: arne@shoogee.com

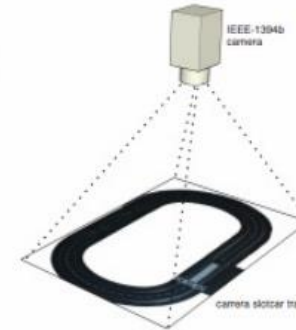
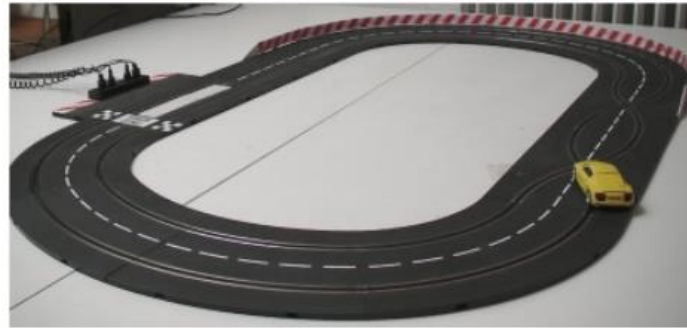


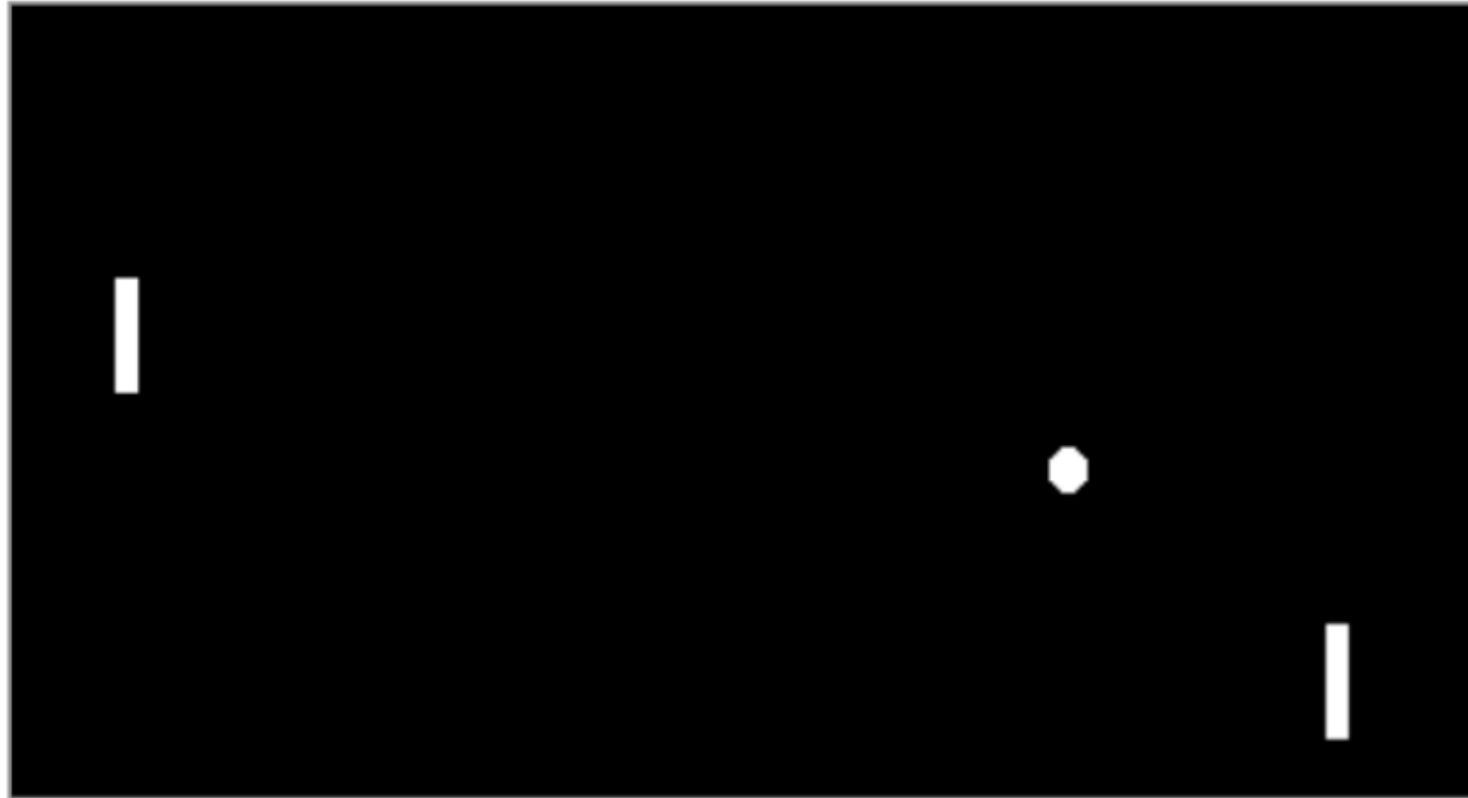
Fig. 1. The visual slot car racer task. The controller has to autonomously learn to steer the racing car by raw visual input of camera images.

Reinforcement Learning Applications



Reinforcement Learning Applications

- Pong Game Learning Example
 - Goal: compute optimal “up” and “down” paddle movements to maximize rewards



Reinforcement Learning Applications

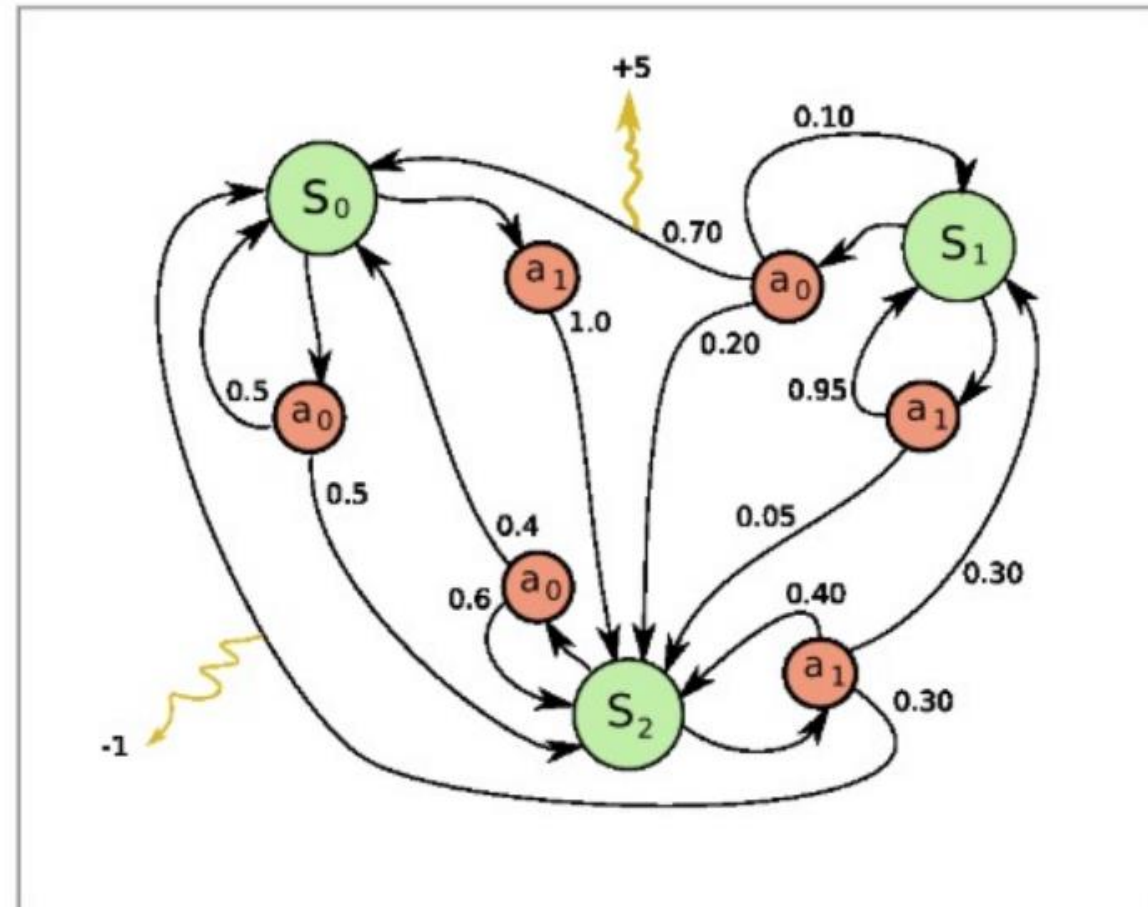
- Pong Game Learning Example

Representation: graph where nodes are game states and edges are possible transitions with rewards

-1 if missed the ball

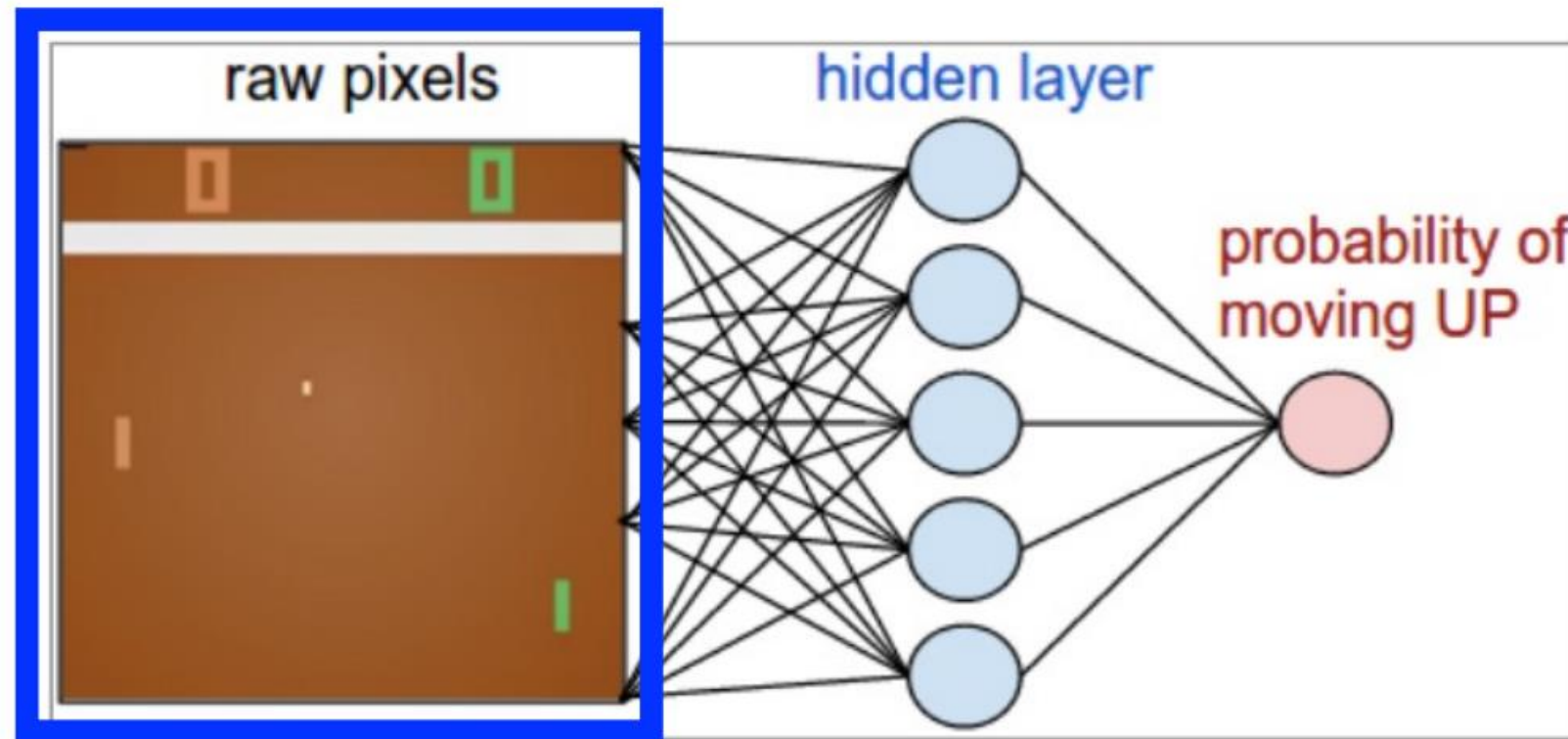
+1 reward if ball goes past opponent

0 otherwise



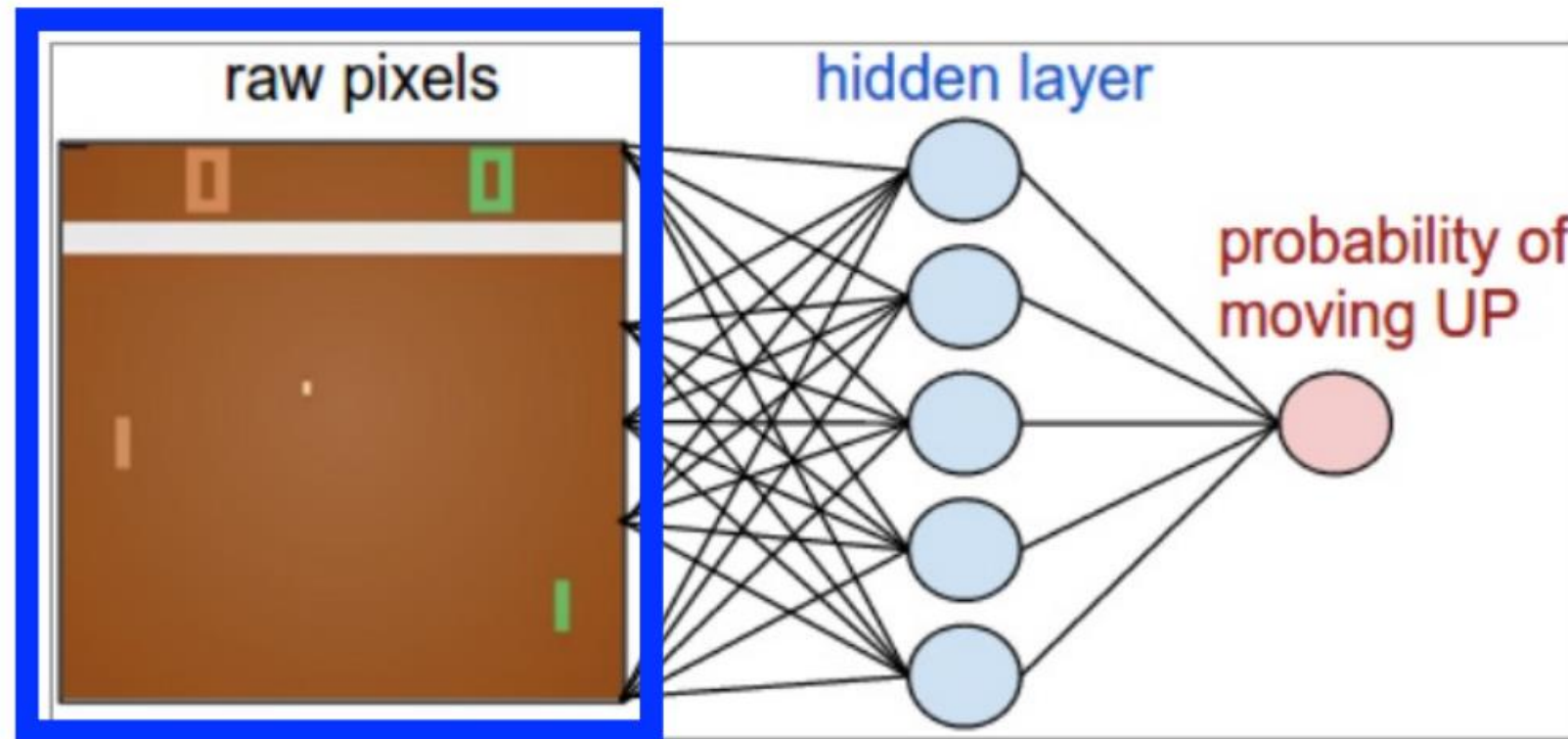
Reinforcement Learning Applications

- Pong Game Learning Example
 - Given game state (as image), decide if to move paddle **up** or **down**



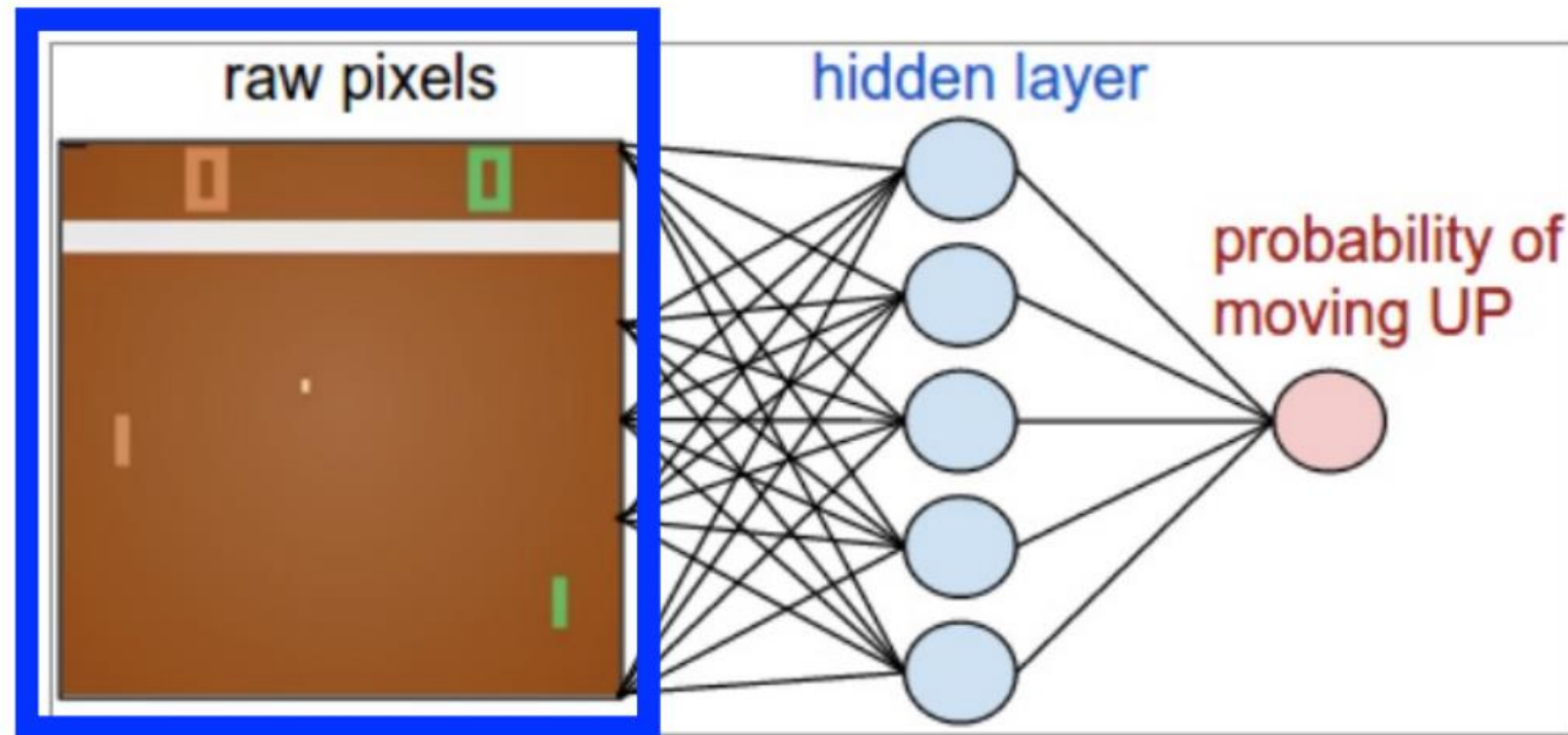
Reinforcement Learning Applications

- Pong Game Learning Example
 - Reward provided after each game state of moving paddle **up** or **down**



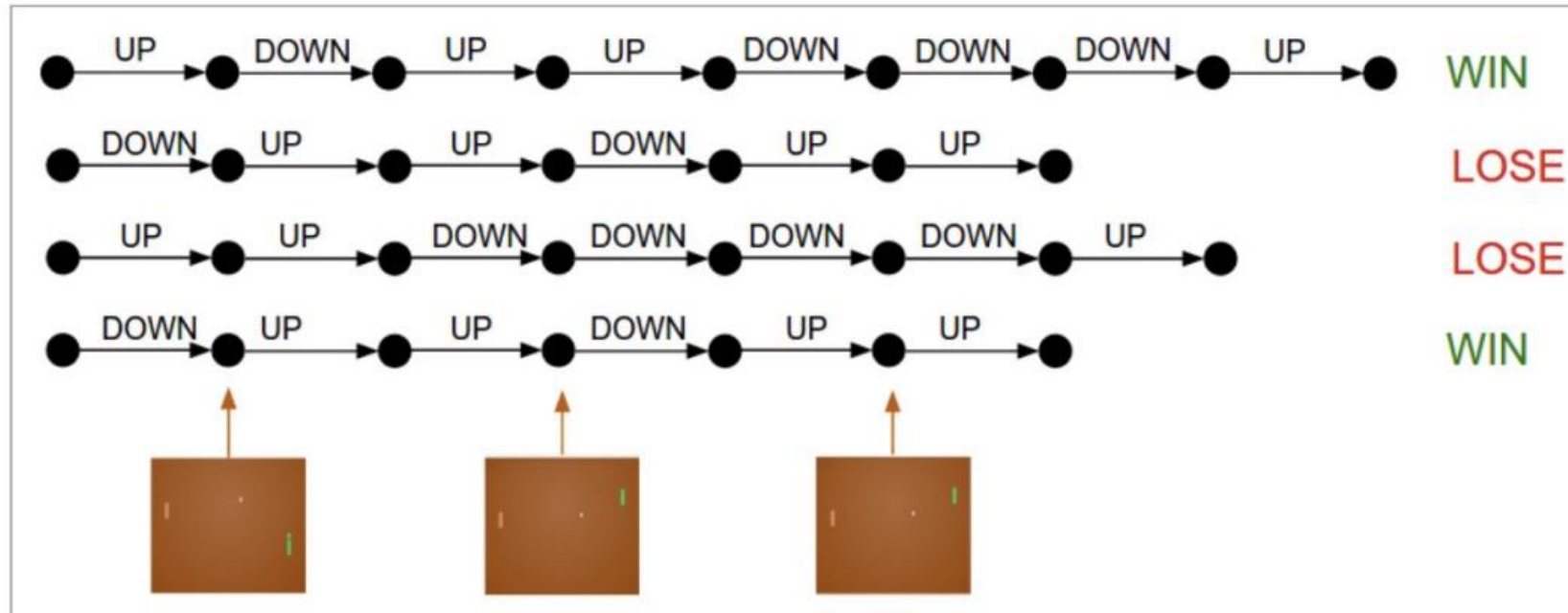
Reinforcement Learning Applications

- Pong Game: Policy Network
 - Problem: reward may be due to good action **many steps ago**



Reinforcement Learning Applications

- Pong Game: Training Protocol



- Encourages actions that eventually lead to good outcomes and discourages actions that eventually lead to bad outcomes by updating gradients accordingly