UNIVERSIDADE CATÓLICA PORTUGUESA

BRAGA

Deep Learning

Session 26

Multi Topic Overview

Applied Data Science

2024/2025

Multimodal Neural Networks

What is Multimodal Learning?

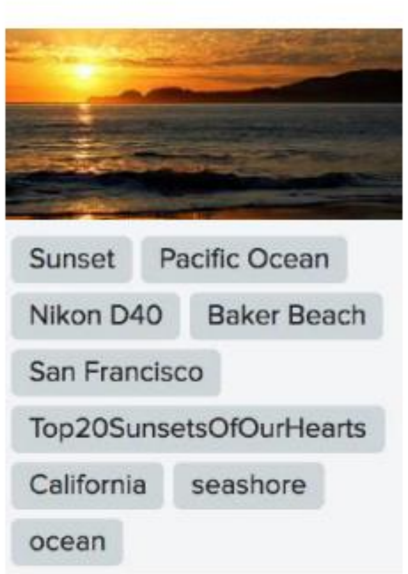
UNIVERSIDADE CATÓLICA PORTUGUESA
BRAGA

- 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

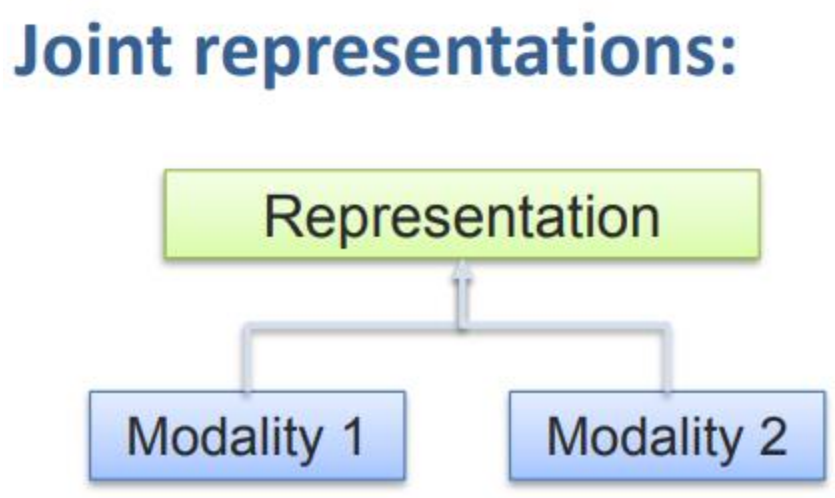
- Data is usually a collection of modalities

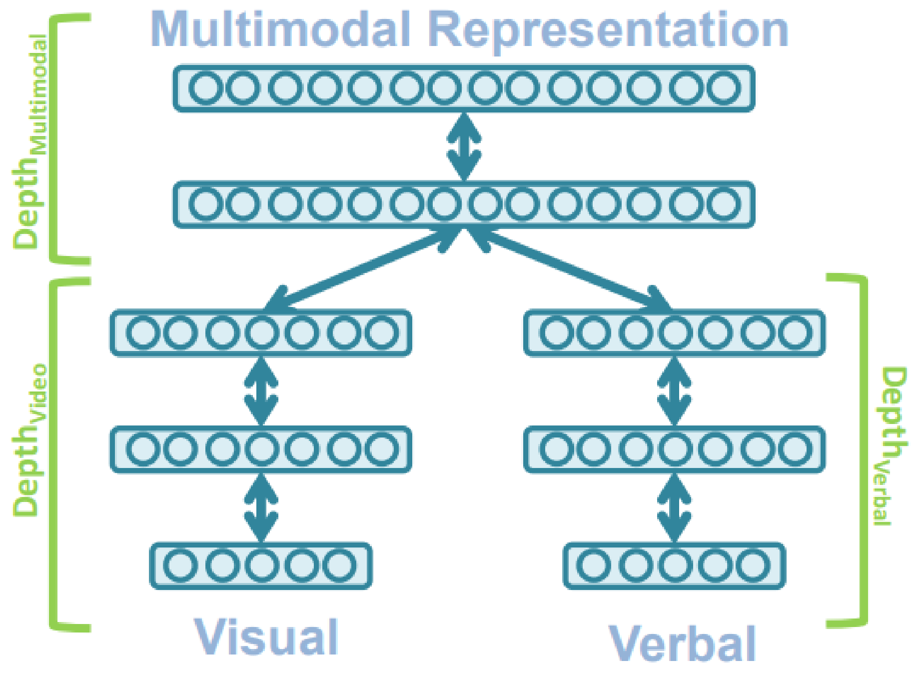


Why is Multimodal Learning Hard?

UNIVERSIDADE CATÓLICA PORTUGUESA
BRAGA

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

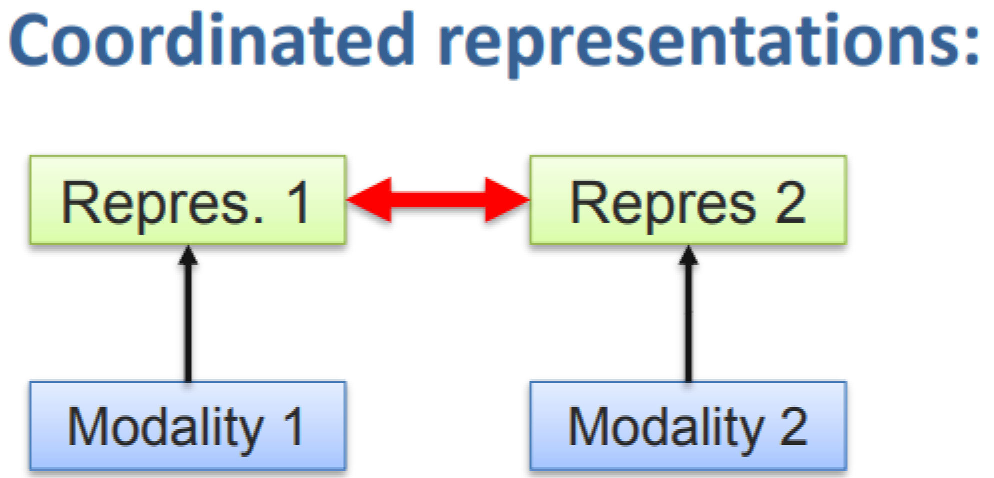


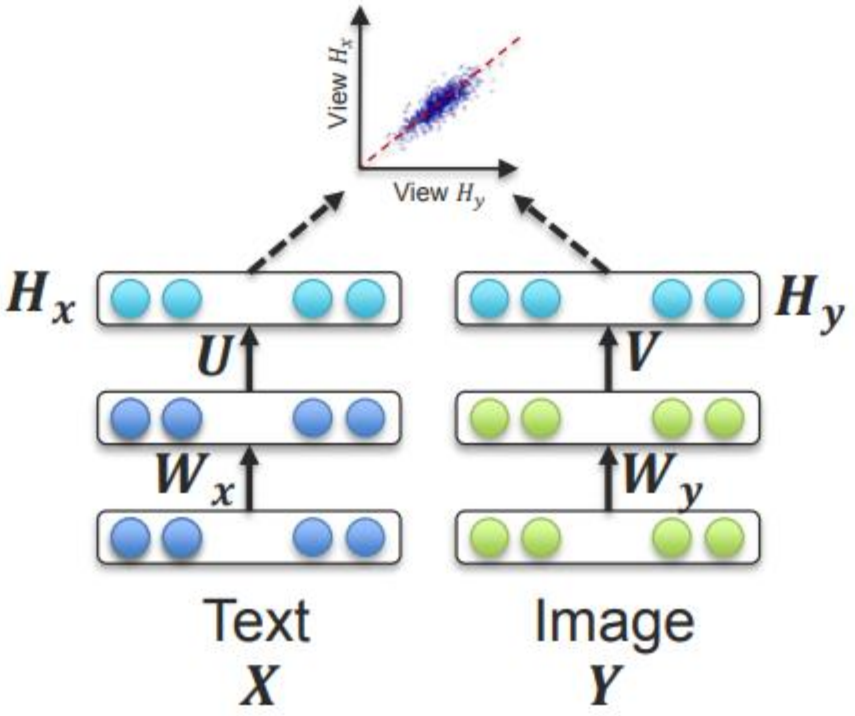


Why is Multimodal Learning Hard?

UNIVERSIDADE CATÓLICA PORTUGUESA
BRAGA

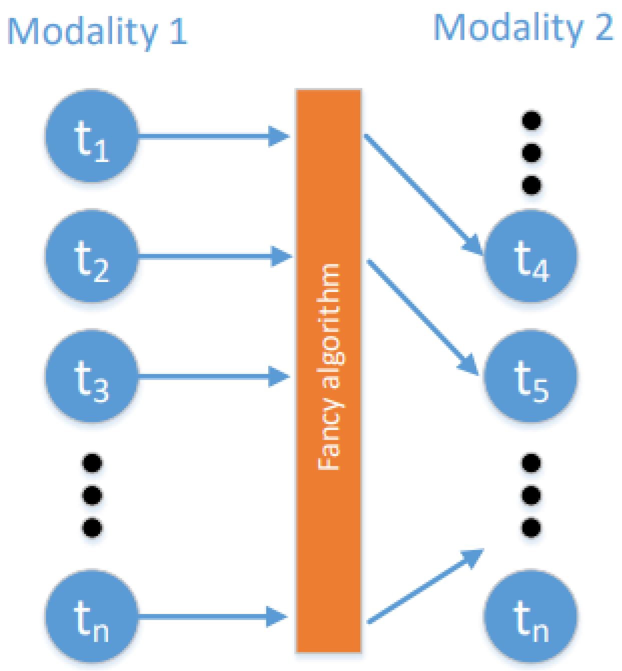
- Different representations
  - Learning how to represent and summarize multimodal data in away that exploits the complementarity and redundancy.





Why is Multimodal Learning Hard?

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



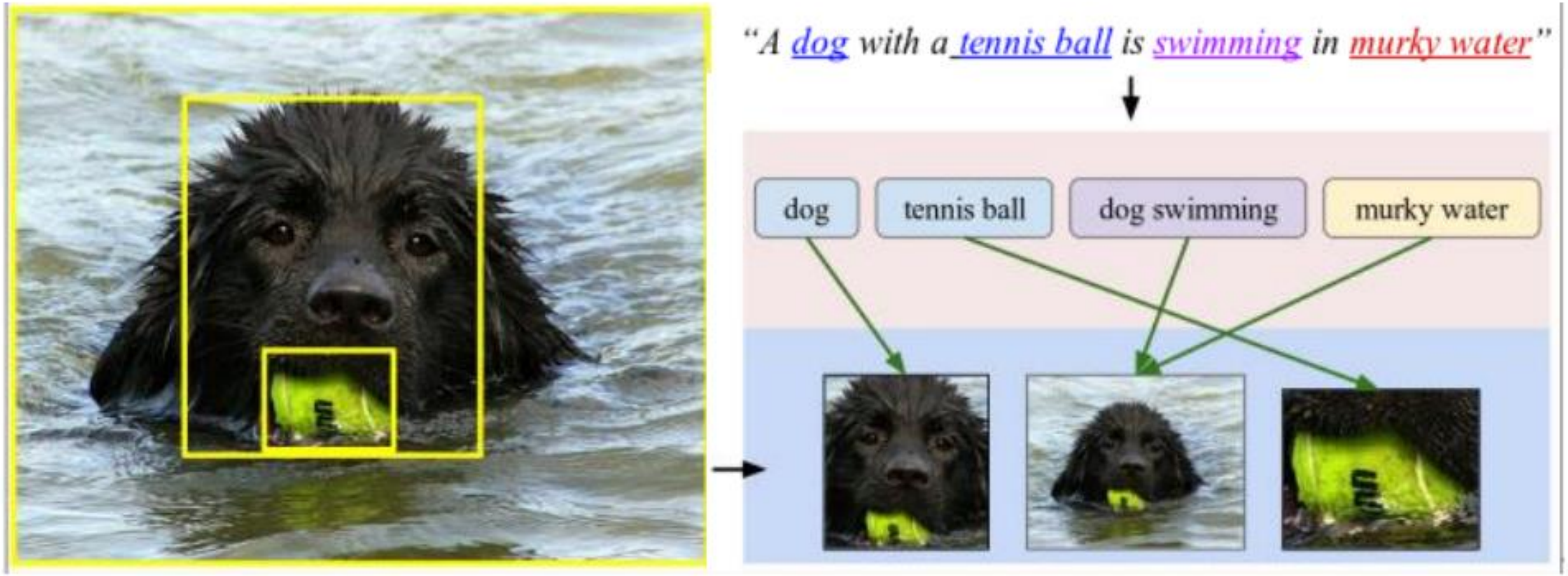
Explicit Alignment

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

Implicit Alignment

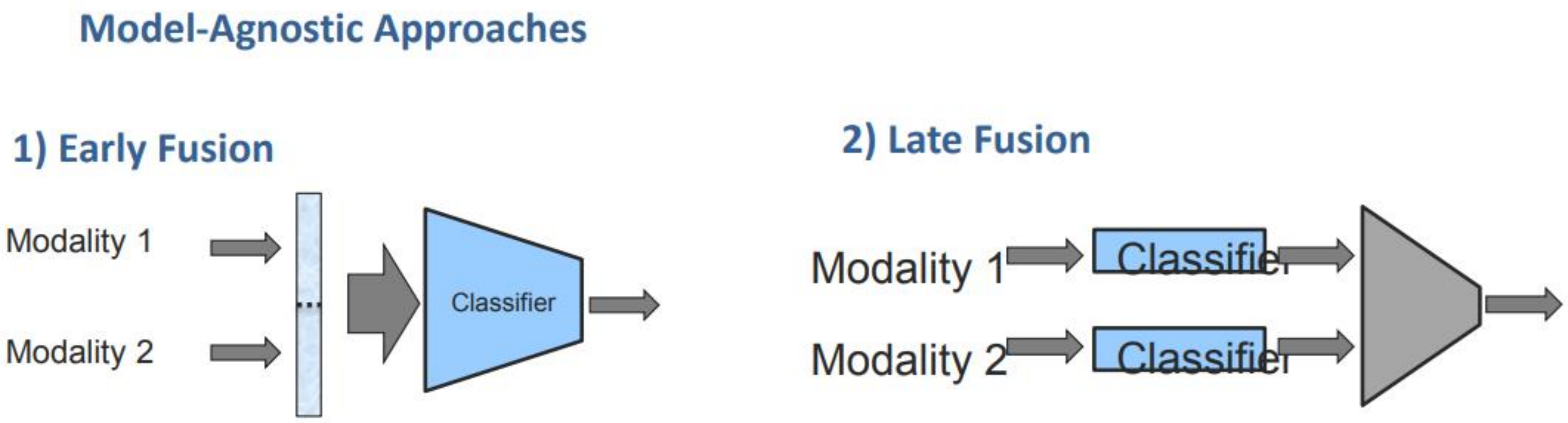
Uses internally latent alignment of modalities in order to better solve a different problem

Implicit Alignment



Why is Multimodal Learning Hard?

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



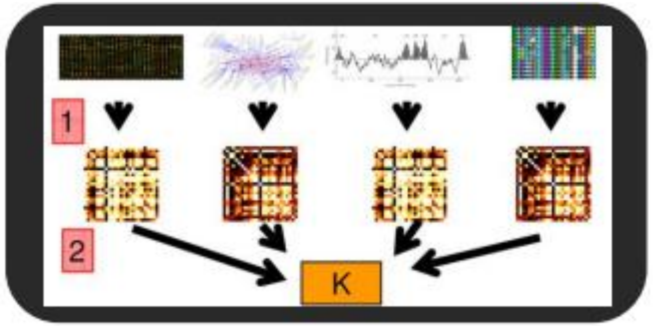
Why is Multimodal Learning Hard?

UNIVERSIDADE CATÓLICA PORTUGUESA
BRAGA

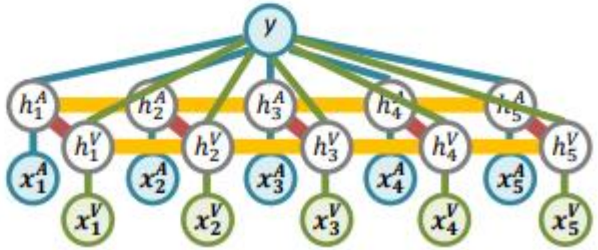
- Fusion
  - To join information from two or more modalities to perform a prediction task.

Model-Based (Intermediate) Approaches

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



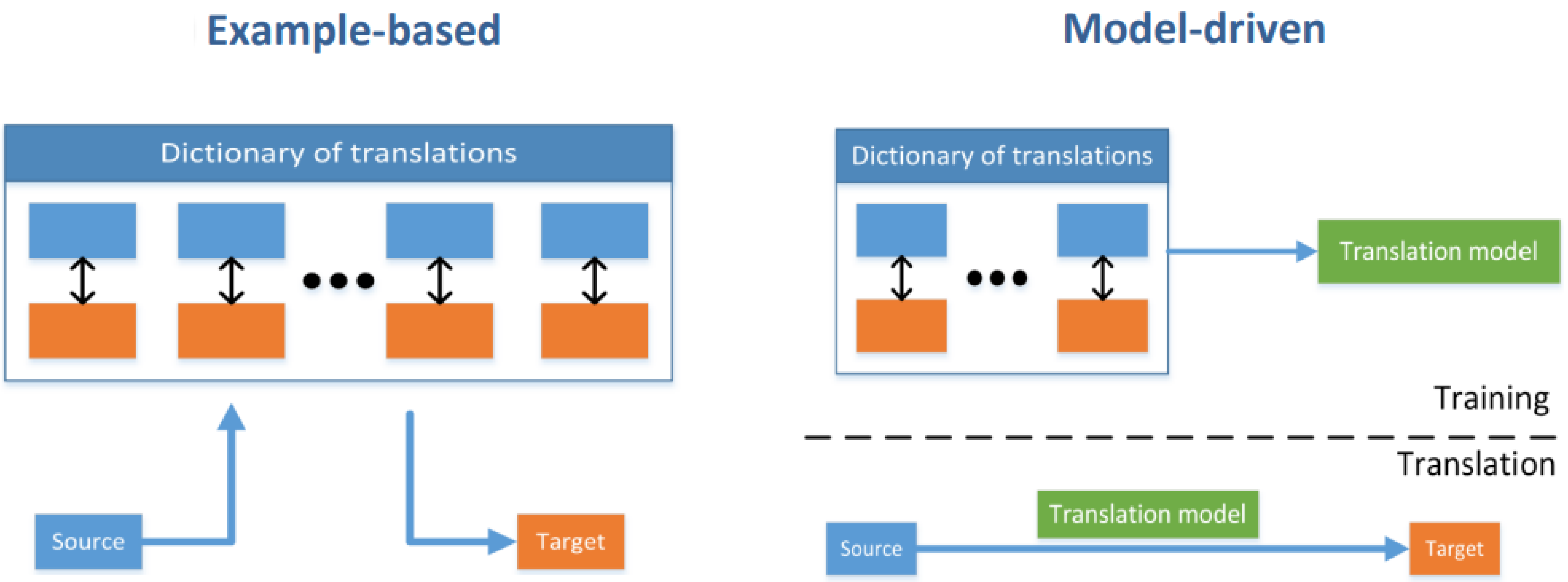
Multiple kernel learning



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.

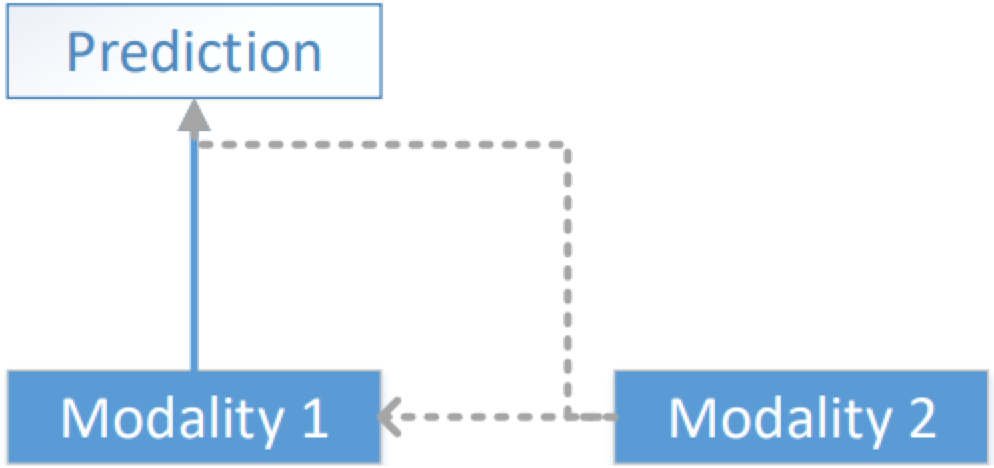


Why is Multimodal Learning Hard?

- 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?

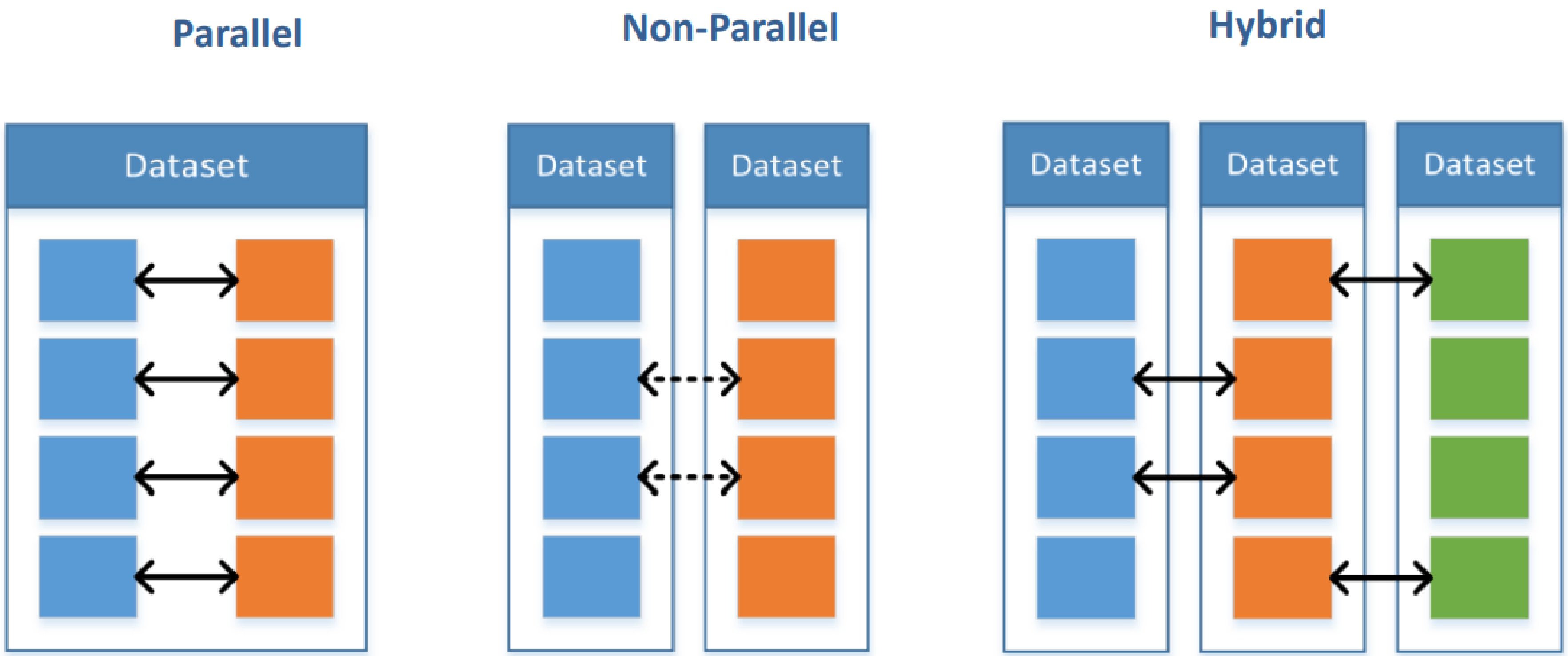
- Co-Learning:
  - Transfer knowledge between modalities, including their representations and predictive models.



Why is Multimodal Learning Hard?

UNIVERSIDADE CATÓLICA PORTUGUESA
BRAGA

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



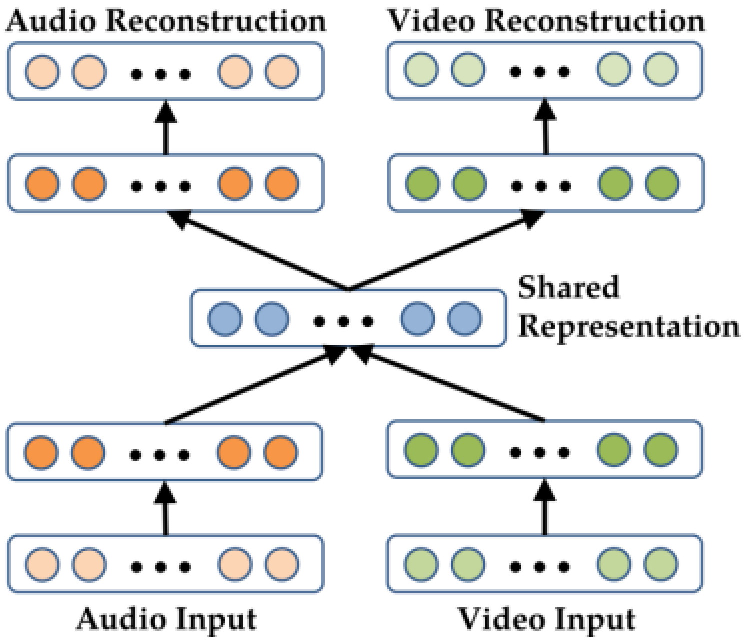
Multimodal Applications

	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

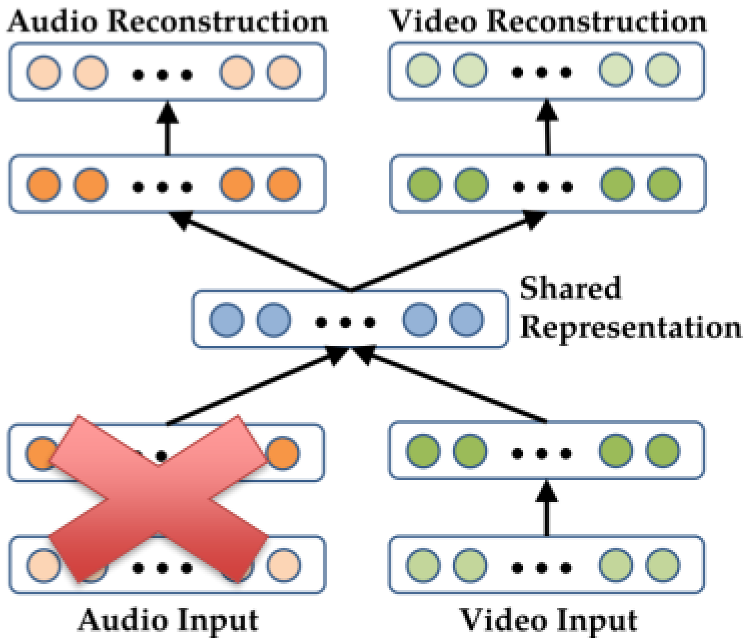
- A bimodal auto-encoder
  - Used for Audio-visual speech recognition



Deep Multimodal Autoencoders

- A deep representation learning approach

- A bimodal auto-encoder
  - Used for Audio-visual speech recognition

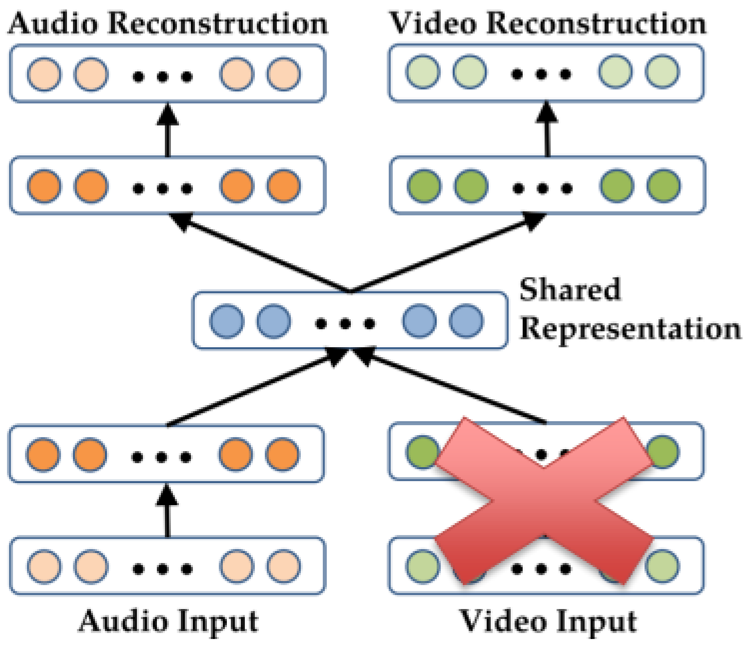


Deep Multimodal Autoencoders

UNIVERSIDADE CATÓLICA PORTUGUESA
BRAGA

- A deep representation learning approach

- A bimodal auto-encoder
  - Used for Audio-visual speech recognition

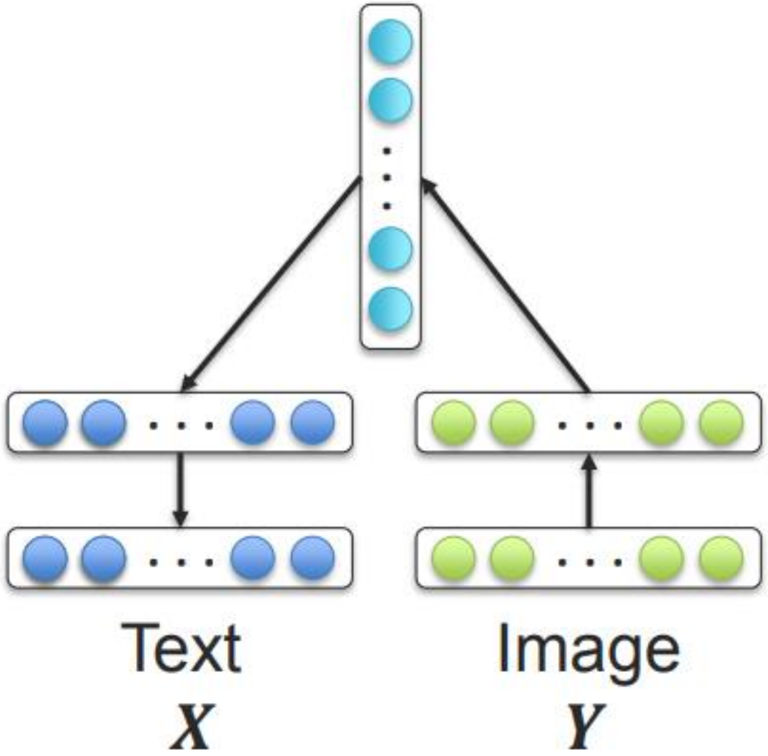


Multimodal Encoder-Decoder

UNIVERSIDADE CATÓLICA PORTUGUESA
BRAGA

- Visual modality often encoded using CNN

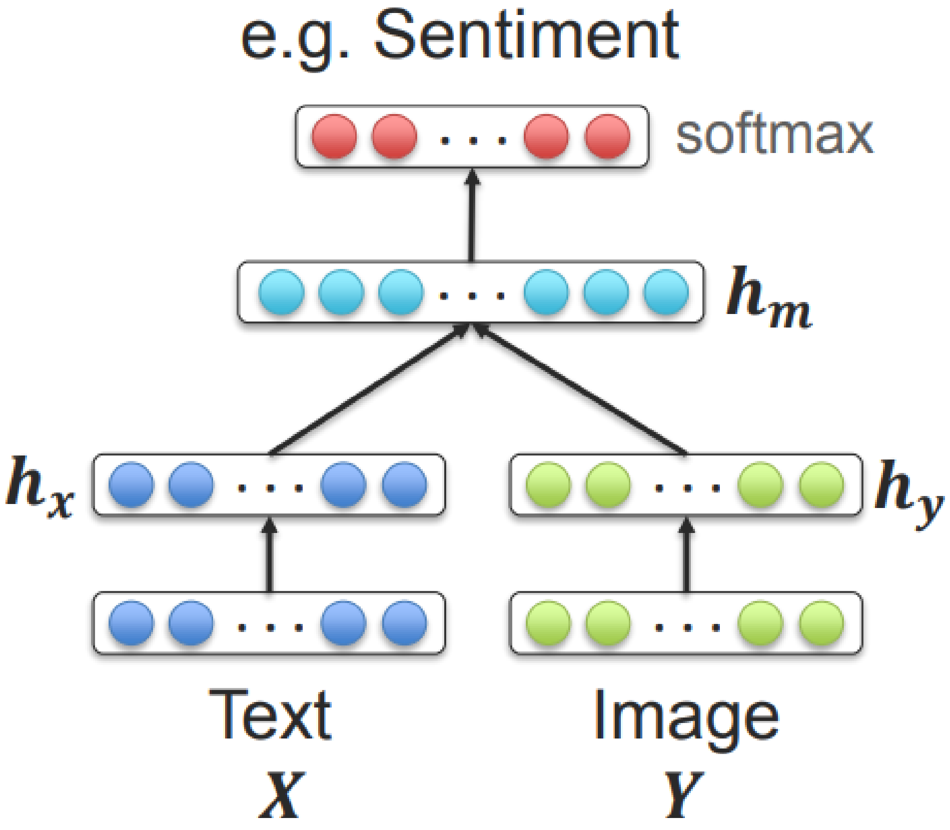
- Language modality will be decoded using LSTM
  - A simple multilayer perceptron will be used to translate from visual (CNN) to language (LSTM)



Multimodal Joint Representation

- For supervised learning tasks

- Joining the unimodal representations
  - Simple concatenation
  - Element-wise multiplication or summation
  - Multilayer perceptron



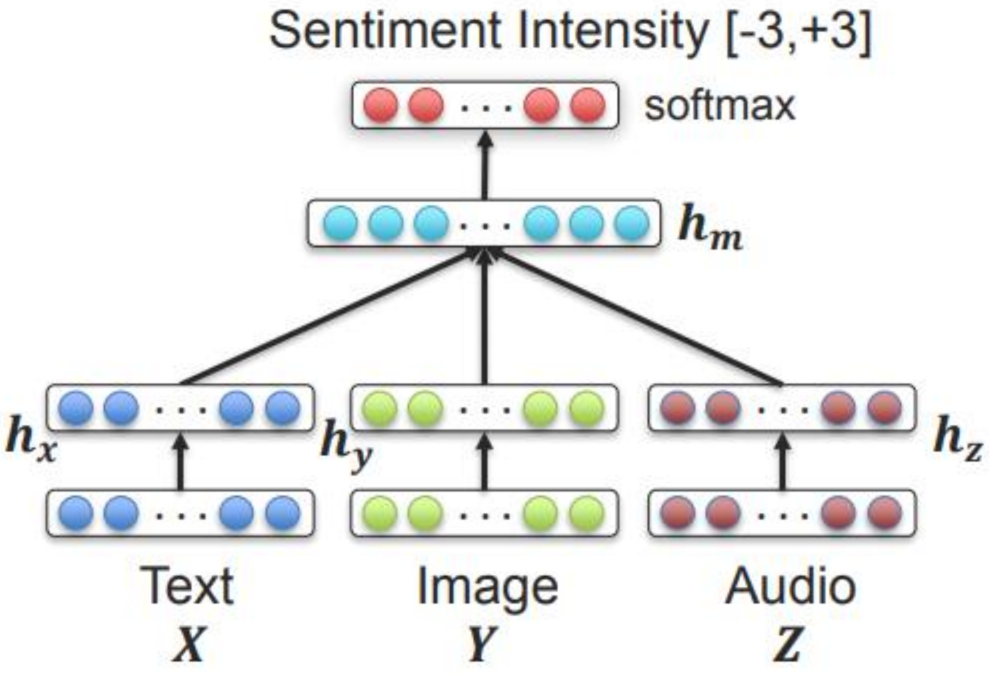
Multimodal Sentiment Analysis

MOSI dataset (Zadeh et al, 2016)

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

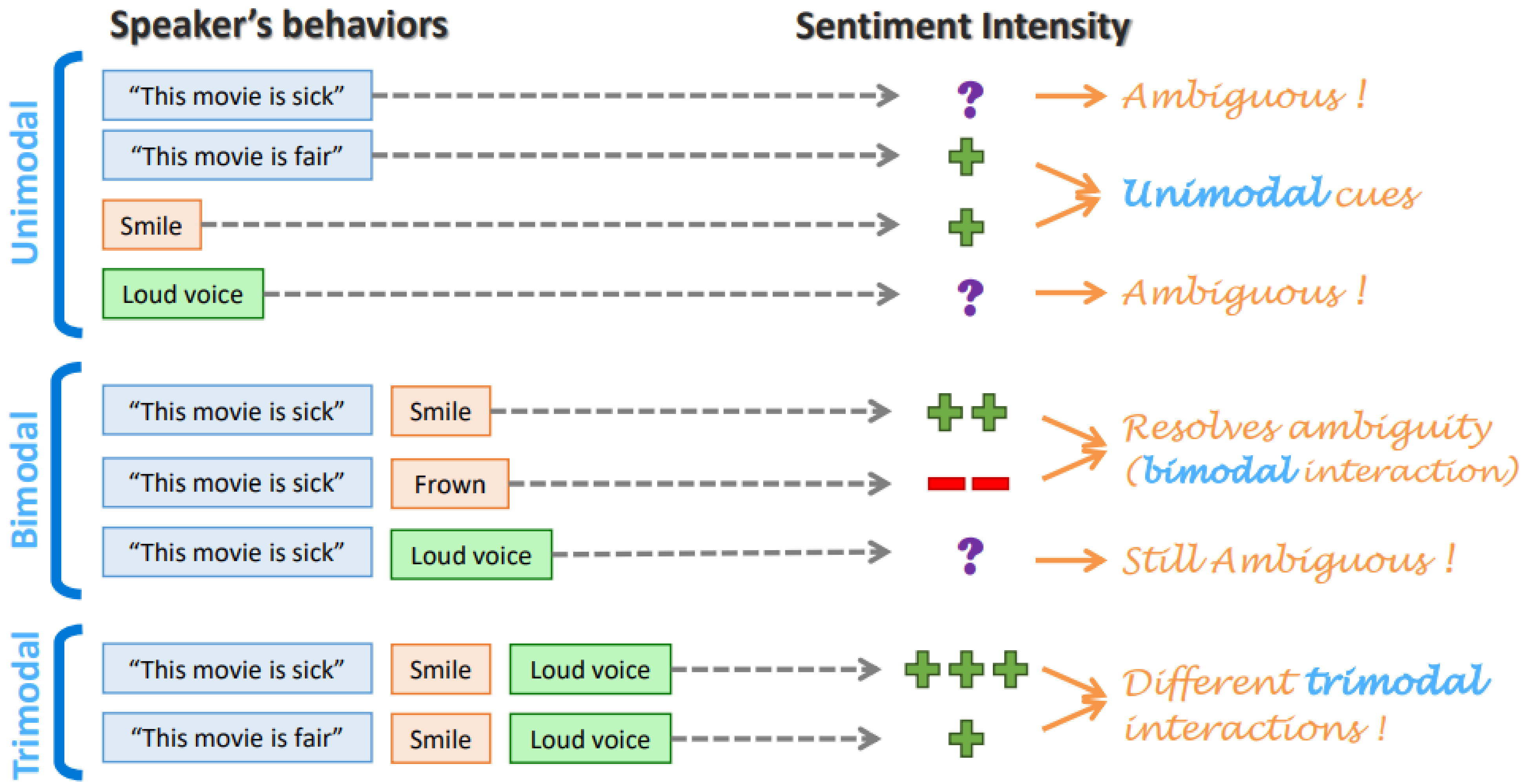
Multimodal joint representation:

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



Unimodal, Bimodal and Trimodal Interactions

UNIVERSIDADE CATÓLICA PORTUGUESA
BRAGA

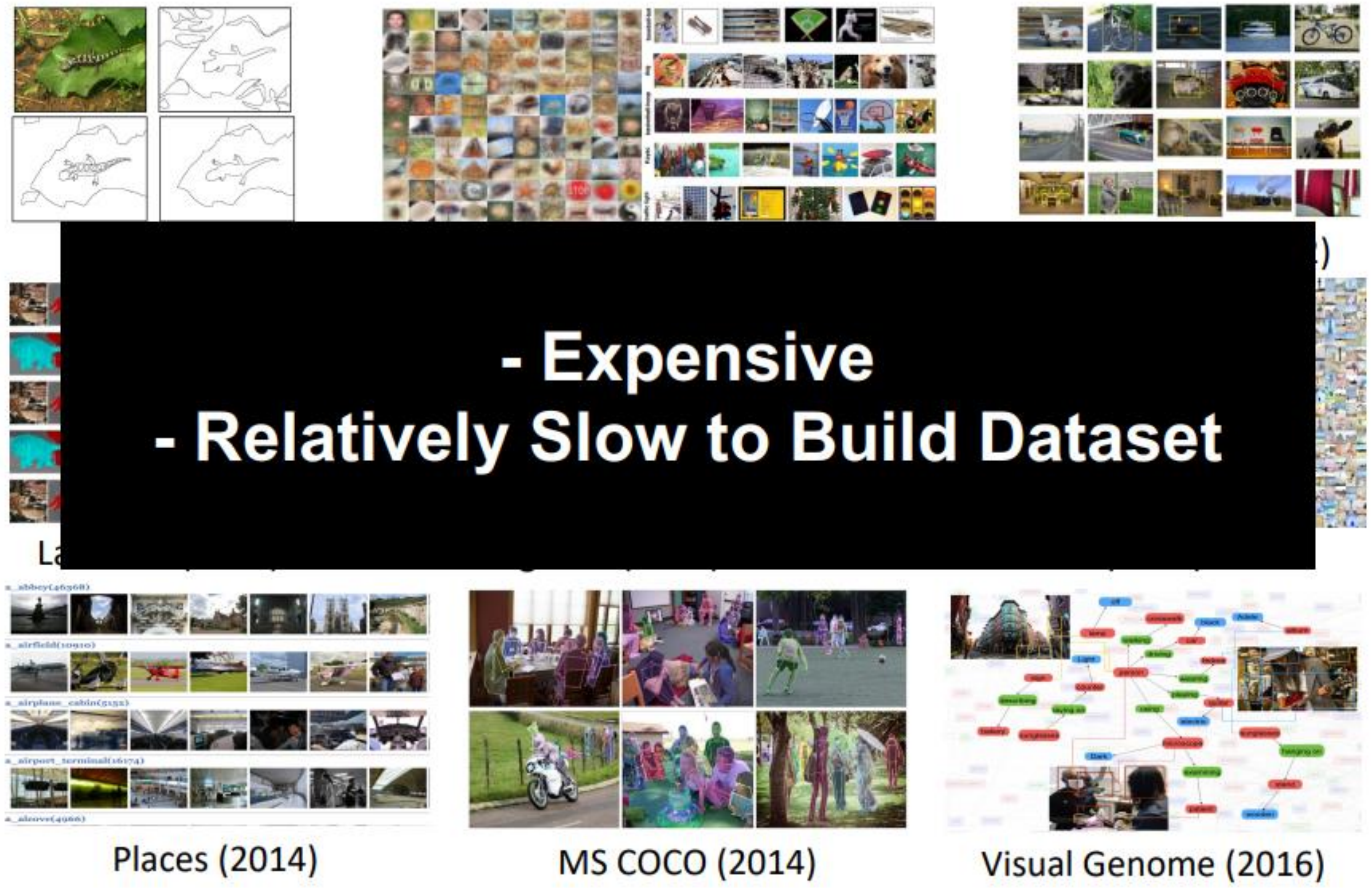


Transfer Learning

Transfer Learning

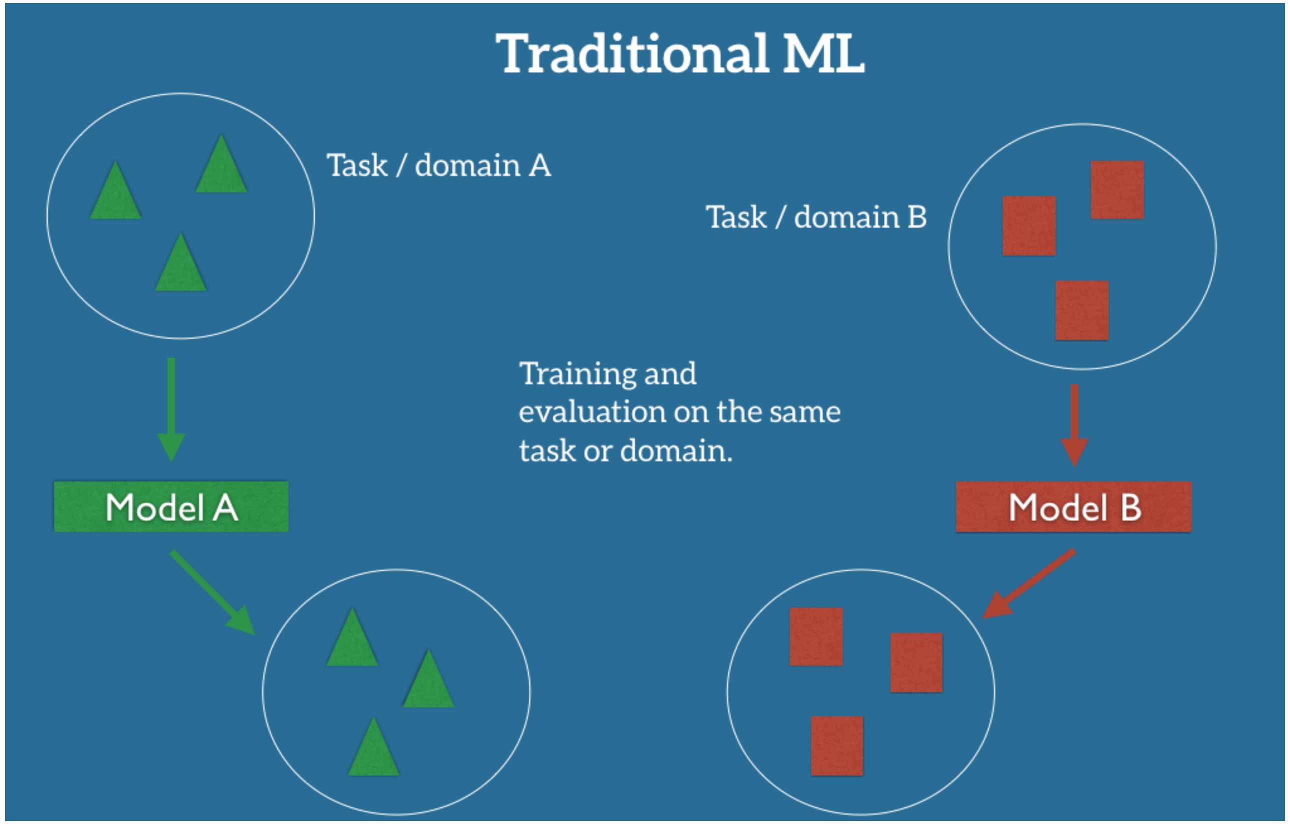
UNIVERSIDADE CATÓLICA PORTUGUESA
BRAGA

- Goal: : Avoid Always Relying on Large Labeled Datasets



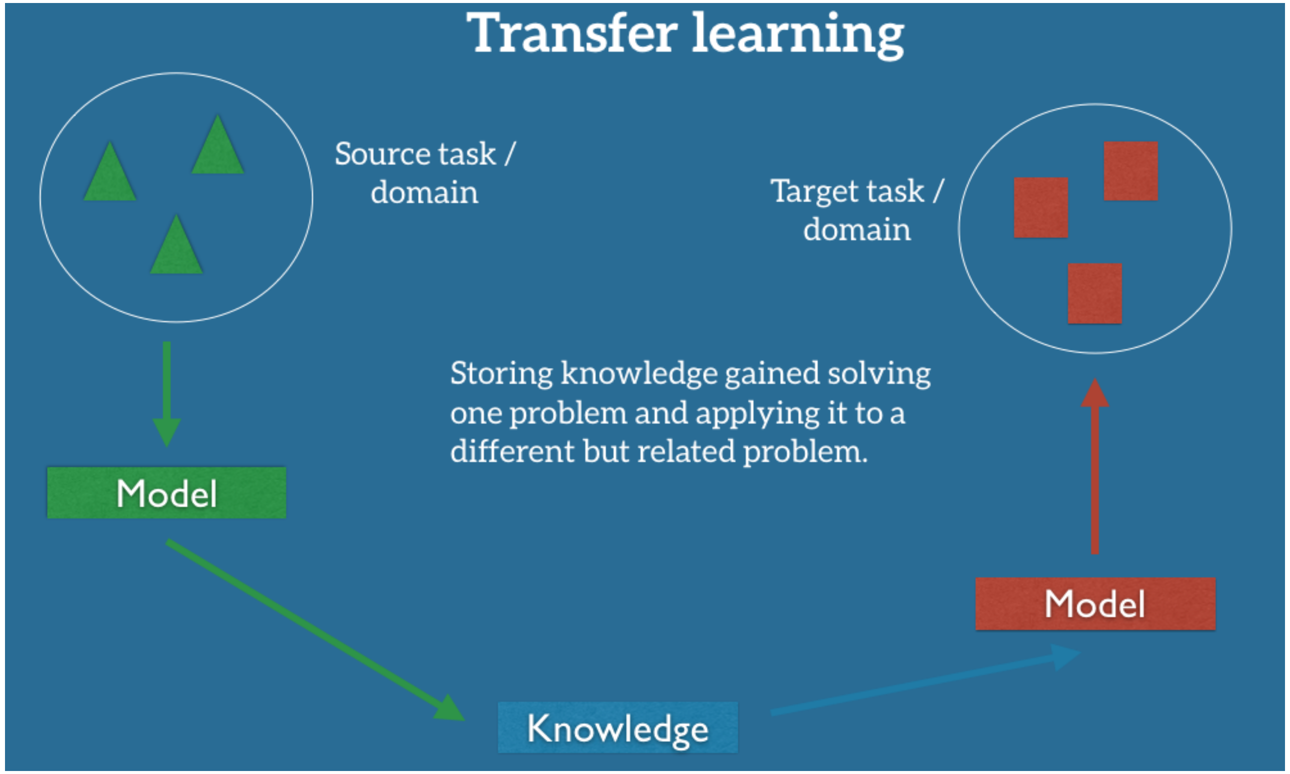
Transfer Learning

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



Transfer Learning

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



Transfer Learning

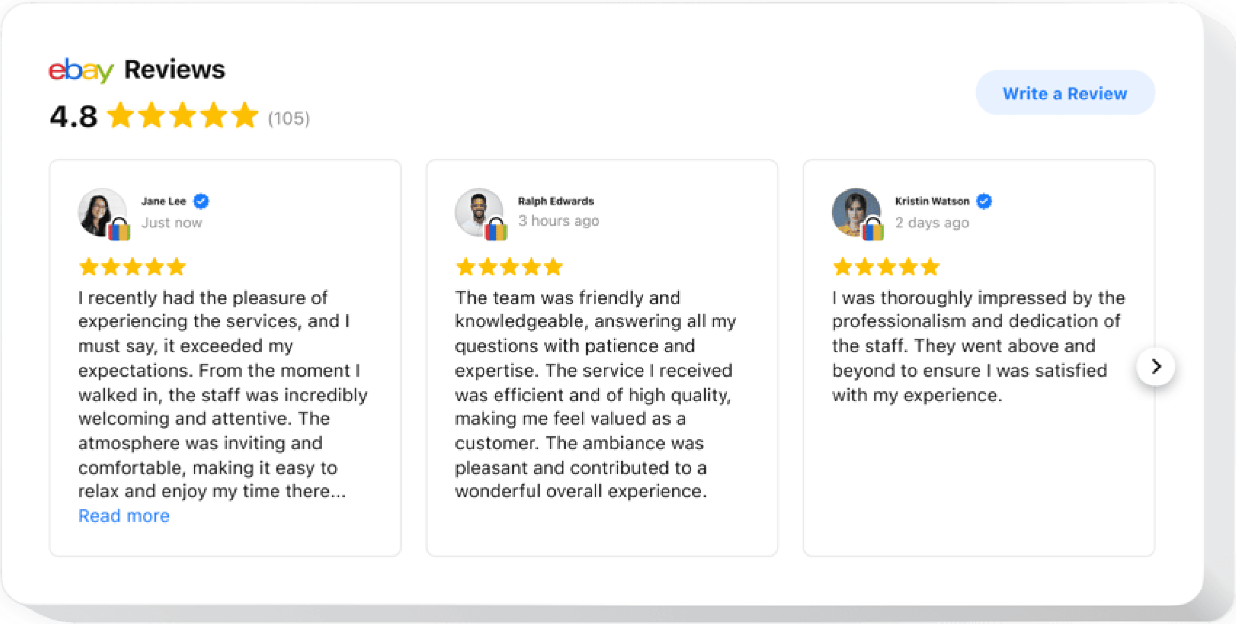
UNIVERSIDADE CATÓLICA PORTUGUESA
BRAGA

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

News (formal and lengthy) **Tweets** (informal and brief)

Transfer Learning

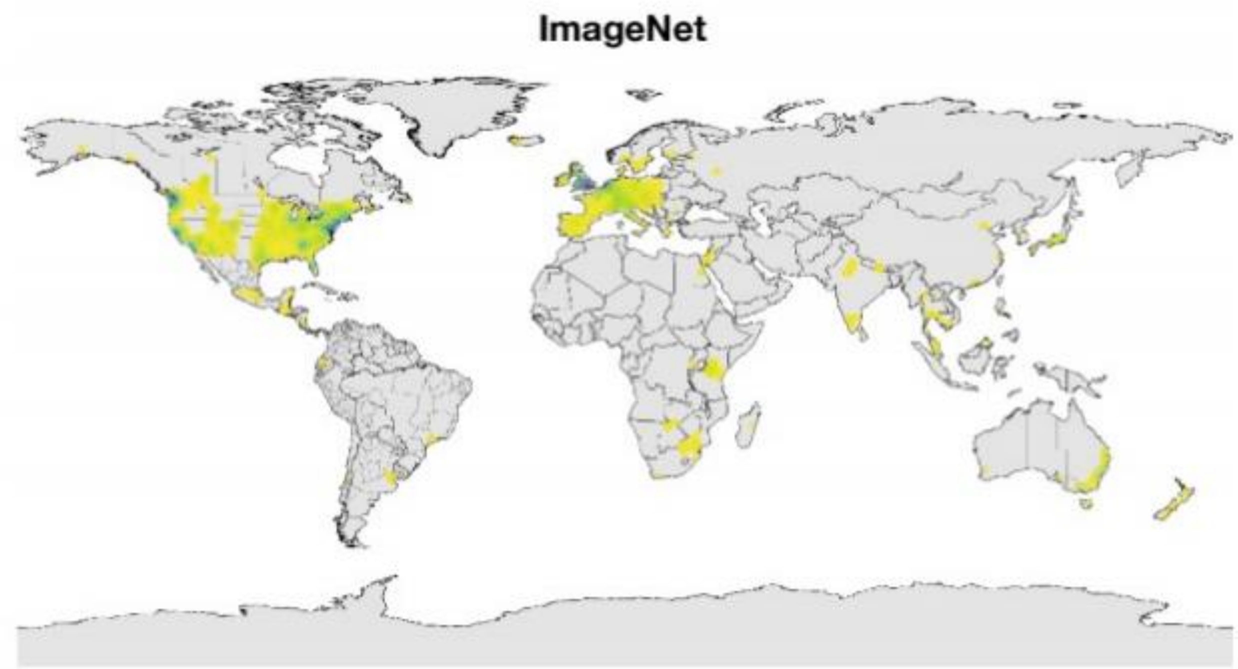
- Transfer Learning When Feature Space Changes (e.g., Sentiment Classification in Different Language)

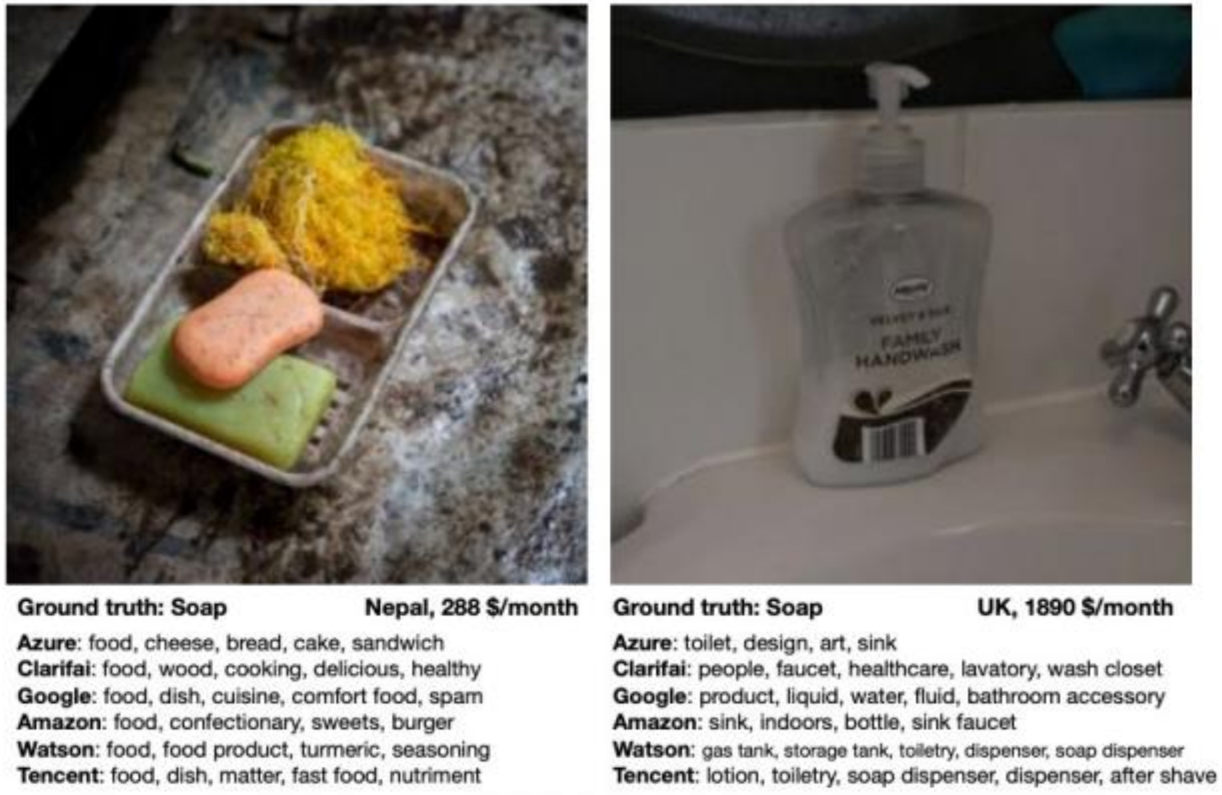


Transfer Learning

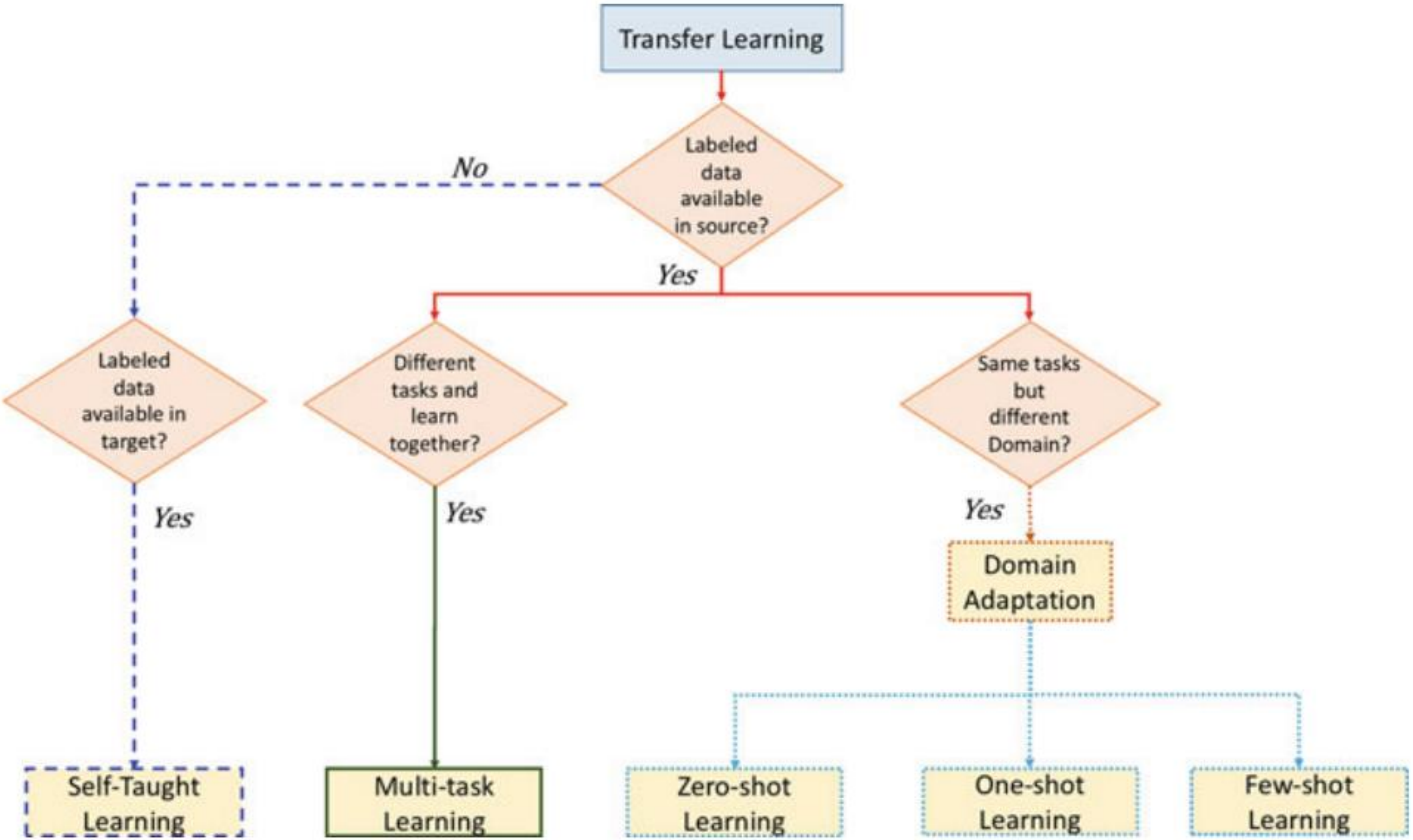
UNIVERSIDADE CATÓLICA PORTUGUESA
BRAGA

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





Transfer Learning Approaches

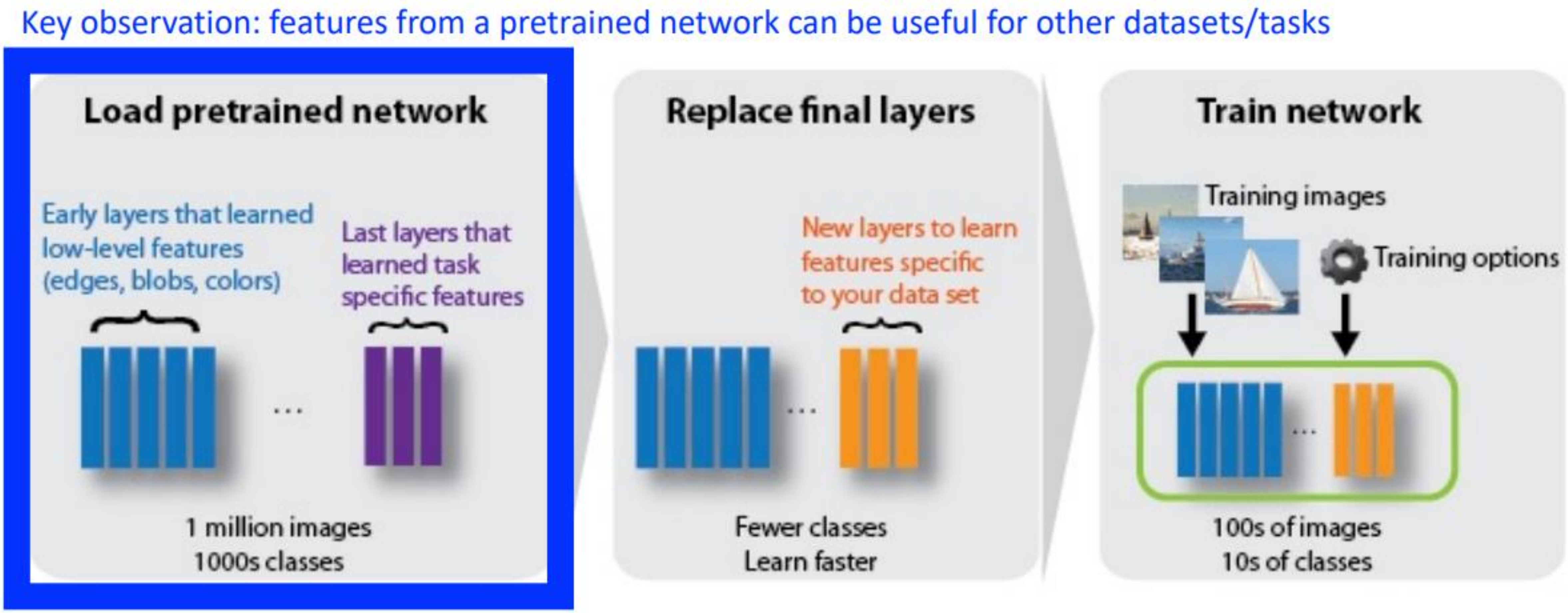


Transfer Learning: Key Challenges

- **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



Transfer Learning: Self-Supervised

- How Do Humans Learn?

With Supervision

Learn from instruction

Unsupervised

Learn from experience

Self-Supervised Learning

UNIVERSIDADE CATÓLICA PORTUGUESA
BRAGA

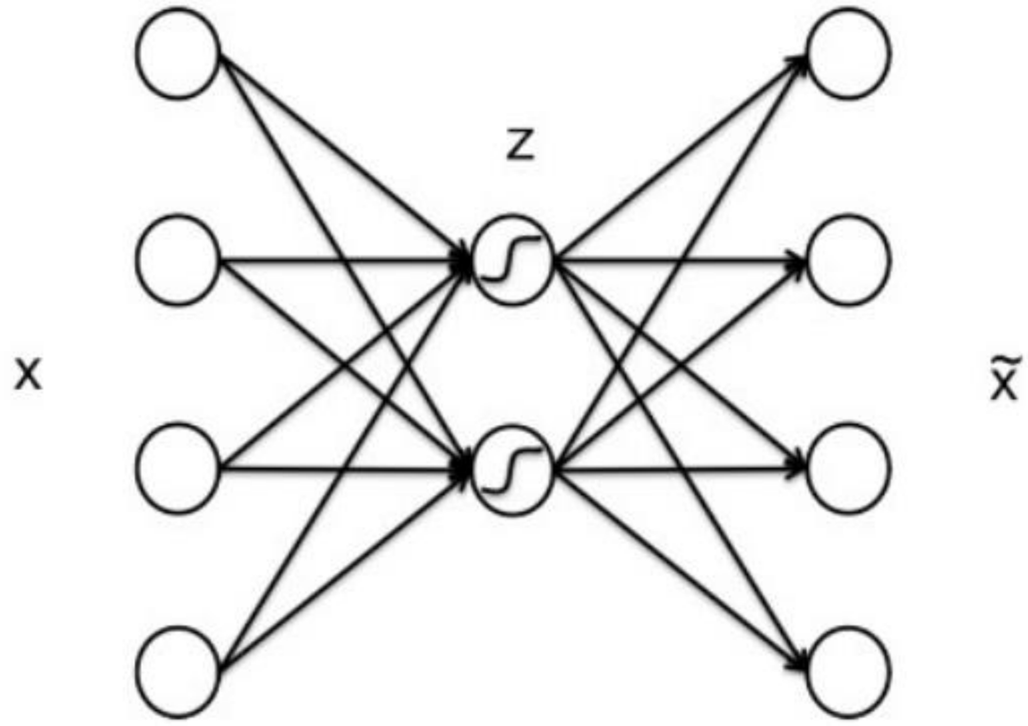
- Data Gives Supervision

- Relatively Cheap
- Can Collect Data Fast

Autoencoders

UNIVERSIDADE CATÓLICA PORTUGUESA BRAGA

- Learn to copy the input to the output



Autoencoders

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

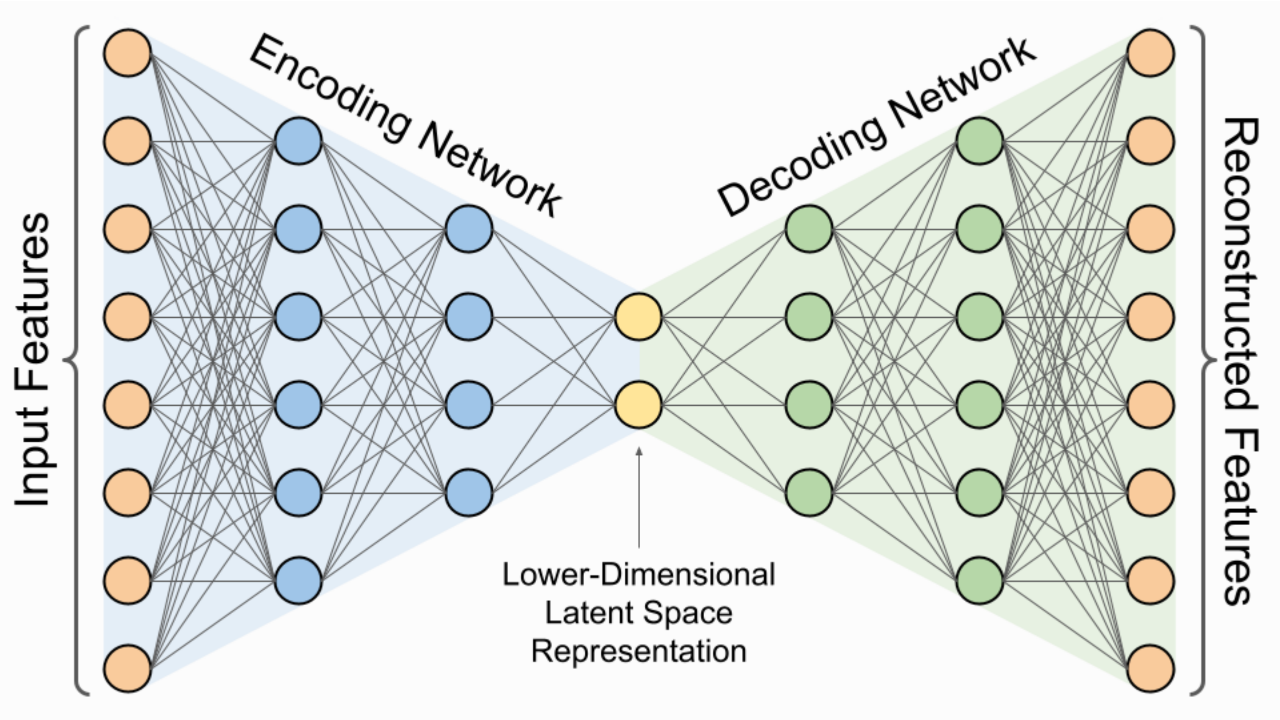
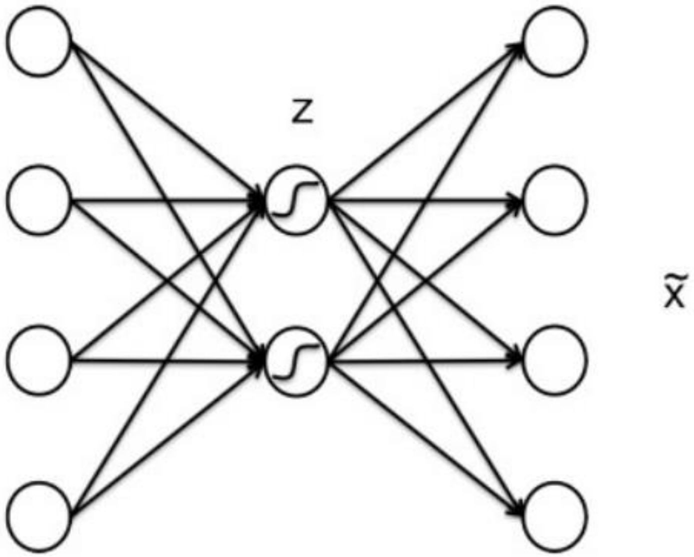


Image Autoencoder Architecture

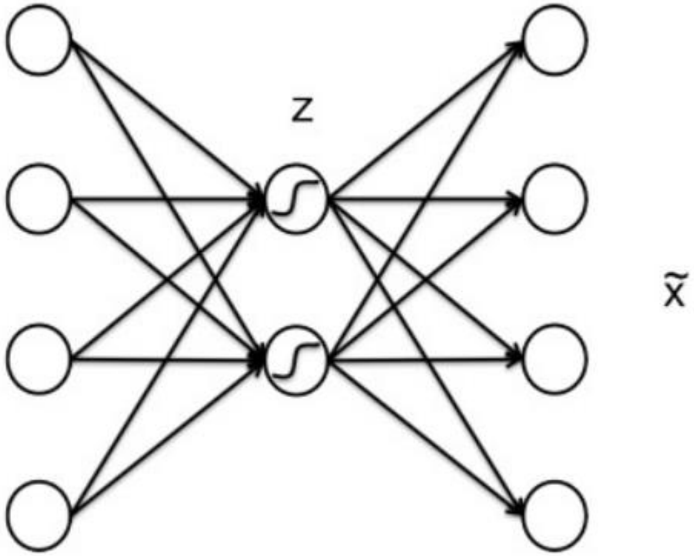
- Given this input 620 x 426 image (264,120 pixels):



- What would a perfect autoencoder predict?

Image Autoencoder Architecture

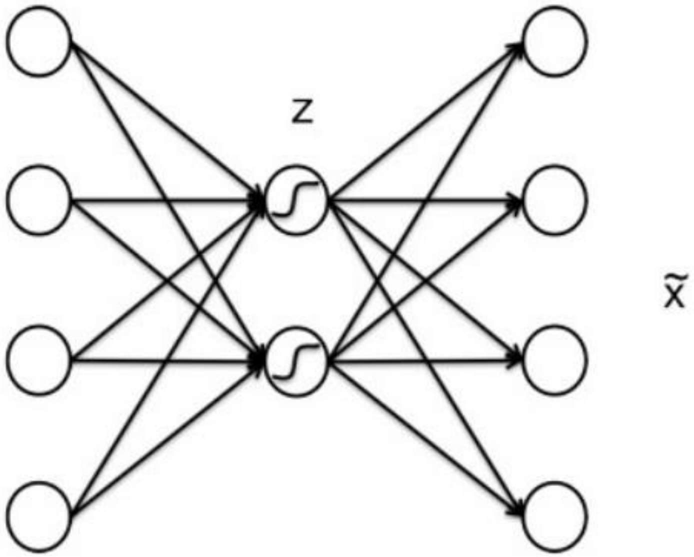
- Given this input 620 x 426 image (264,120 pixels):



- What would a perfect autoencoder predict?
  - itself

Image Autoencoder Architecture

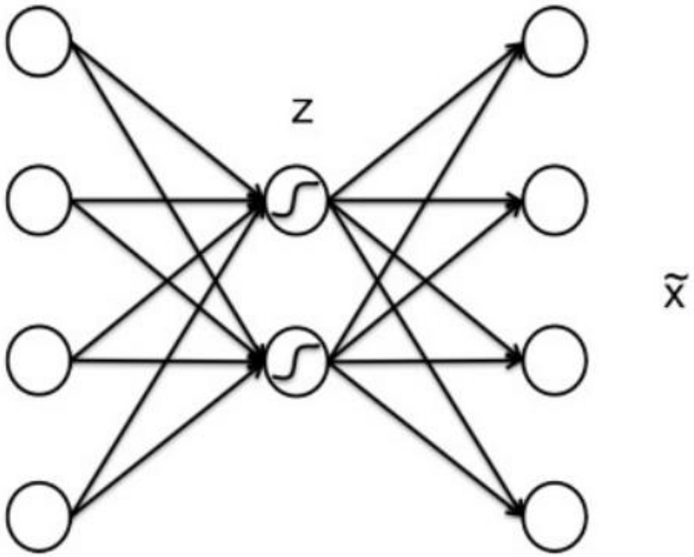
- Given this input 620 x 426 image (264,120 pixels):



- What number of nodes are in the final layer?

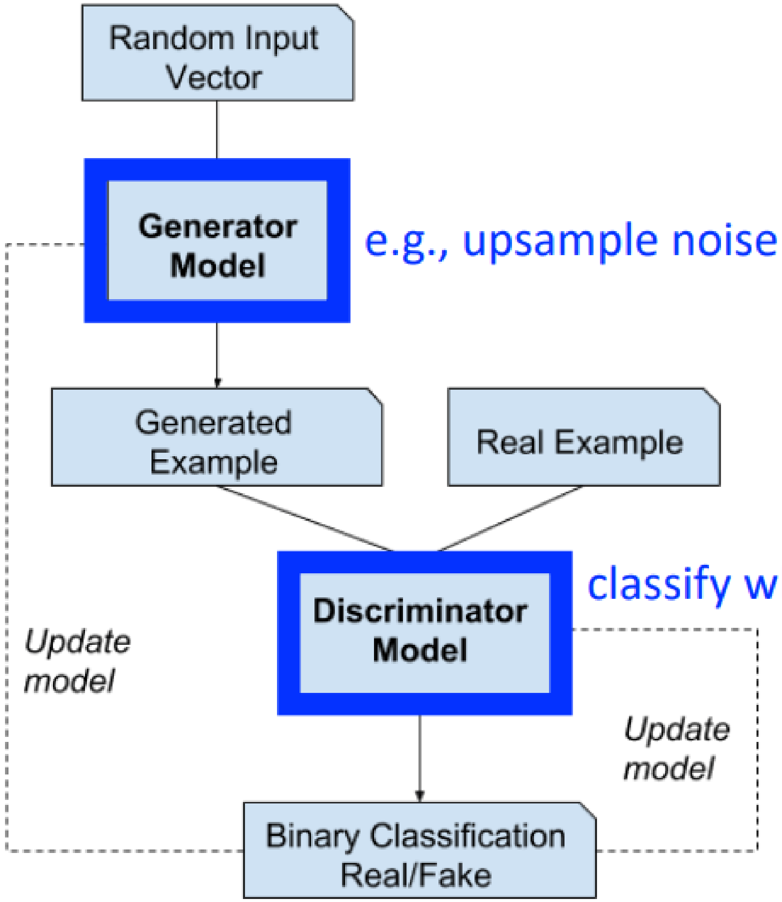
Image Autoencoder Architecture

- 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)

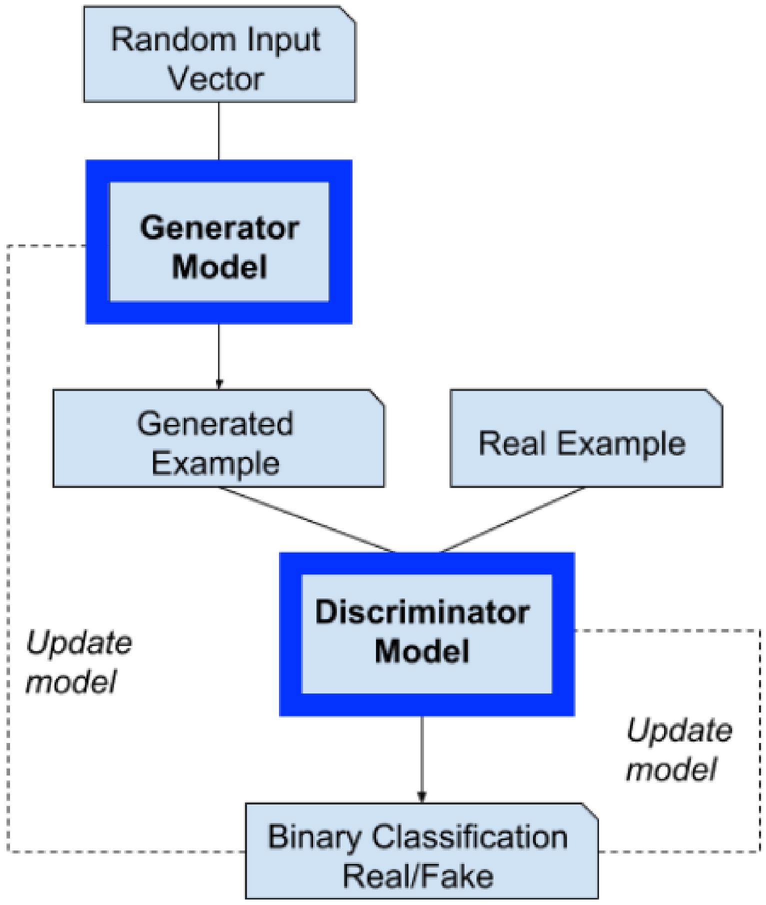


e.g., upsample noise to image of desired size

classify which examples are real

Consists of two models that compete against each other

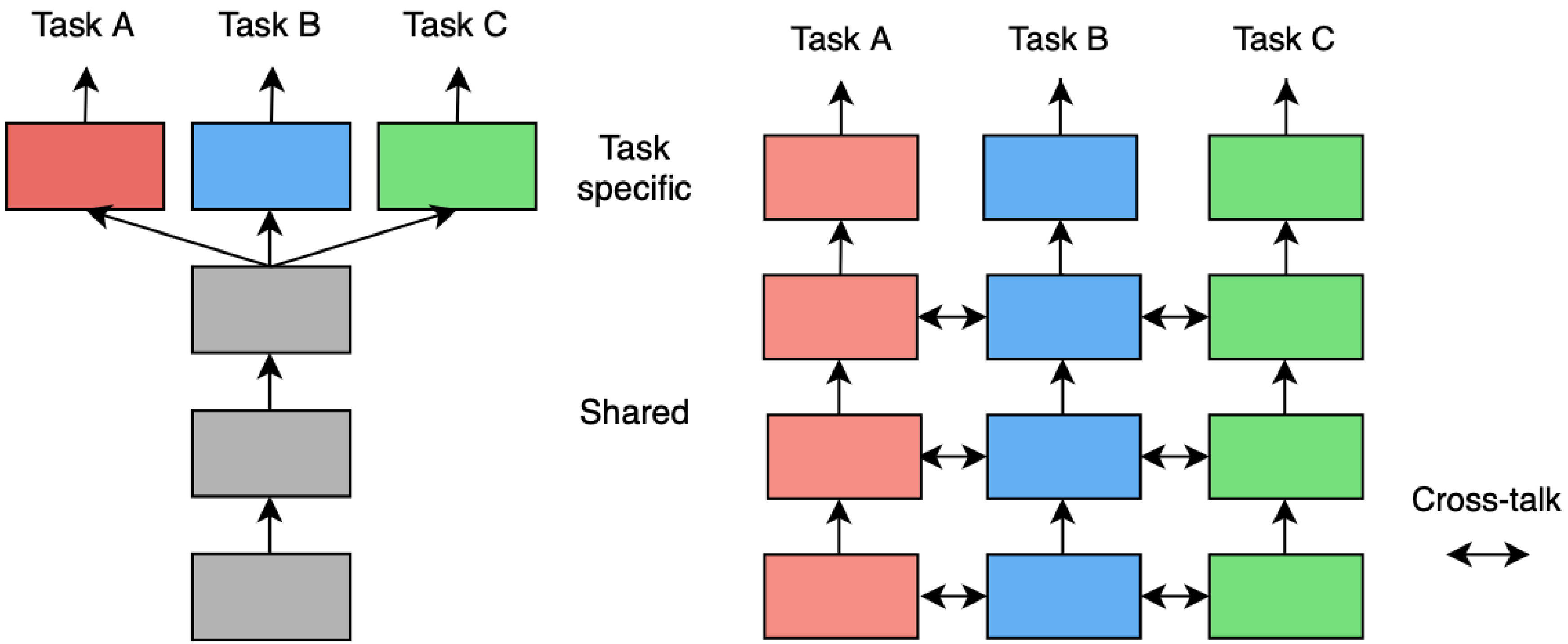
GAN Training



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

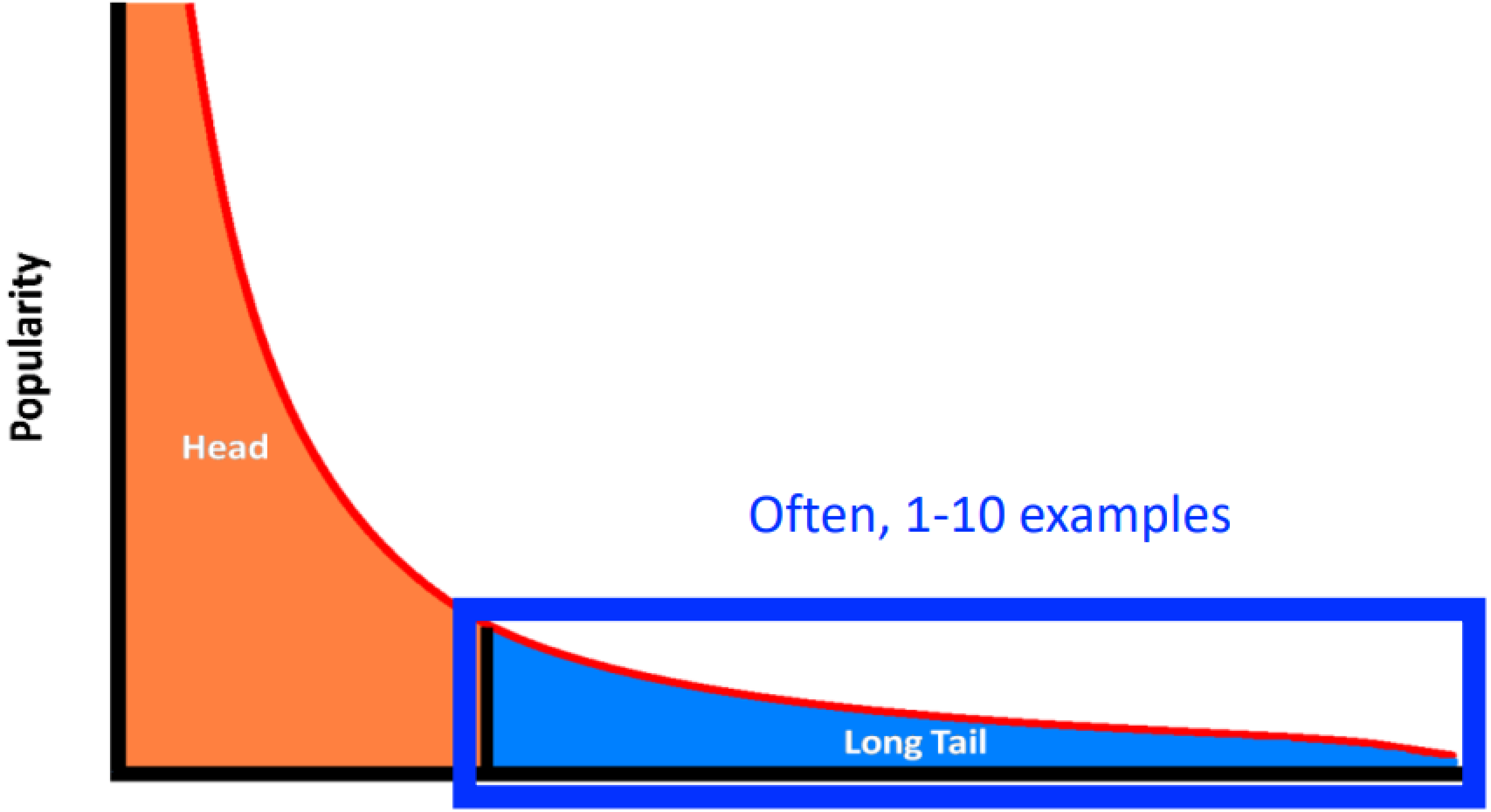
UNIVERSIDADE CATÓLICA PORTUGUESA
BRAGA

- 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

UNIVERSIDADE CATÓLICA PORTUGUESA
BRAGA

- 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

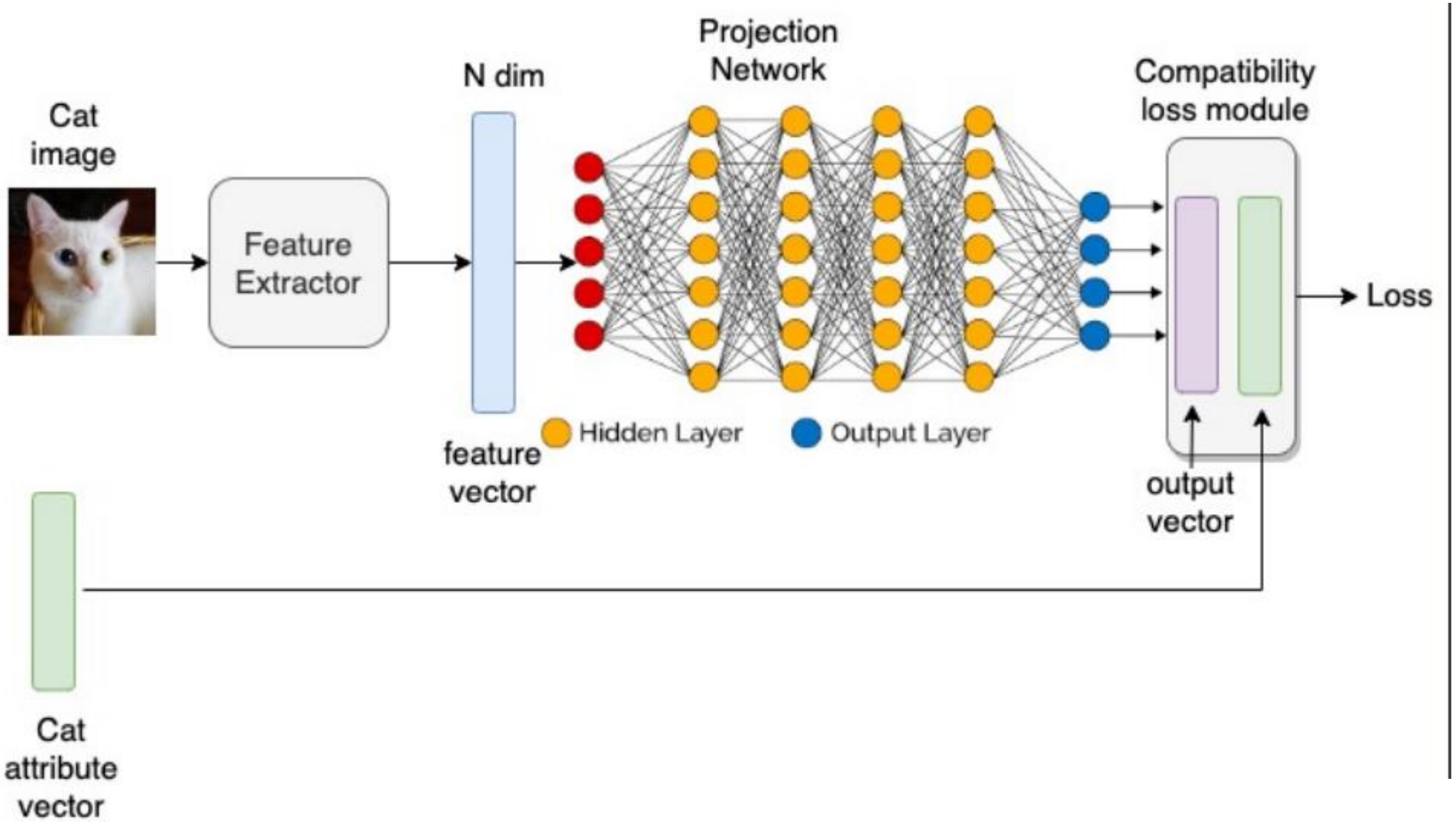
Zero/Few-Shot Learning

- 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

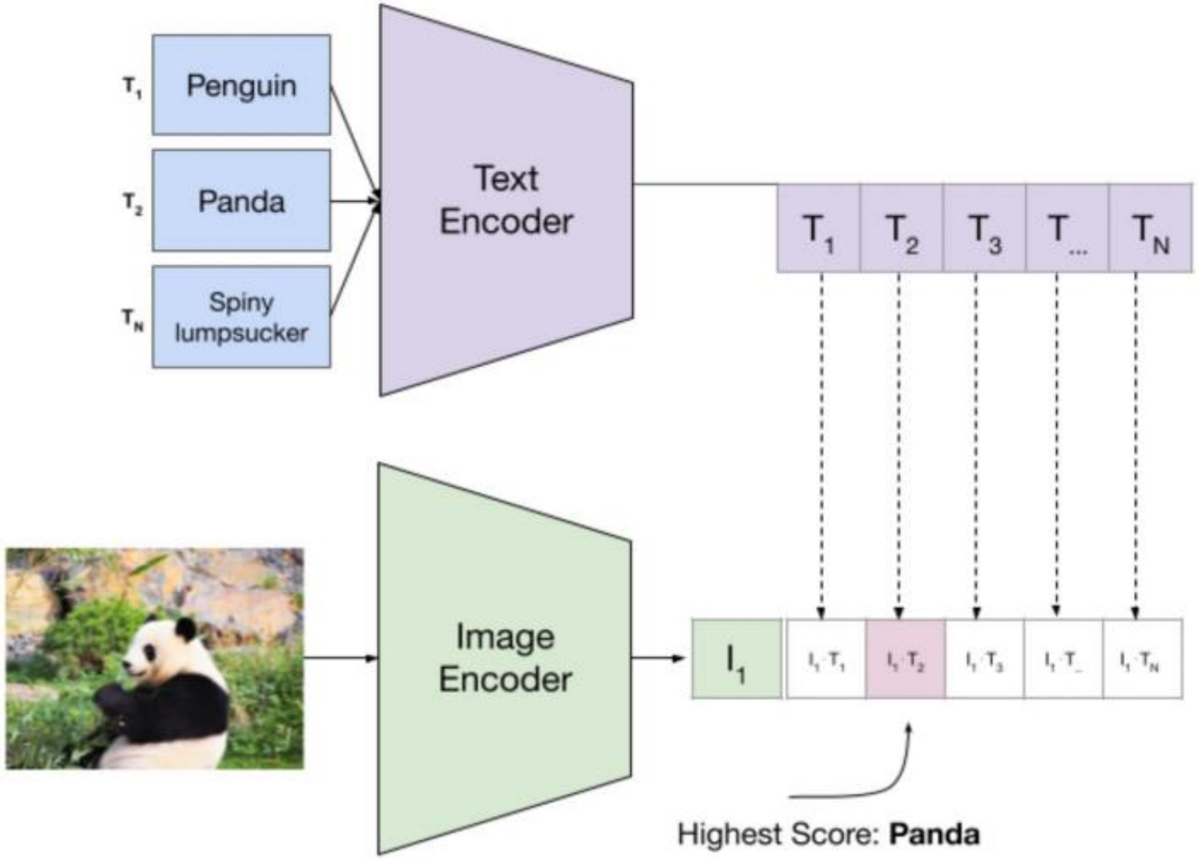
- Key Idea: Learn from Auxiliary Labels How to Perform a Different Task with Zero/Few Training Examples



Learns auxiliary attribute labels

Zero/Few-Shot Learning

- Contrastive Language-Image Pretraining (CLIP)



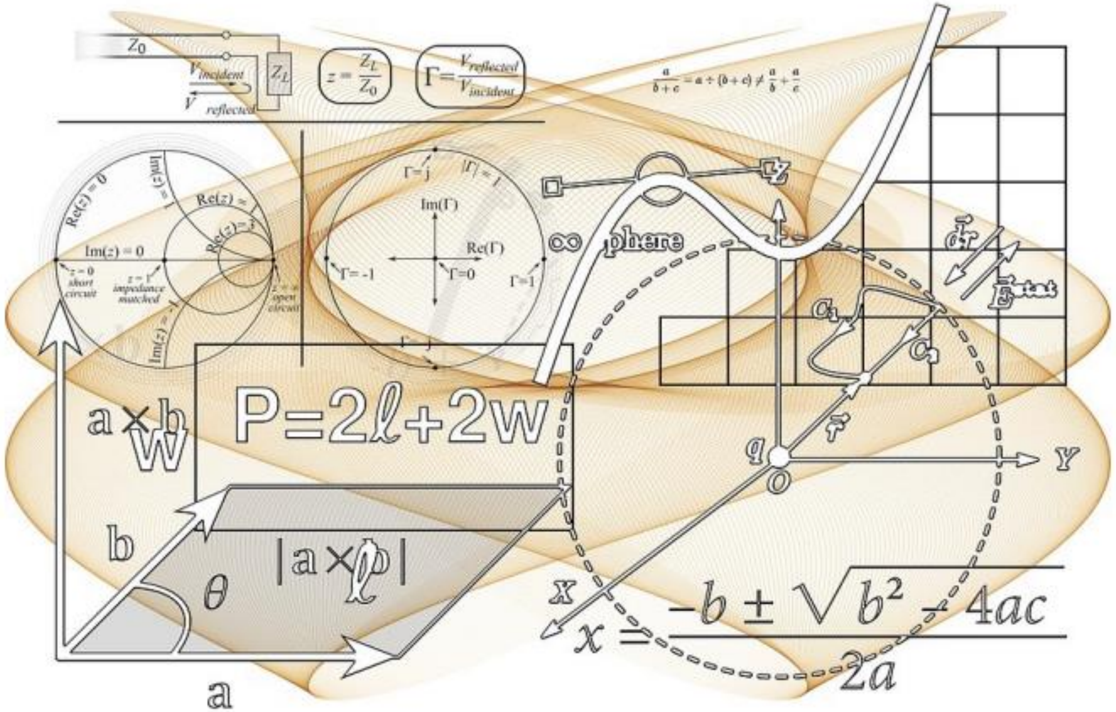
Efficient Learning

UNIVERSIDADE CATÓLICA PORTUGUESA
BRAGA

Efficient Learning

- How to teach machines so they learn faster?

Random Order of Examples



Meaningful Order of Examples


Table of Contents

Efficient Learning

UNIVERSIDADE CATÓLICA PORTUGUESA
BRAGA

- Intuition: How to Teach a Child To Read?

Random Order of Examples

Meaningful Order of Examples

Efficient Learning

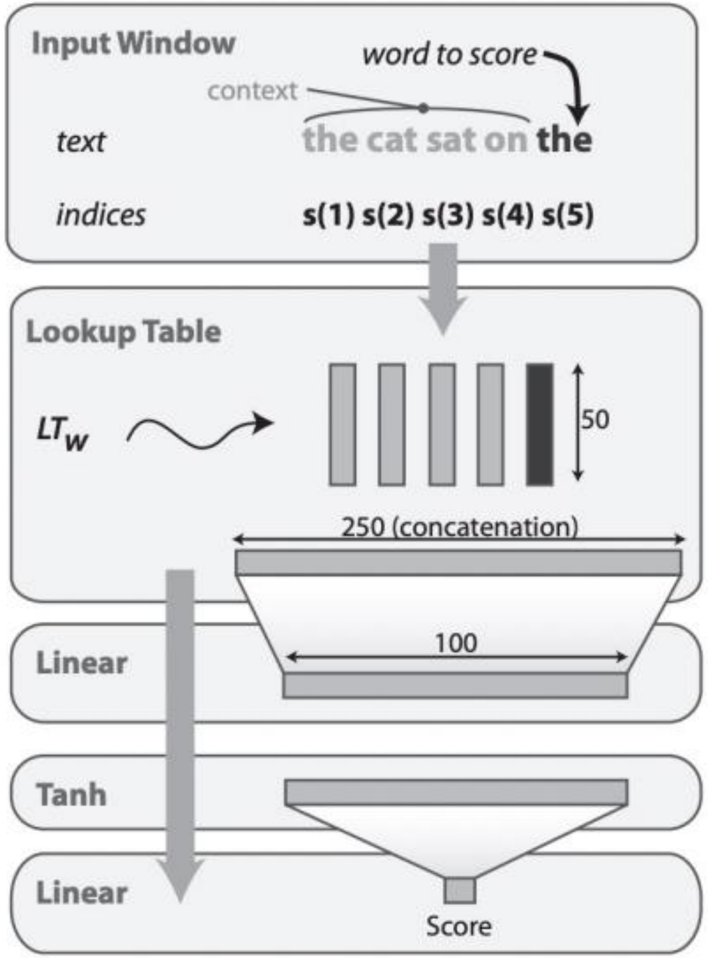
- Curriculum Learning

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

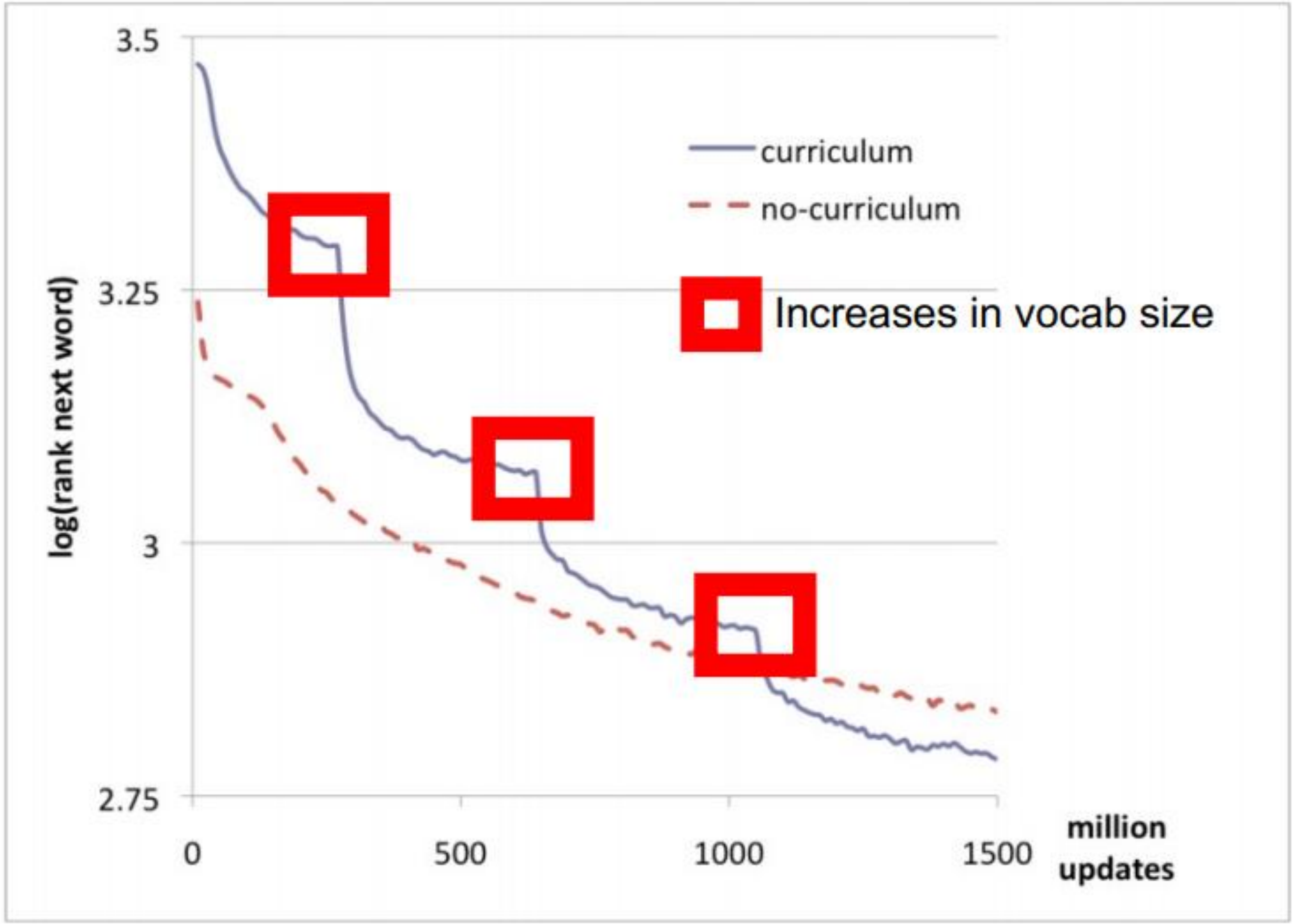


2. Predict the next word

Background music from a ________

Efficient Learning

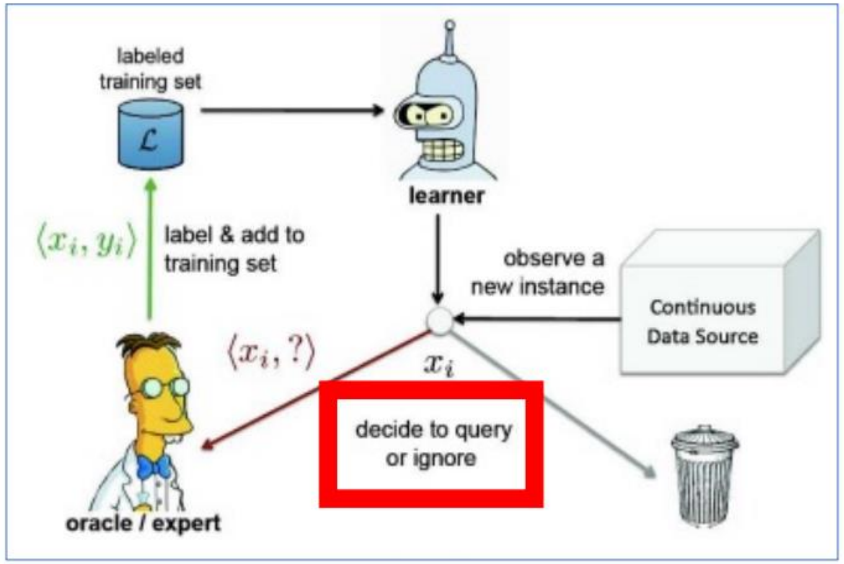
- Curriculum Learning: Next Word Prediction



Efficient Learning

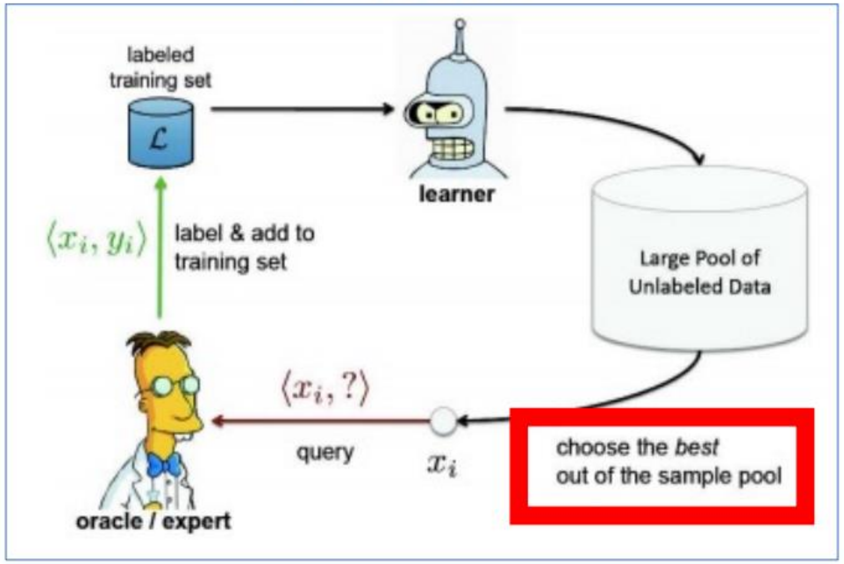
- Active Learning:
  - Actively select the examples to label that would be most effective for learning rather than labelling all available data.
- Types of Active Learning:

Stream-Based



Consider one example at a time

Pool-Based

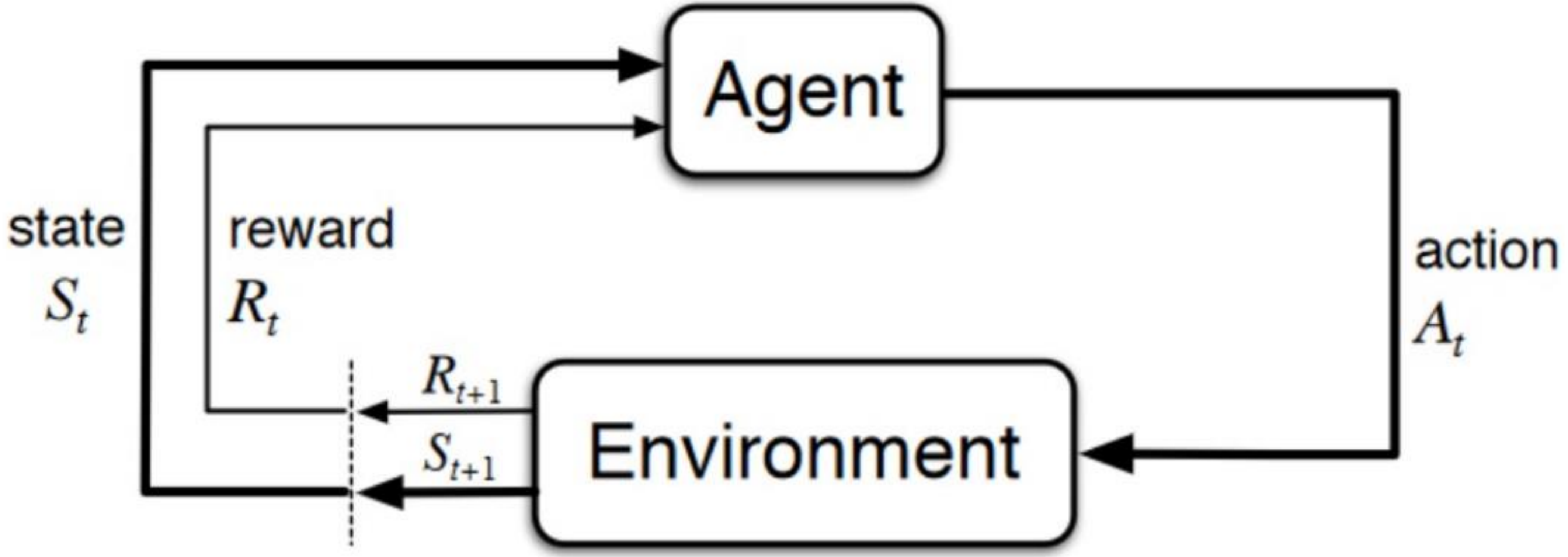


Consider many examples at a time

Reinforcement Learning

Reinforcement Learning

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



Reinforcement Leraning

- 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

UNIVERSIDADE CATÓLICA PORTUGUESA
BRAGA

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

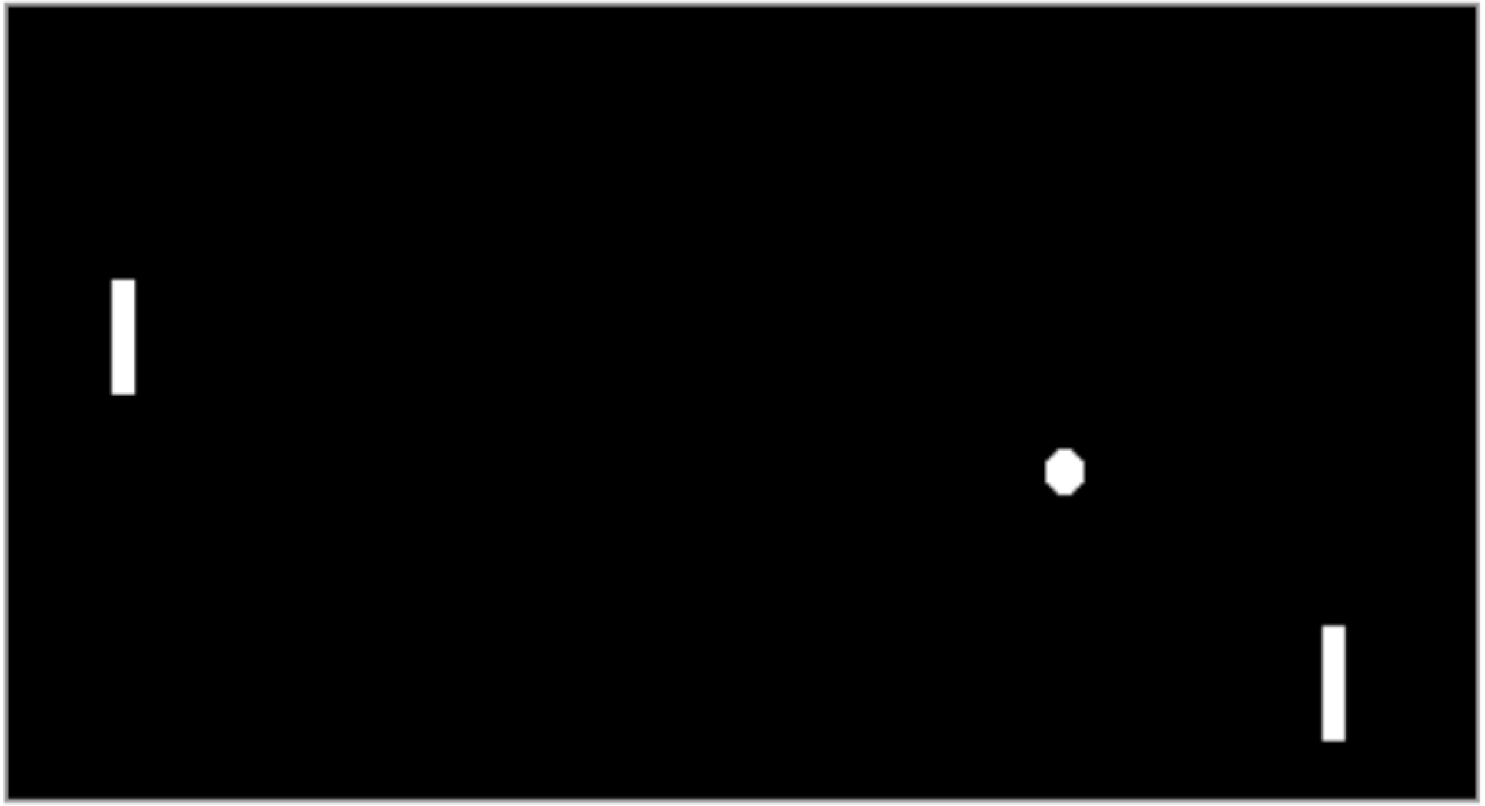
Arne Voigtländer
Shoogee GmbH & Co. KG
Krögerweg 16a
D-48155 Münster, Germany
Email: arne@shoogee.com

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

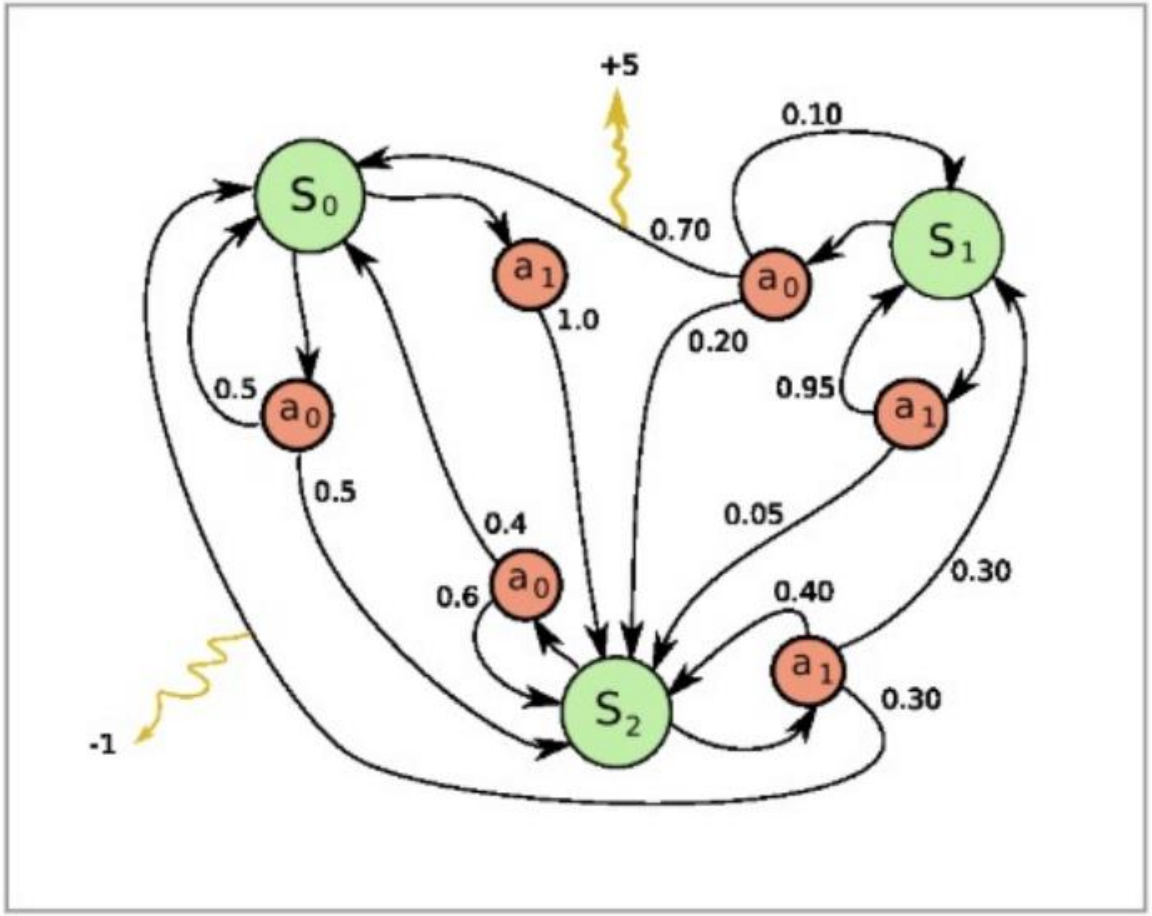
- Pong Game Learning Example

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

-1 if missed the ball

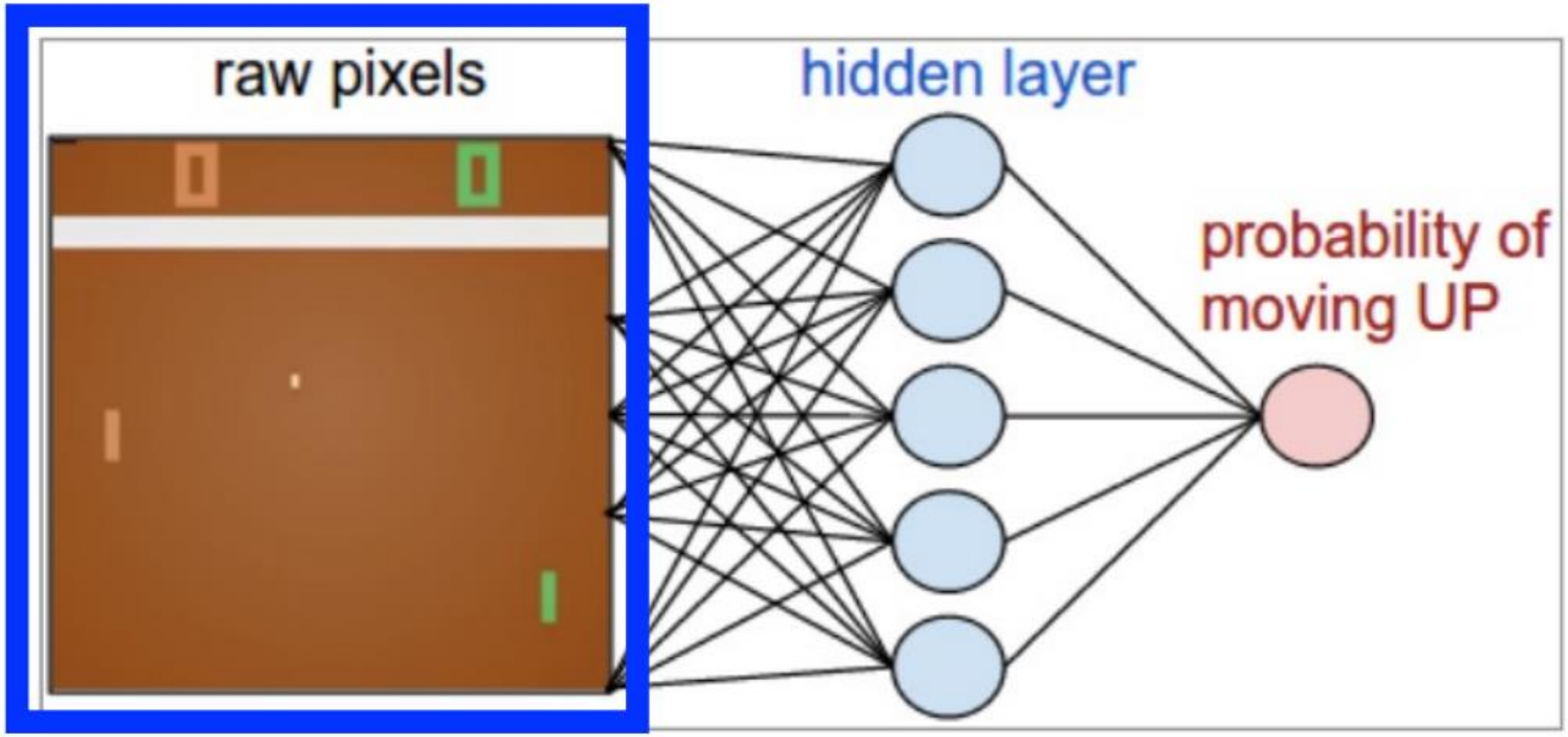
+1 reward if ball goes past opponent

0 otherwise



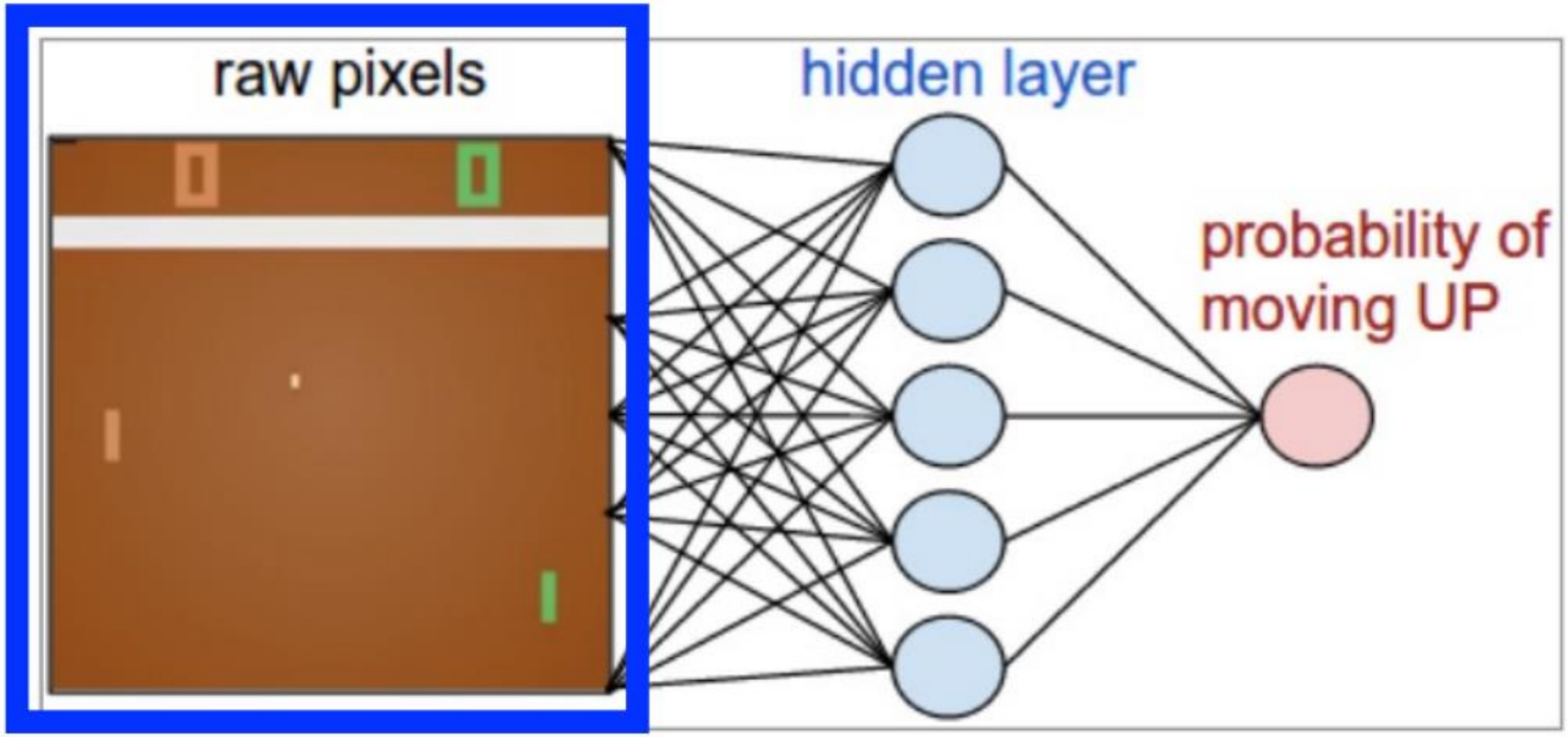
Reinforcement Learning Applications

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



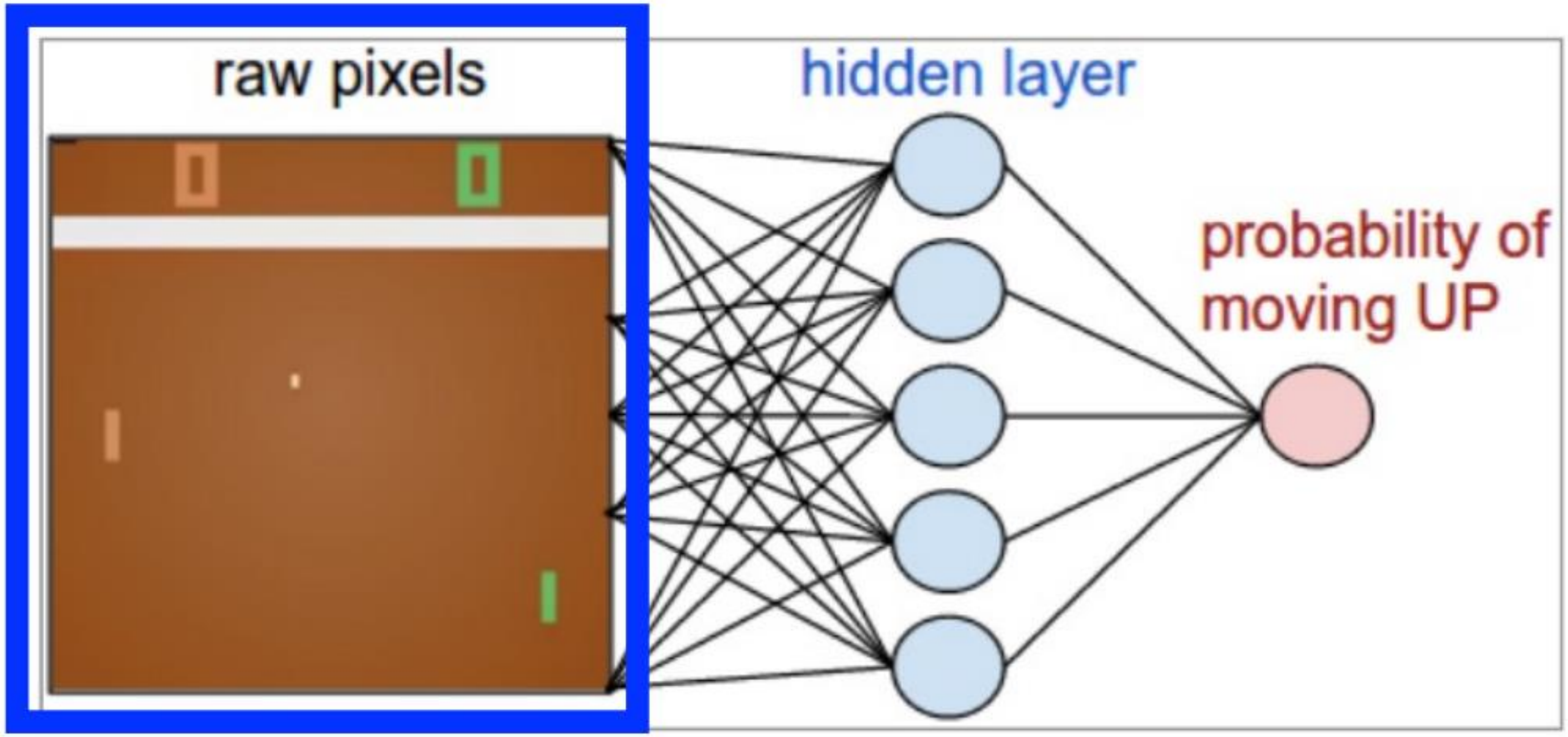
Reinforcement Learning Applications

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



Reinforcement Learning Applications

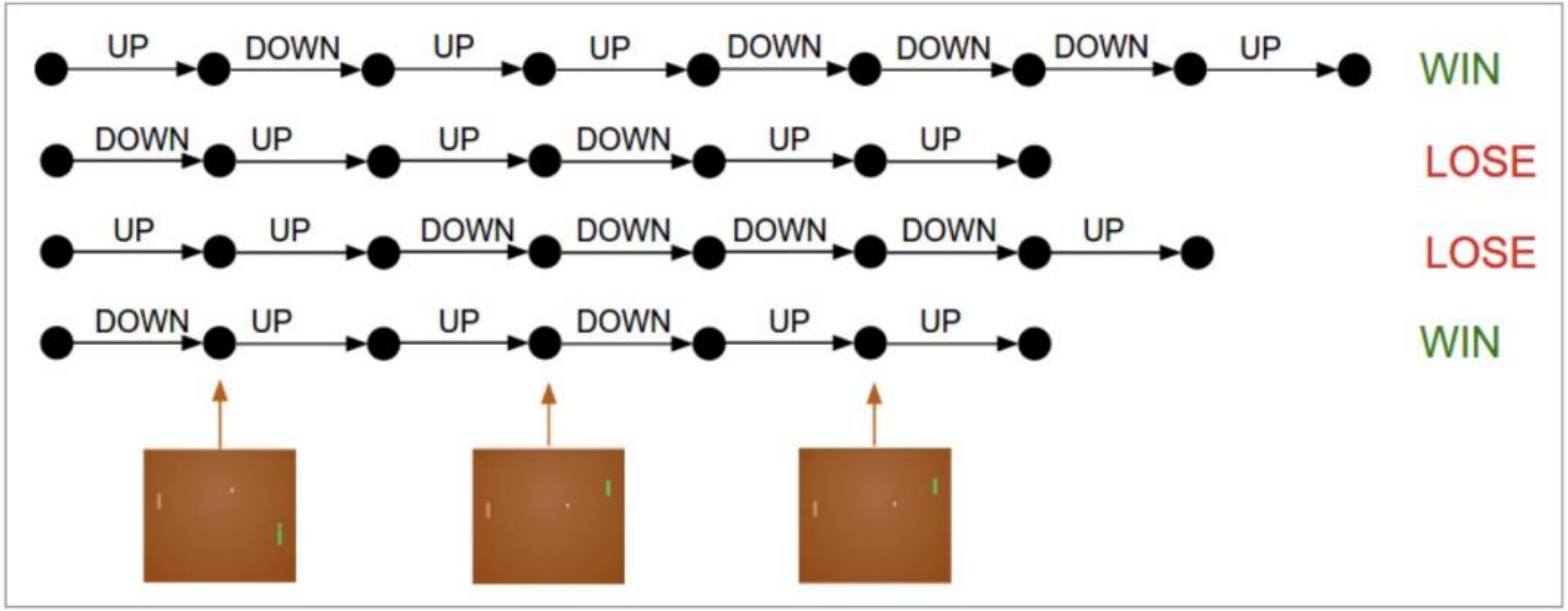
- Pong Game: Policy Network
  - Problem: reward may be due to good action **many steps ago**



Reinforcement Learning Applications

UNIVERSIDADE CATÓLICA PORTUGUESA
BRAGA

- 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