Zero-Shot Learning

Foundations, Methodologies and Applications

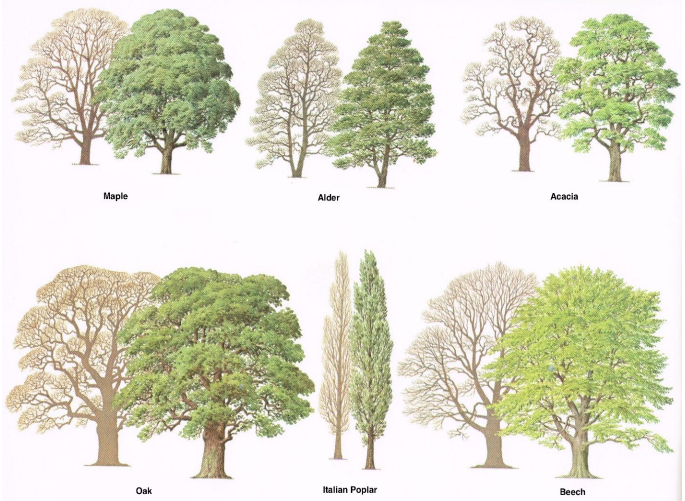
Sandipan Sarma

Introduction

Problems in supervised learning

➔ For any classification task, supervised machine learning approaches need training examples of the objects to classify. But such examples may not be always accessible:

Humans can classify about 30,000 object categories



Labelling all the classes is expensive - requires expert supervision

Target classes can be rare

Possible solution : N-Shot Learning

- ➔ What is meant by a shot ? : A single example that a model can see during training
- ➔ When data is scarce, learn a model with **as low shots (i.e. very low N)** at training examples as possible, from various classes
- ➔ **Achievement** : Our deep model learns to classify an unseen image even after training with very few images ! - contrary to usual supervised DNNs
- ➔ **Variants:**
  - ◆ Zero-Shot Learning : Classify test image without training at a single image of that class !
  - ◆ One-Shot Learning : Classify test image after training at a single image of that class !
  - ◆ Few-Shot Learning : Classify test image after training at a few images (≤ 5) of that class!

Zero-Shot : Intuition

Appearance+Properties+Functionality = Knowledge

Dark + Night sky + Stars + W-like pattern = Constellation

Appearance+Properties+Functionality = Knowledge

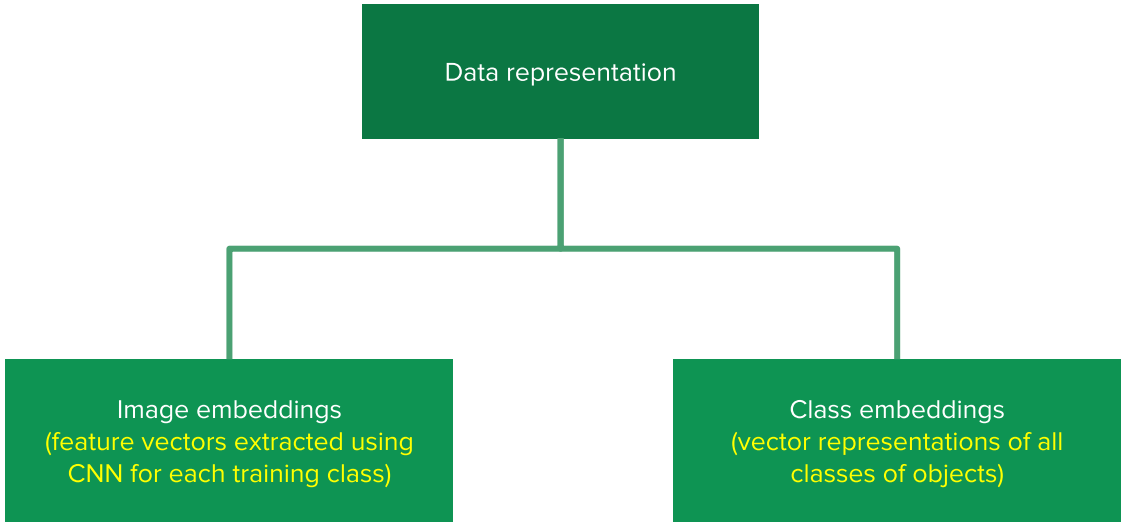
Quadruped + Fast + Horse-like mane + Stripes = Zebra

Diving into Zero-Shot Learning (ZSL)

Zero-Shot Learning - Representation

Sandipan Sarma

- ➔ A paradigm that attempts to classify unseen categories of objects by associating them, using some visual-semantic mappings, with known classes of objects on which training is done
- ➔ We have : training classes and target classes - need to be related using some mapping function



CLASSES		Image Embedding	Class Embedding
TRAINING	A	✔	✔
	B	✔	✔
	C	✔	✔
	D	✔	✔
	E	✔	✔
ZERO SHOT	F	✘	✔
	G	✘	✔
	H	✘	✔

Notations and definition

- ➔ $S = \{ c^s_i : i = 1 \text{ to } N_s \}$ = seen classes
- ➔ $U = \{ c^u_i : i = 1 \text{ to } N_u \}$ = unseen classes

} Disjoint sets

- ➔ X = feature space
- ➔ $D^{tr} = \{ (x^{tr}_i , y^{tr}_i) \in X \text{ ❌ } S ; i = 1 \text{ to } N_{tr} \}$ = training data (from seen classes)
- ➔ $X^{te} = \{ x^{te}_i \in X ; i = 1 \text{ to } N_{te} \}$ = testing data
- ➔ $Y^{te} = \{ y^{te}_i \in U ; i = 1 \text{ to } N_{te} \}$ = testing labels

Task formulation :

Given $D^{tr} \in S$, learn a classifier $f^u(.) : X \rightarrow U$, that can classify instances from X^{te} into a class from Y^{te}

- ➔ Source feature space = feature space of training data
- ➔ Target feature space = feature space of testing data

Auxiliary information in ZSL

- ➔ Some usable information about the seen and unseen classes related to the corresponding feature space is necessary - this is known as the **semantic information or class embedding**.
- ➔ Each class has a corresponding vector representation in the semantic space, called the **class prototype** of that class.
- ➔ Semantic space (T):
  - ➔ $T^s : \{ t^s_i : i = 1 \text{ to } N_s \}$ = seen class prototypes
  - ➔ $T^u : \{ t^u_i : i = 1 \text{ to } N_u \}$ = unseen class prototypes

In ZSL, along with D^{tr}, T^s and T^u are also required to obtain the classifier $f^u(.)$

Dataset example - Animals with Attributes (AwA-2)

Animals with Attributes (AwA-2)

Sandipan Sarma

Classes : 50
Image count : 37,322
Detail : Coarse-grained

Refer : LAMPERT ET AL.: ATTRIBUTE-BASED CLASSIFICATION FOR ZERO-SHOT VISUAL OBJECT CATEGORIZATION

Classes and attributes

Animal Classes of the *Animals with Attributes* Data Set

skunk	polar bear	beaver	giraffe	*leopard*
lion	killer whale	bobcat	wolf	*pig*
fox	grizzly bear	collie	tiger	*hippopotamus*
ox	chihuahua	otter	cow	*seal*
mole	dalmatian	antelope	weasel	*persian cat*
sheep	spider monkey	hamster	mouse	*chimpanzee*
horse	blue whale	squirrel	buffalo	*rat*
bat	siamese cat	elephant	moose	*humpback whale*
zebra	rhinoceros	rabbit	walrus	*giant panda*
deer	german shepherd	dolphin	gorilla	*raccoon*

The 40 classes of the first four columns are used for training, the 10 classes of the last column (in italics) are the test classes.

Eighty-Five Semantic Attributes of the *Animals with Attributes* Data Set in Short Form

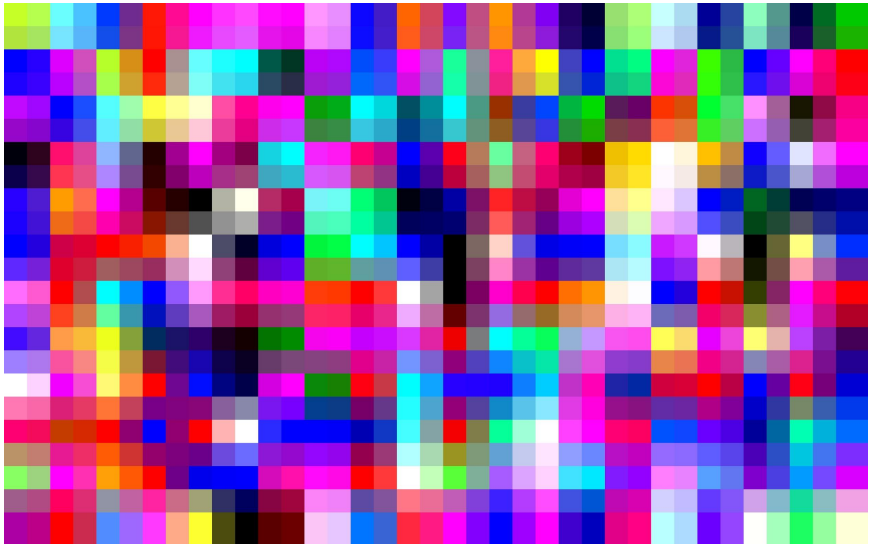
black	toughskin	tail	bipedal	stalker	mountains
white	bulbous	horns	active	skimmer	water
blue	lean	claws	inactive	cave	newworld
brown	flippers	tusks	nocturnal	fierce	oldworld
gray	hands	smelly	hibernate	arctic	timid
orange	hooves	flies	agility	coastal	smart
red	longleg	hops	fish	desert	group
yellow	pads	swims	meat	bush	solitary
patches	paws	tunnels	plankton	plains	nestspot
spots	longneck	walks	vegetation	forest	domestic
stripes	chewteeth	fast	insects	fields	
furry	meatteeth	slow	forager	jungle	
hairless	buckteeth	strong	grazer	tree	
big	strainteeth	weak	hunter	ocean	
small	quadrapedal	muscle	scavenger	ground	

Longer forms given to human subject for annotation were complete phrases, such as has flippers, eats plankton, *or* lives in water.

Refer : LAMPERT ET AL.: ATTRIBUTE-BASED CLASSIFICATION FOR ZERO-SHOT VISUAL OBJECT CATEGORIZATION

Data representations

Image embeddings (from pretrained ResNet-101)



Dimension : num_images x 2048

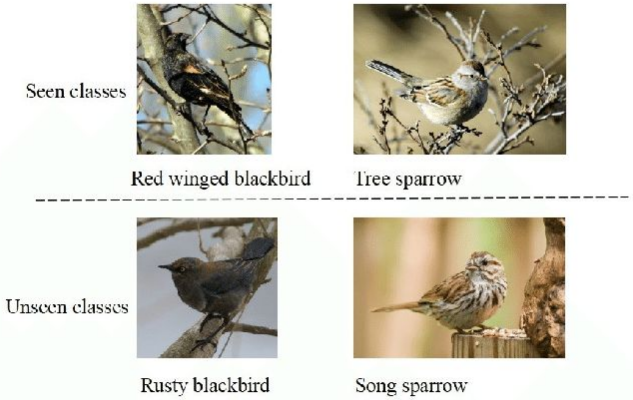
Class embeddings / Semantic matrix

	black	white	blue	brown	gray	orange	red	yellow	patches	spots	...	water	tree	cave	fierce	timid
antelope	-1.00	-1.00	-1.0	-1.00	12.34	0.0	0.0	0.0	16.11	9.19	...	0.00	0.00	1.23	10.49	39.24
grizzly+bear	39.25	1.39	0.0	74.14	3.75	0.0	0.0	0.0	1.25	0.00	...	7.64	9.79	53.14	61.80	12.50
killer+whale	83.40	64.79	0.0	0.00	1.25	0.0	0.0	0.0	68.49	32.69	...	79.49	0.00	0.00	38.27	9.77
beaver	19.38	0.00	0.0	87.81	7.50	0.0	0.0	0.0	0.00	7.50	...	65.62	0.00	0.00	3.75	31.88
dalmatian	69.58	73.33	0.0	6.39	0.00	0.0	0.0	0.0	37.08	100.00	...	1.25	6.25	0.00	9.38	31.67

5 rows × 85 columns

Dimension : 50 x 85

Other datasets



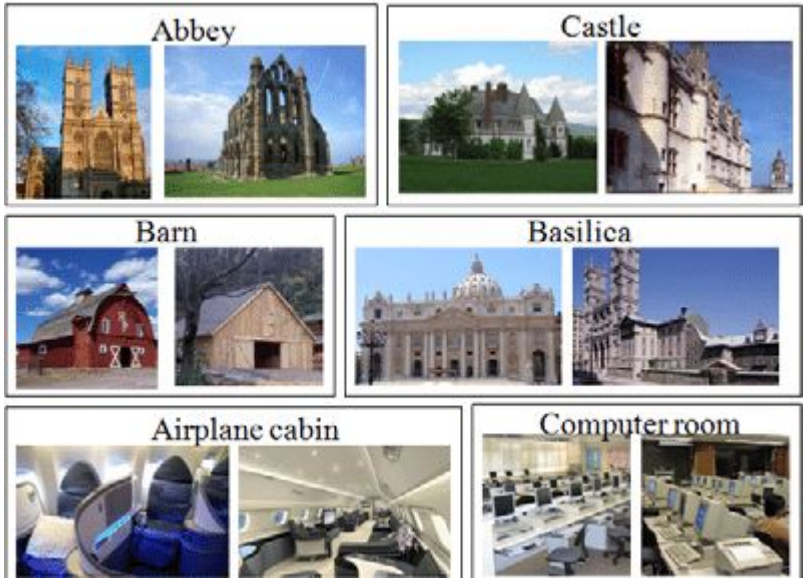
CUB-200

Classes : 200

Image count: 11,788

Attributes : 312

Detail : Fine-grained



SUN-2012

Classes : 717

Image count: 14,340

Attributes : 102

Detail : Fine-grained

Learning settings

The three settings.....

During model training, if information about testing instances is involved, it is called a transductive model.

ZSL settings:

1. Class-inductive Instance-inductive (CIII): Only D^{tr} and T^{s} used for learning
2. Class-transductive Instance-inductive (CTII): Only D^{tr} , T^{u} and T^{s} used for learning
3. Class-transductive Instance-transductive (CTIT): D^{tr} , X^{te}, T^{u} and T^{s} used for learning

CIII setting: more generalized performance but severe problems due to domain shift.

CTII setting: less domain shift, but can't generalize well to unseen classes at times.

CTIT setting: least domain shift, and least generalization ability.

Learning methods

Existing ZSL methods can be divided into:

Classifier-based methods

Focus : How to directly learn a classifier for the unseen class

Approaches:

1. Correspondence methods
2. Relationship methods
3. Combination methods

Instance-based methods

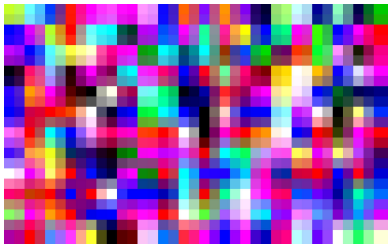
Focus : How to obtain labeled instances of unseen class and use them for classifier learning

Approaches:

1. Projection methods
2. Instance-borrowing methods
3. Synthesizing methods

Classifier-based method : An example

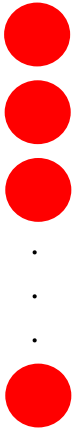
Sandipan Sarma

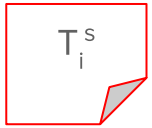


D^{tr}

$Loss_1$

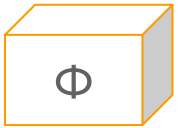
Learn N_s binary one-vs-rest classifiers f^s_i (seen classes)



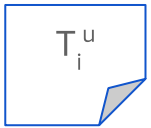


$Loss_2$

Learns mapping from T to w



TRAINING



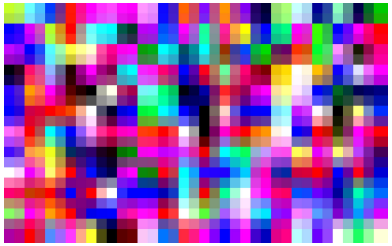
Construct N_u binary one-vs-rest classifiers f^u_i (unseen classes)

TESTING

Illustration of Correspondence method in CIII setting

Instance-based method : An example

Sandipan Sarma



D^{tr}

T_i^s

$Loss_1$

Learn mapping from image embeddings to projection space P

Θ

$Loss_2$

Learn mapping from class embeddings to projection space P

ξ

Illustration of Projection method in CIII setting

TRAINING

X^{te}

T_i^u

Θ

ξ

Project to P space

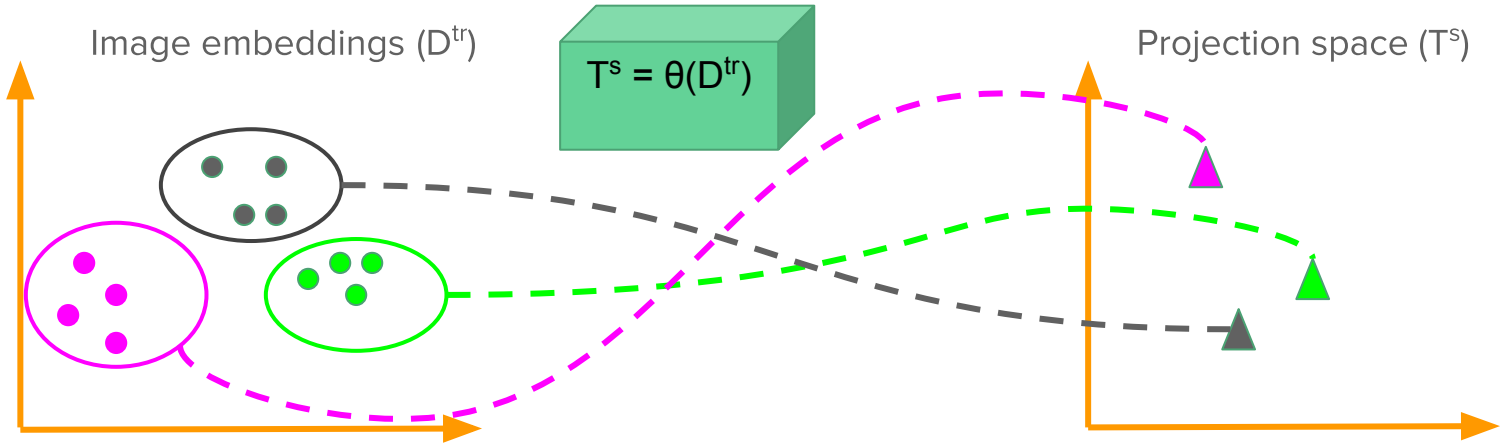
Project to P space

Do nearest neighbor classification in this P space

Y^{te}

TESTING

Projection method: Training (assuming P = T)



Penguin

Cat

Dog

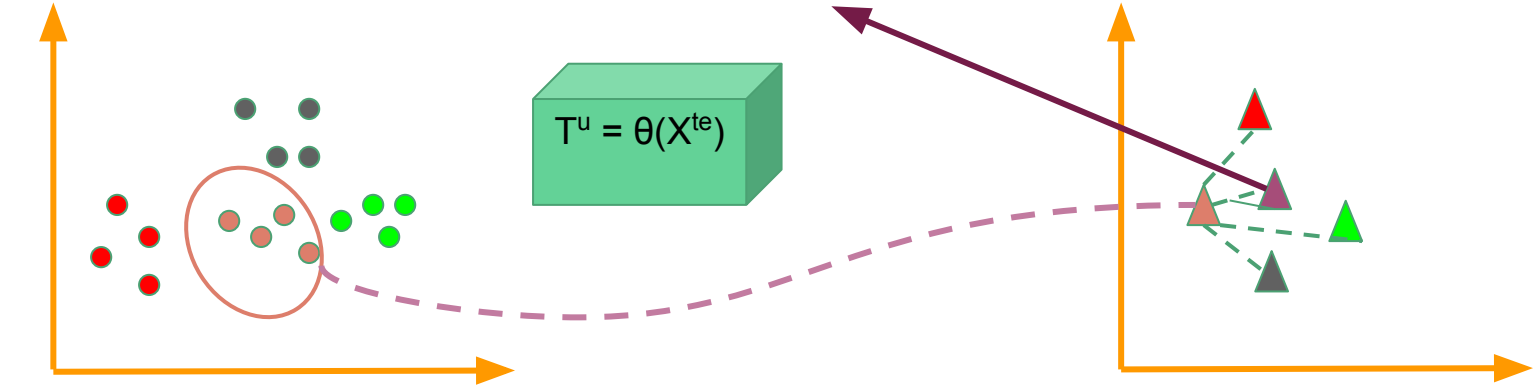
C^S_i :

Projection method: Testing (assuming P = T)

Image embeddings (X^{te})

Class embeddings (T^u)

Assigned Class = Rabbit



Bear

Frog

Gull

Rabbit

c^u_i :

Some special data settings

Generalized Zero-Shot Learning (GZSL)

- ➔ ZSL assumption : Training and testing sets are disjoint - **impractical** !
- ➔ Generalized ZSL assumption: Test set contains examples from both seen classes and unseen classes
- ➔ Problems under this setting are more challenging, as there might be a bias towards the prediction of unseen class examples as seen class examples, since the training was done on seen classes
- ➔ The performance of existing ZSL models under this setting is still not very high - **room for improvement !**

Multi-label Zero -Shot Learning

- ZSL (usual assumption) : An image belongs to only one particular class
- But there are some problems where each instance may belong to multiple classes. Example:

Labels : Sand, Beach, Mountain, Sky

- Existing methods either treat each class label individually, or exploit multi-label correlations

Evaluating Zero-Shot Models

Standard criteria given by Xian et al. (2017)

- ➔ Single label image classification accuracy to be measured with Top-1 accuracy, i.e. **the prediction is accurate when the predicted class is the correct one**
- ➔ Find the recognition accuracy for each class separately and then average it over all classes. This encourages high performance on both sparsely and densely populated classes

Per-class Top-1 accuracy for ZSL:

$$acc_{\mathcal{Y}} = \frac{1}{\|\mathcal{Y}\|} \sum_{c=1}^{\|\mathcal{Y}\|} \frac{\#\text{ correct in c}}{\#\text{ in c}}$$

to insure that all classes will weigh the same

Harmonic Mean for GZSL:

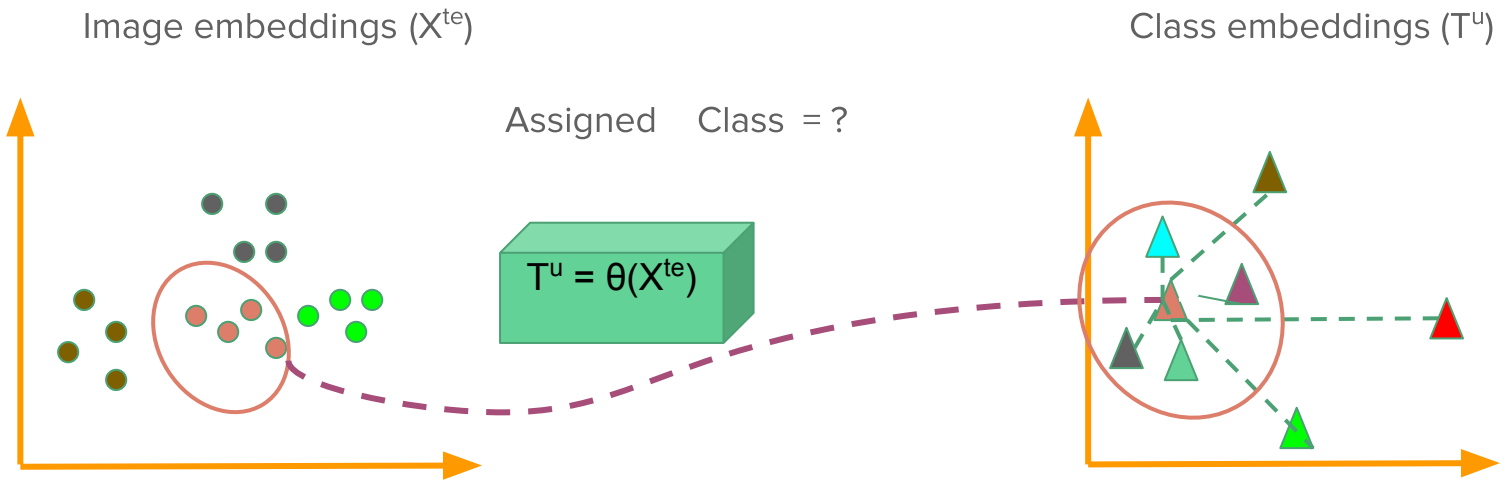
$$H = \frac{2 * acc_{\mathcal{Y}^{tr}} * acc_{\mathcal{Y}^{ts}}}{acc_{\mathcal{Y}^{tr}} + acc_{\mathcal{Y}^{ts}}}$$

to insure that seen and unseen class accuracy will weigh the same

Refer : Xian, Y., Lampert, C. H., Schiele, B., & Akata, Z. (2018). Zero-shot learning—A comprehensive evaluation of the good, the bad and the ugly. IEEE transactions on pattern analysis and machine intelligence, 41(9), 2251-2265.

Challenges in Zero-Shot Learning

Hubness problem



c^u_i :

Domain shift problem

- ➔ A **domain shift / distributional shift** is a change in the data distribution between an algorithm's training dataset, and a dataset it encounters when deployed.
- ➔ Since distribution of training and testing data is different, model performance for testing is generally lesser than training.
- ➔ **This phenomenon is more severe in ZSL since seen and unseen classes are disjoint.**

Example for domain shift. The bags from domain 1 do not have any background while those from domain 2 have complex background. The classifier trained on one domain will suffer from performance degradation when it is tested on the other domain

Bias problem

- ➔ While ZSL training, <image, label> access is limited to seen classes only.
- ➔ Model becomes **biased** towards **predicting seen classes as correct classes** more frequently at test time.
- ➔ Bias problem hurts more in case of generalized ZSL:
  - ◆ Test set has images from both seen and unseen classes
  - ◆ So, a biased model categorizes an unseen class example as a seen class example
  - ◆ Result - accuracy dips drastically !

Cross-domain knowledge transfer problem

- ➔ Prediction power of a ZSL model will depend upon the ability it gained to bridge the gap between seen and unseen classes. This gap is bridged by the semantic space.
- ➔ So, knowledge transfer from visual-to-semantic domain and vice-versa is crucial.
- ➔ Both these domains are very different in nature - one very high-dimensional and another low-dimensional.
- ➔ These mappings should have enough discriminative power. Example* :

Grizzly bear (seen class example)

Polar bear (unseen class example)

During testing, model should understand bear-like properties along with attribute “white” - otherwise may label it as grizzly bear !

* Refer :https://www.learnopencv.com/zero-shot-learning-an-introduction/

Application areas

Zero-shot Localization by Free Text

- Semantic attributes
  - "hat", "white", ...
- Spatial attributes too
  - "right", "on top of", "below", ...
- Global context

[1] Natural Language Object Retrieval, Hu et al., CVPR 2016

Slide credit : Thomas Mensink, Efstratios Gavves, Zeynep Akata, Cees Snoek - University of Amsterdam, in ZERO-SHOT LEARNING FOR COMPUTER VISION: Half-day CVPR 2017 Tutorial - 26th July 2017 (afternoon)

Zero-shot localization in videos, aka *Tracking by Natural Language [1]*

- Define the target not as a bounding box but as a language description?

[1] Tracking by Natural Language Specification, Li et al., CVPR 2017

Slide credit : Thomas Mensink, Efstratios Gavves, Zeynep Akata, Cees Snoek - University of Amsterdam, in ZERO-SHOT LEARNING FOR COMPUTER VISION: Half-day CVPR 2017 Tutorial - 26th July 2017 (afternoon)

Retrieving images from Wikipedia text

Around 850, out of obscurity rose Vijayalaya, made use of an opportunity arising out of a conflict between Pandyas and Pallavas, captured Thanjavur and eventually established the imperial line of the medieval Cholas. Vijayalaya revived the Chola dynasty and his son Aditya I helped establish their independence. He invaded Pallava kingdom in 903 and killed the Pallava king Aparajita in battle, ending the Pallava reign. K.A.N. Sastri, "A History of South India" p 159 The Chola kingdom under Parantaka I expanded to cover the entire Pandya country. However towards the end of his reign he suffered several reverses by the Rashtrakutas who had extended their territories well into the Chola kingdom...

Top 5 Retrieved Images

Slide credit : Thomas Mensink, Efstratios Gavves, Zeynep Akata, Cees Snoek - University of Amsterdam, in ZERO-SHOT LEARNING FOR COMPUTER VISION: Half-day CVPR 2017 Tutorial - 26th July 2017 (afternoon)

Retrieving book excerpts from movies

Slide credit : [Thomas Mensink, Efstratios Gavves, Zeynep Akata, Cees Snoek - University of Amsterdam, in ZERO-SHOT LEARNING FOR COMPUTER VISION: Half-day CVPR 2017 Tutorial - 26th July 2017 (afternoon)]

Retrieving video events from descriptions

Definition: An individual (or more) succeeds in reaching a pre-determined destination before all other individuals, without vehicle assistance or assistance of a horse or other animal. Racing generally involves accomplishing a task in less time than other competitors. The only type of racing considered relevant for the purposes of this event is the type where the task is traveling to a destination, completed by a person(s) without assistance of a vehicle or animal. Different types of races involve different types of human ...

Event Name: Winning a race without a vehicle

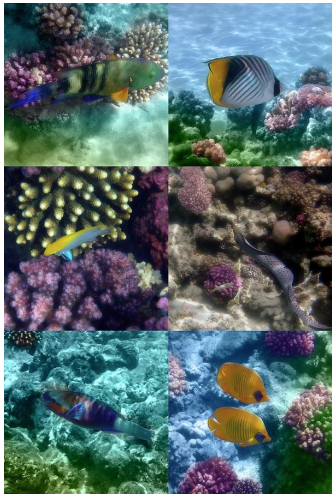
Slide credit : Thomas Mensink, Efstratios Gavves, Zeynep Akata, Cees Snoek - University of Amsterdam, in ZERO-SHOT LEARNING FOR COMPUTER VISION: Half-day CVPR 2017 Tutorial - 26th July 2017 (afternoon)

Robotic vision : Terrain exploration

- ➔ When operating in open world, the exploration terrain might contain new classes of interest that were not available during training. A robot therefore needs the capability to **extend its knowledge and efficiently learn new classes without forgetting the previously learned representations.**
- ➔ This class-incremental learning would preferably be **data-efficient by using zero-shot,** one-shot, or few-shot learning techniques. Semi-supervised approaches that can leverage unlabeled data are of particular interest

An underwater autonomous vehicle (UAV)

Terrain it explores



Questions ?