Zero-Shot Learning

for Vision and Multimedia

Thomas Mensink, Efstratios Gavves, Cees Snoek

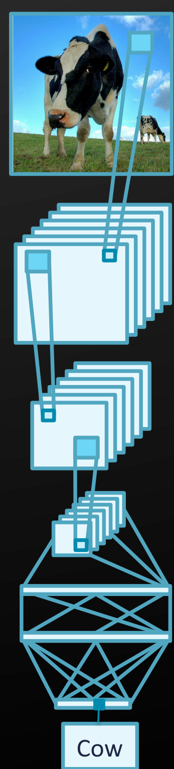
University of Amsterdam

Many-shot learning

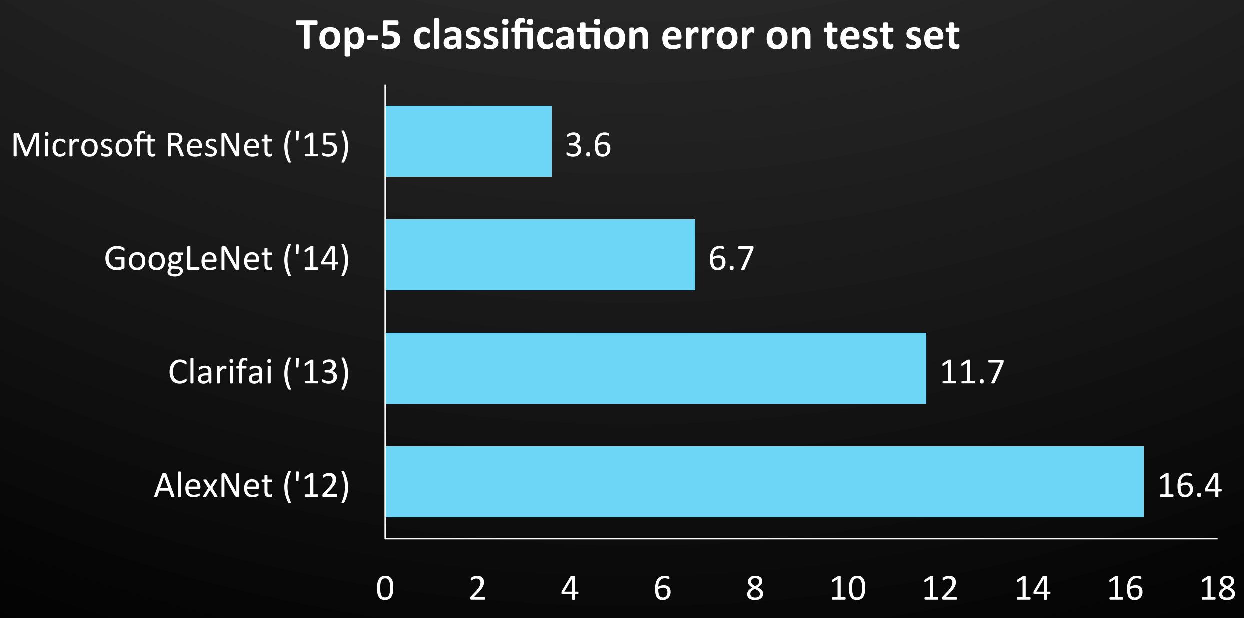
+

Annotations

=

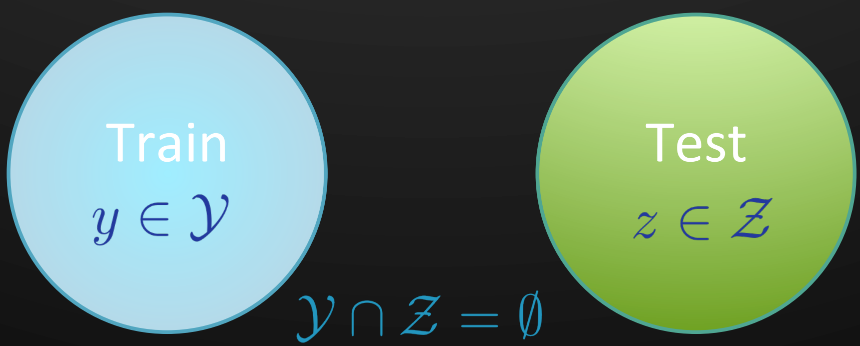


Most popular plot in computer vision



What is zero-shot learning?

Data: $x \in \mathcal{X}$



Objective: $f : \mathcal{X} \rightarrow \mathcal{Z}$

Lampert et al., CVPR09/PAMI13

We have labeled data, why bother?

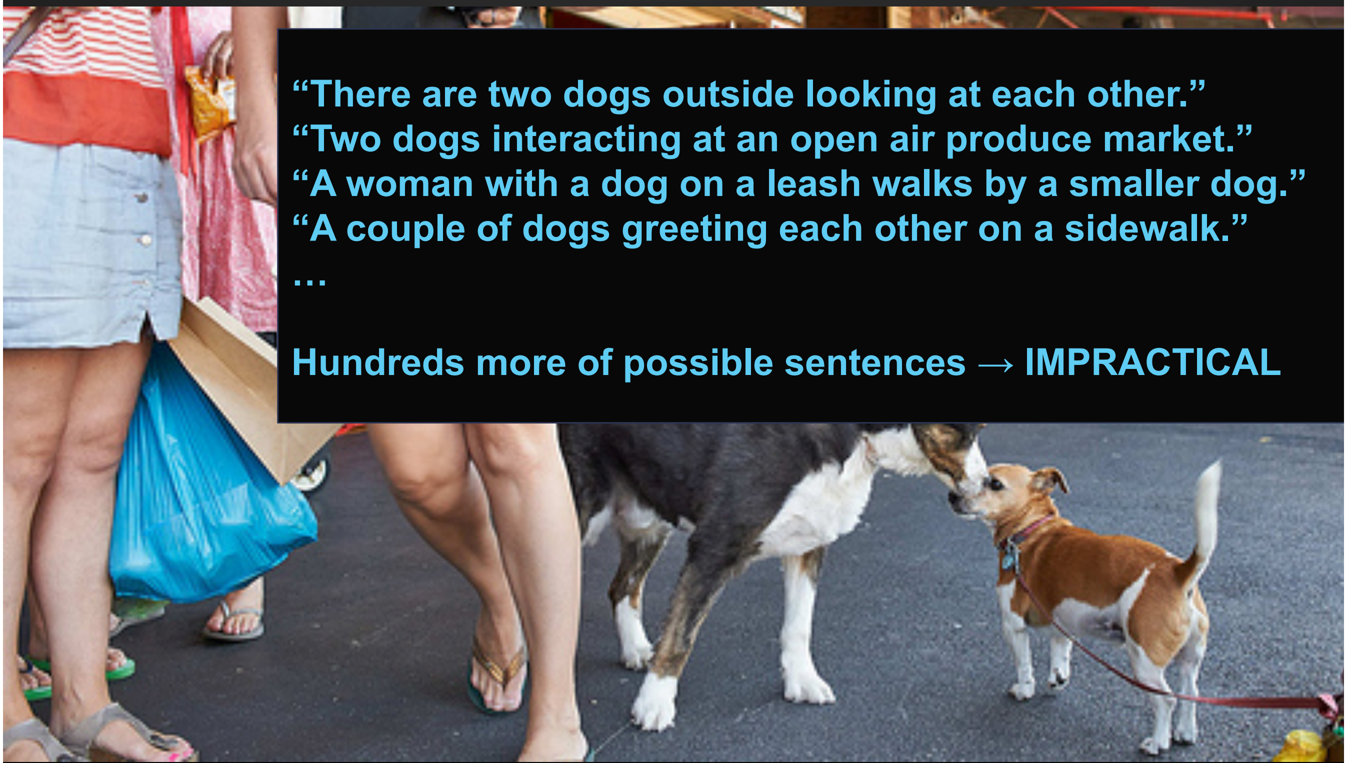
Imagenet: ~15,000,000 images
Open Images: ~9,000,000 images
Places: ~2,500,000 images

An image

Classification

Segmentation

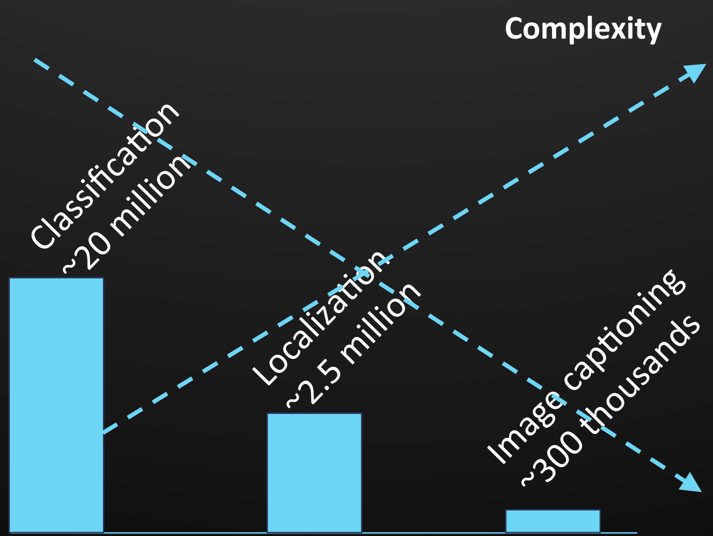
Captioning



"There are two dogs outside looking at each other."
"Two dogs interacting at an open air produce market."
"A woman with a dog on a leash walks by a smaller dog."
"A couple of dogs greeting each other on a sidewalk."
…

Hundreds more of possible sentences → IMPRACTICAL

Annotation vs complexity



Imagenet+Open Images+MS COCO

Why zero-shot learning?

The more complex tasks we target,
the fewer annotations we have,
the more relevant zero shot learning is.

“Man in blue jacket stealing sports bike with crowbar”

Why zero-shot learning?

Privacy-sensitive recognition problems

Why zero-shot learning?

When learning and inference need to be efficient





What is this tutorial about?

Data: $x \in \mathcal{X}$

Knowledge transfer

Train
$y \in \mathcal{Y}$

Test
$z \in \mathcal{Z}$

$\mathcal{Y} \cap \mathcal{Z} = \emptyset$

$\boxed{\text{Objective: } f : \mathcal{X} \to \mathcal{Z}}$

Lampert et al., CVPR09/PAMI13

Today's outline

1. Knowledge transfer
2. Classification
3. Localization

   Break

4. Retrieval
5. Interaction
6. Conclusion and Discussion

Knowledge Transfer

Zero-Shot Learning

for Vision and Multimedia

What is this tutorial about?

Data: $x \in \mathcal{X}$

Knowledge transfer

Train

$y \in \mathcal{Y}$

Test

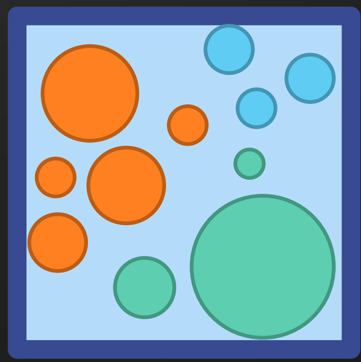
$z \in \mathcal{Z}$

$\mathcal{Y} \cap \mathcal{Z} = \emptyset$

Objective: $f : \mathcal{X} \rightarrow \mathcal{Z}$

Lampert et al., CVPR09/PAMI13

Unsupervised learning



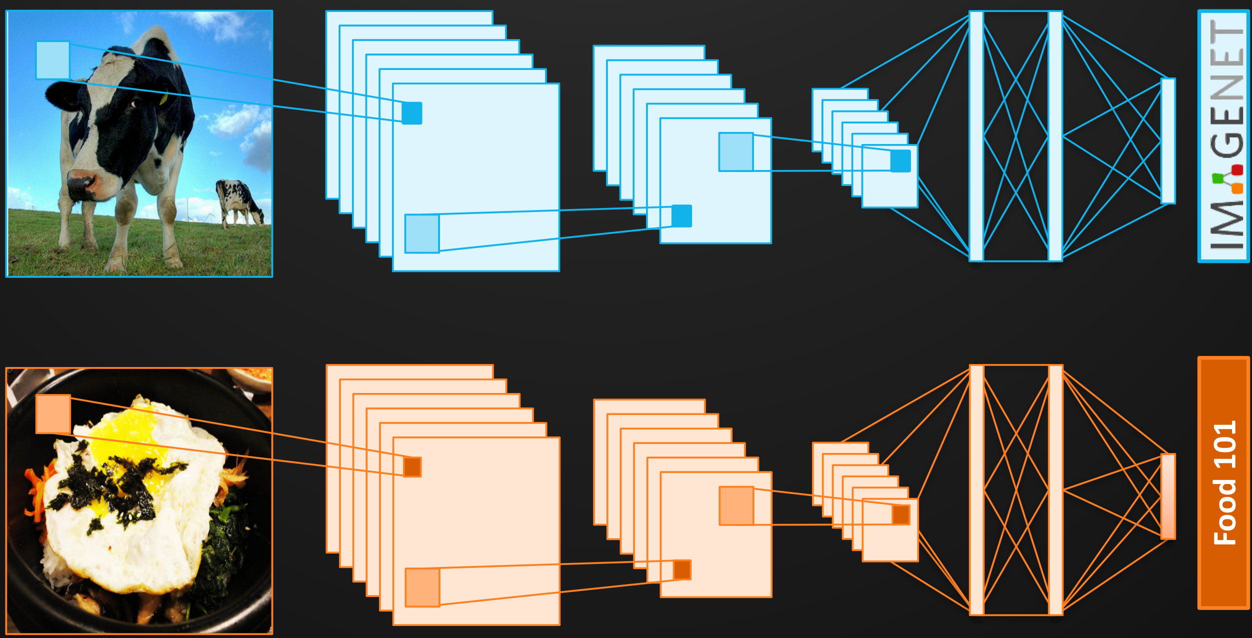
Transfer Learning

+

Pre-trained classifier

(on different dataset)

Transfer Learning: Fine-tuning

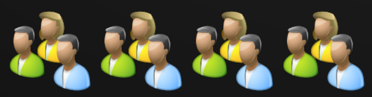


Search Engine Transfer

+

Active Learning

+



Zero-Shot Knowledge Transfer

+
Class Description
+
Background Knowledge

Background knowledge

1. Some visual knowledge
2. Mapping between class description and visual knowledge

Attribute Based Knowledge Transfer

Attributes

Class definitions using a small set of semantic attributes

Extension of standard multi-class annotation

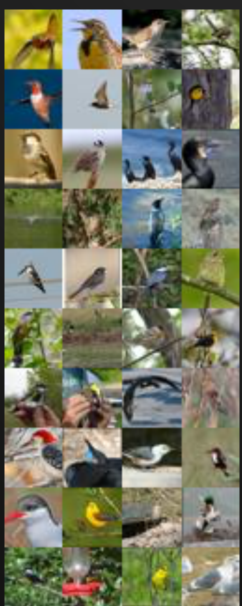
Example: Animals with Attributes

Otter	
black	yes
white	no
brown	yes
stripes	no
water	yes
eat fish	yes

Polar Bear	
black	no
white	yes
brown	no
stripes	no
water	yes
eat fish	yes

Zebra	
black	yes
white	yes
brown	no
stripes	yes
water	no
eat fish	no

Example: CUB Birds



Attributes

Class definitions using a small set of semantic attributes

Attributes

- No formal definition
- Property of object
- Nameable (e.g., color, body part, habitat of animal)
- Not necessarily direct visual meaning (like habitat)
- Semantic (i.e., humans could assign meaning)
- Class discriminative, but not class specific
- Automatically visually detectable

Quiz: What are good attributes?

1. is grey?
2. is made of atoms?
3. lives in Amsterdam?
4. is sunny?
5. eat fish?
6. has a SIFT descriptor with empty bin 3?
7. has 4 wheels?
8. is the only animal with yyy

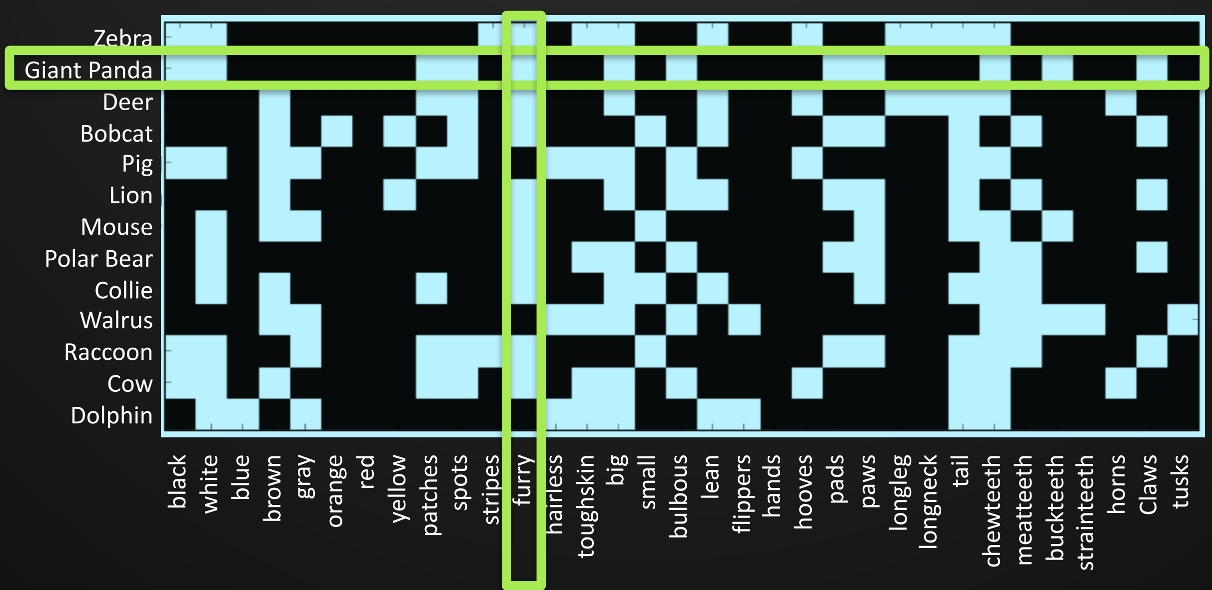
Attribute based transfer

Class definitions using a small set of semantic attributes

Disjoint train and test set, but common set of attributes

16

Class2Attributes mapping



17

Class2Attributes: How to obtain

Manually defined, by

- Experts
- Laymen

Obtained from knowledge sources

- Wikipedia
- Specific websites (eg birdbook)

Obtained from general sources

- Google search
- Flickr tags

Limitations of attributes

1. How to define the attributes of a chair?

2. Unnatural distinction

   classes of interest

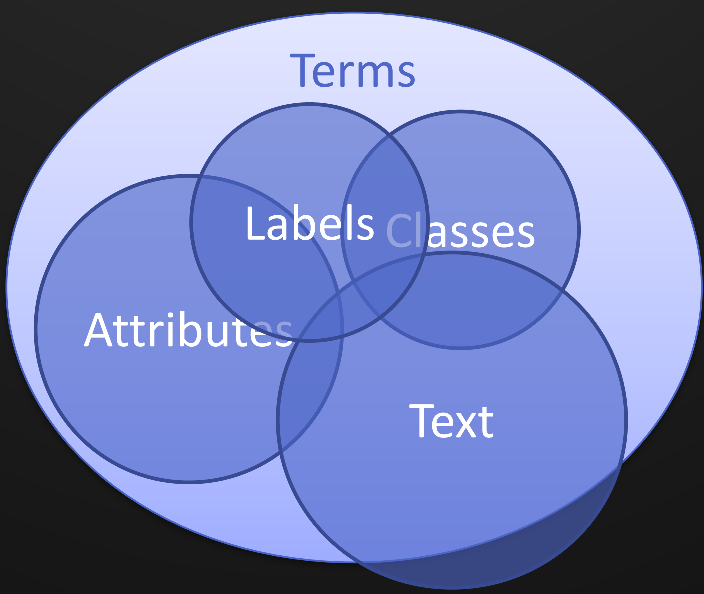
   attributes for recognition

3. Only multi-class

Term based Knowledge Transfer

Terms: What

Terms: any visual concept, label, attribute, or class.



Term based transfer

Represent image with set of visual classifiers scores

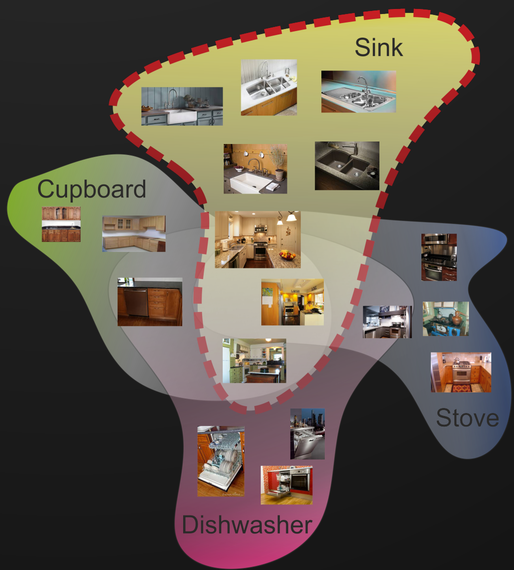
Re-use existing annotation efforts

Relate this set of terms to new concepts/classes

Relate terms using co-occurrences

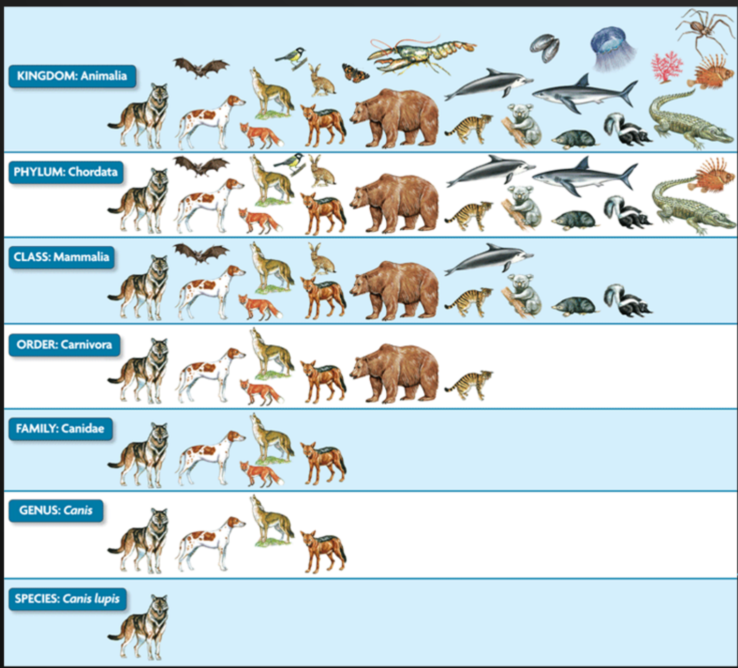
I'm looking for a concept, in a picture with terms:

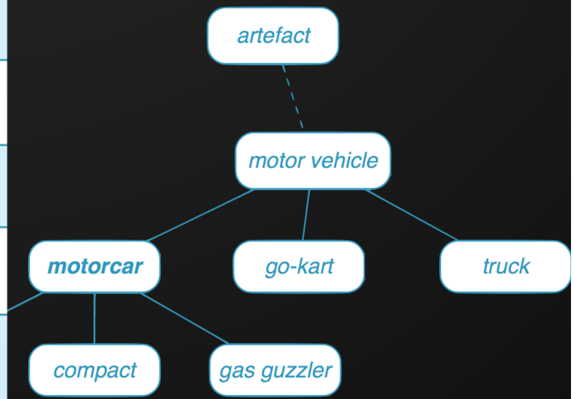
1. Indoor
2. Living room
3. Table
4. Chair
5. …



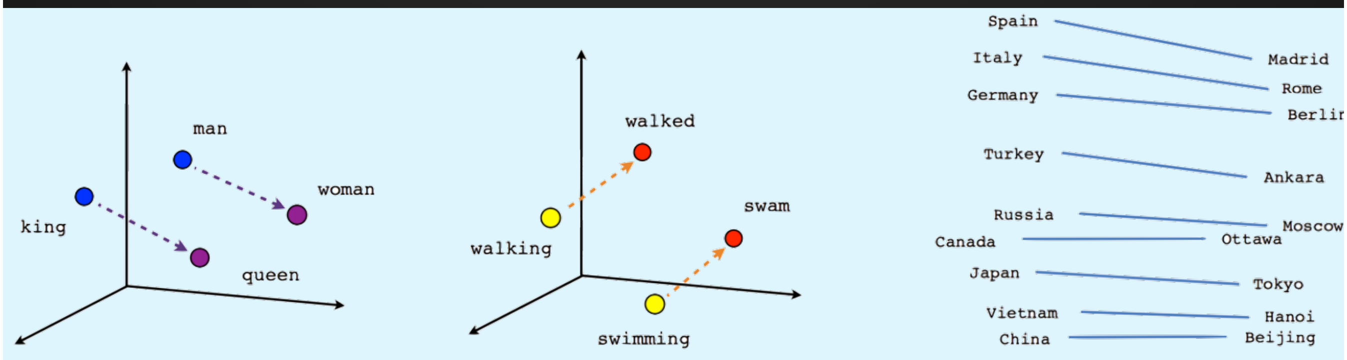
Mensink CVPR'14

Hierarchical Relations





Word2Vec

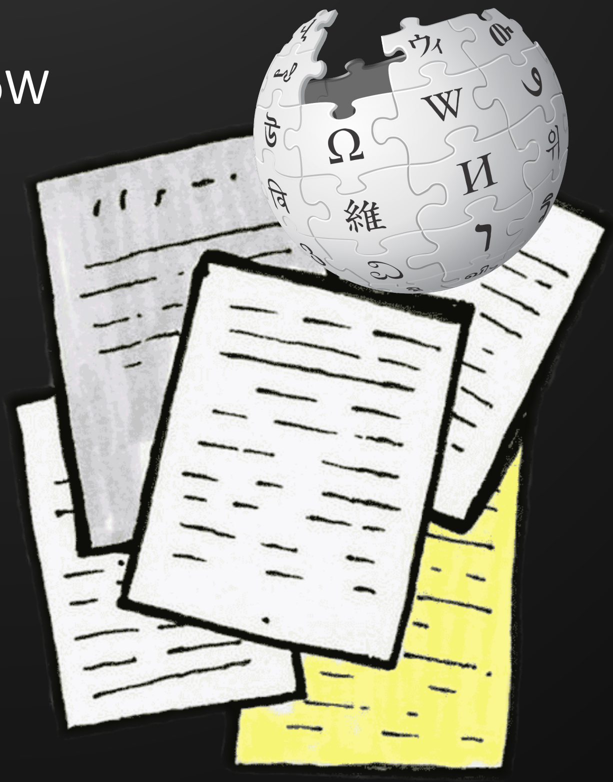


Mikolov ICLR'13

Article Based Knowledge Transfer

Use term scores as image-BoW

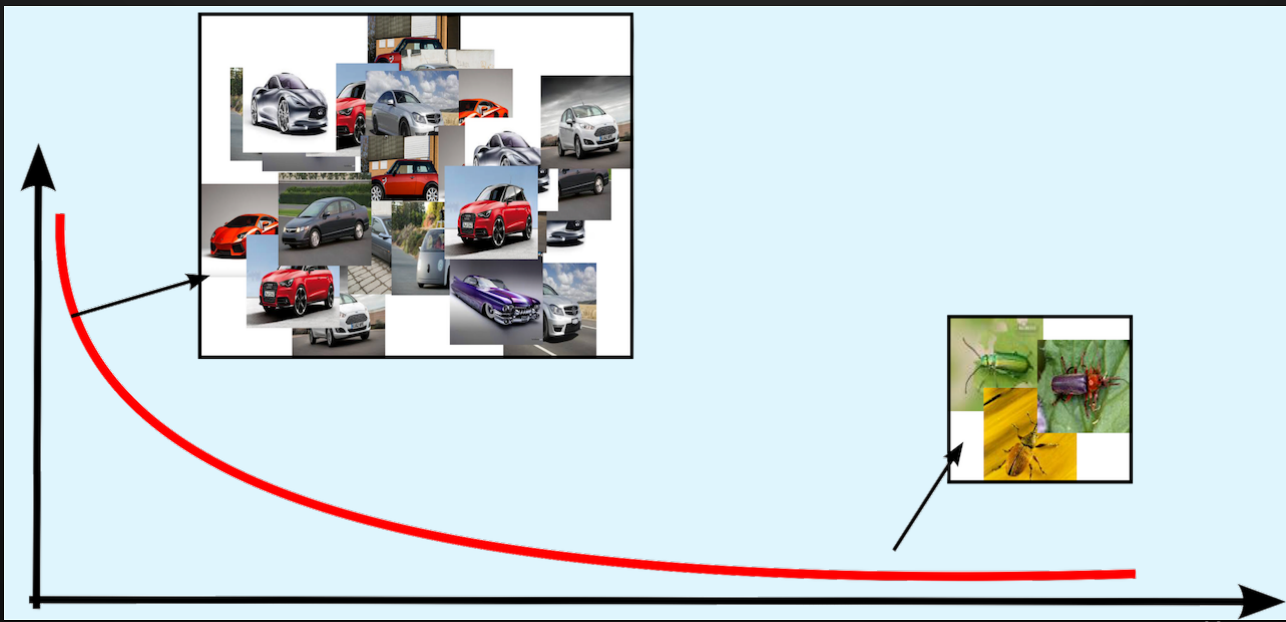
Compute distance between article-BoW and image-BoW



Increasing expressive power of terms

Which terms to use?

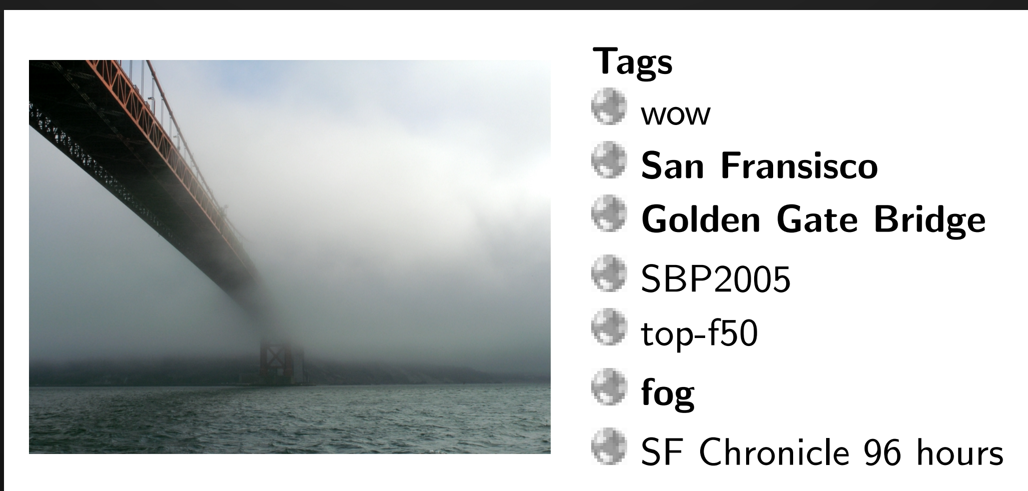
Long tail image distribution



Which terms to use?

Annotation mismatch

User annotates not for training computer vision



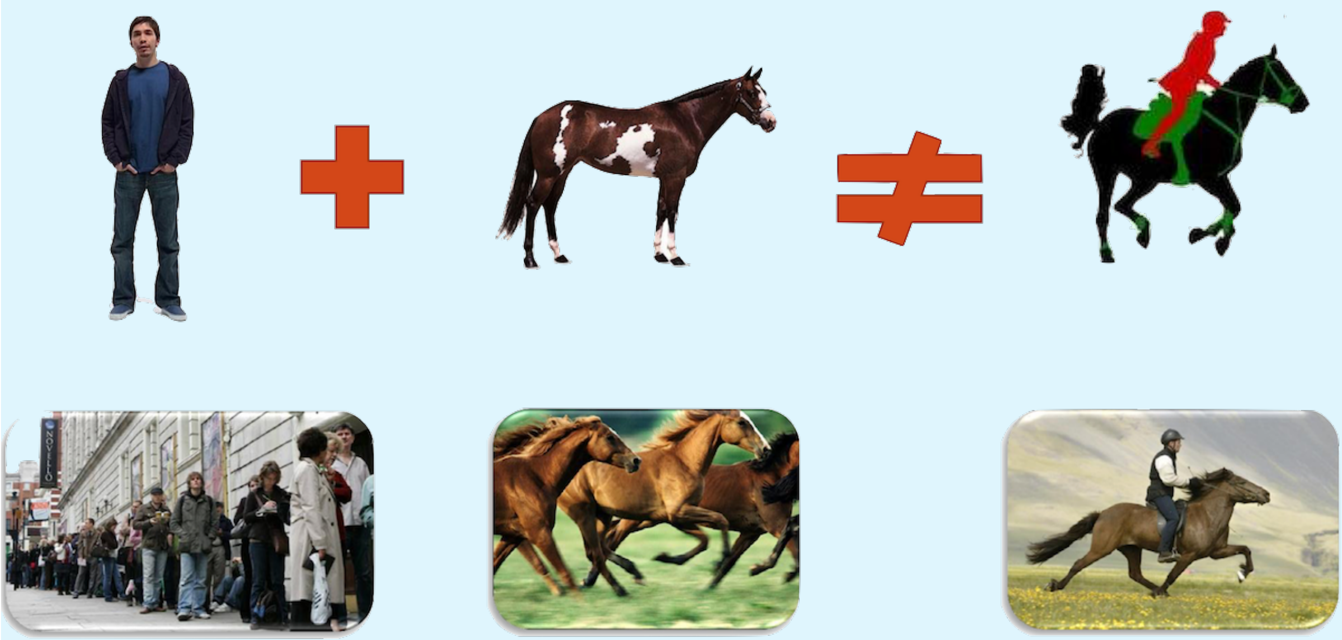
Which terms to use?

Combination semantics

30

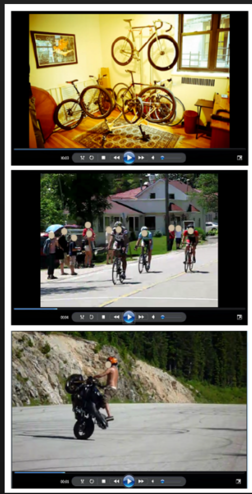
Which terms to use?

Visual coherence of concepts



Term composition trick

Expanding the terms by logical operations



<table>
<tr><th>Ride</th><th>Motor</th><th>Bike</th><th>Ride & Bike</th><th>Bike | Motor</th></tr>
<tr><td>0</td><td>0</td><td>1</td><td>0</td><td>1</td></tr>
<tr><td>1</td><td>0</td><td>1</td><td>1</td><td>0</td></tr>
<tr><td>1</td><td>1</td><td>0</td><td>0</td><td>1</td></tr>
</table>

Habibian et al. ICMR 2014

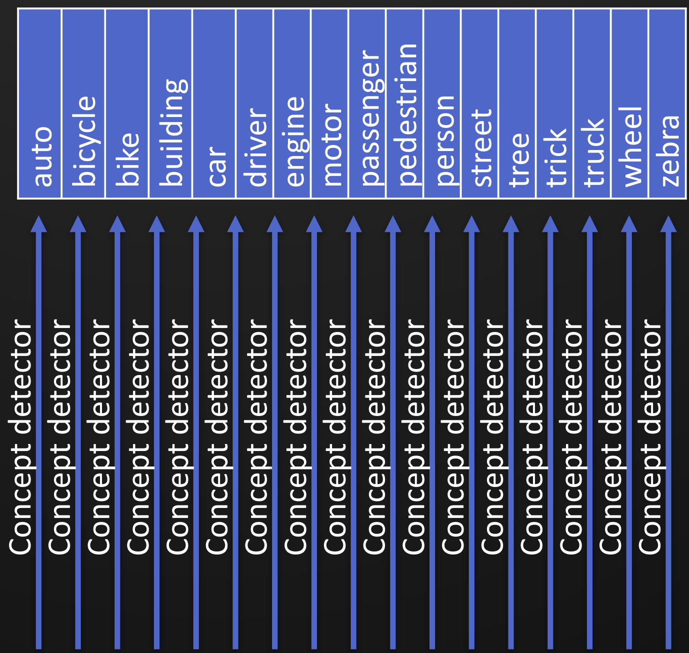
Term composition: motivation

Expanding the vocabulary for *free*

Composite terms can be easier to detect
- •boat-AND-sea
- •bear-AND-cage
- •man-OR-woman

Composite concepts can be more meaningful
- •bike-AND-ride for *attempting a bike trick*

Term Embedding



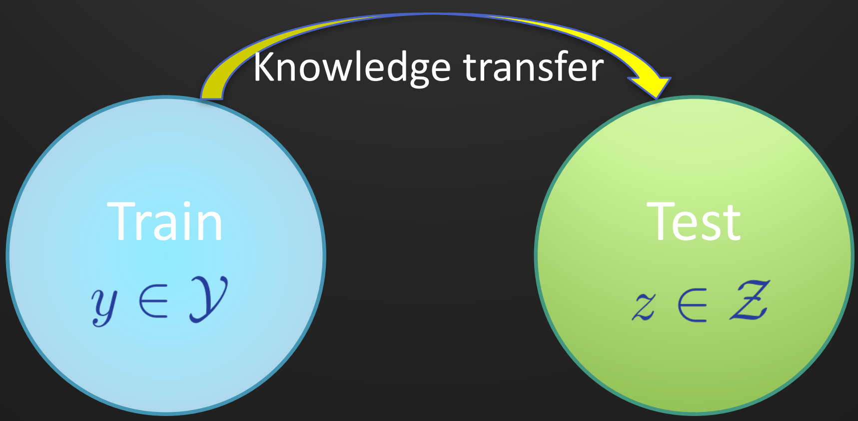
Term Embedding

	auto	bicycle	bike	building	car	driver	engine	motor	passenger	pedestrian	person	street	tree	trick	truck	wheel	zebra
Embedded detector	-0,5	-2,0	0,0	-2,3	0,0	0,0	0,0	-0,4	1,2	1,3	-1,9	-0,4	1,5	0,4	1,4	-0,3	1,3
Embedded detector	-1,7	-1,0	-1,7	-0,1	0,0	0,6	-0,8	0,0	-2,2	0,0	-0,3	0,0	0,0	0,0	0,5	0,0	-1,3
Embedded detector	-0,7	0,0	0,0	0,0	0,0	2,0	0,0	-0,6	-1,5	2,2	0,0	0,0	2,0	1,0	0,0	0,3	0,9
Embedded detector	1,9	-2,5	-1,9	-2,0	0,0	-0,2	0,0	-2,0	0,1	1,7	1,4	2,2	-1,7	2,4	-1,9	-1,9	-0,1
Embedded detector	0,0	0,0	-1,4	0,0	-1,5	0,6	1,2	-0,5	0,0	1,7	0,0	1,6	-0,8	-2,4	0,0	-0,5	2,0
Embedded detector	0,7	-0,6	-2,4	0,0	0,0	-1,5	0,0	0,0	-0,1	-2,1	0,0	2,1	-1,3	-0,2	0,0	-0,5	0,8
Embedded detector	-0,8	-0,4	0,0	0,0	0,0	0,0	1,4	-0,7	0,0	-2,3	-1,9	0,0	0,0	1,8	2,3	1,9	-1,4

Not necessary semantic meaning per detector

Still able to transfer visual meaning for zero-shot

Wrap up



Relations between visual concepts:

Attributes, hierarchical relation

Co-occurences, Word2Vec

Expressive power: combinations of terms and concepts

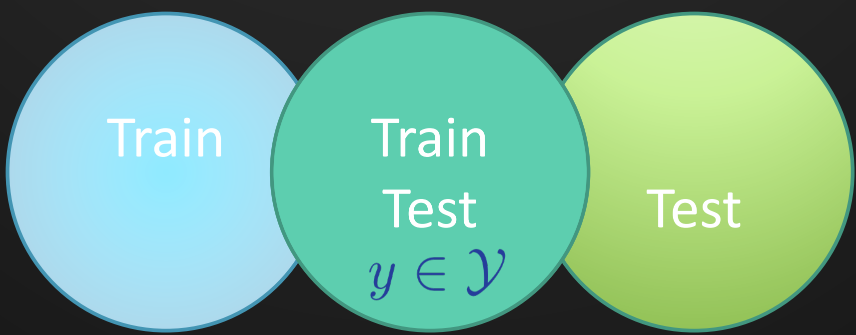
Zero-Shot Classification

Zero-Shot Learning

for Vision and Multimedia

Supervised Learning

Images: $x \in \mathcal{X}$



Classifier: $f : \mathcal{X} \rightarrow \mathcal{Y}$

Zero-shot Classification

Images: $x \in \mathcal{X}$

$$\mathcal{Y} \cap \mathcal{Z} = \emptyset$$

Classifier: $f : \mathcal{X} \rightarrow \mathcal{Z}$

Lampert et al., CVPR09/PAMI13

Attribute Based Classification

Attribute Based Classification

Images: $x \in \mathcal{X}$

Train $y \in \mathcal{Y}$

Test $z \in \mathcal{Z}$

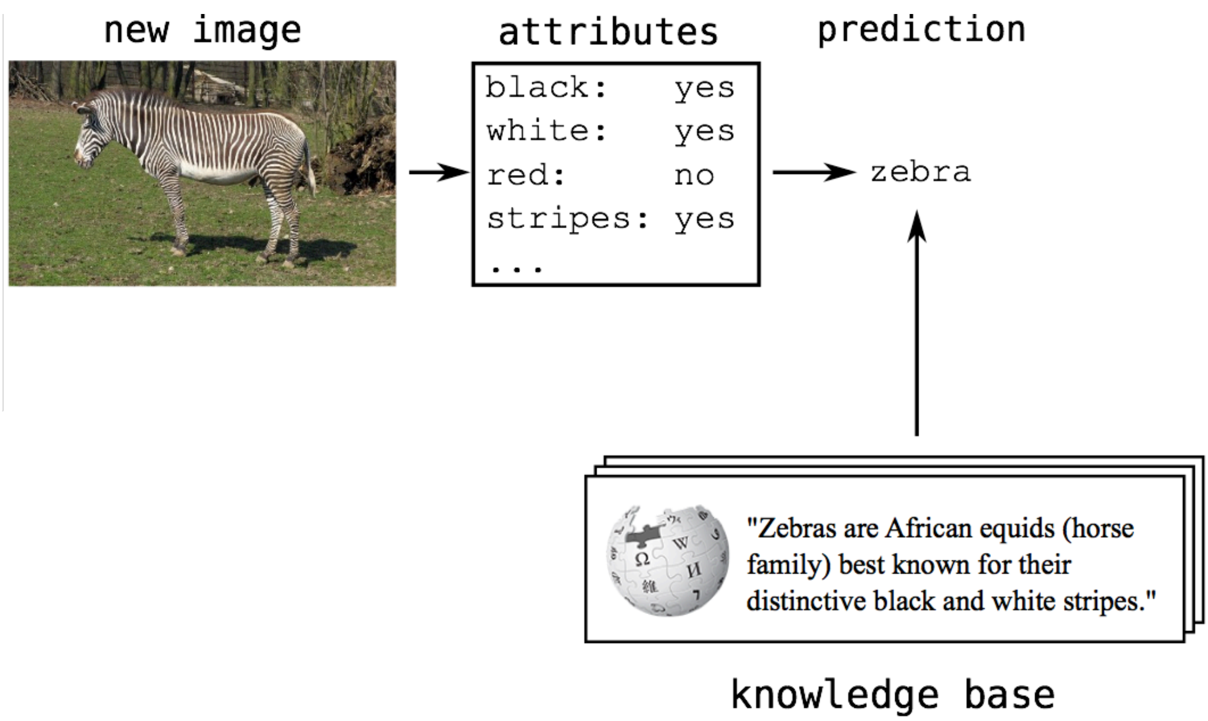
$$\mathcal{Y} \cap \mathcal{Z} = \emptyset$$

$$\forall y \in \mathcal{Y} : a^y \in \mathcal{A} \qquad \forall z \in \mathcal{Z} : a^z \in \mathcal{A}$$

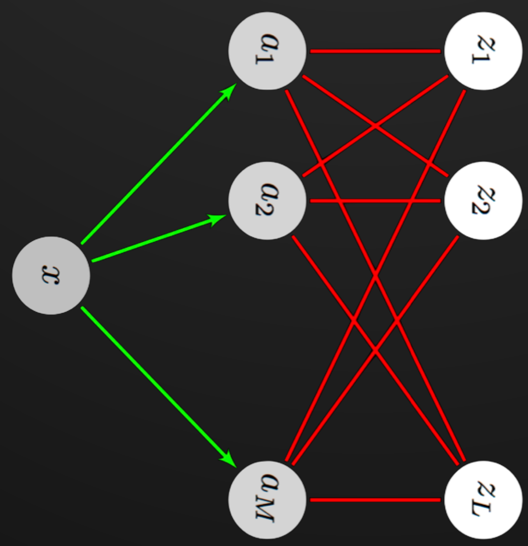
Classifier: $f : \mathcal{X} \rightarrow \mathcal{A} \rightarrow \mathcal{Z}$

Lampert et al., CVPR09/PAMI13

Attribute Based Classification: Example



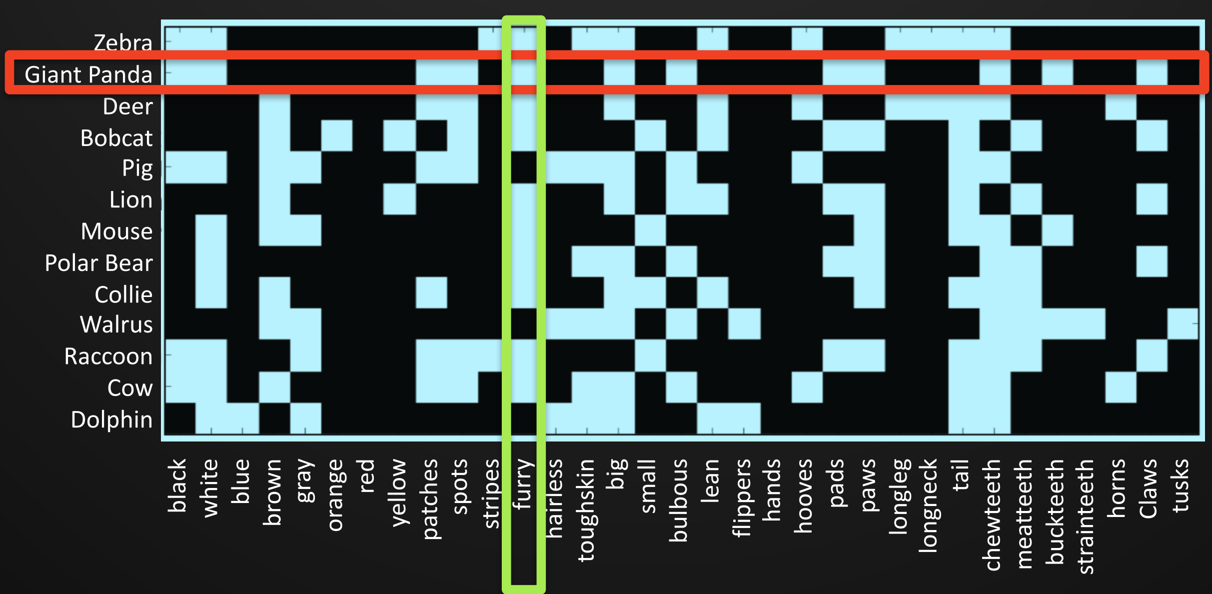
Attribute Based Classification: Graphical



Attribute Predictors
Trained on labeled data

Class Prediction
Based on prior knowledge

Class2Attributes mapping



Quiz: How many attributes?

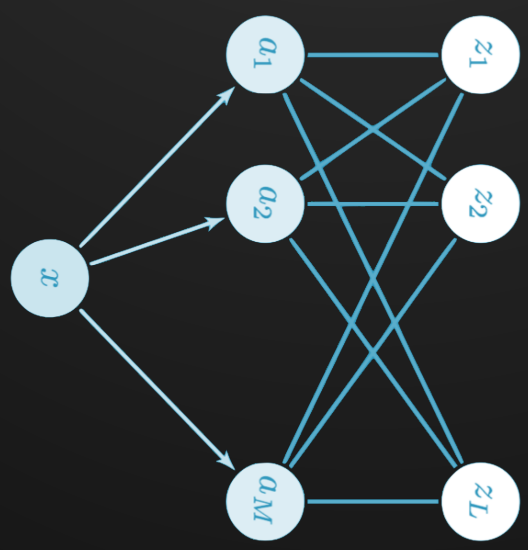
In theory k binary attributes can represent

- 2^k classes

In practice for c classes we need

- Many attributes

Direct Attribute Prediction

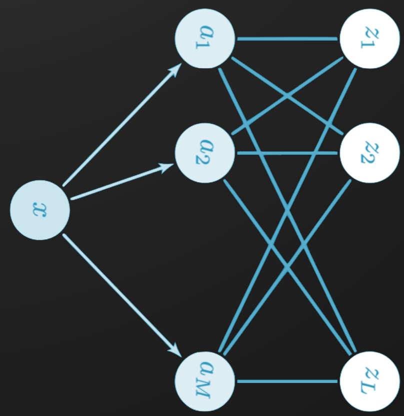


$$p(z|x) = \sum_{a \in \{0,1\}^M} p(z|a)p(a|x) = \frac{p(z)}{p(a^z)} \prod_{m=1}^{M} p(a_m^z|x)$$

Direct Attribute Prediction - Training

Goal:

Optimize attribute prediction

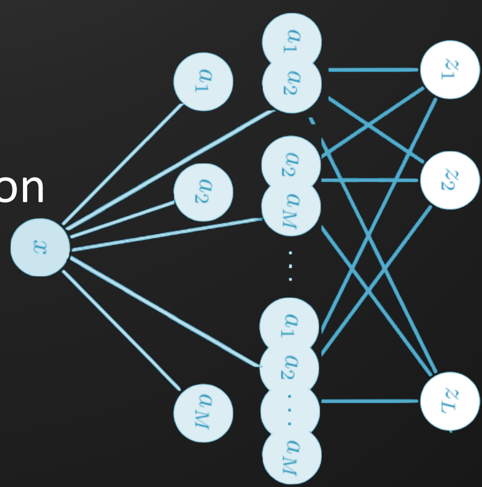


Per attribute, learn a single classifier to maximize p(a_m|**x**) for best AUC/mAP

Structured Attribute Prediction

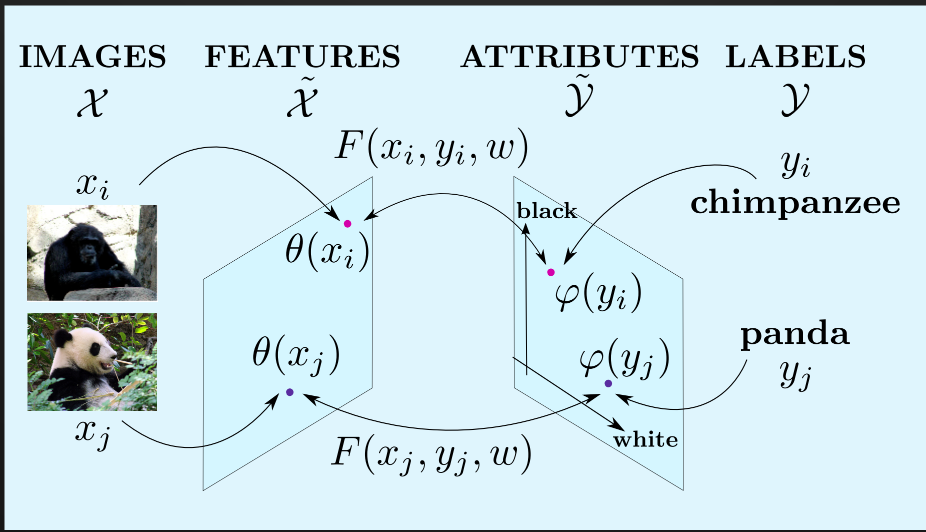
Goal:

Optimize joint attribute prediction



Learn a structured predictor, with links between attributes to predict p(**a**|**x**)

Attribute Label Embedding



$$F(x, y; W) = \theta(x)' W \varphi(y)$$

Akata, CVPR'13/PAMI'15

ALE Mathematics

Comparison DAP and ALE

$$f_{\mathrm{ALE}}(z, \boldsymbol{x}) = \varphi(z)^\top W \boldsymbol{x} = \boldsymbol{a}_z^\top W \boldsymbol{x}$$

$$p(z|\boldsymbol{x}) = \frac{p(z)}{p(\boldsymbol{a}_z)} \prod_m p(a_z^m|\boldsymbol{x}) \underset{\sim}{\propto} \prod_m \exp(a_z^m \boldsymbol{w}_m^\top \boldsymbol{x})$$

$$= \exp\left(\sum_m a_z^m \boldsymbol{w}_m^\top \boldsymbol{x}\right) = \exp(\boldsymbol{a}_z^\top W \boldsymbol{x})$$

Mathematically ALE and DAP are similar

ALE – Training

Objective:

$$L_{\mathrm{ALE}} = \frac{1}{N} \sum_i \max_{\tilde{z} \in \mathcal{Z}} \ell(\tilde{z}, z_i, \boldsymbol{x}_i)$$

ALE directly optimizes image classification

ALE – Generalization

- Non binary attributes

	AWA		CUB	
	$\varphi^{0,1}$	$\varphi^{\mathcal{A}}$	$\varphi^{0,1}$	$\varphi^{\mathcal{A}}$
FV (4K)	36.6	42.3	15.2	19.0
CNN (4K)	45.9	61.9	30.0	40.3
GOOG (1K)	52.0	**66.7**	37.8	**50.1**

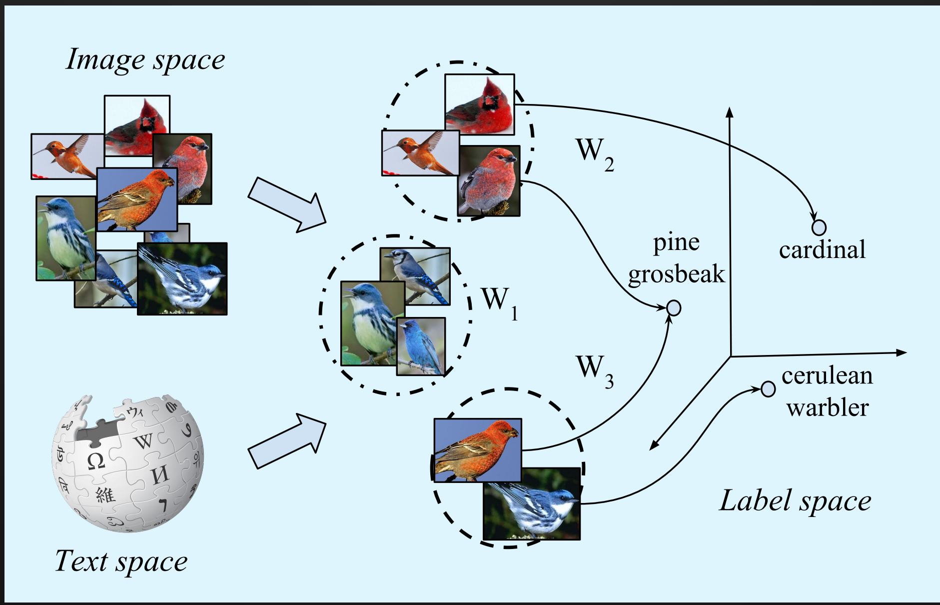
- Integrate other knowledge transfer $\varphi(z)$

  e.g., based on wordnet hierarchy, word2vec, wikipedia

- Few-shot learning: also learn embedding $\varphi(z)$

  With regularization term: $\frac{\mu}{2}||\Phi - \Phi^{\mathcal{A}}||^2$

Latent Attribute Embedding



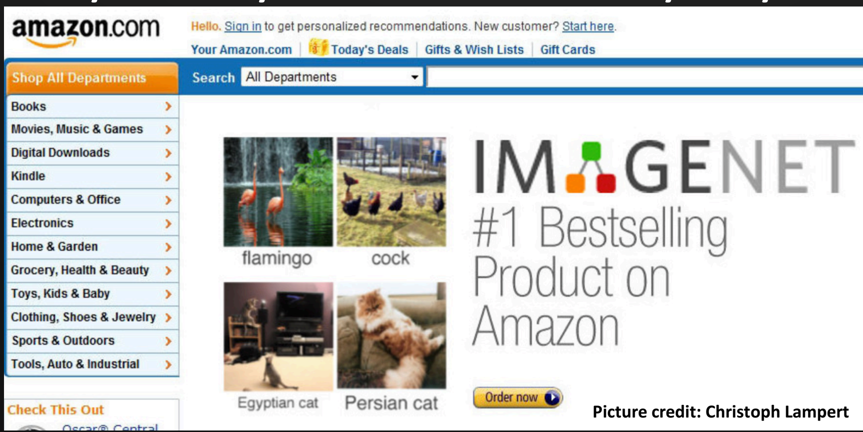
$$f_{\mathrm{ALE}}(z, \boldsymbol{x}) = \boldsymbol{a}_z^\top W \boldsymbol{x}$$

$$f_{\mathrm{LatEm}}(z, \boldsymbol{x}) = \max_k \boldsymbol{a}_z^\top W_k \, \boldsymbol{x}$$

ImageNet

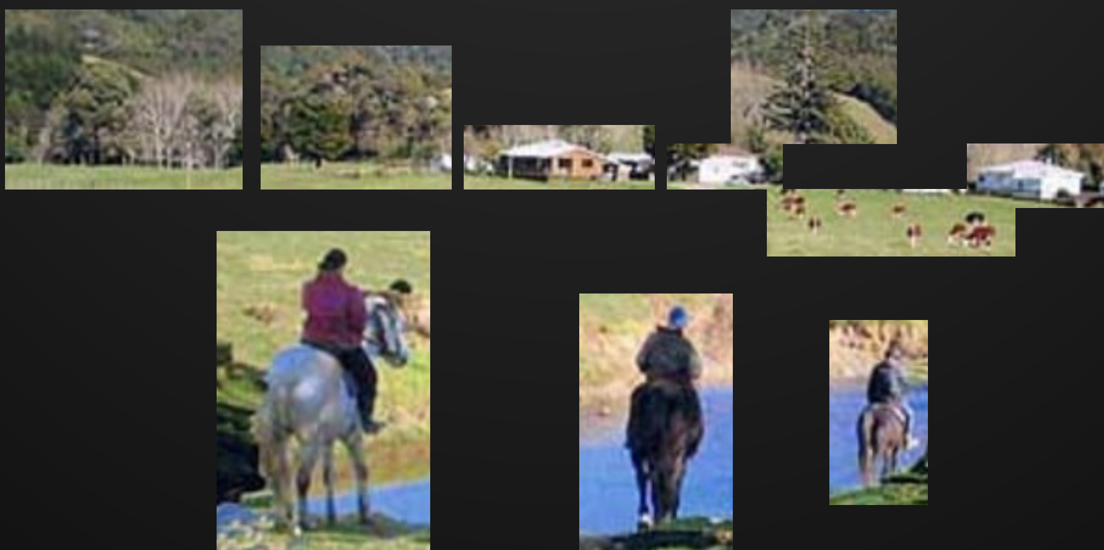
Based Zero-Shot Classification

- 14M images - 22K categories
- Why train your classifier anyway?



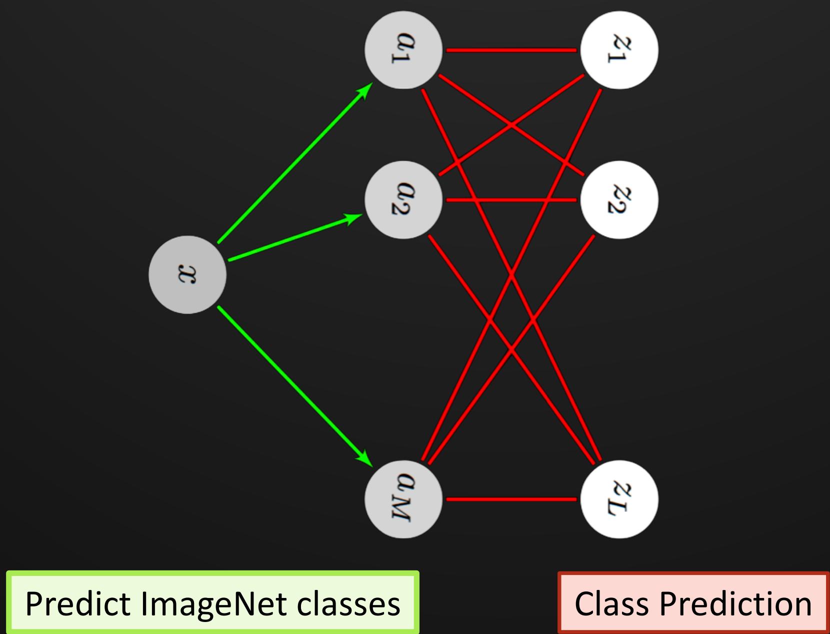
ImageNet limitation: only object classes

What objects tell about…



What objects tell about...

Class Based Classification



Weighted Convex Classifier

Goal: Estimate classifier $\hat{w}_l$ for unseen class

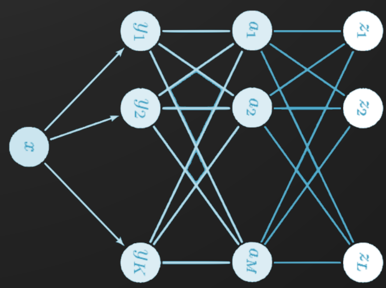
Zero-shot classifier:

$$\hat{w}_l = \sum_k a_k \, w_k \, s_{lk}$$

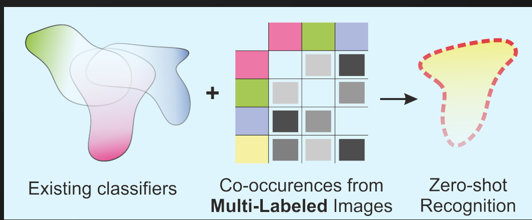
where s_{lk} is similarity between classes; and
where a_k is a weighing term for each known class

Three similarties

1. Indirect Attribute Prediction



2. Using co-occurrence statistics

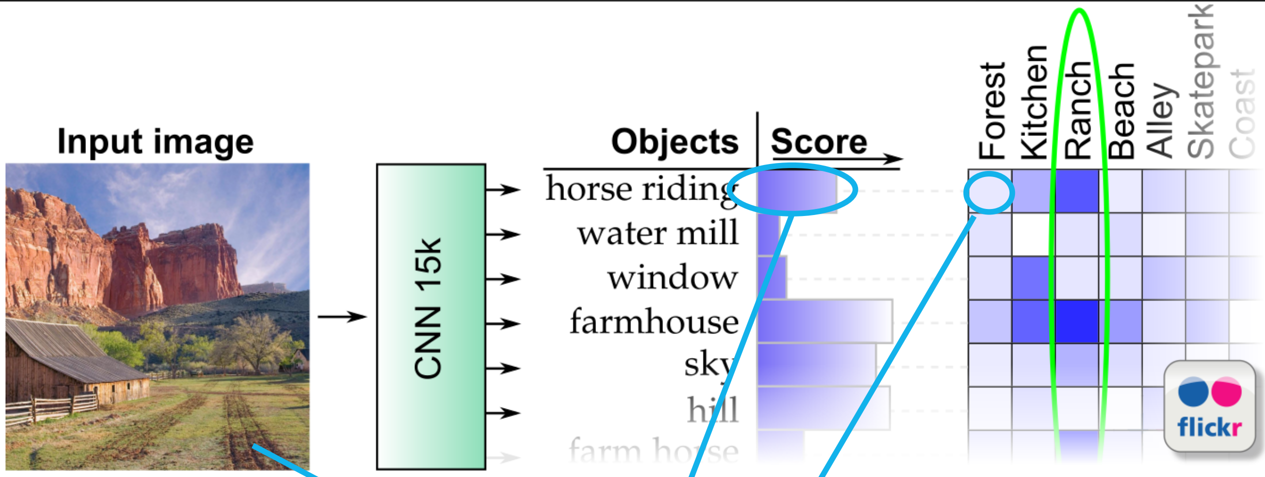


3. Using Word2Vec embeddings

Zero-Shot Learning by Convex Combination of Semantic Embeddings

Mohammad Norouzi*, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg S. Corrado, Jeffrey Dean

Word2Vec: from objects to scenes

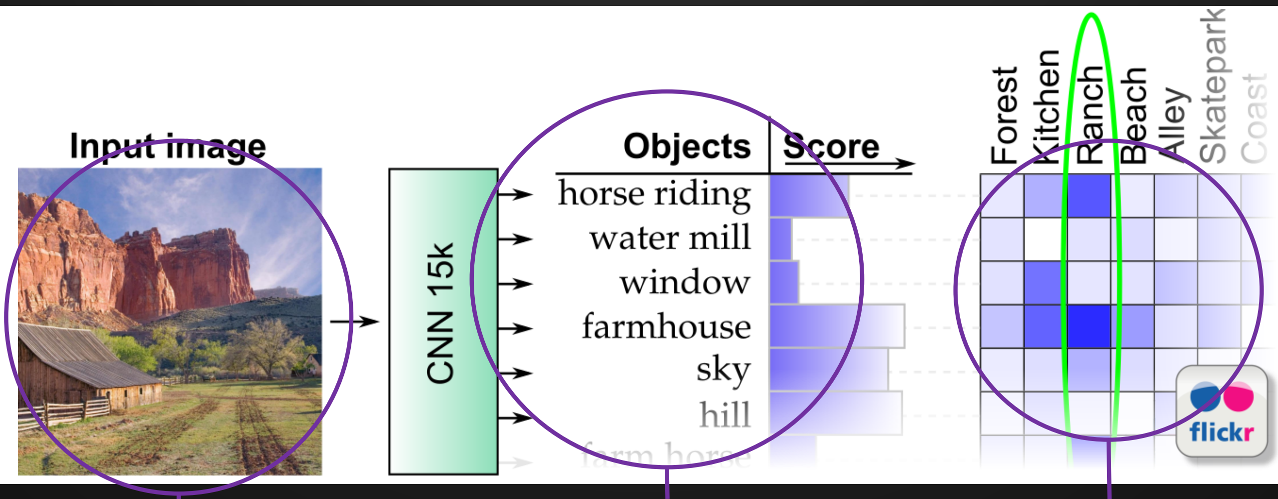


$$C(v) = \underset{z\in Z}{\operatorname{argmax}} \sum_{y\in Y} p(y|v)\, s(z,y) \ ,$$

Scene classes: $z \in Z$
Object classes: $y \in Y$

$$\boxed{s(z,y) = \cos\big(w(y), w(z)\big) = w(y)^T w(z)}$$

Knowledge sources



ImageNet 15,293 object categories

SUN Attributes 717 scene classes

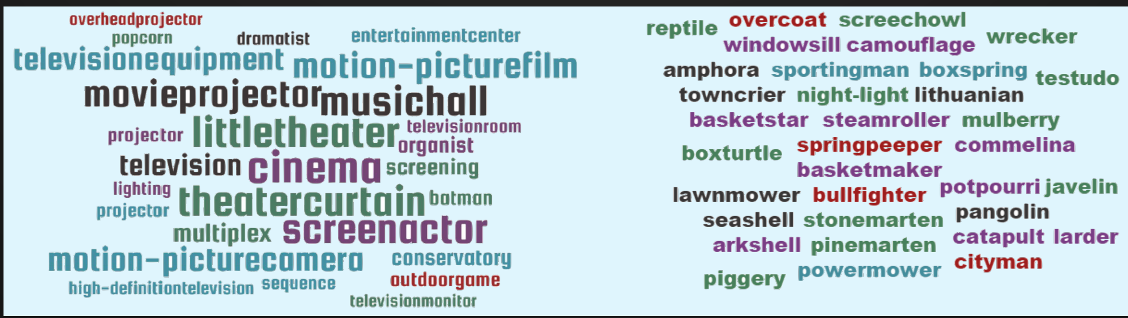
Places2 401 scene classes

YFCC100M 100 million Flickr images with titles, descriptions and tags

Not all 15K classes are relevant

Semantic relevance

Movie theater indoor



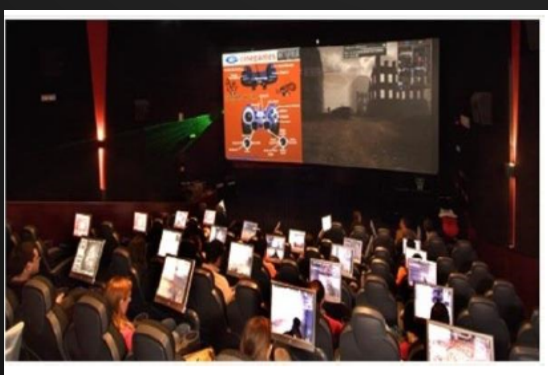
For all scenes $y \in Y$, select the semantically closest objects

$$Y_z = \{y \in Y \mid s(y,z) \; > \; t_s\}$$

Not all 15K classes are relevant

Appearance relevance

Movie theater indoor

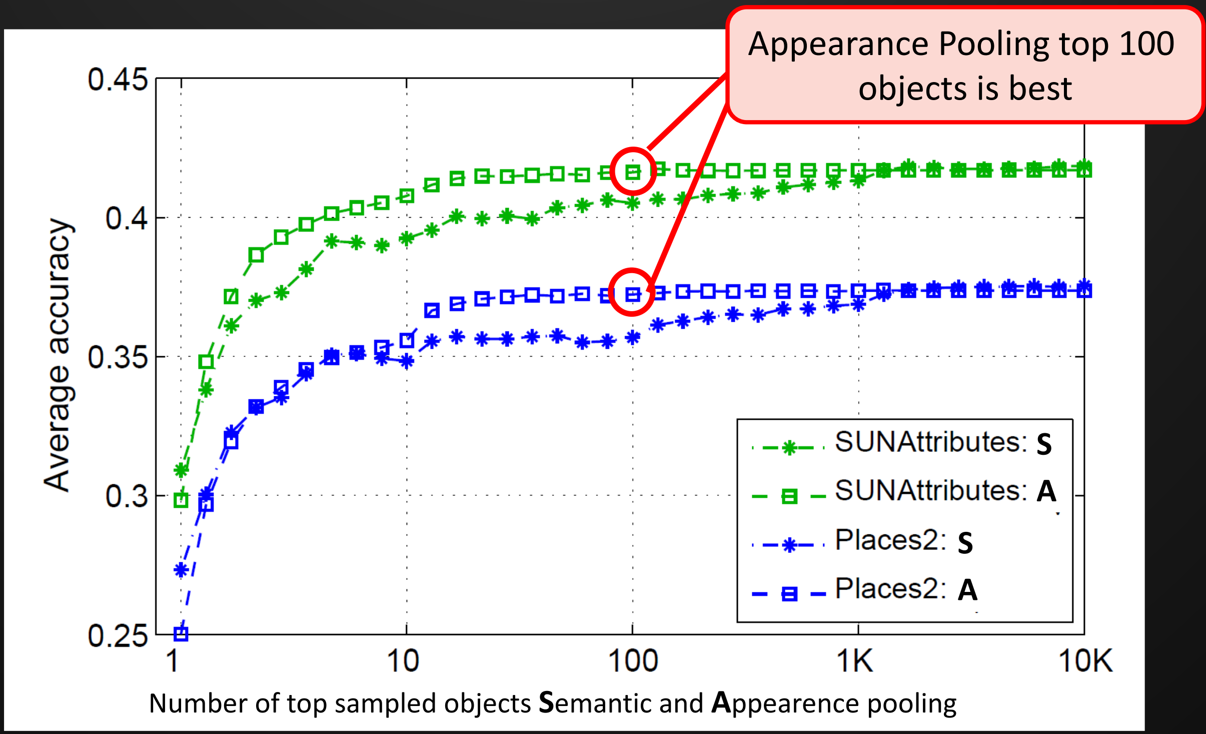


control center	0.97
audiovisual	0.15
assembly hall	0.12
television	0.09
equipment	0.06
orchestra pit	0.05
display	0.05
hall	0.03
lyceum	0.02
speaker	0.00

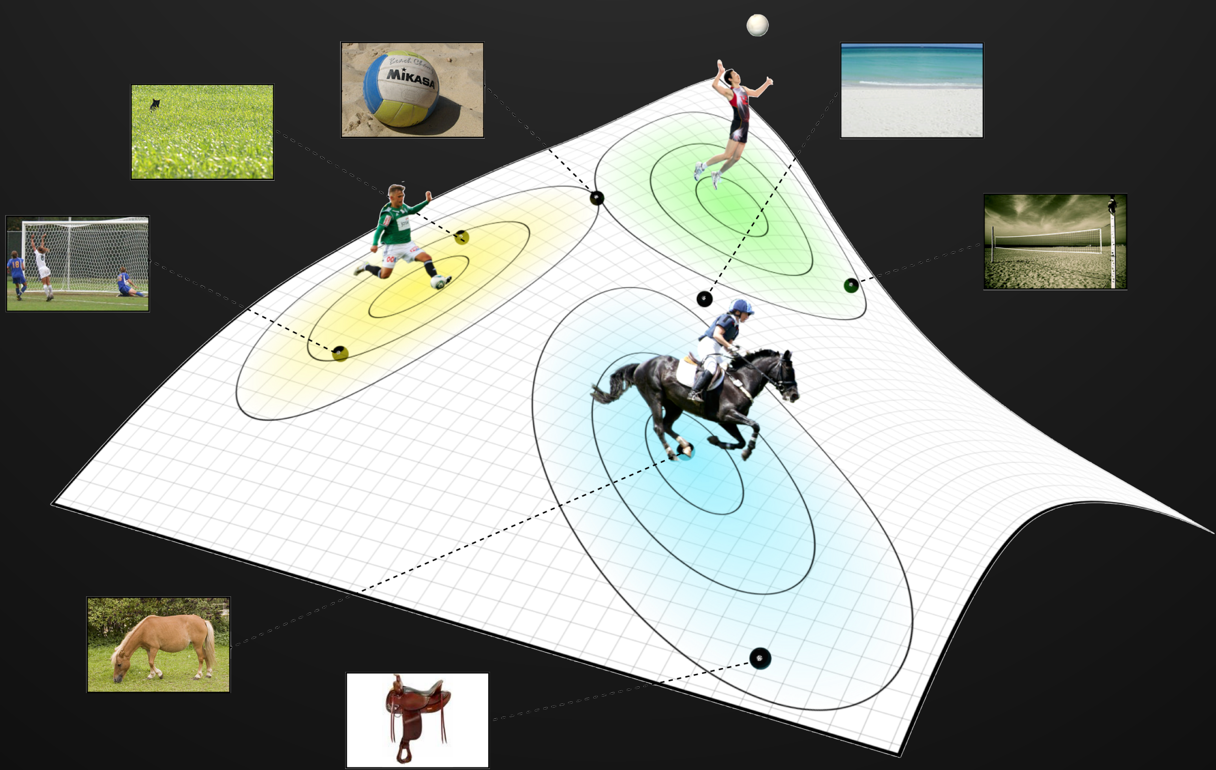
For each image v, select the objects with highest appearance likelihood

$$Y_v = \{y \in Y \mid p(v,y) \; > \; t_p\}$$

Appearance and Semantic Pooling



ImageNet Objects for Video Actions



Object and Action descriptions

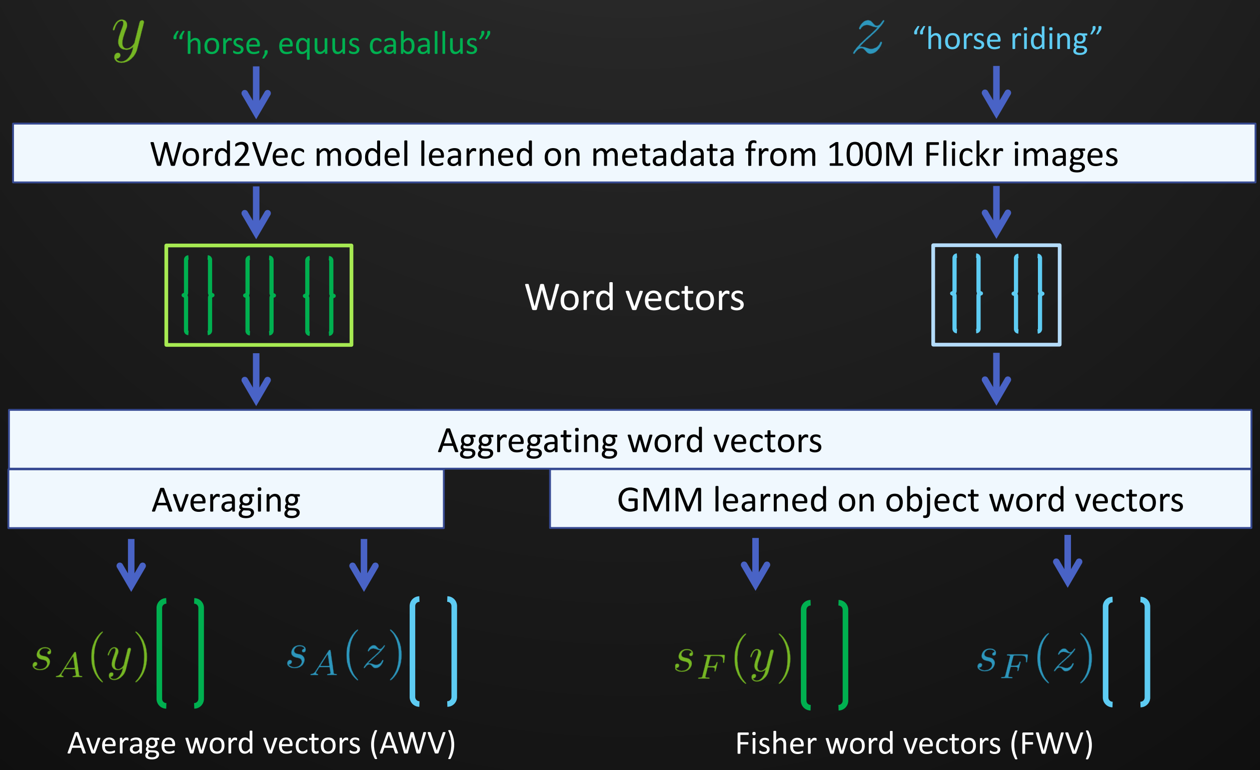
Object and *action* are described by a few words:

Objects: *car, elevator car*

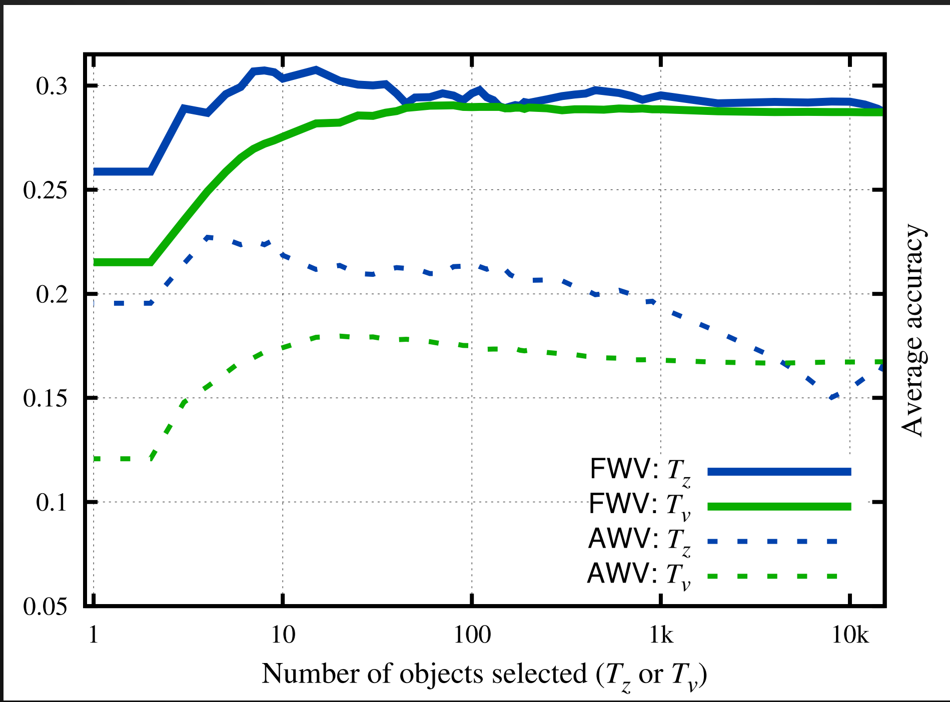
Definition: *where passengers ride up and down*

Actions: *Blow Dry Hair, Handstand Pushups, Ice Dancing*

Aggregating Word Vectors

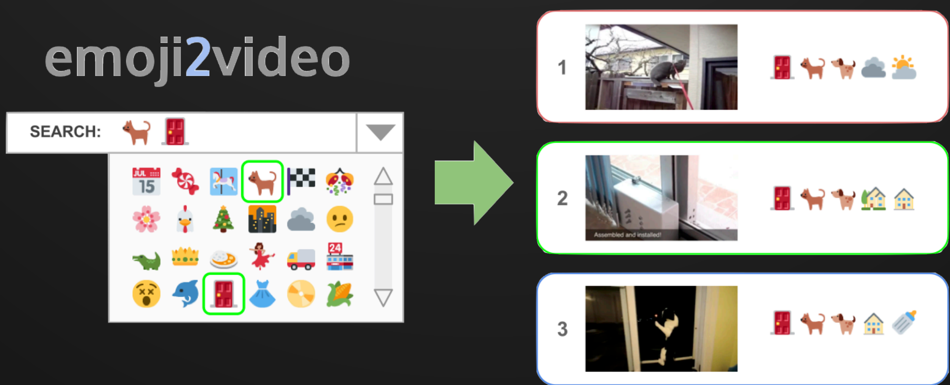


Fisher Word Vectors



Fun: Emoji2video

ImageNet object classifiers to emoji's in videos



Transductive View

Zero-shot: beat the shifts

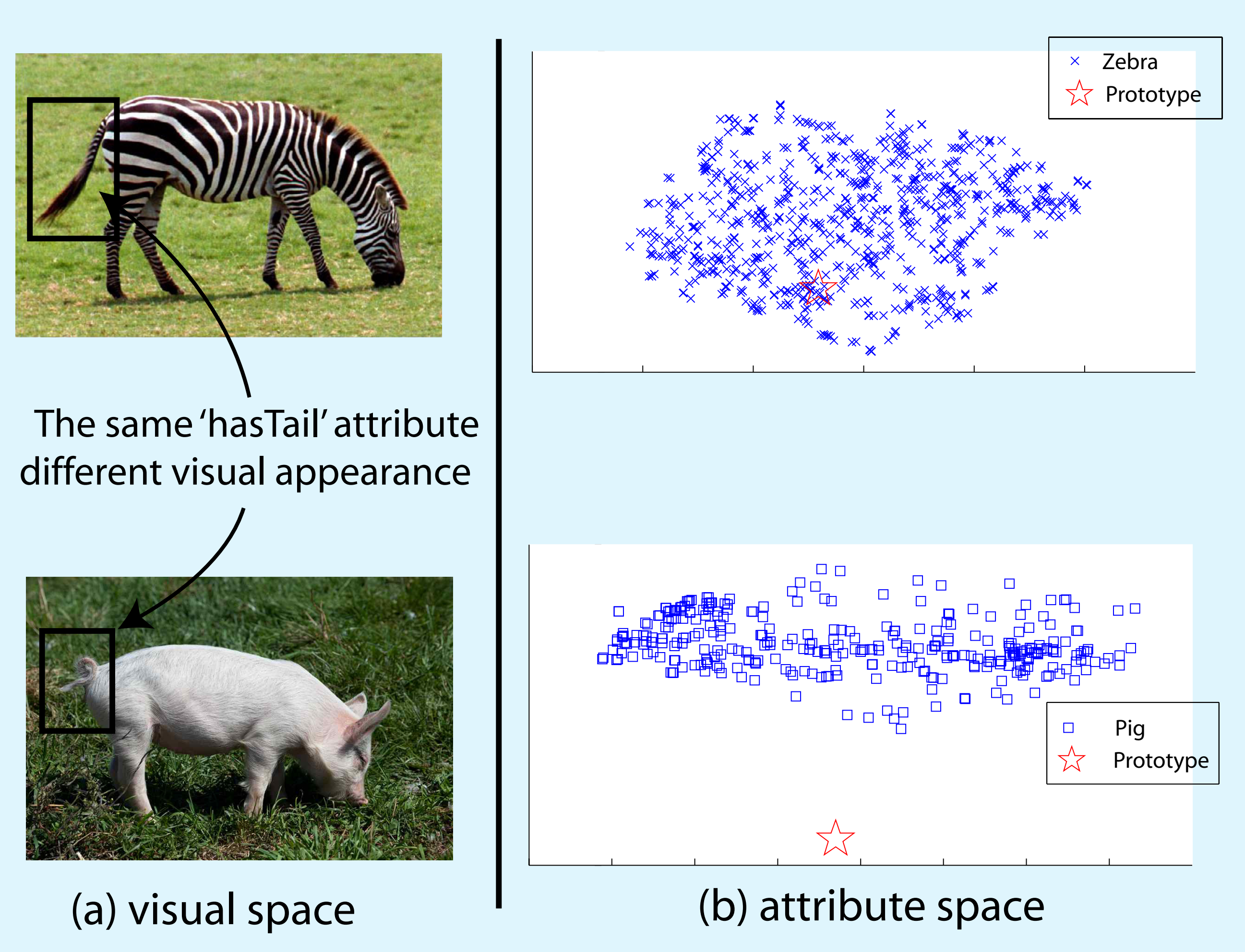
Semantic shift:

Transfer from known classes to unknown classes

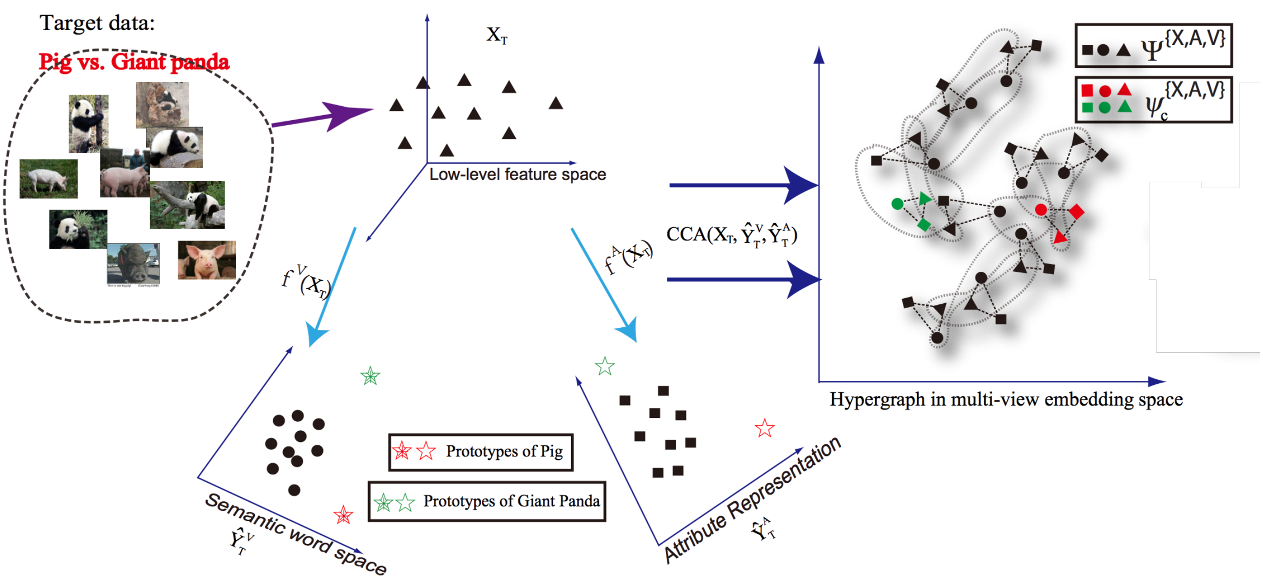
Domain shift:

Agnostic: train and test are both assumed $x \in \mathcal{X}$

Assumption:
attributes and images are iid over test and train set



Multiview Transductive Alignment



Multiview Transductive Alignment

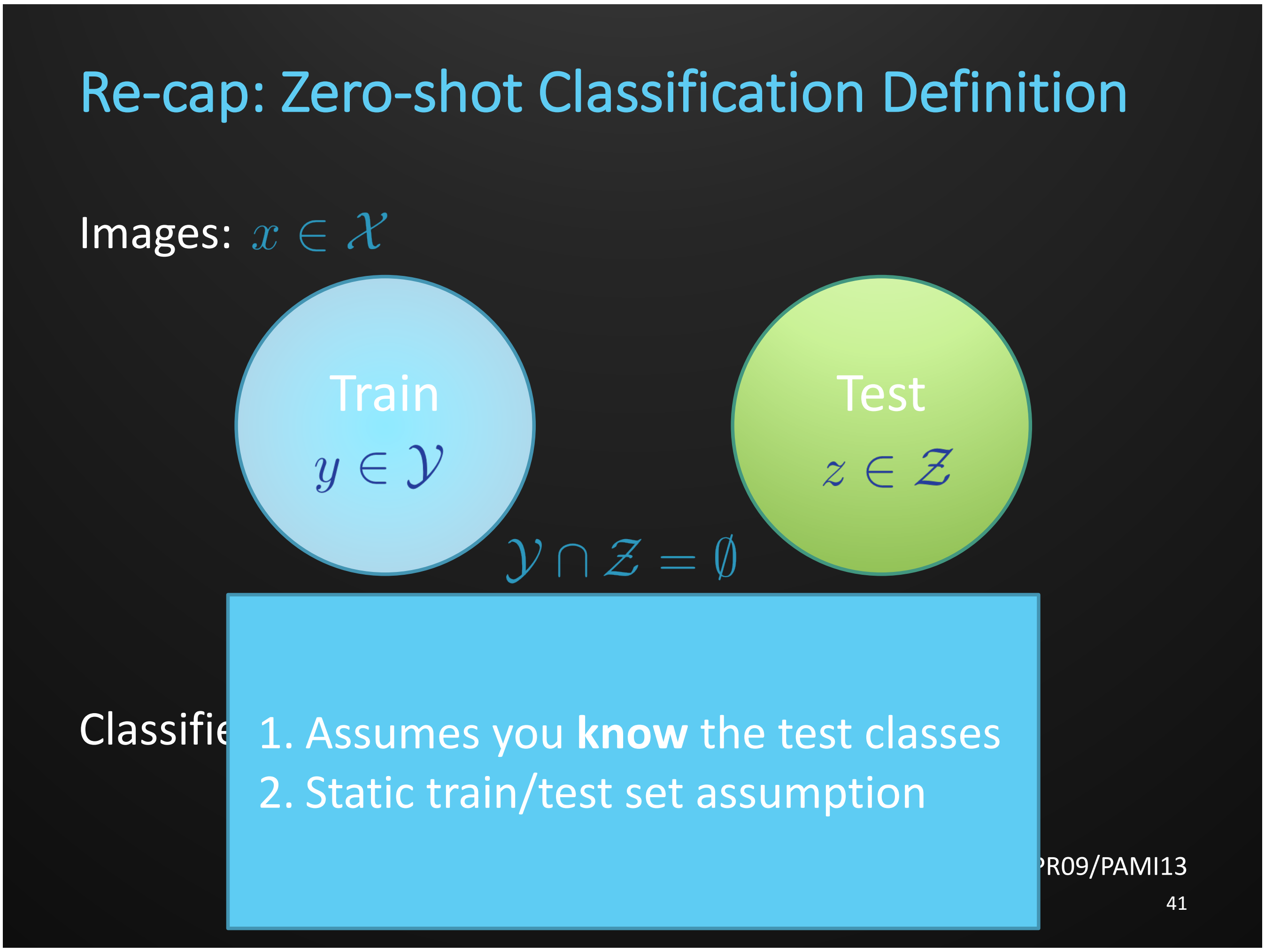
Animals with Attributes

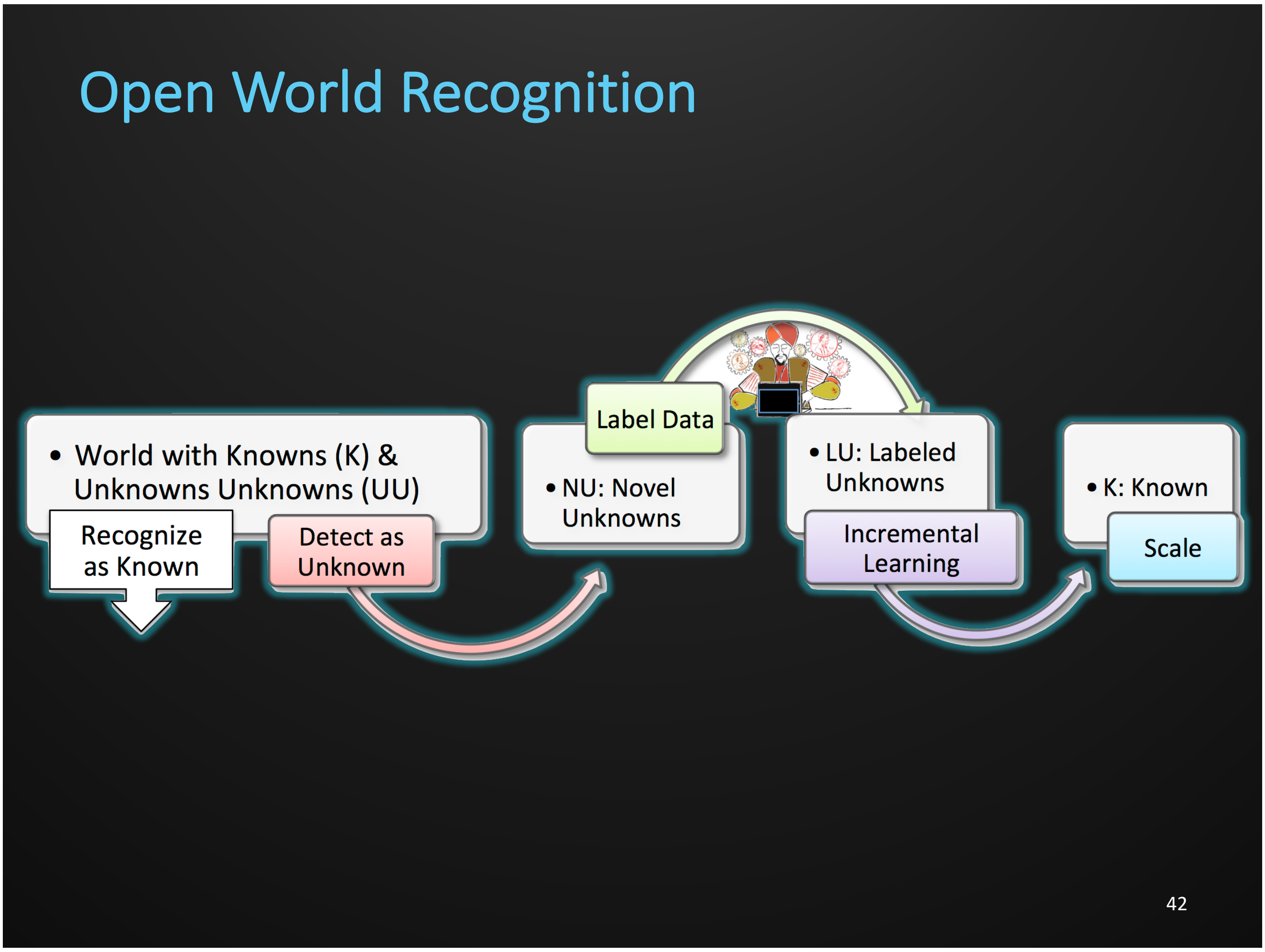
Method	Handcrafted	OverFeat	OverFeat+De Caf
DAP	41.4	51.0	57.1
Transductive	**49.0**	**73.5**	**80.5**

Test set distribution differs from train set

Knowing test set is beneficial for classification

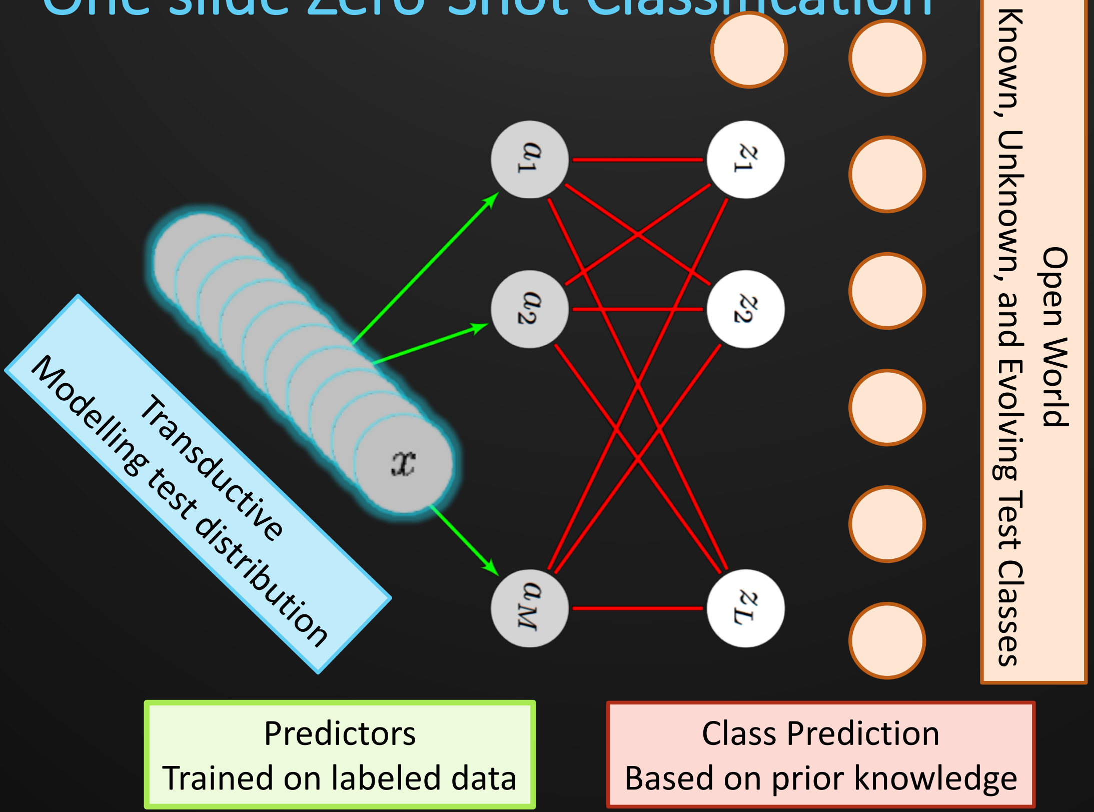
Open World


Re-cap: Zero-shot Classification Definition
Images: $x \in \mathcal{X}$
Train
$y \in \mathcal{Y}$
Test
$z \in \mathcal{Z}$
$\mathcal{Y} \cap \mathcal{Z} = \emptyset$
Classifie
1. Assumes you **know** the test classes
2. Static train/test set assumption
PR09/PAMI13
41


Open World Recognition
• World with Knowns (K) & Unknowns Unknowns (UU)
Recognize as Known
Detect as Unknown
Label Data
• NU: Novel Unknowns
• LU: Labeled Unknowns
Incremental Learning
• K: Known
Scale
42

One slide conclusion

One slide Zero-Shot Classification



Today's outline

1. Knowledge transfer
2. Classification
3. Localization

   Break

4. Retrieval
5. Interaction
6. Conclusion and Discussion

Zero-Shot Learning

with Localization

Efstratios Gavves

Traditional Localization

Training

Inference

Bicyclist

Zero-Shot Localization

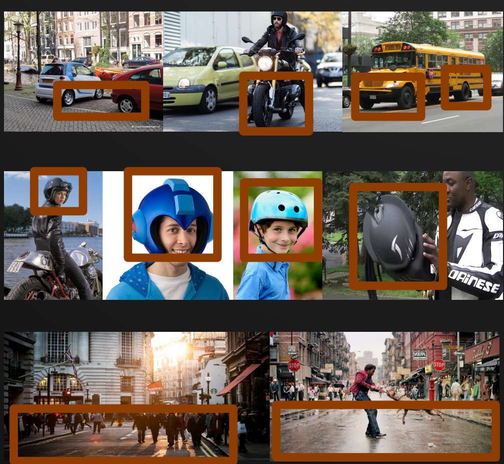
Training

Known visual classes

wheels

helmet

street



Zero-Shot Inference

Bicyclist=“wheels”+”helmet”+”street”

Why Zero-Shot Localization?

Find the object

Brown

Mammal

Curvy beak

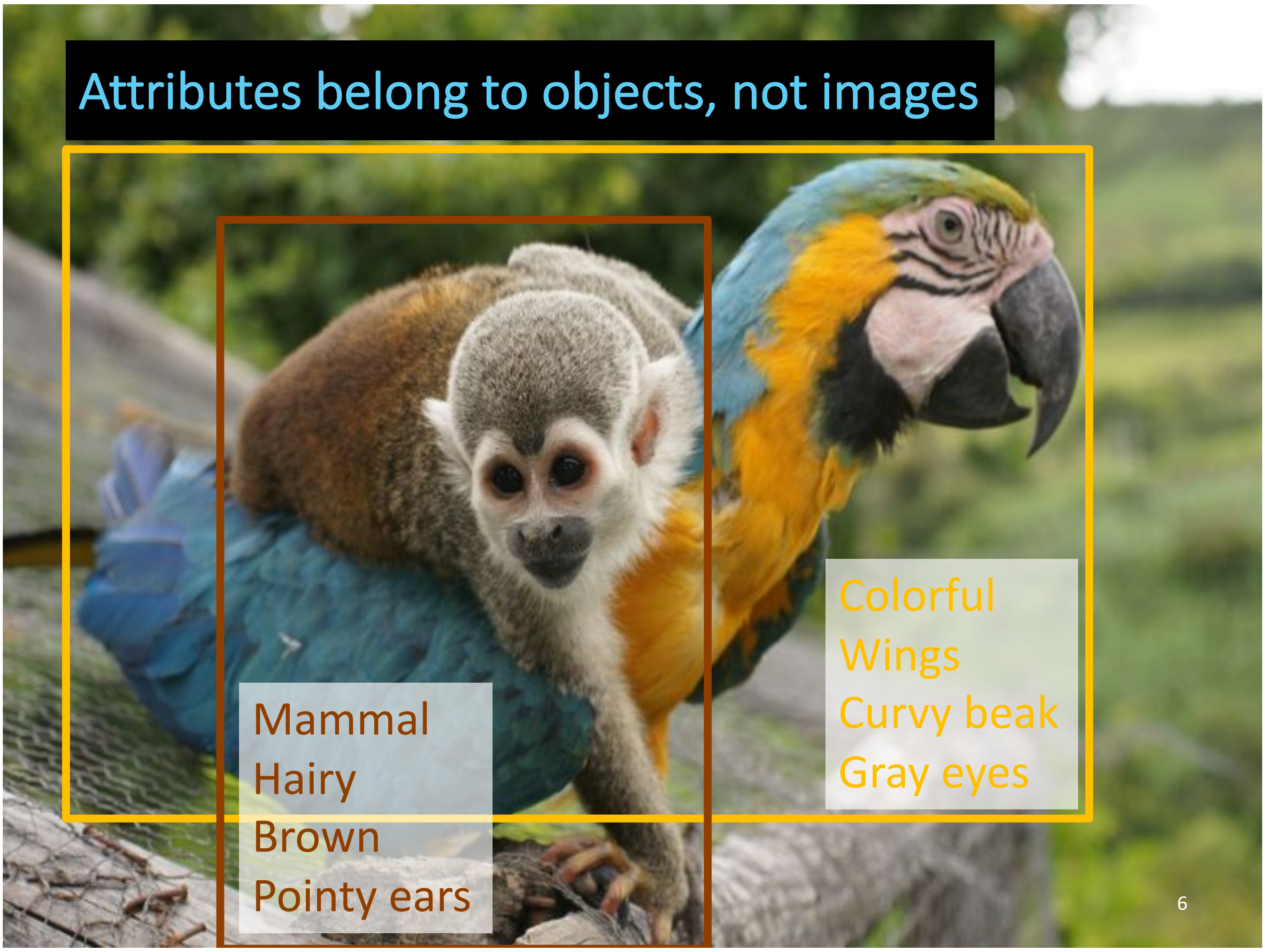
Pointy ears

Hairy

Wings

Colorful

Gray eyes


Attributes belong to objects, not images
Mammal
Hairy
Brown
Pointy ears
Colorful
Wings
Curvy beak
Gray eyes
6


Even more relevant in complex scenes
“Orange shirt”
“Sesame-street”
“Sunglasses”
“White pants”
7

Attributes lost with clutter

Horns
Brown color
White snout

Attributes lost with clutter

Horns
Brown color
White snout

Attribute signal is lost with clutter

Horns
Brown color
White snout

What is the spatial extent of attributes?

Visual details, e.g. “floral patterns”

- Must be discriminative
- Must be repeatable
- Must be salient
- Spatially specific

Regions

- More salient
- Attributes do not have to be visually groundable, e.g., “retro”
- But less specific

At the level of visual details

Learn attributes that are

- discriminative
- machine-detectable

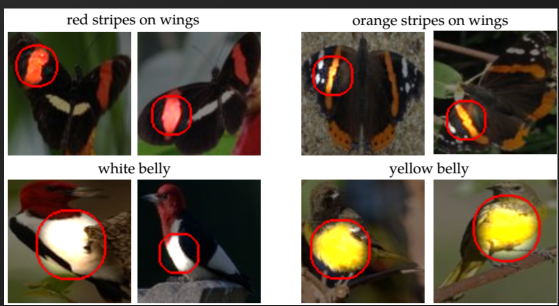
Also, semantically meaningful

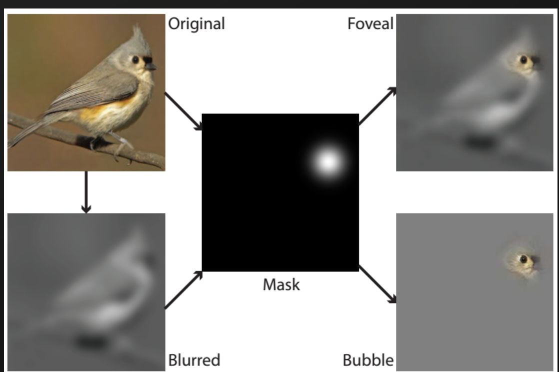
- By design: human in the loop [1]
- By unsupervised clustering [2]

Properties

- Spatially precise
- CNN too invariant (?)

Not explicitly for Zero-Shot





[1] Discovering Localized Attributes for Fine-Grained Recognition, Duan et al., CVPR 2012
[2] BubbLeNet: Foveated Imaging for Visual Discovery, Matzen and Snavely, ICCV 2015

At the level of visual details

$$E(L_k|\mathcal{I}) = \sum_{i=1}^{M} \phi_k(l_i^k|\mathcal{I}_i) + \sum_{i=1}^{M}\sum_{j=1}^{M} \psi_k(l_i^k, l_j^k|\mathcal{I}_i,\mathcal{I}_j)$$

$$E(\mathcal{L}|\mathcal{I}) = \sum_{k=1}^{K} E(L_k|\mathcal{I}) + \sum_{i=1}^{M}\sum_{k,k'} \delta(l_i^k, l_i^{k'}|\mathcal{I}_i)$$

Specific attribute CRF

Set of attributes CRF

Automatically detect discriminative attributes

- Solve CRFs iteratively
- Random attribute initialization

Not necessarily "nameable"

- Convert them to nameable
- Human approves meaningful attributes

[1] Discovering Localized Attributes for Fine-Grained Recognition, Duan et al., CVPR 2012

Zero-shot Localization by Attributes

First to do region-level, attribute based localization [1]

Extract regions localization (CPMC, ~500) [2]

Learn attributes with ALE[3]

$$f(x) = \arg\max_{y\in\mathcal{Y}} \max_{z\in Z(x)} F(z,y)$$

$$F(z,y;W,\phi) = \theta(z)'W\phi(y)$$

$$\min_{W} \frac{\lambda}{2}||W||^2 + R(W,\Phi^{\mathcal{A}})$$

Efficient inference by codemaps [4]

ALE attributes

~500

Per region maximization

[1] Attributes make sense on segmented objects, Li et al., ECCV 2014
[2] Constrained Parametric Min-Cuts for Automatic Object Segmentation, Carreira et al., CVPR 2010
[3] Label-embedding for attribute-based classification, Akata et al., CVPR 2013
[4] Codemaps segment, classify and search objects locally, ICCV, 2013

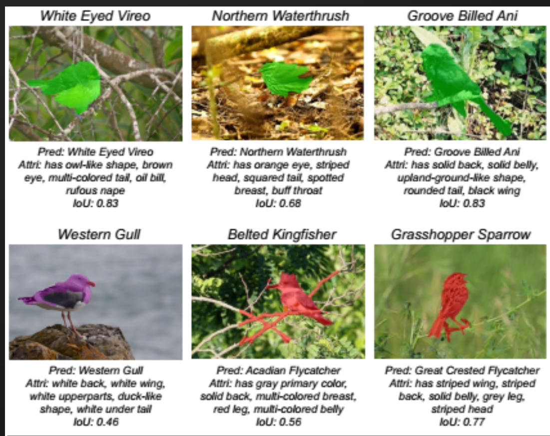
Zero-shot Localization by Attributes

Zero-Shot Localization as Structured Prediction

- Regions are latent variables

Evidence for accidental Zero-Shot recognition

- Mean Class Accuracy (MCA) higher than MCA on well predicted segments (MSO)
- Maybe segment wrong (<50%) but descriptive
- Maybe segment mostly on background

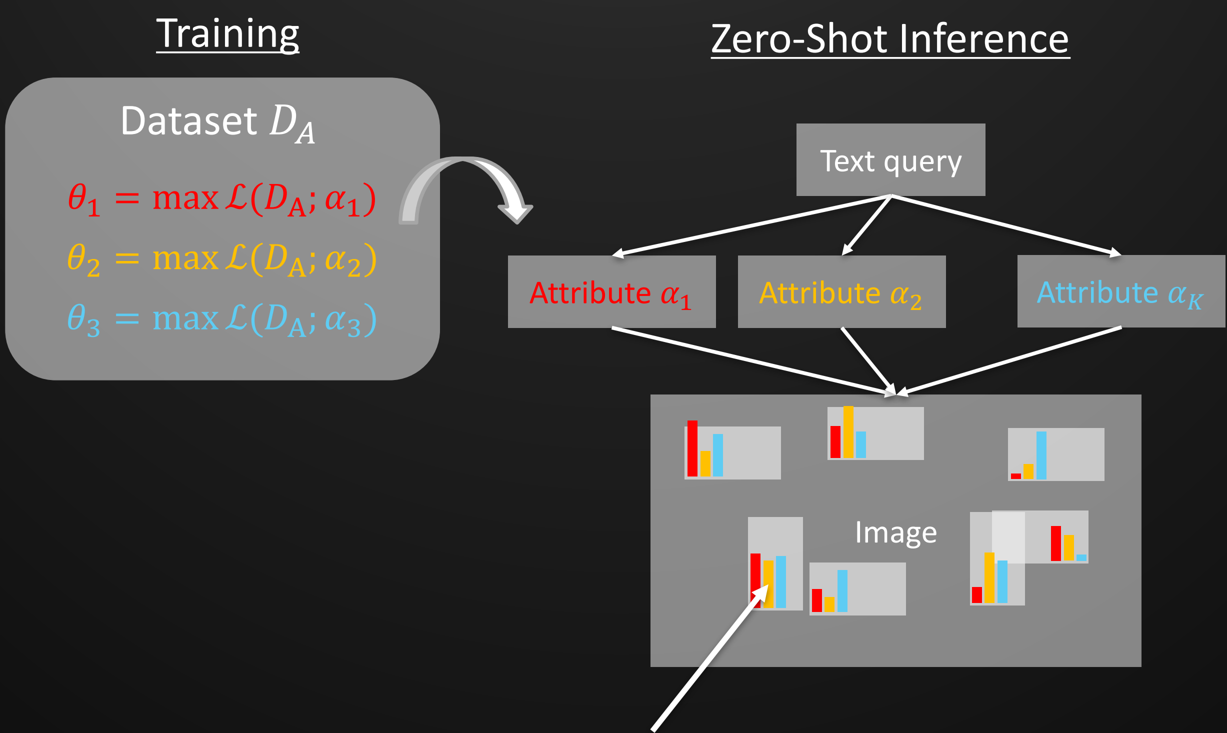


Setting	Codebook	Entire image	Object-level attributes		
		MCA	MCA	MSO	AO
Supervised	$k = 16$	27.1	51.5	43.0	61.8
Zero-shot	$k = 16$	11.3	15.7	12.4	56.3

Accidental Zero-Shot in action

[1] *Attributes make sense on segmented objects*, Li et al., ECCV 2014
[2] *Label-embedding for attribute-based classification*, Akata et al., CVPR 2013
[3] *Codemaps segment, classify and search objects locally*, ICCV, 2013

Zero-shot Localization by Attributes



[1] *Attributes make sense on segmented objects*, Li et al., ECCV 2014

Zero-shot Localization by Attributes

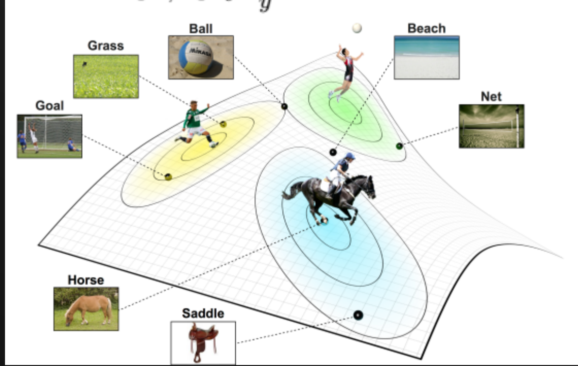
Similar for videos & actions [1]

Instead of CPMC, spatiotemporal action proposals

Replace attributes with Word2Vec

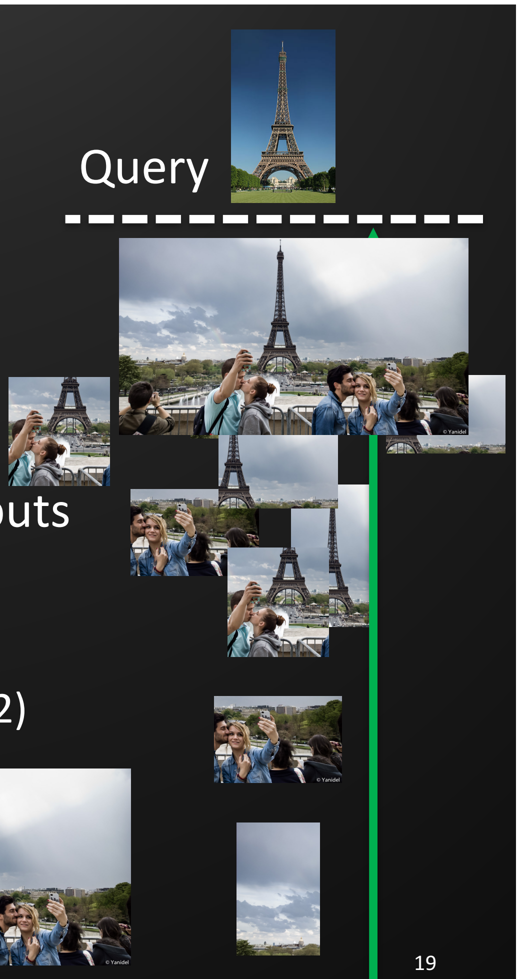
- Aggregate Word2Vec by Fisher vectors

$$\mathcal{C}(v) = \underset{z\in\mathcal{Z},u\in\mathcal{U}_v}{\operatorname{argmax}} \sum_y p_{uy}\, g_{yz}$$



[1] *Objects2action: Classifying and localizing actions without any video example,* Jain et al., ICCV 2015

Localization as Retrieval



Goal: Find the target in the image

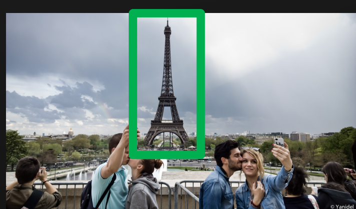
- ranking sliding window images

Sliding window search

- thousands of images generated

Learn scoring function with two inputs

- Input #1: Query image
- Input #2: Sliding image
- Output: Siilarity(Input #1, Input #2)



Zero-shot Localization by Free Text

Similar to Zero-Shot Localization [1]

- #Input 1 is now a text query

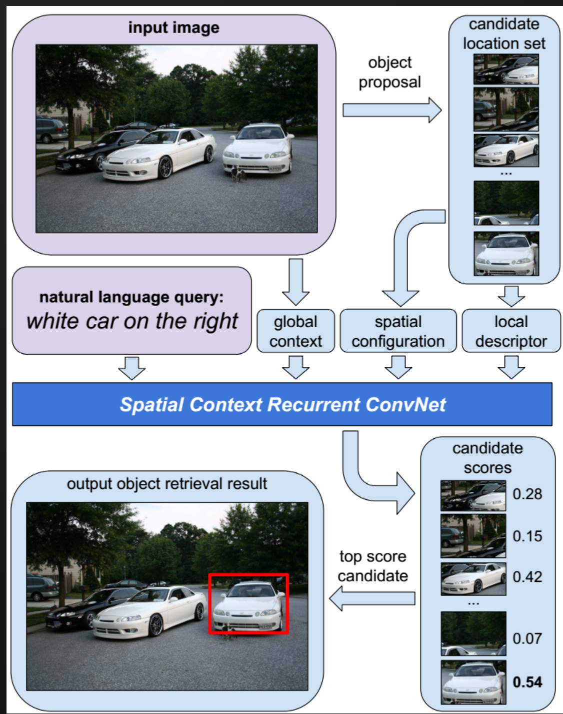
Rank sliding images

- Scoring function measures similarity of image to text

$$p(w_{t+1}|w_t,\cdots,w_1,I_{box},I_{im},x_{spatial})$$
$$= \mathrm{Softmax}(W_{local}h_{local}^{(t)} + W_{global}h_{global}^{(t)} + r)$$

$$\boxed{\begin{aligned} s &= p(S|I_{box},I_{im},x_{spatial}) \\ &= \prod_{w_t \in S} p(w_t|w_{t-1},\cdots,w_1,I_{box},I_{im},x_{spatial}) \end{aligned}}$$

$$L = -\sum_{i=1}^{N}\sum_{j=1}^{M_i}\sum_{k=1}^{K_{i,j}} \log(p(S_{i,j,k}|I_{box\,i,j},I_{im\,i},x_{spatial\,i,j}))$$



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

Zero-shot Localization by Free Text

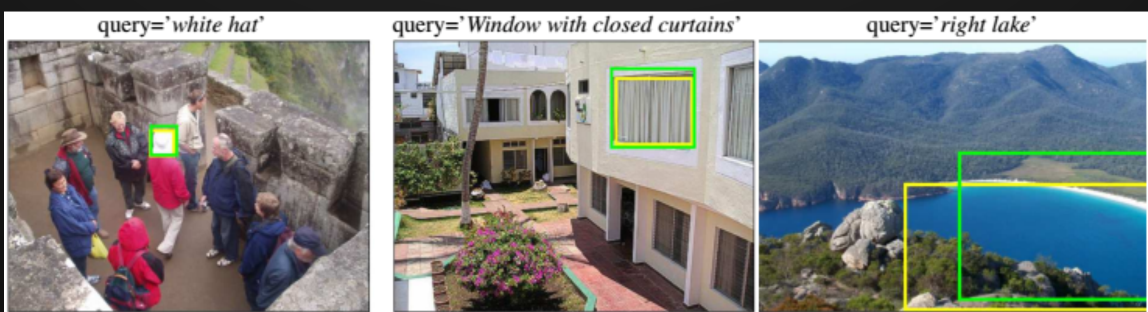
Semantic attributes

- “hat”, “white”, …

Spatial attributes too

- “right”, “on top of”, “below”, …

Global context



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

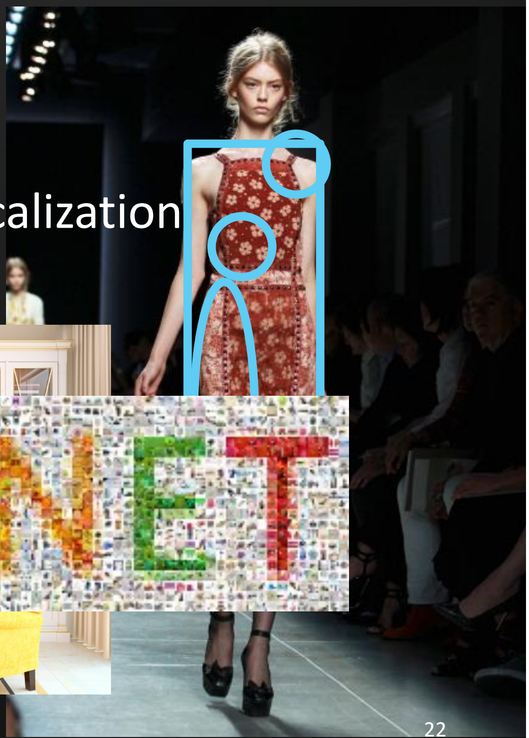
Going to the next level

Detection by context

Very large scale

- Better transfer learning

Joint region- and detail- level of localization



Conclusion

Attributes belong to objects, not images

Zero-Shot localization natural extension

Focus on visual Details or Regions

- Each with their merit, depends on application
- Maybe a smart combination?

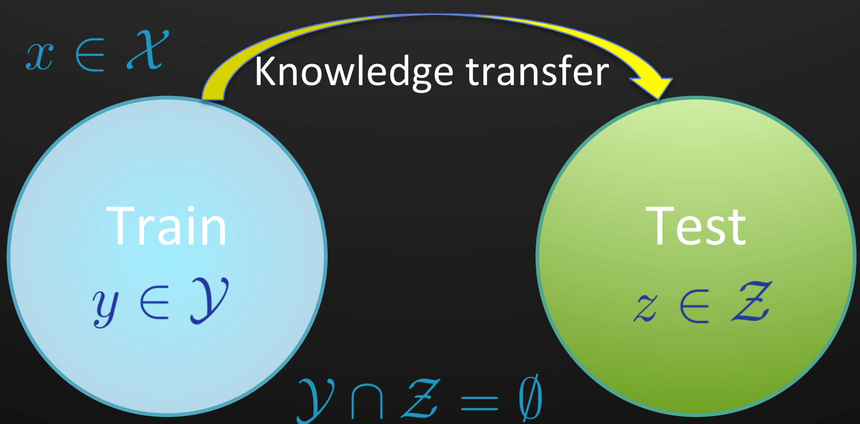
Zero-Shot Learning

for Retrieval

Cees Snoek

What is this tutorial about?

Data: $x \in \mathcal{X}$



Objective: $f : \mathcal{X} \rightarrow \mathcal{Z}$

Lampert et al., CVPR09/PAMI13

Today's outline

1. Knowledge transfer
2. Classification
3. Localization

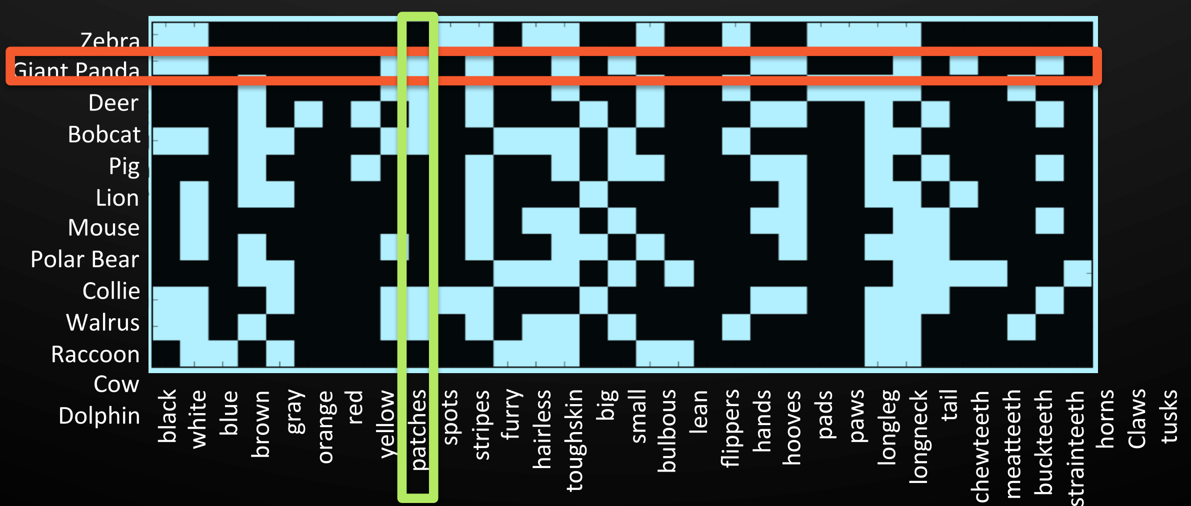
   Break

4. Retrieval
5. Interaction
6. Conclusion and Discussion

Lampert et al PAMI 2013, and many others

Zero-shot classification vs retrieval

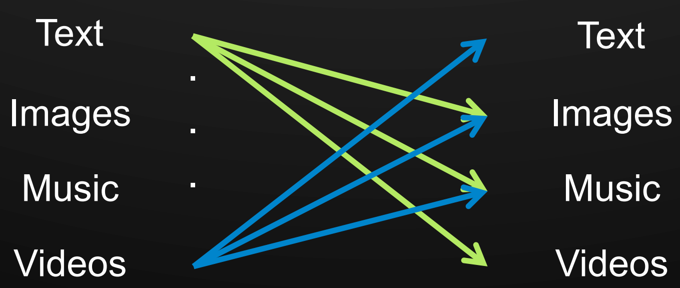
Classify test videos by (predefined) mutual relationship using class-to-attribute mappings



In retrieval we typically rely on a description only

Related work: Cross-modal retrieval

Given query from modality A, retrieve results from modality B, where A!= B.



We focus today on text to visual and vice versa

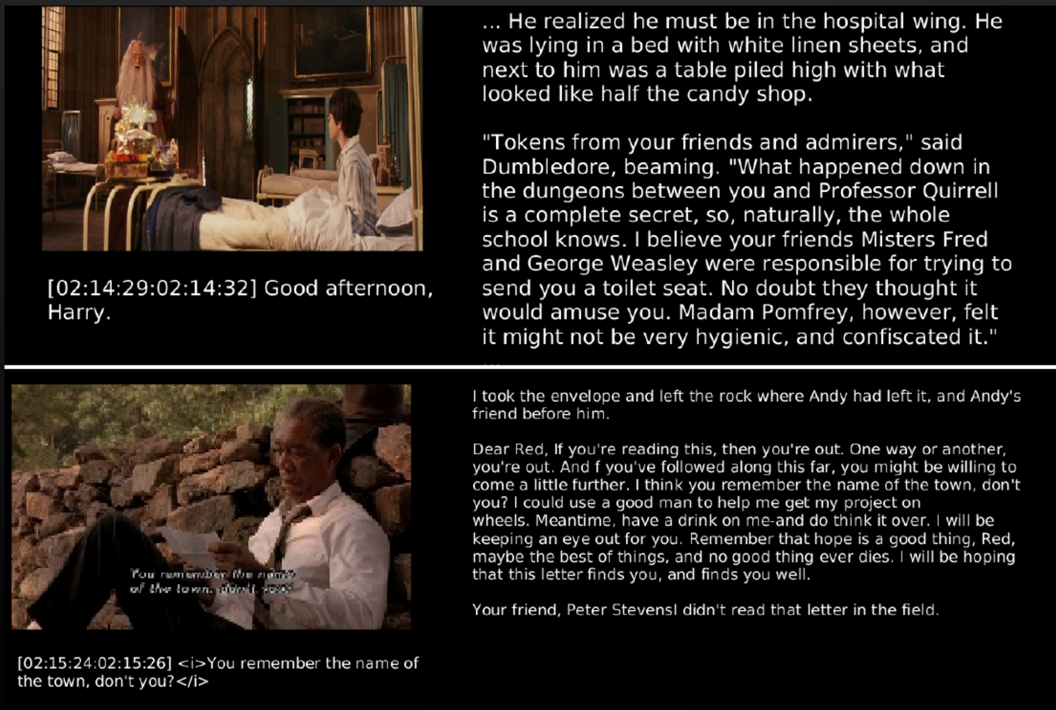
Retrieving images from Wikipedia text

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

Rasiwasia MM'10 / Costa TPAMI'14

Retrieving book excerpts from movies



Zhu ICCV'15

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 ...

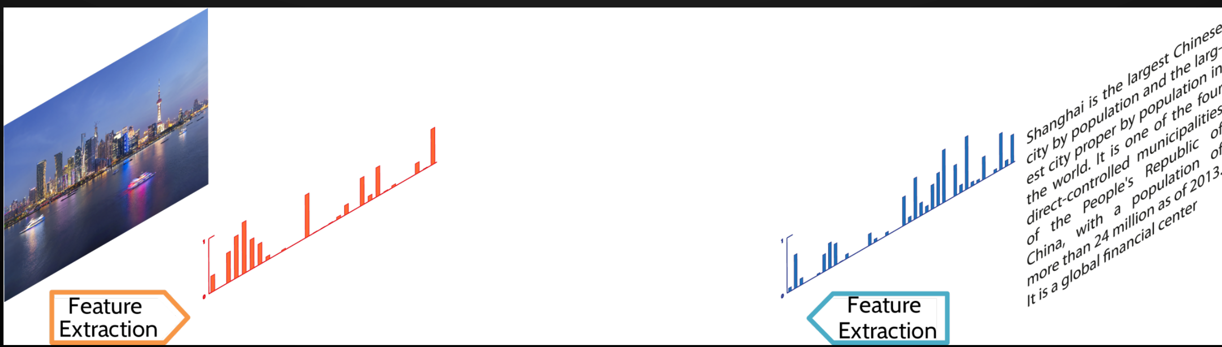
Event Name: Winning a race without a vehicle

NIST TRECVID Multimedia Event Detection Benchmark

Problem statement

How to align visual and textual representations?

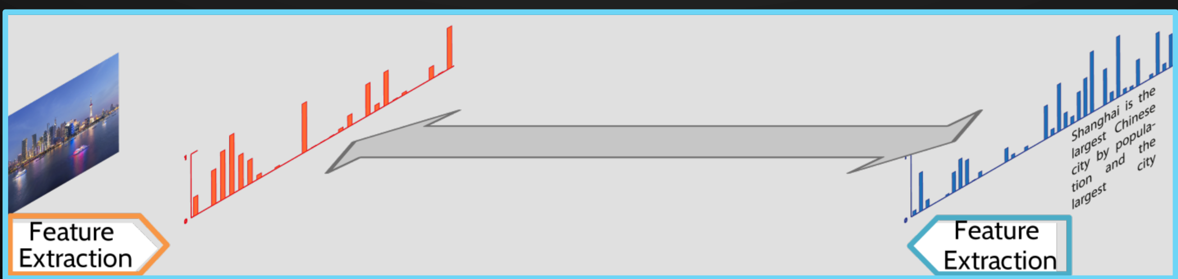
Different dimensionality, distributions, and meaning



Low-level alignment

Aligns two modalities directly at low-level features

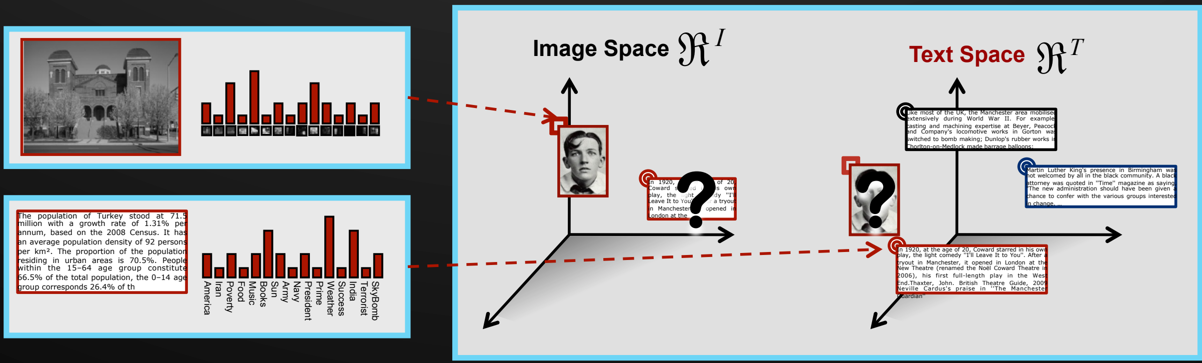
Canonical Correlation Analysis, Cross-Media hashing, …



Not the most effective space to learn the correlations

[Li et al., MM' 03] [Rasiwasia et al., MM'10] [Ballan et al., ICMR'14]

How to compute similarity?

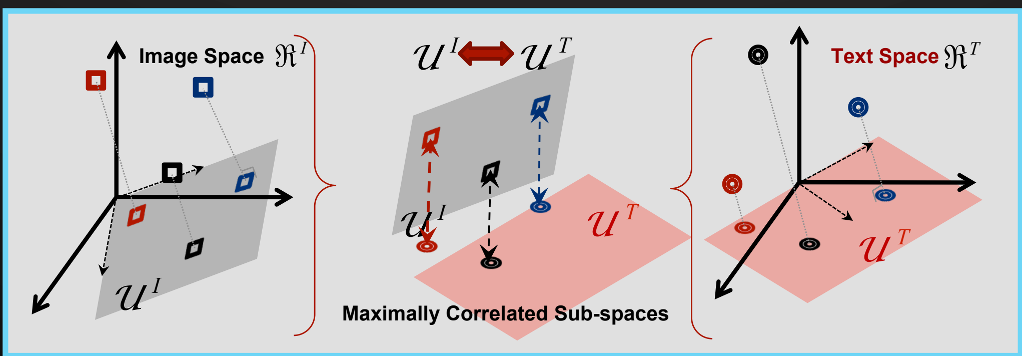


Slide credit: Nikhil Rasiwasia

Slide credit: Nikhil Rasiwasia

Canonical Correlation Analysis

Learn subspaces that maximize correlation between two modalities



Joint dimensionality reduction across two (or more) spaces

$$\max_{w_i \neq 0, w_t \neq 0} \frac{w_i \Sigma_{IT} w_t}{\sqrt{w_i \Sigma_{II} w_i}\sqrt{w_t \Sigma_{TT} w_t}}$$

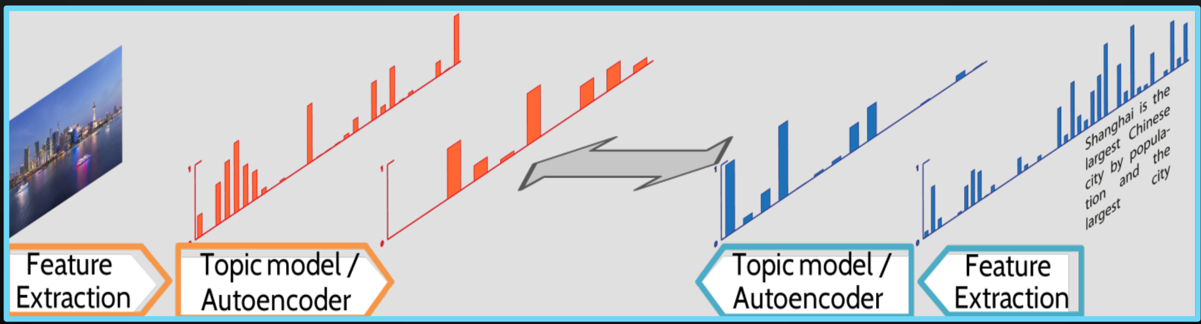
Basis for the maximally correlated space

Empirical covariance for images and text, and their cross covariance.

Mid-level alignment

Aligns two modalities at mid-level features

Extracted by autoencoders, topic models,…



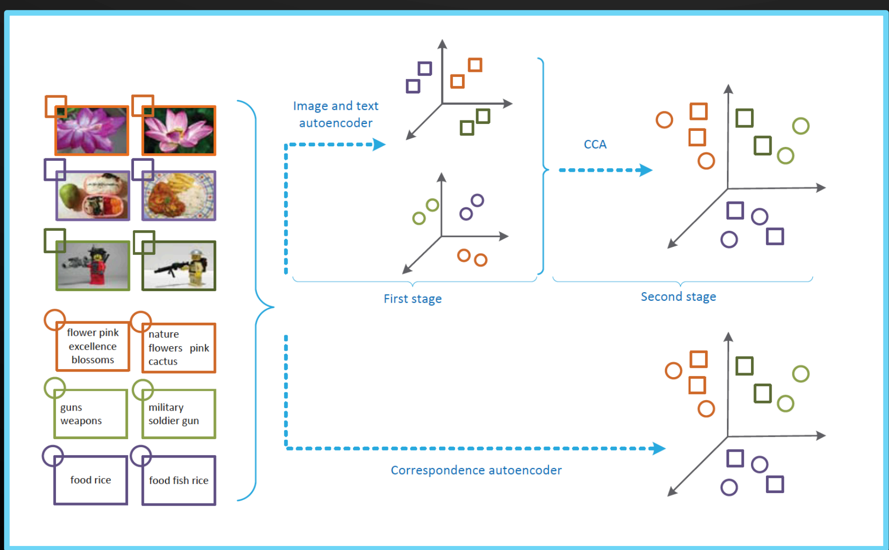
Topic modeling on visual descriptors not straightforward

Deep autoencoders less suited for small datasets

[Blei et al., SIGIR’03] [Wang et al., MM’14] [Feng et al., MM’14] …

Correspondence autoencoder

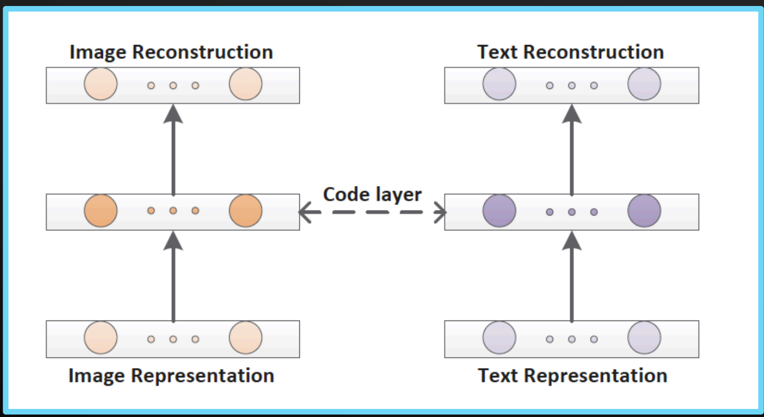
Essentially an end-to-end version of CCA



Feng MM’14

Simplified architecture

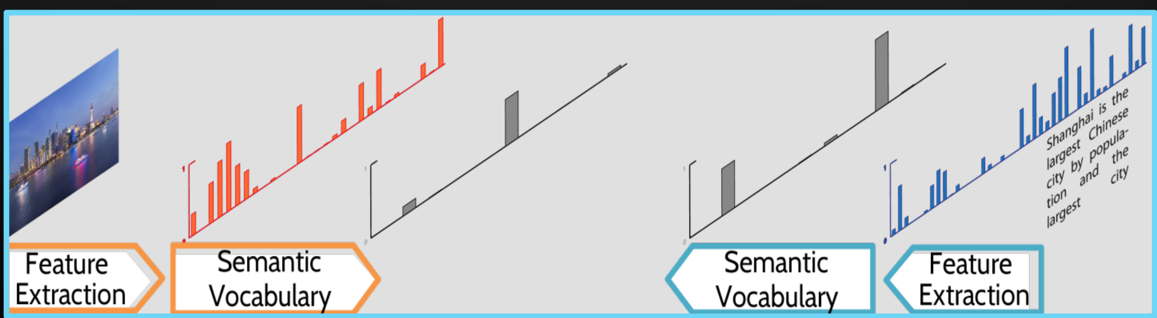
Networks coupled at code layer via similarity measure



Feng MM'14

Semantic alignment

Embeds images and texts into a mutual semantic space
Semantic space is defined by a vocabulary of concepts
Each concept has a visual and a textual classifier



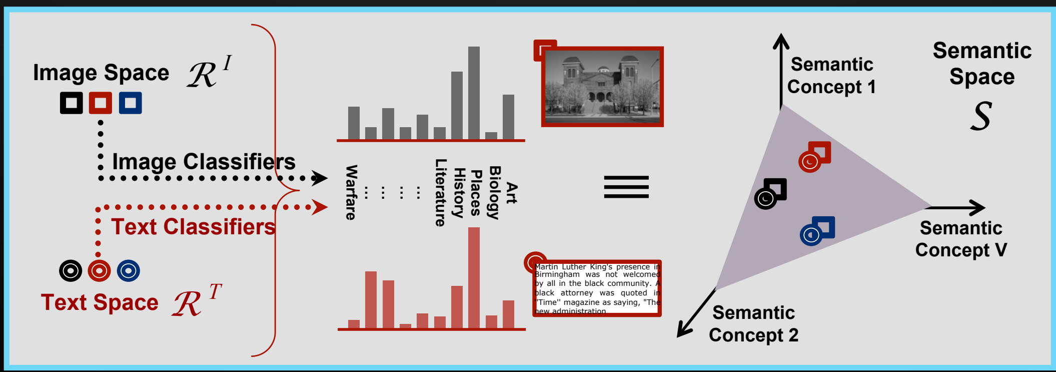
[Smith et al., ICME'03] [Hauptmann et al., TMM'07][Rasiwasia et al., MM'10] ...

Semantic alignment via concepts

Design semantic spaces for both modalities

A space where each dimension is a semantic concept.

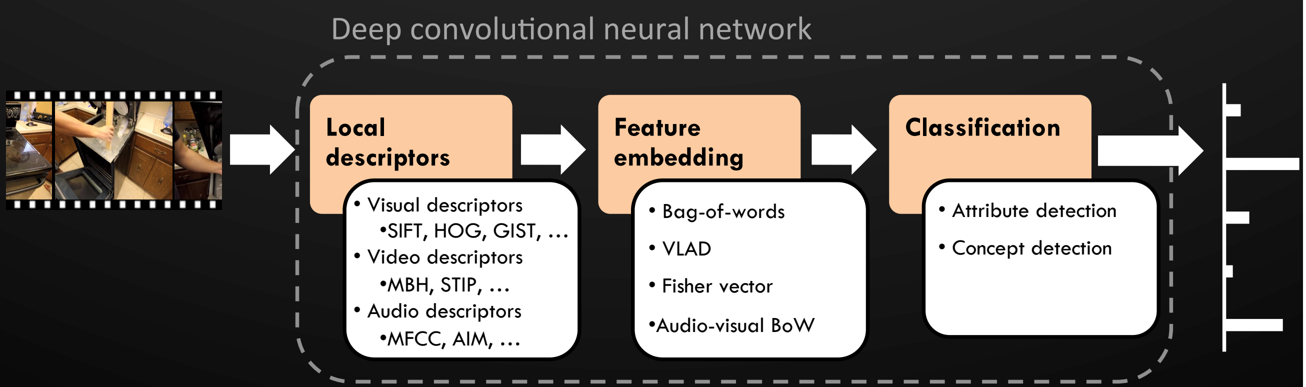
Each point on this space is a weight vector over these concepts



Rasiwasia MM'10 / Costa TPAMI'14

Semantic alignment via concepts

Representing image/video as histogram of concept scores



New problem: define, annotate and train concept classifiers

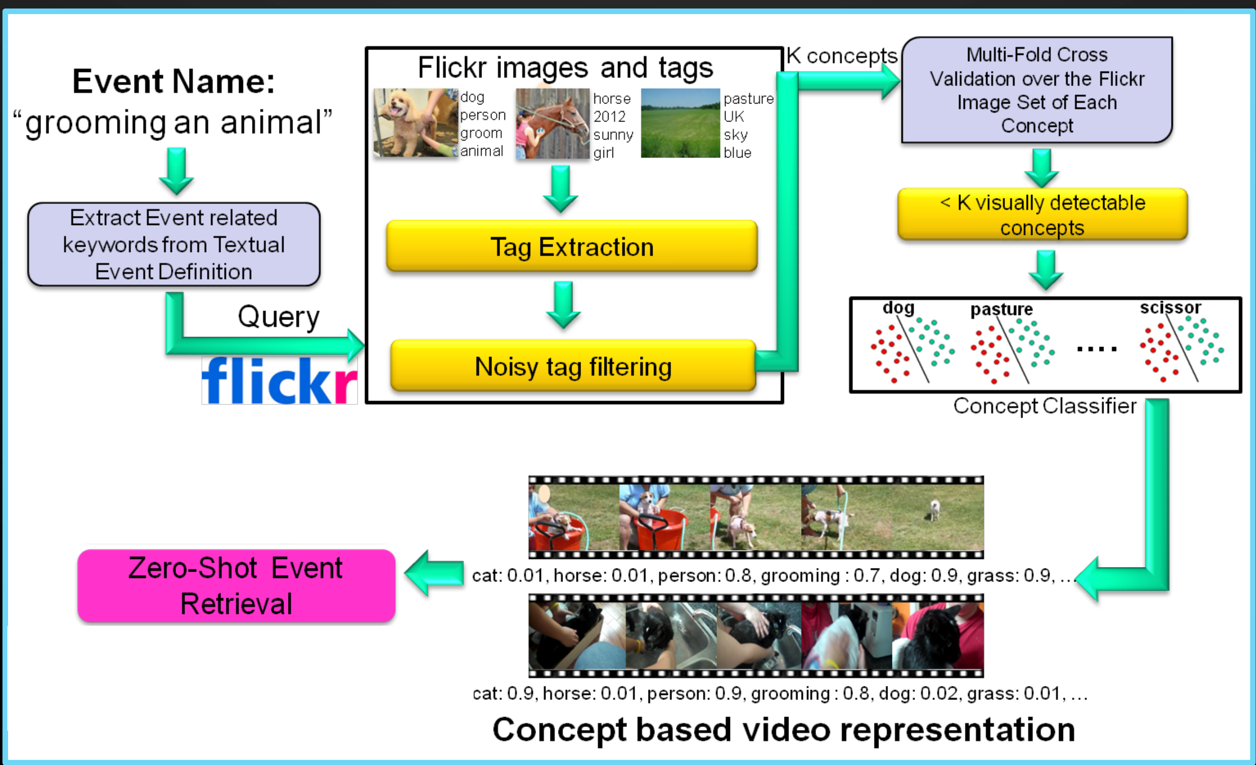
A solution: search engine transfer

+

YouTube flickr

Chen et al. ICMR 2014
Wu et al. CVPR 2014

Discovering concepts from the web



Drawbacks of concept discovery

Representation somewhat ad hoc

Many concepts are rare, insufficient examples to train reliable visual classifiers

Selection is based on visual prediction accuracy only, descriptiveness is ignored

Contextual information is lost, since concepts are learned independently by binary classifiers.

Habibian, TPAMI’17

Zero-Shot Learning with Video2vec

Semantic alignment via multimedia embedding

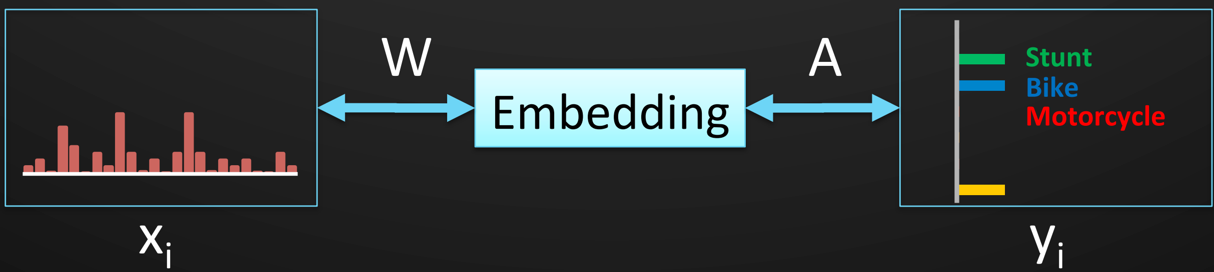
Video

Story

Story usually highlights the key concepts in video

Videos and stories are freely available, *i.e.* YouTube

Traditional embedding



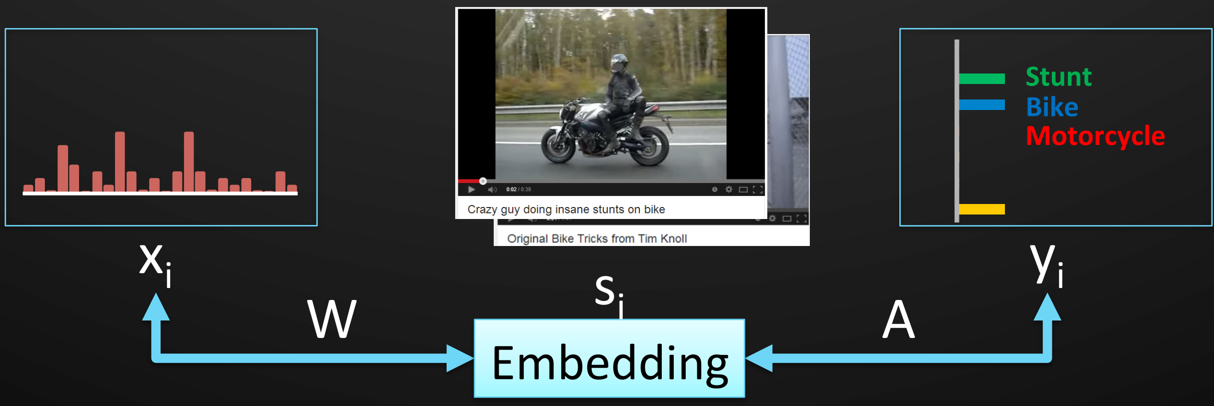
Joint space where $x_i W \approx y_i A$

Explicitly relate training W and A from multimedia

W = Identity matrix	individual term classifiers
A = Projection matrix	select/group terms

[Rasiwasa et al., MM 2010] [Weston et al., IJCAI 2011] [Akata et al., CVPR 2013] [Das et al., WSDM 2013]

Video2vec: Embed the story of a video



Design criteria: learn W and A such that

- *Descriptiveness:* preserve video descriptions
- *Predictability:* recognize terms from video content

Key observation: Compelling forces

Crazy guy doing insane stunts on bike

Why is this important?

Grouping terms:

Number of classes is reduced

Training classifiers per group:

More positive examples available per group

We can train from freely available web data

Key contribution: Joint optimization

Jointly optimize for descriptiveness and predictability

$$L_{\mathrm{VS}}(\boldsymbol{A}, \boldsymbol{W}) = \min_{\boldsymbol{S}} L_d(\boldsymbol{A}, \boldsymbol{S}) + L_p(\boldsymbol{S}, \boldsymbol{W})$$

Hyperparameter: size of the embedding S

L_d Loss function for descriptiveness

L_p Loss function for predictability

Video2vec connects the two loss functions

Video2vec objectives: descriptiveness

Objective 1: The Video2vec embedding should be **descriptive**

$$L_d(\boldsymbol{A}, \boldsymbol{S}) = \frac{1}{N}\sum_{i=1}^{N} \|\boldsymbol{y}_i - \boldsymbol{A}\boldsymbol{s}_i\|_2^2 + \lambda_a \Omega(\boldsymbol{A}) + \lambda_s \Psi(\boldsymbol{S})$$

Original transcriptions   Reconstructed terms   Regularizers

Essentially latent semantic indexing with L2 rather than an L1 norm

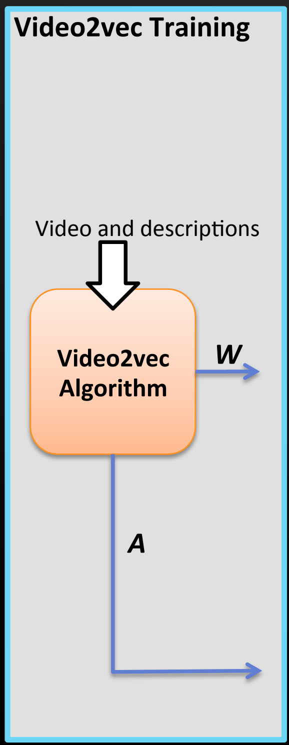
Video2vec objectives: predictability

Objective 2: The Video2vec embedding should be **predictable**

$$L_p(\boldsymbol{S}, \boldsymbol{W}) = \frac{1}{N}\sum_{i=1}^{N} \|\boldsymbol{s}_i - \boldsymbol{W}^\top \boldsymbol{x}_i\|_2^2 + \lambda_w \Theta(\boldsymbol{W})$$

Video2vec embedding   Video feature embedding   Regularizer

Video2vec: Training



Set of videos and their captions

Encode video features x_i

Any feature (combination) will do

Encode video descriptions y_i

Bag-of-words of terms

[Habibian MM 2014]

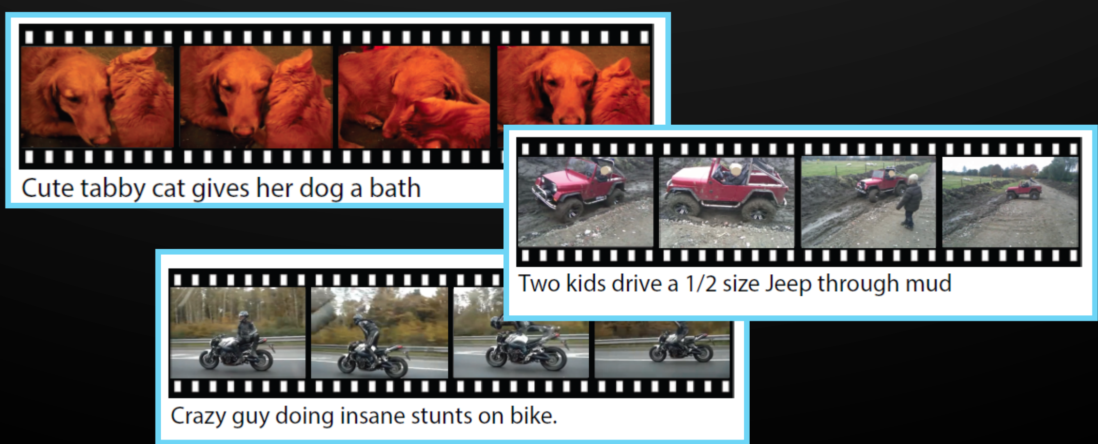
VideoStory46K dataset

Videos and title descriptions from YouTube

46K videos, 19K unique terms in descriptions

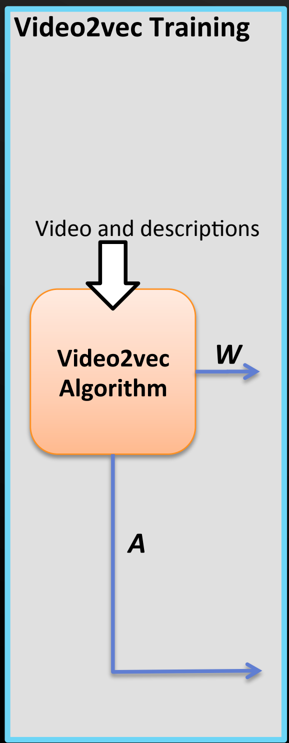
Seeded from video event descriptions

Filters to remove low quality videos



Available for download: www.mediamill.nl

Video2vec: Training (2)



Using *Stochastic Gradient Descent:*

Choose random sample

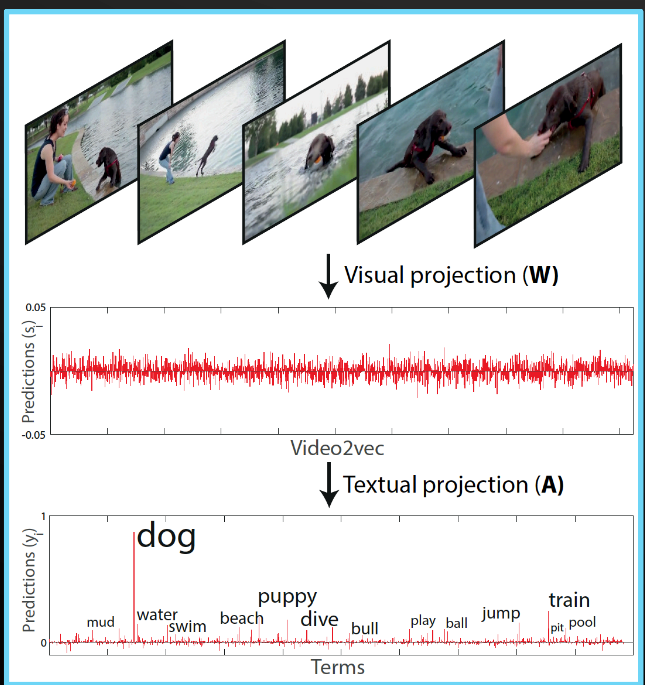
Compute sample gradient wrt objective

$$\nabla_{\boldsymbol{A}} L_{\mathrm{VS}} = -2\left(\boldsymbol{y}_t - \boldsymbol{A}\boldsymbol{s}_t\right)\boldsymbol{s}_t^\top + \lambda_a \boldsymbol{A},$$
$$\nabla_{\boldsymbol{W}} L_{\mathrm{VS}} = -2\,\boldsymbol{x}_t\left(\boldsymbol{s}_t - \boldsymbol{W}^\top \boldsymbol{x}_t\right)^\top + \lambda_w \boldsymbol{W},\ \text{and}$$
$$\nabla_{\boldsymbol{s}_t} L_{\mathrm{VS}} = 2\left[\boldsymbol{s}_t - \boldsymbol{W}^\top \boldsymbol{x}_t - \boldsymbol{A}^\top\left(\boldsymbol{y}_t - \boldsymbol{A}\boldsymbol{s}_t\right)\right] + \lambda_s \boldsymbol{s}_t.$$

Update parameters with step-size η

[Bottou ICCS 2010]

Video2vec at work



1. Project visual features

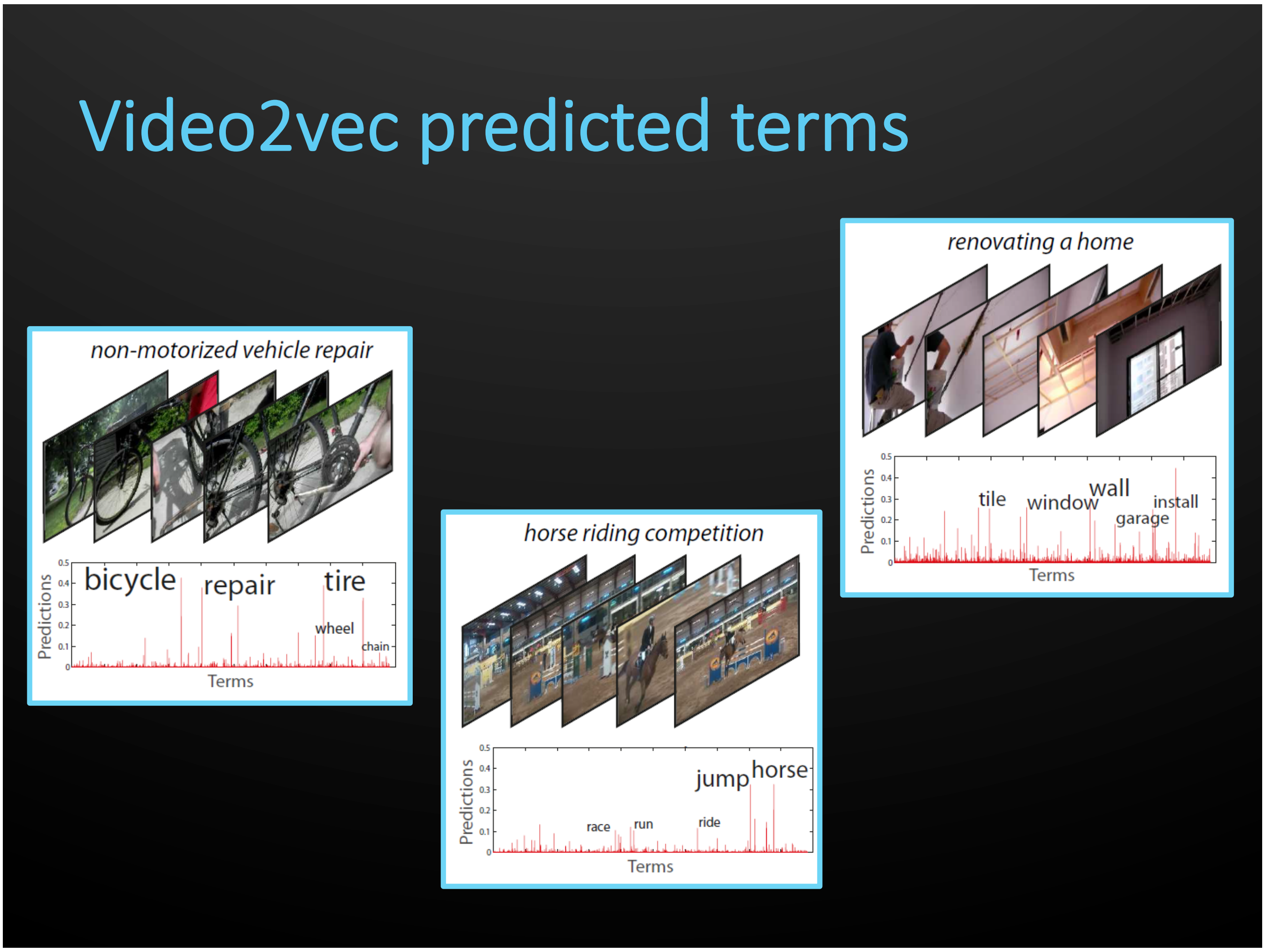
$$\boldsymbol{s}_i = \boldsymbol{W}^\top \boldsymbol{x}_i,$$

2. Translate to text

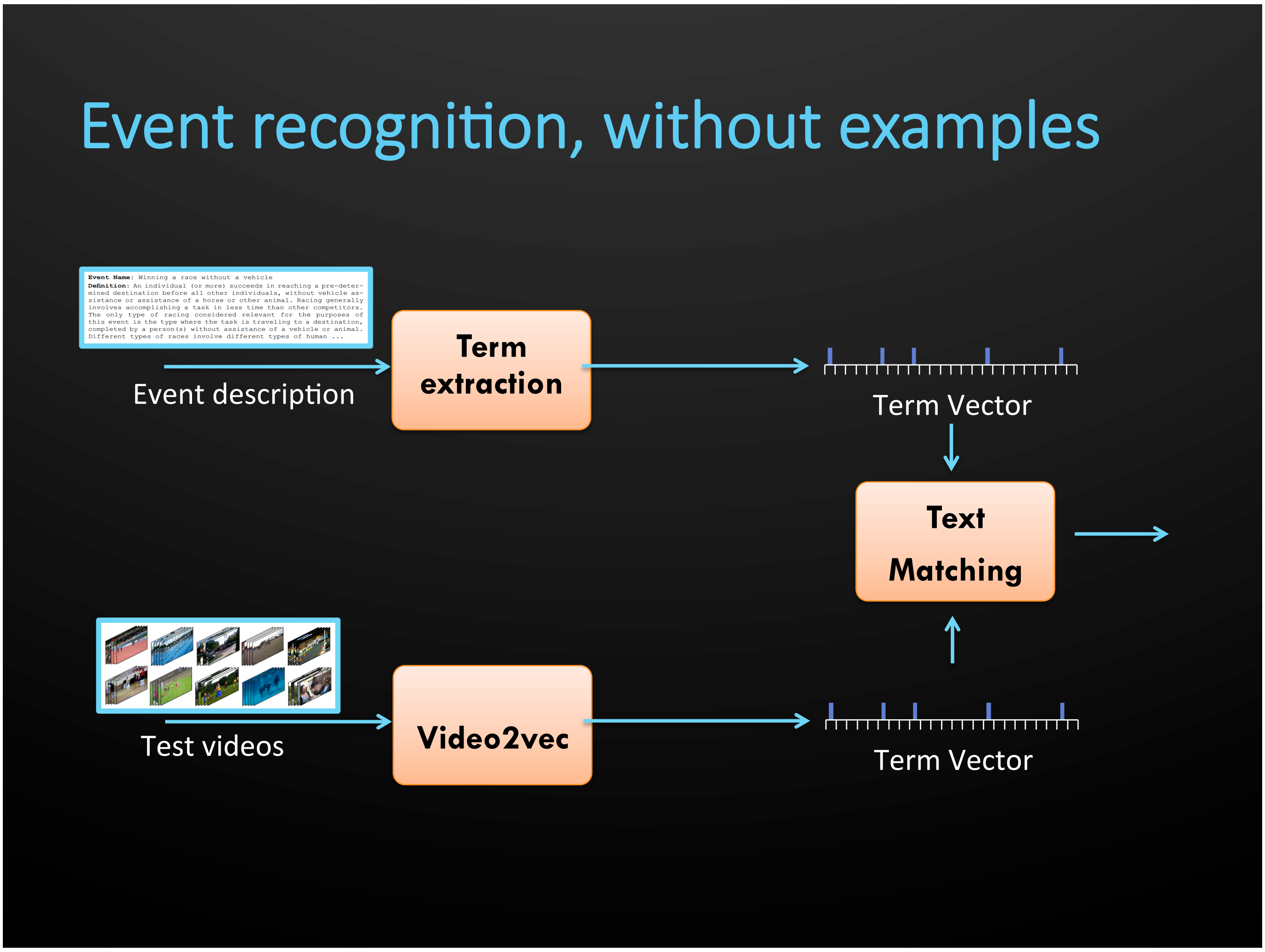
$$\hat{\boldsymbol{y}}_i = \boldsymbol{A}\boldsymbol{s}_i,$$

3. Cosine distance for matching

$$s_e(\boldsymbol{x}_i) = \frac{\boldsymbol{y}^{e\top}\hat{\boldsymbol{y}}_i^e}{||\boldsymbol{y}^e||\ \ ||\hat{\boldsymbol{y}}_i^e||}$$



Video2vec predicted terms
non-motorized vehicle repair
bicycle
repair
tire
wheel
chain
Predictions
Terms
horse riding competition
jump
horse
race
run
ride
Predictions
Terms
renovating a home
tile
window
wall
install
garage
Predictions
Terms



Event recognition, without examples
Event Name: Winning a race without a vehicle
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 description
Term extraction
Term Vector
Text Matching
Test videos
Video2vec
Term Vector

Zero-shot at TRECVID MED2013

Authors	Published	mAP
Habibian et al.	ICMR 2014	6.4
Ye et al.	MM 2015	9.0
Chang et al.	IJCAI 2015	9.6
Mazloom et al.	ICMR 2015	11.9
Wu et al.	CVPR 2014	12.7
Jiang et al.	AAAI 2015	12.9
Mazloom et al.	TMM 2016	12.9
Liang et al.	MM 2015	18.3
Habibian et al.	TPAMI 2017	20.0

Zero-shot at TRECVID MED2013

Authors	Published	mAP
Concept detectors	ICMR 2014	6.4
Ye et al.	MM 2015	9.0
Chang et al.	IJCAI 2015	9.6
Mazloom et al.	ICMR 2015	11.9
Concept discovery	CVPR 2014	12.7
Jiang et al.	AAAI 2015	12.9
Mazloom et al.	TMM 2016	12.9
Liang et al.	MM 2015	18.3
Video2vec	TPAMI 2017	20.0

Open challenges

More precise meaning with adjectives?

Searching video spatiotemporally?

How to handle live video streams?

Cappallo, BMVC'16

Zero-Shot Search for Live Video

Retrieval from live streaming video

Many live stream videos

Services like periscope, facebook, …

Environments like airports, elderly homes, …

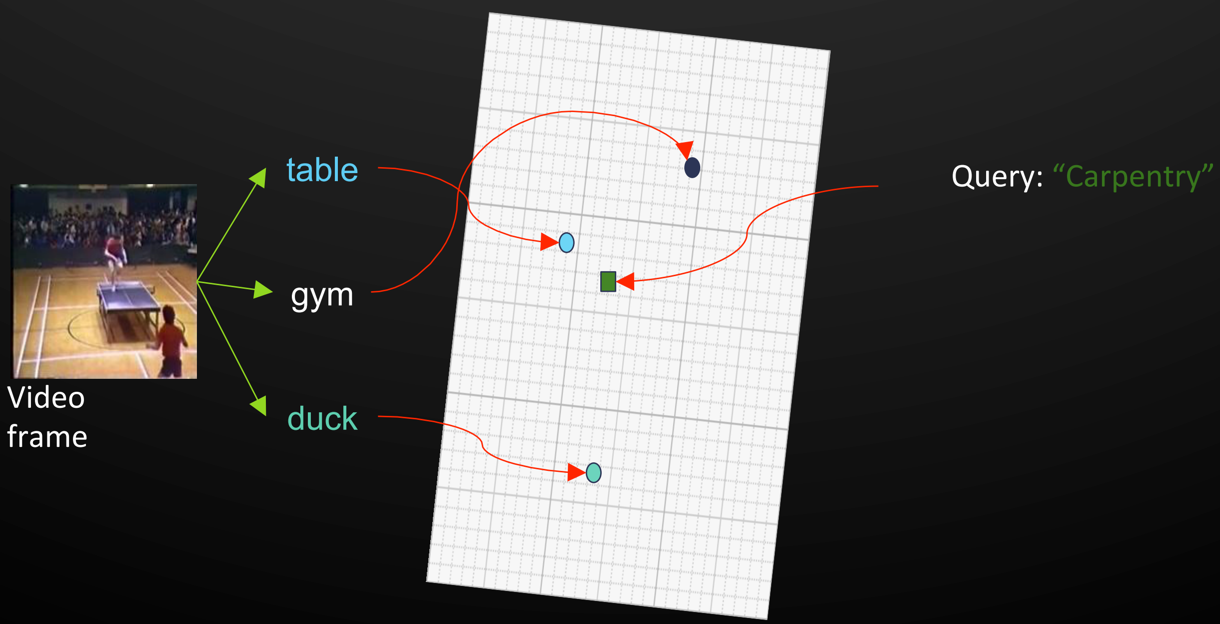
Live means

the future cannot be known

lack of extra metadata or context

Challenging, motivated zero-shot problem

Default: embed concepts per frame



Shared Semantic Embedding

Stream retrieval needs memory

Representation must reflect what is happening *now*

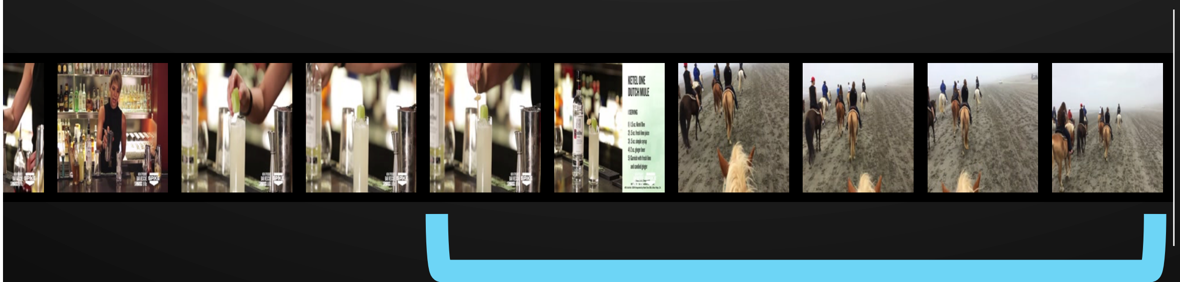
Also requires memory to prioritize recent information

Memory Pooling

Memory Welling

Mean and Max memory pooling

Now



Mean or Max Pooling over memory window

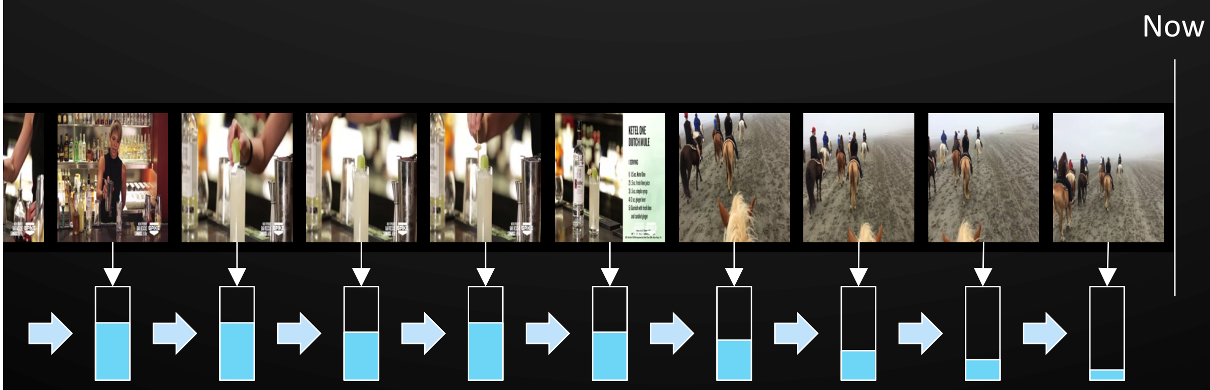
Two parameters:

m amount of memory

n amount of concepts

Memory welling

Instead of temporal pooling, well fills and drains over time...

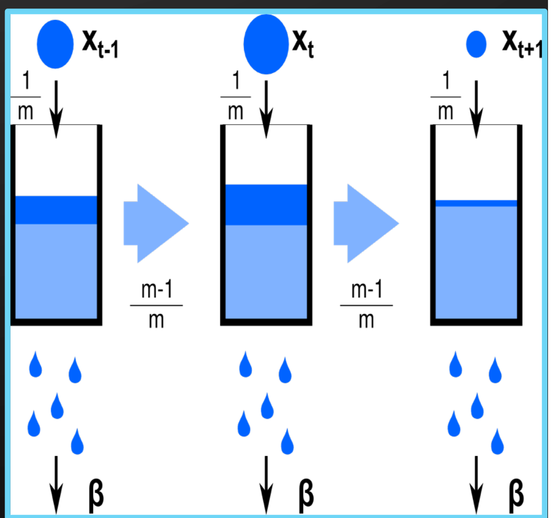


Memory welling

A well is defined by:

$$w(x_t) = \max\left(\frac{m-1}{m} w(x_{t-1}) + \frac{1}{m} x_t - \beta, \quad 0\right)$$

m is memory parameter
β is a constant "leakiness" term



Enforces sparsity
Ensures concept reliability

Welling emphasizes reliable, recent information

Comparing memories

Memory Pooling

- Only uses *m* frames of information

- *m* frames per feature per stream

- Arbitrary selection of top concepts

Memory Welling

+ No hard memory cut-off

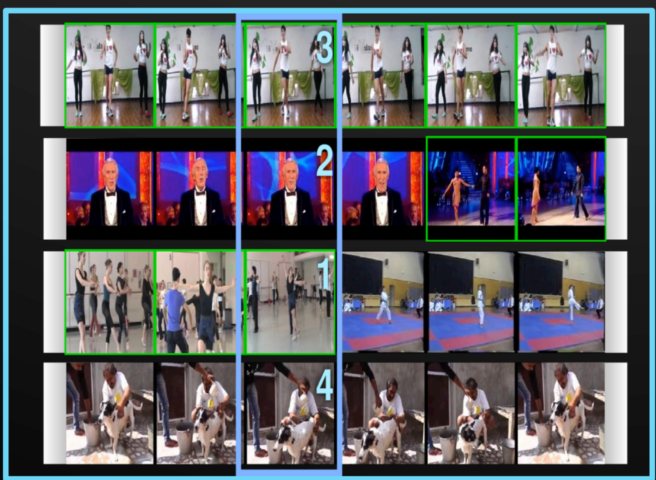
+ Only current state stored

+ Sparsity enforced implicitly

Memory welling addresses limitations of pooling, retains benefits

Live retrieval task 1: Instantaneous search

Which videos are relevant now?



Measure with mean AP across time: $$\frac{1}{\sum_t y^t} \sum_t \mathrm{AP}_t y^t$$

Live retrieval task 2: Continuous search

"Keep showing me relevant content"

e.g., watching dancing videos for thirty minutes

Live retrieval task 2: Continuous search

Reward relevant stream

Penalize needless switches

Temporal consistency

Evaluation metric: $\boxed{\dfrac{z_+ + r_+}{\sum_t y^t}}$

z_+ counts 'zaps' from irrelevant to relevant stream

r_+ rewards consistency on relevant stream

Conclusion

Zero-shot retrieval profits from semantic alignment

- Learnable from freely available online sources
- Better than low- and mid-level alternatives
- Adds meaning and recounting to retrieval results

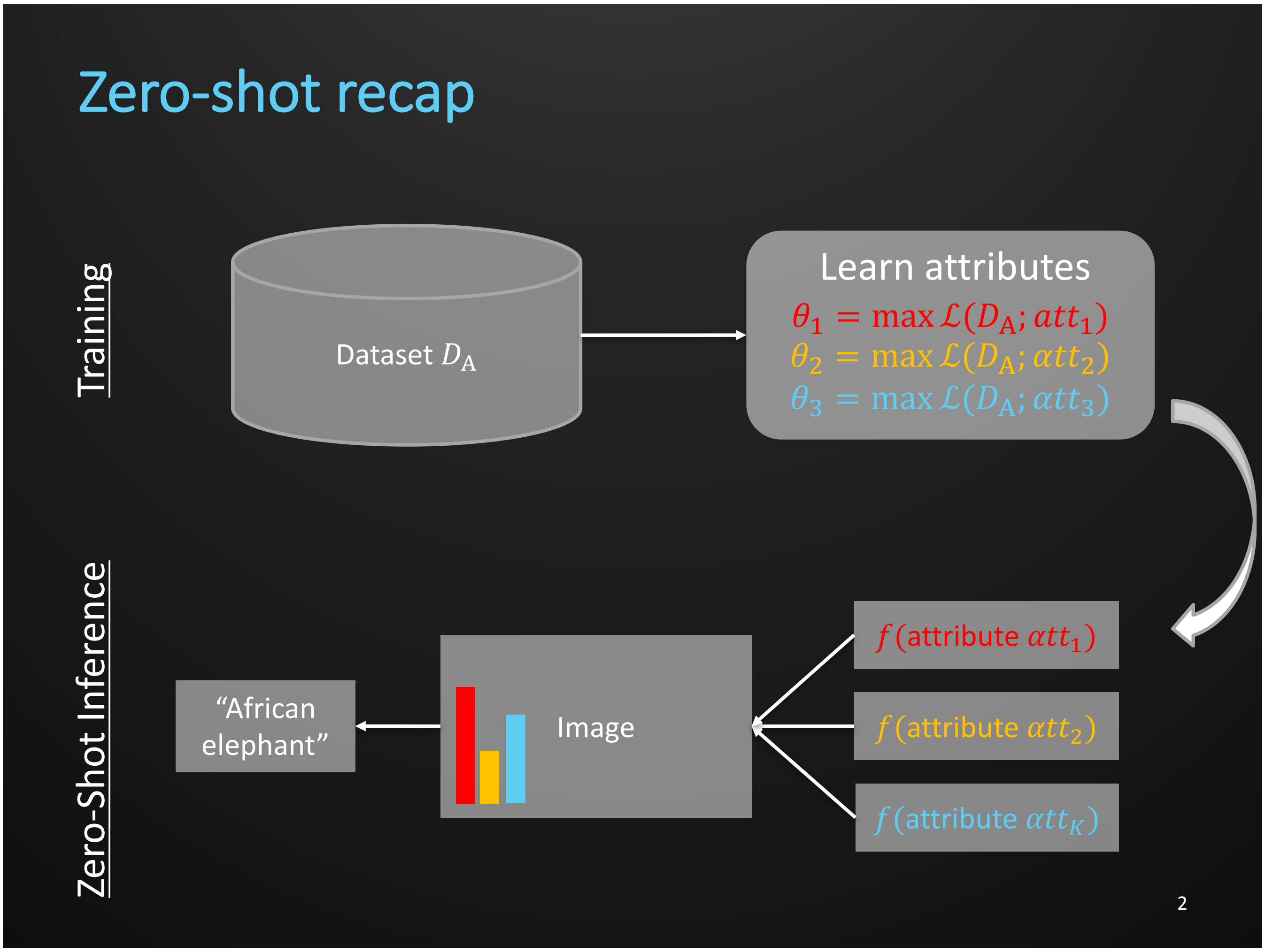
Next challenge:

- Spatiotemporal search and alerts for live video

Zero-Shot Learning

with Interaction

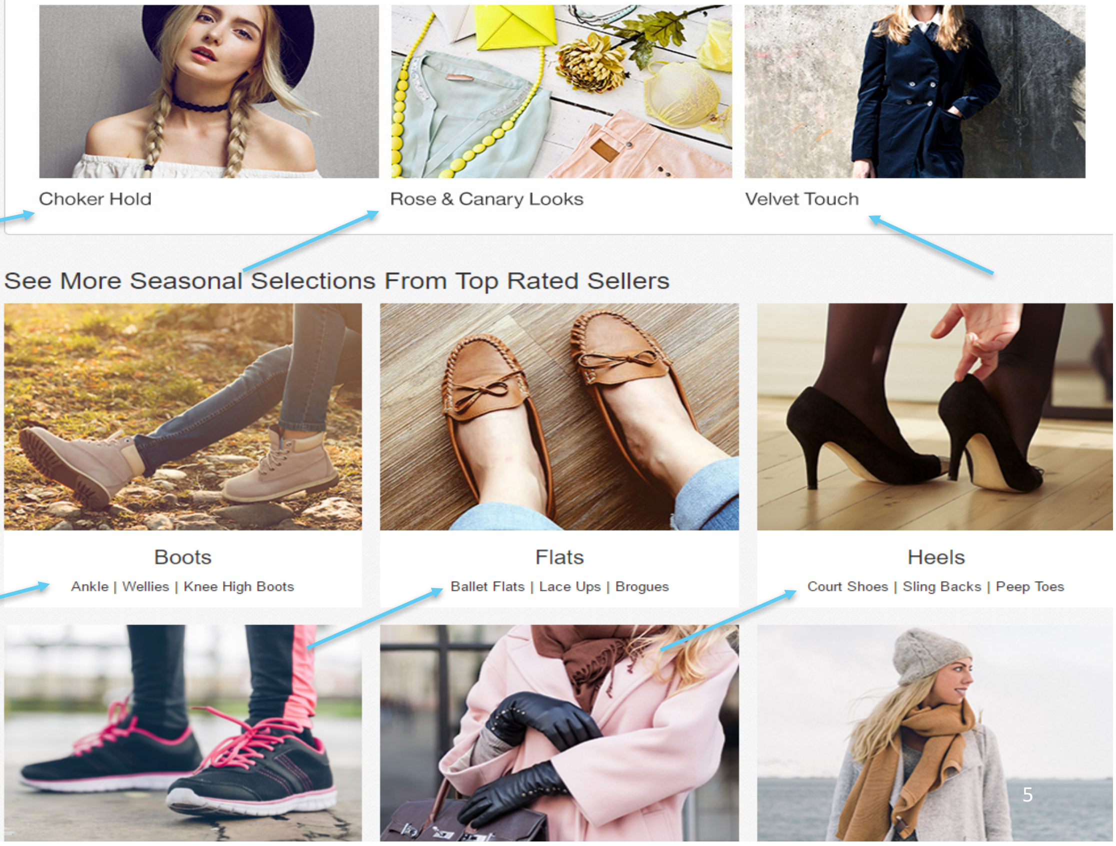
Efstratios Gavves


Zero-shot recap
Training
Dataset D_A
Learn attributes
$\theta_1 = \max \mathcal{L}(D_A; att_1)$
$\theta_2 = \max \mathcal{L}(D_A; att_2)$
$\theta_3 = \max \mathcal{L}(D_A; att_3)$
Zero-Shot Inference
f(attribute att_1)
f(attribute att_2)
f(attribute att_K)
Image
"African elephant"
2

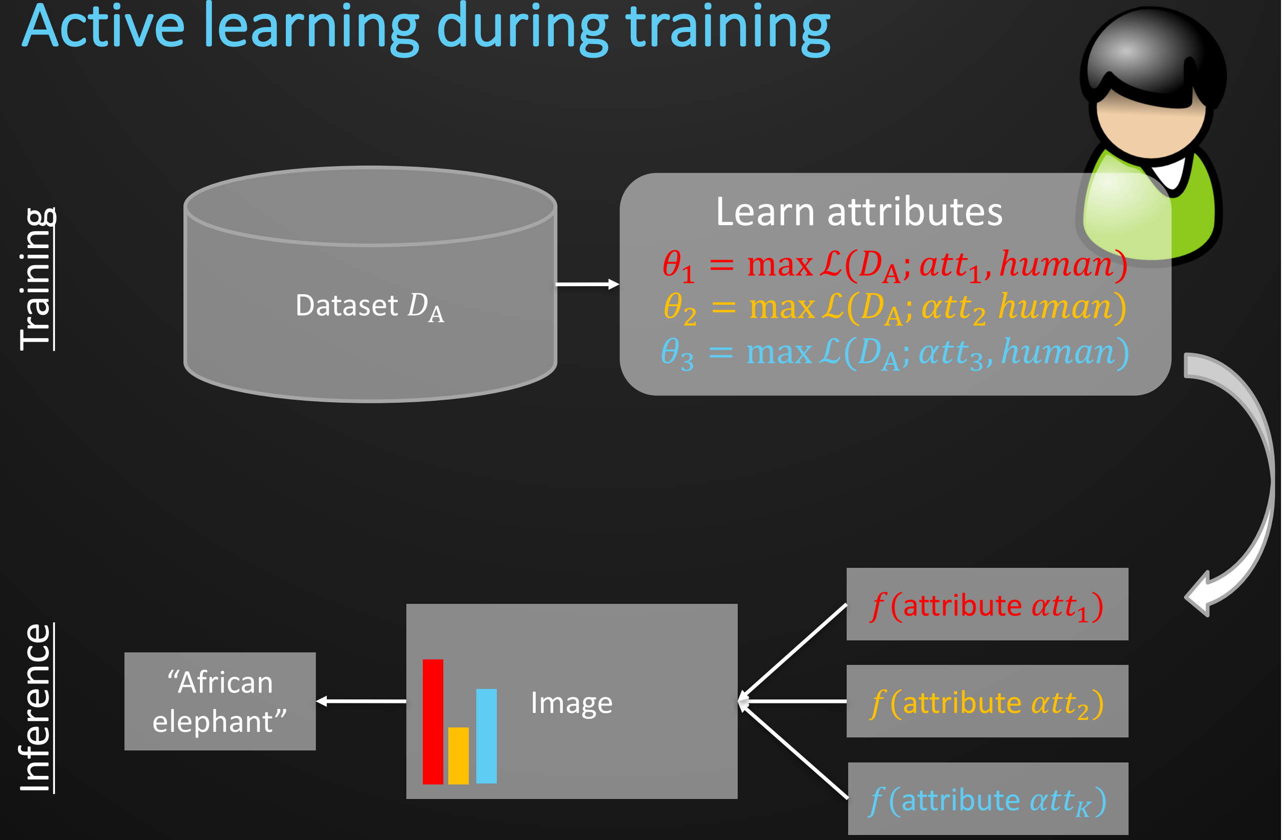

Examples of attributes for birds
"Curved bill"
"Long alula"
"Orange chest"
"Spiky nape"
3

Why Learn Attributes with Interaction?

Attributes are often ad-hoc



Active learning during training



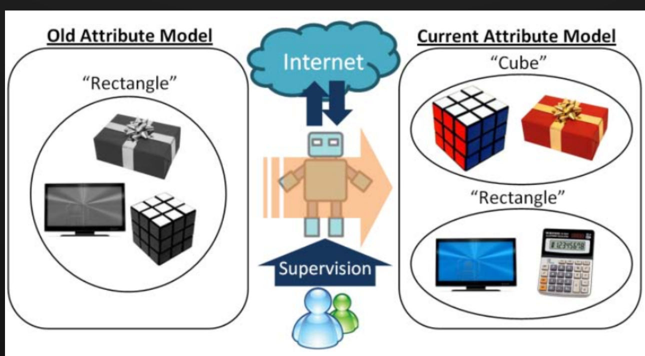
Incrementally learning attributes online

Zero-shot [1] with Independent Attribute Prediction [2]

Online Incremental Learning

- Self Organizing Incremental Neural Networks
- Parse images into positive/negative networks

Linear SVM for learning attribute classifiers



[1] *Online Incremental Attribute-based Zero-Shot Learning*, Kankuekul et al., CVPR 2012
[2] *Attribute-Based Classification for Zero-Shot Visual Object Categorization*, Lampert et al., TPAMI 2013

Interacting with local attributes

Discriminative localized attributes are discovered

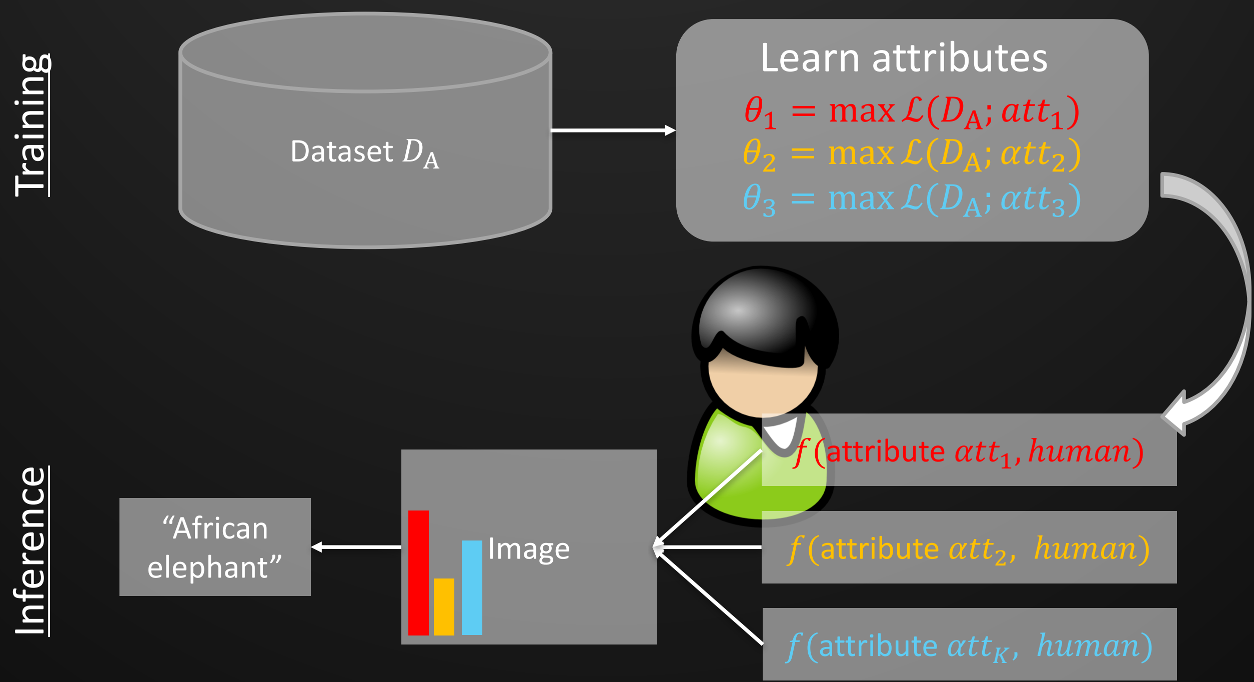
Most discriminative discovered feature shown to user

- If “nameable” → stored
- If not, got to next more discriminative feature

- Recommender system prioritization
  - spatially consistent features shown first

[1] Discovering Localized Attributes for Fine-Grained Recognition, Duan et al., CVPR 2012

Active learning during inference



Interacting with relative attributes

Learn relative attributes

- learning-to-rank

Interactive search

- Learn attributes offline
- At inference rank images according to relevance
- User indicates relative changes in top ranks

Active labelling

$$\text{minimize} \quad \left(\frac{1}{2}\|\boldsymbol{w}_{\boldsymbol{m}}^T\|_2^2 + C\left(\sum \xi_{ij}^2 + \sum \gamma_{ij}^2\right)\right)$$
$$s.t. \quad \boldsymbol{w}_{\boldsymbol{m}}^T(\boldsymbol{x_i} - \boldsymbol{x_j}) \geq 1 - \xi_{ij}; \forall (i,j) \in O_m$$
$$|\boldsymbol{w}_{\boldsymbol{m}}^T(\boldsymbol{x_i} - \boldsymbol{x_j})| \leq \gamma_{ij}; \forall (i,j) \in S_m$$
$$\xi_{ij} \geq 0; \gamma_{ij} \geq 0,$$

[1] Relative Attributes for Enhanced Man-Machine Communication, Parikh et al., AAAI 2012

Predicting unfamiliar classes

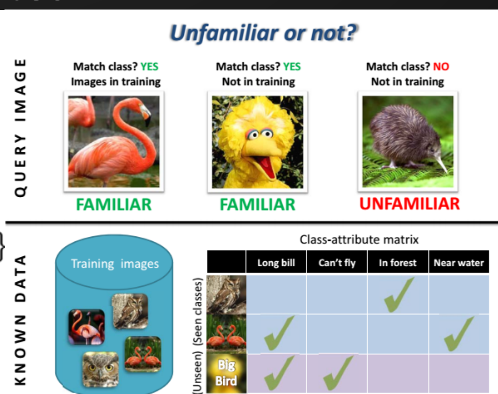
Open set of classes at test time

Slightly different than Zero-Shot

- no known attribute-class mapping
- p(unfamiliar class) = ∏(1 – p(seen class))

User corrects misclassified attributes

$$p(c|U,x) = \frac{p(U|\mathbf{a}^c, x)p(\mathbf{a}^c|x)}{p(U|x)}$$
$$p(U|\mathbf{a}^c, x) = \prod_{\tilde{a}_i \in U} p(\tilde{a}_i|a_i^c)^\gamma = \exp\{\sum_{\tilde{a}_i \in U} \gamma \log p(\tilde{a}_i|a_i^c)\}$$

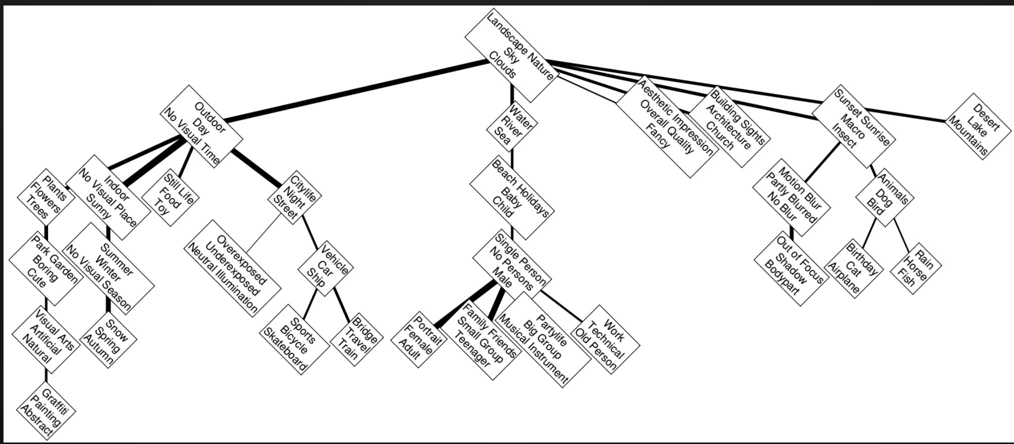


[1] Attribute-Based Detection of Unfamiliar Classes with Humans in the Loop, Wah et al., CVPR 2013

Tree-based Interactive Labelling

Image labels are correlated

- water, river, sea → landscape nature, sky, clouds
- Improved prediction: especially when human-in-the-loop
- Attribute-based image classification: attributes in tree



[1] *Learning Structured Prediction Models for Interactive Image Labelling,* Mensink et al., CVPR 2013

Tree-based Interactive Labelling

Criterion: select attribute that minimizes uncertainty on final class prediction

- select attribute that minimizes conditional class entropy
- new queries are conditioned on the image and the previously selected attributes

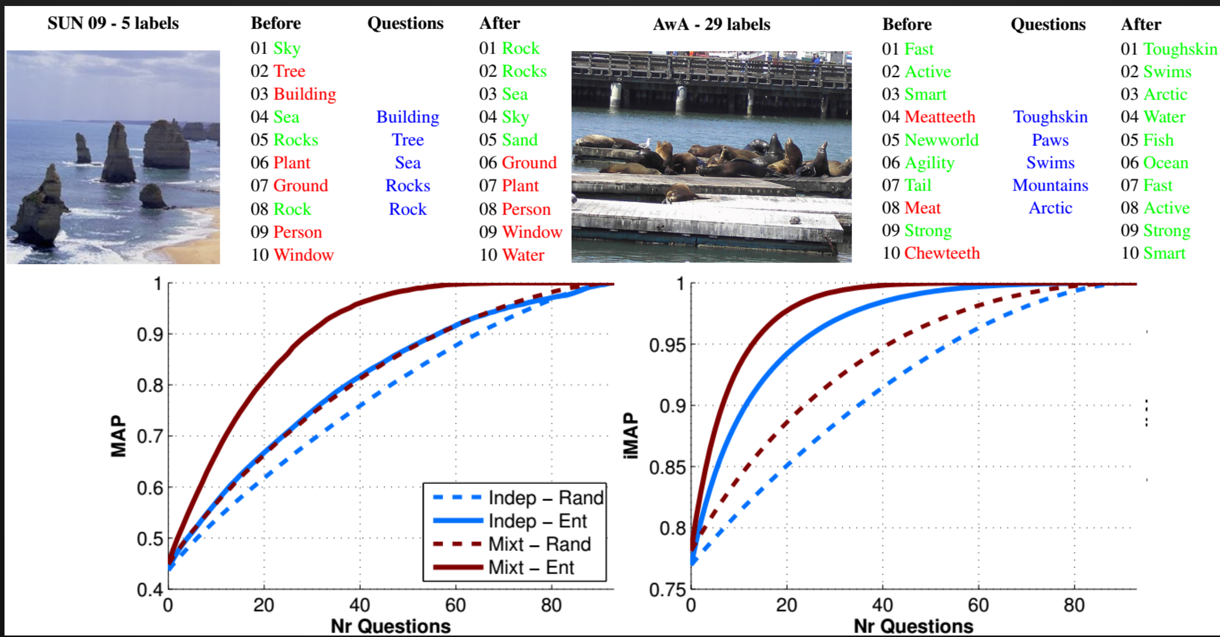
$$H(z, \boldsymbol{y}|\boldsymbol{x}) = H(y_i|\boldsymbol{x}) + H(z|y_i, \boldsymbol{x}) + H(\boldsymbol{y}_{\setminus i}|z, y_i, \boldsymbol{x})$$

$$p(z=c|\boldsymbol{x}) = \frac{p(\boldsymbol{y}_c|\boldsymbol{x})}{\sum_{c'=1}^{C} p(\boldsymbol{y}_{c'}|\boldsymbol{x})} = \frac{\exp -E(\boldsymbol{y}_c, \boldsymbol{x})}{\sum_{c'=1}^{C} \exp -E(\boldsymbol{y}_{c'}, \boldsymbol{x})}$$

$$E(\boldsymbol{y}, \boldsymbol{x}) = \sum_{i=1}^{L} \psi_i(y_i, \boldsymbol{x}) + \sum_{(i,j)\in\mathcal{E}} \psi_{ij}(y_i, y_j)$$

[1] *Learning Structured Prediction Models for Interactive Image Labelling,* Mensink et al., CVPR 2013

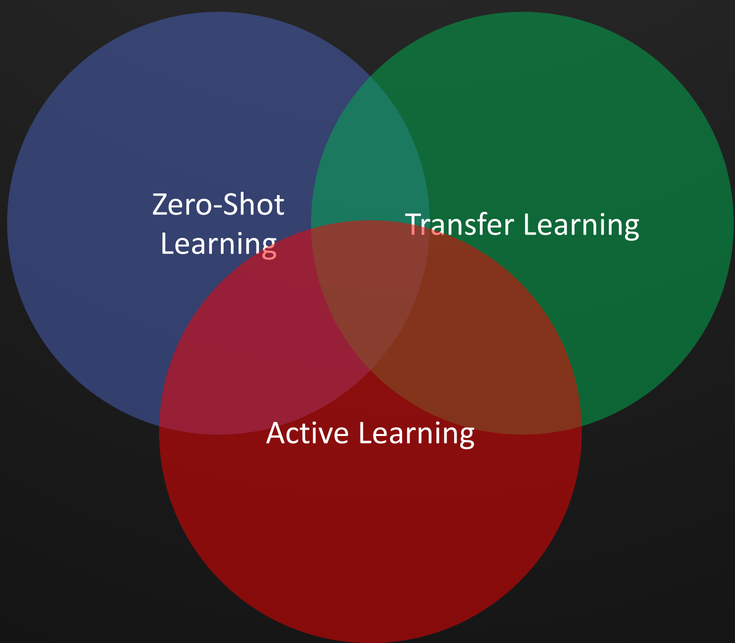
Tree-based Interactive Labelling



[1] Learning Structured Prediction Models for Interactive Image Labelling, Mensink et al., CVPR 2011

Zero-Shot, Transfer and Active Learning overlap!

First to identify & integrate three learning paradigms [1]



*[1] Active Transfer Learning with Zero-Shot Priors: Reusing Past Datasets for Future Tasks,*Gavves, et al., ICCV 2015

Reusing past (unrelated) datasets for future tasks

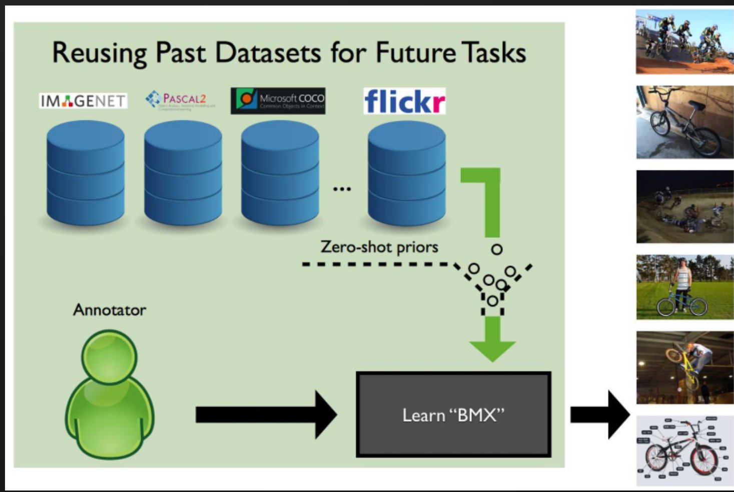
“Recycle” old datasets

ImageNet will not be obsolete in the future

- Open Images [2]

Enrich current datasets

- Segmentation propagation [3]



[1] *Active Transfer Learning with Zero-Shot Priors: Reusing Past Datasets for Future Tasks,* Gavves et al., ICCV 2015
[2] *https://github.com/openimages/dataset*
[3] *Segmentation Propagation in ImageNet,* Kuettel et al., ECCV 2012

How to transfer?

Known class model

- Old datasets
- Google

Class-Attribute mapping, e.g., COSTA [2]

New image

Zero-shot model

$$f^{zs}(\boldsymbol{x}) = \sum_{k \in \mathcal{K}} \beta_{ck}\, \boldsymbol{w}_k \cdot \boldsymbol{x}_i$$

Active updates

$$f^t(\boldsymbol{x}) = \eta^t f^{zs}(\boldsymbol{x}) + \boldsymbol{w}^t \cdot \boldsymbol{x}$$

[1] *Active Transfer Learning with Zero-Shot Priors: Reusing Past Datasets for Future Tasks,*Gavves, et al., ICCV 2015
[2] *COSTA: Co-Occurrence Statistics for Zero-Shot Classification,* Mensink, Gavves, Snoek, CVPR 2014

How to actively learn?

Simply speaking

- Sample from margin
- But make sure positive/ negatives labels balanced
- Keep running log of label sampling likelihoods

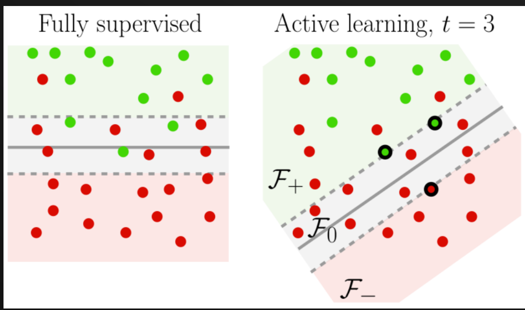
$$\max_{\alpha^t,\gamma^t} \sum_i \gamma_i^t \lambda_i^t \alpha_i^t - \frac{1}{2} \sum_{i,j} \alpha_i^t \alpha_j^t \gamma_i^t \gamma_j^t y_i y_j \boldsymbol{x}_i \cdot \boldsymbol{x}_j \quad (1)$$

$$\text{s.t.} \quad \sum \gamma_i^t \alpha_i^t y_i = 0 \quad (2)$$

$$0 \leq \alpha_i^t \leq C, \ \forall i\,, \quad (3)$$

$$\gamma_i^t \geq \gamma_i^{t-1}, \ \forall i\,, \quad (4)$$

$$\sum_i \gamma_i^t = \sum_i \gamma_i^{t-1} + B\ . \quad (5)$$

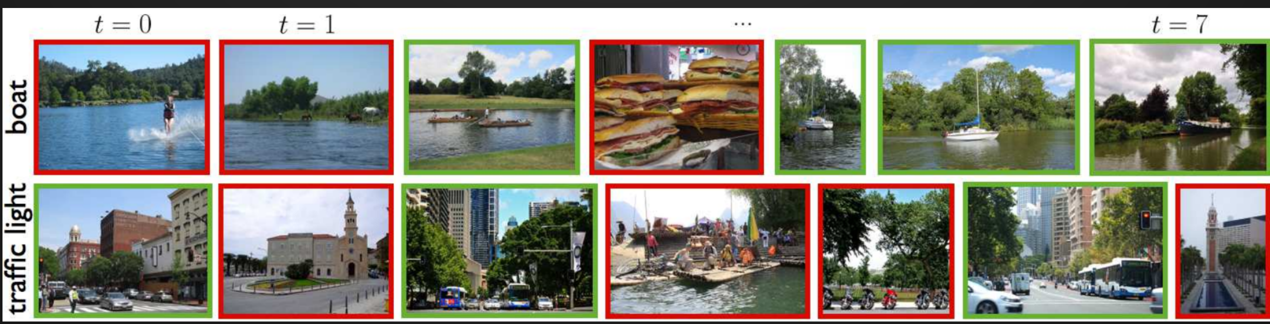


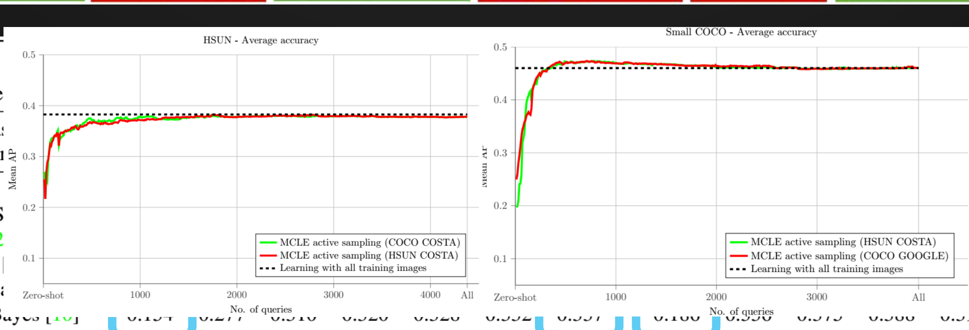
Proposition 1 (Maximum Conflict). *To maximize the objective Eq. (1) at time t, we should query the sample i^* such that* (a) *its label y_{i^*} has an opposite sign from its classification score at* $(t-1)$, *while* (b)) *the classifier score is as high as possible.*

Proposition 2 (Label Equality). *To respect the constraint Eq.* (2) *the number of positive and negative examples in the training set should be balanced,* i.e. $\sum_i \gamma_i^t [y_i = 1] = \sum_i \gamma_i^t [y_i = -1]$.

[1] Active Transfer Learning with Zero-Shot Priors: Reusing Past Datasets for Future Tasks, Gavves, et al., ICCV 2015

Active Transfer Learning with Zero-Shot Priors In Practice





No. of querie	250	300
	.460 mAP)	
MCLE (Data	0.448	0.460
MCLE (Exte	0.442	0.457
BBAL [33]	0.395	0.408
Hiearchical S	0.331	0.365
GP Mean [12	0.431	0.438
GP Variance	0.309	0.326
GP Impact B	0.430	0.436
GP EMOC Bayes	0.399	0.405

CODE https://github.com/stratisgavves/activetransferlearning or www.egavves.com

Going to the next level

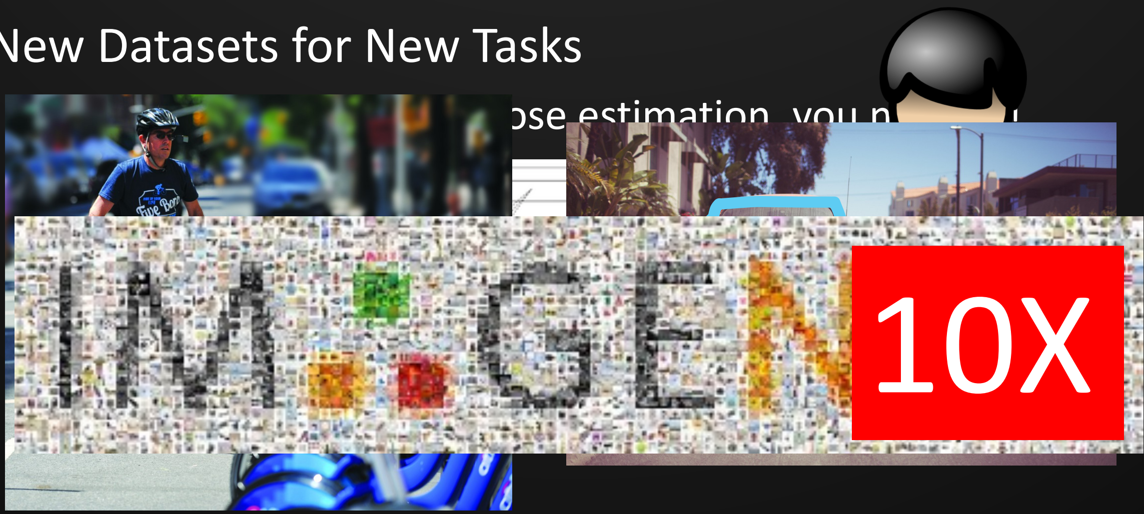
Active Deep Learning for Zero-Shot Recognition

- Deep learning of discriminative, repeatable attributes

Truly diversified transfer from past to future tasks

- Better transfer learning

New Datasets for New Tasks



Conclusion

Attributes not always perfect

- Often there is no good attribute definition for classes
- Often attribute prediction is not that reliable

Interaction remedy to attribute-based classification

- Correct prediction mistakes
- Guide new attribute learning
- Guide classification

Active Transfer Learning

- Don’t waste or throw your old datasets!!
- Much faster active learning than state-of-the-art alternatives

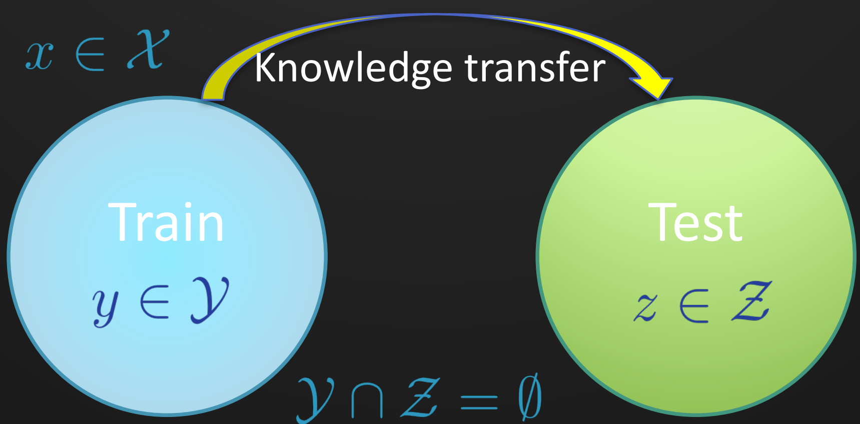
Zero-Shot Learning

for Vision and Multimedia

Conclusion & Discussion

What this tutorial was about?

Data: $x \in \mathcal{X}$



Objective: $f : \mathcal{X} \to \mathcal{Z}$

Lampert et al., CVPR09/PAMI13

Today's outline

1. Knowledge transfer
2. Classification
3. Localization

   Break

4. Retrieval
5. Interaction
6. Conclusion and Discussion

Zero-Shot Classification

Mathematically ALE and DAP are similar

ALE directly optimizes image classification

Focus on visual Details or Regions

- Each with their merit, depends on application
- Maybe a smart combination?

Zero-Shot using pre-trained classifiers

- Indirect attribute prediction
- Co-occurrence statistics
- Word2vec

Zero-Shot with Localization

Attributes belong to objects, not images

Zero-Shot localization natural extension

Focus on visual Details or Regions

- Each with their merit, depends on application
- Maybe a smart combination?

Zero-Shot with Interaction

Attributes not always perfect

- Often there is no good attribute definition for classes
- Often attribute prediction is not that reliable

Interaction remedy to attribute-based classification

- Correct prediction mistakes
- Guide new attribute learning
- Guide classification

Active Transfer Learning → Old datasets no more wasted

- Much faster learning than state-of-the-art alternatives

Zero-Shot Retrieval

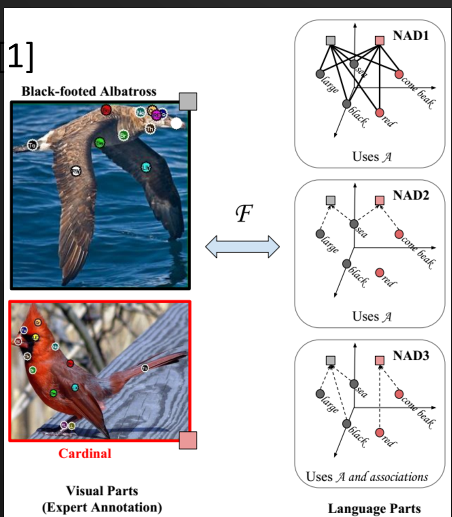
Zero-shot retrieval profits from semantic alignment

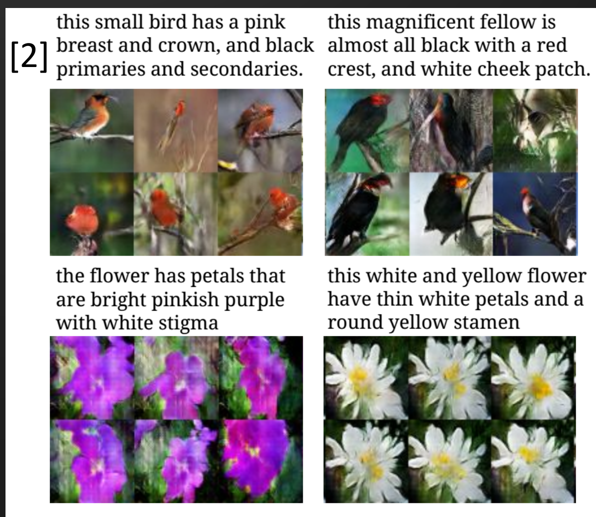
- Learnable from freely available online sources
- Better than low- and mid-level alternatives
- Adds meaning and recounting to retrieval results

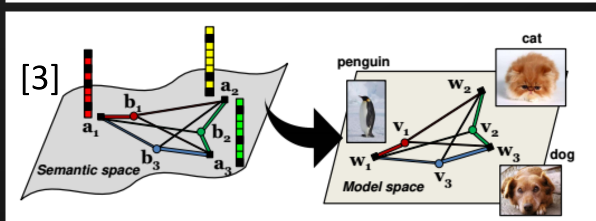
Next challenge:

- Spatiotemporal search and alerts for live video

What's next?







[1] *Multi-Cue Zero-Shot Learning with Strong Supervision*, Akata et al., CVPR 2016
[2] *Generative Adversarial Text to Image Synthesis*, Reed, ICML 2016
[3] *Synthesized Classifiers for Zero-Shot Learning*, Changpinyo, CVPR 2016

Thank you for your attention!

(Slides will be added online later today)

References

References

Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid, Label-Embedding for Image Classification. IEEE TPAMI, 2015

Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid. Label embedding for attribute-based classification. In CVPR, 2013.

Z. Akata, S. Reed, D. Walter, H. Lee and B. Schiele, Evaluation of Output Embeddings for Fine-Grained Image Classification. In CVPR, 2015

A, Bendale and T. Boult, Towards Open World Recognition. In CVPR, 2015.

S. Cappallo, T. Mensink, and C.G.M. Snoek. Image2Emoji: Zero-shot Emoji Prediction for Visual Media. In MM, 2015.

S. Cappallo, T. Mensink, and C.G.M. Snoek. Video Stream Retrieval of Unseen Queries using Semantic Memory. In BMVC, 2016.

J. Chen, Y. Cui, G. Ye, D. Liu, and S.-F. Chang. Event-driven semantic concept discovery by exploiting weakly tagged internet images. In ICMR, 2014.

J. Costa Pereira, E. Coviello, G. Doyle, N. Rasiwasia, G. R. Lanckriet, R. Levy, and N. Vasconcelos. On the role of correlation and abstraction in cross-modal multimedia retrieval. IEEE TPAMI, 2014.

K. Duan, D. Parikh, D. Crandall, K. Grauman, Discovering Localized Attributes for Fine-Grained Recognition. In CVPR 2012

A. Farhadi and M. Sadeghi, Recognition using Visual Phrases. In CVPR, 2011.

F. Feng, X. Wang, R. Li. Cross-modal Retrieval with Correspondence Autoencoder. In MM, 2014.

References

Y. Fu, T. Hospedales, T. Xiang and S. Gong, Transductive Multi-view Zero-Shot Learning. IEEE TPAMI, 2015

E. Gavves, T. Mensink, T. Tommasi, C.G.M. Snoek, and T. Tuytelaars. Active Transfer Learning with Zero-Shot Priors: Reusing Past Datasets for Future Tasks. In ICCV, 2015.

A. Habibian, T. Mensink, and C.G.M. Snoek. Composite Concept Discovery for Zero-Shot Video Event Detection. In ICMR, 2014.

A. Habibian, T. Mensink, and C.G.M. Snoek. Video2vec Embeddings Recognize Events when Examples are Scarce. IEEE TPAMI, 2017.

A. Habibian, T. Mensink, and C.G.M. Snoek. Videostory: A new multimedia embedding for few-example recognition and translation of events. In MM, 2014.

R. Hu, H. Xu, M. Rohrback, J. Feng, K. Saenko, T. Darrell, Natural Language Object Retrieval. In CVPR 2016

M. Jain, J. C. van Gemert, T. Mensink, and C.G.M. Snoek. Objects2action: Classifying and localizing actions without any video example. In ICCV, 2015.

P. Kankuekul, A. Kawewong, S. Tangruamsub, O. Hasegawa. Online Incremental Attribute-based Zero-Shot Learning. In CVPR 2012

S. Kordumova, T. Mensink, and C.G.M. Snoek. Pooling Objects for Recognizing Scenes without Examples. In ICMR, 2016.

C. Lampert, H. Nickisch, and S. Harmeling, Attribute-Based Classification for Zero-Shot Learning of Object Categories. IEEE TPAMI, 2013.

References

C. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In CVPR, 2009.

X. Li, C. Snoek, M. Worring, and A. Smeulders, Harvesting Social Images for Bi-Concept Search. IEEE TMM, 2012.

Z. Li, E. Gavves, K.E.A. van de Sande, C.G.M. Snoek, A.W.M. Smeulders, Codemaps segment, classify and search objects locally. In ICCV, 2013

Z. Li, E. Gavves, T. Mensink, and C.G.M. Snoek. Attributes Make Sense on Segmented Objects. In ECCV, 2014.

K. Matzen, N. Snavely, BubbLeNet: Foveated Imaging for Visual Discovery. In ICCV 2015

T. Mensink, E. Gavves, and C.G.M. Snoek. COSTA: Co-Occurrence Statistics for Zero-Shot Classification. In CVPR, 2014.

T. Mensink, J. Verbeek, F. Perronnin, and G. Csurka, Distance-Based Image Classification: Generalizing to new classes at near-zero cost. IEEE TPAMI, 2013.

T. Mensink, J. Verbeek, F. Perronnin, and G. Csurka, Metric Learning for Large Scale Image Classification: Generalizing to New Classes at Near-Zero Cost. In ECCV, 2012.

T. Mensink, J. Verbeek, and G. Csurka, Learning structured prediction models for interactive image labeling. In CVPR, 2011.

T. Mensink, J. Verbeek, and G. Csurka. Tree-structured CRF Models for Interactive Image Labeling. IEEE TPAMI, 2012.

References

T. Mikolov, K. Chen, G. Corrado, and J. Dean, Efficient Estimation of Word Representations in Vector Space. In ICLR, 2013.

M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. Corrado, and J. Dean, Zero-Shot Learning by Convex Combination of Semantic Embeddings. In ICLR, 2014.

D. Parikh and K. Grauman. Relative attributes. In ICCV, 2011.

D. Parikh, A. Kovashka, A. Parkash, K. Grauman, Relative Attributes for Enhanced Man-Machine Communication. In AAAI 2012

N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G.R.G. Lanckriet, R. Levy, N. Vasconcelos. A New Approach to Cross-Modal Multimedia Retrieval. In MM, 2014.

M. Rohrbach, M. Stark, G. Szarvas, I. Gurevych, and B. Schiele, What Helps Where — And Why? Semantic Relatedness for Knowledge Transfer. In CVPR, 2010.

O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpa- thy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, ImageNet Large Scale Visual Recognition Challenge. IJCV, 2015

C. Wah, S. Belongie, Attribute-Based Detection of Unfamiliar Classes with Humans in the Loop. In CVPR 2013

C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, The Caltech-UCSD Birds-200-2011 Dataset, 2011.

J. Weston, S. Bengio, and N. Usunier, WSABIE: Scaling Up To Large Vocabulary Image Annotation. In IJCAI, 2011.

S. Wu, S. Bondugula, F. Luisier, X. Zhuang, and P. Natarajan. Zero-shot event detection using multi-modal fusion of weakly supervised concepts. In CVPR, 2014.

Y. Xian, Z. Akata, G. Sharma, Q. Nguyen, M. Hein and B. Schiele, Latent Embeddings for Zero-shot Classification. In CVPR, 2016

Lectures

Lectures

Thomas Mensink

Thomas.mensink@uva.nl

Efstratios (Stratis) Gavves

e.gavves@uva.nl

Cees Snoek

c.g.m.snoek@uva.nl