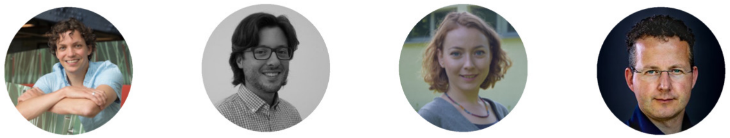
Zero-Shot Learning
for Computer Vision

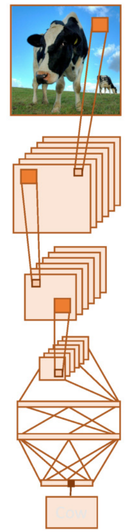


Thomas Mensink, Efstratios Gavves, Zeynep Akata, Cees Snoek
University of Amsterdam

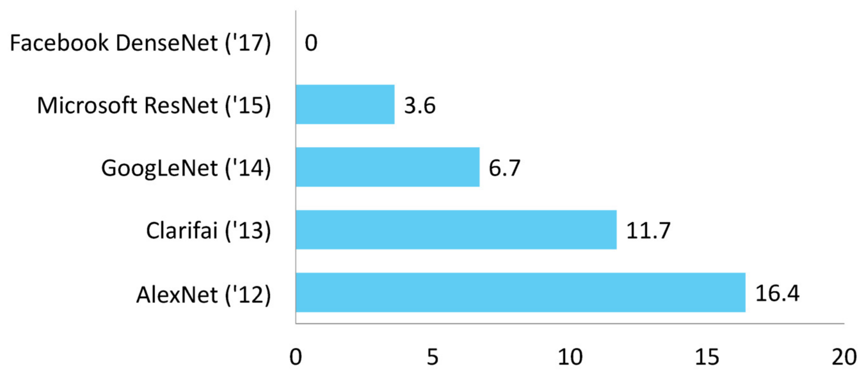
Many-shot learning

+
Annotations

=

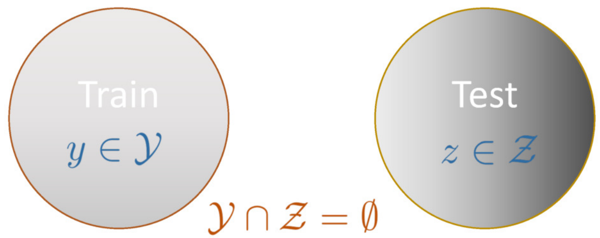


Most popular plot in computer vision



What is zero-shot learning?

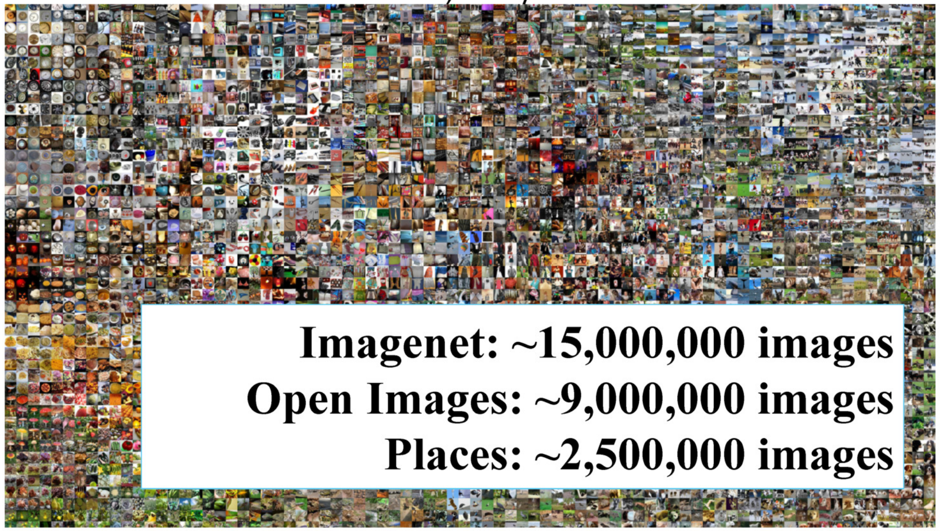
Data: $x \in \mathcal{X}$



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

Lampert et al., CVPR09/PAMI13

We have labeled data, why bother?

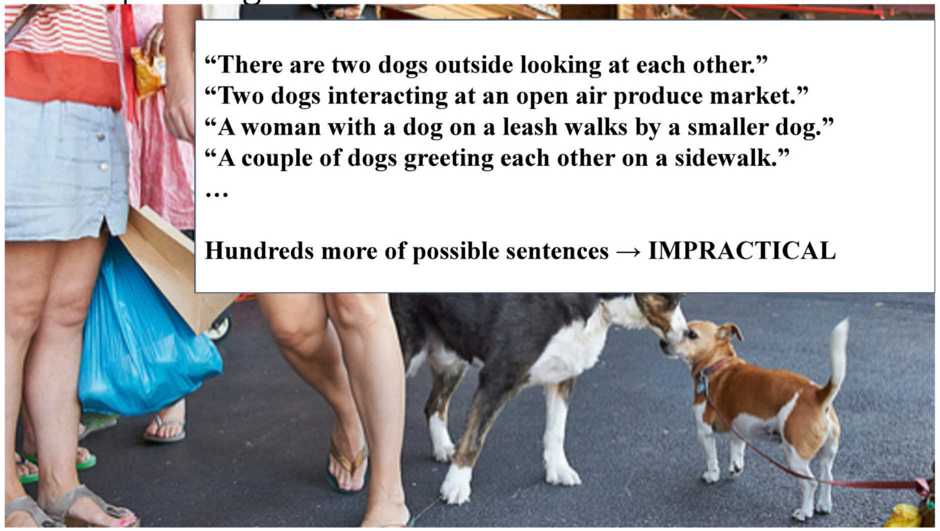


An image

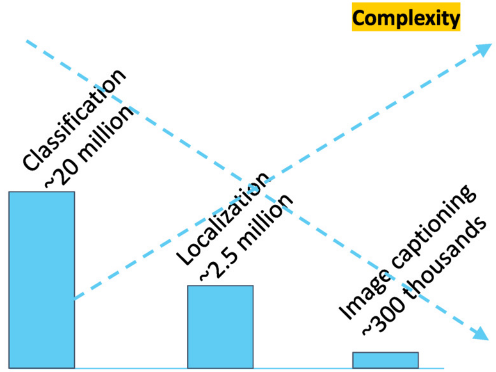
Classification

Segmentation

Captioning



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"

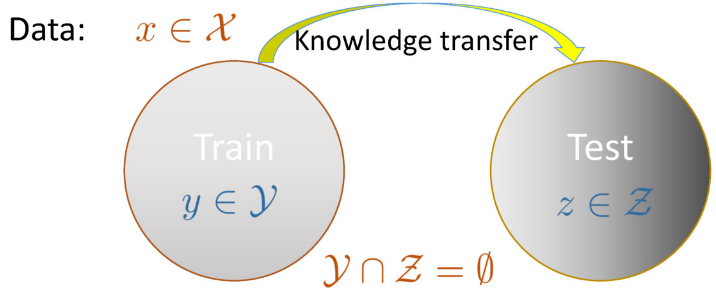
Why zero-shot learning?

- Privacy-sensitive recognition problems

Why zero-shot learning?

- When learning and inference need to be efficient

What is this tutorial about?



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

Lampert et al., CVPR09/PAMI13

TUTORIAL PROGRAM

- 13:30-13:40 | **Introduction** | Efstratios Gavves
- 13:40-14:30 | **Classification** | Zeynep Akata
- 14:30-15:00 | **Localization** | Efstratios Gavves
- 15:00-15:30 | **Retrieval** | Cees G.M. Snoek
- 15:30-16:00 | **Break**
- 16:00-16:40 | **Open problems** | Zeynep Akata, Efstratios Gavves
- 16:40-17:00 | **Conclusion** | Efstratios Gavves

Universiteit van Amsterdam

max planck institut informatik

Zero-Shot Learning for Image Classification

Zeynep Akata

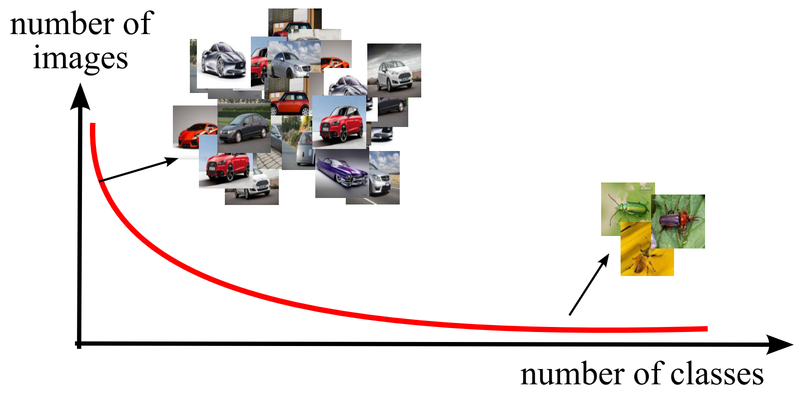
Zero-Shot Learning Tutorial, CVPR 2017

26 July 2017

Outline

Outline

Data Distribution in Large-Scale Datasets



Attributes as Side-Information

images | **attributes** | **class**

black-white
has tail
lives on land
small

zebra

gray
has tail
lives in water
big

whale

[Lampert et.al. CVPR'09, Ferrari et.al. CVPR'09]

Attributes as Side-Information

images | **attributes** | **class**

black-white
has tail
lives on land
small

zebra
[1 0 1 1 0 1]

gray
has tail
lives in water
big

[0 1 1 0 1 0]
whale

[Lampert et.al. CVPR'09, Ferrari et.al. CVPR'09]

Attributes as Side-Information

images | **attributes** | **class**

black-white
has tail
lives on land
small

zebra
[1 0 1 1 0 1]

gray
has tail
lives in water
big

[0 1 1 0 1 0]
whale

[Lampert et.al. CVPR'09, Ferrari et.al. CVPR'09]

Attributes as Side-Information

images | **attributes** | **class**

black-white
has tail
lives on land
small

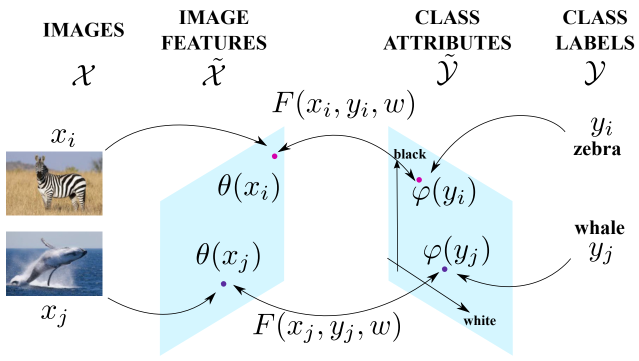
zebra
[1 0 1 1 0 1]

gray
has tail
lives in water
big

[0 1 1 0 1 0]
whale

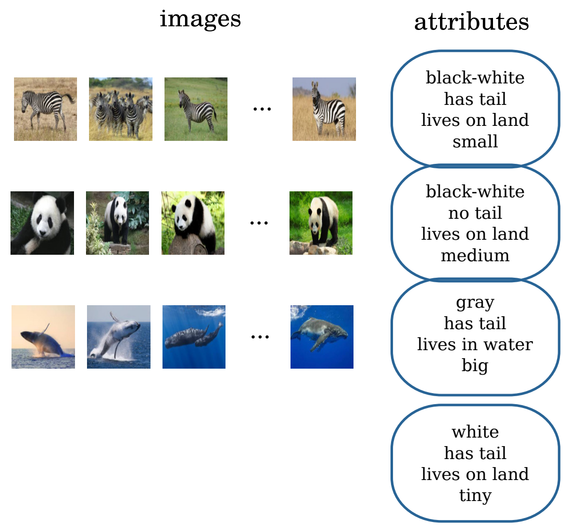
[Lampert et.al. CVPR'09, Ferrari et.al. CVPR'09]

Muldimodal Embeddings for Zero-Shot Learning



[Akata et.al. CVPR'13, CVPR'15, CVPR'16 & TPAMI'16]

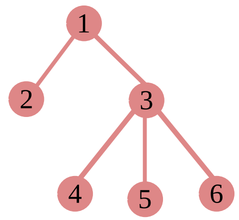
Zero-Shot Learning



Wikipedia and WordNet as Side Information

Wikipedia and Wordnet: object descriptions or hierarchies

Word2Vec [Mikolov et.al. NIPS'13]
GloVe [Pennington et.al EMNLP'14]

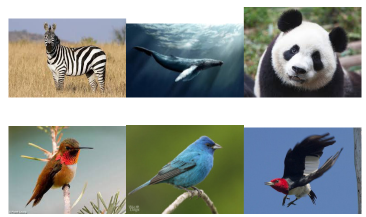


$$2 = \quad [1\ 0\ 2\ 3\ 3\ 3]$$

Hierarchical similarity measures

Experimental Setting

Animals with Attributes (AWA) [Lampert et.al. CVPR'09]	50 cls	85 att
Caltech UCSD-Birds (CUB) [Wah et.al.'11]	200 cls	312 att



Input Embeddings $\theta(x)$: 1K-dim GoogLeNet features

Output Embeddings $\varphi(y)$: `att`, `w2v`, `glo`, `hie`

Evaluation of Class Embeddings

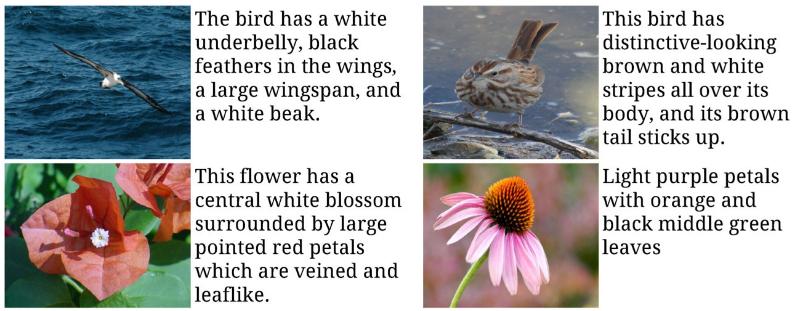
	AWA	CUB
w2v	51.2	28.4
glo	58.8	24.2
hie	51.2	20.6
att-	**60.1**	**29.9**
att+	**73.9**	**51.7**

Evaluation of Class Embeddings

	AWA	CUB
w2v	51.2	28.4
glo	58.8	24.2
hie	51.2	20.6
att-	**60.1**	**29.9**
att+	**73.9**	**51.7**

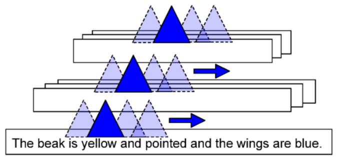
- Attributes & Wikipedia & WordNet are complementary

Detailed Visual Descriptions as Side Information



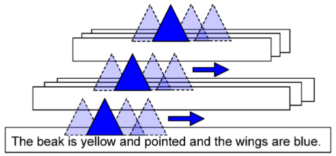
[Reed et.al. CVPR'16, ICML'16, NIPS'16]

Deep Representations of Visual Descriptions

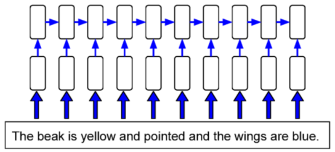


[Zhang and Lecun NIPS'15]

Deep Representations of Visual Descriptions

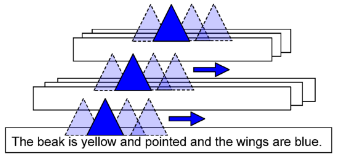


[Zhang and Lecun NIPS'15]

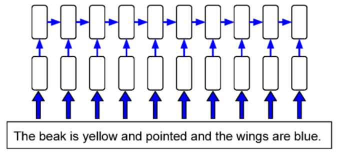


[Hochreiter and Schmidhüber'98]

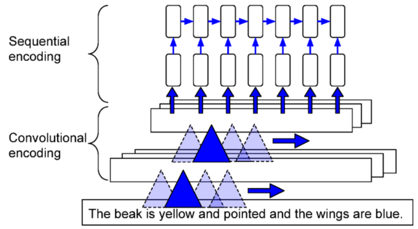
Deep Representations of Visual Descriptions



[Zhang and Lecun NIPS'15]

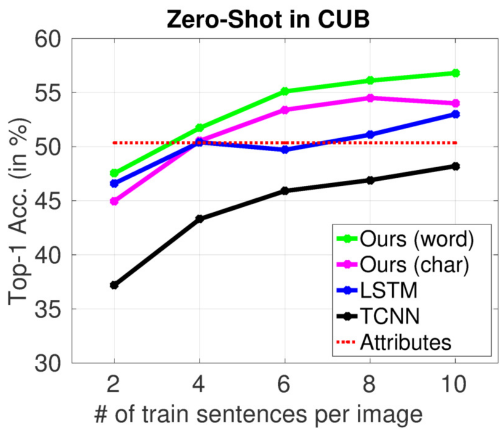


[Hochreiter and Schmidhüber'98]

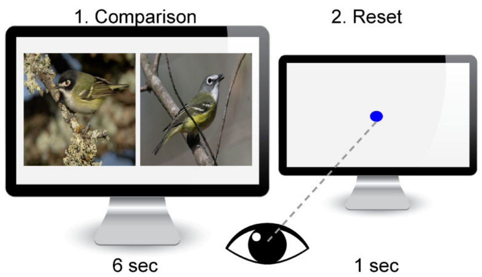


[Reed et.al., CVPR'16]

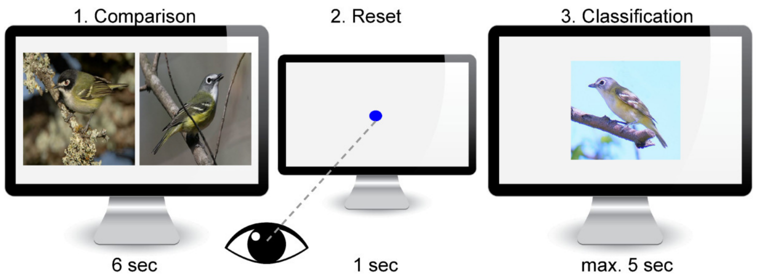
Deep Representations vs Attributes



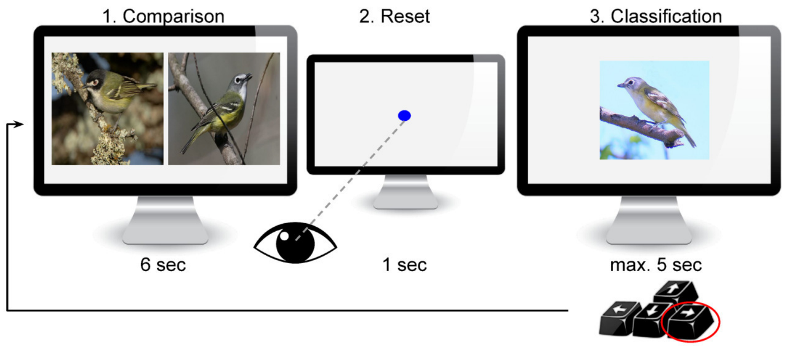
Human Gaze as Side-Information



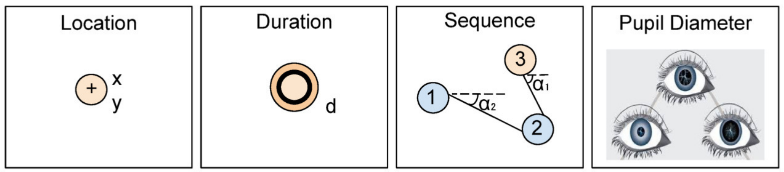
Human Gaze as Side-Information



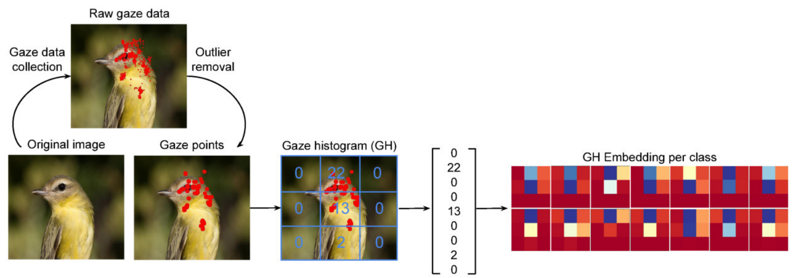
Human Gaze as Side-Information



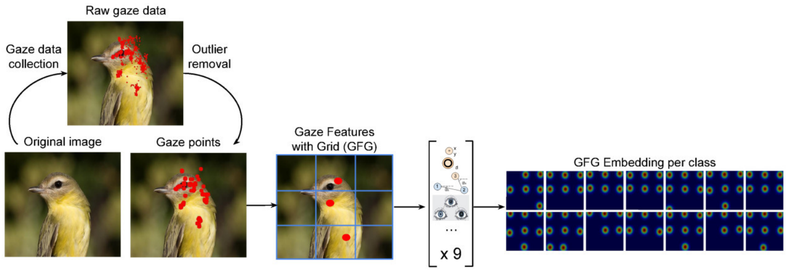
Gaze Features



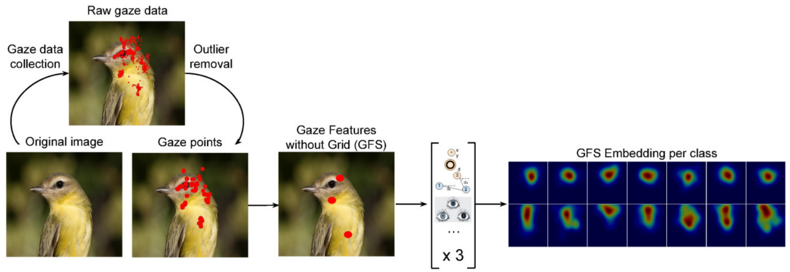
Gaze Embeddings



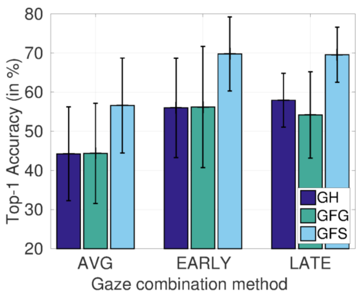
Gaze Embeddings



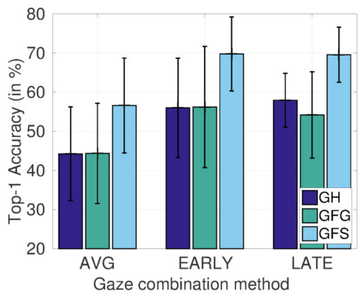
Gaze Embeddings

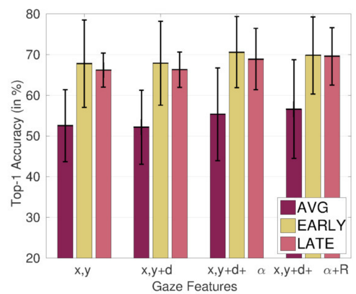


Gaze Embeddings and Gaze Features



Gaze Embeddings and Gaze Features





Gaze Embeddings for Zero-Shot Learning

	CUB-VW
Random points	39.5
Bubbles [Deng et al. CVPR'13]	43.2
Bag of Words from Wikipedia	55.2
Attributes	72.9
Gaze	73.9
Attributes + Gaze	**78.2**

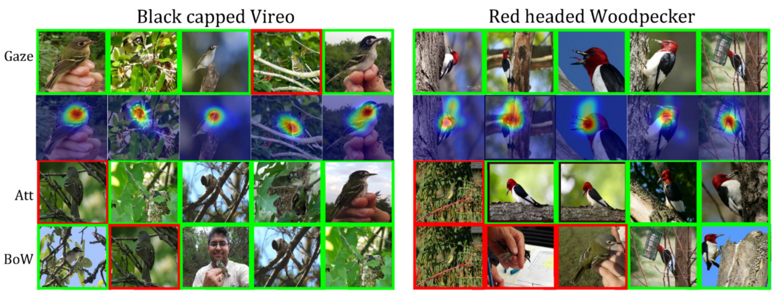
Gaze Embeddings for Zero-Shot Learning

	CUB-VW
Random points	39.5
Bubbles [Deng et al. CVPR'13]	43.2
Bag of Words from Wikipedia	55.2
Attributes	72.9
Gaze	73.9
Attributes + Gaze	**78.2**

Gaze Data → class discriminative + complements attributes

[Karessli et.al. CVPR'17]

Gaze Embeddings



[Karessli et.al. CVPR'17]

Conclusions

Standard image classification models fail with the lack of labels

Conclusions

Standard image classification models fail with the lack of labels

1. Zero-Shot Learning is a challenging task

Conclusions

Standard image classification models fail with the lack of labels

1. Zero-Shot Learning is a challenging task
2. Side information, e.g. attributes, is required

Conclusions

Standard image classification models fail with the lack of labels

1. Zero-Shot Learning is a challenging task
2. Side information, e.g. attributes, is required
3. Several sources of side information exists

[Akata et.al. IEEE CVPR 2013, 2015, 2016, TPAMI 2016] [Reed et.al. IEEE CVPR 2016, ICML 2016, NIPS 2016] [Lampert et.al. IEEE CVPR 2009, TPAMI 2013] [Mikolov et.al. NIPS 2013, Karessli et.al. IEEE CVPR 2017]

Outline

Zero-Shot Learning: Task Formulation

$\mathcal{S} = \{(x_n, y_n), n = 1...N\}$, with $y_n \in \mathcal{Y}^{tr}$

Zero-Shot Learning: Task Formulation

$\mathcal{S} = \{(x_n, y_n), n = 1...N\}$, with $y_n \in \mathcal{Y}^{tr}$

Training: learn $f : \mathcal{X} \to \mathcal{Y}$ by minimizing regularized empirical risk:

$$\frac{1}{N}\sum_{n=1}^{N} L(y_n, f(x_n; W)) + \Omega(W)$$

$L(.)$ = loss function, $\Omega(.)$ = regularization term and

Zero-Shot Learning: Task Formulation

$\mathcal{S} = \{(x_n, y_n), n = 1...N\}$, with $y_n \in \mathcal{Y}^{tr}$

Training: learn $f : \mathcal{X} \to \mathcal{Y}$ by minimizing regularized empirical risk:

$$\frac{1}{N}\sum_{n=1}^{N} L(y_n, f(x_n; W)) + \Omega(W)$$

$L(.)$ = loss function, $\Omega(.)$ = regularization term and

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

Zero-Shot Learning: Task Formulation

$\mathcal{S} = \{(x_n, y_n), n = 1...N\}$, with $y_n \in \mathcal{Y}^{tr}$

Training: learn $f : \mathcal{X} \to \mathcal{Y}$ by minimizing regularized empirical risk:

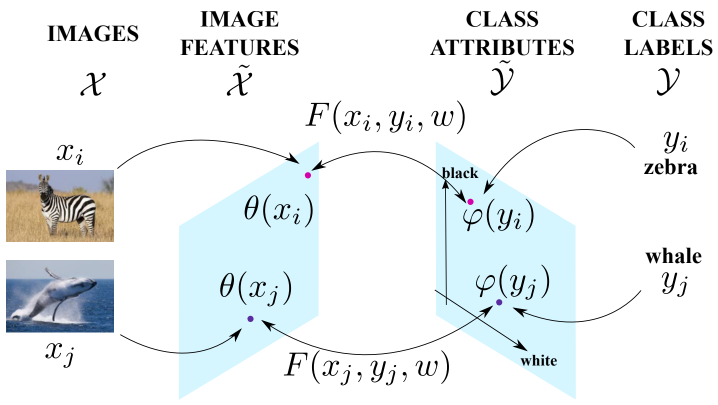
$$\frac{1}{N}\sum_{n=1}^{N} L(y_n, f(x_n; W)) + \Omega(W)$$

$L(.)$ = loss function, $\Omega(.)$ = regularization term and

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

Testing: assign an image to $\mathcal{Y}^{ts} \subset \mathcal{Y}$ with max compatibility

Multimodal Embeddings with Linear Compatibility



$$F(x, y; W) = \theta(x)^T W \phi(y)$$

Deep Visual Semantic Embeddings: DEVISE

Pairwise Ranking: Convex Objective

$$\sum_{y \in \mathcal{Y}^{tr}} [\Delta(y_n, y) + F(x_n, y; W) - F(x_n, y_n; W)]_+$$

Deep Visual Semantic Embeddings: DEVISE

Pairwise Ranking: Convex Objective

$$\sum_{y \in \mathcal{Y}^{tr}} [\Delta(y_n, y) + F(x_n, y; W) - F(x_n, y_n; W)]_+$$

- $\Delta(y_n, y) = 1$ if $y_n = y$, otherwise 0
- Optimized by SGD

[Frome et.al. NIPS 2013]

Attribute Label Embedding: ALE

Weighted Pairwise Ranking Loss:

$$\sum_{y \in \mathcal{Y}^{tr}} l_k[\Delta(y_n, y) + F(x_n, y; W) - F(x_n, y_n; W)]_+$$

Attribute Label Embedding: ALE

Weighted Pairwise Ranking Loss:

$$\sum_{y \in \mathcal{Y}^{tr}} l_k [\Delta(y_n, y) + F(x_n, y; W) - F(x_n, y_n; W)]_+$$

- $\Delta(y_n, y) = 1$ if $y_n = y$, otherwise 0
- $l_k = \sum_{i=1}^{k} \alpha_i$ with $\alpha_i = 1/i$
- Optimized by SGD

[Akata et.al. CVPR 2013 & TPAMI 2016]

Structured Joint Embedding: SJE

Multiclass Objective:

$$[\max_{y \in \mathcal{Y}^{tr}} (\Delta(y_n, y) + F(x_n, y; W)) - F(x_n, y_n; W)]_+$$

Structured Joint Embedding: SJE

Multiclass Objective:

$$[\max_{y \in \mathcal{Y}^{tr}} (\Delta(y_n, y) + F(x_n, y; W)) - F(x_n, y_n; W)]_+$$

- Full weight to the top of the ranked list
- Requires computing score wrt all the classifiers for each sample

[Akata et.al. CVPR 2015 & Reed et.al. CVPR 2016]

Embarassingly Simple Zero-Shot Learning: ESZSL

Additional Regularization Term to SJE Objective:

$$\gamma \|W\phi(y)\|^2 + \lambda \|\theta(x)^T W\|^2 + \beta \|W\|^2$$

where γ, λ, β are regularization parameters

- Euclidean norm of projected attributes in the feature space
- Projected image feature in the attribute space are bounded

[Romera-Paredes and Torr, ICML 2015]

Semantic Autoencoder: SAE

Objective: similar to the linear auto-encoder

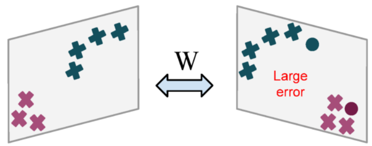
$$\min_{W} ||\theta(x) - W^T\phi(y)||^2 + \lambda||W\theta(x) - \phi(y)||^2,$$

- Learns a linear projection from $\theta(x)$ to $\phi(y)$
- Projection must reconstruct the original image embedding

[Kodirov et.al. CVPR 2017]

Latent Embeddings: LATEM

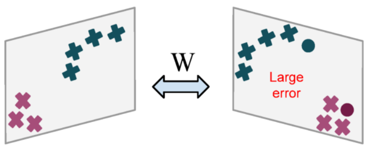
Linear compatibility



$$F(x, y; W) = \theta(x)^T W\phi(y)$$

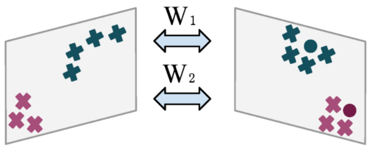
Latent Embeddings: LATEM

Linear compatibility



$$F(x, y; W) = \theta(x)^T W\phi(y)$$

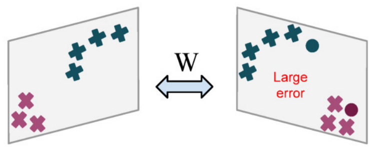
Piecewise-linear compatibility



$$F(x, y; W) = \theta(x)^T W_i\phi(y)$$

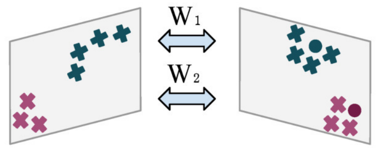
Latent Embeddings: LATEM

Linear compatibility

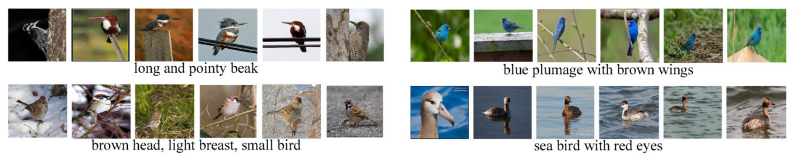


$$F(x, y; W) = \theta(x)^T W\phi(y)$$

Piecewise-linear compatibility



$$F(x, y; W) = \theta(x)^T W_i\phi(y)$$



[Xian et.al. CVPR'16]

Cross-Modal Transfer: CMT

Deep nonlinear embedding objective:

$$\sum_{y \in \mathcal{Y}^{tr}} \sum_{x \in \mathcal{X}_y} \|\phi(y) - W_1 \tanh(W_2.\theta(x))\|^2$$

- (W_1, W_2): weights of the two layer neural network
- Novelty detection: to assign images to unseen or seen classes

[Socher et.al. NIPS'13]

Direct Attribute Prediction: DAP

Two step process

Direct Attribute Prediction: DAP

Two step process

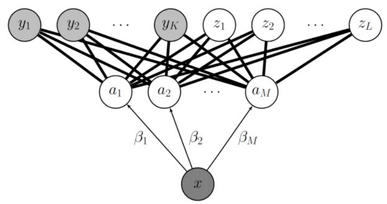
- learn attribute classifiers

Direct Attribute Prediction: DAP

Two step process

- learn attribute classifiers
- combine scores of learned attribute classifiers

$$f(x) = \arg\max_c \prod_{m=1}^{M} \frac{p(a_m^c|x)}{p(a_m^c)}$$



[Lampert et.al. CVPR'09 & TPAMI'13]

Convex Combination of Semantic Emb.: `CONSE`

Probability of a training image belonging to a training class:

$$f(x,t) = \arg\max_{y \in \mathcal{Y}^{tr}} p_{tr}(y|x)$$

Convex Combination of Semantic Emb.: `CONSE`

Probability of a training image belonging to a training class:

$$f(x,t) = \arg\max_{y \in \mathcal{Y}^{tr}} p_{tr}(y|x)$$

Combination of semantic embeddings (s) is used to assign an unknown image to an unseen class:

$$\frac{1}{Z} \sum_{i=1}^{T} p_{tr}(f(x,t)|x), s(f(x,t))$$

- Z = tth most likely label for image x
- T maximum number of semantic embedding vectors

[Norouzi et.al. ICLR'14]

Synthesized Classifiers: `SYNC`

Weighted bipartite graph (s_{cr}): Training (w_c) and Phantom (v_r)

Synthesized Classifiers: `SYNC`

Weighted bipartite graph (s_{cr}): Training (w_c) and Phantom (v_r)

Objective is to minimize distortion error:

$$\min_{w_c} \|w_c - \sum_{r=1}^{R} s_{cr} v_r\|_2^2.$$

Synthesized Classifiers: SYNC

Weighted bipartite graph (s_{cr}): Training (w_c) and Phantom (v_r)

Objective is to minimize distortion error:

$$\min_{w_c} \|w_c - \sum_{r=1}^{R} s_{cr} v_r\|_2^2.$$

Novel class: linear combination of phantom class classifiers

[Changpinyo et.al. CVPR'16]

Co-Occurrence Statistics: COSTA

Uses co-occurrence statistics

- of visual concepts
- between seen and unseen classes

Estimate w_l to classify the unseen label l: $w_l = \sum_k w_k s_{lk}$

[Mensink et.al. CVPR'14]

Summary of Presented ZSL Models

Existing ZSL models can be grouped into 4:

Summary of Presented ZSL Models

Existing ZSL models can be grouped into 4:

1. Linear Compatibility: ALE, DEVISE, SJE, ESZSL, SAE

Summary of Presented ZSL Models

Existing ZSL models can be grouped into 4:

1. Linear Compatibility: `ALE`, `DEVISE`, `SJE`, `ESZSL`, `SAE`
2. Non-linear Compatibilty: `LATEM`, `CMT`

Summary of Presented ZSL Models

Existing ZSL models can be grouped into 4:

1. Linear Compatibility: `ALE`, `DEVISE`, `SJE`, `ESZSL`, `SAE`
2. Non-linear Compatibilty: `LATEM`, `CMT`
3. Two-stage Inference: `DAP`, `CONSE`

Summary of Presented ZSL Models

Existing ZSL models can be grouped into 4:

1. Linear Compatibility: `ALE`, `DEVISE`, `SJE`, `ESZSL`, `SAE`
2. Non-linear Compatibilty: `LATEM`, `CMT`
3. Two-stage Inference: `DAP`, `CONSE`
4. Hybrid Model: `SYNC`

[Akata et.al IEEE CVPR 2013, Frome et.al. NIPS 2013, Akata et. al. 2015, Romera Paredes and Torr ICML 2015, , Kodirov et.al IEEE CVPR 2017, Xian et.al. IEEE CVPR 2016, Socher et.al. NIPS 2013, , Lampert et.al. IEEE CVPR 2009 & TPAMI 2013, Norouzi et.al. ICLR 2014, Changpinyo et.al. IEEE CVPR 2016]

Outline

Zero-Shot Learning: The Good, The Bad, The Ugly

The Good: ZSL is an important direction that has gained interest

Zero-Shot Learning: The Good, The Bad, The Ugly

The Good: ZSL is an important direction that has gained interest

The Bad: No unified evaluation protocol exists

Zero-Shot Learning: The Good, The Bad, The Ugly

The Good: ZSL is an important direction that has gained interest

The Bad: No unified evaluation protocol exists

The Ugly: Test Classes overlap with ImageNet 1K

Benchmark on Attribute Datasets and ImageNet

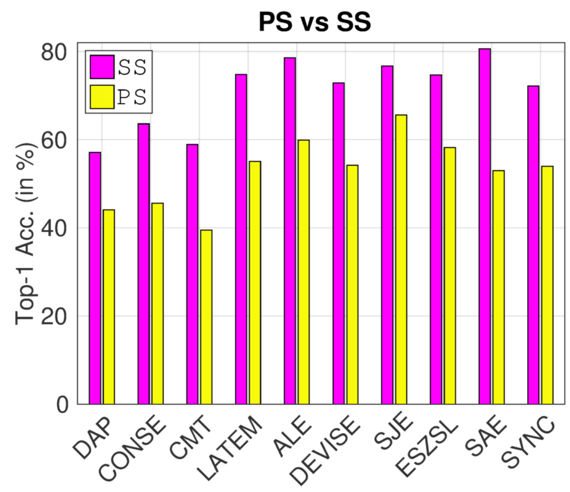
Dataset	Size	$\|\mathcal{Y}\|$	$\|\mathcal{Y}^{tr}\|$	$\|\mathcal{Y}^{ts}\|$
SUN	$14K$	717	$580+65$	72
CUB	$11K$	200	$100+50$	50
AWA1	$30K$	50	$27+13$	10
AWA2*	$37K$	50	$27+13$	10
aPY	$1.5K$	32	$15+5$	12

Benchmark on Attribute Datasets and ImageNet

Dataset	Size	$\lvert\mathcal{Y}\rvert$	$\lvert\mathcal{Y}^{tr}\rvert$	$\lvert\mathcal{Y}^{ts}\rvert$
SUN	$14K$	717	580 + 65	72
CUB	$11K$	200	100 + 50	50
AWA1	$30K$	50	27 + 13	10
AWA2*	$37K$	50	27 + 13	10
aPY	$1.5K$	32	15 + 5	12

ImageNet Split	$\lvert\mathcal{Y}^{ts}\rvert$
ImageNet 21K - $\mathcal{Y}^{tr}$	20345
Within 2/3 hops from $\mathcal{Y}^{tr}$	1509/7678
Most populated classes	$500/1K/5K$
Least populated classes	$500/1K/5K$

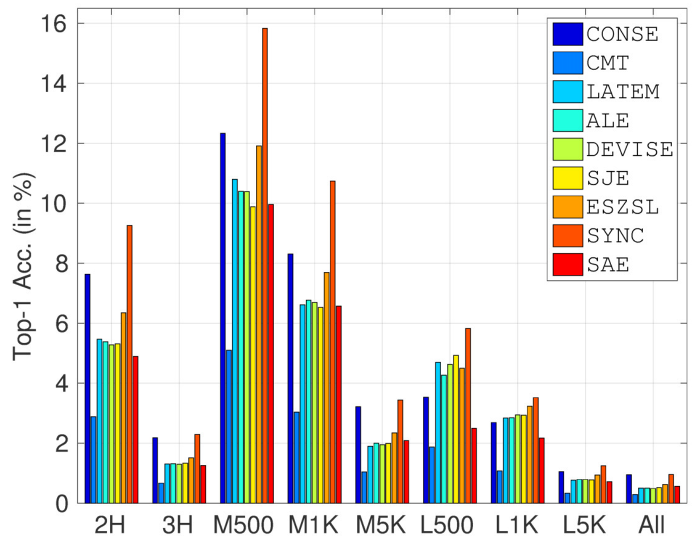
ZSL Results wrt Data Splits on AWA



Ranking Models on Attribute Datasets

Rank	1	2	3	4	5	6	7	8	9	10
ALE [2.0]	5	7	2		1					
DEVISE [2.9]	2	5	4	1	2	1				
SJE [3.4]	5	1	3		1	5				
ESZSL [4.2]	1		3	5	4	1	1			
LATEM [4.5]		1	1	6	5	1	1			
SYNC [5.3]	2	1	2	1		3	3	1	2	
DAP [7.3]				2	1		4	4	4	
SAE [8.4]					1	3	2		1	8
CMT [8.5]						1	3	5		6
CONSE [8.6]							1	5	8	1

Benchmark of ZSL on ImageNet



Conclusions

Benchmark of Zero-Shot Learning

1. Zero-Shot Learning has attracted lots of attention

Conclusions

Benchmark of Zero-Shot Learning

1. Zero-Shot Learning has attracted lots of attention
2. We propose a unified evaluation procedure

Conclusions

Benchmark of Zero-Shot Learning

1. Zero-Shot Learning has attracted lots of attention
2. We propose a unified evaluation procedure
3. Comprehensive evaluation of 12 models on 6 datasets

[Xian et.al. IEEE CVPR 2017 & ArXiv 2017]

Outline

Summary of ZSL for Image Classification

1. Large-scale image classification fails with lack of data
   [Akata et.al. TPAMI'14]

Summary of ZSL for Image Classification

1. Large-scale image classification fails with lack of data
   [Akata et.al. TPAMI'14]
2. Structured Joint Embeddings tackles lack of visual data
   [Akata et.al. CVPR'13, Akata et.al. TPAMI'16]

Summary of ZSL for Image Classification

1. Large-scale image classification fails with lack of data
   [Akata et.al. TPAMI'14]
2. Structured Joint Embeddings tackles lack of visual data
   [Akata et.al. CVPR'13, Akata et.al. TPAMI'16]
3. Attributes, text and gaze provide side information
   [Akata et.al. CVPR'15 & CVPR'16, Xian et.al. CVPR'16 & CVPR'17, Karessli et.al. CVPR'17]

Summary of ZSL for Image Classification

1. Large-scale image classification fails with lack of data
   [Akata et.al. TPAMI'14]
2. Structured Joint Embeddings tackles lack of visual data
   [Akata et.al. CVPR'13, Akata et.al. TPAMI'16]
3. Attributes, text and gaze provide side information
   [Akata et.al. CVPR'15 & CVPR'16, Xian et.al. CVPR'16 & CVPR'17, Karessli et.al. CVPR'17]
4. The Good, the bad and the ugly aspects of zero-shot learning
   [Xian et.al. CVPR'17 & ArXiv'17]

Thank you!

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

Mammal

Brown

Curvy beak

Pointy ears

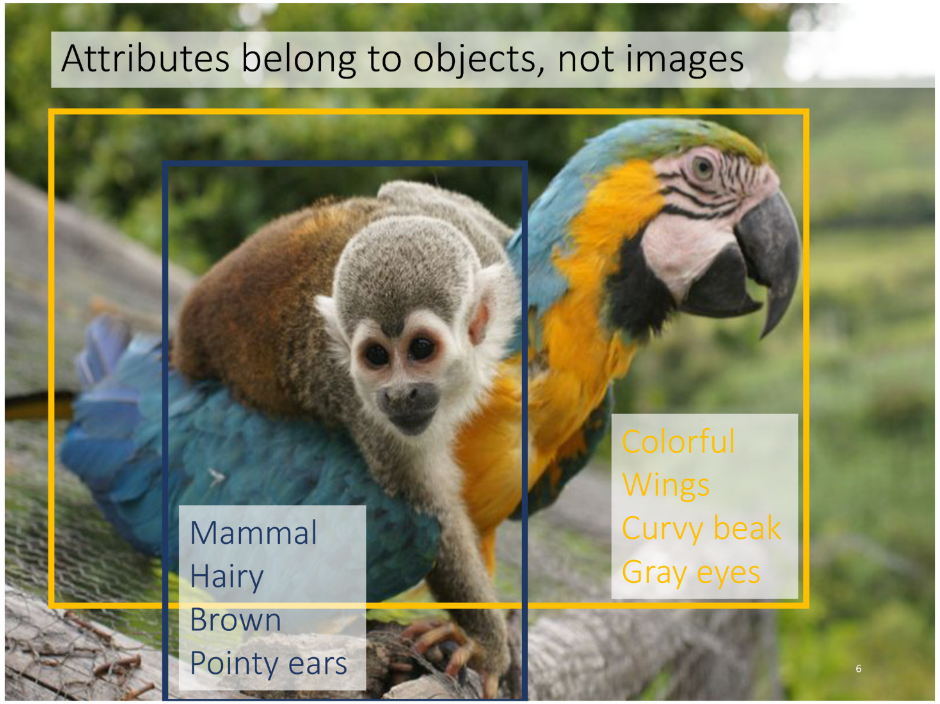
Hairy

Wings

Colorful

Gray eyes

5



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

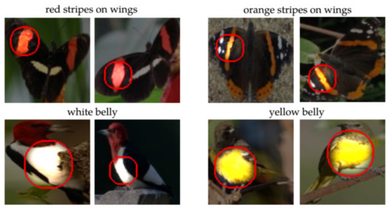
Also, semantically meaningful

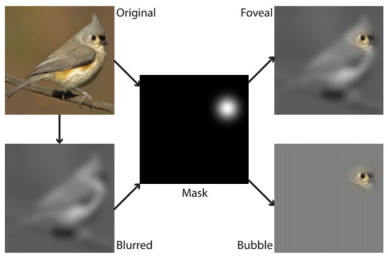
- By design: human in the loop [1]
- By unsupervised clustering [2]

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

Automatically detect discriminative attributes

- Solve CRFs iteratively
- Random attribute initialization

Not necessarily “nameable”

- Convert them to nameable
- Human approves meaningful attributes

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

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

Set of attributes CRF

Specific attribute CRF

[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]

CPMC Regions

ALE attributes

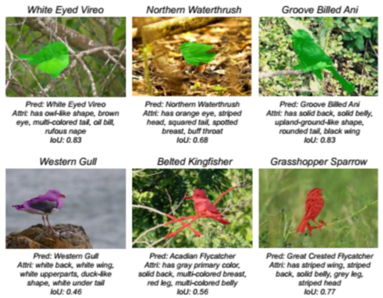
Per region maximization

~500

[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

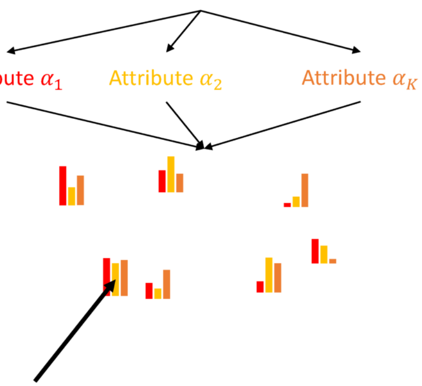
Zero-shot Localization by Attributes

Training

Dataset D_A

$\theta_1 = \max \mathcal{L}(D_A; \alpha_1)$
$\theta_2 = \max \mathcal{L}(D_A; \alpha_2)$
$\theta_3 = \max \mathcal{L}(D_A; \alpha_3)$

Zero-Shot Inference

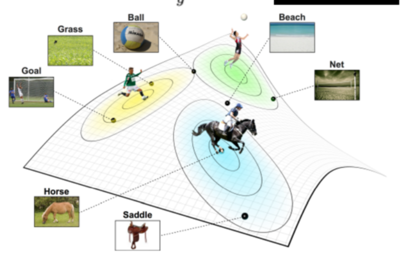


[1] Attributes make sense on segmented objects, Li et al., ECCV 2014

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

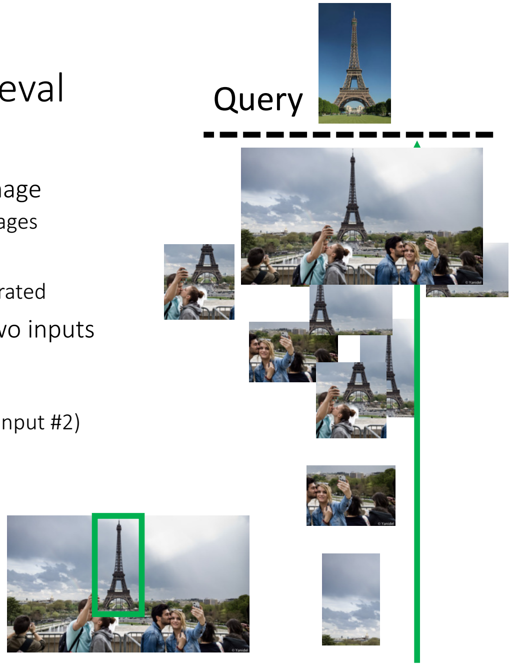
$$\mathcal{C}(v) = \underset{z \in \mathcal{Z}, u \in \mathcal{U}_v}{\text{argmax}} \sum_y p_{uy}\, g_{yz}$$ [redacted]



[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)



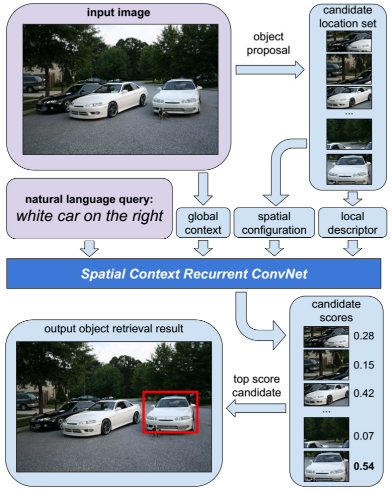
Zero-shot Localization by Free Text

- Similar to Zero-Shot Localization [1]
  - #Input 1 is now a text query
- Rank sliding images
  - Scoring function measures similarity of image to text

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

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

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



[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

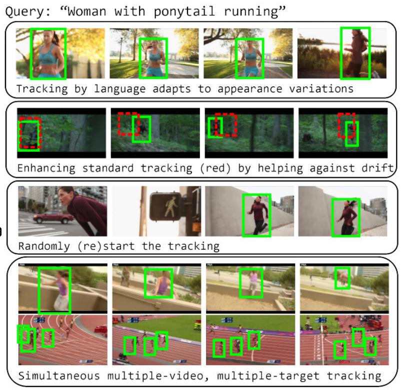
Zero-shot localization in videos, aka *Tracking by Natural Language [1]*

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

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

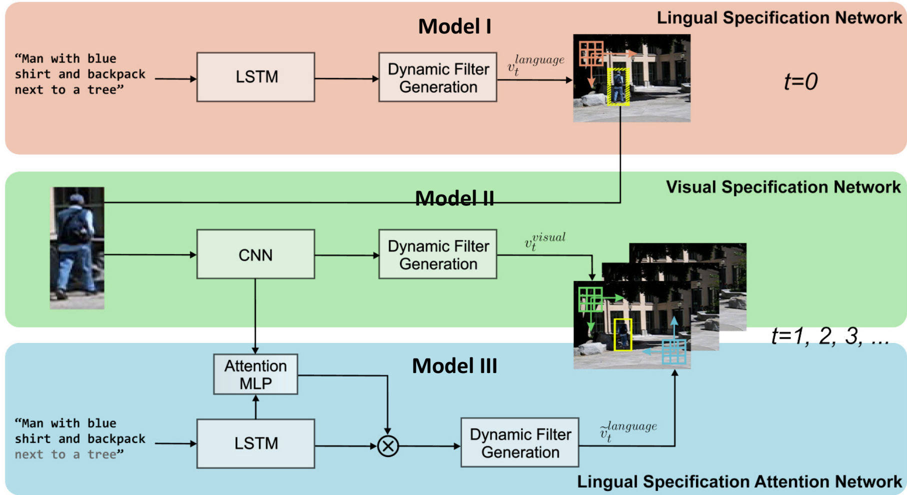
Zero-shot localization in videos, aka *Tracking by Natural Language*

- Novel type of human-machine interaction
  - *"Tesla, follow the red car in the middle lane"*
- Enables novel tracking scenarios
  - No "first-frame" requirement → ideal for "live" or online tracking
  - Multiple-video, multiple-target tracking → ideal for large scale monitoring
- More robust standard tracking
  - Tracker adapts to appearance variations
  - Helping against drift



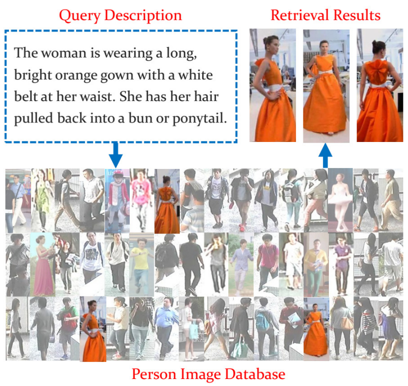
[1] *Tracking by Natural Language Specification*, Li et al., CVPR 2017

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

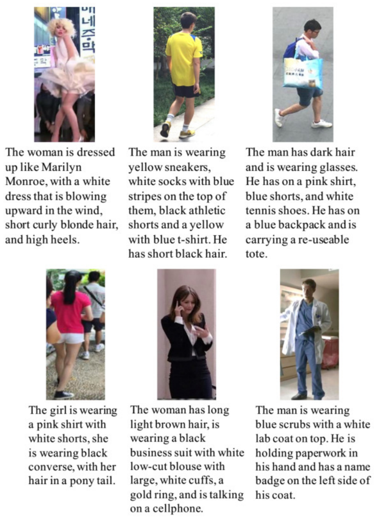


[1] *Tracking by Natural Language Specification*, Li et al., CVPR 2017

Person search with Natural Language

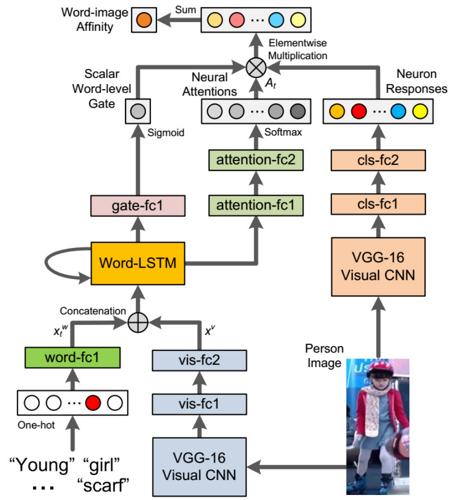


[1] *Person Search with Natural Language Description*, Li et al., CVPR 2017



Person search with Natural Language

- First extract region proposals
- Then compute word specific (dynamic) filters
- Computer Word-Image affinity



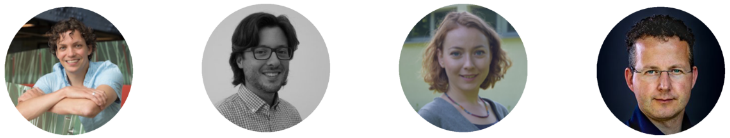
[1] *Person Search with Natural Language Description*, Li et al., CVPR 2017

Conclusion

- Attributes belong to objects, not images
- Zero-Shot localization natural extension
- Object tracking by natural language description is a very novel and relevant direction
  - Also connected to video object detection

Zero-Shot Learning

for Computer Vision



Thomas Mensink, Efstratios Gavves, Zeynep Akata, Cees Snoek

University of Amsterdam

1

- 13:30-13:40 | **Introduction** | Efstratios Gavves
- 13:40-14:30 | **Classification** | Zeynep Akata
- 14:30-15:00 | **Localization** | Efstratios Gavves
- 15:00-15:30 | **Retrieval** | Cees G.M. Snoek
- 15:30-16:00 | **Break**
- 16:00-16:40 | **Open problems** | Zeynep Akata, Efstratios Gavves
- 16:40-17:00 | **Conclusion** | Efstratios Gavves

TUTORIAL PROGRAM

Lampert et al PAMI 2013, and many others

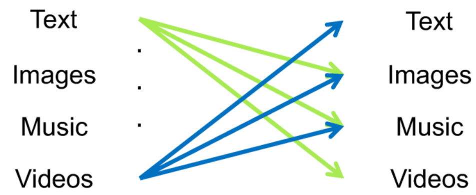
Difference with traditional zero-shot

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

Rasiwasia et al. MM10
Costa et al. TPAMI14

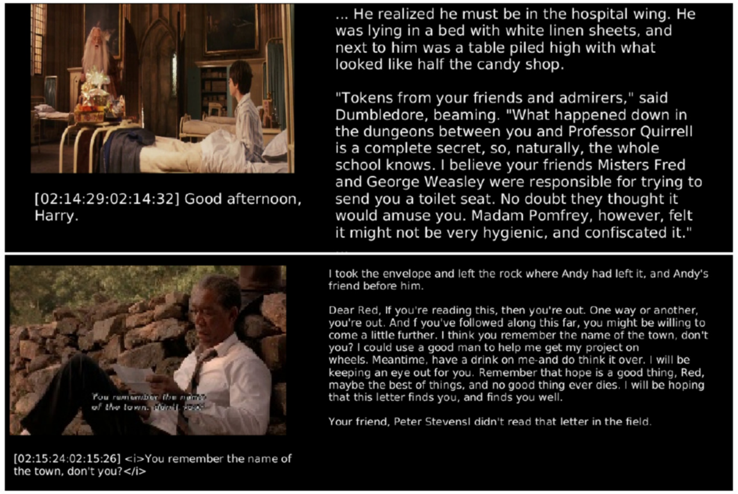
Retrieving images from Wikipedia text

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

Zhu et al. ICCV15

Retrieving book excerpts from movies



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 ...

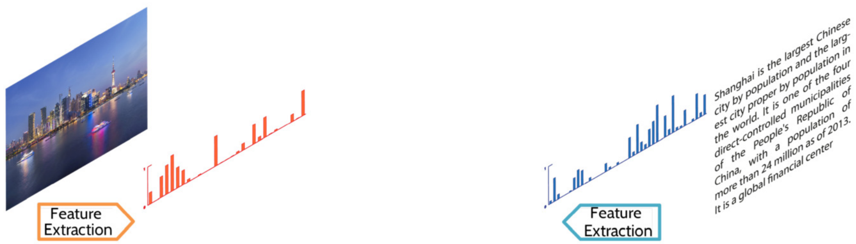
Event Name: Winning a race without a vehicle

NIST TRECVID Multimedia Event Detection Benchmark

Problem statement

How to align visual and textual representations?

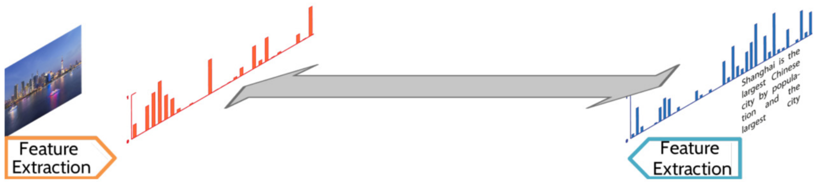
Different dimensionality, distributions, and meaning



Low-level alignment

Aligns two modalities directly at low-level features

Canonical Correlation Analysis, Cross-Media hashing, …

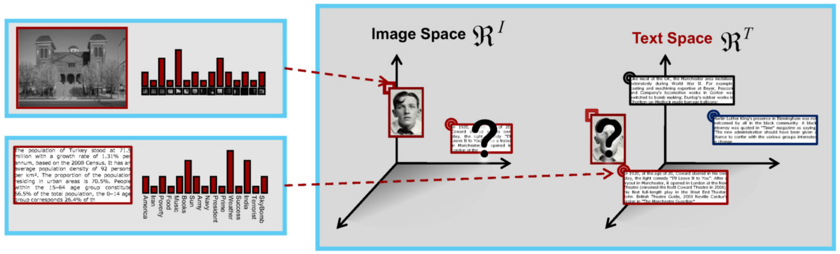


Not the most effective space to learn the correlations

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

Slide credit: Nikhil Rasiwasia

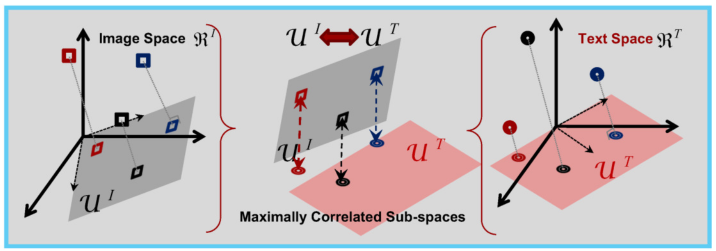
How to compute similarity?



Slide credit: Nikhil Rasiwasia

Canonical Correlation Analysis

Learn subspaces that maximize correlation between two modalities



Joint dimensionality reduction across two (or more) spaces

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

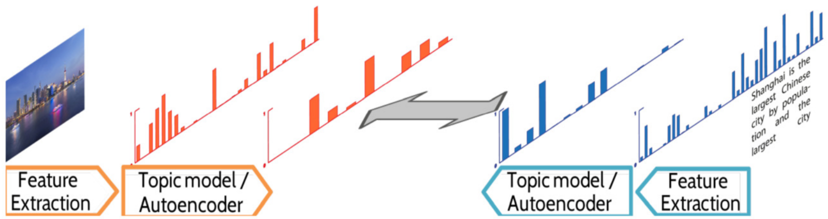
Basis for the maximally correlated space

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

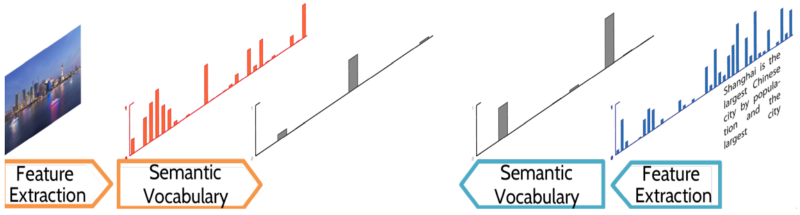
[Blei et al., SIGIR’03] [Wang et al., MM’14] [Feng et al., MM’14] …

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] ...

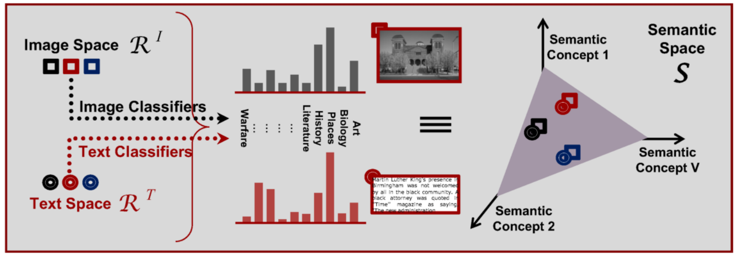
Rasiwasia et al. MM10
Costa et al. TPAMI14

Semantic alignment via concepts

Design semantic spaces for both modalities

A space where each dimension is a semantic concept/attribute.

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



Problem: define, annotate and train concepts

LEARNING THE SEMANTIC ALIGNMENT

Amirhossein Habibian, Thomas Mensink, and Cees G. M. Snoek.
Video2vec Embeddings Recognize Events when Examples are Scarce.
IEEE Transactions on Pattern Analysis and Machine Intelligence. In press.
Previously best paper ACM Multimedia 2014.

Research question

Can we **learn the alignment** from videos and their stories?

Video

Story

Story usually highlights the key concepts in video

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

Multimedia embeddings

Joint space where $x_i W \approx y_i A$

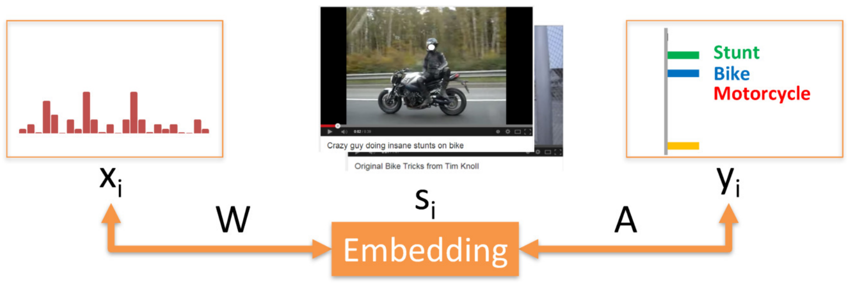
Explicitly relate training W and A from multimedia

W = Visual projection matrix individual term classifiers

A = Textual 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

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

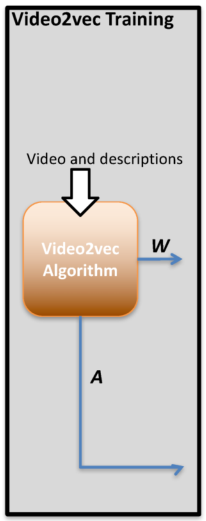
Video2vec objectives: predictability

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]

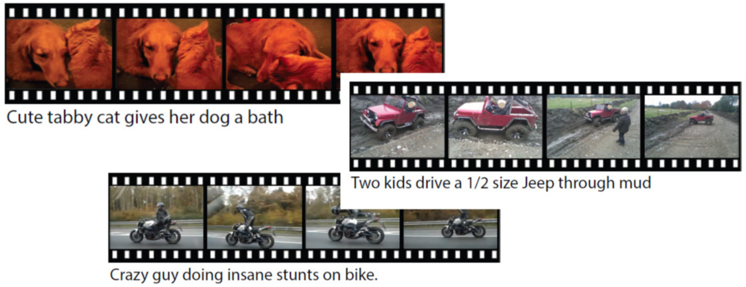
VideoStory46K dataset

Videos and title descriptions from YouTube

46K videos, 19K unique terms in descriptions

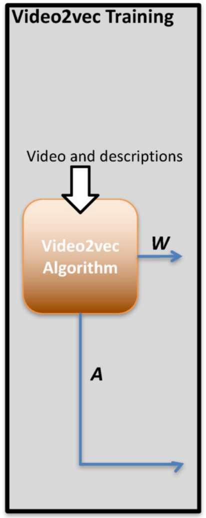
Seeded from video event descriptions

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

$$\nabla_{\boldsymbol{A}} L_{\mathrm{VS}} = -2\left(\boldsymbol{y}_t - \boldsymbol{A}\boldsymbol{s}_t\right)\boldsymbol{s}_t^{\top} + \lambda_a \boldsymbol{A},$$

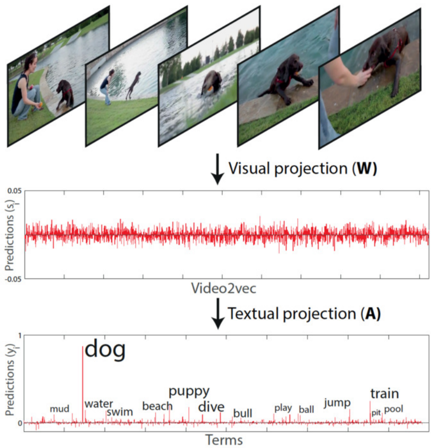
$$\nabla_{\boldsymbol{W}} L_{\mathrm{VS}} = -2\,\boldsymbol{x}_t\left(\boldsymbol{s}_t - \boldsymbol{W}^{\top}\boldsymbol{x}_t\right)^{\top} + \lambda_w \boldsymbol{W}, \text{ and}$$

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

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

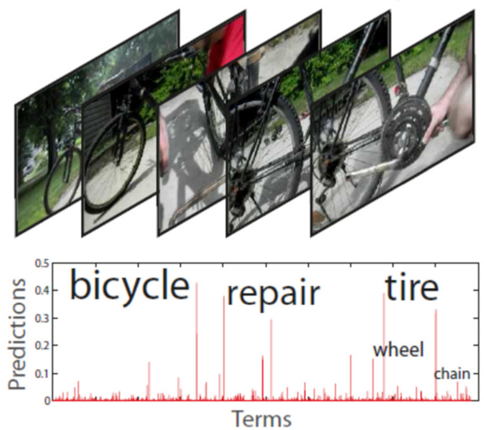
$$\hat{\boldsymbol{y}}_i = \boldsymbol{A}\boldsymbol{s}_i,$$

3. Cosine distance for matching

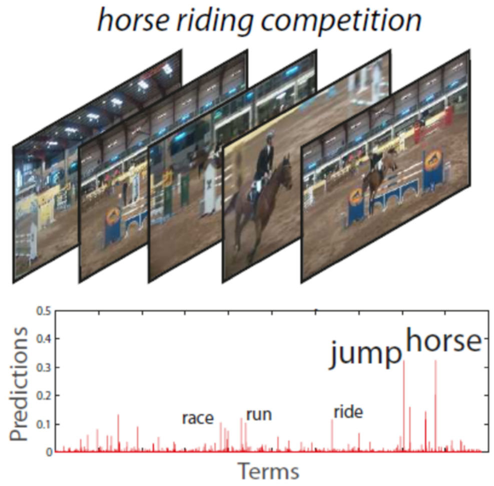
$$s_e(\boldsymbol{x}_i) = \frac{\boldsymbol{y}^{e\top}\hat{\boldsymbol{y}}_i^e}{||\boldsymbol{y}^e|| \quad ||\hat{\boldsymbol{y}}_i^e||}$$

Video2vec predicted terms

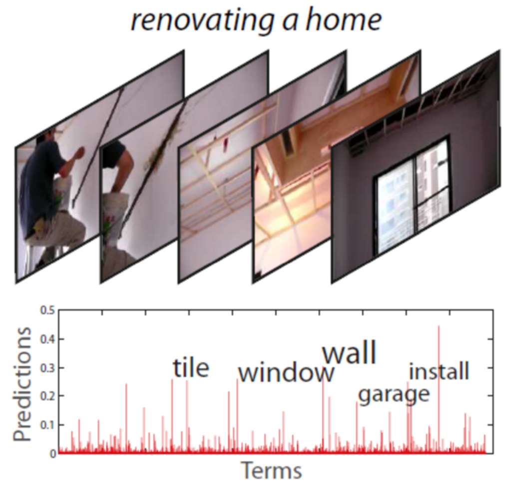
non-motorized vehicle repair



Video2vec predicted terms



Video2vec predicted terms



Zero-shot event retrieval

Authors	Published	mAP
Habibian et al.	ICMR 2014	6.4
Ye et al.	MM 2015	9.0
Chang et al.	IJCAI 2015	9.6
Wu et al.	CVPR 2014	12.7
Jiang et al.	AAAI 2015	12.9
Mazloom et al.	TMM 2016	12.9
Hussein et al.	CVPR 2017	17.9
Liang et al.	MM 2015	18.3
Video2vec embedding	TPAMI 2017	20.0

NIST TRECVID Multimedia Event Detection Benchmark, edition 2013

ADDING LOCALIZATION

Pascal Mettes and Cees G. M. Snoek.
Spatial-Aware Object Embeddings for Zero-Shot Localization and Classification of Actions. *To appear in ICCV 2017.*

Related: zero-shot action recognition

Focus on classification.

Position is irrelevant.

Emphasis:
Are **relevant** objects present.

Merler *et al.* TMM 2012
Liu *et al.* WACV 2013
Jain *et al.* CVPR 2015, ICCV 2015
Gan *et al.* CVPR 2016
Xu *et al.* ICIP 2015, ECCV 2016, IJCV 2017

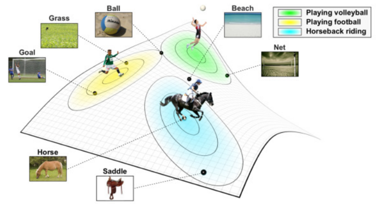
Jain *et al.* ICCV 2015

Related: Objects2action

Simple convex combination of known classifiers

$$\mathcal{C}(v) = argmax_z \sum_y p_{vy}\, g_{yz}$$

Test video
Object representation
Object/action affinities



$$g_{yz} = s(y)^T s(z),$$

where *s()* = word2vec

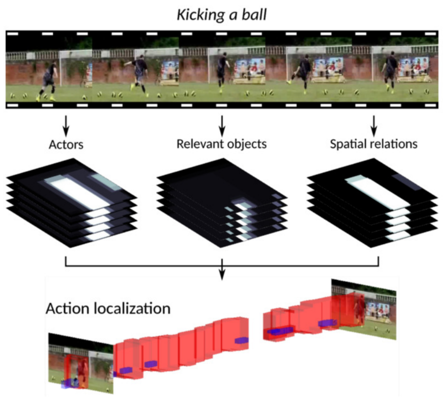
Our proposal

Two types of objects.

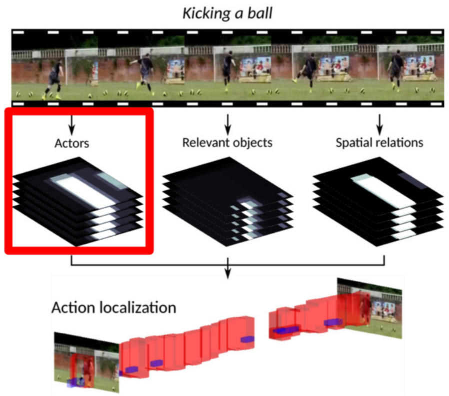
Interacting objects.
Background objects.

Position key for relevance.

Spatial-aware object embeddings

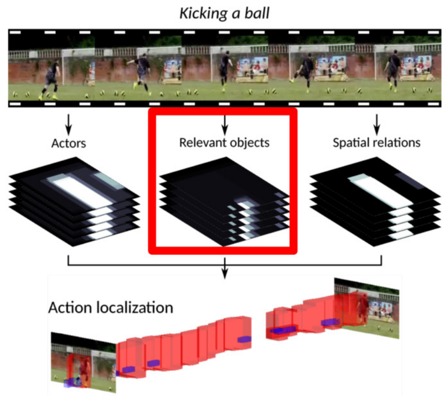


Spatial-aware object embeddings



Where are actors occurring?

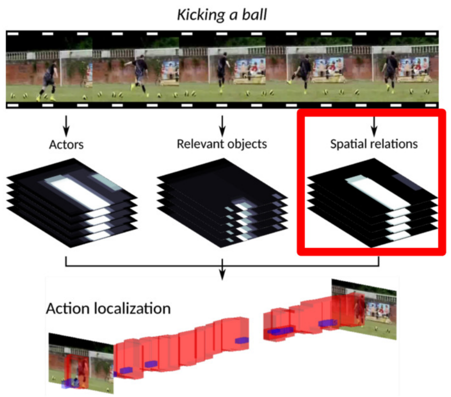
Spatial-aware object embeddings



Where are actors occurring?

Where are the <u>relevant</u> objects?

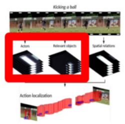
Spatial-aware object embeddings



Where are actors occurring?

Where are the <u>relevant</u> objects?

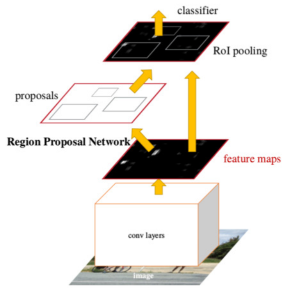
Are objects located as expected?



Actors and objects

Faster R-CNN for bounding boxes and scores.

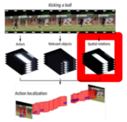
Pre-trained on ImageNet and MS-COCO.



Ren *et al.* NIPS 2015

Actor: Use person class.

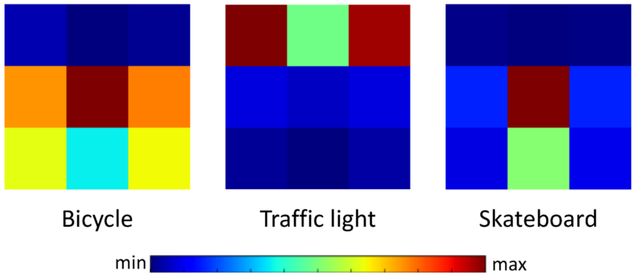
Object: Select objects with highest word2vec similarity.



Spatial relations

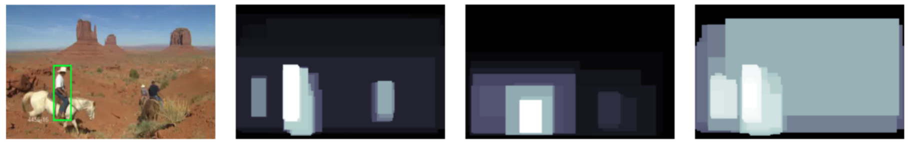
Relative positions of object and actor mined from MS-COCO.

3x3 grid used: Left of, Right of, On, Above, Below, Above left of, ...



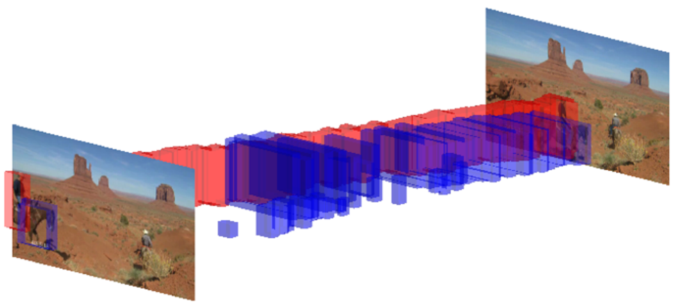
Scoring actor-object interactions

$$s(b, F, \mathcal{Z}) = \boxed{p(\texttt{actor}|b)} + \sum_{o \in O_{\mathcal{Z}}} w(o, \mathcal{Z}) \cdot \left(\max_{f \in F_n} \boxed{p(o|f)} \cdot \boxed{(1 - JSD_2(d(\texttt{actor}, o)||d(b, f)))} \right)$$



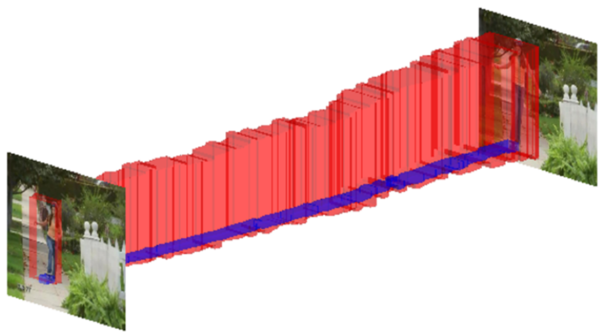
Link boxes over time that have both high scores and high overlaps

Qualitative results



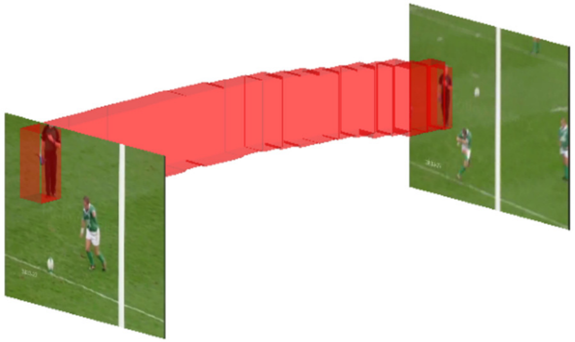
Riding horse (*horse*)

Qualitative results



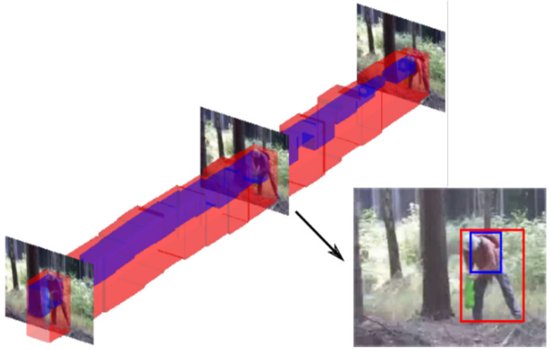
Skateboarding (*skateboard*)

Qualitative results



Kicking (*tie*)

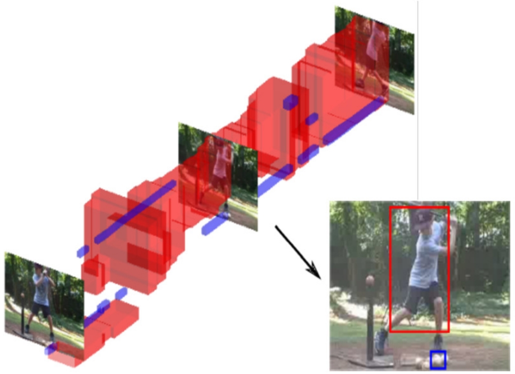
Spatiotemporal action retrieval



Backpack *on* actor

Spatiotemporal action retrieval

Desired object size can also be incorporated in query.



Sports ball (0.10) *RIGHT OF* actor

Conclusion on zero-shot retrieval

Semantic embeddings align visual and textual modality

Learn embedding from webly-supervised classifiers

Off-the-shelf object detectors add spatial-awareness

Spatiotemporal retrieval in video with position and size.

Zero-Shot Learning
Open Problems

Efstratios Gavves

Zero-shot recap

Training

Dataset D_{A}

Learn attributes
$\theta_1 = \max \mathcal{L}(D_{\mathrm{A}}; att_1)$
$\theta_2 = \max \mathcal{L}(D_{\mathrm{A}}; att_2)$
$\theta_3 = \max \mathcal{L}(D_{\mathrm{A}}; att_3)$

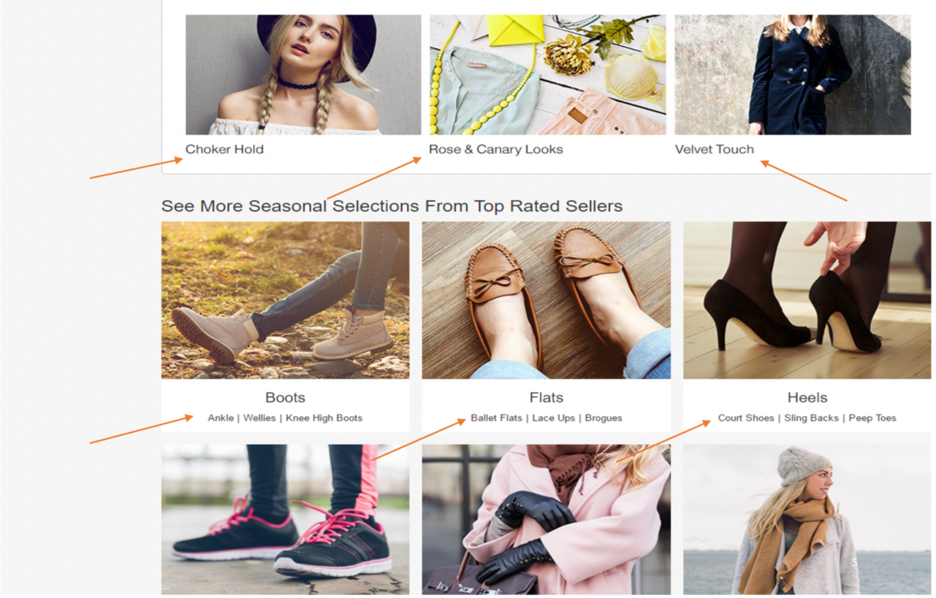
Zero-Shot Inference

"African elephant"

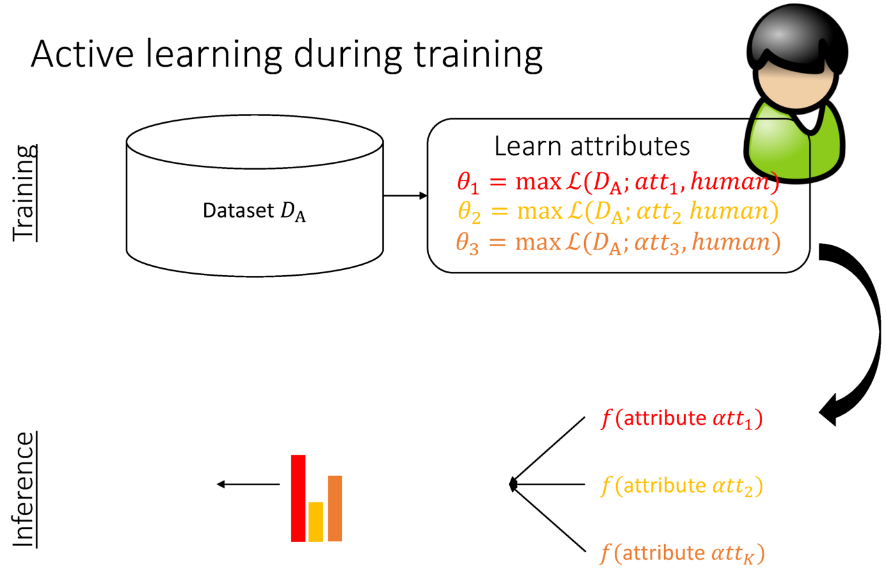
$f(\text{attribute } att_1)$
$f(\text{attribute } att_2)$
$f(\text{attribute } att_K)$

Why not Knowledge Transfer with Interaction?

Attributes are often ad-hoc

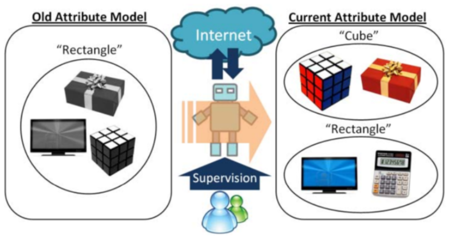


Active learning during training



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

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

$$\text{minimize} \quad \left(\frac{1}{2}||\boldsymbol{w_m^T}||_2^2 + C\left(\sum \xi_{ij}^2 + \sum \gamma_{ij}^2\right)\right)$$

$$\begin{aligned} s.t. \quad & \boldsymbol{w_m^T}(\boldsymbol{x_i}-\boldsymbol{x_j}) \geq 1-\xi_{ij}; \forall(i,j)\in O_m \\ & |\boldsymbol{w_m^T}(\boldsymbol{x_i}-\boldsymbol{x_j})| \leq \gamma_{ij}; \forall(i,j)\in S_m \\ & \xi_{ij}\geq 0; \gamma_{ij}\geq 0, \end{aligned}$$

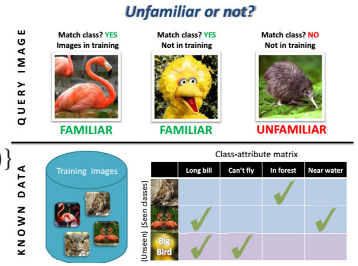
[1] Relative Attributes for Enhanced Man-Machine Communication, Parikh et al., AAAI 2012

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

$$p(c|U,x) = \frac{p(U|\mathbf{a}^c,x)p(\mathbf{a}^c|x)}{p(U|x)}$$

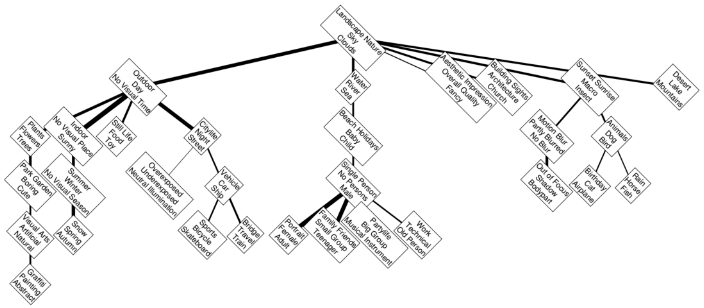
$$p(U|\mathbf{a}^c,x) = \prod_{\tilde{a}_i\in U} p(\tilde{a}_i|a_i^c)^\gamma = \exp\{\sum_{\tilde{a}_i\in U}\gamma\log p(\tilde{a}_i|a_i^c)\}$$



[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

Knowledge is not static

- Every year new and large datasets pop up
- Few out of the ~90 new datasets in 2016-2017
  - Kinetics
  - M2CAI
  - ScanNet
  - Oxford RobotCar
  - Cityscapes
  - LabelMeFacade
- Wikipedia expands by 10 edits per second, 750 new articles per day

- Should we discard old datasets & knowledge when new ones appears?
- Can we actively engage with external knowledge sources such as Wikipedia, so that QA is not constrained to whatever dataset we trained?

External data sources?

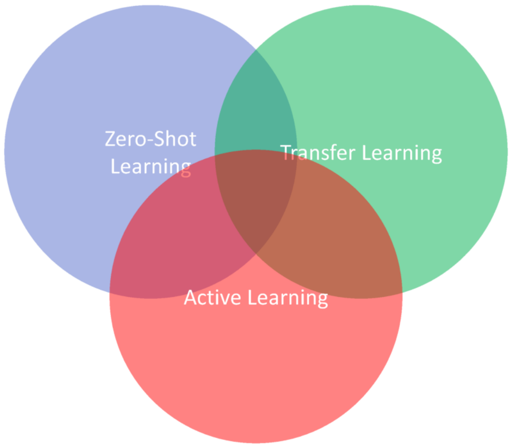
- A few only external data sources one can rely on
  - Wikihow
  - Wikipedia
  - Wikitravel
  - DBPedia
  - EventNet

- A few ways only one can exploit external data sources
  - Active Learning [1]
  - Parsing knowledge graphs [2]
  - Avoiding catastrophic forgetting [3]

*[1] Active Transfer Learning with Zero-Shot Priors: Reusing Past Datasets for Future Tasks,*Gavves, et al., ICCV 2015
[2] The More You Know: Using Knowledge Graphs for Image Classification, Marino et al., CVPR 2017
[3] Overcoming catastrophic forgetting in neural networks, Kirkpatrick et al., arXiv 2016

Zero-Shot, Transfer and Active Learning overlap!

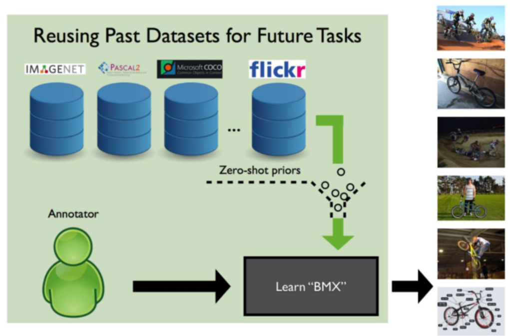
- What if we integrate the three learning paradigms [1]



*[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

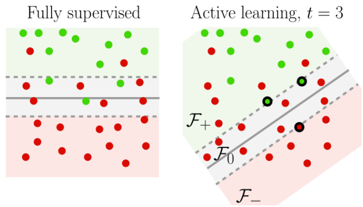
$$\max_{\alpha^t,\gamma^t} \sum_i \gamma_i^t \lambda_i^t \alpha_i^t - \frac{1}{2}\sum_{i,j} \alpha_i^t \alpha_j^t \gamma_i^t \gamma_j^t y_i y_j \boldsymbol{x}_i \cdot \boldsymbol{x}_j \quad (1)$$

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

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

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

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

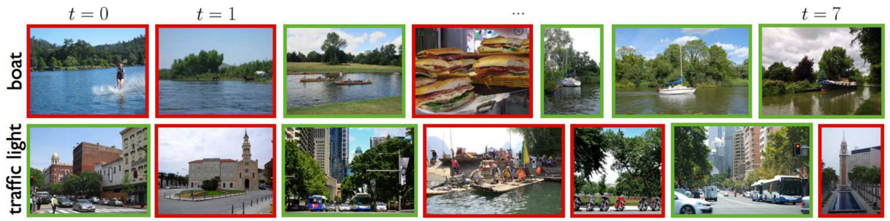


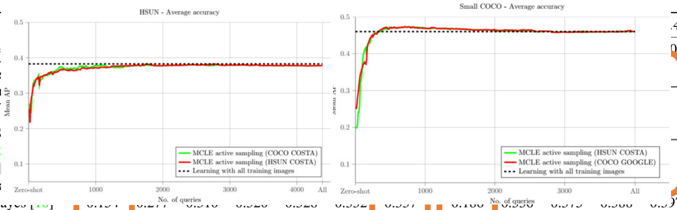
Proposition 1 (Maximum Conflict). *To maximize the objective Eq.* (1) *at time t, we should query the sample i^* such that* (a) *its label y_{i^*} has an opposite sign from its classification score at* $(t-1)$, *while* (b)) *the classifier score is as high as possible.*

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 <u>In Practice</u>

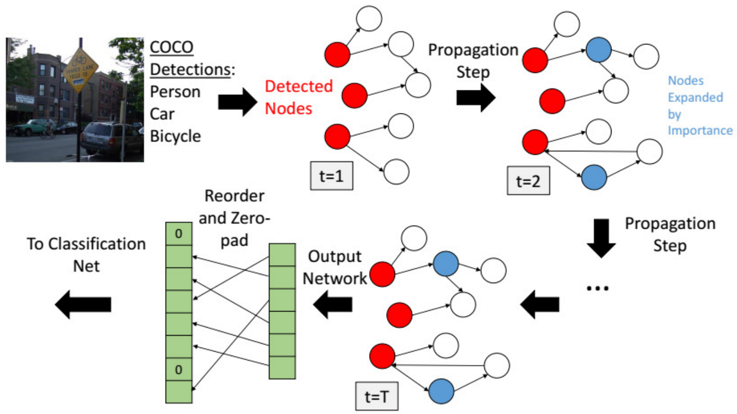




460 mAP)		
No. of querie	250	300
MCLE (Data:	0.448	0.460
MCLE (Exter	0.442	0.457
BBAL [33]	0.395	0.408
Hiearchical S	0.331	0.365
GP Mean [12	0.431	0.438
GP Variance	0.309	0.326
GP Impact B:	0.430	0.436
GP EMOC Bayes	0.399	0.405

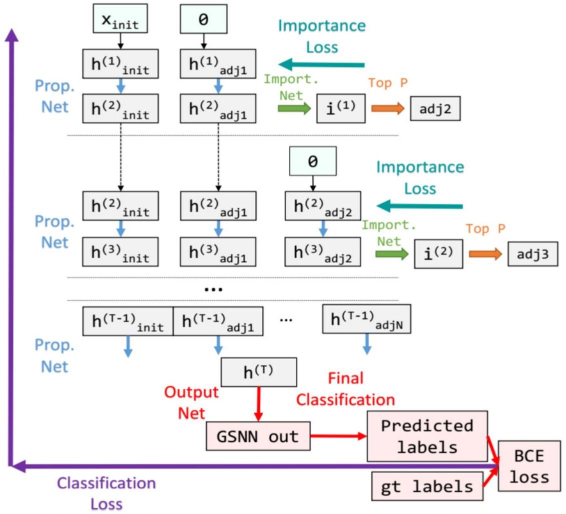
C{}DE https://github.com/stratisgavves/activetransferlearning or www.egavves.com

Using Knowledge Graphs for Novel QA



[1] The More You Know: Using Knowledge Graphs for Image Classification, Marino et al., CVPR 2017

Knowledge Graph QA: Model



[1] *The More You Know: Using Knowledge Graphs for Image Classification,* Marino et al., CVPR 2017

Knowledge Graph QA: Example

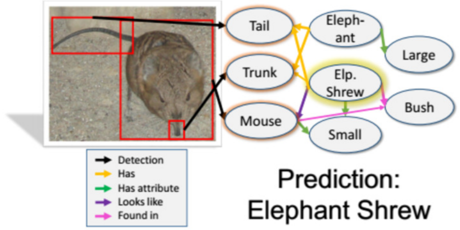
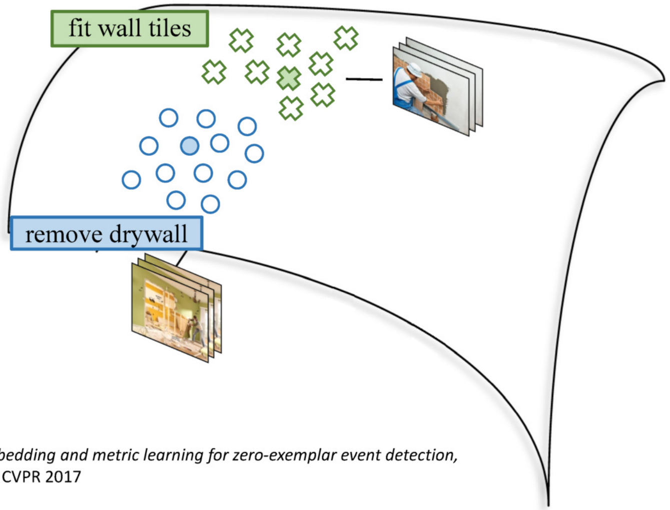


Figure 1. Example of how semantic knowledge about the world aids classification. Here we see an elephant shrew. Humans are able to make the correct classification based on what we know about the elephant shrew and other similar animals.

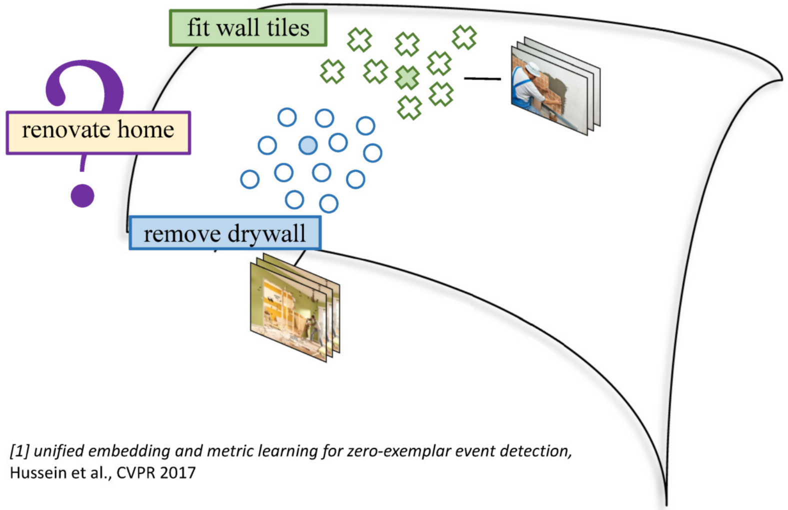
[1] *The More You Know: Using Knowledge Graphs for Image Classification,* Marino et al., CVPR 2017

Zero-exemplar Event Detection



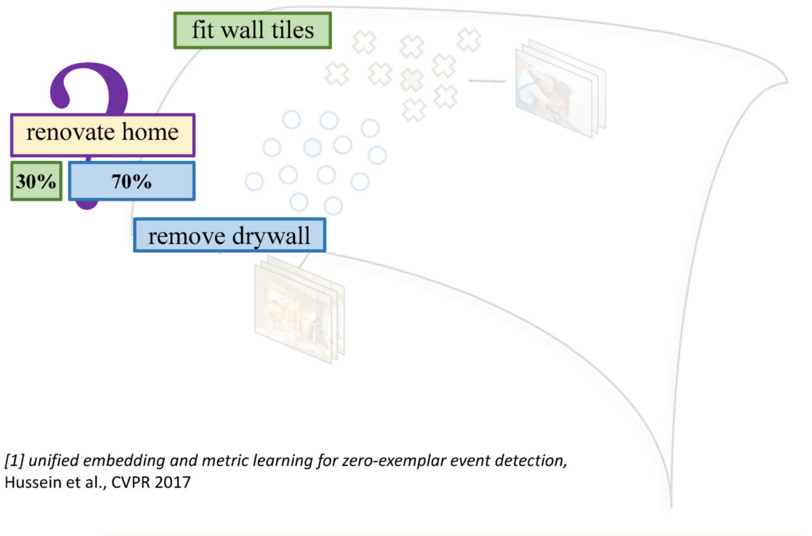
[1] *unified embedding and metric learning for zero-exemplar event detection,* Hussein et al., CVPR 2017

Zero-exemplar Event Detection



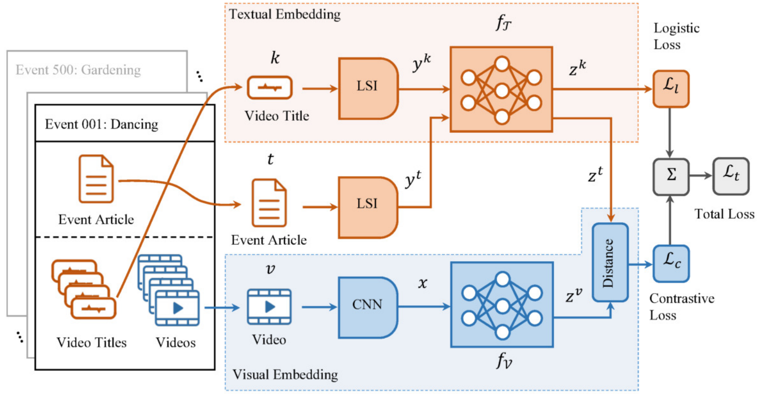
[1] *unified embedding and metric learning for zero-exemplar event detection,* Hussein et al., CVPR 2017

Zero-exemplar Event Detection



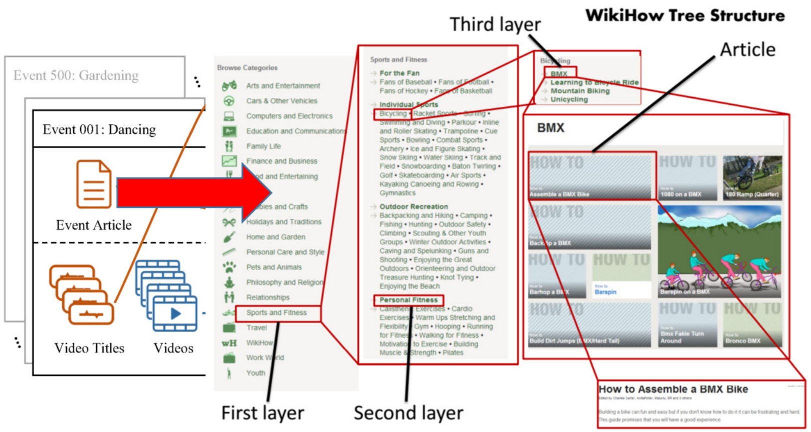
[1] unified embedding and metric learning for zero-exemplar event detection,
Hussein et al., CVPR 2017

Zero-exemplar Event Detection



[1] unified embedding and metric learning for zero-exemplar event detection,
Hussein et al., CVPR 2017

Zero-exemplar Event Detection

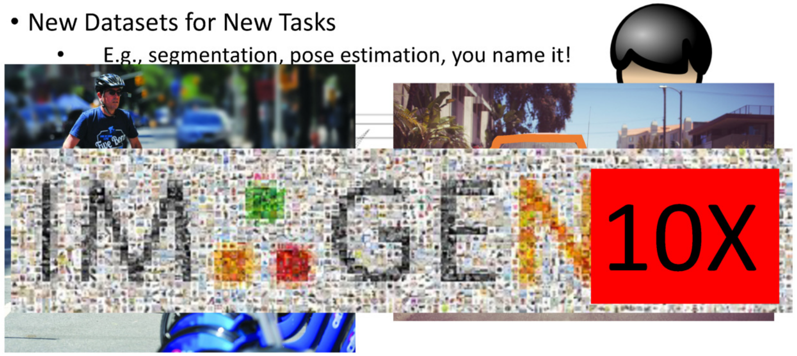


Conclusion

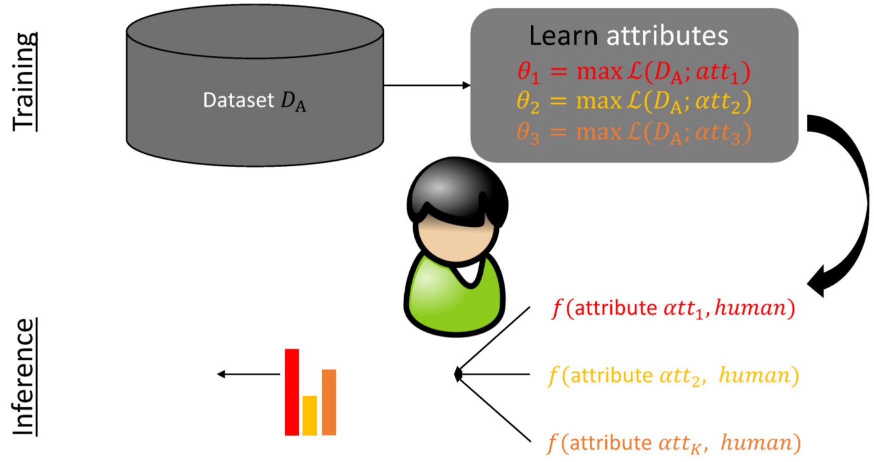
- Attributes not always perfect
  - Often there is no good attribute definition for classes
  - Often attribute prediction is not that reliable
- Knowledge transfer via external knowledge sources
  - Complex inferences about open-world questions
  - Make inferences beyond what static datasets can teach
  - Feature sharing via knowledge sharing
- Active interaction for practical zero-shot classification
  - Correct prediction mistakes through active learning
  - Guide novel attribute learning and knowledge transfer
  - Active Transfer Learning: Don't waste or throw your old datasets!!

Going to the next level

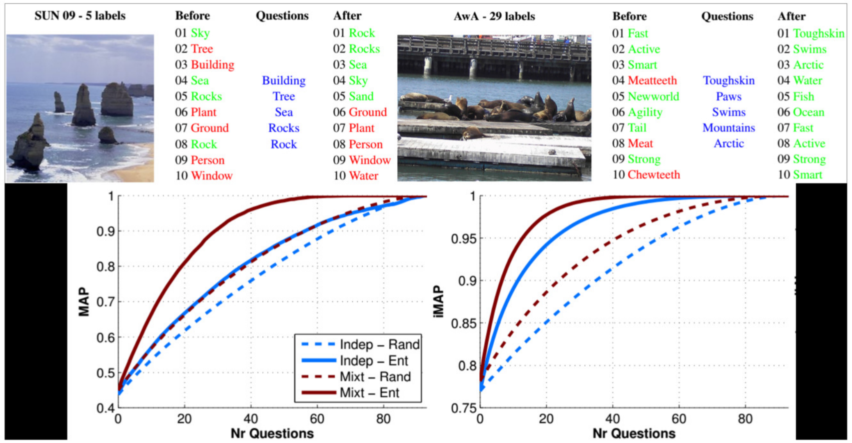
- Active Deep Learning for Zero-Shot Recognition
  - Deep learning of discriminative, repeatable attributes
- Truly diversified transfer from past to future tasks
  - Better transfer learning
- New Datasets for New Tasks
  - E.g., segmentation, pose estimation, you name it!



Active learning during inference



Tree-based Interactive Labelling



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

Universiteit van Amsterdam

max planck institut informatik

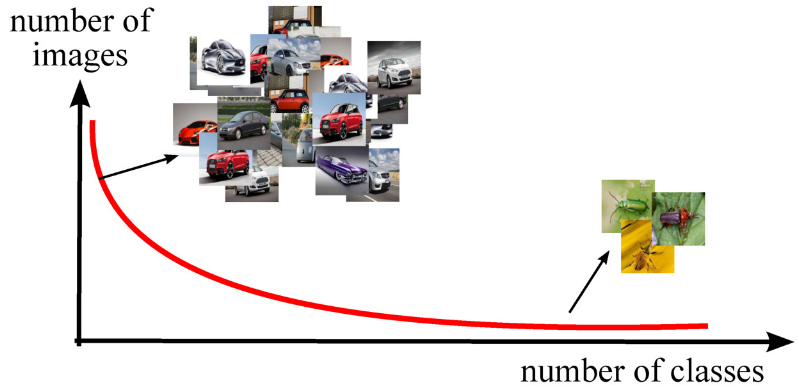
Open Problems in Zero-Shot Learning

Zeynep Akata

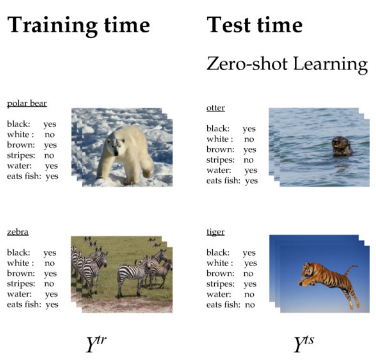
Zero-Shot Learning Tutorial, CVPR 2017

26 July 2017

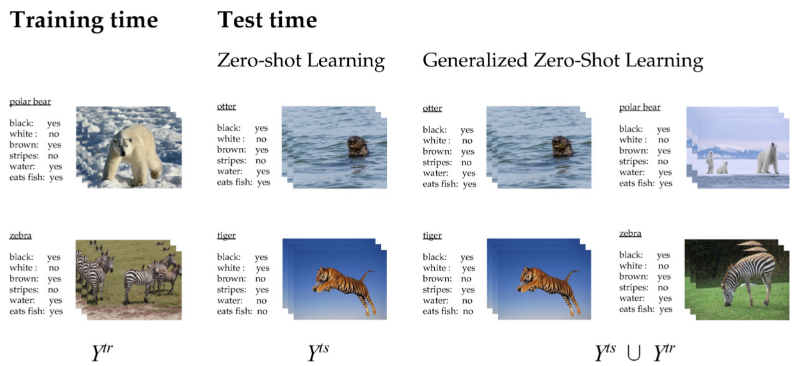
Data Distribution in Large-Scale Datasets



(Generalized) Zero-Shot Learning Setting



(Generalized) Zero-Shot Learning Setting



Evaluating GZSL

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

Evaluating GZSL

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

Zero-Shot Learning Models

Existing ZSL models can be grouped into 4:

Zero-Shot Learning Models

Existing ZSL models can be grouped into 4:

1. Linear Compatibility: `ALE`, `DEVISE`, `SJE`, `ESZSL`, `SAE`

Zero-Shot Learning Models

Existing ZSL models can be grouped into 4:

1. Linear Compatibility: `ALE`, `DEVISE`, `SJE`, `ESZSL`, `SAE`
2. Non-linear Compatibilty: `LATEM`, `CMT`

Zero-Shot Learning Models

Existing ZSL models can be grouped into 4:

1. Linear Compatibility: `ALE`, `DEVISE`, `SJE`, `ESZSL`, `SAE`
2. Non-linear Compatibilty: `LATEM`, `CMT`
3. Two-stage Inference: `DAP`, `CONSE`

Zero-Shot Learning Models

Existing ZSL models can be grouped into 4:

1. Linear Compatibility: `ALE`, `DEVISE`, `SJE`, `ESZSL`, `SAE`
2. Non-linear Compatibilty: `LATEM`, `CMT`
3. Two-stage Inference: `DAP`, `CONSE`
4. Hybrid Model: `SYNC`

[Akata et.al IEEE CVPR 2013, Frome et.al. NIPS 2013, Akata et. al. 2015, Romera Paredes and Torr ICML 2015, , Kodirov et.al IEEE CVPR 2017, Xian et.al. IEEE CVPR 2016, Socher et.al. NIPS 2013, , Lampert et.al. IEEE CVPR 2009 & TPAMI 2013, Norouzi et.al. ICLR 2014, Changpinyo et.al. IEEE CVPR 2016]

Datasets Used for Evaluation

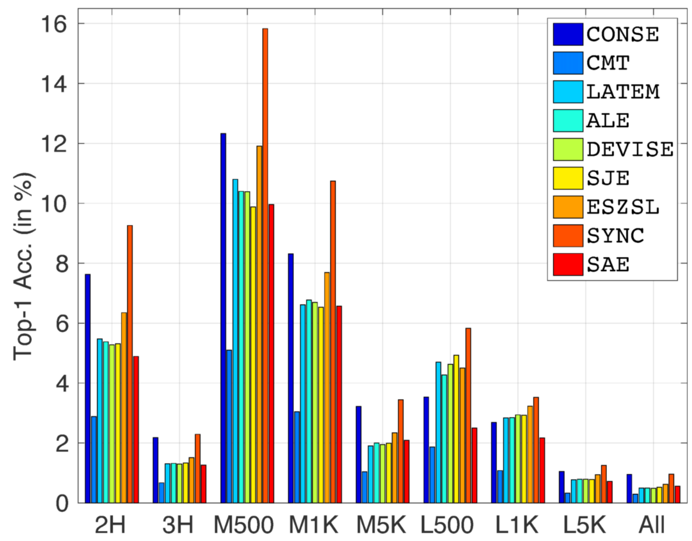
Dataset	Size	$\|\mathcal{Y}\|$	$\|\mathcal{Y}^{tr}\|$	$\|\mathcal{Y}^{ts}\|$
SUN	$14K$	717	$580+65$	72
CUB	$11K$	200	$100+50$	50
AWA1	$30K$	50	$27+13$	10
AWA2*	$37K$	50	$27+13$	10
aPY	$1.5K$	32	$15+5$	12

Datasets Used for Evaluation

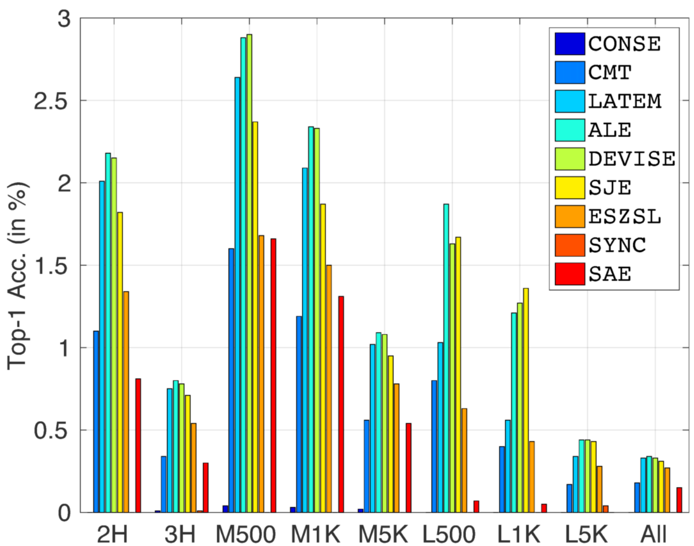
Dataset	Size	$\lvert\mathcal{Y}\rvert$	$\lvert\mathcal{Y}^{tr}\rvert$	$\lvert\mathcal{Y}^{ts}\rvert$
SUN	$14K$	717	$580 + 65$	72
CUB	$11K$	200	$100 + 50$	50
AWA1	$30K$	50	$27 + 13$	10
AWA2*	$37K$	50	$27 + 13$	10
aPY	$1.5K$	32	$15 + 5$	12

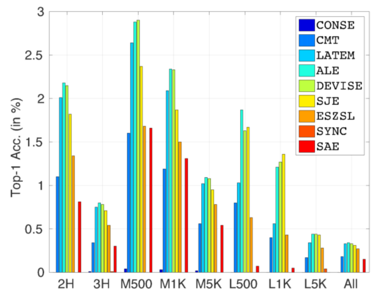
ImageNet Split	$\lvert\mathcal{Y}^{ts}\rvert$
ImageNet 21K - $\mathcal{Y}^{tr}$	20345
Within 2/3 hops from $\mathcal{Y}^{tr}$	1509/7678
Most populated classes	$500/1K/5K$
Least populated classes	$500/1K/5K$

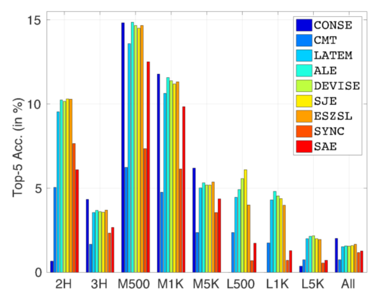
Motivating GZSL Setting on ImageNet

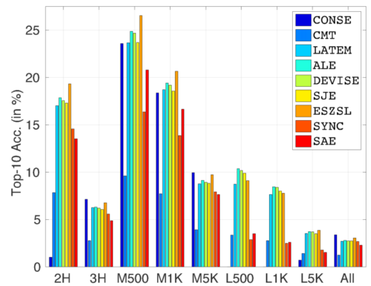


GZSL Results on ImageNet

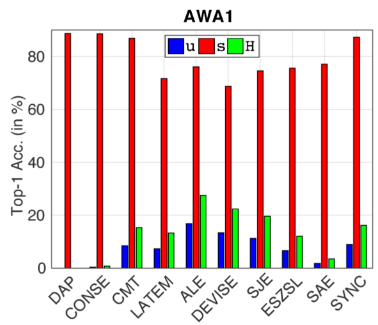


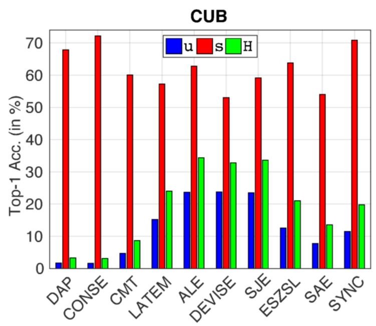






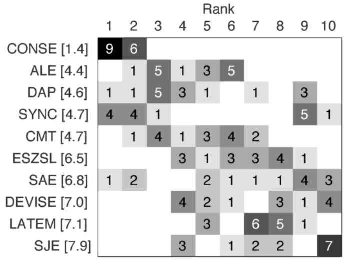
Zooming Into GZSL Performance



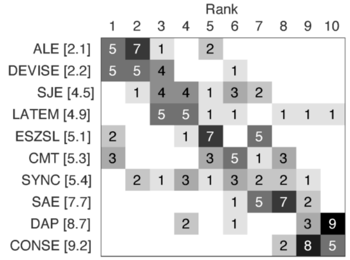


$$\mathbf{u} = \frac{1}{\|\mathcal{Y}^{ts}\|} \sum_{c=1}^{\|\mathcal{Y}^{ts}\|} \frac{\#\text{ correct in c}}{\#\text{ in c}} \text{ and } \mathbf{s} = \frac{1}{\|\mathcal{Y}^{tr}\|} \sum_{c=1}^{\|\mathcal{Y}^{tr}\|} \frac{\#\text{ correct in c}}{\#\text{ in c}}$$

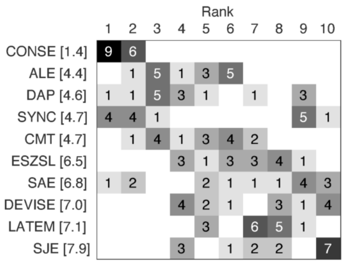
$$\mathbf{H} = \frac{2 * acc_{\mathcal{Y}^{tr}} * acc_{\mathcal{Y}^{ts}}}{acc_{\mathcal{Y}^{tr}} + acc_{\mathcal{Y}^{ts}}} = \frac{2 * \mathbf{s} * \mathbf{u}}{\mathbf{s} + \mathbf{u}}$$



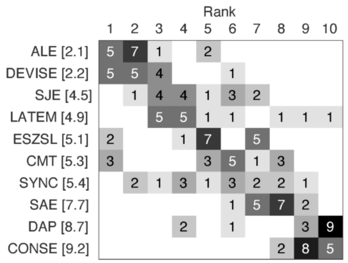
Seen Class Accuracy



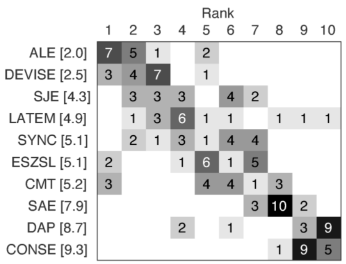
Unseen Class Accuracy



Seen Class Accuracy



Unseen Class Accuracy



Harmonic mean

Conclusions

In Generalized Zero-Shot Learning

1. The setup is challenging but more practical

Conclusions

In Generalized Zero-Shot Learning

1. The setup is challenging but more practical
2. Unseen images embedded close to seen classes

Conclusions

In Generalized Zero-Shot Learning

1. The setup is challenging but more practical
2. Unseen images embedded close to seen classes
3. Results much lower than ZSL: Room for improvement

[Xian et.al. IEEE CVPR 2017 & ArXiv 2017]

Thank you!

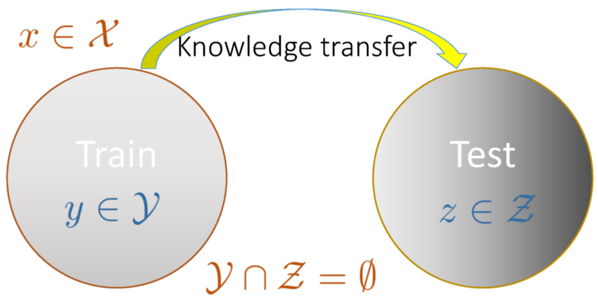
Zero-Shot Learning

for Computer Vision

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. Classification
2. Localization
3. Retrieval
4. Open Problems
4. Conclusion

Zero-Shot Classification

- Mathematically ALE and DAP are similar
- ALE directly optimizes image classification
- Zero-Shot using pre-trained classifiers
  - Indirect attribute prediction
  - Word2vec, Co-occurrence statistics
  - ALE, DEVISE, SJE, ESZSL, SAE, LATEM, CMT, CONSE, SYNC
- Evaluate, evaluate, evaluate!

Zero-Shot with Localization

- Attributes belong to objects, not images
- Zero-Shot localization is a natural extension to the problem
- Focus on visual Details or Regions
  - Each with their merit, depends on application
  - Maybe a smart combination?
- Localization in <u>images and videos</u> using natural language queries is possible and promising
  - Offers also a great evaluation framework for image captioning, visual question answering

Zero-Shot Retrieval

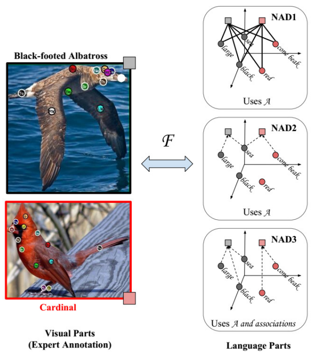
- Zero-shot retrieval profits from semantic alignment
- Learnable from freely available online sources
- Better than low- and mid-level alternatives
- Adds meaning and recounting to retrieval results

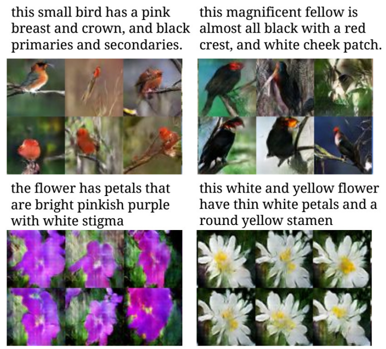
- Next challenge:
- Spatiotemporal search and alerts for live video

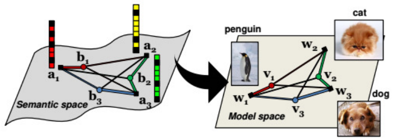
Open Problems

- The evaluation of zero-shot classifiers is very important!
  - Thankfully, now there is a benchmark to compare against
  - Zero-Shot Learning - The Good, the Bad and the Ugly, Xian et al., CVPR 2017
  - 12 models compared in 6 datasets
- Generalized Zero-Shot Learning
  - More challenging, more practical!
  - Unseen images embedded close to seen classes
- How to optimally exploit knowledge graphs to answer novel QA?
- 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

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!

Slides will be added online later at the website:
https://staff.fnwi.uva.nl/t.e.j.mensink/zsl2017/