Emerging Topics in Learning from Noisy and Missing Data: Zero Shot Learning

Zero-Shot Learning

Aka: *How to recognize/retrieve concepts for which all visual data is missing.*

Timothy Hospedales

University of Edinburgh

Queen Mary University of London

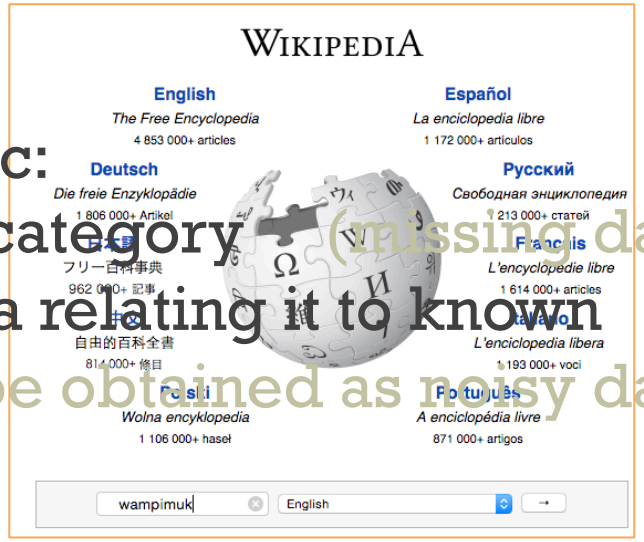
Emerging topics in learning from noisy and missing data.
Tutorial @ ACM Multimedia 2016

What is Zero-Shot Learning?

What is Zero-Shot Learning?

Live Demonstration!

- Audience Task: Recognise the Wampimuk.
  - Impossible?
- Solution: Semantic Transfer
  - Domain Ontology:
    - Wampimuk := small, horns, furry, cute
  - Wikipedia Page:



Zero-Shot problem spec:

1. Recognize an novel category (missing data!)
2. …Via some metadata relating it to known categories (can be obtained as noisy data!)

Why Care about Zero-Shot Machine Learning?

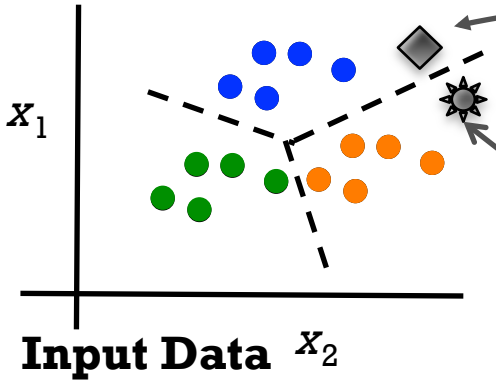
- Vast and growing number of categories: Depth and Breadth.
  - Collecting and particularly annotating examples is impossible
    - Especially deep learning scale
    - New categories always emerging

- Examples:
  - Object Recognition
  - fMRI mind reading
    - Every annotation is a brain-scan
    - Every word is a category

- Humans can do it, but (classic) supervised learning can't!

From Supervised to Zero-Shot Pattern Recognition

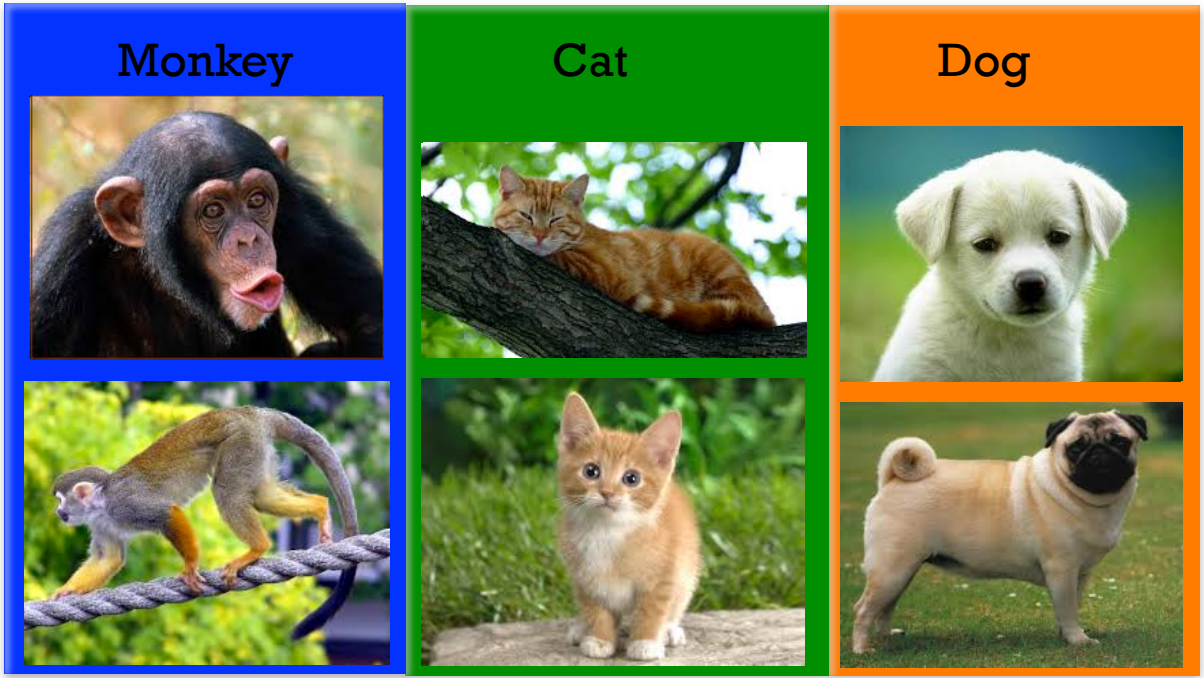
Output Labels:
Dog
Cat
Monkey



Wampimuk

Unicorn

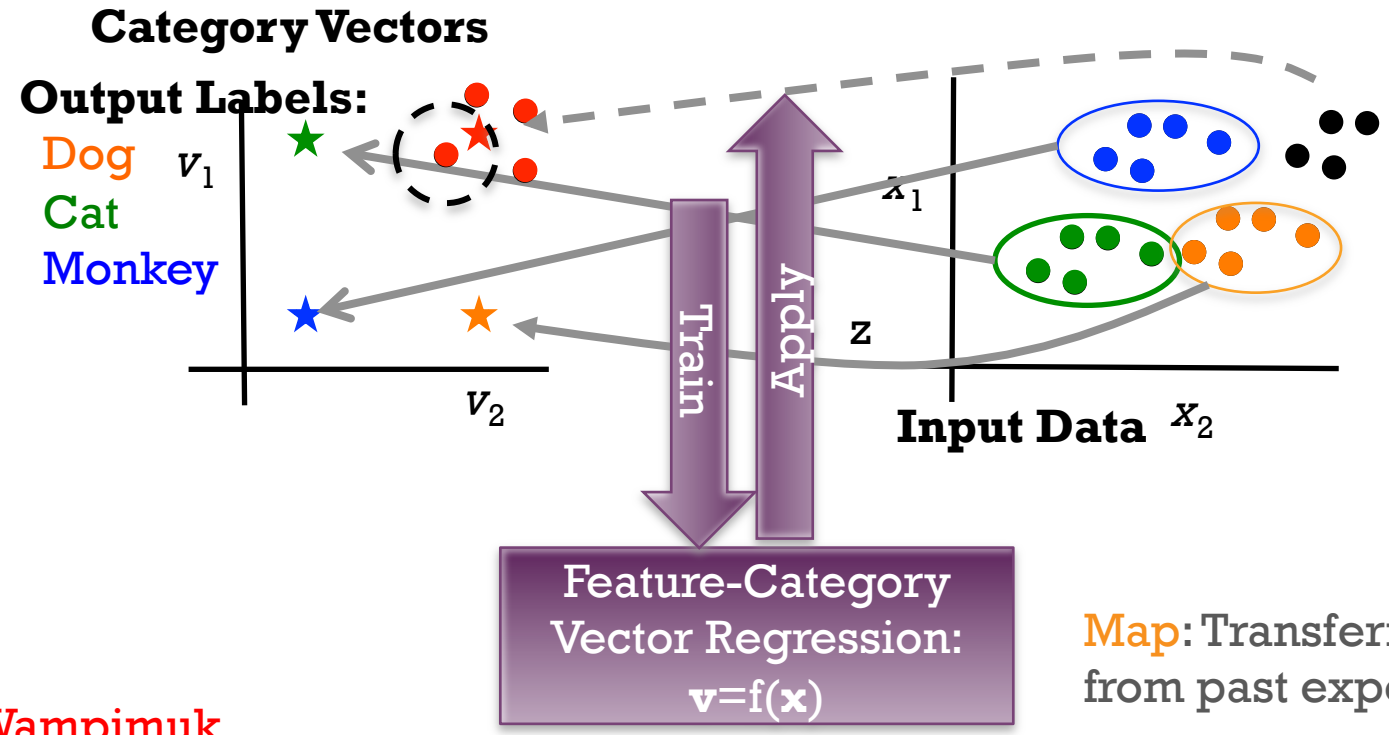
???



From Supervised to Zero-Shot Pattern Recognition

Key is to embed categories as vectors

Task: Transferred from external metadata source



Map: Transferred from past experience

Wampimuk

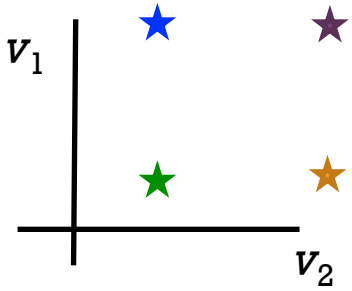
Feature-category vector map can generalize to new categories.

Where to get Category Vectors?

Lampert, CVPR'09
Akata, PAMI'16
Rohrbach, CVPR'11

- "Supervised" sources:
  - Class-attribute vector
  - Vector encoded location in a taxonomic class hierarchy
    - Pros: +Accurate if informative by design
    - Cons: -Manual annotation effort

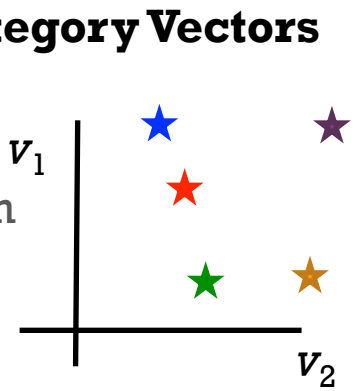
Category Vectors

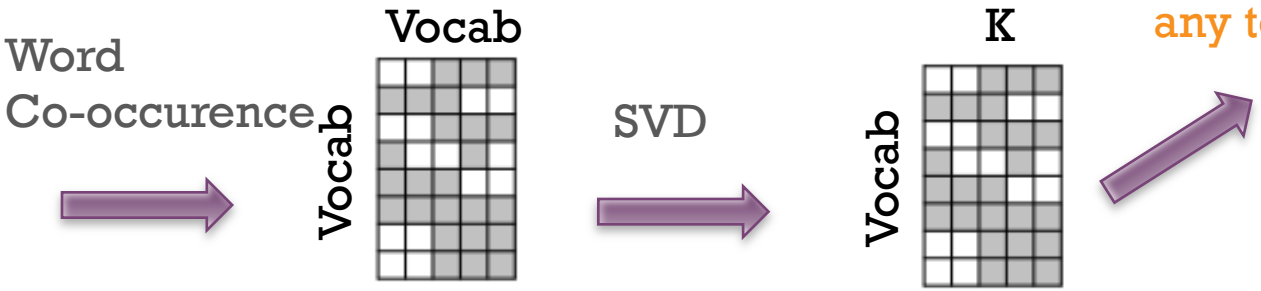


Where to get Category Vectors?

Akata, PAMI'16
Frome, NIPS'13
Rohrbach, CVPR'11

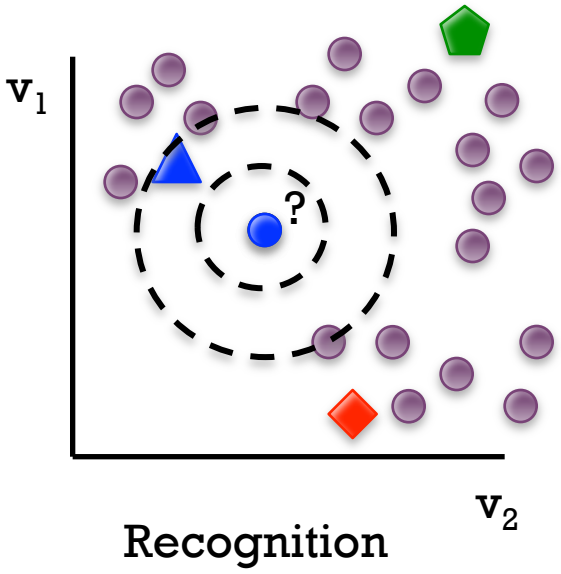
- "Unsupervised" sources: Existing unstructured data
  - Word (token) co-occurrence.
    - OR: word2vec: Representation of adjacent word prediction DNN.
    - => Automatic vector for any nameable category.
  - Pros/Cons:
    - +Automatic/Free. –Maybe less informative than attributes.
    - Can be better if trained on a task relevant corpus.
  - Search-Engine hit co-occurrence, etc.

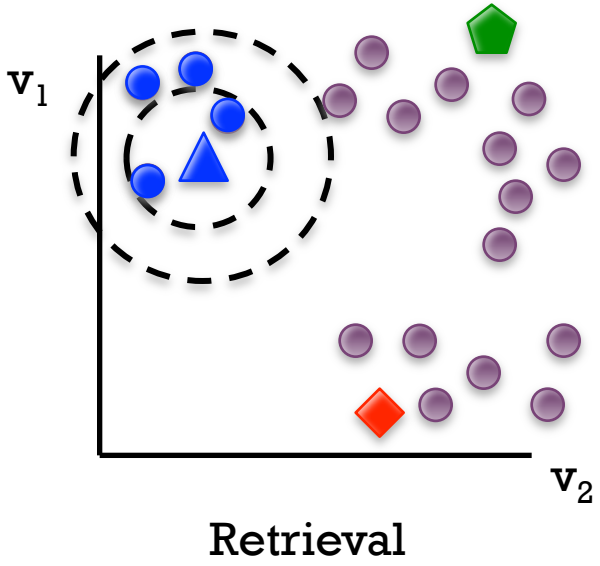




Zero Shot: Recognition versus Retrieval

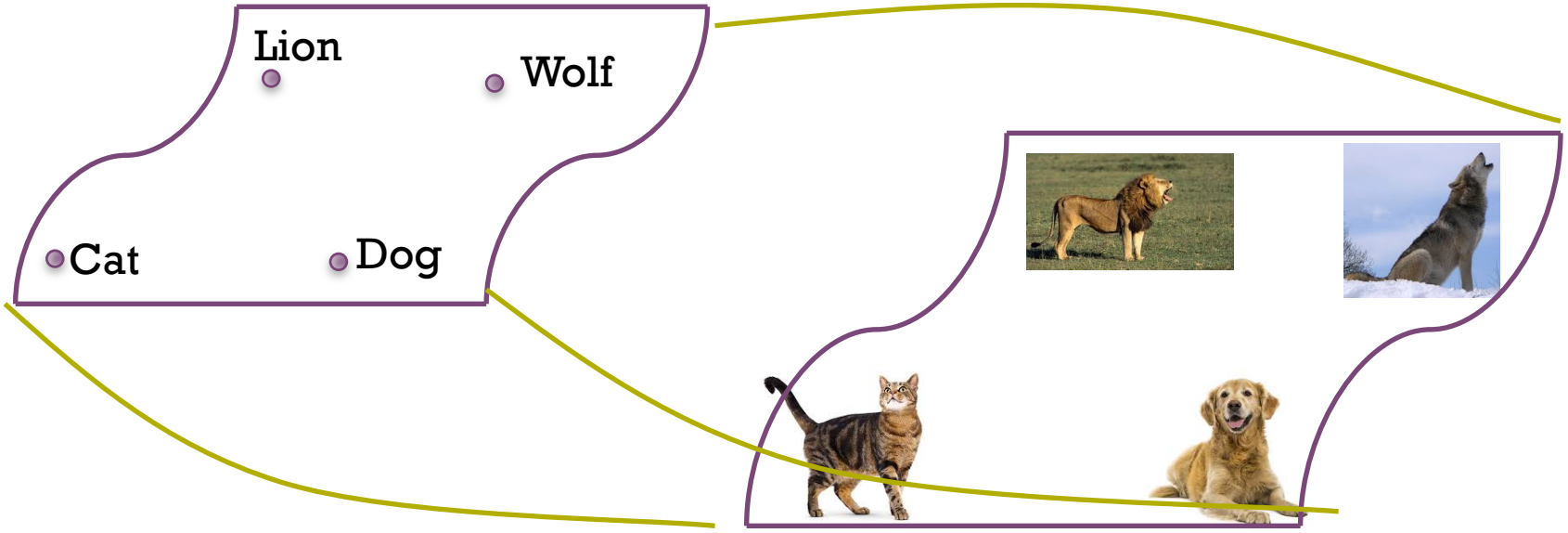
- Zero-shot Learning can address both Recognition + Retrieval
  - Once images are projected into category vector space…
    - Recognition: Given an image, what's the nearest category vector?
    - Retrieval: Given a category vector, what's the nearest image?





Why Does it Work?
Any theory/intuition?

- Very little theory.
- Intuition:
  - IF training category vectors and image vectors lie in the same relative positions on their respective manifolds….
  - Then a few examples can establish correspondence between the two spaces
  - And any new category can also be recognized.

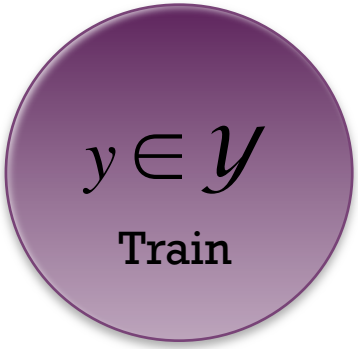


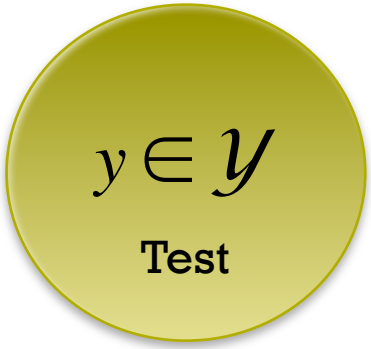
More formal problem statement

- Supervised Learning

Given Data: $x \in \mathcal{X}$, $y \in \mathcal{Y}$

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





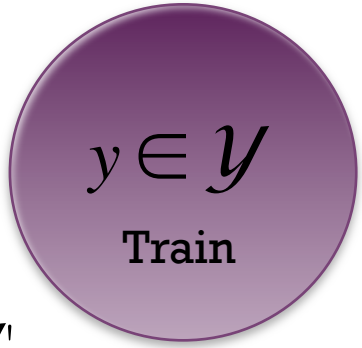
- Zero-Shot Learning

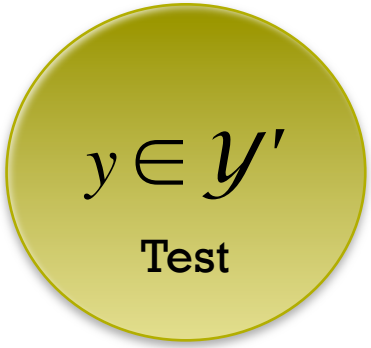
Given Data: $x \in \mathcal{X}$, $y \in \mathcal{Y}$

Learn: $f : \mathcal{X} \rightarrow \mathcal{Y}'$

Or: $f : \mathcal{X} \rightarrow \mathcal{Y} \cup \mathcal{Y}'$

Where: $\mathcal{Y} \cap \mathcal{Y}' = \varnothing$





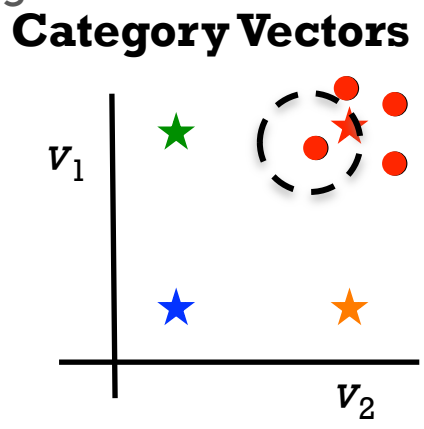
Assuming: $\mathbf{v}_y \in \Re^D$

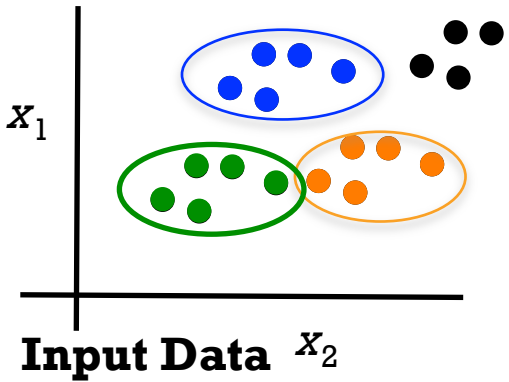
How Do I Implement Zero-Shot Learning?

ZSL Formalisation #1/3: Regression/Classification

- 1. Obtain category vectors, **v**.
  - Attribute description (Wampimuk := small, cute, furry, horns)
  - Word-vector (Digested wikipedia co-occurrence count)
- 2. Train:
  - Given some known class-category vectors **v** and images **x**:
  - Learn image -> category vector classifiers/regressors **v=f(x)**.
    - E.g., SVM / OLS,SVR. Deep Neural Net.
- 3. Test:
  - Specify vec **v*** for new class to recognise
  - Map test data **f(x*)** to cat vec space
  - NN matching of **v*** vs **f(x*)**
- Pros/Cons:
  - + Easy and fast!
  - - Category separability

Lampert, CVPR'09
Socher, NIPS'13
Xu, ECCV'16





ZSL Formalisation #2/3: Energy Function Ranking

- Given: Training image data **x**, category vectors **v**.
- 1. Train an energy function $E_W(\mathbf{x},\mathbf{v})$.
  - E.g. Bilinear: $E_W(\mathbf{x},\mathbf{v})=\mathbf{x}^T W\mathbf{v}$.
  - W such that $E_W(\mathbf{x},\mathbf{v})$ is large when data and category match. **x=v**.
  - W such that $E_W(\mathbf{x},\mathbf{v})$ is small when data and category mismatch. **x!=v**.
  - E.g., Max margin ranking objective.
- 2. Testing: Classify example **x*** that may be a novel class.
  - Consider vectors **v*** for classes to recognize.
  - Evaluate $E(\mathbf{x}^*,\mathbf{v}^*)$ for each **v***.
  - Max response gives classification. $\text{argmax}_{v^*}\ E(x^*,v^*)$

Frome, NIPS'13
Akata, PAMI'16
Yang, ICLR'15
Romera-Paredes, ICML'15

- Pros/Cons
  - + Train for separability => higher accuracy.
  - - More complex and slower optimisation (Except [Romera-Paredes, ICML'15])

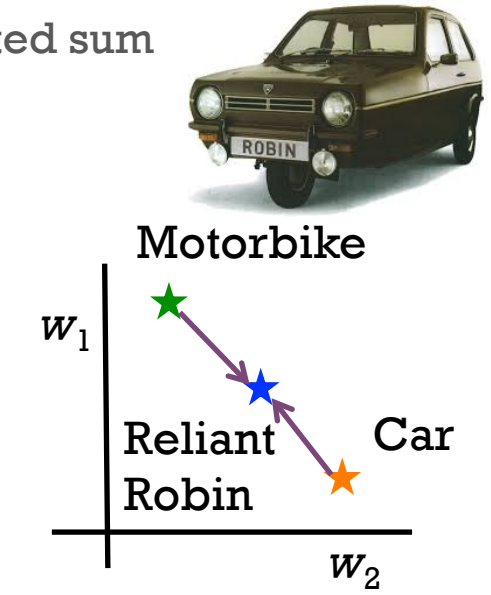
ZSL Formalisation #3/3: Model Averaging

Norouzi, ICLR'14
Mensink, CVPR'14

- Given: Training image data **x,y.** Category vectors **v**.
- 1. Train a recognition model for each known category
  - E.g., Linear model $\mathbf{w}_c$ for each category **c**.
- 2. Testing
  - To recognize a novel category **c'**
  - Generate a recognition model $\mathbf{w}_{c'}$ as the similarity weighted sum of known models. E.g.,

$$\mathbf{w}_{c'} = \sum_{c} s(c,c')\mathbf{w}_c$$

  - Makes use of inter-category similarity matrix **s**.
    - Can be determined by category vector similarity.
- Pros/Cons:
  - + Easy and fast!
  - - Linear models only?



Summary So Far…

- We have methodology to recognize visual concepts with completely missing training data.
  - Class embeddings from noisy metadata replace manual annotation.
- Asides:
  - Straightforward application to other data types: video, text, audio, etc.
  - Applicable to other settings:
    - Single -> Multi-label annotation
    - Zero-shot classification -> zero-shot regression

Summary So Far…

- We have methodology to recognize new concepts
- But:
  - Not comparable accuracy to state of the art supervised learning
- => Research directions: Improving performance:
  - Cross-dataset transfer / data augmentation
  - Transductive inference + Semi-supervised learning
  - Deep zero-shot learning
  - Beyond vector class embeddings
- => Research directions: Related Problem Variants
  - Zero-shot domain adaptation

Current Research Directions

Cross-Dataset Transfer / Data Augmentation

- Consider zero-shot recognition of a given dataset. E.g., HMDB51/UCF101/TRECVID video datasets.

Cross-Dataset Transfer / Data Augmentation

- Consider zero-shot recognition of a given dataset. E.g., HMDB51/UCF101/TRECVID video datasets.
  - In supervised learning, performance boosts by knowledge sharing through cross-class transfer learning methodologies. (Since different datasets have different classes)
- In zero-shot learning, the Feature=>Embedding=>Class breakdown means no special methodology necessary.
  - Simply aggregate datasets when learning the embedding.
- Possible due to no requirement for shared label space in ZSL.

Cross-Dataset Transfer / Data Augmentation

Xu, arXiv'16
Xu, ECCV'16

- Cross dataset transfer by Data Augmentation

HMDB51 Dataset

UCF101 Dataset

Regression Model

Vec('Dribble') Vec('Ride Horse') Vec('Archery') Vec('Apply Makeup')

Train

Olympic Sport Dataset

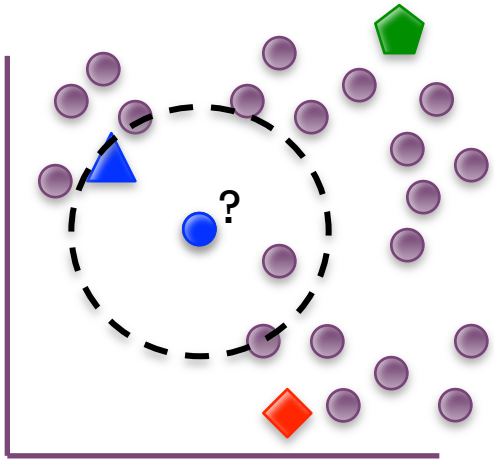
Vec('Diving')

Zero-Shot Test

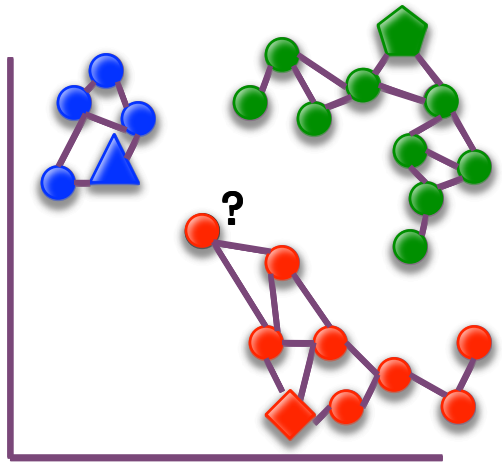
- Aggregate multiple train datasets => Better embedding.
  - More data => Bigger performance bump than fancy methodology improvements
  - Complementary to other methodology improvements

Transductive Inference

- At test time ZSL has the challenge of 1-shot-learning.
  - NN matching in label-vector space: category vector is one labeled point
- Solution: Borrow transductive inference methods.
  - Transductive label propagation on the manifold of unlabeled test data.



1-NN Recognition



Manifold Label Propagation

Transductive Inference

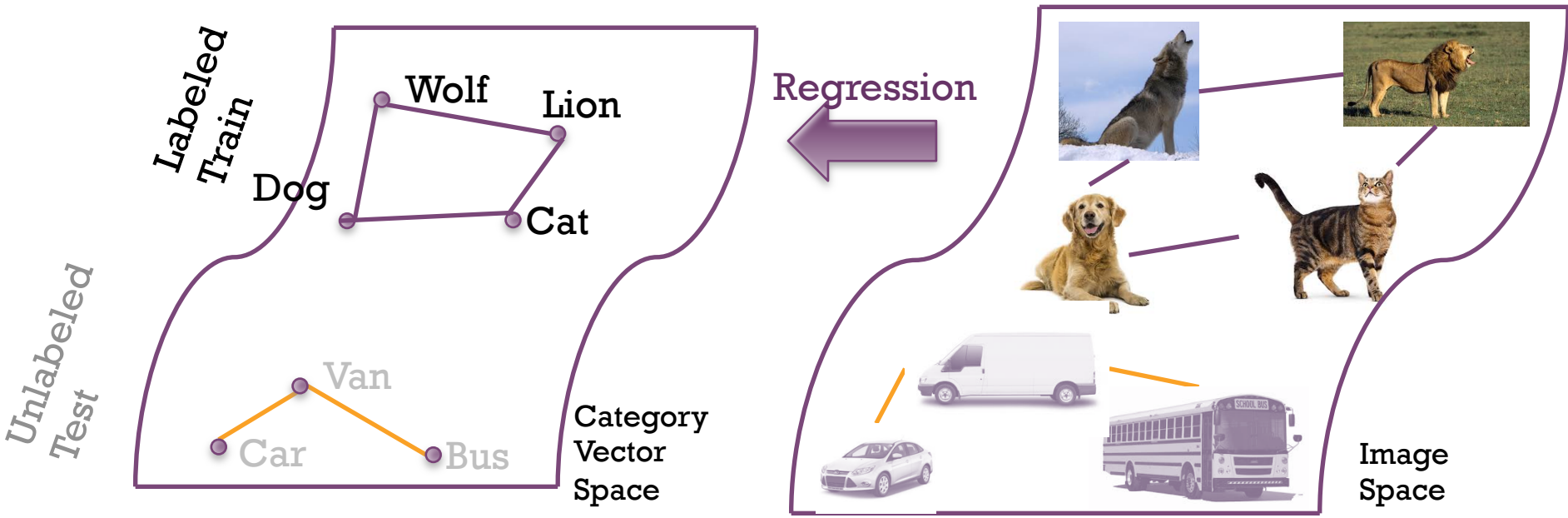
Rohrbach, NIPS'13
Fu, PAMI'15
Xu, arXiv'16

- Given:
  - Initial label posterior for all N test points belonging to class *y*: L_y
  - KNN graph matrix of size N^2, W and its graph Laplacian S
- Iterate: $L_y = (I - \alpha S)^{-1} L_y$
- Results:
  - Key contributor in our big win on the AwA ZSL benchmark.
    - From ~44% (2009-2014) to 80% accuracy (2015). [Fu, PAMI'15]
  - Transductive post-processing is complementary to other strategies [Xu, arXiv'16]
- Caveat: Transductive requirement. Need to have a batch of test data.

Transductive Semi-Supervised Learning

Xu, arXiv'16
Belkin, JMLR'06

- Previously we exploited unlabeled data at test time.
  - Can also exploit unlabeled data when training the visual-category map.
  - => Semi-supervised manifold regression
  - Intuition: For a good regression, image/video instances that are close, should make (category embedding) predictions that are close.
    - Unlabeled data can impact the regression with this regulariser.



Transductive Semi-Supervised Learning

Xu, arXiv'16
Belkin, JMLR'06

- Kernel regression from image, x to category vector y.
  - With NN graph among images, W.

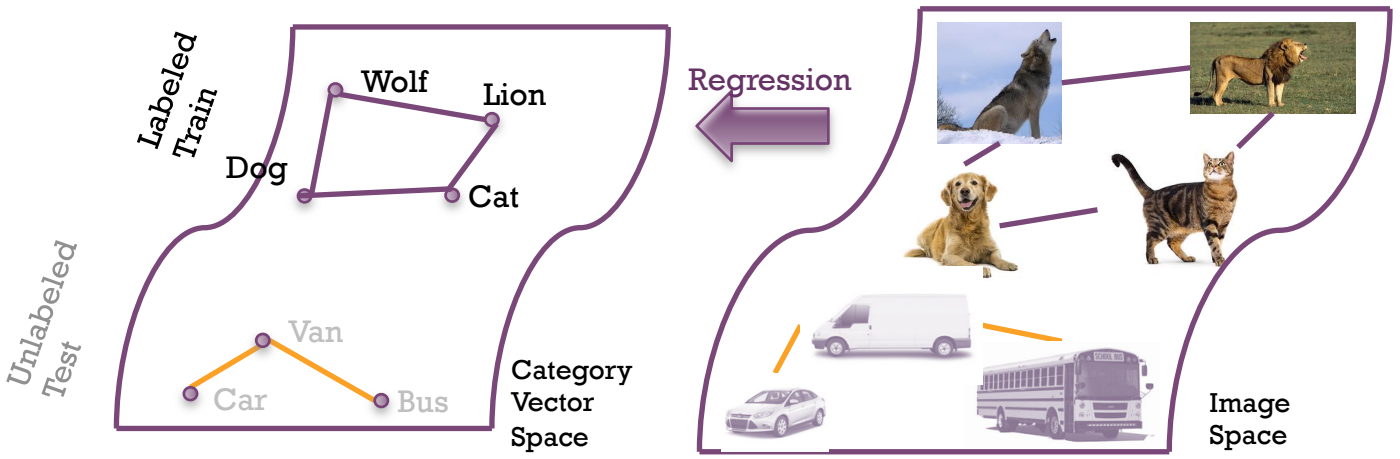
Ridge Regulariser

Manifold Regulariser

$$f^* = \arg\min_f \frac{1}{L}\sum_{i=1}^{L} L(\mathbf{x}_i, \mathbf{y}_i; f) + \lambda \|f\|^2$$

$$f^* = \arg\min_f \frac{1}{L}\sum_{i=1}^{L} L(\mathbf{x}_i, \mathbf{y}_i; f) + \lambda \|f\|^2 + \frac{\gamma}{(L+U)^2}\sum_{i,j=1}^{L+U} W_{ij}(f(x_i) - f(x_j))^2$$

  - => Closed form KRR solution as Laplacian Regularized Least Squares



Zero-Shot as a Prior for Few-Shot Learning

- Sometimes we have sparse but non-zero data
  - How to synergistically combine sparse data and a category vector?
- Use ZSL methods that can generate a linear classifier **w'** in $f : \mathcal{X} \rightarrow \mathcal{Y}'$
  - E.g., [Mensink, CVPR'14, Yang ICLR'15]
  - Use the ZSL linear classifier **w'** as a regularization target for few-shot learning

$$\min_{\mathbf{w}} L_{\mathbf{w}}(\mathbf{x}, y) + \|\mathbf{w}\| \quad \Rightarrow \quad \min_{\mathbf{w}} L_{\mathbf{w}}(\mathbf{x}, y) + \|\mathbf{w} - \mathbf{w}'\|$$

- Better performance than ZSL or few-shot learning alone.

Akata, PAMI'16
Mensink, CVPR'14

Deep Zero-Shot Learning

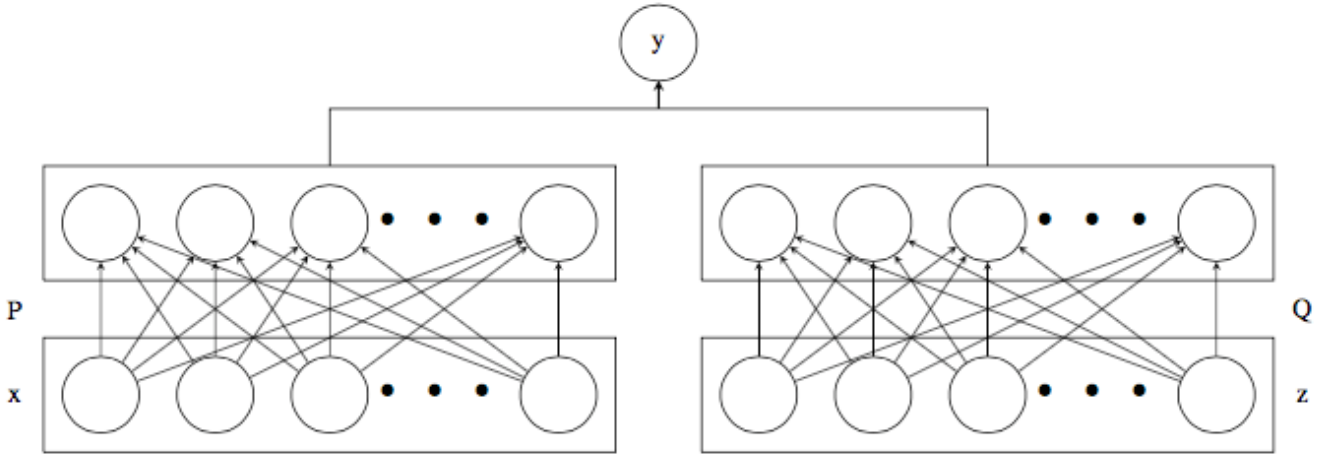
Frome, NIPS'13
Yang, ICLR'15
Ba, ICCV'15

- So far we've talked about shallow models.
  - How to do end-to-end zero-shot learning in deep models?
- Recall the energy function approach
  - Train $E(x,z)$ to be large for matching pairs, small for mismatching pairs.
  - Let $E(x,z)$ be a deep network rather than bilinear model.
- Simplest realization:
  - Concatenate $[x,z]$ and feed into a deep network.
  - Better: Do some representation learning on x and z, then inner product
    [Frome NIPS'13, Yang ICLR' 15]

Deep Zero-Shot Learning

Frome, NIPS'13
Yang, ICLR'15
Ba, ICCV'15

- A simple deep network for ZSL.



Data Input

Category Vector

- Train a max-margin ranker. Or y={1,0} for {matching,mismatching} pairs.

$$y = \sigma(P\mathbf{x})\sigma(Q\mathbf{z})^T$$

Another interpretation

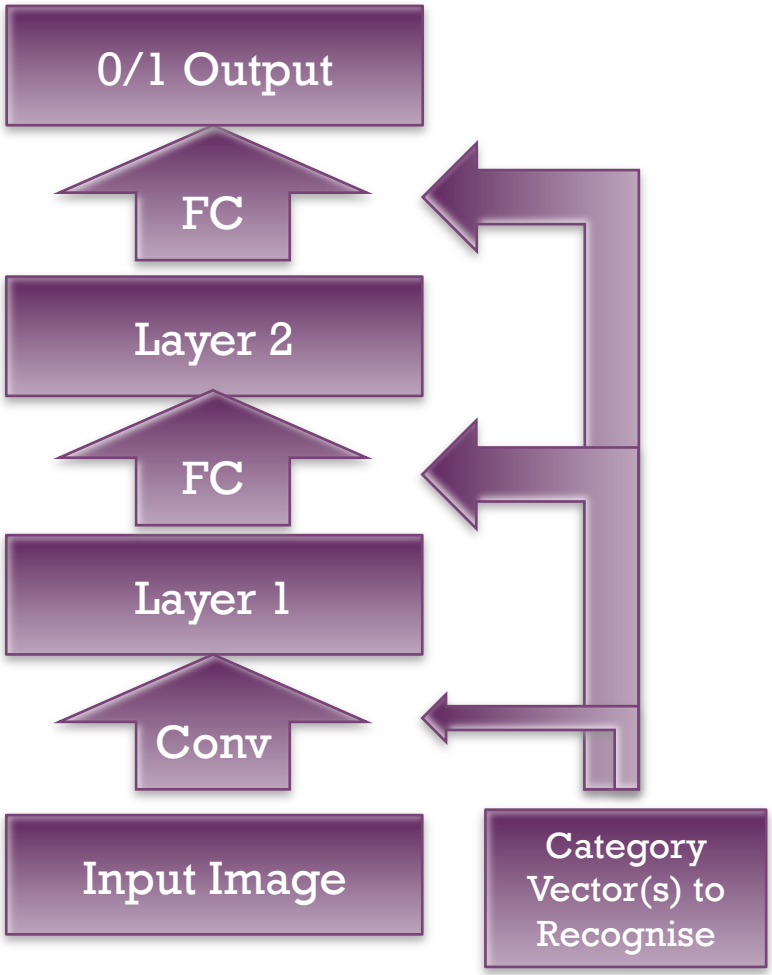
$$y = \mathbf{w}^T\sigma(P\mathbf{x})$$
$$\mathbf{w} = \sigma(Q\mathbf{z})$$

Dynamically synthesise weights to recgonise a concept given its category vector

Deeper Zero-Shot Learning

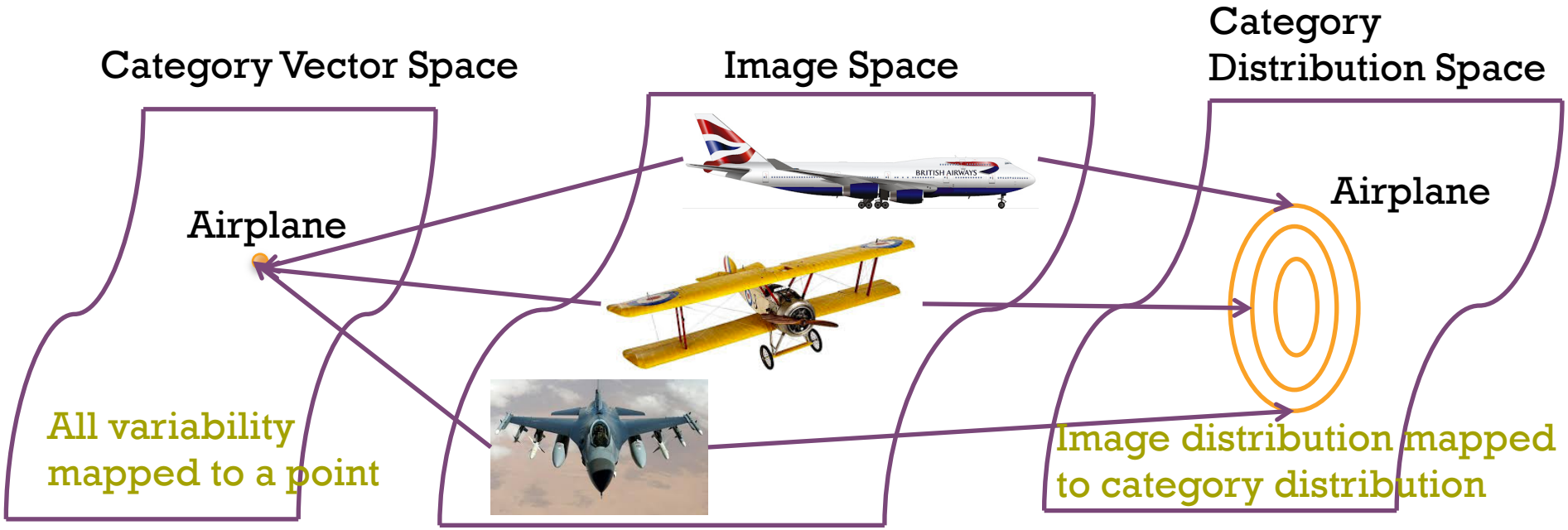
- Previously: Synthesise one weight vector for final layer.
- Now: Synthesise weight matrix (FC layers) and tensor (Conv layers) at every layer for CNN recognition of required categories.

Yang, arXiv'16a
Yang, arXiv'16b



Beyond Vector Embeddings

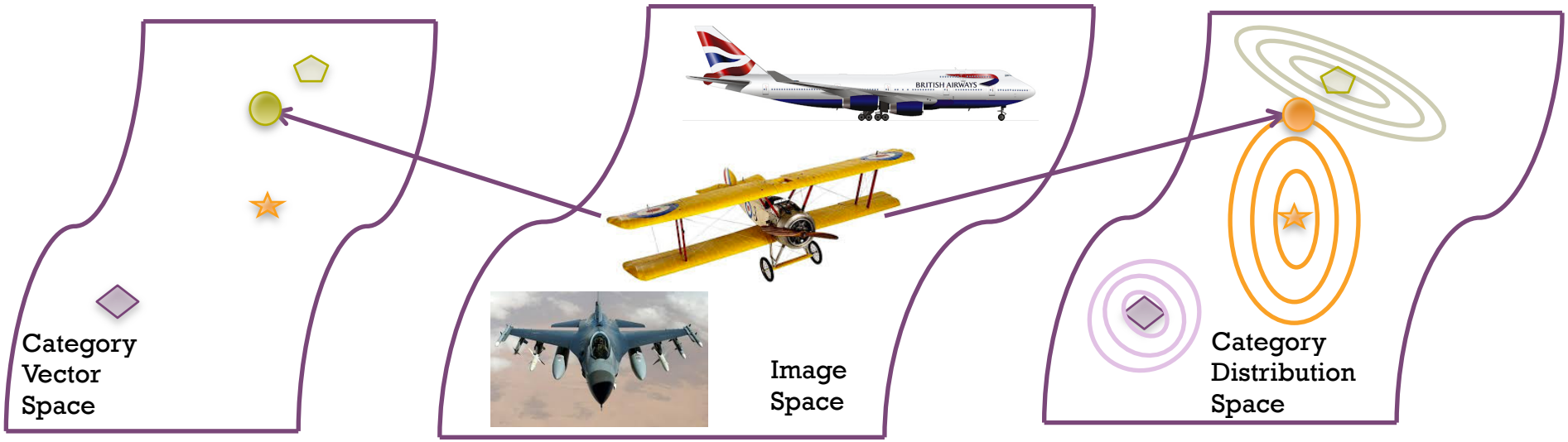
- So far we embedded each category as a vector.
  - Word co-occurrence, attribute vectors, etc.
  - But… other representations may provide better embeddings
- Consider embedding categories as distributions.
  - Categories now have width and shape to model intra-category variability.



Beyond Vector Embeddings

Mukherjee, EMNLP'16
Ren, ACM MM'16

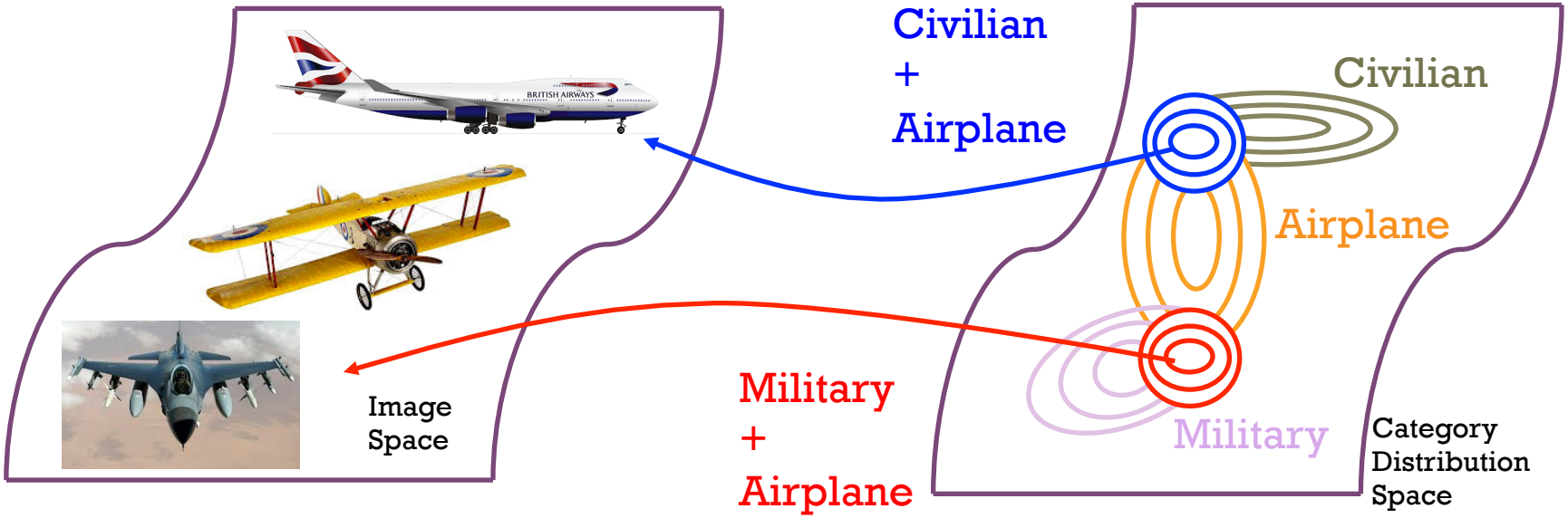
- Embedding categories as distributions.
  - Obtain category distributions from word context variability.
    - Conventional category vector can be seen as mean of category distribution.
  - Train by usual max-margin ranking.
    - But replace bilinear energy function with probability product kernel.
  - Recognition finding the maximum likelihood Gaussian rather than NN vector
    - Because a category embedding now has **shape**, the label of an image may not be its nearest category mean.



Beyond Vector Embeddings

Mukherjee, EMNLP'16
Ren, ACM MM'16

- Embedding categories as distributions.
- For retrieval, rather than recognition, can also exploit category shape for multi-tag querying.



From Zero-Shot Recognition to Zero-Shot Domain Adaptation

From Zero Shot Recognition to Zero-Shot Domain Adaptation

- Zero-Shot Recognition:
  - When entire categories are missing examples and/or labels.
- Zero-Shot Domain Adaptation:
  - When entire domains are missing examples…
    - Directly applying models across domains => Poor performance
    - If the domain has examples but not labels => Unsupervised domain adaptation.
    - If missing examples, then even (unsupervised) domain adaptation is not a solution.
  - …Zero-Shot Domain Adaptation may help here.
- Caveat:
  - Where Zero-Shot Recognition needs a category vector embedding.
  - Zero-Shot Domain Adaptation needs a domain vector embedding.

From Zero Shot Recognition to Zero-Shot Domain Adaptation

- Domain Vector Embedding

Vector Domain Embedding

[[0,0,1], [1,0,0]]
Back, 2010

[[0,1,0], [0,1,0]],
Side, 1950

[[1,0,0], [0,0,1]]
Front, 1940

Car Make Recognition

Domain = 1

Domain = 2

Domain = 3

Categorical Domain

Surveillance Vehicle Detection

Evening, Summer
6PM, Weekday

Night, Summer
1AM, Weekend

Day, Winter
6PM, Weekend

Deep Zero-Shot Domain Adaptation: Surveillance Car Detection

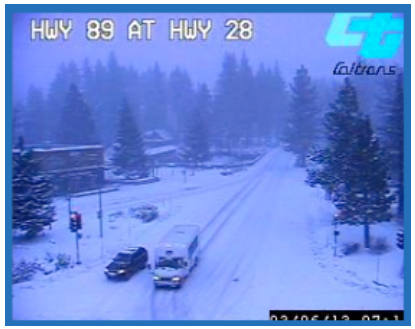
Yang, CVPR'16
Yang, arXiv'16b

Train Domains

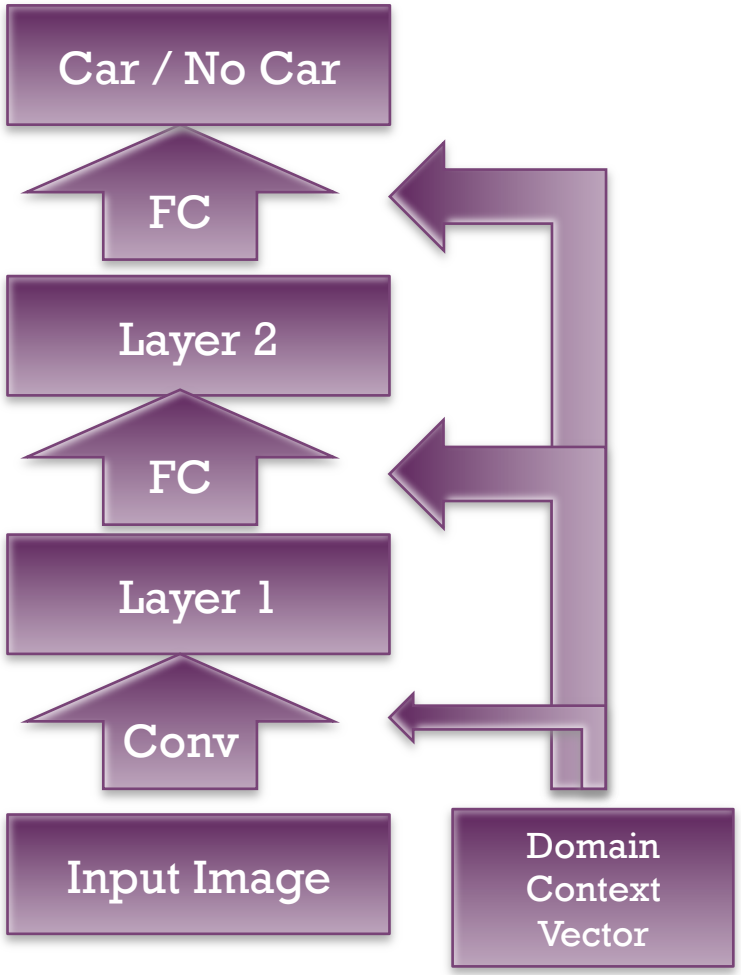
Summer, Weekend, 6AM

Summer, Weekday, 11PM

Test Domain



Winter Weekday 6AM



Deep Zero-Shot Domain Adaptation: Machine Listening

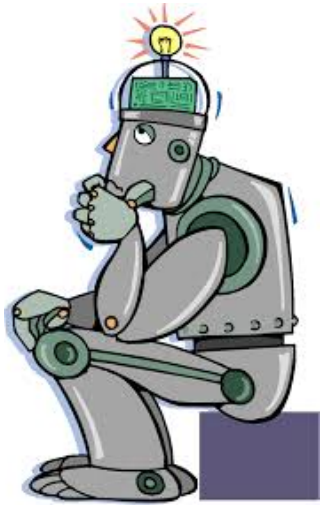
- Audio recognition. [Yang, ICLR' 15]
  - Another problem with many covariates/domains….

Source

Environment

Sensor

Recognition



Too Many Combinatorial Conditions!

Solution:

- Class Vectors => Domain vectors.
- Zero-shot generalization to any new context is a context vector.

Summary

- Zero-Shot Learning
  - Allows recognition/retrieval of previously unseen categories
  - …Assuming both new and old categories have a vector embedding
- Zero-Shot Domain Adaptation
  - Allows immediately applying model in previously unseen domain.
    - Without waiting for either labeled or unlabeled data.
  - …Assuming domains have a vector embedding
  - (Xiaogang will talk next about DA with data)
- Emerging Issues Discussed
  - Cross-dataset transfer
  - Transductive Learning & Inference
  - Zero-shot as a Few-shot prior
  - Deep + Deeper ZSL
  - From Vector to Distribution Embeddings

Summary

- Further Emerging Issues Not Discussed.
  - Diverse ways to obtain category/image embeddings
    - Objects [Kordumova, ICMR'16], Label Hierarchy, Wikipedia [Rohrbach, CVPR'11; Gan, AAI'15]
  - Train/Test Domain Shift & Hubness problem
    - [Lazardiou, ACL'16; Fu PAMI'15]
  - Exploiting Unused Words (cf: unlabeled instances)
    - [Fu, CVPR' 16]
  - ZSL for other problems than recognition
  - Missing Theory of ZSL

References

- Akata et al, PAMI, Label-Embedding for Image Classification, 2015
- Ba et al, ICCV, Predicting Deep Zero-Shot Convolutional Neural Networks using Textual Descriptions, 2015
- Belkin et al, JMLR, Manifold Regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples, 2006
- Frome et al, NIPS, DeViSE: A Deep Visual-Semantic Embedding Model, 2013
- Fu et al, IEEE T PAMI, Transductive Multi-View Zero-Shot Learning, 2015.
- Fu & Sigal, CVPR, Semi-supervised Vocabulary-informed Learning, 2016.
- Kordumova et al, ICMR, Pooling Objects for Recognizing Scenes Without Examples, 2016
- Gan et al, AAAI, Exploring Semantic Inter-Class Relationships (SIR) for Zero-Shot Action Recognition, 2015
- Lazardiou et al, ACL, Hubness and Pollution: Delving into Cross-Space Mapping for Zero-Shot Learning, 2016
- Mensink et al, CVPR, COSTA: Co-Occurrence Statistics for Zero-Shot Classification, 2014
- Mukherjee & Hospedales, EMNLP, Gaussian Visual-Linguistic Embedding for Zero-Shot Recognition, 2016
- Norouzi et al, ICLR, Zero-Shot Learning by Convex Combination of Semantic Embeddings, 2014

References

- Ren et al, ACM MM, Joint image-text representation by gaussian visual semantic embedding, 2016
- Rohrbach et al, NIPS, Transfer Learning in a Transductive Setting, 2013
- Rohrbach et al, CVPR, Evaluating Knowledge Transfer and Zero-Shot Learning in a Large-Scale Setting, 2011
- Romera-Paredes & Torr, ICML, An embarrassingly simple approach to zero-shot learning, 2015
- Socher et al, NIPS, Zero-Shot Learning Through Cross-Modal Transfer, 2013
- Xu et al, ECCV, Multi-task zero-shot action recognition with prioritized data augmentation, 2016
- Xu et al, arXiv:1511.04458, Transductive Zero-Shot Action Recognition by Word-Vector Embedding
- Yang & Hospedales, ICLR, A Unified Perspective on Multi-Domain and Multi-Task Learning, 2015
- Yang & Hospedales, CVPR, Multivariate Regression on the Grassmannian for Predicting Novel Domains, 2016
- Yang & Hospedales, arXiv:1605.06391, Deep Multi-task Representation Learning: A Tensor Factorisation Approach, 2016
- Yang & Hospedales, arXiv, Unifying Multi-Domain Multi-Task Learning: Tensor and Neural Network Perspectives, 2016