Sıfır Örnek ile Nesne Tanıma, Nesne Tespiti ve Görüntü Alt-yazılama

Sabancı Üniversitesi – Veri Bilimi Yaz Okulu

Gökberk Cinbiş
Department of Computer Engineering
METU

July 2018

Machine Learning in Nutshell

- Tens of thousands of machine learning algorithms
  - Hundreds new every year

- Decades of ML research oversimplified:
  - All of Machine Learning:
  - Learn a mapping from input to output $f: X \rightarrow Y$
    - e.g. X: emails, Y: {spam, notspam}

Slide by Dhruv Batra

Supervised Learning

- Input: x (images, text, emails…)
  Output: y (spam or non-spam…)
- (Unknown) Target Function
  - $f: X \to Y$ (the "true" mapping / reality)
- Training dataset: $(x_1,y_1), (x_2,y_2), \ldots, (x_N,y_N)$
- Model / Hypothesis Class
  - $g: X \to Y$
- Learning = Search in hypothesis space
  - Find best g in model class.

Slide adapted from Dhruv Batra

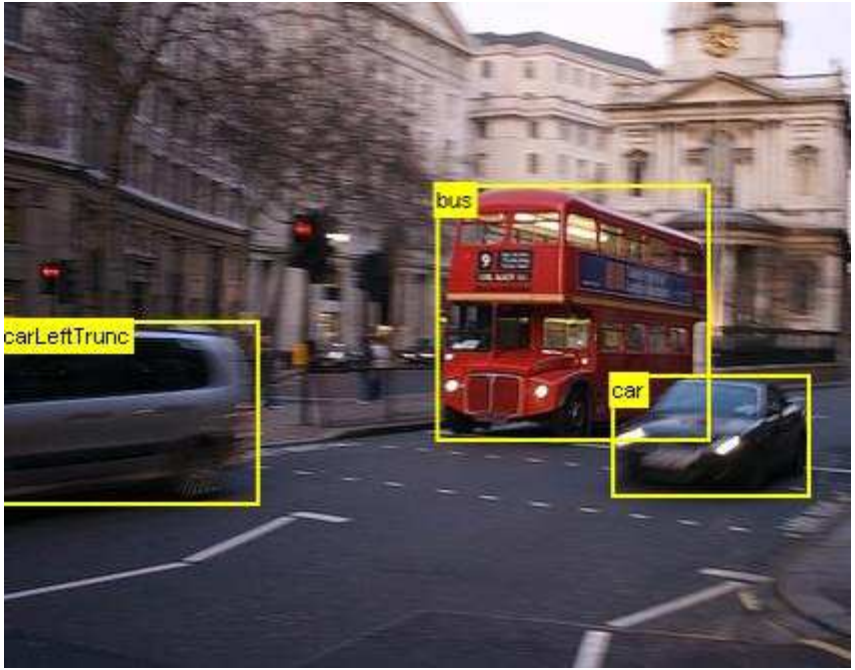
Supervised Learning

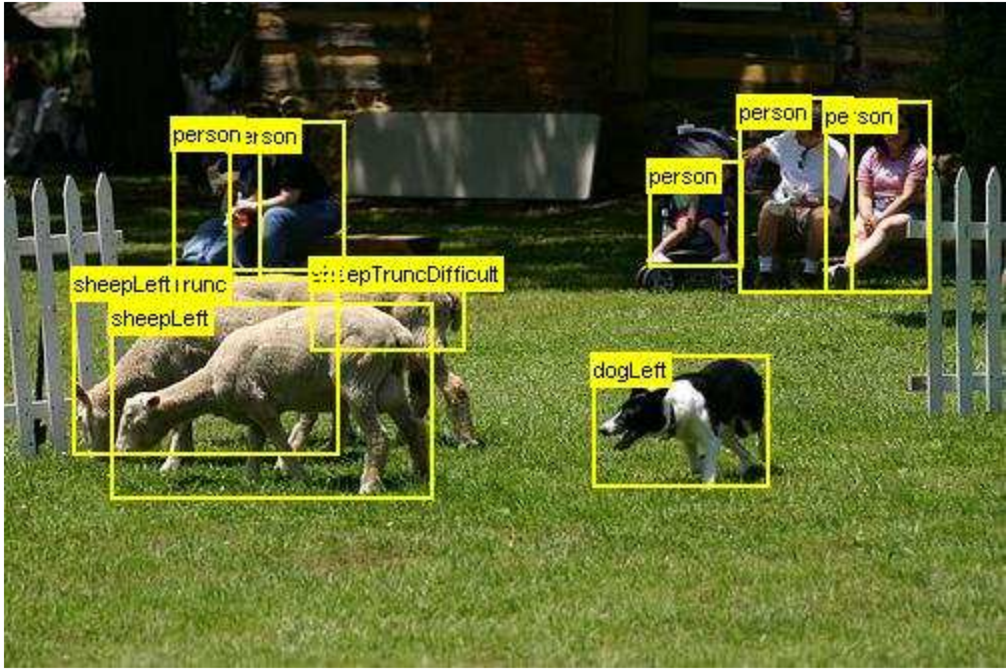
- Input: x (images, text, emails…)
  Output: y (spam or non-spam…)
- (Unknown) Target Function
  - $f: X \rightarrow Y$ (the "true" mapping / reality)
- Training dataset: $(x_1,y_1), (x_2,y_2), \ldots, (x_N,y_N)$
- Model / Hypothesis Class
  - $g: X \rightarrow Y$
- Learning = Search in hypothesis space
  - Find best g in model class.

Slide adapted from Dhruv Batra

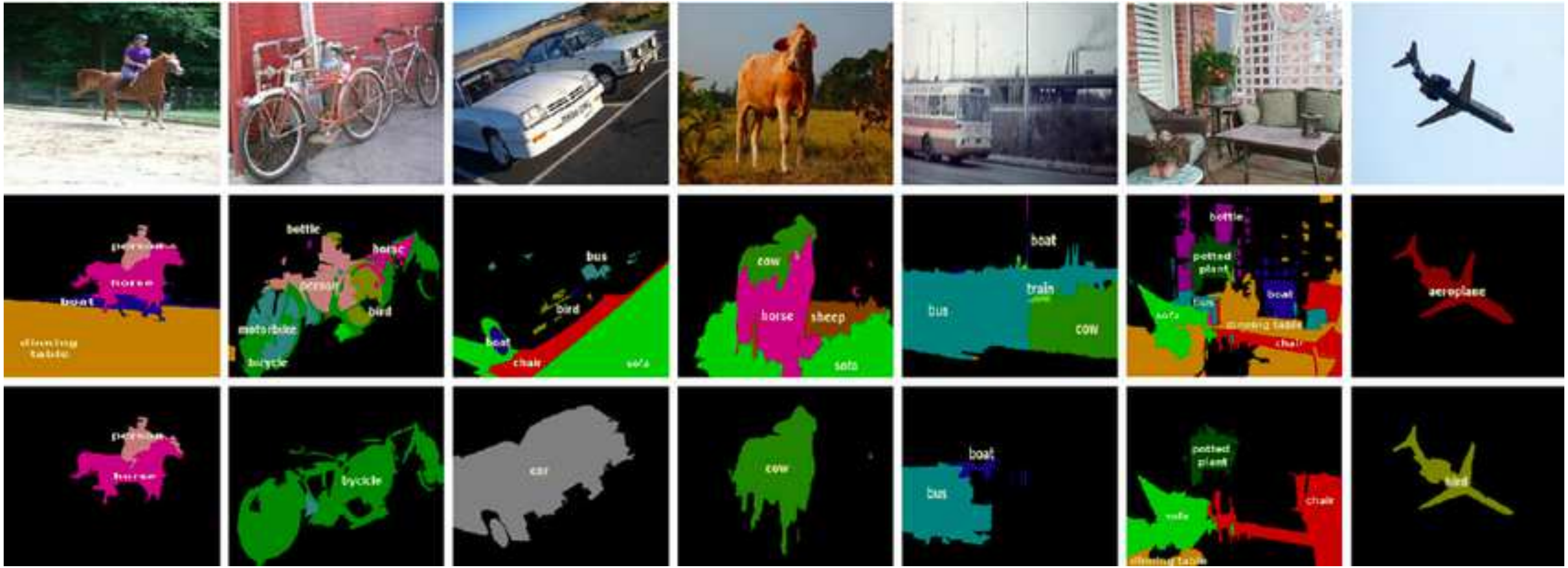
Supervised training - Image classification

Supervised training - Object detection





Supervised training - Semantic segmentation



How many training examples do we need?

- 75.000 non-abstract nouns from WordNet*, some of which are *rare*

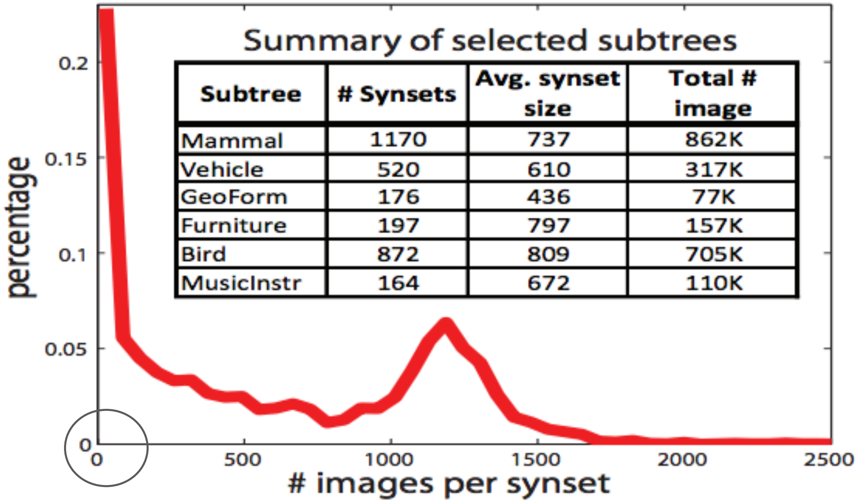
* Torralba, et al. 2008.

How many training examples do we need?

- 75.000 non-abstract nouns from WordNet*, some of which are *rare*







Summary of selected subtrees

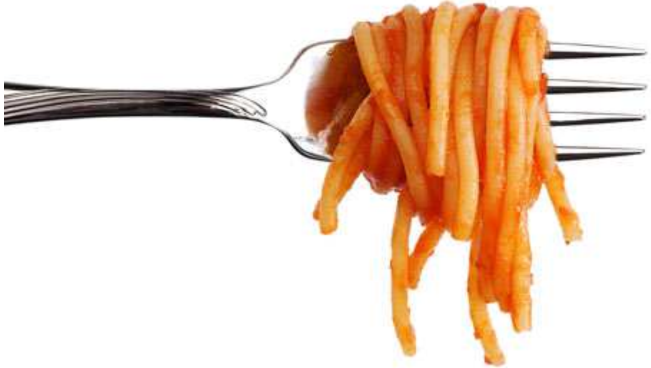
Subtree	# Synsets	Avg. synset size	Total # image
Mammal	1170	737	862K
Vehicle	520	610	317K
GeoForm	176	436	77K
Furniture	197	797	157K
Bird	872	809	705K
MusicInstr	164	672	110K

* Torralba, et al. 2008.

How many training examples do we need?

- … plus object combinations, scenes

A tennis player hitting a ball



Fork with spaghetti

Wedding car

- It is not feasible to collect several fully annotated samples per "class"
- (... and *categorization* is a questionable paradigm)

Deng et al. CVPR 2009

Learning with Incomplete Supervision

- The main goal: minimize the data collection and/or annotation effort

- Between the two extremes of *supervised* and *unsupervised* learning

- Some examples that we focus in our research group:
  - **Semi-supervised learning** (supervised+unsupervised)
  - **Transductive learning** (unsupervised test examples)
  - **Weakly-supervised localization** (training images with labels only)
  - **Zero-shot learning** (learning novel classes based on auxiliary knowledge only)
  - **One-shot learning** (learning from a single example)

Learning with Incomplete Supervision

- The main goal: minimize the data collection and/or annotation effort

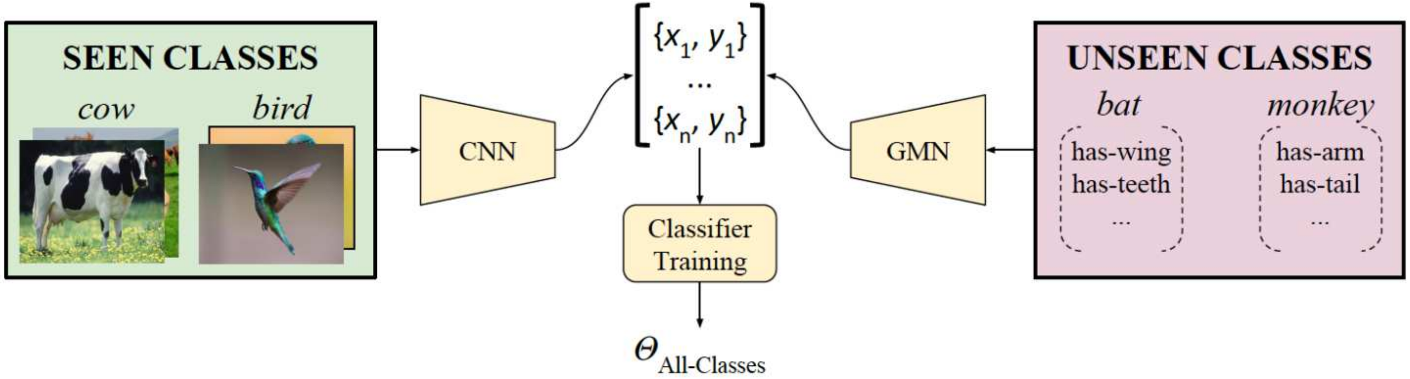
- Between the two extremes of *supervised* and *unsupervised* learning

- Some examples that we focus in our research group:
  - **Semi-supervised learning** (supervised+unsupervised)
  - **Transductive learning** (unsupervised test examples)
  - **Weakly-supervised localization** (training images with labels only)
  - **Zero-shot learning** (learning novel classes based on auxiliary knowledge only)
  - **One-shot learning** (learning from a single example)

Learning with Incomplete Supervision

- The main goal: minimize the data collection and/or annotation effort
- Between the two extremes of *supervised* and *unsupervised* learning
- Some examples that we focus in our research group:
  - **Semi-supervised learning** (supervised+unsupervised)
  - **Transductive learning** (unsupervised test examples)
  - **Weakly-supervised localization** (training images with labels only)
  - **Zero-shot learning** (learning novel classes based on auxiliary knowledge only)
  - **One-shot learning** (learning from a single example)

Part 1: Gradient Matching Networks



IEEE / CVF Conf. on Computer Vision and Pattern Recognition (CVPR), June 2019

Part 2: Zero-shot Object Detection

British Machine Vision Conference (BMVC), September 2018

Part 3: Image Captioning with Unseen Objects

♦: A yellow and black **train** traveling down the road.
★: A yellow and black **bus** driving down a road.

♦: A couple of **elephants** standing next to each other.
★: A couple of **zebra** standing next to each other.

♦: A piece of **cake** on a white plate.
★: A piece of **pizza** on a white plate.

British Machine Vision Conference (BMVC), September 2019

Outline

- Introduction
- Gradient Matching Networks
- Zero-Shot Object Detection by Hybrid Region Embedding
- Image Captioning with Unseen Objects
- Conclusions

Zero-shot **object recognition**

Seen Classes

cow *bird*


Unseen Classes

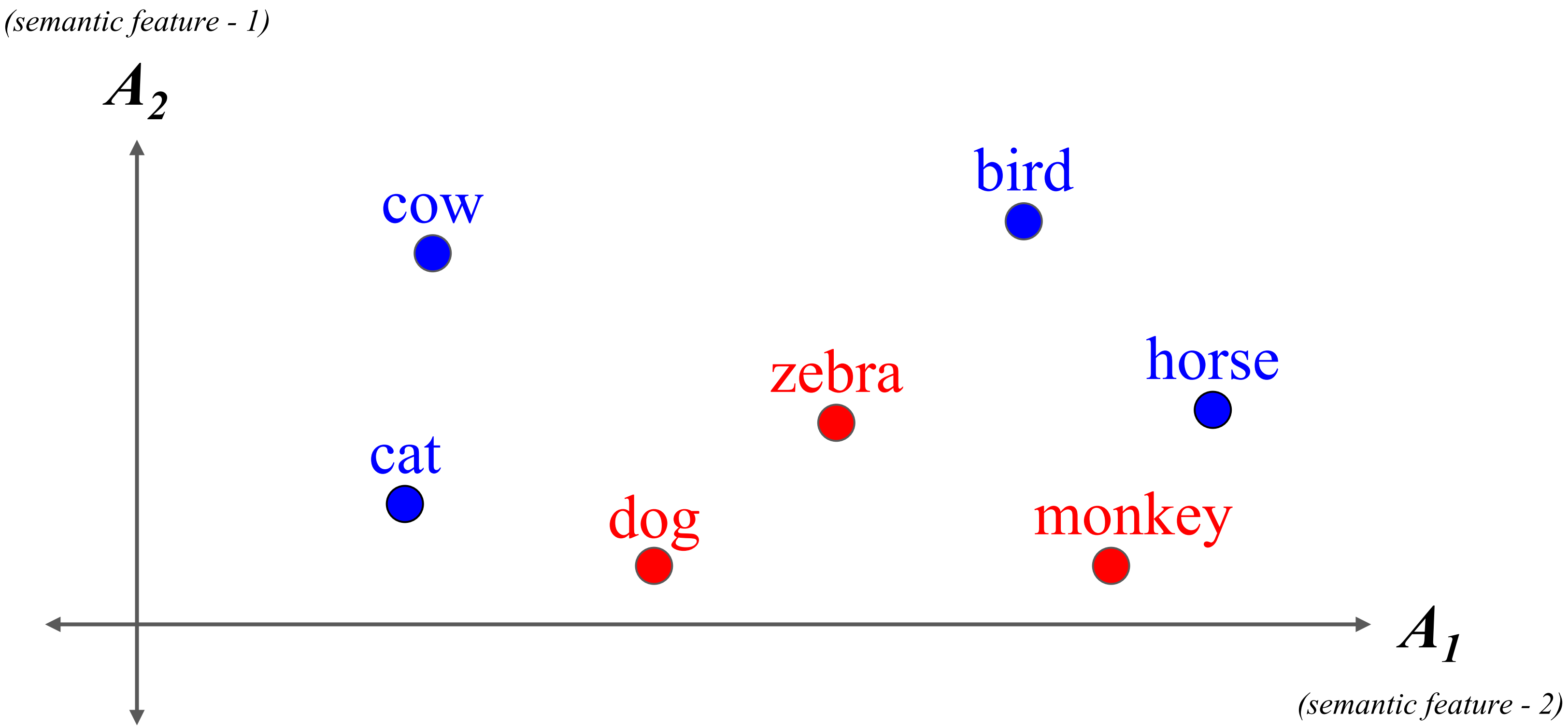
bat *monkey*

~~Training samples~~

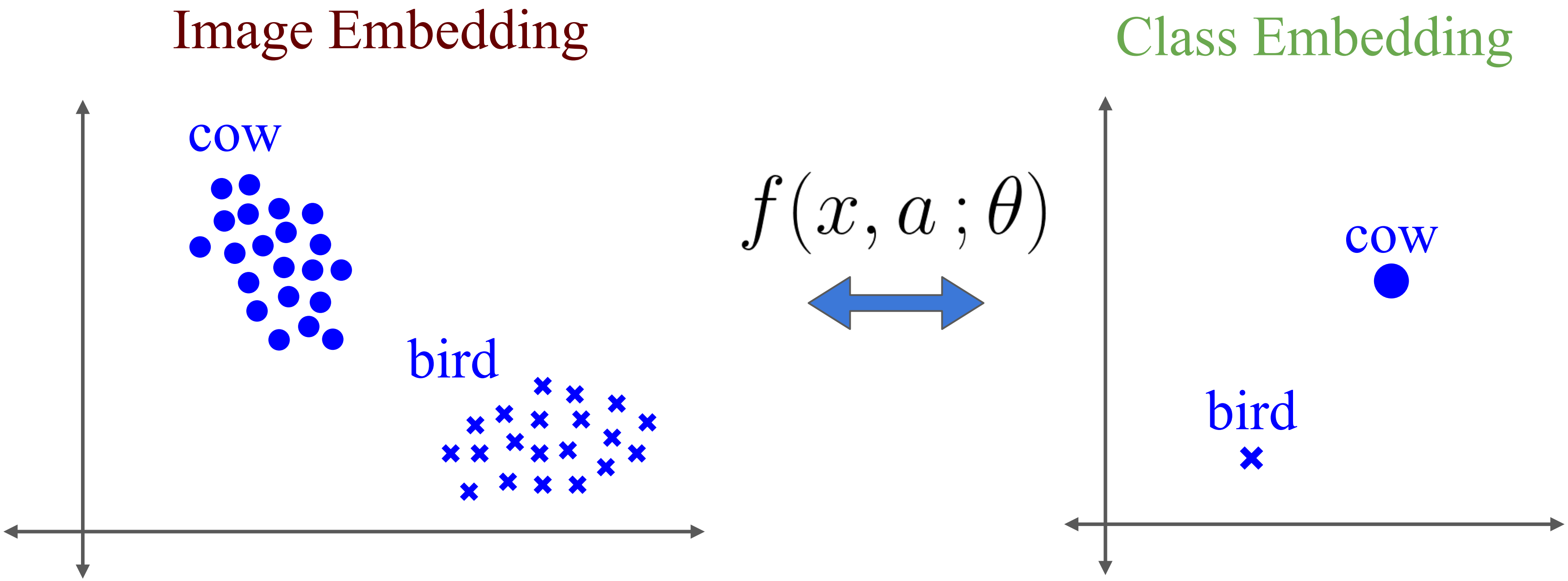
i - Learn a classification model on **seen** classes

ii - Use the model for **both** sets

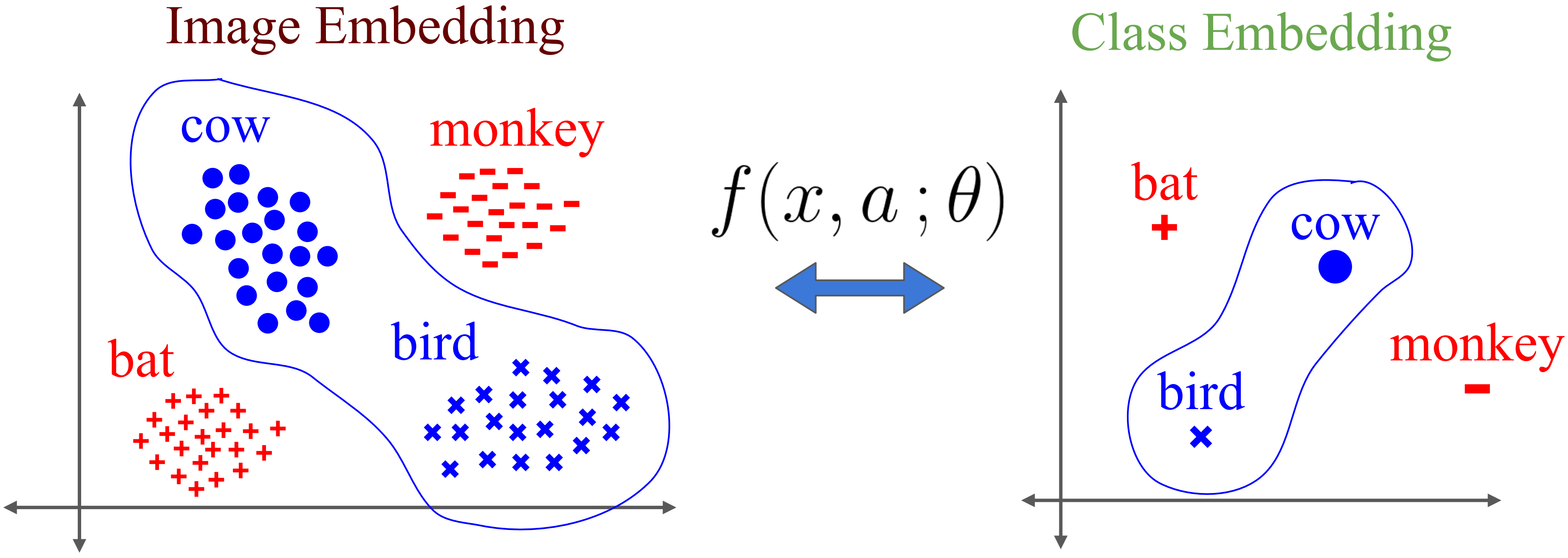
Semantic Class Embedding Space



Mainstream approach

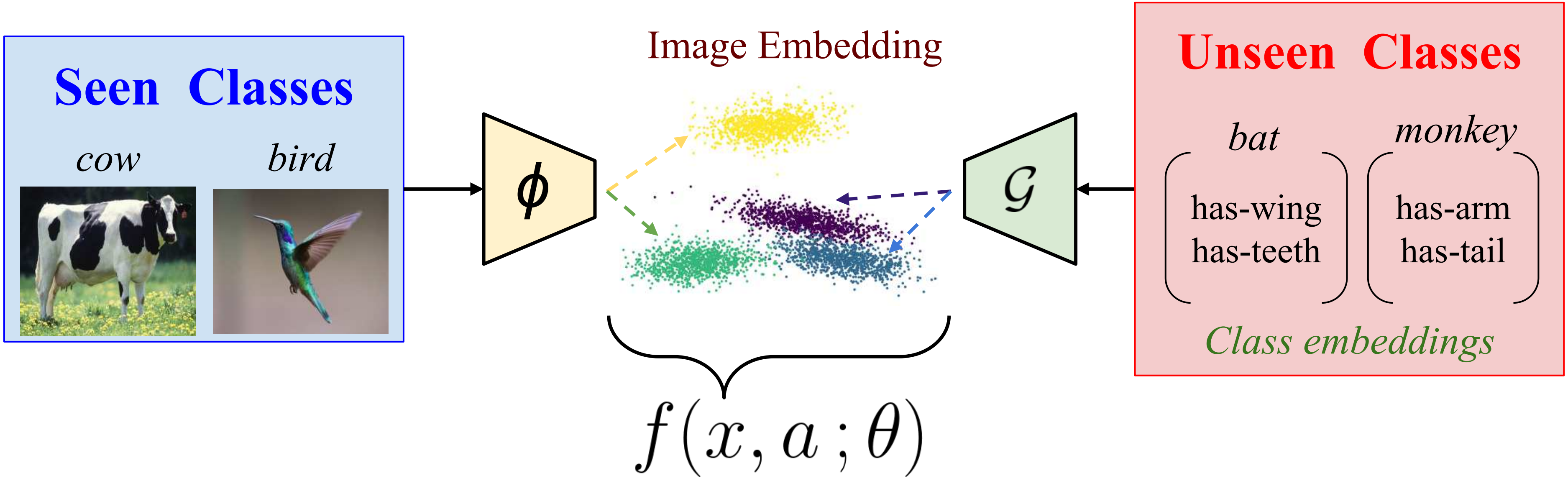


A weakness in purely discriminative approaches



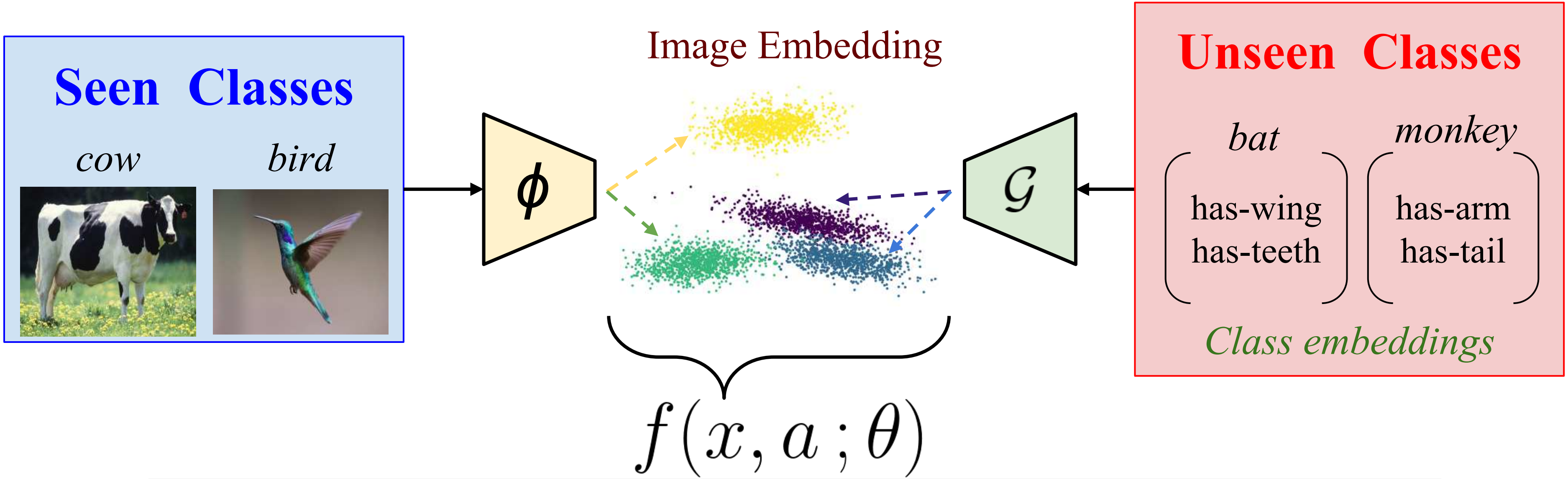
Akata et al. "Label-embedding for attribute-based classification." CVPR 2013.

Generative-model-based approaches



Xian et al. "Feature generating networks for zero-shot learning." CVPR 2018.
Verma et al. "Generalized zero-shot learning via synthesized examples." CVPR 2018.

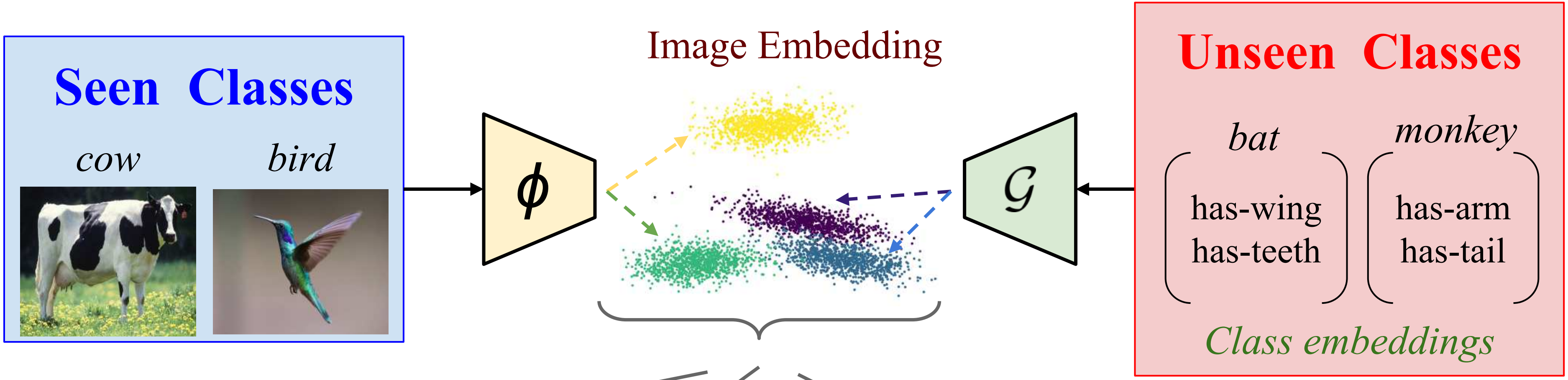
Generative-model-based approaches



Three important inter-connected challenges:
- **Semantics:** How do we enforce producing samples that truly belong to the target class?
- **Variance:** How do we enforce producing a variety of samples for a given embedding?
- **Data quality:** How do we make sure that the resulting training examples is actually useful? (ie. will the classifier trained over them be accurate?)

Xian et al. "Fe
Verma et al. "

Training with real and generated samples



$$\theta^{(0)} \xrightarrow{-\nabla_\theta \mathcal{L}} \theta^{(1)} \xrightarrow{-\nabla_\theta \mathcal{L}} \cdots \xrightarrow{-\nabla_\theta \mathcal{L}} f(x, a; \theta^{(N)})$$

Gradient matching loss

$$\mathcal{L}_{\text{GM}} = \mathbb{E}_{\theta}\left[1 - \frac{g_r(\theta)^T g_f(\theta)}{||g_r(\theta)||_2 \; ||g_f(\theta)||_2}\right]$$

Gradient by real

$$g_r(\theta) = \mathbb{E}_{(x,a)\sim p_{\text{data}}}\left[\nabla_{\theta}\mathcal{L}(x, a, f_{\theta})\right]$$

Gradient by generated

$$g_f(\theta) = \mathbb{E}_{\tilde{x}\sim\mathcal{G}(z,a),\, a\sim p_{\text{data}}}\left[\nabla_{\theta}\mathcal{L}(\tilde{x}, a, f_{\theta})\right]$$

To approximate the expectation over θ

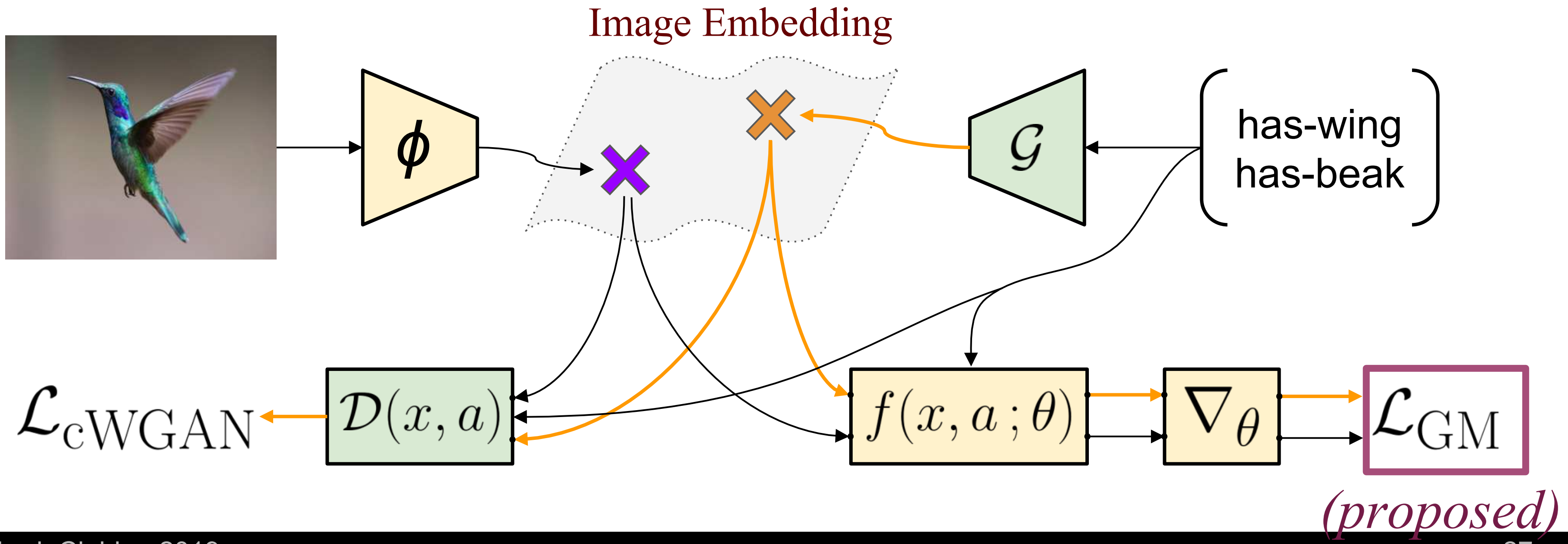
$$\mathcal{L}_{\mathrm{GM}} = \underline{\mathbb{E}_\theta} \left[1 - \frac{g_r(\theta)^T g_f(\theta)}{||g_r(\theta)||_2 \; ||g_f(\theta)||_2} \right]$$

Repeatedly:

- train the classification model **N** epochs,
- re-initialize all parameters and reset the optimizer state.

Gradient matching network (GMN)

Gradient matching loss
+ adversarial loss (allows unsupervised learning)



Experiments - Datasets

- Caltech-UCSD Birds-200-2011 (**CUB**) - 200 bird species - 12k

- SUN Attribute (**SUN**) - 717 scene categories - 14k

- Animals with Attributes (**AWA**) - 50 animal categories - 30k

Wah et al. “The Caltech-UCSD Birds-200-2011 Dataset”, 2011.
Patterson et al. “Sun attribute database: Discovering, annotating, and recognizing scene attributes” CVPR, 2012.
Lampert et al. “Attribute-based classification for zero-shot visual object categorization” TPAMI, 2013.

Evaluation Metrics

Normalized score (NS) : average of the top-1 per-class scores

- **T-1** : NS of unseen classes in ZSL setting
- **u**: NS of unseen classes in GZSL setting
- **s**: NS of seen classes in GZSL setting
- **h:** harmonic mean of **u** and **s** $\frac{2 \times \mathbf{u} \times \mathbf{s}}{\mathbf{u} + \mathbf{s}}$

Zero-shot prediction (unseen classes)

					CUB	SUN	AWA
					T-1	T-1	T-1
1	*Zhang et al.* '18				52.6	61.7	67.4
2	*Bucher et al.* '17				57.8	60.4	66.3
3	*Xian et al.* - `DEVISE` '18				60.3	60.9	66.9
4	*Xian et al.* - `ALE` '18				61.5	62.1	68.2
5	*Xian et al.* - `Softmax` '18				57.3	60.8	68.2
6	*Verma et al.* '18				59.6	63.4	69.5
7	*Felix et al.* - `cycle-WGAN` '18				57.8	59.7	65.6
8	*Felix et al.* - `cycle-CLSWGAN` '18				58.4	60.0	66.3
9	Bilinear	LN	$\mathcal{L}^{\mathrm{S}}_{\mathrm{cWGAN}}$		61.7	62.7	67.3
10	Bilinear	LN	$\mathcal{L}^{\mathrm{S}}_{\mathrm{cWGAN}} + \mathcal{L}_{\mathrm{CLS}}$		61.9	62.7	66.4
11	Bilinear	LN	$\mathcal{L}^{\mathrm{S}}_{\mathrm{cWGAN}} + \mathcal{L}_{\mathrm{CYCLE}}$		62.2	62.7	68.2
12	Bilinear	LN	$\mathcal{L}^{\mathrm{S}}_{\mathrm{cWGAN}} + \mathcal{L}_{\mathrm{GM}}$	*(Ours)*	**67.0**	**63.6**	**72.0**
13	Linear	LN	$\mathcal{L}^{\mathrm{S}}_{\mathrm{cWGAN}} + \mathcal{L}_{\mathrm{GM}}$	*(Ours)*	63.1	58.9	70.1
14	Bilinear	AC	$\mathcal{L}^{\mathrm{S}}_{\mathrm{cWGAN}} + \mathcal{L}_{\mathrm{GM}}$	*(Ours)*	65.7	62.6	69.7
15	Linear	AC	$\mathcal{L}^{\mathrm{S}}_{\mathrm{cWGAN}} + \mathcal{L}_{\mathrm{GM}}$	*(Ours)*	63.8	61.1	66.8

Generalized zero-shot prediction (seen + unseen classes)

				CUB			SUN			AWA		
				u	s	h	u	s	h	u	s	h
1	*Zhang et al.* '18			31.5	40.2	35.3	41.2	26.7	32.4	38.7	74.6	51.0
2	*Bucher et al.* '17			28.8	55.7	38.0	40.5	37.2	38.8	2.3	**90.2**	4.5
3	*Xian et al.* - DEVISE '18			52.2	42.4	46.7	38.4	25.4	30.6	35.0	62.8	45.0
4	*Xian et al.* - ALE '18			40.2	59.3	47.9	41.3	31.1	35.5	47.6	57.2	52.0
5	*Xian et al.* - Softmax '18			43.7	57.7	49.7	42.6	36.6	39.4	57.9	61.4	59.6
6	*Verma et al.* '18			41.5	53.3	46.7	40.9	30.5	34.9	56.3	67.8	61.5
7	*Felix et al.* - cycle-WGAN '18			46.0	60.3	52.2	48.3	33.1	39.2	56.4	63.5	59.7
8	*Felix et al.* - cycle-CLSWGAN '18			45.7	61.0	52.3	49.4	33.6	40.0	56.9	64.0	60.2
9	Bilinear	LN	$\mathcal{L}^{S}_{cWGAN}$	45.6	59.2	51.5	50.6	30.3	37.3	53.5	72.0	61.4
10	Bilinear	LN	$\mathcal{L}^{S}_{cWGAN} + \mathcal{L}_{CLS}$	45.5	58.9	51.4	50.6	30.3	37.3	52.7	71.0	60.5
11	Bilinear	LN	$\mathcal{L}^{S}_{cWGAN} + \mathcal{L}_{CYCLE}$	51.1	54.9	52.9	50.6	30.3	37.3	55.4	70.1	61.8
12	Bilinear	LN	$\mathcal{L}^{S}_{cWGAN} + \mathcal{L}_{GM}$ *(Ours)*	**54.7**	58.4	**56.5**	42.5	35.5	38.7	**61.1**	71.3	65.8
13	Linear	LN	$\mathcal{L}^{S}_{cWGAN} + \mathcal{L}_{GM}$ *(Ours)*	48.5	**62.8**	54.7	42.0	**39.3**	40.7	57.1	81.3	**67.1**
14	Bilinear	AC	$\mathcal{L}^{S}_{cWGAN} + \mathcal{L}_{GM}$ *(Ours)*	53.8	58.2	55.9	43.2	36.2	39.4	54.8	74.1	63.0
15	Linear	AC	$\mathcal{L}^{S}_{cWGAN} + \mathcal{L}_{GM}$ *(Ours)*	45.8	65.5	53.9	**53.2**	33.0	**42.8**	46.8	84.8	60.3

In summary

- a **novel** proxy loss for **zero-shot learning**
    - better estimation of class distributions
- **state of the art** on CUB, AWA and SUN

Source code: https://mbsariyildiz.github.io/

Outline

- Introduction
- Gradient Matching Networks
- Zero-Shot Object Detection by Hybrid Region Embedding
- Image Captioning with Unseen Objects
- Conclusions

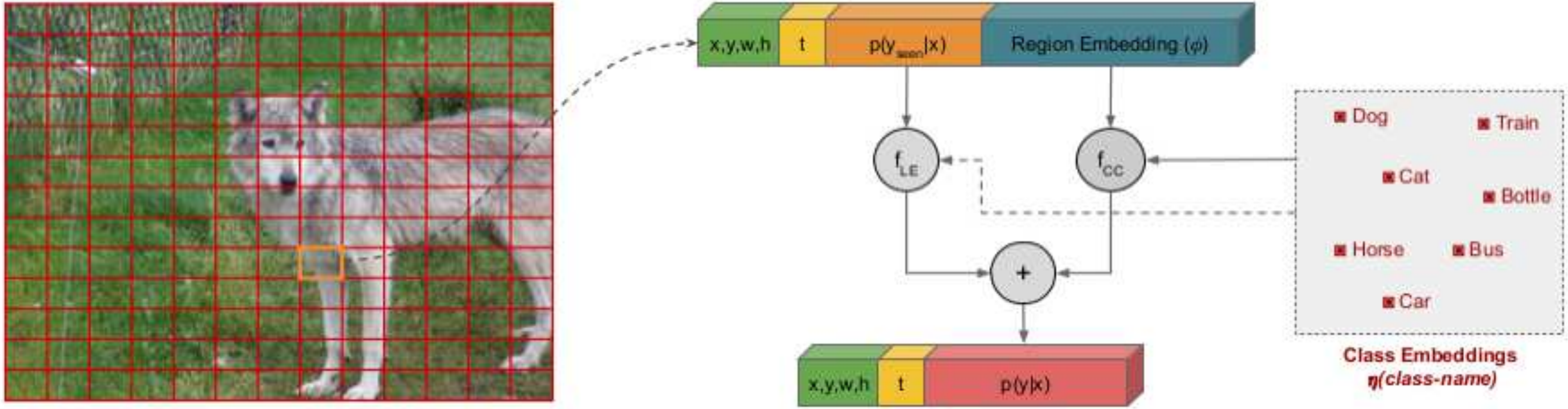
Motivation

Detection in the Wild
using text-based queries

Robotic

Our approach

- ➔ Our method consists of two components:
  - ◆ (i) utilize a convex combination of class embeddings,
  - ◆ (ii) directly learn to map regions to the space of class embeddings.
- ➔ Zero-shot object detection within the YOLO detection framework.



Convex Combination of Class Embeddings

- Represent a given image region (i.e. a bounding box) as the convex combination of training class embeddings.

Class embedding

$$f_{\mathrm{CC}}(x,b,y) = \frac{\phi_{\mathrm{CC}}(x,b)^{\mathrm{T}}\eta(y)}{\|\phi_{\mathrm{CC}}(x,b)\|\|\eta(y)\|}$$

$$\phi_{\mathrm{CC}}(x,b) = \frac{1}{\sum_{y\in\mathcal{Y}_s} p(y|x,b)} \sum_{y\in\mathcal{Y}_s} p(y|x,b)\eta(y)$$

Convex Combination of Class Embeddings

- Represent a given image region (i.e. a bounding box) as the convex combination of training class embeddings.

Sum of class embeddings, weighted by posterior probability

$$f_{\mathrm{CC}}(x,b,y) = \frac{\phi_{\mathrm{CC}}(x,b)^{\mathrm{T}}\eta(y)}{\|\phi_{\mathrm{CC}}(x,b)\|\|\eta(y)\|}$$

$$\phi_{\mathrm{CC}}(x,b) = \frac{1}{\sum_{y\in\mathcal{Y}_s} p(y|x,b)} \sum_{y\in\mathcal{Y}_s} p(y|x,b)\eta(y)$$

Region Scoring by Label Embedding

- The goal is to directly model the compatibility between the visual features of image regions and class embeddings.
- The equation can be interpreted as a dot product between L2-normalized image region descriptors and class embeddings.

Region descriptor

$$f_{\mathrm{LE}}(x,b,y) = \frac{\phi_{\mathrm{LE}}(x,b)^{\mathrm{T}}\eta(y)}{\|\phi_{\mathrm{LE}}(x,b)\|\,\|\eta(y)\|}$$

Hybrid region embedding

- The two scores are accumulated within the loss function:

$$L_{\mathrm{LE}}(x,b,y) = \frac{1}{|\mathcal{Y}_s|-1} \sum_{y' \in \mathcal{Y}_s \setminus \{y\}} \max\left(0, 1 - f_{\mathrm{LE}}(x,b,y) + f_{\mathrm{LE}}(x,b,y')\right)$$

Experimental Results on PASCAL VOC

- Select 16 of the 20 classes as the training set.
- Remaining 4 classes as the test set. These test classes are car, dog, sofa and train respectively.
- Class-attribute relations of aPaY dataset are used for semantic descriptions.

Experimental Results on PASCAL VOC

- Select 16 of the 20 classes as the training set.
- Remaining 4 classes as the test set. These test classes are car, dog, sofa and train respectively.
- Class-attribute relations of aPaY dataset are used for semantic descriptions.
- 65.6% mAP on seen classes, 54.6% mAP on unseen ones.

Method	Test split	aeroplane	bicycle	bird	boat	bottle	bus	cat	chair	cow	dining table	horse	motorbike	person	potted plant	sheep	tvmonitor	car	dog	sofa	train	mAP (%)
LE	v	.46	.50	.44	.28	.12	.59	.44	.20	.11	.38	.35	.47	.65	.16	.18	.53	-	-	-	-	36.8
	t	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	.54	.79	.45	.12	47.9
	v+t	.34	.48	.40	.23	.12	.34	.28	.12	.09	.32	.28	.36	.60	.15	.13	.50	.27	.26	.20	.05	27.4
CC	v	.69	.74	.72	.63	.43	.83	.73	.43	.43	.66	.78	.80	.75	.41	.62	.75	-	-	-	-	65.0
	t	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	.60	.85	.44	.27	53.8
	v+t	.67	.73	.70	.59	.41	.61	.58	.32	.32	.65	.74	.68	.72	.39	.57	.72	.49	.24	.10	.15	52.0
H	v	.70	.73	.76	.54	.42	.86	.64	.40	.54	.75	.80	.80	.75	.34	.69	.79	-	-	-	-	**65.6**
	t	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	.55	.82	.55	.26	**54.2**
	v+t	.68	.72	.74	.48	.41	.61	.48	.25	.48	.73	.75	.71	.73	.33	.59	.57	.44	.25	.18	.15	**52.3**

Example detections

Outline

- Introduction
- Gradient Matching Networks
- Zero-Shot Object Detection by Hybrid Region Embedding
- Image Captioning with Unseen Objects
- Conclusions

Problem Statement

- **Motivation:** Overcome the data collection bottleneck in image captioning.
- **Task:** Define a new paradigm for generating captions of unseen classes.
- **Key Idea:** Use zero-shot object detector with template based sentence generator.

Zero-shot Image Captioning

	Image Captioning
Visual Input	
Textual Input	"a **person** riding a **horse**"

Zero-shot Image Captioning

{person, horse} ∈ unseen classes

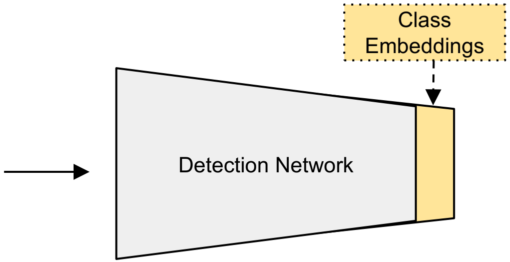
	Image Captioning	Partial Zero-Shot Image Captioning
Visual Input		
Textual Input	"a **person** riding a **horse**"	~~"a **person** riding a **horse**"~~

Zero-shot Image Captioning

{person, horse} ∈ unseen classes

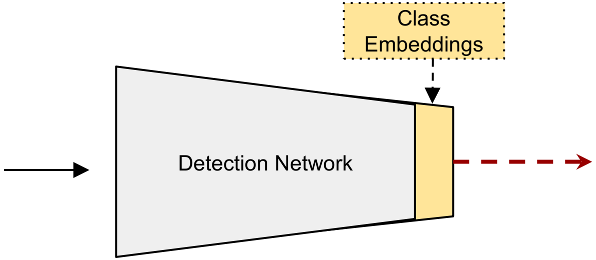
	Image Captioning	Partial Zero-Shot Image Captioning	True Zero-Shot Image Captioning
Visual Input			
Textual Input	“a person riding a horse”	~~“a person riding a horse”~~	~~“a person riding a horse”~~

Framework - Fully Zero-shot Image Captioning



Zero-Shot Object Detector

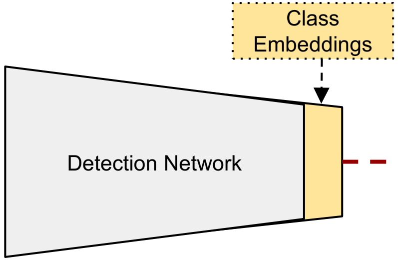
Framework - Fully Zero-shot Image Captioning



$$L_{cls}(x) = \sum_{i=0}^{S^2} \mathbb{1}^{i}_{obj} \sum_{c \in Y_s} (t(x,c,i) - f(x,c,i))^2$$

Zero-Shot Object Detector

Framework - Fully Zero-shot Image Captioning



$$L_{cls}(x) = \sum_{i=0}^{S^2} \mathbb{1}_{obj}^{i} \sum_{c \in Y_s} \left(t(x,c,i) - f(x,c,i)\right)^2$$

$$f(x,c,i) = \frac{\Omega(x,i)^T \Psi(c)}{\| \Omega(x,i) \| \| \Psi(c) \|}$$

Zero-Shot Object Detector

Framework - Fully Zero-shot Image Captioning

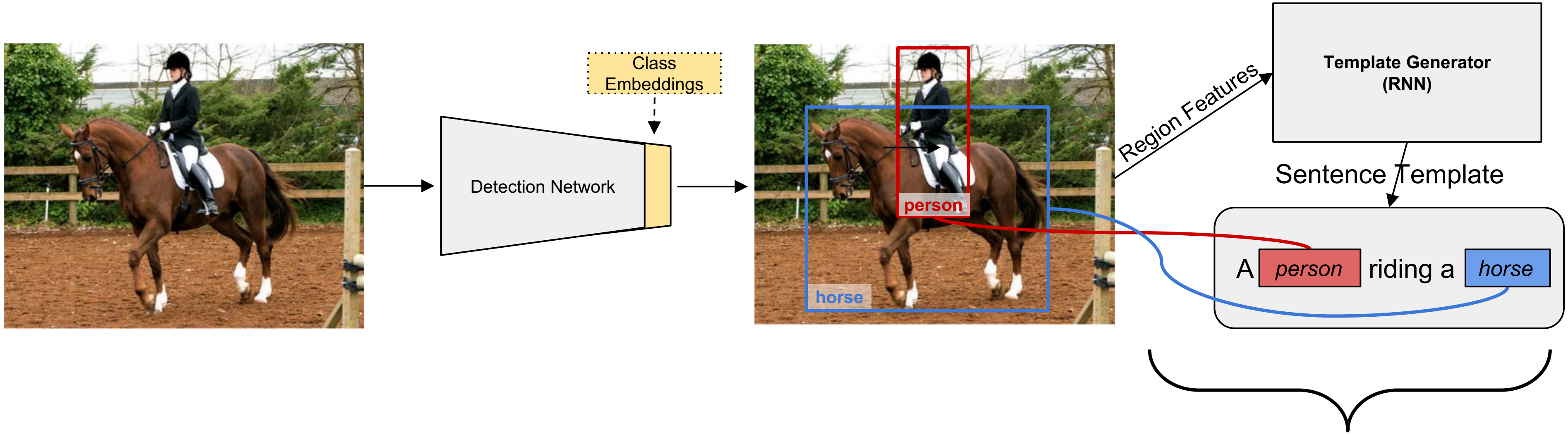


Image Caption Generation*

* Lu, Jiasen, et al. *"Neural baby talk."* CVPR 2018.

Generalized Zero-shot Detection

- There can still be a significant bias towards the seen classes.
- Aim to overcome this problem by introducing a scaling coefficient:

$$f(x,c,i) = \begin{cases} \alpha f(x,c,i), & if\ c \in \hat{Y}_s \\ f(x,c,i), & otherwise \end{cases}$$

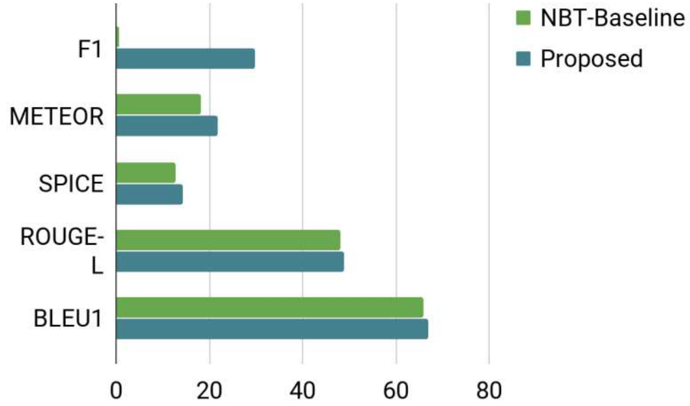
Experimental Setup

- **Dataset**: MSCOCO splits for evaluating zero-shot image captioning.
- **Evaluation**: F1 score, METEOR, SPICE, ROUGE-L, BLEU metrics.
- **Class Embeddings:** Use 300-dim word2vec of class embeddings.
- **ZSD Evaluation:** COCO validation images consist of only unseen objects.
- **GZSD Evaluation:** Use COCO val5k split, which contains both seen and unseen class instances.

Generalized-ZSD results

Classes	GZSD w/o α	GZSD
Bottle	0	0.8
Bus	0	21.4
Couch	2.7	4.9
Microware	0	1.2
Pizza	0	4.8
Racket	0	0.7
Suitcase	0	9.1
Zebra	0	15.8
U-mAP(%)	0.3	**7.3**
S-mAP(%)	27.4	**19.2**
Harmonic Mean	0.7	**10.6**

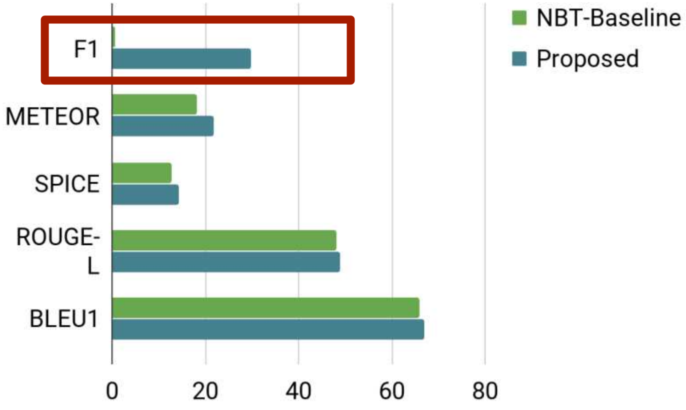
Image Captioning Results



Comparison Results

NBT- Baseline	A piece of **cake** on a white plate.	A yellow and black **train** traveling down the road.
Proposed	A piece of **pizza** on a white plate.	A yellow and black **bus** driving down a road.

Image Captioning Results

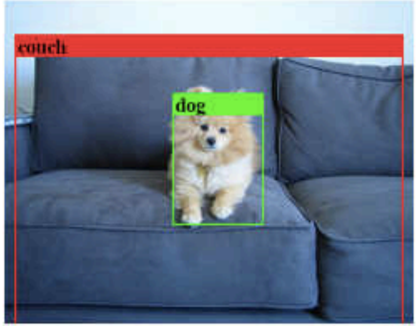


Comparison Results

NBT- Baseline	A piece of **cake** on a white plate.	A yellow and black **train** traveling down the road.
Proposed	A piece of **pizza** on a white plate.	A yellow and black **bus** driving down a road.

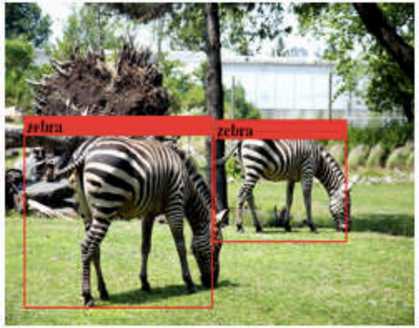
Qualitative Results

Image captioning results of images which consist of seen and unseen classes:



A small white dog is sitting on a couch.

A red bus is driving down the street.



A couple of zebra standing in a field.

A tennis player is about to hit a racket.

A white plate topped with a piece of pizza.

A kitchen with a microwave and a counter.

In summary,

- a **new** paradigm for generating captions of **unseen classes**.
- a **novel** approach for generalized zero-shot object detection problem.

Conclusions

- Towards semantically rich recognition systems, build models that are
  - more flexible
  - more tightly integrated with language
  - requires less supervision
- Presented:
  - Gradient Matching Networks
    - GMN can be used for **semi-supervised / transductive training** not only for ZSL but also in traditional classification and few-shot learning settings
  - A zero-shot object detection approach
  - A approach for Captioning with Unseen Objects

Thank you!