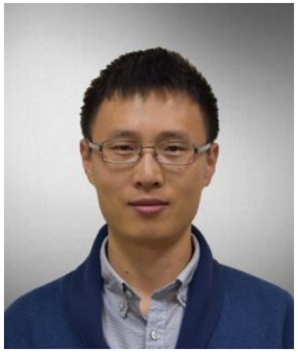
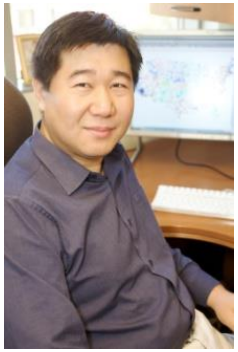
An Empirical Study and Analysis of **Generalized Zero-Shot Learning** for Object Recognition in the Wild

Wei-Lun (Harry) Chao[*1]

Soravit (Beer) Changpinyo[*1]



Boqing Gong[2]



Fei Sha[1,3]

1 USC

2 UCF

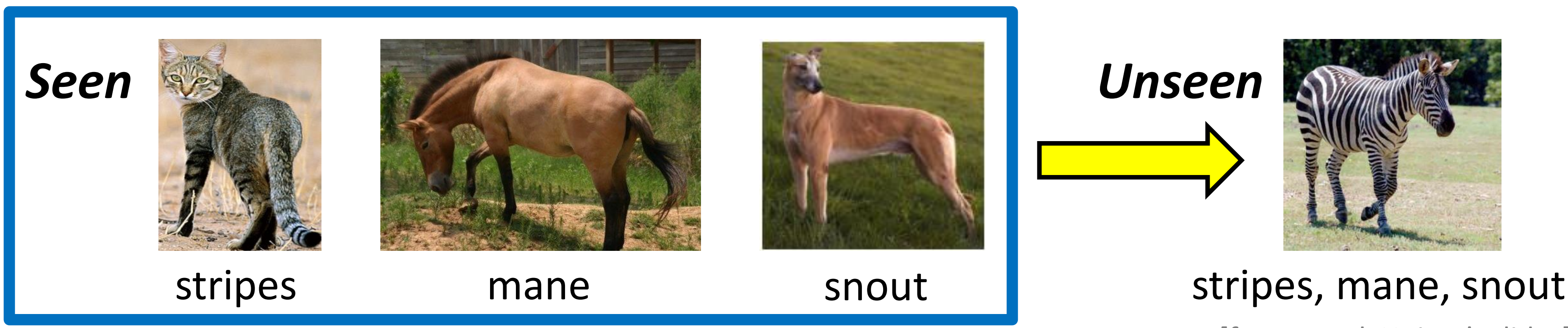
3 UCLA

Challenges of recognition in the wild:

- large-scale labeling space with a long-tail distribution

Zero-shot learning (ZSL):

- expand classifiers beyond ***Seen*** objects to ***Unseen*** objects using **semantic embeddings** (e.g., attributes, WORD2VEC)



[from Derek Hoiem's slides]

Training of ZSL:

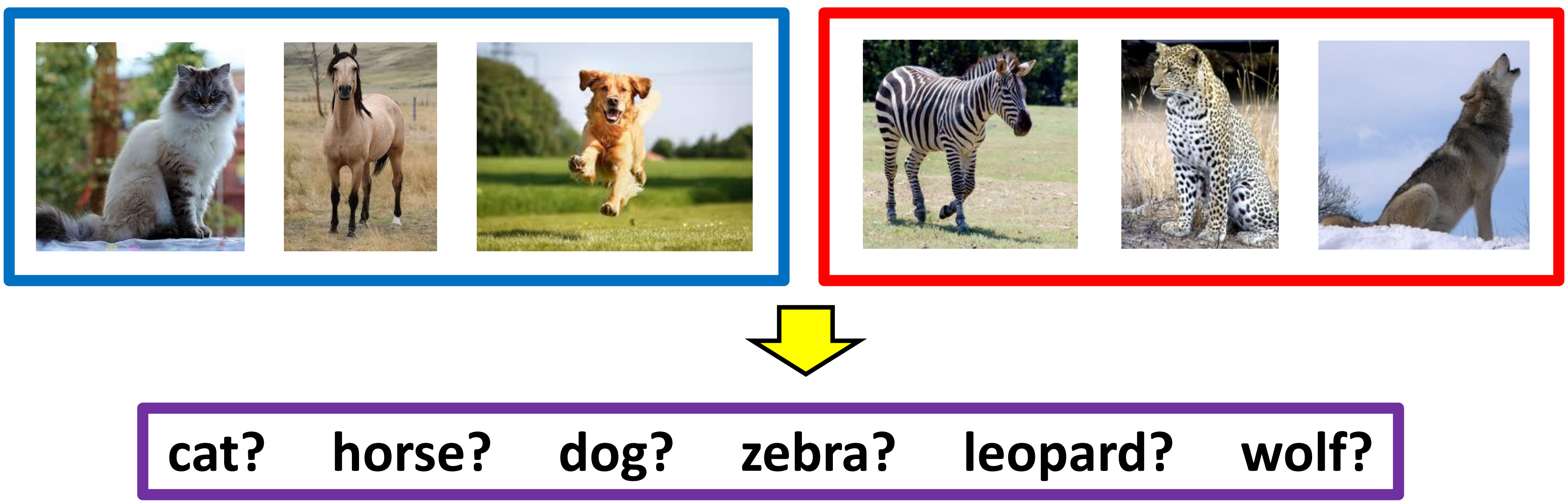
- learn from ***Seen*** classes' images and semantic embeddings

Testing of *"conventional"* ZSL:

- classify images from ***Unseen*** classes into ***Unseen*** classes, ***unrealistically*** assuming the absence of ***Seen*** classes

Testing of *"generalized"* ZSL:

- classify images from **BOTH** ***Seen*** & ***Unseen*** classes into the space of **BOTH** ***Seen*** & ***Unseen*** classes



Generalized ZSL (GZSL) is *nontrivial*!

- joint labeling space ***T*** **=** ***(S)een*** **+** ***(U)nseen***
- scoring function of each class $f_c(\boldsymbol{x})$ ➡ **direct stacking** $\hat{y} = \underset{c \in \boldsymbol{T}}{\operatorname{argmax}} f_c(\boldsymbol{x})$
- accuracy on ***Unseen*** classes suffers in GZSL

CUB dataset	$A_{U \to U}$	$A_{S \to S}$	$A_{U \to T}$	$A_{S \to T}$
SynC [Changpinyo et al., 2016]	54.4	73.0	**13.2**	72.0

$A_{P \to Q}$: accuracy of classifying images from ***P*** into the space of ***Q***

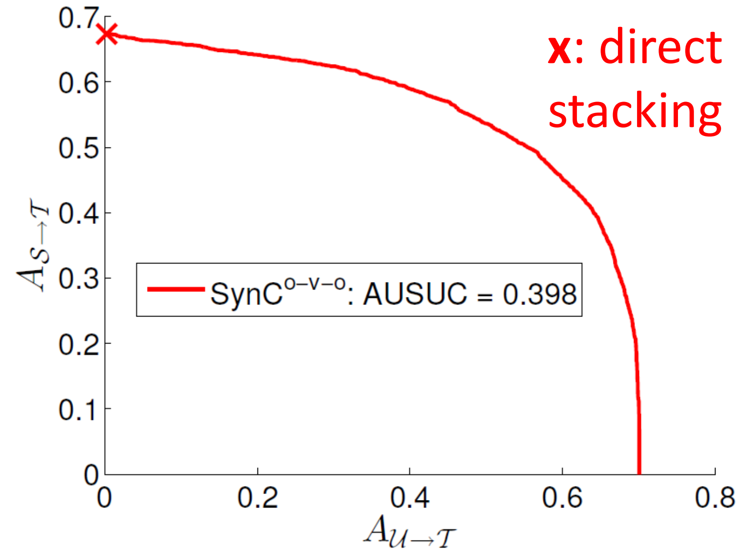
Calibrated stacking:

$$\hat{y} = \underset{c \in \boldsymbol{T}}{\operatorname{argmax}} f_c(\boldsymbol{x}) - \boldsymbol{\gamma} \mathbb{I}[c \in \boldsymbol{S}]$$

- effect: $\gamma \to \infty$: all into $\boldsymbol{U}$ $\quad \gamma \to -\infty$: all into $\boldsymbol{S}$
  $\gamma = 0$: direct stacking

*A*rea *U*nder *S*een *U*nseen *A*ccuracy *C*urve (AUSUC):

- varying γ leads to the **seen unseen accuracy curve (SUC)** of ($A_{\mathcal{U}\rightarrow\mathcal{T}}$, $A_{\mathcal{S}\rightarrow\mathcal{T}}$)
- **Area Under SUC (AUSUC)** to characterize the tradeoff



Extensive empirical studies

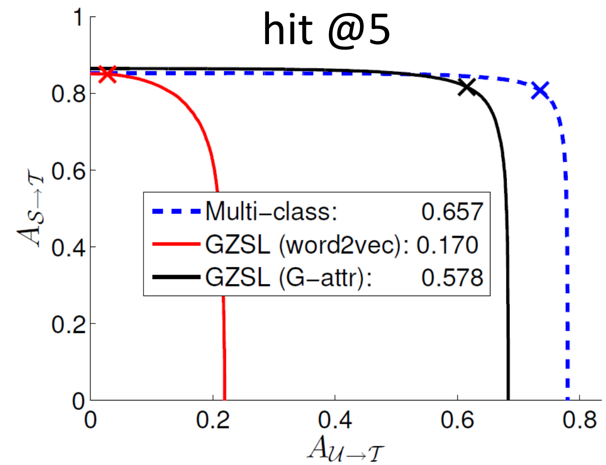
- **Datasets:** AwA, CUB, **ImageNet (|*S*| = 1K, |*U*| = 21K)**
- **Comparing ZSL algorithms:** DAP, IAP [Lampert et al., 2009], ConSE [Norouzi et al., 2014], SynC [Changpinyo et al., 2016]
- **Calibrated stacking** outperforms **novelty detection** [Socher et al., 2013] in adapting ZSL algorithms to GZSL

How far are we from ideal multi-class & GZSL performance?

- ImageNet-2K (1K ***Seen*** + 1K subsampled ***Unseen***)
- **multi-class classifiers** trained on data from ***S*** + ***U***
- **semantic embeddings of GZSL:**
  (1) WORD2VEC
  (2) G-attr: average visual features of each class of ***S*** + ***U***

Method		hit @1	hit @5
GZSL	WORD2VEC	0.04	0.17
	G-attr	**0.25**	**0.58**
multi-class classifiers		0.35	0.66

[measured in **AUSUC**]



- ***High quality semantic embeddings*** **is vital to GZSL!**

Poster ID 8