

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

Wei-Lun Chao\*<sup>1</sup>, Soravit Changpinyo\*<sup>1</sup>, Boqing Gong<sup>2</sup>, and Fei Sha<sup>1,3</sup>  
<sup>1</sup>U. of Southern California, <sup>2</sup>U. of Central Florida, <sup>3</sup>U. of California, Los Angeles



NSF IIS-1566511, 1065243, 1451412, 1513966, 1208500, CCF-1139148, USC Graduate Fellowship, a Google Research Award, an Alfred P. Sloan Research Fellowship and ARO# W911NF-12-1-0241 and W911NF-15-1-0484.

## Highlights

- Study *generalized zero-shot learning* (GZSL) Test data & possible labels from BOTH **Seen** + **Unseen** classes, not just from **Unseen** ones.
- Propose an effective *calibration* method to adapt ZSL algorithms to perform well in GZSL
- Develop a metric **AUSUC** for GZSL evaluation
- Establish a performance upper bound of GZSL via *idealized* semantic embeddings

## ZSL vs. Generalized ZSL

- Seen** classes come with labeled examples. **Unseen** classes come without.
  - Goal:** Expand classifiers and label space from Seen classes to Unseen ones = dealing with long-tailed object distributions and recognition in the wild
  - Relate Seen and Unseen classes with *Semantic embeddings* (attributes, word vectors, etc.)
- 
- From Derek Hoiem's slides

- Training:** Learn from **Seen** classes' images and semantic embeddings
- Testing:** (Conventional) Zero-Shot Learning (ZSL) Classifying images from **Unseen** into the label space of **Unseen**  
**Generalized Zero-Shot Learning (GZSL)** Classify images from BOTH **Seen** + **Unseen** into the label space of BOTH **Seen** + **Unseen**  
**Much more challenging!**

## ZSL algorithms in GZSL setting

- Joint labeling space of Seen (**S**) and Unseen (**U**):  
 $\mathcal{T} = \mathcal{S} \cup \mathcal{U}$
- Scoring function for each class  $f_c(\mathbf{x}), \forall c \in \mathcal{T}$ 
  - DAP [Lampert et al., CVPR 09]:  $f_u(\mathbf{x}) = \mathbf{w}(\mathbf{a}_u)^T \mathbf{x}$
  - ConSE [Norouzi et al., ICLR 14]:  $f_u(\mathbf{x}) = \cos(\mathbf{s}(\mathbf{x}), \mathbf{a}_u)$
  - SynC [Changpinyo et al., CVPR 16]:  $f_u(\mathbf{x}) = P(\mathbf{a}_u | \mathbf{x})$
- Classification by **Direct Stacking**  
 $\hat{y} = \arg \max_{c \in \mathcal{T}} f_c(\mathbf{x})$

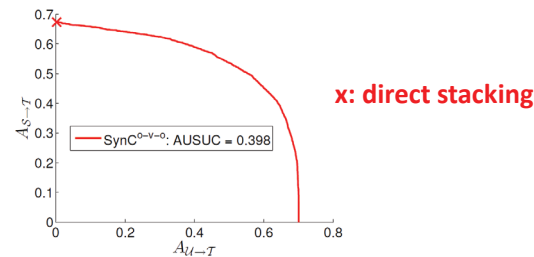
Method	AwA				CUB			
	$A_{U \rightarrow U}$	$A_{S \rightarrow S}$	$A_{U \rightarrow \mathcal{T}}$	$A_{S \rightarrow \mathcal{T}}$	$A_{U \rightarrow U}$	$A_{S \rightarrow S}$	$A_{U \rightarrow \mathcal{T}}$	$A_{S \rightarrow \mathcal{T}}$
DAP	51.1	78.5	2.4	77.9	38.8	56.0	4.0	55.1
ConSE	63.7	76.9	9.5	75.9	35.8	70.5	1.8	69.9
SynC	73.4	81.0	0.4	81.0	54.4	73.0	13.2	72.0

$A_{Z \rightarrow Y}$ : Accuracy of classifying images from **Z** into the space of **Y**

## Proposed Calibration Method & Metric

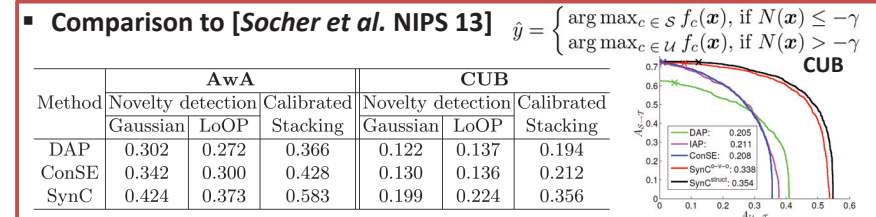
- Classification by **Calibrated Stacking**  
 $\hat{y} = \arg \max_{c \in \mathcal{T}} f_c(\mathbf{x}) - \gamma \mathbb{I}[c \in \mathcal{S}]$   
 $\gamma \rightarrow +\infty$  All into **U**     $\gamma \rightarrow -\infty$  All into **S**     $\gamma = 0$  Direct stacking

- Area Under Seen Unseen accuracy Curve (AUSUC)**
  - Varying the calibration factor leads to Seen-Unseen Accuracy Curve (SUC) of ( $A_{U \rightarrow \mathcal{T}}, A_{S \rightarrow \mathcal{T}}$ )
  - Area Under SUC (AUSUC) as the metric for GZSL



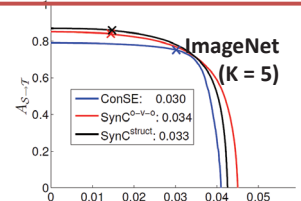
## Experiments & Analysis

- Datasets** ( $|\mathcal{S}|/|\mathcal{U}|$ ): AwA (40/10), CUB (150/50), ImageNet (1,000/20,842)
- Semantic embeddings:** attributes for AwA/CUB, word vectors for ImageNet
- Visual features:** 1,024-dim GoogLeNet features
- Evaluation:** **AUSUC** on (class-normalized) classification accuracy or Flat Hit@K  
AwA /CUB: also test on reserved 20% of data from the **S** seen classes  
ImageNet : also test on validation set



- Which ZSL method is more robust to GZSL?**

Unseen classes	Method	Flat hit@K			
		1	5	10	20
<i>2-hop</i> (1,509)	ConSE	0.042	0.168	0.247	0.347
	SynC	0.044	0.218	0.338	0.466
<i>All</i> (20,345)	ConSE	0.007	0.030	0.048	0.073
	SynC	0.006	0.034	0.059	0.097



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

**Analysis on ImageNet-2K:**  $|\mathcal{U}| = 1000$

- Multi-class classifiers trained on data from **S** & **U**
- Idealized** semantic embeddings (**G-attr**) = Average of visual features for each class

Method	Flat hit@K				
	1	5	10	20	
GZSL (SynC)	WORD2VEC	0.04	0.17	0.27	0.38
	G-attr from 1 image	0.08	0.25	0.33	0.42
	G-attr from 10 images	0.20	0.50	0.62	0.72
Multi-class classification	G-attr from all images	0.25	0.58	0.69	0.79
		0.35	0.66	0.75	0.82

**Analysis on ImageNet All:** 80% of **U** for G-attr

Flat hit@K ( $K = 1/5$ )  
WORD2VEC: 0.006/0.034  
G-attr from 1 image: 0.018/0.071  
G-attr from all images 0.067/0.236

