Zero-Shot Learning - The Good, the Bad and the Ugly

G29: Mainak Ghosh, Shayoni Halder, Smith Agarwal, Aadhithya Sankar, Shyam Arumugaswamy

Technical University of Munich

Zero Shot Learning (ZSL)

An approach to recognize objects whose instances may not have been seen during training, e.g. Classifying snake without seeing snake in training data.

Proposed Method

- **Image and Class Embedding**
  - Entire Images used to extract features with no pre-processing.
  - Image Embedding Units: 2048-dim top layer pooling units of ResNet-101.
  - Class Attributes: Per-class attributes for aPY, AWA, CUB and SUN. and Word2Vec embeddings for ImageNet.
- **Dataset Split**

				Number of Classes			Number of Images							
							At Training Time				At Evaluation Time			
							SS		PS		SS		PS	
Dataset	Size	Detail	Att	y	y^{tr}	y^{ts}	Total	y^{tr}	y^{ts}	y^{tr}	y^{ts}	y^{tr}	y^{ts}	y^{tr}	y^{ts}
SUN [28]	medium	fine	102	717	580 + 65	72	14K	12900	0	10320	0	0	1440	2580	1440
CUB [38]	medium	fine	312	200	100 + 50	50	11K	8855	0	7057	0	0	2933	1764	2967
AWA [22]	medium	coarse	85	50	27 + 13	10	30K	24295	0	19832	0	0	6180	4958	5685
aPY [10]	small	coarse	64	32	15 + 5	12	15K	12695	0	5932	0	0	2644	1483	7924

  - The proposed dataset split(PS) does not contain any overlap with the ImageNet classes.
  - Standard Split(SS) and PS has no image from test classes during training. SS does not have images from training classes in test split but PS has.
  - SS considers all classes that are 2-hops and 3-hops away from the 1k ImageNet classes to test generalisation ability of the models.
  - PS considers 500, 1k and 5k most populated classes among the other 21k ImageNet classes.
  - The 2nd PS also considers the 500, 1k and 5k least populated classes and the final PS considers all the 20k classes of ImageNet.
- **Evaluation Criteria**
  - Standard Evaluation method averages Top-1 accuracy for all images to get accuracy of model. This promotes high performance on dense classes.
  - The proposed method measures average per-class top-1 accuracy.
  - For generalised ZSL, harmonic mean of accuracy on training and test classes are used:

$$H = \frac{2 * (acc_{y^{tr}} * acc_{y^{ts}})}{(acc_{y^{tr}} + acc_{y^{ts}})}$$

Zero Shot Learning Results

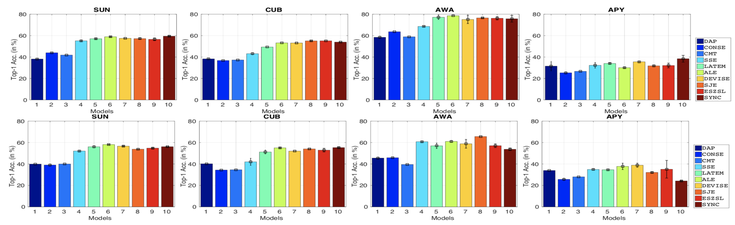
- **Reproducing Results**
  - Re-evaluating methods on the dataset indicate deviation from original results mainly due to:
    - Non convexity and sensibility to initialization(LATEM)
    - Random sampling in SGD(SJE)
    - Different hyperparameters due to random validation set(ESZSL) or hardcoded values(SSE)

	SUN		AWA	
Model	R	O	R	O
DAP [22]	22.1	22.2	41.4	41.4
SSE [42]	83.0	82.5	64.9	76.3
LATEM [39]	–	–	71.2	71.9
SJE [4]	–	–	67.2	66.7
ESZSL [32]	64.3	65.8	48.0	49.3
SYNC [7]	62.8	62.8	69.7	69.7

- **Standard and Proposed Splits**
  - Experiments with deep ResNet features reveal significantly lower results(DAP) due to differences in test and standard split; and usage of pre-trained MIT Places model.
  - New dataset splits insuring no test class belong to ImageNet1K, leading to observations varied from the original standard split.

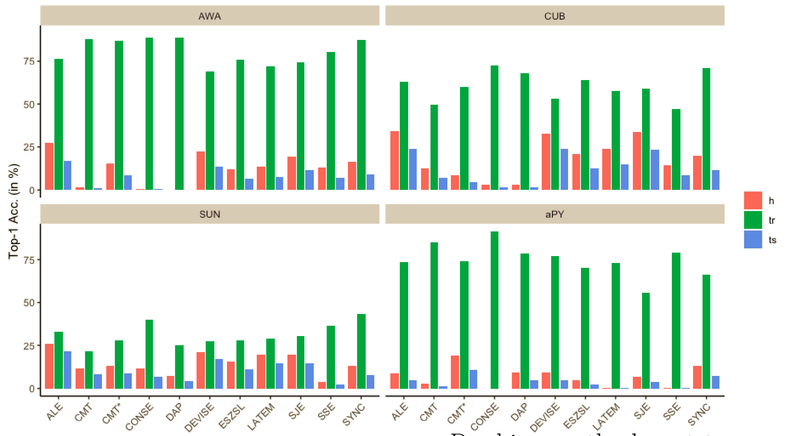
	SUN		CUB		AWA		aPY	
Method	SS	PS	SS	PS	SS	PS	SS	PS
DAP [22]	38.9	39.9	37.5	40.0	57.1	44.1	35.2	33.8
CONSE [26]	44.2	38.8	36.7	34.3	63.6	45.6	25.9	26.9
CMT [34]	41.9	39.9	37.3	34.6	58.9	39.5	26.9	28.0
SSE [42]	54.5	51.5	43.7	43.9	68.8	60.1	31.1	34.0
LATEM [39]	56.9	55.3	49.4	49.3	74.8	55.1	34.5	35.2
ALE [3]	59.1	**58.1**	53.2	54.9	**78.6**	59.9	30.9	39.7
DEVISE [11]	57.5	56.5	53.2	52.0	72.9	54.2	35.4	**39.8**
SJE [4]	57.1	53.7	**55.3**	53.9	76.7	**65.6**	32.0	32.9
ESZSL [32]	57.3	54.5	55.1	53.9	74.7	58.2	34.4	38.3
SYNC [7]	**59.1**	56.3	54.1	**55.6**	72.2	54.0	**39.7**	23.9

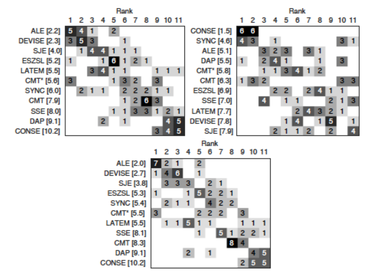
- **Robustness Method Ranking**
  - Evaluating robustness of 10 methods using 3 different validation splits with test split being intact.
  - Method ranking on their per-class top-1 accuracy.



Generalised ZSL Results

- In real world applications, image classification systems do not know whether an image belongs to a seen or unseen class in advance.
- Here, we use same models trained on zero-shot learning setting on our proposed splits (PS).
- Therefore, we evaluate performance on both training and test classes using held out images from test set.





- Ranking methods wrt ts (left), tr (right) and H (bottom)
- Methods with very high seen class accuracy have low unseen class accuracy
- Harmonic mean is a good measure that balances seen and unseen class accuracy
- Generalized zero-shot results are significantly lower than zero-shot results as training classes are included in the search space
- The results clearly suggest that methods should be optimized for test class accuracy and also for training class accuracy when evaluating zero-shot learning as both are equally important in real world

Conclusion

Compatibility learning framework have an edge over learning independent object, attribute classifier. Disjoint training & validation class split is an important factor of parameter tuning in ZSL setting.