Performance Variability in Zero-shot Classification

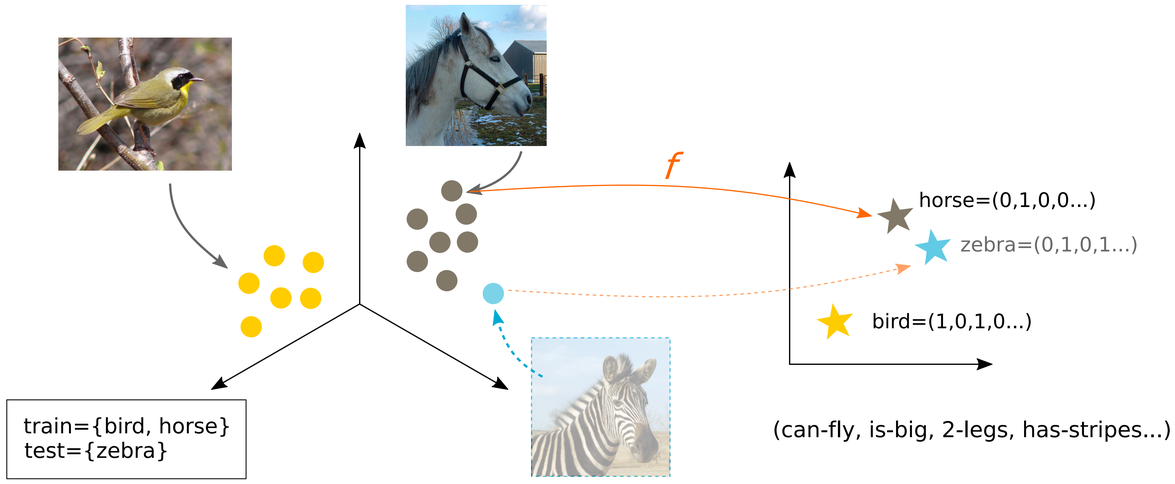
Matías Molina[1] · Jorge Sánchez[1,2]

[1] Universidad Nacional de Córdoba, Argentina. [2]CONICET

Motivation

Zero-shot classification (ZSC) is the task of learning predictors for classes not seen during training.

- How much does the ZSC performance vary over different class partitions?
- Is it average accuracy enough when choosing one method or another? Under which setup?



Problem setup

Given a training set

$$\mathcal{D}^{tr} = \{(x_i, y_i) \mid x_i \in \mathcal{X}^{tr}, y_i \in \mathcal{Y}^{tr}\}$$

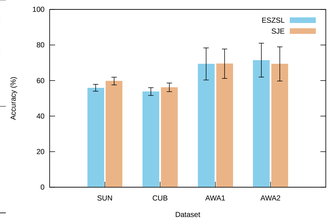
Goal:

- Learn $f : \mathcal{X} \rightarrow \mathcal{Y}$ from $\mathcal{D}^{tr}$
- Use f to classify images from a different set of categories $\mathcal{Y}^{ts}$. Where $\mathcal{Y}^{tr} \cap \mathcal{Y}^{ts} = \emptyset$

Variability Problem

Two ZSC methods: SJE and EZSL over 20+ train-test random partitions:

		SUN	CUB	AWA1	AWA2
Avg. acc.	ESZSL	55.90 (1.95)	53.49 (2.10)	69.66 (9.94)	71.10 (10.94)
	SJE	59.16 (2.37)	56.08 (3.03)	68.85 (7.96)	68.84 (11.16)
	p-value	0.000001	0.0012	0.7024	0.5028
Avg. per-class acc.	ESZSL	55.92 (1.94)	53.81 (2.20)	69.34 (9.02)	71.48 (9.54)
	SJE	59.73 (2.17)	56.19 (2.44)	69.48 (8.27)	69.34 (9.63)
	p-value	0.0000005	0.0000024	0.8736	0.1762



1 Strong performance variability, specially in coarse-grained data (AWA1 and AWA2)).

2 The accuracy difference might bias the selection between methods.

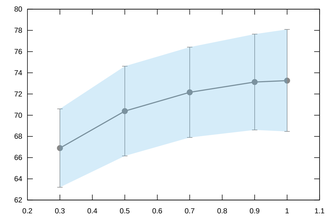
3 Are SJE and ESZSL the same? the p-values(*) show that for the fine-grained cases we can reject this hypothesis.

* We use the Wilcoxon test based on the fact that the data is paired w.r.t. each train-test partition.

Ensemble learning

Ensemble of n ESZSL models trained with a proportion s of the original training set. For $n = 90$:

	s 0.3	0.5	0.7	0.9	baseline
SUN	55.61 (2.16)	56.81 (2.02)	56.77 (1.98)	57.03 (1.73)	56.91 (1.63)
CUB	50.89 (2.92)	53.45 (2.84)	54.39 (2.84)	54.83 (2.72)	54.80 (2.82)
AWA1	65.35 (6.52)	68.38 (7.49)	69.70 (7.63)	70.52 (7.31)	70.62 (7.32)
AWA2	66.90 (3.70)	70.39 (4.23)	72.16 (4.26)	73.13 (4.52)	73.26 (4.81)



1 As the proportion s increases, the result approaches the baseline.

2 The standard deviation decreases marginally but with a considerable loss in performance (more noticeable in coarse-grained cases).

3 The use of ensemble does not lead to an increase on the overall ZSC performance.

References

[1] Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele. Evaluation of output embeddings for fine-grained image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[2] B. Romera-Paredes and P. Torr. An embarrassingly simple approach to zero-shot learning. In *International Conference on Machine Learning*, pages 2152–2161, 2015.

[3] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata. Zero-shot learning-a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*, 2018.

Conclusions

1 The ZSC task suffers the problem of performance variability w.r.t the class partitions.

2 The accuracy difference might bias the selection between one model or another.

3 It is important to consider the variability to compare different methods.

4 The ensemble learning is not enough to reduce the variability without losing precision.

5 As general conclusion, we suggest to incorporate the variability to obtain a more comprehensive evaluation protocol in ZSC.