Performance Variability in Zero-Shot Classfication

Matías Molina

Universidad Nacional de Córdoba, Argentina.

LXAI Workshop - NeurIPS 2020

FAMAF UNC

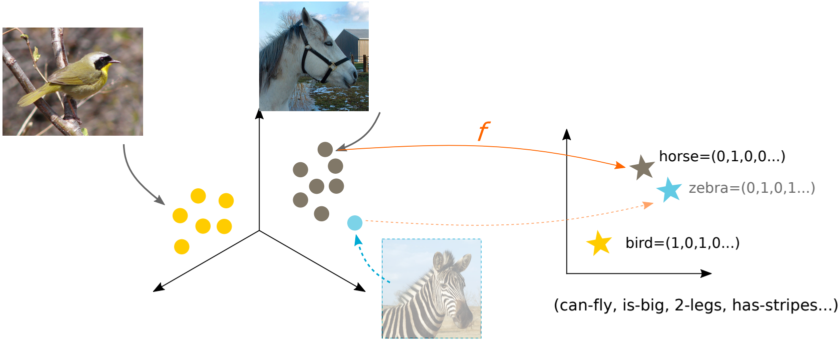
CONICET

ZSC Problem

Zero-shot classification is the task of learning predictors for samples not seen during training:

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

Learn $f : \mathcal{X} \to \mathcal{Y}$ from $\mathcal{D}^{tr}$ to be used on $\mathcal{Y}^{ts} \subset \mathcal{Y}$, where $\mathcal{Y}^{ts} \cap \mathcal{Y}^{tr} = \emptyset$

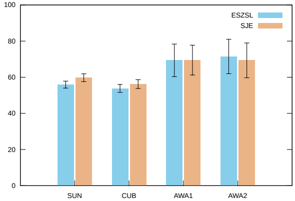


train={bird, horse}
test={zebra}

Variability

- *Xian et.al.*[1] propose a train-test partition to compare different ZSC methods. But, *How much does the ZSC performance vary over different class partitions?*
- Two popular methods, SJE and ESZSL, over different class partitions randomly created:

		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



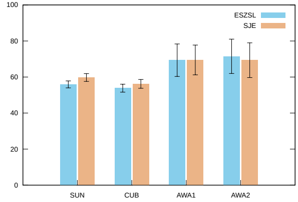
- Strong variability (higher in coarse-grained datasets: AWA1,AWA2).
- The variability is not dependent to the class imbalance.

[1] Zero-shot learning - A comprehensive evaluation of the good, the bad and the ugly, 2018

Variability

- *Xian et.al.*[1] propose a train-test partition to compare different ZSC methods. But, *How much does the ZSC performance vary over different class partitions?*
- Two popular methods, SJE and ESZSL, over different class partitions randomly created:

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



- The accuracy might bias the selection of one method even when their difference is not statistically significant:
- For the fine-grained cases, the p-value (*Wilcoxon signed-rank test*) is fairly low, we can reject the null hypothesis. → SJE > ESZSL.

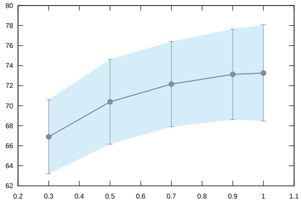
Ensemble learning

- We adapt the Bootstrap Aggregation(Bagging) technique by generating different subset of categories to train each predictor.
- We ensemble n ESZSL models. Each model is trained using a training set generated with a proportion s of the original set of categories.
- Predictors are combined by some voting scheme, e.g., $\hat{f}(x) = \arg\max_y \{\sum_i p_i(y|x)\}$

Ensemble learning

- Ensemble of $n = 90$ ESZSL models, of size s:

	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)



Enseble results for AWA2 dataset

- As the proportion s increases, the result approaches the baseline.
- The standard deviation may marginally decrease but with a considerable loss in performance (more noticeable in coarse-grained cases)

Conclusions

- ZSC strongly suffers from the variability problem w.r.t. the class partitioning.
- The variability is higher for coarse-grained dataset and lower for the fine-grained.
- It is important to consider the variability to obtain a more comprehensive evaluation process.
- The ensemble learning is not enough to reduce the variability without losing precision.
- In summary, we suggest that is important to complement the evaluation of ZSC by considering the performance variability.