# Performance Variability in Zero-Shot Classification

**Matías Molina**
Universidad Nacional de Córdoba
matias.molina@unc.edu.ar

**Jorge Sánchez**
CONICET
jorge.sanchez@unc.edu.ar

## Abstract

Zero-shot classification (ZSC) is the task of learning predictors for classes not seen during training. Although the different methods in the literature are evaluated using the same class splits, little is known about their stability under different class partitions. In this work we show experimentally that ZSC performance exhibits strong variability under changing training setups. We propose the use ensemble learning as an attempt to mitigate this phenomena.

## 1 Motivation

Training classifiers on specific non-generic domains requires a non negligible effort on data annotation. Although this might be easy for some of the target categories, it might become too costly for others due to the long-tail distribution of samples. This has motivated the development of models that can be trained on little data (few-shot learning) or no data at all (zero-shot learning). In zero-shot classification (ZSC) [5] we are given a set of labeled samples from a known set of categories and the goal is to learn a model that is able to cast predictions over a set of categories not seen during training. Despite having identified the difficulties and particularities in the evaluation of different approaches [9], little attention has been paid to the effect of considering different training class partitions for a given problem. Given the large number and diversity of models proposed in the literature in the recent years, we believe this is an important factor to be considered when choosing between competing approaches. In this work, we start exploring this problem. Our preliminary experiments using different datasets of varying granularity and two simple baselines confirm our hypothesis: performance differences observed in the literature might be not as significant as it seems due to the large variability observed across different subsets of training classes.

## 2 Experiments and Discussion

In ZSC we are given a training set $\mathcal{D}^{tr} = \{(x_i, y_i) \mid x_i \in \mathcal{X}, y_i \in \mathcal{Y}^{tr}\}$. The goal is to learn a mapping $f : \mathcal{X} \to \mathcal{Y}$ from $\mathcal{D}^{tr}$ that can be used to classify samples over a different set $\mathcal{Y}^{ts} \subset \mathcal{Y}$. We consider the standard ZSC setting, where $\mathcal{Y}^{tr} \cap \mathcal{Y}^{ts} = \varnothing$. Given a representation $z_y \in \mathbb{R}^E$ for each $y \in \mathcal{Y}$, a common approach [1, 7] is to learn a function $F : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$ to reflect the degree at which $x$ and $z$ agree on a given concept. Given a test sample $x$, its class is predicted as $\hat{y} = \mathrm{argmax}_{y \in \mathcal{Y}^{ts}} F(x, z_y)$.

The work of Xian *et al.* [9] identified several problems in the evaluation methodology used in the ZSC literature. One key contribution of their work was the proposal of fixed set of train/test class splits for different datasets. Although this addresses many of the evaluation problems identified in [9], it does not considers the effect of varying training class partitions. We believe analyzing not only the mean but also the variability of the zero-shot predictive performance under changing training configurations is an important factor towards a more thoughtful evaluation of the different methods. Our work is a first step in that direction.

Table 2 shows the mean and standard deviation over different training partitions for two simple baselines, ESZSL[7] and SJE[1], on two fine-grained (SUN[6] and CUB[8]) and two coarse-grained

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

Table 1: Soft ensemble results. Top-1 average per-class accuracy and its std. deviation, for $n$ = 90.

| $s$ | 0.3 | 0.5 | 0.7 | 0.9 |
|---|---|---|---|---|
| SUN | 55.61 (2.16) | 56.81 (2.02) | 56.77 (1.98) | 57.03 (1.73) |
| CUB | 50.89 (2.92) | 53.45 (2.84) | 54.39 (2.84) | 54.83 (2.72) |
| AWA1 | 65.35 (6.52) | 68.38 (7.49) | 69.70 (7.63) | 70.52 (7.31) |
| AWA2 | 66.90 (3.70) | 70.39 (4.23) | 72.16 (4.26) | 73.13 (4.52) |

Table 2: Top-1 accuracy and average top-1 per-class accuracy, std. deviation and p-value for 22 splits.

| | | 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 |

(AWA1[5] and AWA2[9]) datasets, using different class partitions sampled at random. We observe a great deal of variability whether the sample per class imbalance is considered (avg. acc.) or not (avg. per-class acc.). We observe that the difference in performance (as reported in the literature) might bias the selection of one method over the other even when their difference is not statistically significant. The table also show p-values of a Wilcoxon signed-rank test computed from 22 different partitions chosen at random. We see that while for the fine-grained cases we can reject the null hypothesis for a fairly low confidence level, this is not the case in the coarse-grained data regime. Although the difference in mean values seems high (for the standards observed in the literature), the variability observed in the experiments warns against choosing one method over the other.

**Ensemble learning for ZSC.** Beyond the identification of the variability problem, we ran experiments using standard ensemble techniques as an attempt to mitigate its effect. The idea is that by combining more than one predictor into a single model, it is possible to reduce the variance by averaging [3]. One popular approach is the *Bootstrap Aggregation* or *Bagging* meta-algorithm [2]. It is based on learning different predictors using different subsets of training samples and aggregating them via a suitable voting scheme. The *hard* voting scheme assigns the class predicted by the majority, *i.e.* $\hat{f}(x) = \mathrm{mode}\{f_1(x), ..., f_n(x)\}$. In *soft* voting, prediction is given by the highest score over all the models , *i.e.* $\hat{f}(x) = \mathrm{argmax}_y\{\sum_i F_i(x, z_y)\}$.

In the context of ZSC, we use different (random) subsets of training categories to generate the set of base predictors, *i.e.* we learn a set $n$ predictors using a proportion $s$ of randomly chosen classes from the original training set. We use hard and soft ensembles considering $n = \{10, 30, 50, 70, 90\}$ and $s = \{0.3, 0.5, 0.7, 0.9\}$, *e.g.* $(n, s) = (10, 0.3)$ means training 10 different models using $30\%$ of the full set of training categories. Each sub-problem is trained on a different subset of training classes. We sample 4 different sub-problems for each $(n, s)$ combination. Baseline performances are as follows: $56.91\ (1.63)$ on SUN, $54.80\ (2.82)$ on CUB, and $70.62\ (7.32)$, $73.26\ (4.81)$ on AWA1 and AWA2 respectively. We use the ResNet101 features and continuous attribute vectors from [9] and normalize both to unit norm. We found that as the proportion $s$ increases performance approaches the baseline, which is to be expected since the set of training categories tends to resemble the original set. The standard deviation may marginally decrease but with a considerable loss in performance. This situation is more noticeable in the case of AWA1 and AWA2, both coarse-grained datasets, compared to the others. Table 1 shows the ensemble results for $n$ = 90. [1] Beyond these observations, the use of ensemble does not lead to an increase on the overall ZSC performance. Alternatives to this formulation is the topic of our current research.

[1] Different combinations of voting schemes and accuracy metrics lead to similar conclusions.

## 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*, pages 2927–2936, 2015.

[2] L. Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[3] T. G. Dietterich. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer, 2000.

[4] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[5] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 951–958. IEEE, 2009.

[6] G. Patterson and J. Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2751–2758. IEEE, 2012.

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

[8] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[9] 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.