Evaluating the Evaluators: Are Current Few-Shot Learning Benchmarks Fit for Purpose?

THE UNIVERSITY of EDINBURGH informatics

Luísa B. Shimabucoro [1], Timothy M. Hospedales [2] and Henry Gouk [2]

[1]University of Sao Paulo [2]University of Edinburgh

Aims

While many Few-shot learning benchmarks have been developed over the years all of them focus exclusively on performance averaged over many tasks, thus neglecting the question of how to reliably evaluate and tune models trained for individual tasks, which often results in the inability to deploy models that rely on this type of evaluation [1]. We perform an investigation into task-level evaluation to answer the following questions:

Q1: How accurately are we able to estimate the performance of task-level models trained in the FSL regime?

Q2: Are estimates from existing evaluators well-correlated with the true performance of models?

Q3: By how much could performance in FSL be improved by incorporating accurate model selection procedures?

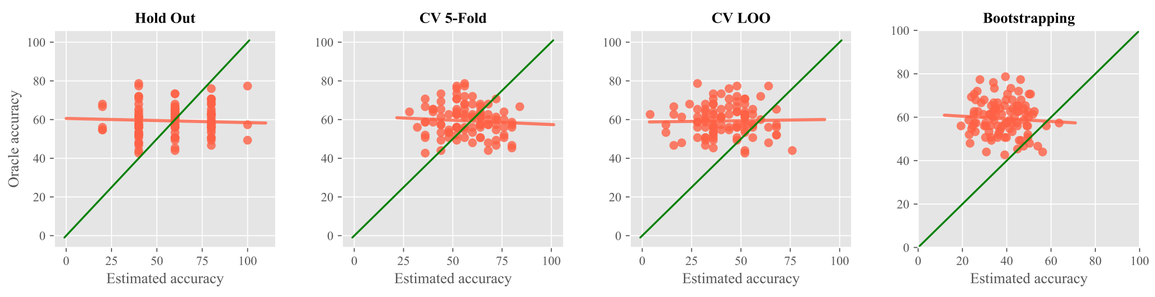

Figure 1. Scatter plots comparing the accuracy of a model on a query set (oracle) with the accuracy estimated by alternative evaluation methods that use only the support set. The ideal estimator would have the points almost co-linear and lying approximately on the diagonal line (green). Estimators with high bias and variance will exhibit a lack of co-linearity and be centred off the diagonal, respectively. We can see that all estimators have very high bias and variance, indicating that they do not provide reliable estimates of actual few-shot learning performance.

Few-Shot Model Validation and Selection

The evaluation methods used to perform the experiments are the following:

- **Oracle:** a dedicated query set is used to measure the model's accuracy.
- **Hold-out:** the support set is split into N folds and use one of them to test and the rest to train the model.
- **Cross-validation:** the support set is split into N folds and use one of them to test and the rest to train the model and we do that for N iterations, each one using a different chunk to estimate performance. At the end results are averaged.
- **Bootstrapping:** a new support set is constructed out of the original by sampling for the original set (with replacement) and the out-of-bag set is used as the query set.

Method

To answer these questions we analyse task-level performance of classic FSL methods (Baseline(++) [2], ProtoNet [3], MAML [4], R2D2 [5]) on miniImageNet, CIFAR-FS [6] and Meta Album [7]. For each meta-test episode, all the evaluators are given the support set only to estimate model performance for the aforementioned FSL algorithms, except for the oracle estimator, which uses a dedicated query set to give its estimate. With these results we calculate the mean absolute difference in accuracy between each estimator and the oracle accuracy, which is used to produce the plots shown.

Performance Estimation

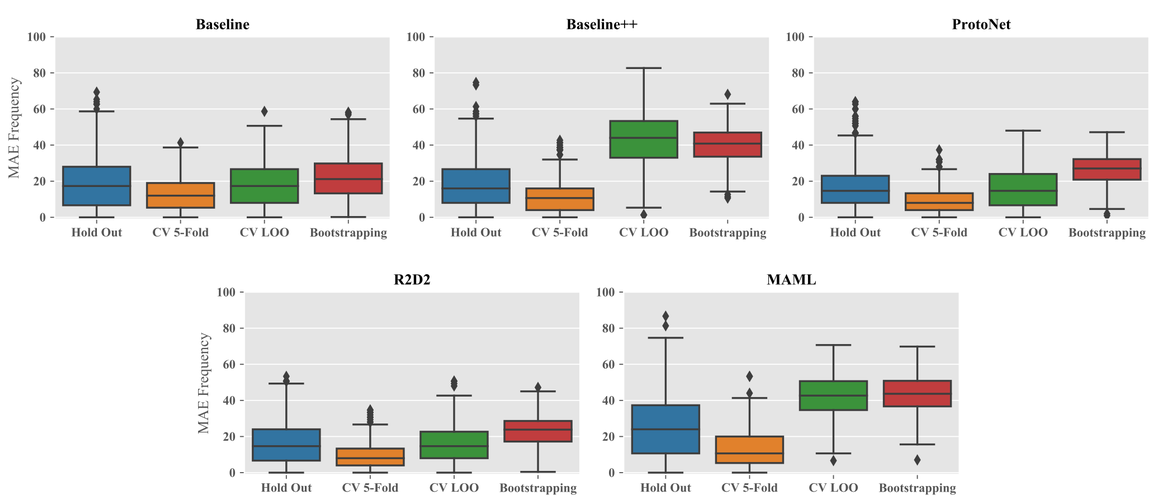

Figure 2. Box plots showing the distribution of absolute differences between the estimated accuracy and oracle accuracy on the meta-test episodes of miniImagenet. Distributions should be ideally concentrated as close to zero as possible, but we can see that a substantial proportion of the mass is far away from zero. This indicates that many of the performance estimates are unreliable.

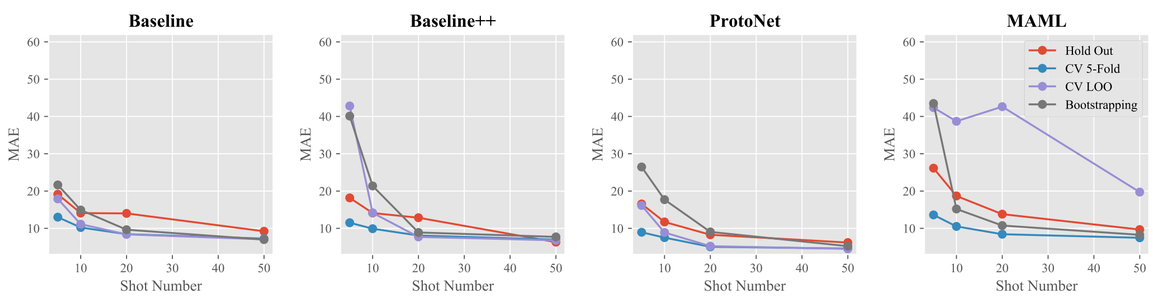

Figure 3. Dependence of estimator-oracle error on shot number. Estimator error is substantial in the few-shot regime.

Model Selection

We investigate how well these inaccurate performance estimates can be used to rank models, rather than provide precise estimates of performance.

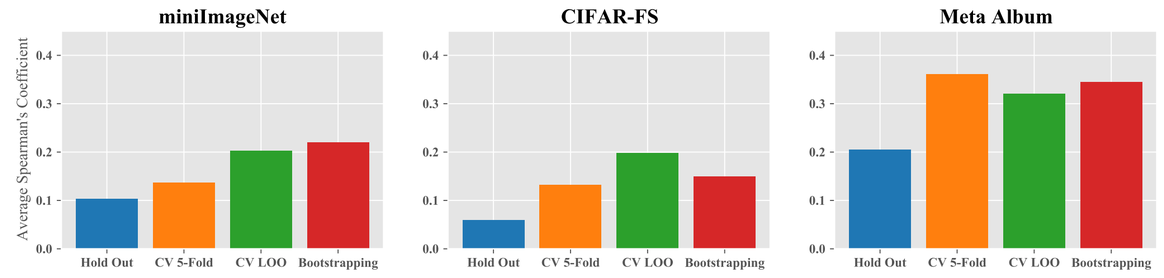

Figure 4. Mean Spearman correlation between the rankings produced by the oracle and the different performance estimators, computed across all the meta-test episodes in each dataset. A correlation coefficient of 1 indicates the same rankings, -1 indicates the opposite rankings, and 0 indicates that the rankings are unrelated.

Further Analysis of Cross-Validation

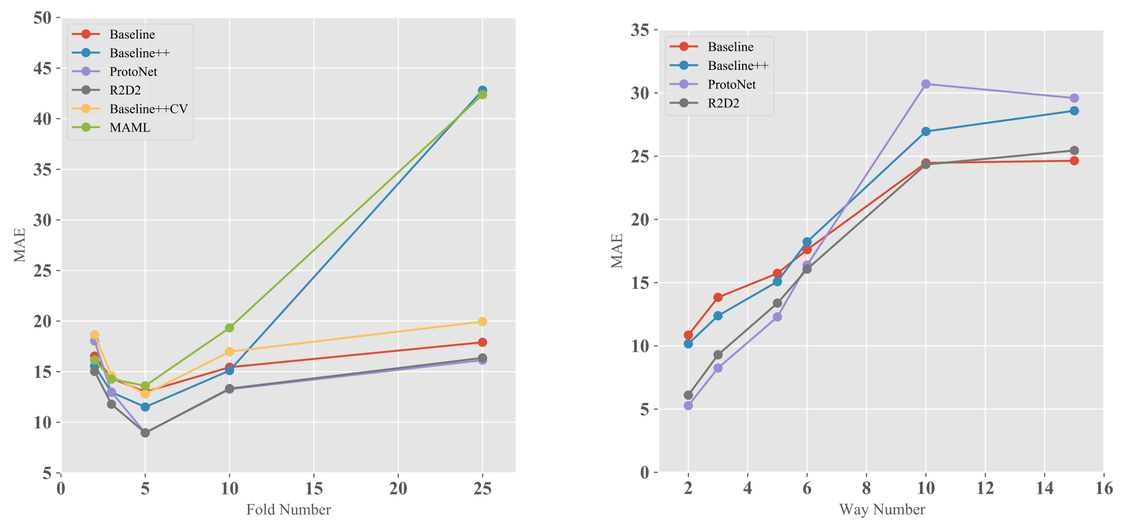

Figure 5. Further analysis of the estimation accuracy of cross validation performed on the miniImageNet dataset. Left: MAE between k-fold CV estimates and the oracle estimates, as a function of k. Right: MAE of LOO-CV, relative to the oracle, as the number of ways (and therefore class imbalance) is increased.

How can we improve FSL in practice?

Table 1. Aggregated Accuracy of the different baseline models. BaselineCV indicates that the ridge regularisation hyperparameter is tuned on a task-level basis using 5-fold cross validation.

Model	CIFAR-FS	miniImageNet	Meta-Album
Baseline	71.17 ± 0.727	59.36 ± 0.646	59.36 ± 1.688
BaselineCV	72.89 ± 0.737	62.11 ± 0.698	58.46 ± 1.745

Table 2. Aggregated Accuracy of Task-Level Model Selection using each of the performance estimators.

Model	CIFAR-FS	miniImageNet	Meta-Album
Oracle	80.11 ± 0.495	71.40 ± 0.464	63.53 ± 0.932
Hold-Out	71.65 ± 0.749	62.79 ± 0.710	56.30 ± 1.048
5-Fold CV	72.38 ± 0.726	63.73 ± 0.734	58.58 ± 1.005
LOO-CV	73.29 ± 0.738	64.34 ± 0.717	58.53 ± 1.014
Bootstrapping	73.44 ± 0.724	64.05 ± 0.737	58.62 ± 1.011

Conclusions

Q1: There are no combinations of learning algorithms and evaluators that are able to produce reliable performance estimates, but we find that 5-fold cross-validation is the best of all the bad options.

Q2: Current model evaluation procedures do not provide reliable rankings at the per-episode level, but methods based on re-sampling with a large number of iterations are most reliable.

Q3: Our results show that there is still a lot of room for improvement in the case of model selection, as evidenced by how well current model selection methods compare to models selected using the test set can perform.

References

[1] Gaël Varoquaux and Veronika Cheplygina.
Machine learning for medical imaging: methodological failures and recommendations for the future.
NPJ digital medicine, 5(1):48, 2022.

[2] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang.
A closer look at few-shot classification.
ICLR, 2019.

[3] Jake Snell, Kevin Swersky, and Richard S. Zemel.
Prototypical networks for few shot learning.
In *NIPS*, 2017.

[4] Chelsea Finn, Pieter Abbeel, and Sergey Levine.
Model-agnostic meta-learning for fast adaptation of deep networks.
In *ICML*, 2017.

[5] Luca Bertinetto, João F. Henriques, Philip H.S. Torr, and Andrea Vedaldi.
Meta-learning with differentiable closed-form solvers.
In *ICLR*, 2019.

[6] Alex Krizhevsky.
Learning multiple layers of features from tiny images, 2009.

[7] Ihsan Ullah, Dustin Carrión-Ojeda, Sergio Escalera, Isabelle Guyon, Mike Huisman, Felix Mohr, Jan N van Rijn, Haozhe Sun, Joaquin Vanschoren, and Phan Anh Vu.
Meta-album: Multi-domain meta-dataset for few-shot image classification.
Advances in Neural Information Processing Systems, 35:3232–3247, 2022.