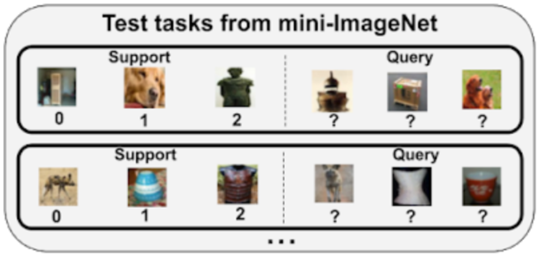
Meta-Dataset: A Dataset of Datasets for Learning to Learn from Few Examples

Chen Zhao

ICLR 2020, `https://arxiv.org/pdf/1903.03096.pdf`

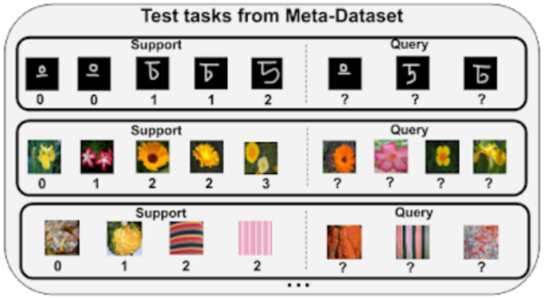
1 Comparison of Meta-Dataset with Previous Benchmarks

A popular dataset for studying few-shot classification is mini-ImageNet, a downsampled version of a subset of classes from ImageNet. This dataset contains 100 classes in total that are divided into training, validation and test class splits. While classes encountered at test time in benchmarks like mini-ImageNet have not been seen during training, they are still substantially similar to the training classes visually. Recent works reveal that this allows a model to perform competitively at test time simply by re-using features learned at training time, without necessarily demonstrating the capability to learn from the few examples presented to the model in the support set. In contrast, performing well on Meta-Dataset requires absorbing diverse information at training time and rapidly adapting it to solve significantly different tasks at test time that possibly originate from entirely unseen datasets.



Test tasks from mini-ImageNet. Each task is a classification problem between previously unseen (test) classes. The model can use the support set of a few labeled examples of the new classes to adapt to the task at hand and then predicts labels for the query examples of these new classes. The evaluation metric is the query set accuracy, averaged over examples within each task and across tasks.

While other recent papers have investigated training on mini-ImageNet and evaluating on different datasets, Meta-Dataset represents the largest-scale organized benchmark for cross-dataset, few-shot image classification to date. It also introduces a sampling algorithm for generating tasks of varying characteristics and difficulty, by varying the number of classes in each task, the number of available examples per class, introducing class imbalances and, for some datasets, varying the degree of similarity between the classes of each task. Some example test tasks from Meta-Dataset are shown below.



Test tasks from Meta-Dataset. Contrary to the mini-ImageNet tasks shown above, different tasks here originate from (the test classes of) different datasets. Further, the number of classes and the support set sizes differ across tasks and the support sets might be class-imbalanced.

2 Initial Investigation and Findings on Meta-Dataset

We benchmark two main families of few-shot learning models on Meta-Dataset: pre-training and meta-learning.

Pre-training simply trains a classifier (a neural network feature extractor followed by a linear classifier) on the training set of classes using supervised learning. Then, the examples of a test task can be classified either by fine-tuning the pre-trained feature extractor and training a new task-specific linear classifier, or by means of nearest-neighbor comparisons, where the prediction for each query example is the label of its nearest support example. Despite its “baseline” status in the few-shot classification literature, this approach has recently enjoyed a surge of attention and competitive results.

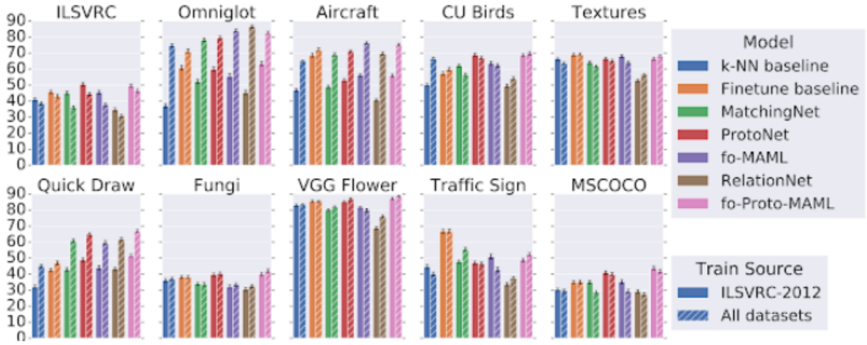
On the other hand, meta-learners construct a number of “training tasks” and their training objective explicitly reflects the goal of performing well on each task’s query set after having adapted to that task using the associated support set, capturing the ability that is required at test time to solve each test task. Each training task is created by randomly sampling a subset of training classes and some examples of those classes to play the role of support and query sets.

Below, we summarize some of our findings from evaluating pre-training and meta-learning models on Meta Dataset:

2.1 Existing approaches have trouble leveraging heterogeneous training data sources

We compared training models (from both pre-training and meta-learning approaches) using only the training classes of ImageNet to using all training classes from the datasets in Meta-Dataset, in order to measure the generalization gain from using a more expansive collection of training data. We singled out ImageNet for this purpose, because the features learned on ImageNet readily transfer to other datasets. The evaluation tasks applied to all models are derived from a held-out set of classes from the datasets used in training, with at least two additional datasets that are entirely held-out for evaluation (i.e., no classes from these datasets were used for training).

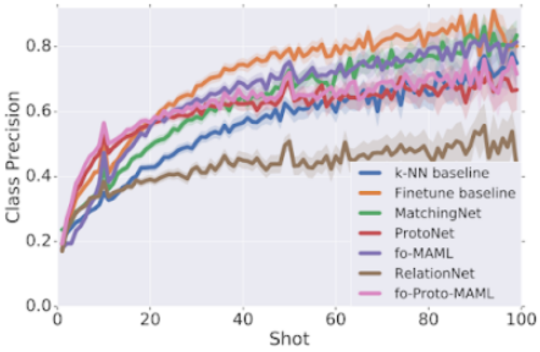
One might expect that training on more data, albeit heterogeneous, would generalize better on the test set. However, this is not always the case. Specifically, the following figure displays the accuracy of different models on test tasks of Meta-Dataset's ten datasets. We observe that the performance on test tasks coming from handwritten characters / doodles (Omniglot and Quickdraw) is significantly improved when having trained on all datasets, instead of ImageNet only. This is reasonable since these datasets are visually significantly different from ImageNet. However, for test tasks of natural image datasets, similar accuracy can be obtained by training on ImageNet only, revealing that current models cannot effectively leverage heterogeneous data towards improving in this regard.



Comparison of test performance on each dataset after having trained on ImageNet (ILSVRC-2012) only or on all datasets.

2.2 Some models are more capable than others of exploiting additional data at test time

We analyzed the performance of different models as a function of the number of available examples in each test task, uncovering an interesting trade-off: different models perform best with a particular number of training (support) samples. We observe that some models outshine the rest when there are very few examples ("shots") available (e.g., ProtoNet and our proposed fo-Proto-MAML) but don't exhibit a large improvement when given more, while other models are not well-suited for tasks with very few examples but improve at a quicker rate as more are given (e.g., Finetune baseline). However, since in practice we might not know in advance the number of examples that will be available at test time, one would like to identify a model that can best leverage any number of examples, without disproportionately suffering in a particular regime.



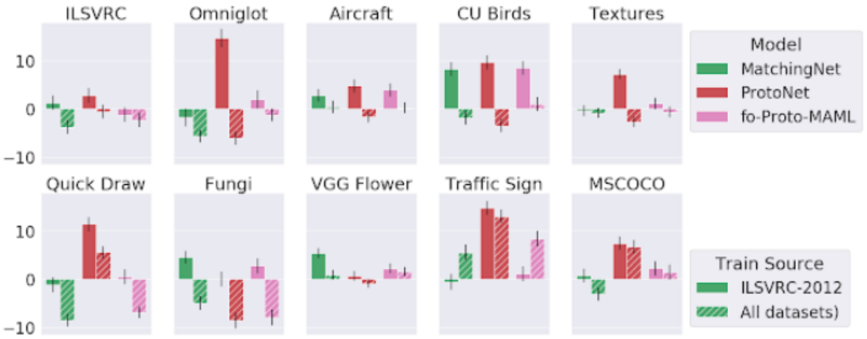
Comparison of test performance averaged across different datasets to the number of examples available per class in test tasks ("shots"). Performance is measured in terms of class precision: the proportion of the examples of a class that are correctly labeled, averaged across classes.

2.3 The adaptation algorithm of a meta-learner is more heavily responsible for its performance than the fact that it is trained end-to-end (i.e. meta-trained)

We developed a new set of baselines to measure the benefit of meta-learning. Specifically, for several meta-learners, we consider a non-meta-learned counterpart that pre-trains a feature extractor and then, at evaluation time only, applies the same adaptation algorithm as the respective meta-learner on those features. When training on ImageNet only, meta-training often helps a bit or at least doesn't hurt too much, but when training on all datasets, the results are mixed. This suggests that further work is needed to understand and improve upon meta-learning, especially across datasets.

3 Conclusion

Meta-Dataset introduces new challenges for few-shot classification. Our initial exploration has revealed limitations of existing methods, calling for additional research. Recent works have already reported exciting results on Meta-Dataset, for example using cleverly-designed task conditioning, more sophisticated hyper-parameter tuning, a 'meta-baseline' that combines the benefits of pre-training



Comparison of three different meta-learner variants to their corresponding inference-only baselines, when training on ImageNet (ILSVRC-1012) only or all datasets. Each bar represents the difference between meta-training and inference-only, so positive values indicate improved performance from meta-training.

and meta-learning, and finally using feature selection to specialize a universal representation for each task. We hope that Meta-Dataset will help drive research in this important sub-field of machine learning.