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

Eleni Triantafillou, Tyler Zhu, Vincent Dumoulin, Pascal Lamblin, Kelvin Xu,
Ross Goroshin, Carles Gelada, Kevin Swersky, Pierre-Antoine Manzagol, Hugo Larochelle
Google

Abstract

Few-shot classification refers to learning a classifier for new classes given only a few examples of them. While a plethora of models have emerged to tackle this recently, we find the current procedure and datasets that are used to systematically assess progress in this task lacking. To address this, we propose META-DATASET: a new benchmark for training and evaluating few-shot classifiers that is large-scale, consists of multiple datasets, and presents more natural and realistic tasks. The aim is to measure the ability of state-of-the-art models to leverage diverse sources of data to achieve higher generalization, and to evaluate that generalization ability in a more challenging and realistic setting. We additionally measure robustness to variations in the number of available examples and the number of classes.

1 Introduction

Few-shot learning refers to learning new concepts from few examples, an ability that humans naturally possess, but machines desperately lack. Improving on this aspect would lead to more efficient agents that can flexibly expand their knowledge as necessary without requiring large labeled datasets, that are usually unavailable. We focus on few-shot classification: classifying unseen examples into one of N new classes, given only a few reference examples of each new class. The key to success is efficiently leveraging other training data towards this goal, albeit originating from different classes. Recent progress in this direction can be naturally described in meta-learning terms: though we are not interested in learning about any training class in particular, we can exploit the training classes for the purpose of *learning to learn new classes from few examples*. The acquired learning procedure can then be directly applied to few-shot learning problems on new classes.

This intuition has inspired numerous models of increasing complexity for this problem. However, the manner in which they are evaluated is lacking. Notably, generalization is only examined between different classes of the same dataset. This setup is very constrained compared to the ultimate goal of few-shot learning "in the wild" where different distributions are encountered. We also elaborate on other unrealistic aspects of the previous setup in the following section. Our contribution is therefore to pinpoint these shortcoming and compose a new benchmark to address them. Through META-DATASET, we offer a large-scale, diverse environment for training and evaluating meta-learners that aims to expose and understand their limitations in operating in a realistic setting.

2 Background

Task Formulation The end-goal of few-shot classification is to produce a model which, given a new learning *episode* with N classes and a few labeled examples (k_c per class, $c \in 1, \ldots, N$), is able to generalize to unseen examples for that episode. In other words, the model learns from a training (*support*) set $\mathcal{S} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_K, y_K)\}$ (with $K = \sum_c k_c$) and is evaluated on

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

a held-out test (*query*) set $\mathcal{Q} = \{(\mathbf{x}_1^*, y_1^*), (\mathbf{x}_2^*, y_2^*), \dots, (\mathbf{x}_T^*, y_T^*)\}$. Each example $(\mathbf{x}, y)$ is formed of an input vector $\mathbf{x} \in \mathbb{R}^D$ and a class label $y \in \{1, \dots, N\}$. Episodes with balanced training sets (i.e., $k_c = k \ \forall c$) are usually described as "N-way, k-shot" episodes.

These evaluation episodes are constructed by sampling their N classes from a larger set $\mathcal{C}_{test}$ of classes and sampling the desired number of examples per class. A disjoint set $\mathcal{C}_{train}$ of classes is used to train the model. Few-shot learning does not prescribe a specific training procedure, but a common approach [1] involves matching the conditions in which the model is trained and evaluated. In other words, training often (but not always) proceeds in an episodic fashion.

Standard Datasets Two datasets have emerged as *de facto* benchmarks for few-shot learning. Omniglot [2] is a dataset of 1623 handwritten characters from 50 different alphabets and contains 20 examples per class (character). From a few-shot learning perspective, Omniglot is essentially considered solved: most recent methods obtain very high accuracies on the various standard flavours of ways and shots and leave very little room for improvement. The second benchmark, *mini*ImageNet [1], is formed out of 100 ImageNet [3] classes (64/16/20 for train/validation/test) and contains 600 examples per class. *mini*ImageNet, albeit harder than Omniglot, has the same property that most recent methods trained on it present similar accuracies when controlling for model capacity, and we believe the dataset is approaching its limit in terms of allowing to discriminate between the merits of competing approaches. We hypothesize this can be due to an artificially constrained setup. In particular, current benchmarks:

- Consider a fixed number of shots and ways. In contrast, real-life episodes are heterogeneous: they vary in terms of their number of classes and examples per class, and are unbalanced.
- Measure only within-dataset generalization. However, we really care equally about classifying different input distributions, e.g. abstract shapes and natural images.
- Ignore the relationships between classes when forming episodes. The coarse-grained classification of dogs and chairs presents different difficulties than the fine-grained classification of dog breeds, and current benchmarks do not establish a distinction between the two.

3 Approaches to Few-Shot Classification

Non-episodic Baselines A non-episodic approach is to first train a ConvNet classifier to predict all classes in $\mathcal{C}_{train}$ using standard batch training and treat the ConvNet until the penultimate layer as an embedding function $g(\mathbf{x})$. Then, given a test episode of new classes, we can compute embeddings for the support and query examples using g and adopt either one of the following methods to predict classes for the query examples in the test episode: 1. 'k-NN" baseline: perform nearest neighbour lookup in that embedding space (each query example is classified as its 'nearest' support example) 2. "Finetune" baseline: Construct a fully-connected layer for the new classes on top of those embeddings, train it on the support set and then use it to predict class labels for the query examples.

Episodic Models We experiment with three popular episodic models: Matching Networks [1], Prototypical Networks [4] and Model Agnostic Meta-Learning (MAML) [5]. In each training episode, these models compute for each query example $x^* \in \mathcal{Q}$, the distribution for its label $p(y^*|x^*, \mathcal{S})$ conditioned on the support set $\mathcal{S}$ and allow training this differentiably-parameterized conditional distribution end-to-end via gradient descent. The different models are distinguished by the manner in which this conditioning on the support set is realized. Prototypical Networks use the support examples to construct a prototype for each class as the average of the "embeddings" of the corresponding support points. $p(y^*|x^*, \mathcal{S})$ is then defined in terms of distances between the query x^* and each prototype, with x^* adopting the label of the closest prototype. Matching Networks (in its simplest form) labels each query point as a weighted linear combination of the support labels, where the weight of a label is again given by an estimate of similarity to the corresponding support point. MAML uses the support set to perform a small number of training steps for adapting the classifier's weights to the new classes and then predicts the class distribution for query examples using the adapted weights. In effect, the representation thus learns to constitute a good initialization that can be rapidly adapted for a new task. In all cases, the performance on the query set drives the update of the meta-learner's weights, which in these cases are solely the embedding weights.

4 META-DATASET: A New Few-Shot Classification Benchmark

META-DATASET offers a more realistic environment for assessing few-shot learning performance on a more realistic version of the task. Our approach therefore is twofold: 1) changing the data and 2) changing the formulation of the task (i.e., how episodes are generated). In this workshop contribution, we present early results on META-DATASET. We plan to make our code public upon finalizing the benchmark.

First, the data we propose to use is much larger in size than any previous benchmark, and is comprised of *multiple different existing datasets*. This allows to examine a more challenging generalization problem, to new datasets altogether. Specifically, META-DATASET leverages data from the following 10 datasets: ILSVRC-2012 (ImageNet) [3], Omniglot [2], Aircraft [6], CUB-200-2011 (Birds) [7], Describable Textures [8], Quick Draw [9], Fungi[1], VGG Flower [10], Traffic Signs [11], and MSCOCO [12], that are further described in Appendix A. However, to ensure that episodes correspond to realistic classification problems, all episodes generated in META-DATASET use classes from a single dataset at a time only. Two of these datasets, Traffic Signs and MSCOCO, are fully reserved for evaluation, meaning that no classes from them participate in the training set. The remaining ones contribute some classes to each of training, validation and test splits of classes.

Second, we modify the task formulation by introducing a new sampling procedure for episodes that more closely resembles realistic learning scenarios. Contrary to the usual setup, we vary the number of classes and training sizes of different episodes. Further, within a given episode, the "shot" varies from class to class, introducing imbalance. To obtain realistic imbalance ratios, we sample the number of examples of each class from a distribution derived from the relative class frequencies in the overall dataset. The exact sampling procedure is described in Appendix B.

Regarding class relationships, generally, the specific classes of an episode are selected uniformly at random as usual. However, for ILSVRC-2012, we leverage the class ontology to obtain a hierarchically-aware episode sampling procedure that shapes how likely certain classes may co-exist in an episode, aiming at more natural tasks. For example, it is arguably fairly unusual to classify frogs from laptops. A more natural task would instead be to classify desks from laptops, since these are more often observed simultaneously. For this, we first sample an internal node uniformly at random from the ontology and then select all the leaves of that node as the classes of the episode. We additionally make use of ImageNet's hierarchy for creating the class splits in a way that ensures that the test classes are indeed substantially different than the training classes: they belong to a different and disjoint sub-tree of the overall ontology. More details on this can be found in Appendix C.

5 Experiments

We present our experimental results in Table 1 and Table 2. In both tables we evaluate all methods on the test set of classes of each of the 10 datasets. The difference between the tables is the training source, i.e., the data that the models were trained on, which is (the training classes of) ILSVRC-only, and all datasets, respectively. To perform this all-dataset training on the non-episodic baselines, we consider a large classification problem between the training classes of all datasets. Meta-learners on the other hand first sample a dataset uniformly at random and then receive an episode from that dataset that is generated according to our sampling procedure described above.

For each meta-learner, the best performance was obtained by initializing its embedding weights from the weights that the k-NN baseline trained on ILSVRC-2012 converged to, and then "fine-tuning" them episodically. We found this pre-training really improved performance. Another interesting finding is that while Finetune does best with 100 adaptation steps, MAML does best with 7, highlighting the benefit of end-to-end training. In these preliminary results, under the constraint of memory requirement and training time, we use the first-order variant of MAML. While this has been shown to perform as well as full-MAML for classification [5], we will use full-MAML in our finalized table, for fair comparison. Model selection was performed on the validation set, episodically on either ILSVRC-2012 only or different datasets, depending on the corresponding training setting.

Tables 1 and 2 demonstrate the competency of meta-learners for few-shot classification across different datasets compared to the two standard baselines. Notably, the Prototypical Network model

[1] `https://sites.google.com/view/fgvc5/competitions/fgvcx/fungi`

Table 1: Results on META-DATASET using models trained on ILSVRC-2012 only.

Test Source	Method: Accuracy ± confidence				
	k-NN	Finetune	MatchingNet	ProtoNet	MAML
ILSVRC	34.70±0.95	38.34±1.12	40.89±1.08	**43.37±1.17**	38.10±1.13
Omniglot	59.84±0.96	59.19±1.18	61.85±1.00	**66.18±1.12**	54.00±1.47
Aircraft	36.47±0.93	**41.18±1.07**	**41.91±0.96**	**42.14±0.97**	**42.52±1.16**
Birds	40.38±1.09	45.82±1.25	54.26±1.16	**57.85±1.23**	50.78±1.32
Textures	56.45±0.78	58.06±0.88	**61.70±0.84**	**60.95±0.80**	**61.26±0.93**
Quick Draw	36.09±1.19	38.43±1.39	38.52±1.12	**44.02±1.35**	30.71±1.51
Fungi	23.70±0.97	22.20±0.92	27.21±0.97	**31.18±1.15**	20.35±0.87
VGG Flower	66.16±0.99	69.32±1.13	75.05±0.91	**79.89±0.90**	65.12±1.15
Traffic Signs	**44.81±1.47**	39.36±1.28	**45.36±1.31**	**44.04±1.24**	31.10±1.20
MSCOCO	29.69±1.00	30.25±1.17	32.32±1.08	**36.44±1.23**	25.17±1.15
Avg. rank	4	3.4	2.2	**1.35**	4.05

Table 2: Results on META-DATASET using models trained on All datasets.

Test Source	Method: Accuracy ± confidence				
	k-NN	Finetune	MatchingNet	ProtoNet	MAML
ILSVRC	25.88±0.83	25.84±0.83	35.88±0.98	**38.51±1.01**	30.56±1.00
Omniglot	**92.45±0.41**	85.20±0.73	90.21±0.46	91.32±0.50	78.05±0.98
Aircraft	54.60±0.97	58.22±1.02	**70.71±0.78**	**71.54±0.84**	68.62±0.90
Birds	36.74±1.01	38.56±1.08	59.28±1.06	**61.81±1.13**	54.59±1.24
Textures	50.06±0.77	48.37±0.82	**60.61±0.82**	**59.31±0.75**	**59.25±0.80**
Quick Draw	**59.54±1.08**	54.05±1.30	57.44±1.17	**60.99±1.21**	44.48±1.41
Fungi	24.60±0.95	22.90±0.95	31.10±1.04	**35.96±1.25**	21.12±0.88
VGG Flower	62.49±0.91	59.72±1.17	76.72±0.83	**81.06±0.87**	66.05±1.09
Traffic Signs	**41.68±1.46**	30.02±1.13	**43.20±1.33**	39.95±1.18	30.23±1.24
MSCOCO	23.55±0.99	23.01±0.96	26.87±1.00	**30.81±1.13**	21.13±1.06
Avg. rank	3.4	4.3	2.15	**1.4**	3.75

performs consistently well. To further understand the advantages and weaknesses of the various models, we analyzed their performance for different "shots" and "ways" (the relevant plots can be found in Appendix D). In particular, we computed how the classification accuracy changes as a function of the way, and how a class' precision (proportion of correctly classified examples of that class) changes as a function of its shot. Our analysis indicates that, as expected, in general all models' performance improves when the way decreases and the shot increases. Interestingly, we also found that Prototypical Networks' performance plateaus as the shot increases, contrary to the other two meta-learning models, a fact that we think could be attributed to the representation capacity of the prototypes. Also, the Finetune baseline and MAML more greatly benefit from the increase of the shot compared to other models, which is expected since they perform further fine-tuning based on the support set.

Finally, we examine the model's ability to leverage data of diverse sources. This can be assessed by comparing the values between Tables 1 and 2, and from the visualization of the same data in Appendix E. We do not always observe an improvement from training on all datasets over training on ILSVRC-2012 only which suggests that this direction requires further research.

6 Conclusion

We have introduced a new large-scale, diverse, and realistic environment for training and testing meta-learners for the task of few-shot classification. Our preliminary results suggest that popular meta-learners are indeed performing favorably to baselines for most datasets that we consider. A direction that evidently requires further research is how to better leverage data from multiple sources at training time: our preliminary results do not (always) demonstrate a gain from training on all datasets as opposed to training only on ILSVRC-2012.

References

[1] Oriol Vinyals, Charles Blundell, Tim Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems*, pages 3630–3638, 2016.

[2] Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

[3] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[4] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, pages 4077–4087, 2017.

[5] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. *arXiv preprint arXiv:1703.03400*, 2017.

[6] S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. Technical report, 2013.

[7] Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-ucsd birds 200. 2010.

[8] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[9] Jonas Jongejan, Henry Rowley, Takashi Kawashima, Jongmin Kim, and Nick Fox-Gieg. The Quick, Draw! – A.I. experiment, 2016.

[10] M-E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing*, 2008.

[11] Sebastian Houben, Johannes Stallkamp, Jan Salmen, Marc Schlipsing, and Christian Igel. Detection of traffic signs in real-world images: The German Traffic Sign Detection Benchmark. In *International Joint Conference on Neural Networks*, number 1288, 2013.

[12] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.