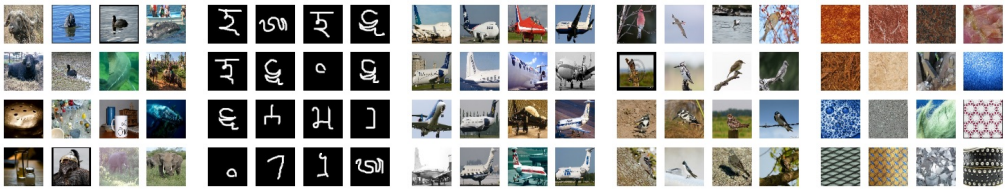
Appendix for Meta-Dataset: A Dataset of Datasets for Learning to Learn from Few Examples

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

Appendix A Datasets

META-DATASET is formed of data originating from 10 different image datasets, all of whose images we resize to 84 x 84. A complete list of the datasets we use is the following.



(a) ImageNet (b) Omniglot (c) Aircraft (d) Birds (e) DTD

(f) Quick Draw (g) Fungi (h) VGG Flower (i) Traffic Signs (j) MSCOCO

Figure 1: Training examples taken from the various datasets forming META-DATASET.

ILSVRC-2012 (ImageNet) [1] A dataset of natural images from 1000 categories (Figure 1a).

Omniglot [2] A dataset of images of 1623 handwritten characters from 50 different alphabets, with 20 examples per class (Figure 1b). While recently [3] proposed a new split for this dataset, we instead make use of the original intended split [2] which is more challenging since the split is on the level of alphabets (30 training alphabets and 20 evaluation alphabets), not characters from those alphabets, therefore posing a more challenging generalization problem. Out of the 30 training alphabets, we hold out the 5 smallest ones (i.e. with the least number of character classes) to form our validation set, and use the remaining 25 for training.

Aircraft [4] A dataset of images of aircrafts spanning 102 model variants, with 100 images per class (Figure 1c).

CUB-200-2011 (Birds) [5] A dataset for fine-grained classification of 200 different bird species. (Figure 1d).

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

Describable Textures (DTD) [6] A texture database, consisting of 5640 images, organized according to a list of 47 terms (categories) inspired from human perception. (Figure 1e).

Quick Draw [7] A dataset of 50 million black-and-white drawings across 345 categories, contributed by players of the game Quick, Draw! (Figure 1f).

Fungi [8] A large dataset of approximately 100K images of nearly 1,500 wild mushrooms species (Figure 1g).

VGG Flower [9] A dataset of natural images of 102 flower categories. The flowers chosen to be ones commonly occurring in the United Kingdom. Each class consists of between 40 and 258 images. (Figure 1h).

Traffic Signs [10] A dataset of 50,000 images of German road signs in 43 classes (Figure 1i).

MSCOCO [11] A dataset of images collected from Flickr with 1.5 million object instances belonging to 80 classes labelled and localized using bounding boxes. We choose the train2017 split and create images crops from original images using each object instance's groundtruth bounding box. (Figure 1j).

Appendix B Algorithm for Episode Sampling

The algorithm we use to generate an episode can be broken-down into the following steps: 1) Sample the classes of the episode, 2) Compute the size of the query set, 3) Compute the total size of the support set, 4) Compute the shot of each class. We describe each of these below.

Sampling the number of classes. For ILSVRC-2012, the classes are sampled using the algorithm described in Section C. For all other datasets, the number of classes is sampled uniformly at random from the closed interval $[5, \text{MAX-CLASSES}]$ where MAX-CLASSES is either 50 or as many as there are available if that number is less than 50. Note that we assume that we always have at least 5 classes available in each class split (train, validation, test) of each given dataset.

Computing the size of the query set. The query set is class-balanced, reflecting the fact that we care equally to perform well on all classes of an episode. The number of query images *per class* is computed as:

$$q = \min\{10, (\min_{c \in \mathcal{C}} 0.5 * |Im(c)|)\}$$

where $\mathcal{C}$ is the set of selected classes and $Im(c)$ denotes the set of images belonging to class c. The min over classes ensures that each class has at least q images to add to the query set (otherwise we could not necessarily create a balanced query set). The 0.5 multiplier is in place to allow sufficiently many images to remain for the support set too, and the minimum with 10 is used to preclude from having too small query sets.

Computing the size of the support set. The total support set size is computed as:

$$|\mathcal{S}| = \min\{500, \beta \sum_{c \in \mathcal{C}} \min\{100, |Im(c)| - q\}\}$$

where β is sampled uniformly from the closed interval $[0, 1]$. Intuitively, each class on average contributes either all its remaining examples (after placing q of them in the query set) if they're less than 100 or 100 otherwise, to avoid having too large support sets. The multiplication with β enables the potential generation of smaller support sets even when multiple images are available, since we are also interested in examining that end of the spectrum. Finally, we cap the total support set size to 500.

Computing the shot of each class. The proportion of the support set that will be devoted to class c is computed as:

$$R_c = \frac{\exp(\alpha_c)|Im(c)|}{\sum_{c' \in \mathcal{C}} \exp(\alpha'_c)|Im(c')|}$$

where α_c is sampled uniformly from the closed interval $[\log(0.5), \log(2)]$. Intuitively, the unnormalized proportion of the support set that will be occupied by class c is a noisy version of the total number of images of that class in the dataset $Im(c)$. This design choice is made in the hopes of obtaining realistisc class ratios, under the hypothesis that the dataset class statistics are a reasonable approximation of the real-world statistics of appearances of the corresponding classes.

The shot of a class c is then set to $k_c = R_c|\mathcal{S}|$, or as many examples are available of that class at this point (after having removed q for the query set), if they are less than that. If the shot of a class is computed to be 0 (which can happen because of the β multiplier above), we set it to 1 instead.

Appendix C The Hierarchy of ImageNet and How we Exploit it

ImageNet is a dataset comprised of 82,115 'synsets'. A synset is a concept that belongs to a larger ontology (ImageNet's synsets are based on the WordNet ontology). ImageNet provides IS-A relationships for the synsets it contains, therefore defining a DAG over its synsets. In this benchmark, we only use the 1000 synsets that were chosen for the ILSVRC 2012 classification challenge as classes that can appear in our episodes. However, we leverage the ontology DAG for defining a sampling procedure that determines which of these 1000 classes should co-occur in the each episode.

For this purpose, we consider a sub-graph of the overall DAG that consists of only the 1000 synsets of ILSVRC-2012 and their ancestors. In particular, these 1000 synsets are all and only the leaves of the DAG. We then further 'cut' this sub-graph into three pieces, for the training, validation, and test splits, such that there is no overlap between the leaves of any of these pieces. For this, we select the synset 'carnivore' as the root of the validation sub-graph, and the synset 'device' as the root of the test sub-graph. All the leaves that are reachable by 'carnivore' and 'device' form the sets of validation and test classes respectively. All remaining leaves constitute the training classes. This separation leads to training taking place on animals that are not carnivores, validation taking place on carnivores and testing taking place on inanimate devices, such as various tools and instruments. These splits were chosen for the objective of splitting the classes into approximately 70 / 15 / 15 (%) for training / validation / testing, and ensuring that the three groups are comprised of substantially different semantic classes. This leads to 712 training, 202 validation and 188 test classes.

The procedure for sampling the classes for an episode of a given split is then as follows: sample an internal node of that split's sub-graph uniformly at random, and use all leaves spanned by that node as the classes of the episode. We limit the possible number of classes of an episode to 50, and therefore any internal nodes that span more than 50 leaves are excluded from the pool that we sample from. For example, if the sampled internal node is close to the 'bottom' of the sub-graph, the resulting classification task will be finer-grained than if the sampled node is 'higher' in the tree and this spans more general concepts. In future work we plan to investigate the effect of fine versus coarse grained tasks on performance.

Appendix D Analysis of Performance Across Shots and Ways

Figure 2: Train on ILSVRC-2012 and Evaluate on All Datasets

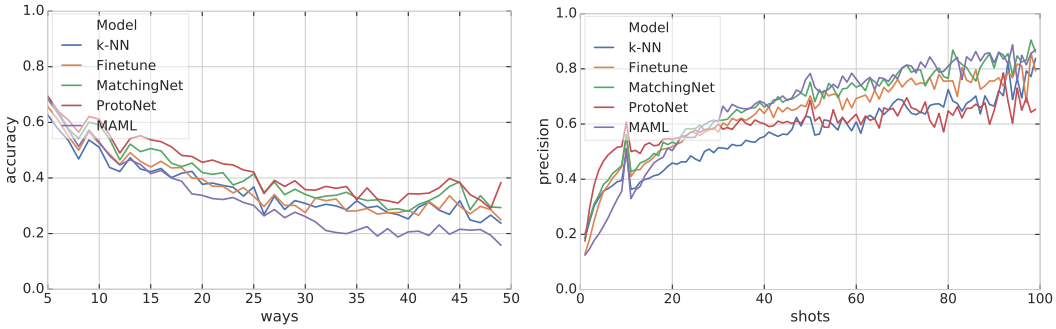


Figure 3: Train on ILSVRC-2012 and Evaluate on ILSVRC-2012

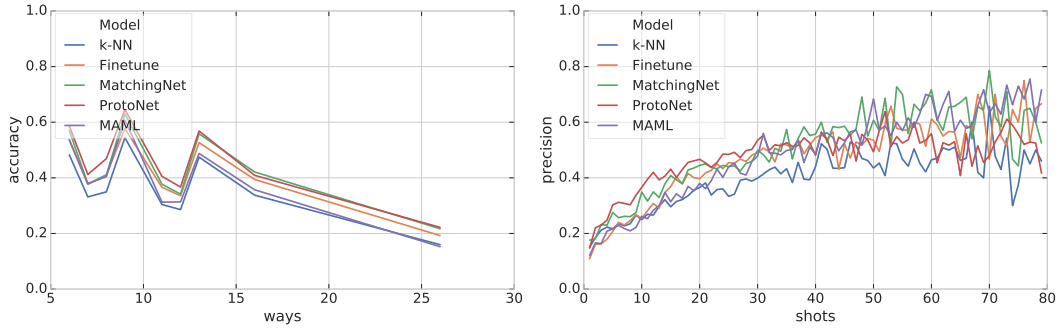


Figure 4: Train on All Datasets and Evaluate on All Datasets

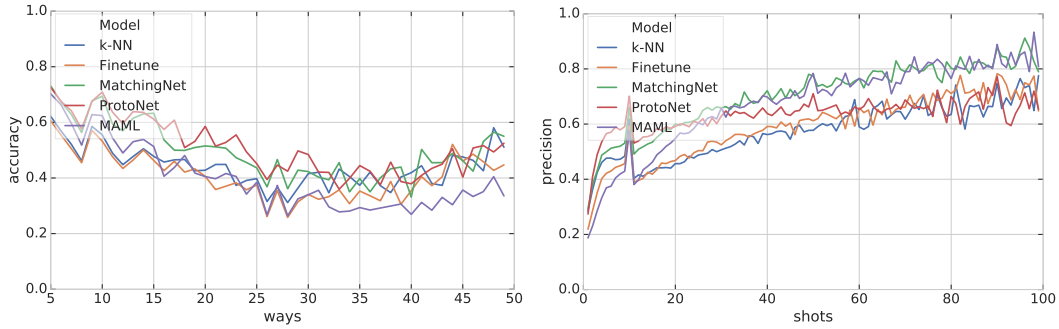
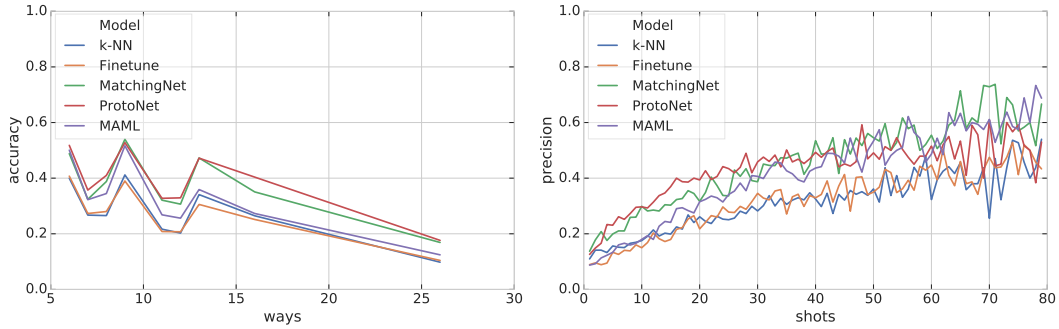


Figure 5: Train on All Datasets and Evaluate on ILSVRC-2012



Appendix E Training on more datasets than ILSVRC-2012

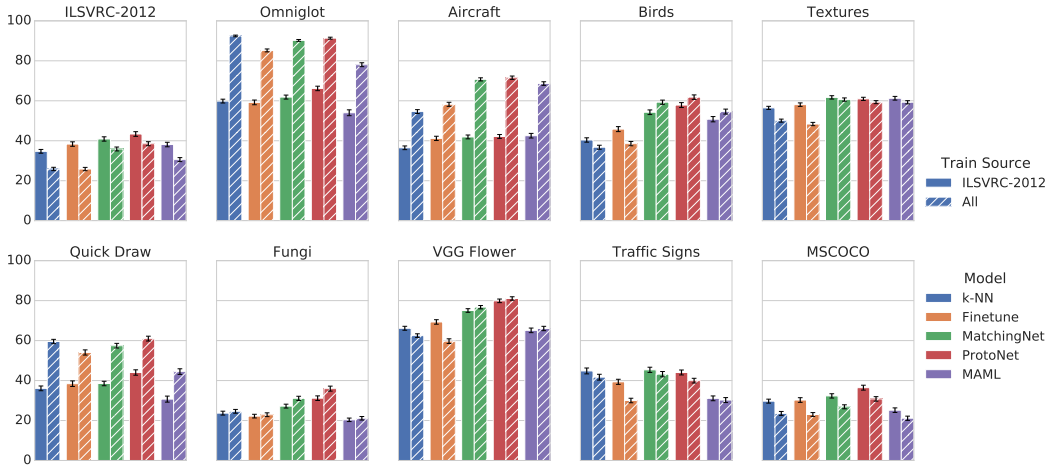
The results in the following table represent the 'element-wise' difference between the results in the tables presented in Figure 6, for the purpose of assessing to what degree the performance of each model on each test dataset is affected by training it on all datasets, as opposed to only on ImageNet.

These results do not show a clear generalization advantage in training from a wider collection of image datasets. While some of the datasets that were added to the meta-training phase did see an improvement across all models, in particular for Omniglot and Quick Draw, this did not translate to an improvement on the datasets which were only used for evaluation: performance on Traffic Signs and MSCOCO decreased slightly.

Table 1: The improvement on META-DATASET obtained by training on All Datasets instead of ILSVRC only.

Test Source	Method: Accuracy ± confidence				
	k-NN	Finetune	MatchingNet	ProtoNet	MAML
ILSVRC	-8.82±1.26	-12.5±1.39	-5.01±1.46	-4.86±1.55	-7.54±1.51
Omniglot	32.61±1.04	26.01±1.39	28.36±1.1	25.14±1.23	24.05±1.77
Aircraft	18.13±1.34	17.04±1.48	28.8±1.24	29.4±1.28	26.1±1.47
Birds	-3.64±1.49	-7.26±1.65	5.02±1.57	3.96±1.67	3.81±1.81
Textures	-6.39±1.1	-9.69±1.2	-1.09±1.17	-1.64±1.1	-2.01±1.23
Quick Draw	23.45±1.61	15.62±1.9	18.92±1.62	16.97±1.81	13.77±2.07
Fungi	0.9±1.36	0.7±1.32	3.89±1.42	4.78±1.7	0.77±1.24
VGG Flower	-3.67±1.34	-9.6±1.63	1.67±1.23	1.17±1.25	0.93±1.58
Traffic Signs	-3.13±2.07	-9.34±1.71	-2.16±1.87	-4.09±1.71	-0.87±1.73
MSCOCO	-6.14±1.41	-7.24±1.51	-5.45±1.47	-5.63±1.67	-4.04±1.56

Figure 6: Accuracy on the test datasets, for each model. The difference between the plain-colored and hacked bars show the effect of training the model on ILSVRC-2012 only, vs. all the datasets.



References

[1] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[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] 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.

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

[5] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

[6] 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.

[7] David Ha and Douglas Eck. A neural representation of sketch drawings. *arXiv preprint arXiv:1704.03477*, 2017.

[8] 2018 FGVCx fungi classification challenge. https://github.com/visipedia/fgvcx_fungi_comp.

[9] M-E. Nilsback and A. Zisserman. A visual vocabulary for flower classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 1447–1454, 2006.

[10] Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. The German Traffic Sign Recognition Benchmark: A multi-class classification competition. In *IEEE International Joint Conference on Neural Networks*, pages 1453–1460, 2011.

[11] 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.