Studying and Improving Extrapolation and Generalization of Non-Parametric and Optimization-Based Meta-Learners

CS330 Project

Axel Gross-Klussmann
SCPD / Department of Computer Science
Stanford University
agk1982@stanford.edu, SUNet agk1982
Code: see appendix

Abstract

In case of few-shot learners, the ability of the algorithms to generalize beyond the training tasks and datasets it sees is of utmost importance. However, little is known about individual extrapolation properties of non-parametric few-shot learning algorithms on unknown out-of-distribution tasks. Next to examining the generalization and extrapolation performance of the raw and unmodified non-parametric few-shot learners we explore and propose bespoke regularization methods that can help reduce initial training task overfitting. In this context we examine the corresponding out-of-distribution performances of regularized learners.

Finn and Levine [2018] compare the gradient based-optimization learners (MAML) to black-box (recurrent) meta-learners on out-of-distribution tasks. Yet it is still unclear how generalization and extrapolation properties of non-parametric few-shot learners compare to MAML and recurrent learners. Given the competitive performance of proto-nets on, e.g., the Omniglot dataset (see Snell et al. [2017] and Vinyals et al. [2017]), it is important to extend the analysis to non-parametric-type models.

Further, recent research shows that 'in-domain' task distributions are typically well captured by state of the art few-shot learning approaches such that the models are robust towards mild in-domain distributional shifts (Finn and Levine [2018]). To assess the cross-domain robustness of meta-learners, Triantafillou et al. [2020] introduce a meta-dataset comprised of several sub-datasets suitable for few-shot learning. The authors show that the performance of meta-learners trained on a subset of datasets and tested on different datasets deteriorates considerably. Hence, few-shot learning models do not generalize well to out-of-domain data and in this respect fall short of the ambition to match humanoid performance. Various approaches were proposed to mitigate the initial overfitting in meta-learning. Recent examples of regularization procedures for meta-learning are given in Zintgraf et al. [2019], Lee et al. [2020], Zhang et al. [2018], Guiroy et al. [2019] and Jamal et al. [2018]. In contrast to the traditional approaches, Finn and Levine [2018] address the problem of meta-overfitting in the context of non-mutually exclusive tasks.

The primary aim of our study is to empirically assess the out-of-distribution (OOD) generalization properties of non-parametric and optimization-based meta-learning. We assess the generalization on both perturbed test data from the domain under consideration (in-domain generalization) and also on tasks drawn from other unrelated datasets (out-of-domain) taken from a meta dataset. The latter test can certainly be seen as the tougher stress test for few-shot learners. Second, our approach builds on the premise that few meta-learning studies are explicitly concerned with regularization for generalizing across domains. To better control the trade-off between in sample / in-domain fit and the ability of a model to generalize we investigate the use of regularization techniques for this purpose.

Our main findings show that non-parametric meta-learners as represented by ProtoNets generalize better than optimization-based meta-learners such as MAML and ANIL. This finding holds for both distributional shifts of in-domain data and cross-domain data taken from a meta-dataset, i.e. a dataset of datasets. In addition, our experiments suggest that regularizing the image embedding layers is beneficial to both MAML/ANIL and ProtoNets. In specific, we find that the meta-regularization proposed in Yin et al. [2020] can slightly improve the accuracies achieved with ANIL while facilitating trainability. Further, a combination of the meta-regularization and dropout seems well suited for regularizing ProtoNets.

Future work can directly adapt to our results. A promising path is to assess the generalization properties of the ProtoMAML model introduced by Triantafillou et al. [2020] which was shown to be successful in cross-domain exercises. All regularization techniques applied in our study directly carry over to the ProtoMAML. Apart from regularizing the image embedding layers, Zhou et al. [2020] is a recent approach in meta-learning aiming to learn equivariances from data. Equivariances are immune to certain shifts of the data input and hence a model based on encoded equivariances should generalize well.

1 Introduction

Although huge leaps were made for neural architectures to better mimic the human mind, the ability of human beings to quickly adapt to new tasks is unparalleled. In this vein, few-shot learning aims to improve the ability of (machine) learners to learn new concepts quickly based on new examples. Much research is devoted to K-shot N-way classification exercises, where a model must learn to classify N classes with a small number of K examples or features per class. Ultimately, the few-shot learners proposed in these studies train global parameters of neural models in a meta training loop which can be refined quickly when new K-shot N-way classification tasks are presented to the model. The accuracy achieved for supervised tasks based on just one example has been impressive in recent works like Snell et al. [2017] as well as Finn et al. [2017]. Models like prototypical networks (ProtoNet henceforth) or MAML generalize well to new tasks from a task distribution sampled from datasets like Omniglot.

Recent research shows that 'in-domain' task distributions are typically well captured by state of the art few-shot learning approaches such that the models are robust towards mild in-domain distributional shifts (Finn and Levine [2018]). To assess the cross-domain robustness of meta-learners, Triantafillou et al. [2020] introduce a meta-dataset comprised of several sub-datasets suitable for few-shot learning. The authors show that the performance of meta-learners trained on a subset of datasets and tested on different datasets deteriorates considerably. Hence, few-shot learning models do not generalize well to out-of-domain data and in this respect fall short of the ambition to match humanoid performance.

The primary aim of our study is to empirically assess the out-of-distribution (OOD) generalization properties of two major meta-learning approaches, non-parametric and optimization-based meta-learning. We assess the generalization on both perturbed test data from the domain under consideration (in-domain generalization) and also on tasks drawn from other unrelated datasets (out-of-domain) taken from a meta dataset. The latter test can certainly be seen as the tougher stress test for few-shot learners. Second, our approach builds on the premise that few meta-learning studies are explicitly concerned with regularization for generalizing across domains. To better control the trade-off between in sample / in-domain fit and the ability of a model to generalize we investigate the use of regularization techniques for this purpose. In this respect we consider traditional regularization means like dropout layers, weight decay and data augmentation as well as the recently proposed meta-regularization (Yin et al. [2020]).

Our contributions are as follows.

1. We conduct an empirical comparison of the in-domain OOD and cross-domain OOD performance of optimization based meta-learners and non-parametric meta-learners.
2. We give an empirical assessment of (meta-) regularization as it pertains to the generalization of few-shot learning models.
3. Motivated by theoretical results in Lee et al. [2020] showing that reducing the inner-loop expressivity for MAML-type learners improves generalization properties, we propose to use a meta-regularized ANIL (almost no inner loop, see Raghu et al. [2019]) model. The meta-regularized ANIL exhibits attractive properties such as faster training in addition to slightly better generalization when compared to the standard MAML.

2 Related Work

Several works have addressed the problem of cross-domain few-shot learning. Triantafillou et al. [2020] put forward a meta dataset and uncover shortcomings of state-of-art meta-learning approaches trained on heterogenous data. In a similar vein, Guo et al. [2019] show that earlier meta-learning models outperform more recent state-of-art models in cross-domain exercises. Interestingly, some of the modern approaches even underperform networks with random weights. Apart from the tougher cross-domain test, Finn and Levine [2018] compare the performance of MAML to black-box meta-learners on out-of-distribution tasks generated based on the Omniglot dataset. The results achieved in this in-domain analysis are more favorable to the more recent meta-learners like MAML.

In an attempt to improve the generalization of few-shot learners, Lee et al. [2019] apply traditional regularization techniques and manage to improve the performances of the meta-learners under consideration. Recognizing that meta-learners are prone to overfit the embedding of, e.g., images in

image classification task, Lee et al. [2020] introduce an alternative encoding of the embeddings which regularizes the meta-learning models. While Yin et al. [2020] primarily address the memorization problem in non-mutually exclusive few-shot learning tasks, the authors show that meta-learners can benefit from their meta-regularization also in mutually exclusive task environments. Yin et al. [2020] pick up ideas from Tishby et al. [2000] as well as Alemi et al. [2018] on the information bottleneck problem. In terms of regularizing MAML, Guiroy et al. [2019] show that reducing the expressivity of the inner loop updates acts as regularization which is beneficial for the generalization properties of MAML. In this sense, Raghu et al. [2019] document good results for a MAML-type model which applies the inner loop update only to the output layers of the underlying architecture. The latter model is called ANIL (almost no inner loop).

3 Problem setup

3.1 Data

To assess the generalization and extrapolation properties of optimization-based and non-parametric few-shot learners we utilize a meta-dataset, closely inspired by the meta-dataset in Triantafillou et al. [2020]. Our dataset is comprised of the german traffic signs data put forward by Houben et al. [2013], the omniglot data (Lake et al. [2011]) as well as the miniImagenet data used by, e.g., Ravi and Larochelle [2017]. To increase classes (by a factor 1.5) I augment the german traffic data by the chinese traffic sign database, `http://www.nlpr.ia.ac.cn/pal/trafficdata/recognition.html`. Likewise, I add 120 image classes from the Imagenet database, namely the task 3 data for the Large Scale Visual Recognition Challenge 2012 (ILSVRC2012). Figure 1 illustrates examples from the three datasets.

Figure 1: One example each from traffic signs, miniImagenet and omniglot.

All of the datasets contain distinct classes of images, traffic signs or characters, respectively, which will be relied upon for few-shot classification exercises in our study. As the datasets differ in size, we always down-sample the larger datasets when conducting cross-domain analyses. Further, to keep the computational burden moderate, we reshape each image to 28 x 28 pixels on a grayscale.

3.2 Near and far out-of-distribution data

Works like Finn and Levine [2018] mimic distributional shifts by shearing and scaling images from Omniglot. The advantage of this approach is that varying degrees of the out-of-distribution character can be simulated. Figure 2 illustrates the approach. Our study follows the same path as we present results for in-domain distributional shifts based on Omniglot.

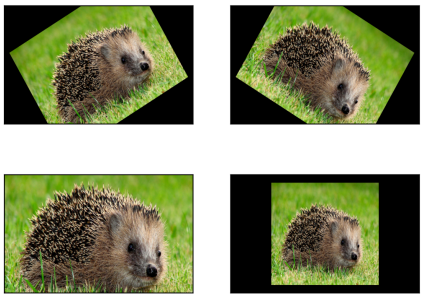

Figure 2: Shearing (top) and scaling (bottom)

Our cross-domain analysis is based on the meta-dataset described in subsection 3.1. In terms of distributional closeness, we conjecture that the traffic signs are somewhat close to the character data from Omniglot. The data from miniImagenet, however, supposedly represents a large distributional shift from both Omniglot and the traffic signs data.

3.3 The meta-learning problem

The meta-learning problem is based on the premise that we are given task data $\mathcal{T}_1 = (\mathbf{x}, \mathbf{y})_1 ..., \mathcal{T}_N = (\mathbf{x}, \mathbf{y})_N$, where the bold-faced symbols denote vectors. We typically sample disjoint tasks from the overall task set $\mathcal{T}$ and construct datasets of tasks, the meta training data D_i^{train} and meta test data D_i^{test} per

sample query i. In specific, we have

$$\begin{aligned}
&\text{Inputs:} \quad \mathcal{D}^{train} = \{(\mathbf{x}, \mathbf{y})_{1:N}\}, \quad \mathbf{x}_{test} \\
&\text{Outputs:} \quad \mathbf{y}_{test} \\
&\text{Meta-learning problem: } \mathbf{y}_{test} = h(\mathcal{D}^{train}, \mathbf{x}_{test}; \theta),
\end{aligned}$$

where y will contain class labels in our case. The examples per class can be quite low. The common terminology has $K =$ number of examples per class, and N way $=$ number of classes per task such that the typical 1-shot 5-way classification means we are given only one example per class to classify into 5 classes.

The main problem in meta-learning consists in finding a form of $h(\mathcal{D}^{train}, \mathbf{x}_{test}; \theta)$ and subsequently in optimizing the parameters governing the form. Several algorithms have been proposed for this purpose.

3.3.1 Optimization-based meta-learning: standard MAML and ANIL

The MAML algorithm can in principle be applied to numerous neural architectures. It consists of a fine-tuning step based on D^{train} which aims at learning an initialization for the parameters of the underlying architecture that can quickly adapt to new tasks during meta-test time.

$$\begin{aligned}
&\text{Fine-tuning at test time} \qquad \phi \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}(\theta, \mathcal{D}^{train}) \\
&\text{Meta-learning} \qquad \min_\theta \sum_{\mathcal{T}_i} \mathcal{L}(\theta - \alpha \nabla_\theta \mathcal{L}(\theta, \mathcal{D}_i^{train}), \mathcal{D}_i^{test})
\end{aligned}$$

For each task i, MAML first computes inner gradient updates on D_i^{train} and evaluates on D_i^{test}. This is the inner loop of MAML. The outer loop collects the post-update losses and updates the model parameters.

A typical MAML implementation for image classification is based on a (4-layer) CNN architecture parameterized with θ^e yielding image embeddings $f_{\theta^e}(\mathbf{x})$. The last layer on top of the embedding layer f_{θ^e} can be a simple linear layer which is subsequently fed into a softmax and cross entropy loss,

$$p(y|\mathbf{x}, \theta) = \text{softmax}(\mathbf{b} + \mathbf{W} f_{\theta^e}(\mathbf{x})). \tag{1}$$

In its original form, the inner and outer loop gradient updates operate on the full parameter set $\{\theta^e, \mathbf{W}, \mathbf{b}\}$ which can be computationally intensive due to the higher order differentiation needed in the inner loop.

However, Raghu et al. [2019] show that inner gradient updates only for the parameters of the last layer(s) such as given in equation (1) give results similar to vanilla MAML models which fine-tune the full parameter set. The model variant is called ANIL (almost no inner loop). Given the ease of training we consider it here, too. Further, the ANIL has a theoretical underpinning as it essentially regularizes MAML. It should therefore be helpful in our attempt to improve generalization properties of basic meta-learners.

3.3.2 Non-parametric meta-learning: Prototypical Networks

As the prime example for non-parametric meta-learning we consider prototypical networks, ProtoNets henceforth. ProtoNets first compute a representative, i.e. a prototype, per class before non-parametric methods like nearest neighbors are applied. As for many MAML architectures for image processing, typical ProtoNets are based on an image embedding $f_\theta(\mathbf{x})$. In our case the embedding is given by a 4 layer CNN. Equation 2 give the core steps of the ProtoNet.

Given support examples for each class, per-class prototypes $\mathbf{c}_k$ are computed as averages of the embedded examples,

$$\begin{aligned}
\mathbf{c}_k &= \frac{1}{|\mathcal{D}_i^{train}|} \sum_{(\mathbf{x},\mathbf{y}) \in \mathcal{D}_i^{train}} f_\theta(\mathbf{x}), \\
p_\theta(y = k|\mathbf{x}^q) &= \frac{\exp(-d(f_\theta(\mathbf{x}^q), \mathbf{c}_k))}{\sum_{k'} \exp(-d(f_\theta(\mathbf{x}^q), \mathbf{c}_{k'}))}.
\end{aligned} \tag{2}$$

The probability $p_\theta(y = k|\mathbf{x}^q)$ gives the probability of a query example $\mathbf{x}^q$ belonging to class k. The distance function d is the Euclidean distance.

3.3.3 (Meta-) Regularizing meta-learners

Several studies show that regularizing meta-learners is crucial for them to generalize well beyond the in-domain task distributions. As for MAML, the Theorem 1 in Lee et al. [2020] shows that the vanilla MAML benefits from reducing the expressivity of the inner gradient loop. The ANIL version of the MAML (see subsection 3.3.1) can be seen as a regularized MAML: Instead of altering all parameters during meta-test time the ANIL concentrates on inner loop initialization for the last layers only. The ANIL version of MAML can hence be considered a first means of regularization.

A second successful strand of regularization for meta-learners focuses on the embedding layers (see Lee et al. [2019]). When the application is few-shot image classification, both the MAML/ANIL and the ProtoNet operate on CNN embedding layers. In specific, models in our study are based on 4 CNN layers with a filter size of 32, each with a batch normalization layer and ReLu activation. We consider three different approaches to regularizing the CNN embeddings. First, we include dropout layers in the CNN architecture. Second, we use weight decay. We concentrate our efforts on L1-regularization of weights. L1-losses of the embedding layers are simply added to the standard loss function in order to regularize the expressiveness of the embeddings. Third, while originally proposed to address the memorization problem in non-mutually-exclusive few-shot learning, Yin et al. [2020] show that meta-regularizing MAML yields improvements in the standard mutually exclusive setting, too. In contrast to dropout and L1-regularization, the meta-regularization is based on a variational inference building block. The meta-regularization aims to control the information flow between x and y better. Due to its Bayesian underpinnings, it does not impose hard restrictions on parameters and hence leaves the expressiveness of the network fully intact.

The meta-regularization can conveniently be implemented in the ANIL setting. Let $\tilde{\theta}$ denote the adaptation parameters of the last layers. We leave these parameters unregularized. The CNN embedding layers, however, are replaced by their reparameterization layer counterparts. The Kullback-Leibler distance of the embedding distributions against their normal priors are added as losses to the loss function. As such, the weight distributions are always pulled a bit towards their prior. The full algorithm of Yin et al. [2020] works as follows.

 # Algorithm: Meta-Regularized ANIL
Input: Embedding weights distribution $q(\theta;\tau) = N(\theta;\tau)$ with Gaussian parameters $\tau = (\theta_\mu, \theta_\sigma)$, prior distribution $r(\theta)$ and Lagrangian multiplier β. Stepsizes $\alpha,\ \alpha'$
Output: Network parameters $\tau, \tilde{\theta}$

Initialize $\tau, \tilde{\theta},$ randomly
while not converged:
  Sample a mini-batch of $\{\mathcal{T}_i\}$ from $\mathcal{T}$
  Sample $\theta \sim q(\theta;\tau)$ with reparameterization
  for $\mathcal{T}_i \in \{\mathcal{T}_i\}$:
    Sample $\mathcal{D}_i^{train} = (\mathbf{x}_i, \mathbf{y}_i),\ \mathcal{D}_i^{test} = (\mathbf{x}_i^{ts}, \mathbf{y}_i^{ts})$ from $\mathcal{T}_i$
    Encode / embed the observations $\mathbf{z}_i = f_\theta(\mathbf{x}_i^{ts}),\ \mathbf{z}_i^{ts} = f_\theta(\mathbf{x}_i^{ts})$
    Compute task specific parameter $\phi_i = \tilde{\theta} + \alpha' \nabla_{\tilde{\theta}} \log\ q(\mathbf{y}_i|\mathbf{z}_i, \tilde{\theta})$
  Update $\tilde{\theta} \leftarrow \tilde{\theta} + \alpha \nabla_{\tilde{\theta}} \sum_{\mathcal{T}_i} \log\ q(\mathbf{y}_i^{ts}|\mathbf{z}_i^{ts}, \phi)$
  Update $\tau \leftarrow \tau + \alpha \nabla_\tau \left[\sum_{\mathcal{T}_i} \log\ q(\mathbf{y}_i^{ts}|\mathbf{z}_i^{ts}, \phi) - \beta D_{KL}(q(\theta;\tau)||r(\theta))\right]$

We note that the above algorithm idea straightforwardly carries over to the case of ProtoNet, too: We replace the CNN layers of ProtoNet with the reparameterization layers and add the Kullback-Leibler loss before the gradient update.

4 Experiments and results

4.1 Generalization of vanilla few-shot learners on Omniglot

We follow Finn and Levine [2018] and artificially construct OOD data from Omniglot by shearing and scaling the character images. Figure 3 shows meta test accuracies for a 1-shot, 5-way exercise for four meta-learning models trained on the unsheared, unscaled vanilla Omniglot dataset. After

training we compute the meta-test accuracies on the artificially constructed OOD Omniglot data. The performance for the unscaled and unsheared Omniglot is shown for $scale = 1$ and $shear = 0$.

At this initial stage we consider four models. First, the ProtoNet (see subsection 3.3.2). Second, the vanilla MAML (see subsection 3.3.1). Third, the ANIL variant of MAML (see subsection 3.3.1). Finally, we drop the embedding layers of the ProtoNet and operate solely with the final nearest neighbor layer of the ProtoNet. Ultimately, this last model, called 'KNN baseline' in the plots, serves as baseline.

We train the MAML-type models for 6K steps which seems sufficient for convergence on Omniglot as can be seen from the validation set accuracy plots in the Appendix. To improve the convergence properties of the MAML-type models we always implement the inner learning rate, i.e. the learning rate of the inner loop, as a learnable parameter which is optimized during training. Further, Finn and Levine [2018] show that the inner loop can handle up to 100 inner gradient steps on Omniglot without a drop in test accuracy. In light of this we train all MAML models and variants thereof with 2 gradient steps to balance test accuracy improvements and computational burden. Our ProtoNet models are trained on only 2K steps due to their fast convergence. All models are written in Tensorflow and trained on the Google cloud platform.[1] An exception are the ProtoNets which could be trained locally.

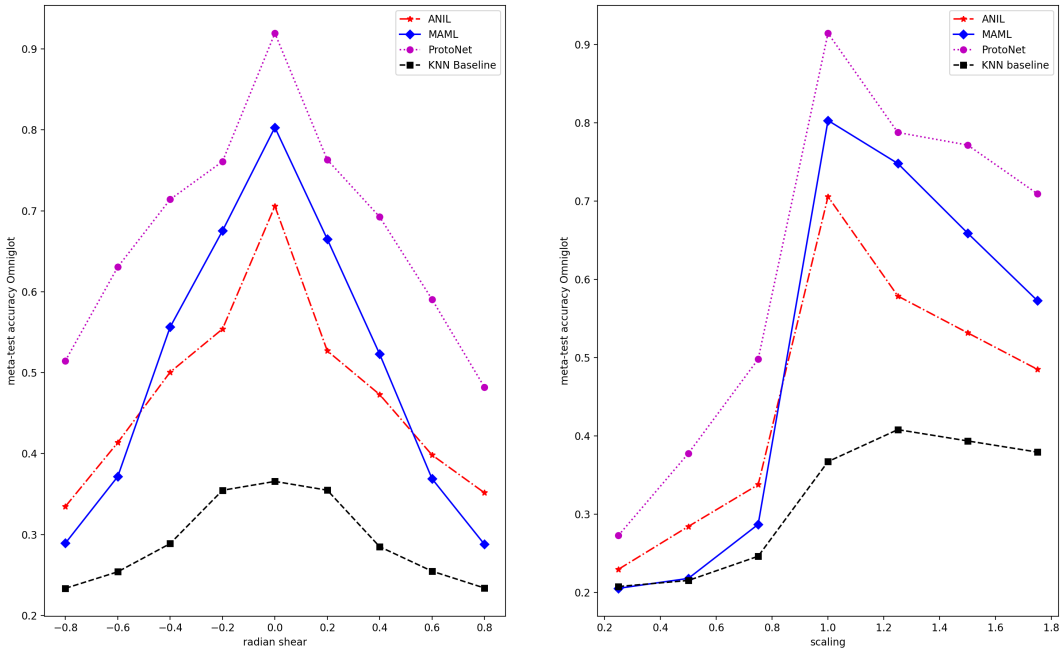

Figure 3: Meta accuracy of meta-learners for different degrees of scaling and shearing of Omniglot images. 5 way, 1 shot setting.

Finn and Levine [2018] illustrate that MAML performs robustly for out-of-distribution tasks based on sheared and scaled Omniglot data. In specific, MAML outperforms the blackbox meta-learners such as given by Mishra et al. [2018] and Santoro et al. [2016]. MAML is shown to attain above 95% test accuracy for all kinds of digit shears of Omniglot images. It further attains above 75% accuracy for scaled Omniglot images.

In our case, however, MAML is outperformed by ProtoNet, taken as representative of non-parametric learners. Further, the test accuracy attained by all models drops quickly as the degree of shearing and scaling is increased. The MAML works better than the ANIL for 'near' distributional shifts. As we shear the images more, the vanilla MAML accuracy falls behind that of ANIL. Given the discussion in subsection 3.3.3, the ANIL is a regularized version of MAML. We therefore take the

[1]The code can be found here: `https://drive.google.com/file/d/1MvHm2sxUmVPH9ez7ybLkPE6SLlwLPBn9/view?usp=sharing`

improved accuracy of ANIL at the boundaries of the shearing experiment as indicative of the benefits of regularization for generalization. The robust performance of ProtoNet is due to the ease of training on just 2K steps. Moreover, the nearest-neighbor-type head of the ProtoNet model attains above random accuracies even without the CNN backbone as exemplified by the 'KNN baseline'.

We attribute the performance differences of our MAML to the one in Finn and Levine [2018] to three key changes we make to their setup. First and most importantly, we operate on a slightly reduced resolution of Omniglot images (28 x 28) which amplifies the shearing and scaling effect on the models. Second, the constrained number of training steps (6K). Third, the lower number of inner loop gradient steps at two.

4.2 Generalization of (meta-) regularized few-shot learners on Omniglot

We first calibrate the Lagrangian multiplier β (see subsection 3.3.3) of the meta-regularization for both the ProtoNet and ANIL variants. Following Yin et al. [2020] our choice is based on the end-of-training validation accuracies. Figure 5 in appendix 6.2 gives details. Likewise we calibrate the L1-penalty as well as the dropout probability. We report findings for $\beta = 5e - 5$ and a dropout probability of 10%.

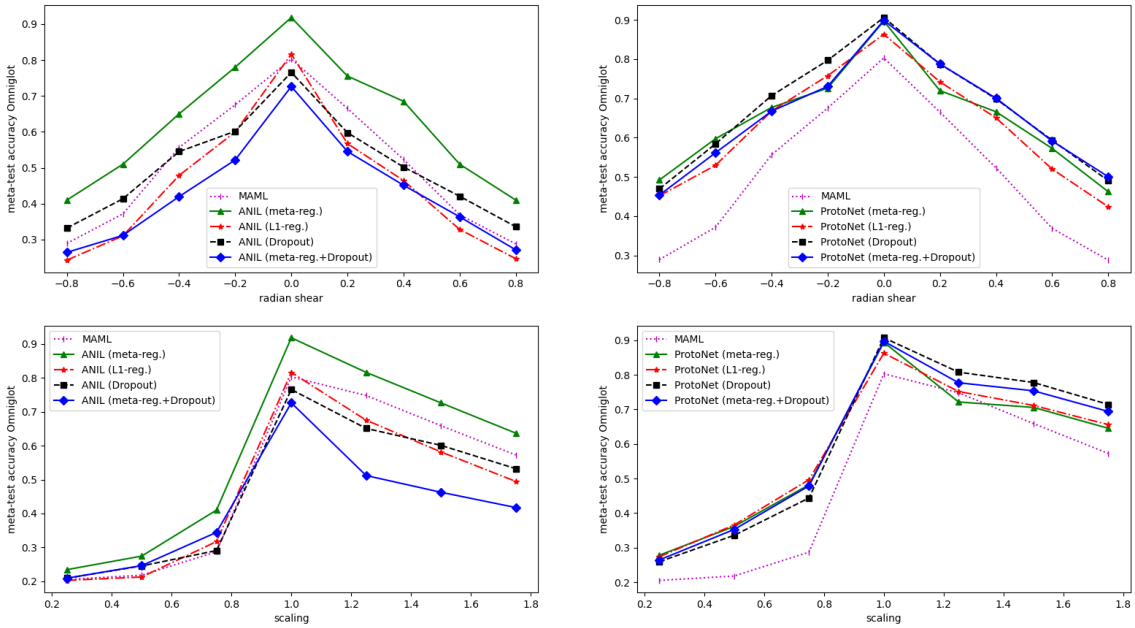

Figure 4: Meta accuracy of regularized meta-learners for different degrees of scaling and shearing of Omniglot images. 5 way, 1 shot setting. Left panels: optimization-based. Right panels: ProtoNets.

In the same spirit as in section 4.1, Figure 4 shows meta test accuracies for a 1-shot, 5-way exercise for regularized meta-learning models trained on the unsheared, unscaled vanilla Omniglot dataset. We always include the vanilla MAML for comparison purposes. In the analysis for MAML-type models, the meta-regularized ANIL outperforms all MAML variants for all degrees of shearing and scaling. However, a caveat is that its accuracy decreases at a rate similar to vanilla MAML when distorting the distributions more and more. Altogether we attribute the high test accuracies observed for the meta-regularized ANIL rather to the ease of training than to the regularization effect. We observe that the rate of accuracy decline as the distribution gets distorted more is best offset by the ANIL with an added dropout layer. A combination of meta-regularization and dropout deteriorates the performance. The results are somewhat at odds with the large percentage accuracy gains reported by Lee et al. [2019] for various regularization approaches and combinations thereof. However, while the setting itself is different, Lee et al. [2019] are moreover not concerned with OOD scenarios.

Inspecting the pre-inner loop accuracies (see Table 1) of the optimization-based models we find that the various (meta-) regularization techniques succeed in reducing the expressiveness of the network

prior to fine-tuning steps. However, the accuracy margin is very slim, indicating that only little regularization is applied based on the penalty parameters.

	ANIL (meta-reg.+Dropout)	ANIL (+Dropout)	ANIL (meta-reg.)	ANIL	MAML
accuracy	0.194	0.199	0.198	0.208	0.202

Table 1: Meta-validation pre-inner-loop training accuracy. End of training, smoothed value.

As for the ProtoNets, we find that both dropout and meta-regularization can best support the accuracy on out-of-distribution tasks. However, similar to the findings for optimization-based models, the actual accuracy increases over the vanilla ProtoNet are small. Interestingly, the meta-regularized ANIL matches the performances of the ProtoNet variants while the vanilla MAML remains about 10% short of the ProtoNet across all OOD scenarios. Overall we conjecture that the applied regularization is too ineffective for both the optimization-based models and the non-parametric meta-learners.

4.3 A cross-domain exercise based on Meta-Dataset

A tougher out-of-distribution test is given by cross-domain meta-datasets. In light of the previous in-domain (but still out-of-distribution) analysis we take the most promising models to our mini-meta-dataset (see subsection 3.1). Table 2 show accuracies for models trained on a part of the meta-dataset and tested on another part. Due to the composition of our mini-meta-dataset (traffic sign data, Omniglot and miniImagenet), we can hence test true out-of-domain scenarios. Note that results presented here deviate from, e.g., those in Triantafillou et al. [2020], especially for miniImagenet. The reason is that we constrain all images to be 28x28 in order to cross-apply models.

train/validation (row): / test (col):	omniglot	traffic	mImagenet	∅
MAML:				
omniglot	0.80	0.21	0.21	0.41
meta-regularized ANIL:				
omniglot	0.92	0.23	0.21	0.45
traffic	0.22	0.52	0.31	0.35
mImagenet	0.21	0.35	0.26	0.27
ProtoNet:				
omniglot	0.96	0.51	0.29	0.59
traffic	0.65	0.81	0.27	0.58
mImagenet	0.71	0.56	0.40	0.56
meta-regularized ProtoNet+Dropout:				
omniglot	0.96	0.50	0.29	0.58
traffic	0.61	0.83	0.28	0.57
mImagenet	0.72	0.59	0.42	0.58
traffic ∪ mImagenet	0.74	0.79	0.39	0.64
omniglot ∪ mImagenet	0.93	0.57	0.36	0.62
omniglot ∪ traffic	0.92	0.70	0.32	0.65
omniglot ∪ mImagenet ∪ traffic	0.92	0.70	0.37	0.66

Table 2: Meta-test accuracies for models trained on the dataset given in the row names and evaluated on the test set given by the column names. 1 shot, 5-way. Rightmost column gives the average.

The following findings emerge from our analysis. First, as was already observed in the previous sections, the optimization-based models have more trouble generalizing to distorted distributions than the non-parametric models like ProtoNets. In the tougher cross-domain exercise, both the vanilla MAML and the meta-regularized ANIL completely fail to generalize to other domains, rarely surpassing 20% accuracy on out-of-domain test data. We conjecture that this finding is due to the instability of the fine-tuning via gradients. Gradients are notoriously noisy.

In contrast, the ProtoNet with and without regularization generalizes better to out-of-domain data according to our results. As before, there are virtually no differences between the regularized and unregularized ProtoNet.

Data augmentation is a form of regularization and can be directly applied based on the meta-dataset. When sampling from different datasets, we have to make sure that the probabilities of samples being picked are equal. We achieve this by down- or up-sampling data. Table 2 shows that, e.g., training regularized ProtoNet on a combined dataset of traffic signs and miniImagenet still gives a competitive 74% meta test accuracy on Omniglot. Likewise, training on Omniglot and miniImagenet data and testing on traffic sign data yields 57% accuracy which surpasses the accuracy of a meta-regularized ANIL model trained in-domain on traffic sign data.

Not surprisingly, the highest average test accuracy across all sub-datasets are attained when training on the combination of the three corresponding training datasets. However, this amounts to an in-domain analysis then.

5 Conclusions

Our study conducts few-shot image classification experiments analyzing the generalization of optimization based and non-parametric meta-learners. Upon observing that the meta-learners under consideration have trouble maintaining high classification accuracies when faced with out-of-distribution data, we explore several techniques of regularizing the models.

Our main findings show that non-parametric meta-learners as represented by ProtoNets generalize better than optimization-based meta-learners such as MAML and ANIL. This finding holds for both distributional shifts of in-domain data and cross-domain data taken from a meta-dataset, i.e. a dataset of datasets. In addition, our experiments suggest that regularizing the image embedding layers is beneficial to both MAML/ANIL and ProtoNets. In specific, we find that the meta-regularization proposed in Yin et al. [2020] can slightly improve the accuracies achieved with ANIL while facilitating trainability. Further, a combination of the meta-regularization and dropout seems well suited as a building block for regularizing ProtoNets. Notably, the accuracy gains achievable through regularization are small in our setting, which renders the question of how to design a model that generalizes well out-of-distribution and out-of-domain still open.

Future work can directly adapt to our results. A promising path is to assess the generalization properties of the ProtoMAML model introduced by Triantafillou et al. [2020] which was shown to be successful in cross-domain exercises. All regularization techniques applied in our study directly carry over to the ProtoMAML. Apart from regularizing the image encodings, Zhou et al. [2020] is a recent approach in meta-learning aiming to learn equivariances from data. Equivariances are immune to certain shifts of the data input and hence a model based on encoded equivariances should generalize well.

References

Chelsea Finn and Sergey Levine. Meta-learning and universality: Deep representations and gradient descent can approximate any learning algorithm, 2018.

Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning, 2017.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning, 2017.

Eleni Triantafillou, Tyler Zhu, Vincent Dumoulin, Pascal Lamblin, Utku Evci, Kelvin Xu, Ross Goroshin, Carles Gelada, Kevin Swersky, Pierre-Antoine Manzagol, and Hugo Larochelle. Meta-dataset: A dataset of datasets for learning to learn from few examples. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rkgAGAVKPr.

Luisa Zintgraf, Kyriacos Shiarli, Vitaly Kurin, Katja Hofmann, and Shimon Whiteson. Fast context adaptation via meta-learning. volume 97 of *Proceedings of Machine Learning Research*, pages 7693–7702, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL http://proceedings.mlr.press/v97/zintgraf19a.html.

Yoonho Lee, Wonjae Kim, Wonpyo Park, and Seungjin Choi. Discrete infomax codes for supervised representation learning, 2020.

Ruixiang Zhang, Tong Che, Zoubin Ghahramani, Yoshua Bengio, and Yangqiu Song. Metagan: An adversarial approach to few-shot learning. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 2365–2374. Curran Associates, Inc., 2018. URL http://papers.nips.cc/paper/7504-metagan-an-adversarial-approach-to-few-shot-learning.pdf.

Simon Guiroy, Vikas Verma, and Christopher Pal. Towards understanding generalization in gradient-based meta-learning, 2019.

Muhammad Abdullah Jamal, Guo-Jun Qi, and Mubarak Shah. Task-agnostic meta-learning for few-shot learning, 2018.

Mingzhang Yin, George Tucker, Mingyuan Zhou, Sergey Levine, and Chelsea Finn. Meta-learning without memorization, 2020.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks, 2017.

Aniruddh Raghu, Maithra Raghu, Samy Bengio, and Oriol Vinyals. Rapid learning or feature reuse? towards understanding the effectiveness of MAML. *CoRR*, abs/1909.09157, 2019. URL http://arxiv.org/abs/1909.09157.

Yunhui Guo, Noel C. F. Codella, Leonid Karlinsky, John R. Smith, Tajana Rosing, and Rogério Schmidt Feris. A new benchmark for evaluation of cross-domain few-shot learning. *CoRR*, abs/1912.07200, 2019. URL http://arxiv.org/abs/1912.07200.

Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, and Stefano Soatto. Meta-learning with differentiable convex optimization, 2019.

Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. In *Proc. of the 37-th Annual Allerton Conference on Communication, Control and Computing*, pages 368–377, 2000. URL https://arxiv.org/abs/physics/0004057.

Alexander A. Alemi, Ian Fischer, and Joshua V. Dillon. Uncertainty in the variational information bottleneck. *CoRR*, abs/1807.00906, 2018. URL http://arxiv.org/abs/1807.00906.

S. Houben, J. Stallkamp, J. Salmen, M. Schlipsing, and C. Igel. Detection of traffic signs in real-world images: The german traffic sign detection benchmark. In *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2013. doi: 10.1109/IJCNN.2013.6706807.

M. Brenden Lake, Ruslan Salakhutdinov, Jason Gross, and B. Joshua Tenenbaum. One shot learning of simple visual concepts. *CogSci*, 2011.

S. Ravi and H. Larochelle. Optimization as a model for few-shot learning. In *ICLR*, 2017.

Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel. A simple neural attentive meta-learner, 2018.

Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. One-shot learning with memory-augmented neural networks, 2016.

Allan Zhou, Tom Knowles, and Chelsea Finn. Meta-learning symmetries by reparameterization, 2020.

6 Appendix

6.1 Zipped code repository

`https://drive.google.com/file/d/1MvHm2sxUmVPH9ez7ybLkPE6SLlwLPBn9/view?usp=sharing`

6.2 Calibrating the meta regularization β

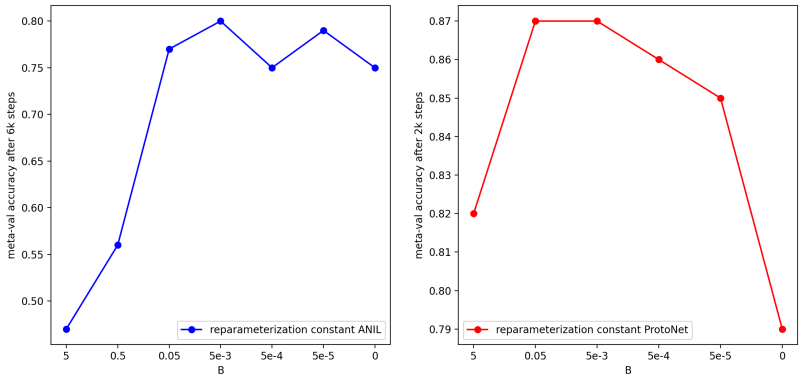

Figure 5: Meta validation accuracies per parameter value.

6.3 Tensorboard plots

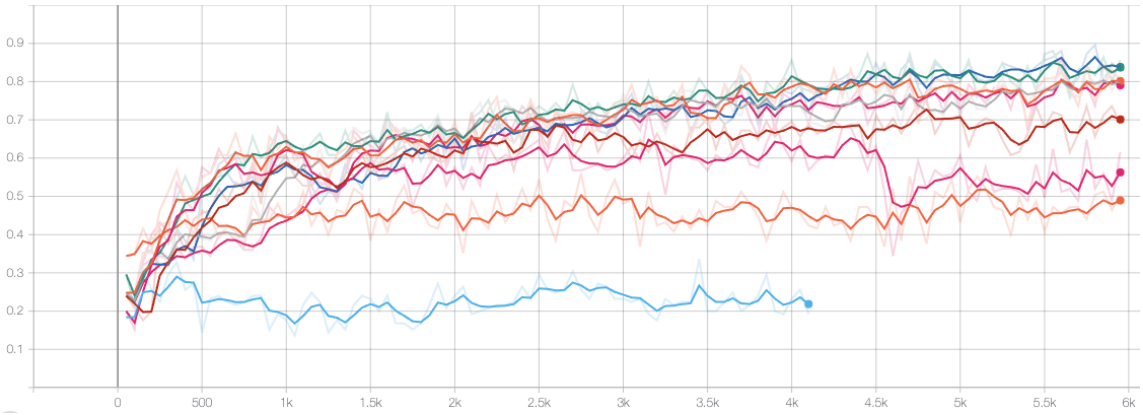

Figure 6: Snapshot maml-type model training

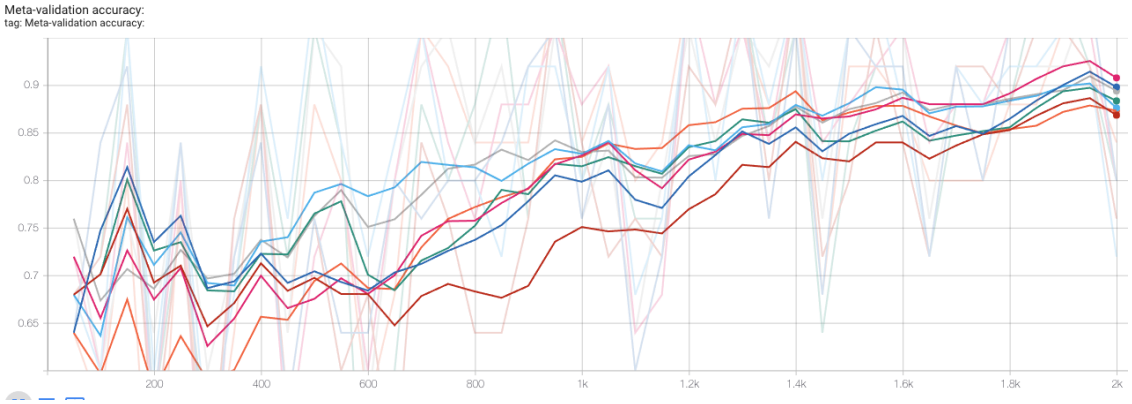

Figure 7: Snapshot ProtoNet training.