Towards an Unsupervised Method for Model Selection in Few-Shot Learning

Simon Guiroy [1 2] **Vikas Verma** [3] **Christopher Pal** [1 4 5]

Abstract

The study of generalization of neural networks in gradient-based meta-learning has recently great research interest. Previous work on the study of the objective landscapes within the scope of few-shot classification empirically demonstrated that generalization to new tasks might be linked to the average inner product between their respective gradients vectors (Guiroy et al., 2019). Following that work, we study the effect that meta-training has on the learned space of representation of the network. Notably, we demonstrate that the global similarity in the space of representation, measured by the average inner product between the embeddings of meta-test examples, also correlates to generalization. Based on these observations, we propose a novel model-selection criterion for gradient-based meta-learning and experimentally validate its effectiveness.

1. Introduction

To address the problem of the few-shot learning, many meta-learning approaches have been proposed recently (Finn et al., 2017; Ravi & Larochelle, 2017; Rothfuss et al., 2018; Oreshkin et al., 2018; Snell et al., 2017) , among others. In this work, additional steps towards understanding the characteristics of learned space of representation and its relation to generalization, in the context of gradient-based few-shot meta-learning. We focus our experimental work here within a setup that follows the recently proposed Model Agnostic Meta-Learning (MAML) (Finn et al., 2017). MAML is a good candidate for studying gradient-based meta-learning because of its independence from the underlying network architecture, and because of its reasonable success as a few-shot image classification algorithm. One practical motivation for understanding the space of representation and its relation to generalization is to mitigate meta-overfitting, i.e. when the average target accuracy to meta-test tasks, after it has peaked, starts to decrease. In this scope, we are notably interested in following the evolution of this space, as meta-training progresses.

Our main insights and contributions can be summarized as follows:

• In an attempt to provide an intuitive explanation for the correlation between generalization to new tasks and the similarity of meta-test gradients in inner product, we suggest that MAML, in the standard case of few-shot image classification, might be learning a representation space based on the inner product. We provide empirical evidence showing the correlation between the average inner product between the representation vectors, produced by the model at meta-train solution, for the meta-test data taken as input, and the ability of the model to generalize to the meta-test tasks.

• From this last observation, we propose an early stopping criterion, which can be used on a single meta-test task basis, with similar performance to early stopping based on meta-validation, without using the meta-valid data, which can be Incorporated into the training data. This criterion is an unsupervised method, as it doesn't reply on information on the class labels and is measured before adaptation to new tasks.

2. Gradient-Based Meta-Learning

We consider the meta-learning scenario where we have a distribution over tasks $p(\mathcal{T})$, and a model f parametrized by θ, that must learn to adapt to tasks $\mathcal{T}_i$ sampled from $p(\mathcal{T})$. The model is trained on a set of training tasks $\{\mathcal{T}_i\}^{train}$ and evaluated on a set of testing tasks $\{\mathcal{T}_i\}^{test}$, all drawn from $p(\mathcal{T})$. In this work we only consider classification tasks, with $\{\mathcal{T}_i\}^{train}$ and $\{\mathcal{T}_i\}^{test}$ using disjoint sets of classes to constitute their tasks. Here we consider the setting of k-shot learning, that is, when f adapts to a task $\mathcal{T}_i^{test}$, it only has access to a set of few support samples $\mathcal{D}_i = \{(\mathbf{x}_i^{(1)}, \mathbf{y}_i^{(1)}), ..., (\mathbf{x}_i^{(k)}, \mathbf{y}_i^{(k)})\}$ drawn from $\mathcal{T}_i^{test}$, where k is the number of examples per class. We then evaluate the model's performance on $\mathcal{T}_i^{test}$ using a new set of target samples $\mathcal{D}_i'$. By gradient-based meta-learning, we imply that f is trained using information about the gradient of a certain loss function $\mathcal{L}(f(\mathcal{D}_i; \theta))$ on the tasks. Throughout this work the loss function is the cross-entropy between the

[1]Mila [2]Université de Montréal [3]Aalto University, Finland [4]'Ecole Polytechnique de Montréal [5]ElementAI, Montréal. Correspondence to: Simon Guiroy <simon.guiroy@umontreal.ca>.

Lifelong Learning Workshop - 37th International Conference on Machine Learning, Copyright 2020 by the author(s).

predicted and true class.

Throughout this work, the main algorithm we focus on is Model Agnostic Meta-Learning (MAML) (Finn et al., 2017). We also study a variant of MAML, namely its first-order version, where the second-order derivatives are omitted.

2.1. Model Agnostic Meta-Learning

MAML learns an initial set of parameters θ such that on average, given a new task $\mathcal{T}_i^{test}$, only a few samples are required for f to learn and generalize well to the new task. During a meta-training iteration s, where the current parametrization of f is θ^s, a batch of n training tasks is sampled from $p(\mathcal{T})$. For each task $\mathcal{T}_i$, a set of support samples $\mathcal{D}_i$ is drawn and f adapts to $\mathcal{T}_i$ by performing T steps of full batch gradient descent on $\mathcal{L}(f(\mathcal{D}_i;\theta))$ w.r.t. θ, obtaining the adapted solution $\tilde{\theta}_i$:

$$\tilde{\theta}_i = \theta^s - \alpha \sum_{t=0}^{T-1} \nabla_\theta \mathcal{L}(f(\mathcal{D}_i;\theta_i^{(t)})) \tag{1}$$

where $\theta_i^{(t)} = \theta_i^{(t-1)} - \alpha \nabla_\theta \mathcal{L}(f(\mathcal{D}_i;\theta_i^{(t-1)}))$ and adaptation trajectories for all $\mathcal{T}_i$ are independent and start from θ^s, i.e. $\theta_i^{(0)} = \theta^s, \forall i$. Then from each $\mathcal{T}_i$, a set of target samples $\mathcal{D}'_i$ is drawn, and the adapted meta-training solution θ^{s+1} is obtained by minimizing the loss on the target samples $\mathcal{D}'_i$, across all task $\mathcal{T}_i$ as follows:

$$\theta^{s+1} = \theta^s - \beta \frac{1}{n} \sum_{i=1}^{n} \nabla_\theta \mathcal{L}(f(\mathcal{D}'_i;\tilde{\theta}_i)) \tag{2}$$

As one can see in Eq.1 and Eq.2, deriving the meta-gradients implies computing second-order derivatives, which can come at a significant computational expense. The authors introduced a first-order approximation of MAML, where these second-order derivatives are ommited, and we refer to that other algorithm as First-Order MAML.

3. Background and Related Works

More recently, some works have started to analyze theoretical aspects of gradient-based meta-learning.

The authors of (Finn et al., 2019) introduced the Online Meta-Learning setting, where in online learning the agent faces a sequence of tasks, and provided a theoretical upper bound for the regret of MAML.

The work of (Guiroy et al., 2019) empirically study the objective landscapes of gradient-based meta-learning, with a focus on few-shot classification. They notably show that average generalization to new tasks appears correlated with the average inner product between their gradient vectors. In other words, as gradients appear more similar in inner product, the model will, on average, better generalize to new tasks, after following a step of gradient descent.

In this work, in an attempt to provide an intuitive interpretation, we extend this analysis by showing empirically that for different settings of gradient-based meta-learning of few-shot classification tasks, the average inner product among representation vectors, for meta-test examples, also correlates with generalization to new tasks.

Prior to our work, the authors of (Raghu et al., 2019), while not explicitly studying generalization, showed that for MAML at meta-test time, the embeddings representing the new task inputs, thus the outputs of the feature network, barely change during finetuning of the model, as opposed to the outputs of the linear classifier, a phenomenon they name "feature reuse".

In (Denevi et al., 2019), the authors study meta-learning through the perspective of biased regularization, where the model adapts to new tasks by starting from a biased parameter vector, which we refer in this work as the meta-training solution. For simple tasks such as linear regression and binary classification, they prove the advantage of starting from the meta-training solution, when learning new tasks via SGD. They use an assumption on the task similarity where the weight vectors parameterizing the tasks are assumed to be close to each other.

Working in the framework for Online Convex Optimization where the model learns from a stream of tasks, (Khodak et al., 2019) make an assumption that the optimal solution for each task lies in a small subset of the parameter space and use this assumption to design an algorithm such that the "Task-averaged-regret (TAR)" scales with the diameter of this small subset of the parameter space, when using Reptile (Nichol et al., 2018), a first-order meta-learning algorithm.

4. Analysis

In the context of gradient-based meta-learning, we define generalization as the model's ability to reach a high accuracy on a testing task $\mathcal{T}_i^{test}$, evaluated with a set of target samples $\mathcal{D}'_i$, for several testing tasks. This accuracy is computed after f, starting from a given meta-training parametrization θ^s, has optimized its parameters to the task $\mathcal{T}_i^{test}$ using only a small set of support samples $\mathcal{D}_i$, resulting in the adapted solution $\tilde{\theta}_i^{test}$. We thus care about the expected accuracy $\mathbb{E}_{\mathcal{T}_i^{test} \sim p(\mathcal{T})}[Acc(f(\mathcal{D}'_i;\tilde{\theta}_i^{test}))]$.

We consider the space of representation that is learned by the model, and follow its evolution as meta-training progresses, after each epoch, which results in a different parametrization θ^s, and compare with $\mathbb{E}[Acc(f(\mathcal{D}'_i;\tilde{\theta}_i^{test}))]$. This approach is motivated by the aim of defining a metric that would reflect meta-overfitting, which could be use to stop the meta-training or alternatively, select the saved model that would achieve the best generalization.

The authors of (Guiroy et al., 2019) notably observe that the average inner product between the gradient vectors $\mathbf{g}_i$ of different meta-test tasks $\mathcal{T}_i$, for a model with a given parametrization θ^s, correlates with the average target accuracy that the model reaches, for those tasks, after following a step of gradient descent. They define this coherence between meta-test gradients as:

$$\mathbb{E}_{\mathcal{T}_i,\mathcal{T}_j\sim p(\mathcal{T})}[\ \mathbf{g}_i{}^T\mathbf{g}_j\ ] \tag{3}$$

The learned representation space In Section 6 we further investigate the link between inner product of gradient vectors and generalization, by analysing the effect that meta-training has on the learned space of representation. More concretely, we characterize the similarity among test embedding in that space, where similarity is measured by the inner product. In general, the network $f(x)$ can be expressed as $f_{lin}(f_{feat}(x))$ where f_{lin} is as linear classifier, with of a weight matrix W and a bias vector b followed by a softmax, and f_{feat} is a feature network, that outputs a representation vector h, such that $h = f_{feat}(x)$.

We define the global similarity in the space of representation, among representation vectors h_i, produced by a model f_{feat} parametrized by θ^s, of test examples $x_i \sim p^{test}(X)$, as:

$$\mathbb{E}_{h_i,h_j\sim p(f_{feat}(x;\theta^s)}[\ \mathbf{h}_i{}^T\mathbf{h}_j\ ] \tag{4}$$

Interestingly, we observe that this global similarity in the space of representation correlates to generalization, while this metric is computed before adaptation to the new tasks and is independent of the class labels, thus being an unsupervised method.

5. Experiments Setup

5.1. Model Architectures

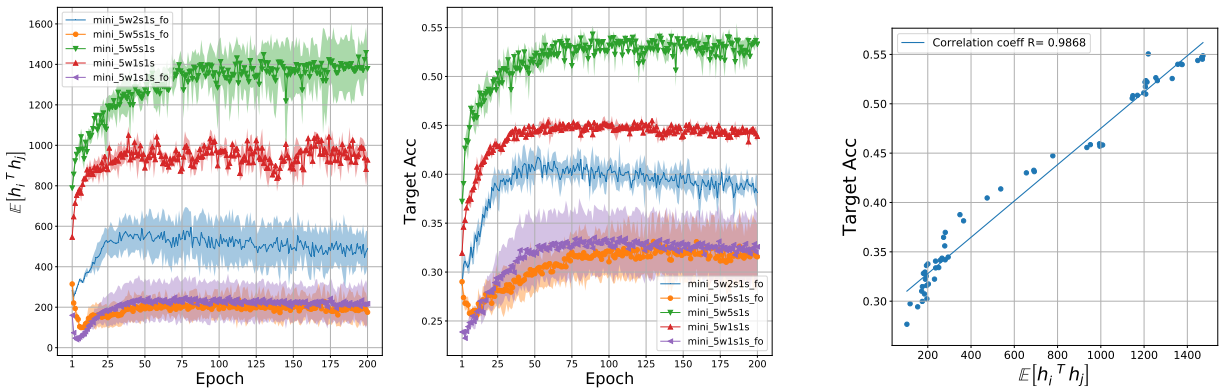
We use the architecture proposed by (Vinyals et al., 2016) which is used by (Finn et al., 2017), consisting of 4 modules stacked on each other, each being composed of 64 filters of of 3×3 convolution, followed by a batch normalization layer, a ReLU activation layer, and a 2×2 max-pooling layer. With Omniglot, strided convolution is used instead of max-pooling, and images are downsampled to 28×28. With MiniImagenet, we used fewer filters to reduce overfitting, but used 48 while MAML used 32. As a loss function to minimize, we use cross-entropy between the predicted classes and the target classes.

5.2. Hyperparameters used in meta-training and meta-testing for few-shot classification

We follow the same experimental setup as (Finn et al., 2017) for training and testing the models using MAML and First-Order MAML. During meta-training, the inner loop updates are performed using a fixed learning rate α of 0.1 for Omniglot and 0.01 for MiniImagenet, while ADAM is used as the optimizer for the meta-update, without any learning rate scheduling, using a meta-learning rate β of 0.001. At meta-test time, adaptation to meta-test task is always performed using a fixed number of steps, same as for the meta-training inner loop updates. We use either one or five steps, depending on the experiment. With Omniglot, we use batches of 16 and 8 tasks for the 1-shot and 5-shot settings respectively, while for the MiniImagenet experiments, we use batches of 4 and 2 tasks for the 1-shot and 5-shots settings respectively. Let's also precise that, in *k-shot* learning for an *m-way* classification task $\mathcal{T}_i$, the set of support samples $\mathcal{D}_i$ comprises $k \times m$ samples. Each meta-training epoch comprises 500 meta-training iterations.

6. The learned space of representations

The observations of (Guiroy et al., 2019) on the coherence of meta-test gradients, and their relation to generalization to new tasks, are surprising yet hard to interpret intuitively. In this section, we attempt to provide such interpretation, which is an informal hypothesis rather than a theoretical claim, but which we further verify empirically. Our intuition is that MAML, in order to represent the data and classify from few examples, might be learning a representation space in which similarity among embeddings is captured by their inner product, as in Eq 4. Recently, the authors of (Raghu et al., 2019), while not explicitly studying generalization, showed that for MAML at meta-test time, the vectors h representing the new task inputs, barely change during fine-tuning of the model, as opposed to the outputs of f_{lin}, a phenomenon they name "feature reuse". In a related way, a Prototypical Network (Snell et al., 2017) learns a metric space in which an example x is classified based on a softmax on the Euclidean distances between its vector h and the learned cluster mean vectors. In MAML, an example x is classified according the highest score from the logits of f_{lin}, scores which are proportional to the inner product of h with their respective row of the weight matrix W. Our intuition is that vectors h that have higher mutual inner products could lead to higher inner product between task gradients, on average, and that MAML would learn a representation space learned where similarity between embeddings is based on the inner product. To empirically verify this, we computed the average inner product between the vectors h_i from all images of the meta-test data, produced by f_{feat} at θ^s, after each meta-train epoch. In Figure 1a, we observe that meta-overfitting reflects $\mathbb{E}[\mathbf{h}_i^T\mathbf{h}_j]$, and in Figure 1b we show the correlation between $\mathbb{E}[\mathbf{h}_i^T\mathbf{h}_j]$ and generalization. These results suggest that this interpretation is plausible, while further theoretical work is required to further validate it. Thus according to this proposed interpretation, for MAML, meta-training would learn a representation space in which



(a) Comparison between average inner product between representation vectors and average target accuracy on meta-test tasks. The metric captures the difference in performance between different settings (number of shots, first vs. second order MAML). Qualitatively, we observe that it reflects meta-overfitting

(b) Correlation between peak value of the metric (across meta-train epochs), its and related target accuracy, MAML and First-Order MAML, with k varying between 1 and 5

Figure 1. Comparison between average inner product between representation vectors, generated by the feature network at meta-train solution, and average target accuracy on meta-test tasks, for different regimes of MAML and First-Order MAML on MiniImagenet.

the embeddings for the new, previously unseen data will gradually appear more similar to each other according to their inner product. The model would gradually learn general features, which are able to represent new data, more similarly, but as meta-overfitting occurs, they become too specific to the training classes less general w.r.t. new data.

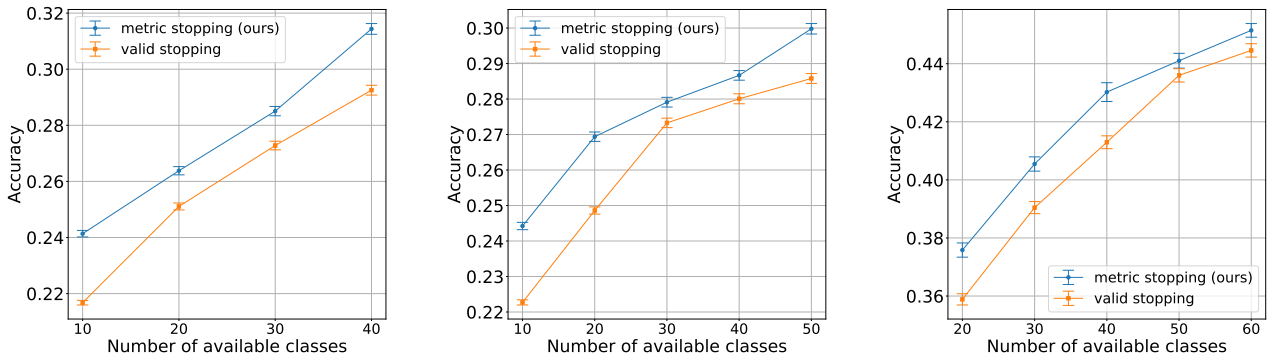
7. Unsupervised Model Selection for Gradient Based Meta Training

Based on the analysis presented so far in this work, we propose an "unsupervised" model selection criterion for gradient-based meta learning based on Eq 4. In gradient-based meta-training (or in other general meta-learning frameworks), the model selection is typically based on the performance of the model on the average target accuracy of the meta-validation tasks, after adaptation. This has two pitfalls: (1) The distribution of support and target samples in the meta-validation tasks may significantly differ from that of the meta-test task, potentially resulting in the sub-optimal model selection. (2) Using the held-out meta-validation set reduces the number of samples and classes that can be used in the meta-training.

Our proposed method address both of the above limitations. In particular, for any given meta-test task, we propose to use its support set for model selection as follows: Let us assume that we train a model using a gradient-based meta-learning algorithm for 1 to N epochs. For a given meta-test task, for each of the N epochs, we compute a metric which is the average inner product of the representation vectors of n samples in its support set: $\frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} \mathbf{h}_i^T \mathbf{h}_j$, and select the model at the epoch where the aforementioned metric has highest value. To validate the effectiveness of this method, for a given meta-test task, we compute the target accuracy of the selected model. We perform this over 5 independent runs, each with 500 meta-test tasks. We perform this analysis with MAML and First-Order MAML on MiniImagenet and Omniglot for different number of ways and shots.

As a baseline, we perform the model selection by fine-tuning the model on the support set of the meta-validation tasks and measuring the accuracy of this fine-tuned model on the target set. In Table 1, we see that our selection criterion offer only slightly lower performance than selection based on meta-validation tasks. Despite being comparable to the meta-validation tasks based model selection, our method has a big practical advantage that it does not require held-out validation data, thus the meta-validation data could be incorporated in the meta-training data, potentially leading to better generalization on the unseen tasks. Furthermore, since we do not use the labels of the support set for this criterion of model selection, it is effectively an unsupervised method. Thus we can leverage the potentially large amount of unlabeled samples from the meta-test task to measure this criterion.

To further validate the effectiveness of our proposed method, we demonstrate that incorporating meta-validation classes into the meta-training set, and relying on our metric for selecting the epoch at which to test the model. See Figure 2. We observe that our method, using same number of available classes, achieves better generalization to new test classes, compared to the standard meta-training where a portion of those classes are used for validation. These results suggest



(a) 5-way, 1-shot, 1-step, First-Order (b) 5-way, 5-shot, 1-step, First-Order (c) 5-way, 1-shot, 5-step, First-Order

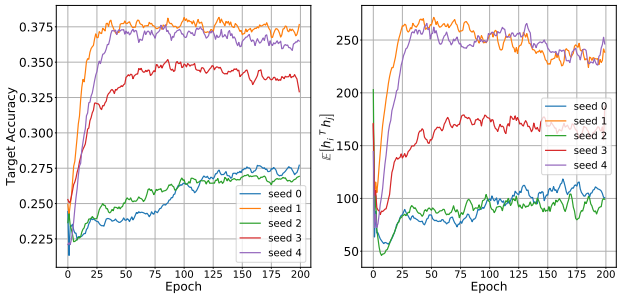
Figure 2. Model Selection: For our method, we use all available classes to train the model, and rely on $\mathbb{E}[\mathbf{h}_i^T \mathbf{h}_j]$ at meta-test time to perform model selection. For the standard validation based model selection method, 10 out of the available classes are used for validation, while the rest is used for training (unless when a total of 10 classes is available, then 5 are used for validation). In each setting, we observe that our method achieves better generalization to new classes than the standard validation based method, demonstrating the advantage of using the extra number of classes while using $\mathbb{E}[\mathbf{h}_i^T \mathbf{h}_j]$ to choose which at epoch to test the model. Results include 95% confidence intervals.

	MiniImagenet			**Omniglot**			
	5-way, 1-shot		5-way, 5-shot	5-way, 5-shot		20-way, 5-shot	
	MAML	First-Order MAML	MAML	MAML	First-Order MAML	MAML	First-Order MAML
$\mathbb{E}\left[\mathbf{h}_i{}^T \mathbf{h}_j\right]$ **(ours)**	45.8	45.0	60.5	98.6	98.2	92.5	93.9
Meta-val	47.1	45.4	61.5	98.6	98.7	93.4	94.2

Table 1. Model selection. For each meta-test task, performance is measured as target accuracy at selected model (epoch). The performance of our method is on average, across settings, 0.9% lower than the meta-validation selection, on MiniImagenet, while for Omniglot, the average difference is 0.5%.

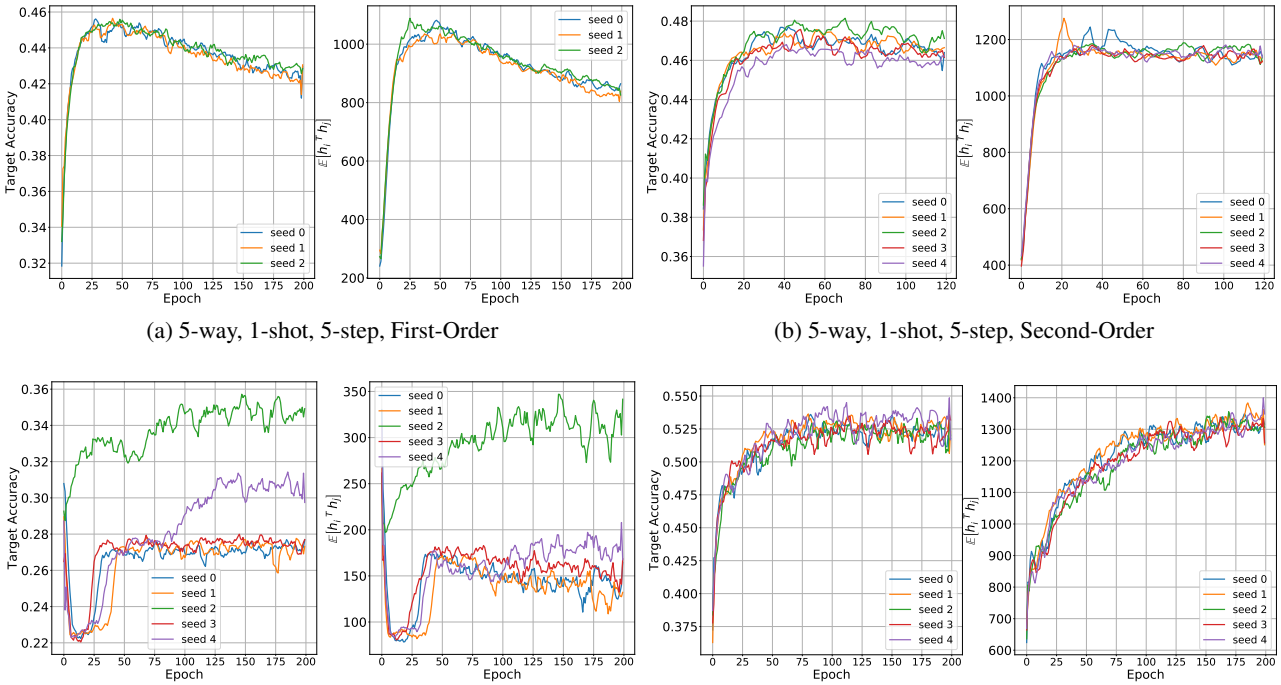
the advantage of using the extra number of classes while using $\mathbb{E}[\mathbf{h}_i^T \mathbf{h}_j]$ to choose which at epoch to test the model.

In addition, we present qualitative evaluations of our model selection method, on MiniImagenet, These results suggest that our metric reflects the target accuracy of meta-test tasks. For multiple runs, we plot the target accuracy averaged over 500 tasks, against the model selection metric computed on the support set of those tasks. In Figure 3, we first demonstrate the method where the model uses one step of adaptation at meta-test time (also during meta-training), when the model uses only one shot of support samples. In Figure 4, we show results where the model uses multiple steps of adaptation, again using of shot of support examples. Finally, in Figure 3, the model uses multiple shots of support examples.



(a) 5-way, 1-shot, 1-step, First-Order

Figure 3. Model Selection: 1 step of adaptation. MiniImagenet



(a) 5-way, 1-shot, 5-step, First-Order

(b) 5-way, 1-shot, 5-step, Second-Order

(c) 5-way, 5-shot, 1-step, First-Order

(d) 5-way, 5-shot, 1-step, Second-Order

Figure 4. Model Selection: 5 shots of support samples. MiniImagenet. As it can be seen in most settings, the results suggest that our metric reflects the target accuracy of meta-test tasks.

8. Conclusion

Using the few-shot image classification setting, we provide empirical evidence that when using gradient-based meta-learning algorithms, generalization to new tasks is correlated with global similarity within the learned space of representation, measured by the average inner product between the embeddings of the meta-test examples. Based on these observations, we proposed a model-selection criterion.

References

Denevi, G., Ciliberto, C., Grazzi, R., and Pontil, M. Learning-to-learn stochastic gradient descent with biased regularization. *CoRR*, abs/1903.10399, 2019. URL http://arxiv.org/abs/1903.10399.

Finn, C., Abbeel, P., and Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. *CoRR*, abs/1703.03400, 2017. URL http://arxiv.org/abs/1703.03400.

Finn, C., Rajeswaran, A., Kakade, S. M., and Levine, S. Online meta-learning. *CoRR*, abs/1902.08438, 2019. URL http://arxiv.org/abs/1902.08438.

Guiroy, S., Verma, V., and Pal, C. J. Towards understanding generalization in gradient-based meta-learning. *CoRR*, abs/1907.07287, 2019. URL http://arxiv.org/abs/1907.07287.

Khodak, M., Balcan, M., and Talwalkar, A. Provable guarantees for gradient-based meta-learning. *CoRR*, abs/1902.10644, 2019. URL http://arxiv.org/abs/1902.10644.

Nichol, A., Achiam, J., and Schulman, J. On first-order meta-learning algorithms. *CoRR*, abs/1803.02999, 2018. URL http://arxiv.org/abs/1803.02999.

Oreshkin, B. N., López, P. R., and Lacoste, A. TADAM: task dependent adaptive metric for improved few-shot learning. *CoRR*, abs/1805.10123, 2018. URL http://arxiv.org/abs/1805.10123.

Raghu, A., Raghu, M., Bengio, S., and Vinyals, O. Rapid learning or feature reuse? towards understanding the effectiveness of maml, 09 2019.

Ravi, S. and Larochelle, H. Optimization as a model for few-shot learning. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017. URL https://openreview.net/forum?id=rJY0-Kcll.

Rothfuss, J., Lee, D., Clavera, I., Asfour, T., and Abbeel, P. Promp: Proximal meta-policy search. *CoRR*, abs/1810.06784, 2018. URL http://arxiv.org/abs/1810.06784.

Snell, J., Swersky, K., and Zemel, R. S. Prototypical networks for few-shot learning. *CoRR*, abs/1703.05175, 2017. URL http://arxiv.org/abs/1703.05175.

Vinyals, O., Blundell, C., Lillicrap, T. P., Kavukcuoglu, K., and Wierstra, D. Matching networks for one shot learning. *CoRR*, abs/1606.04080, 2016. URL http://arxiv.org/abs/1606.04080.