MATH 6380O Project1 Report
Investigation on Model-Agnostic Meta-Learning

Lanqing Xue
20361709
lxueaa@connect.ust.hk

Feng Han
20407369
fhanab@connect.ust.hk

Jianyue Wang
20550738
jwangen@connect.ust.hk

Zhiliang Tian
20526153
ztianac@connect.ust.hk

October 6, 2019

1 Introduction

In this project, we follow MAML [2], a milestone paper in meta-learning area, which proposed a fast adaptation algorithm as a meta-learning and applied it on some supervised learning tasks and reinforcement learning tasks. In that paper, the algorithm for meta-learning that is model-agnostic, in the sense that it is compatible with many models that trained with gradient descent and applicable to a variety of different learning problems, including classification, regression, and reinforcement learning in the few-shot scenarios.

The primary contribution of this MAML paper is a simple model and task-agnostic algorithm for meta-learning that trains a model's parameters such that a small number of gradient updates will lead to fast learning on a new task. We demonstrate the algorithm on different model types, including fully connected and convolutional networks. MAML paper evaluate the algorithm on several distinct tasks, including few-shot image classification and reinforcement learning. The meta-learning algorithm of MAML compares favorably to state-of-the-art one-shot learning methods designed specifically for supervised classification, while using fewer parameters, but that it can also be readily applied to regression and can accelerate reinforcement learning in the presence of task variability, substantially outperforming direct pretraining as initialization.

We choose that paper to conduct our investigation since 1. it is a well-known and significant paper in meta-learning with 509 citations and was published in ICML 2017. 2. It involves many latest and interesting techniques of machine learning, including meta-learning, few shot classification. and it applies meta-learning on reinforcement learning. Thus, this paper can let us study and understand a lot about these new techniques.

In this project, we conduct an empirical study and investigation on MAML. we will implement following ideas and experiments:

1. Reproduce MAML on image classification and study how to make it work.
2. Conduct ablation & hyper-parameter study on image classification tasks.
3. Try MAML on some new datasets.
4. Try to improve MAML with some practical strategies.

In following parts, we will introduce the MAML model (on both few shot classification), the experiments (reproduce and some more studies on different setting and dataset), and further improvements.

2 MAML model

Meta-learning is a class of methods used for solving few-shot learning problem, that is learning how to get highest performance using only a few of training examples available. That is achieved not with learning how to learn the network, while doing it from scratch, but of learning the best start to learn from, which are initial parameters for each other new task, few-shot learning is available to. Thus that approach is helpful for training either classification tasks on small datasets or using active learning with reduced number of training examples from oracle or even with reinforcement learning, which always suffers from sample inefficiency.

Few-shot learning requires of quickly adapting of model f, which maps observations x to outputs a, to perform efficiently on a new task (seen or unseen). Thus during meta-learning the model is trained on a large or infinite number of tasks $\mathcal{T}_j$ sampled from distribution of tasks $p(\mathcal{T})$, in order to adapt itself quickly from the trained parameters of model f to a new task $\mathcal{T}_i$. Samples to learn on are sampled from tasks as $\mathcal{T}_j = \{\mathcal{L}_{\mathcal{T}_j}(x_1, a_1, \cdots, x_H, a_H), q_j(x_1), q_j(x_{t+1}|x_t, a_t), H\}$, which consists of loss $\mathcal{L}$, distribution over initial observations $q(x_1)$, a transition distribution $q(x_{t+1}|x_t, a_t)$ and episodic length H. While evaluating on new task $\mathcal{T}_i$ K-shot learning setting model is adapting using only K samples x_1 drawn from q_i and then lower the feedback $\mathcal{L}_{\mathcal{T}_i}$ and checking performance on new samples from $\mathcal{T}_i$. For the meta-training process these feedbacks $\mathcal{L}_{\mathcal{T}_j}$ on the testing samples are used as test errors for task $\mathcal{T}_j$ as well as the training errors of the meta-learning process. Meta-performance is measured as the model's performance on tasks generated from $\mathcal{T}$ and trained using K-samples from that task. Generally, tasks used for meta-testing are held out during meta-training.

2.1 MAML for image classification

Few-shot learning is an important domain of machine learning tasks, where the goal is to learn a new function from only a few input or output pairs for that task, while using prior data from similar tasks for meta-learning. For instance,

the goal might be to classify images of a Segway after seeing only one or a few examples, with a model that has previously seen many other types of objects.

To formalize the supervised few-shot classification problems in the context of the meta-learning definitions in [2] as following: we can define the horizon $H = 1$ and drop the timestep subscript on x_t, since the model accepts a single input and produces a single output, rather than a sequence of inputs and outputs. The task τ_i generates K i.i.d. observations x from q_i, and the task loss is represented by the error between the model's output for x and the corresponding target values y for that observation and task.

The detail algorithm of MAML can be found in 1 (Algorithm 2)

Algorithm 2 MAML for Few-Shot Supervised Learning

Require: $p(\mathcal{T})$: distribution over tasks
Require: α, β: step size hyperparameters
1: randomly initialize θ
2: **while** not done **do**
3: Sample batch of tasks $\mathcal{T}_i \sim p(\mathcal{T})$
4: **for all** $\mathcal{T}_i$ **do**
5: Sample K datapoints $\mathcal{D} = \{\mathbf{x}^{(j)}, \mathbf{y}^{(j)}\}$ from $\mathcal{T}_i$
6: Evaluate $\nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)$ using $\mathcal{D}$ and $\mathcal{L}_{\mathcal{T}_i}$ in Equation (2) or (3)
7: Compute adapted parameters with gradient descent: $\theta_i' = \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)$
8: Sample datapoints $\mathcal{D}_i' = \{\mathbf{x}^{(j)}, \mathbf{y}^{(j)}\}$ from $\mathcal{T}_i$ for the meta-update
9: **end for**
10: Update $\theta \leftarrow \theta - \beta \nabla_\theta \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}_{\mathcal{T}_i}(f_{\theta_i'})$ using each $\mathcal{D}_i'$ and $\mathcal{L}_{\mathcal{T}_i}$ in Equation 2 or 3
11: **end while**

MAML for Few-Shot Supervised Learning

3 Experiments

In this part, we will show our experiments on reproduction, ablation study and some attempts on new datasets and the strategies to improve MAML.

3.1 Reproduction on Classification tasks

We evaluate MAML by applying our MAML to few-shot image recognition task on the Omniglot dataset and MiniImagenet dataset. We follow the experimental protocol proposed by [6], which involves fast learning of classification with 1 or 5 shots. The classification task is set up as follows: 1. select N unseen classes, 2. provide the model with K different instances of each of the N classes, 3.evaluate the model's ability to classify new instances within N classes.

For the Omniglot dataset, we randomly select 1200 characters for training, irrespective of alphabet while using the remaining for testing. The Omniglot

dataset is augmented with rotations by multiples of 90 degrees. For MiniImagenet, we use the same setting as done by [2, 5]. We verify our performance by accuracy as in original paper [2].

3.1.1 Mini-Imagenet Dataset

The MiniImagenet dataset was proposed by [5], and involves 64 training classes, 12 validation classes, and 24 test classes. The Omniglot and MiniImagenet image recognition tasks are the most common recently used few-shot learning benchmarks used in many few-shot learning papers.

Experiments & results
Exp1 is the experiment to reproduce the original MAML paper in 5-shot.

MAML Paper	Ours(reproduced)
63.51	64.87

Exp1: The performance on reproducing MAML

Exp2 is based on all the setting of Exp1 except the shot numbers, which indicating the more examples in shot can increase the performance.

	1-shot	5-shot
MAML(ours)	48.56	64.87

Exp2: The study on shot number

3.1.2 Omniglot Dataset

For Omniglot dataset, we shuffle all character classes and randomly select 1150 for the training set and from the remaining classes. We use 50 for validation and 423 for testing. In most few-shot learning papers, the first 1200 classes are used for training and the remaining for testing. However, having a small validation set to choose the best model is crucial, so we choose to use a small set of 50 classes as validation set. For each class, we use all available 20 samples in the sets. Furthermore, for the Omniglot dataset, data augmentation is used on the images in the form of rotations of 90 degree increments. Class samples that are rotated are considered new classes, e.g. a 180 degree rotated character C is considered a different class from a non rotated C, thus effectively having 1623 x 4 classes in total. However, the rotated classes are generated dynamically after the character classes have been split into the sets such that rotated samples from a class reside in the same set.

Experiments & results
Exp3 is the experiment to reproduce the performance in the original MAML paper in 5-way for 1-shot and 5-shot learning.

	1 shot	5-shot
MAML Paper	98.7	99.9
MAML (reproduced)	98.75	99.72

Exp3: The performance on reproducing MAML

3.2 Ablation Studies

3.2.1 Mini-Imagenet Dataset

Exp4 is based on all the setting of Exp1 except the inner loop learning rate (α in Figure 1) (0.01 is default one)

α	0.1	0.05	**0.01**	0.005	0.001
MAML(ours)	64.48	65.10	64.87	64.80	22.76

Exp4: The study on the inner loop learning rate (α in Figure 1)

Exp5 is based on all the setting of Exp1 except the meta learning rate (out loop) (β in Figure 1) (0.001 is default one)

β	0.01	0.005	**0.001**	0.0005	0.0001
MAML(ours)	20.08	24.75	64.87	68.07	45.02

Exp5: The study on the meta learning rate (out loop) (β in Figure 1)

Exp6 is based on all the setting of Exp1 except the number of stage for convolutional layers (4 is default one)

# stages	2	**4**	8
MAML(ours)	50.00	64.87	62.32

Exp6: The study on the number of stage

Exp7 is based on all the setting of Exp1 except whether use batch normalization or layer normalization (batch normalization is the default).

	batchNorm	LayerNorm
MAML(ours)	64.87	53.69

Exp7: The study on batch normalization or layer normalization

Exp8 is based on the random seed, which indicates the algorithm is sensitive to the random seed.

3.2.2 Omniglot Dataset

Exp9 and Exp10 are based on all the setting of Exp3 except the shot numbers, which indicating the more examples in shot can increase the performance. We

random seed	**0**	2	4
MAML(ours)	64.87	65.22	64.70

Exp8: The study on the random seed

checked it for 5-way and 20-way few-shot learning tasks.

	1-shot	3-shot	5-shot
MAML (5-way)	98.75	99.39	99.72

Exp9: The study on K-shot number for 5-way few-shot classification

	1-shot	3-shot
MAML (20-way)	91.50	96.84

Exp10: The study on K-shot number for 20-way few-shot classification

3.2.3 Discussion

From Exp1 to Exp3, we can infer that our implementation of MAML is accurate enough. Based on it we made other experiments, so it was important to check if main result is reproducible.

From Exp4 to Exp8, we know that MAML is more sensitive to the meta learning rate (outer loop) rather than the inner loop learning rate, which control the learning strategies and determind the learning quality among different subtasks. MAML is also sensitive to the random seed, showing the algorithm is not so robust in terms of the variance. The depth of convolutional neural network (stage) and the batch normalization trick is crucial to the performance.

Exp9 and Exp10 turned to be that including new K-shot parameter, with $K = 3$ makes monotonic improvement from $K = 1$ and is bounded with performance on $K = 5$, which shows reasonable dependence from the amount of available information. Most significant difference is seen from 20-way few-shot classification, shown in Exp3, which indicates the complexity of learning considerable amount of data form banch of different task simultaneously to find fast-adaptation parameters. Moreover, enlarging number of tasks to train on and evaluate means, that more.

3.3 Adaption on new datasets

After applying our MAML on the classification tasks on Omniglot and MiniImagenet, we try to adapt MAML on a new image classification dataset, cifar-10 in a similar way with the above two datasets and report the result.

We design three different experiments to examine the performance of MAML on cifar-100, and we use the same experimental setting as [4] (including the split of train, test, valid)

1. Cifar100(TADAM). Apply TADAM [4] on cifar-100, which is implemented by [4]. We can treat it as a competitive baseline since that is proposed after MAML.
2. Cifar100(MAML). Apply MAML on cifar-100 as the same setting with [4].
3. mimg(MAML). Train MAML on mini-imagenet and test on cifar100.
4. Cifar100+mimg(MAML). Use MAML, training on both mini-imagenet and cifar100, and testing on cifar100

Cifar100(TADAM)	Cifar100(MAML)	mimg(MAML)	Cifar100+mimg(MAML)
56.10	49.81	50.76	52.74

The study on new dataset

3.4 Strategies to improve MAML

We find some strategies to improve MAML and verify them by experiments.

- MSL. Due to gradient instability we considered the **Multi-Step Loss Optimization (MSL):** MAML works by minimizing the target set loss computed by the base-network after it has completed all of its inner-loop updates towards a support set task. Instead was proposed minimizing the target set loss computed by the base-network after every step towards a support set task [1]. More specifically, we propose that the loss minimized is a weighted sum of the target set losses after every support set loss update. More formally:

$$\theta = \theta - \beta \nabla_{\theta} \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \sum_{l=0}^{N_{\text{inner}}} w_l \mathcal{L}_{\mathcal{T}_i}(f_{\theta_l'}) \tag{1}$$

  This method is widely used for classification tasks, while for reinforcement learning task it is redundant due to only one inner gradient updates, which can be computed for each task in reinforcement learning setting.

- LSLR. Together with the problem of shared inner loop learning rate (across step and across parameters), which is called **layer-wise learning rate (LSLR)** trick [4]. The work in [3] demonstrated that learning a learning rate and gradient direction for each parameter in the base- network improved the generalization performance of the system. However, that had the consequence of increased number of parameters and increased computational overhead. So instead, it was proposed in [1], learning a learning rate and direction for each layer in the network as well as learning different learning rates for each adaptation of the base-network as it takes steps.

- BNRS. BNRS was used Per-step Batch Normalization weights and Biases trick, which lead to **batch normalization trick statistic** trick [1]. To fix the dissimilarity of feature distributions from each other step in the inner loop, in the paper [1] was propose learning a set of biases per-step within the inner-loop update process. Doing so, means that batch normalization will learn biases specific to the feature distributions seen at each set, which should increase convergence speed, stability and generalization performance.

Exp11 is to test the layer-wise learning rate trick (LSLR) [1] and the batch normalization statistic trick (BNRS). The layer-wise learning rate trick [1] means assign different learnable learning rate per-layer and learn to get the learning rate.

MAML	MAML+LSLR	MAML+BNRS
64.87	65.44	66.10

Exp11: The ablation studies of LSLR and BNRS on LSLR dataset

In Exp12, we made ablation studies on Omniglot with MSL, LSLR and BNRS improvement tricks:

	5-way,1-shot
MAML	98.14
MAML+LSLR	98.41
MAML+MSL	98.34
MAML+BNRS	98.51
MAML+LSLR+MSL	98.47
MAML+BNRS+LSLR	98.75
MAML+MSL+BNRS	98.51
MAML+All	98.81

Exp12: The ablation study of MSL, LSLR and BNRS on Omniglot dataset

Ablation study on using different strategies (Exp11 Exp12) shows, that using all proposed tricks in "MAML+All" is much favourable, as it can increase more than half of the percent of accuracy measure, while having the same variance in the output. While BNstat trick have the biggest influence, MSL trick is of less importance.

4 Contribution

We follow MAML [2], a fast adaptation algorithm for meta-learning and apply this algorithm on some supervised learning tasks. Besides, we conduct a series of investigation including the ablation & hyper-parameter study, application on some new tasks, test some new strategies and propose some suggestions for it.

5 Division of labor

Lanqing Xue, CSE 20361709 : reproduce and conduct an empirical on Omniglot dataset.

Jianyue Wang, ISOM 20550738: reproduce and conduct an empirical on MiniImagenet dataset.

Han Feng, CSE 20407369 : Apply MAML on new datasets; Help Zhiliang to imporve MAML by LSLR and MSL on MiniImagenet.

Zhiliang Tian, CSE 20526153: Imporve MAML by BNRS, LSLR and MSL on Omniglot and MiniImagenet dataset.

We divided the labor to implement separately and write the report together.

References

[1] Antreas Antoniou, Harrison Edwards, and Amos Storkey. How to train your maml. *arXiv preprint arXiv:1810.09502*, 2018.

[2] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1126–1135. JMLR. org, 2017.

[3] Ke Li and Jitendra Malik. Learning to optimize. *CoRR*, abs/1606.01885, 2016.

[4] Boris Oreshkin, Pau Rodríguez López, and Alexandre Lacoste. Tadam: Task dependent adaptive metric for improved few-shot learning. In *Advances in Neural Information Processing Systems*, pages 721–731, 2018.

[5] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. 2016.

[6] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Advances in neural information processing systems*, pages 3630–3638, 2016.