Meta-Learning with Network Pruning (Supplementary Materials)

Hongduan Tian[1], Bo Liu[2], Xiao-Tong Yuan[1], and Qingshan Liu[1]

[1] B-DAT Lab, Nanjing University of Information Science and Technology, Nanjing, 210044, China
[2] JD Finance America Corporation, Mountain View, CA 94043, USA
{hongduan.tian,kfliubo,xtyuan1980}@gmail.com, qsliu@nuist.edu.cn

A Proofs of Results

A.1 Proof of Theorem 1

We need the following lemma which guarantees the uniform convergence of $\mathcal{R}_S(\theta)$ towards $\mathcal{R}(\theta)$ for all θ when the loss function is Lipschitz continuous and smooth, and the optimization is limited on a bounded domain.

Lemma 1. *Assume that the domain of interest $\Theta \subseteq \mathbb{R}^p$ is bounded by R and the loss function $\ell(f_\theta(\boldsymbol{x}), y)$ is G-Lipschitz continuous and H-smooth with respect to θ. Also assume that $0 \leq \ell(f_\theta(\boldsymbol{x}), y) \leq B$ for all $\{f_\theta(\boldsymbol{x}), y\}$. Then for any $\delta \in (0, 1)$, the following bound holds with probability at least $1 - \delta$ over the random draw of sample set S for all $\theta \in \Theta$,*

$$|\mathcal{R}(\theta) - \mathcal{R}_S(\theta)| \leq \mathcal{O}\left(B\sqrt{\frac{\log(1/\delta) + p\log(\sqrt{M}GR(1+\eta H)/B)}{M}}\right).$$

Proof. For any task T, let us denote $\tilde{\ell}(\theta; T) := \mathcal{L}_{\mathcal{D}_T^{query}}\left(\theta - \eta\nabla_\theta \mathcal{L}_{\mathcal{D}_T^{supp}}(\theta)\right)$. Since $\ell(f_\theta(\boldsymbol{x}), y)$ is G-Lipschitz continuous with respect to θ, we can show that

$$\begin{aligned}
|\tilde{\ell}(\theta; T) - \tilde{\ell}(\theta'; T)| \leq & G\|\theta - \eta\nabla_\theta \mathcal{L}_{\mathcal{D}_T^{supp}}(\theta) - \theta' + \eta\nabla_\theta \mathcal{L}_{\mathcal{D}_T^{supp}}(\theta')\| \\
\leq & G\left(\|\theta - \theta'\| + \eta\|\nabla_\theta \mathcal{L}_{\mathcal{D}_T^{supp}}(\theta) - \nabla_\theta \mathcal{L}_{\mathcal{D}_T^{supp}}(\theta')\|\right) \\
\leq & G(1 + \eta H)\|\theta - \theta'\|,
\end{aligned}$$

which indicates that $\tilde{\ell}(\theta; T)$ is $G(1 + \eta H)$-Lipschitz continuous for any task T.

As a subset of an L_2-sphere, it is standard that the covering number of Θ with respect to the L_2-distance is upper bounded by

$$\mathcal{N}(\epsilon, \Theta, L_2) \leq \mathcal{O}\left(\left(1 + \frac{R}{\epsilon}\right)^p\right).$$

Since the task-level loss function $\tilde{\ell}(\theta; T)$ is $G(1 + \eta H)$-Lipschitz continuous as shown above, it can be verified that the covering number of the class of functions

$\tilde{\mathcal{L}} = \left\{ T \mapsto \tilde{\ell}(\theta; T) \mid \theta \in \Theta \right\}$ with respect to L_∞-distance $L_\infty(\tilde{\ell}(\theta_1;\cdot), \tilde{\ell}(\theta_2;\cdot)) := \sup_T |\tilde{\ell}(\theta_1; T) - \tilde{\ell}(\theta_2; T)|$ is given by

$$\mathcal{N}(\epsilon, \tilde{\mathcal{L}}, L_\infty) \leq \mathcal{N}\left(\frac{\epsilon}{G(1+\eta H)}, \Theta, L_2\right) \leq \mathcal{O}\left(\left(1 + \frac{GR(1+\eta H)}{\epsilon}\right)^p\right).$$

Therefore, there exists a set of points $\Omega \subseteq \mathbb{R}^p$ with cardinality at most $\mathcal{N}(\epsilon, \tilde{\mathcal{L}}, L_\infty)$ such that the following bound holds for any $\theta \in \Theta$:

$$\min_{\omega \in \Omega} |\tilde{\ell}(\theta; T) - \tilde{\ell}(\omega; T)| \leq \epsilon, \ \forall T.$$

For an arbitrary $\omega \in \Omega$, based on Hoeffdings inequality (note that $\ell(\cdot,\cdot) \leq B$ implies $\tilde{\ell}(\cdot,\cdot) \leq B$) we have

$$\mathbb{P}\left(|\mathcal{R}_S(\omega) - \mathcal{R}(\omega)| > t\right) \leq \exp\left\{-\frac{Mt^2}{2B^2}\right\}.$$

For any $\theta \in \Theta$, based on triangle inequality we can show that there exits $\omega_\theta \in \Omega$ such that

$$\begin{aligned}|\mathcal{R}_S(\theta) - \mathcal{R}(\theta)| =& |\mathcal{R}_S(\theta) - \mathcal{R}_S(\omega_\theta) + \mathcal{R}_S(\omega_\theta) - \mathcal{R}(\omega_\theta) + \mathcal{R}(\omega_\theta) - \mathcal{R}(\theta)| \\ \leq& 2\epsilon + |\mathcal{R}_S(\omega_\theta) - \mathcal{R}(\omega_\theta)| \leq 2\epsilon + \max_{\omega \in \Omega} |\mathcal{R}_S(\omega) - \mathcal{R}(\omega)|.\end{aligned}$$

Applying uniform bound we know that

$$\begin{aligned}&\mathbb{P}\left(\sup_{\theta \in \Theta} |\mathcal{R}(\theta) - \mathcal{R}_S(\theta)| \geq 2\epsilon + t\right) \\ \leq& \mathcal{N}(\epsilon, \mathcal{L}, \ell_\infty) \exp\left(-\frac{Mt^2}{2B^2}\right) \leq \mathcal{O}\left(\left(1 + \frac{GR(1+\eta H)}{\epsilon}\right)^p \exp\left(-\frac{Mt^2}{2B^2}\right)\right).\end{aligned}$$

Let us choose $\epsilon = B/\sqrt{M}$ and

$$t = \sqrt{2}B\sqrt{\frac{\log(1/\delta) + p\log(GR(1+\eta H)/\epsilon)}{M}}$$

such that the right hand side of the previous inequality equals δ. Then we obtain that with probability at least $1 - \delta$

$$\sup_{\theta \in \Theta} |\mathcal{R}(\theta) - \mathcal{R}_S(\theta)| \leq \mathcal{O}\left(B\sqrt{\frac{\log(1/\delta) + p\log(\sqrt{M}GR(1+\eta H)/B)}{M}}\right).$$

This proves the desired result.

Based on this lemma, we can readily prove the main result in the theorem.

Proof (Proof of Theorem 1). For any fixed supporting set $J \in \mathcal{J}$, by applying Lemma 1 we obtain that the following uniform convergence bound holds for all θ with $\text{supp}(\theta) \subseteq J$ with probability at least $1-\delta$ over S:

$$|\mathcal{R}(\theta) - \mathcal{R}_S(\theta)| \leq \mathcal{O}\left(B\sqrt{\frac{\log(1/\delta) + k\log(\sqrt{M}GR(1+\eta H)/B)}{M}}\right).$$

Since by constraint the parameter vector θ is always k-sparse, we thus have $\text{supp}(\theta) \in \mathcal{J}$. Then by union probability we get that with probability at least $1-\delta$, the following bound holds for all θ with $\|\theta\|_0 \leq k$:

$$|\mathcal{R}(\theta) - \mathcal{R}_S(\theta)| \leq \mathcal{O}\left(B\sqrt{\frac{\log(|\mathcal{J}|) + \log(1/\delta) + k\log(\sqrt{M}GR(1+\eta H)/B)}{M}}\right).$$

It remains to bound the cardinality $|\mathcal{J}|$. From [3, Lemma 2.7] we know $|\mathcal{J}| = \binom{p}{k} \leq \left(\frac{ep}{k}\right)^k$, which then implies the desired generalization gap bound. This completes the proof.

A.2 Proof of Corollary 1

Proof. Let $\mathcal{R}_\gamma$ be a population version of $\mathcal{R}_{\gamma,S}$ with margin-based loss function ℓ_γ used for computing both $\mathcal{L}_{\mathcal{D}_T^{supp}}$ and $\mathcal{L}_{\mathcal{D}_T^{query}}$. Since ℓ_γ is a surrogate of the binary loss as used by $\tilde{\mathcal{R}}$ for query classification error evaluation, we must have $\tilde{\mathcal{R}} \leq \mathcal{R}_\gamma$. Then the desired bound follows directly by invoking Theorem 1 to the considered margin loss.

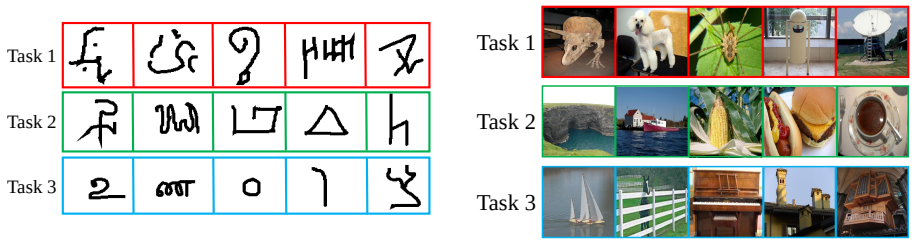
B Detailed Experimental Settings

B.1 Model

The model used in our experiments is consistent with that considered for Reptile[2]. The model used throughout the experiment contains 4 sequential modules. Each module has a convolutional layer with 3×3 kernel, followed by a batch normalization and a ReLU activation. Additionally for the experiments on Mini-ImageNet, a 2×2 max-pooling pooling is used on the batch normalization layer output while for Omniglot a stride of 2 is used in convolution. The above network structure design is consistent with those considered for Reptile in [2]. We test with varying channel number $\{32, 64, 128, 256\}$ in each convolution layer to show the robustness of our algorithms to meta-overfitting.

B.2 Datasets

There are three popular benchmark datasets used in our experiments.



(a) 5-way 1-shot tasks generated from Omniglot

(b) 5-way 1-shot tasks generated from MiniImageNet or TieredImageNet dataset

Fig. 1. Tasks used in our experiments. (a). Tasks generated from Omniglot. (b). Tasks generated from MiniImageNet or TieredImageNet dataset.

Omniglot The Omniglot dataset has 1623 characters from 50 alphabets. Each character contains 20 instances drawn by different individuals. The size of each image is 28×28. We randomly select 1200 characters for meta training and the rest are used for meta testing. Following [5], we also adopt a data augmentation strategy based on image rotation to enhance performance.

MiniImageNet The MiniImageNet dataset consists of 100 classes from the ImageNet dataset [1] and each class contains 600 images of size $84 \times 84 \times 3$. There are 64 classes used for training, 12 classes for validation and 24 classes for testing.

TieredImageNet The TieredImageNet dataset consists of 608 classes from the ILSVRC-12 dataset [4] and each image is scaled to $84 \times 84 \times 3$. There are 351 classes used for training, 97 classes for validation and 160 classes used for testing.

B.3 Detailed Experimental Settings

The experimental details of DSD-based Reptile and IHT-based Reptile can respectively be seen in Table 1 and Table 2. There are two points of hyperparameter settings that should be highlighted.

- The outer learning rate has an initial value 1.0 which will decay with iteration added.
- For MiniImageNet [6] with DSD-based Reptile, the iteration number of pruning phase for 32-channel case is 5×10^4 and for 64/128/256-channel case is 6×10^4. Correspondingly, the iteration number of retraining phase for 32-channel case is 2×10^4 and for 64/128/256-channel case is 10^4.
- For study of complex networks in Section 5.3, since our experiments are conducted on 4 RTX 2080Ti GPUs(11GB) while MetaOptNet is trained on 4 Titan X GPUs(12GB), we have to reduce the training shots from 15 to 10 in our experiments. For fair comparison, we rerun the baseline on the same model as in that paper with 10 training shots. We respectively set the number

Table 1. Detailed experimental settings for Omniglot, MiniImageNet, TieredImageNet datasets with DSD-based Reptile.

Hyperparameters	**Omniglot**	**MiniImageNet**	**TieredImageNet**
classes	5	5	5
shot	1 or 5	1 or 5	1 or 5
inner batch	10	10	6
inner iterations	5	8	8
outer learning rate	1	1	1
meta batch	5	5	5
meta iterations	10^4	10^4	10^4
evaluation batch	5	5	5
evaluation iterations	50	50	50
inner learning rate	0.001	0.001	0.001
pre-train iterations	3×10^4	3×10^4	3×10^4
pruning iterations(32c)	5×10^4	5×10^4	5×10^4
retrain iterations(32c)	2×10^4	2×10^4	2×10^4
pruning iterations(64/128/256c)	5×10^4	6×10^4	5×10^4
retrain iterations(64/128/256c)	2×10^4	10^4	2×10^4

of iterations of pre-training, pruning and retraining as 5 epochs, 20 epochs and 15 epochs. The learning rate is 0.1 in the first 30 epochs, 0.006 in next 5 epochs and 0.0012 in the final 5 epochs.

– For CAVIA, in DSD-based CAVIA case, the numbers of the iterations for pre-training, pruning and retraining phase are respectively 20K, 20K and 20K. In IHT-based CAVIA case, the iteration number of pre-training is 20K, and the iterative phase include 2 sparse-dense processes. Each sparse-dense process contains 20K iterations in which 16K iterations are for pruning fine-tuning and 4K iterations are for dense retraining.

C Additional Experimental Results

This appendix contains complete experimental results for Omniglot, MiniImageNet and TieredImageNet datasets. We performed our methods on 4-layer CNNs with varying channel number $\{32, 64, 128, 256\}$ as mentioned in Section B.

C.1 Results on Omniglot dataset

The baselines and all the results of Omniglot dataset are reported in Table 3. For each case, both DSD-based Reptile approach and IHT-based Reptile approach are evaluated on various pruning rates. The settings are the same as proposed in Section B.3.

For 32-channel case and 64-channel cases, which is less prone to be overfitting, both DSD-based Reptile approach and IHT-based Reptile approach tend to

Table 2. Detailed experimental settings for Omniglot, MiniImageNet, TieredImageNet datasets with IHT-based Reptile.

Hyperparameters	**Omniglot**	**MiniImageNet**	**TieredImageNet**
classes	5	5	5
shot	1 or 5	1 or 5	1 or 5
inner batch	10	10	6
inner iterations	5	8	8
outer learning rate	1	1	1
meta batch	5	5	5
meta iterations	10^4	10^4	10^4
evaluation batch	5	5	5
evaluation iterations	50	50	50
inner learning rate	0.001	0.001	0.001
epoch numbers	5	5	5
iteration numbers per interval	2×10^4	2×10^4	2×10^4
pruning iterations	1.5×10^4	1.5×10^4	1.5×10^4
retrain iterations	5×10^3	5×10^3	5×10^3

achieve comparable performance to baselines. When the channel size increases to 128 and 256, slightly improved performance can be observed. This is consistent with our analysis that overfiting is more likely to happen when channel number is relatively large and weight pruning helps alleviate such phenomenon to improve the generalization performance, which then leads to accuracy improvement with retraining operation.

C.2 Results on MiniImageNet dataset

In this section, we report the detailed results of experiments on MiniImageNet dataset.

From the table, it can be obviously observed that our method achieves remarkable performance consistently. For one thing, with the number of channels increasing, the accuracies of our methods keep being improved while the baselines perform oppositely. For example, in the 32-channel setting in which the model is less prone to overfit, when applying DSD-based Reptile with 10% and 40% pruning rate, the accuracy gain is 0.35% and 0.5% on 5-way 1-shot tasks and 1.02% and 1% on 5-way 5-shot tasks. In the 64-channel setting, DSD-based Reptile respectively achieves 0.83%, 0.83%, 0.88% improvements over 5-way 1-shot baseline and 1.75%, 1.77%, 1.18% improvements over 5-way 5-shot baseline with pruning rates 20%, 30%, 40%. Meanwhile our IHT-based Reptile approach respectively improves about 1.15%, 1.05%, 1.51% on 5-way 1-shot tasks and 0.62%, 1.32% and 1.95% on 5-way 5-shot tasks with pruning rates 10%, 20%, 40%. In the setting of 128-channel, all the cases of our method outperform the baseline remarkably, and the best accuracy of DSD-based Reptile on 5-way 1-shot tasks is nearly 3% higher than the baseline while on 5-way 5-shot tasks the gain is about 4.47%.

CAVIA [7] is also an effective approach to alleviate overfitting. In CAVIA, additional context parameters are introduced to be updated in task-specific phase while the network parameters are updated during outer loop. In our experiment, we also compare our method with CAVIA. As we can see in Table 4, our method can outperform CAVIA in all cases when the networks have the same number of channels.

C.3 Results on TieredImageNet dataset

In this section, we present the detailed results of experiments on TieredImageNet dataset in Table 5.

From the table, we can observe that our method achieves good performance on 5-way 1-shot classification tasks. For example, in 32-channel settings, the accuracy of DSD-based Reptile with 10% pruning rate is $\sim 0.5\%$ higher than baseline; in 64-channel settings, both DSD-based Reptile and IHT-based Reptile improve the performance evidently, respectively are 0.64% and 1.24%; and in 256-channel settings, the best performance achieves 0.44% improvement over the baseline.

However, in most 5-way 5-shot classification tasks, the performance of our method drops. We conjecture that the reason is TieredImageNet dataset, compared with MiniImageNet dataset, contains more classes.

Table 3. Few Shot Classification results on Omniglot dataset for 4-layer convolutional network with different channels on 5-way 1-shot and 5-way 5-shot tasks. The "±" shows 95% confidence intervals over tasks. The evaluation baselines are run by us.

Methods	**Backbone**	**Rate**	**5-way 1-shot**	**5-way 5-shot**
Reptile baseline	32-32-32-32	0%	96.63±0.17%	99.31±0.07%
	64-64-64-64	0%	97.68±0.10%	99.48±0.06%
	128-128-128-128	0%	97.99±0.11%	99.60±0.13%
	256-256-256-256	0%	98.05±0.13%	99.65±0.06%
DSD-based Reptile	32-32-32-32	10%	96.42±0.17%	**99.38±0.07%**
		20%	95.98±0.18%	99.33±0.07%
		30%	96.22±0.17%	99.23±0.08%
		40%	96.53±0.17%	99.37±0.07%
	64-64-64-64	10%	**97.64±0.02%**	**99.50±0.05%**
		20%	97.60±0.07%	99.49±0.04%
		30%	97.47±0.05%	99.49±0.05%
		40%	97.43±0.01%	99.45±0.03%
	128-128-128-128	10%	**98.04±0.10%**	99.61±0.10%
		20%	97.99±0.10%	99.62±0.12%
		30%	97.96±0.12%	**99.63±0.12%**
		40%	97.99±0.10%	99.61±0.10%
	256-256-256-256	10%	**98.12±0.12%**	**99.68±0.05%**
		20%	98.02±0.13%	99.66±0.05%
		30%	97.96±0.13%	99.67±0.05%
		40%	97.99±0.10%	99.63±0.06%
IHT-based Reptile	32-32-32-32	10%	**96.65±0.16%**	99.49±0.06%
		20%	96.54±0.17%	**99.57±0.06%**
		30%	96.45±0.17%	99.52±0.06%
		40%	96.21±0.18%	99.48±0.07%
	64-64-64-64	10%	97.63±0.14%	99.49±0.06%
		20%	97.60±0.13%	**99.57±0.06%**
		30%	**97.77±0.15%**	99.52±0.06%
		40%	97.51±0.1%	99.48±0.07%
	128-128-128-128	10%	98.12±0.12%	99.63±0.06%
		20%	**98.22±0.12%**	99.64±0.05%
		30%	98.01±0.13%	**99.65±0.05%**
		40%	98.06±0.12%	99.63±0.06%
	256-256-256-256	10%	**98.16±0.12%**	99.66±0.05%
		20%	98.08±0.13%	**99.69±0.05%**
		30%	98.05±0.13%	99.64±0.05%
		40%	97.90±0.13%	99.65±0.05%

Table 4. Few Shot Classification results on MiniImageNet dataset for 4-layer convolutional network with different channels on 5 way setting. The "±" shows 95% confidence intervals over tasks. The evaluation baselines are run by us.

Methods	**Backbone**	**Rate**	**5-way 1-shot**	**5-way 5-shot**
Reptile baseline	32-32-32-32	0%	50.30±0.40%	64.27±0.44%
	64-64-64-64	0%	51.08±0.44%	65.46±0.43%
	128-128-128-128	0%	49.96±0.45%	64.40±0.43%
	256-256-256-256	0%	48.60±0.44%	63.24±0.43%
CAVIA baseline	32-32-32-32	0%	47.24±0.65%	59.05±0.54%
	128-128-128-128	0%	49.84±0.68%	64.63±0.54%
	512-512-512-512	0%	51.82±0.65%	65.85±0.55%
DSD-based Reptile	32-32-32-32	10%	**50.65±0.45%**	**65.29±0.44%**
		20%	49.94±0.43%	64.65±0.43%
		30%	50.18±0.43%	**65.78±0.41%**
		40%	**50.83±0.45%**	**65.24±0.44%**
	64-64-64-64	10%	51.12±0.45%	65.80±0.44%
		20%	**51.91±0.45%**	**67.21±0.43%**
		30%	**51.91±0.45%**	**67.23±0.43%**
		40%	**51.96±0.45%**	**67.17±0.43%**
	128-128-128-128	30%	51.98±0.45%	68.16±0.43%
		40%	52.15±0.45%	68.19±0.43%
		50%	52.08±0.45%	**68.87±0.42%**
		60%	**52.27±0.45%**	68.44±0.42%
	256-256-256-256	60%	**53.00±0.45%**	**68.04±0.42%**
IHT-based Reptile	32-32-32-32	10%	**50.45±0.45%**	63.91±0.46%
		20%	50.26±0.47%	63.63±0.45%
		30%	50.21±0.44%	**65.05±0.45%**
		40%	49.74±0.46%	64.15±0.45%
	64-64-64-64	10%	**52.23±0.45%**	**66.08±0.43%**
		20%	**52.13±0.46%**	**66.78±0.43%**
		30%	51.98±0.45%	66.14±0.43%
		40%	**52.59±0.45%**	**67.41±0.43%**
	128-128-128-128	30%	51.64±0.45%	67.05±0.43%
		40%	52.73±0.45%	**68.69±0.42%**
		50%	52.76±0.45%	67.63±0.43%
		60%	**52.95±0.45%**	68.04±0.42%
	256-256-256-256	60%	**49.85±0.44%**	**66.56±0.42%**

Table 5. Few Shot Classification results on TieredImageNet dataset for 4-layer convolutional network with different channels on 5 way setting. The "±" shows 95% confidence intervals over tasks. The evaluation baselines are run by us.

Methods	**Backbone**	**Rate**	**5-way 1-shot**	**5-way 5-shot**
Reptile baseline	32-32-32-32	0%	50.52±0.45%	64.63±0.44%
	64-64-64-64	0%	51.98±0.45%	67.70±0.43%
	128-128-128-128	0%	53.30±0.45%	69.29±0.42%
	256-256-256-256	0%	54.62±0.45%	68.06±0.42%
DSD-based Reptile	32-32-32-32	10%	**50.94±0.46**%	64.65±0.44%
		20%	49.85±0.46%	63.72±0.44%
	64-64-64-64	10%	**52.62±0.46**%	66.69±0.43%
		20%	51.95±0.45%	66.05±0.43%
	128-128-128-128	10%	53.39±0.46%	67.22±0.43%
		20%	52.61±0.46%	66.39±0.43%
	256-256-256-256	10%	54.55±0.45%	**68.60±0.43**%
		20%	**54.98±0.45**%	67.98±0.43%
IHT-based Reptile	32-32-32-32	10%	**50.58±0.46**%	63.09±0.45%
		20%	50.19±0.46%	63.42±0.44%
	64-64-64-64	10%	51.75±0.45%	65.20±0.44%
		20%	**53.22±0.46**%	66.15±0.44%
	128-128-128-128	10%	**53.48±0.45**%	69.36±0.42%
		20%	52.98±0.45%	66.22±0.43%
	256-256-256-256	10%	**55.06±0.45**%	67.60±0.43%
		20%	54.38±0.45%	**69.36±0.42**%

References

1. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems. pp. 1097–1105 (2012)
2. Nichol, A., Achiam, J., Schulman, J.: On first-order meta-learning algorithms. arXiv preprint arXiv:1803.02999 (2018)
3. Rigollet, P.: 18. s997: High dimensional statistics. Lecture Notes, Cambridge, MA, USA: MIT Open-CourseWare (2015)
4. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**(3), 211–252 (2015)
5. Santoro, A., Bartunov, S., Botvinick, M., Wierstra, D., Lillicrap, T.: Meta-learning with memory-augmented neural networks. In: International Conference on Machine Learning. pp. 1842–1850 (2016)
6. Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al.: Matching networks for one shot learning. In: Advances in Neural Information Processing Systems. pp. 3630–3638 (2016)
7. Zintgraf, L., Shiarli, K., Kurin, V., Hofmann, K., Whiteson, S.: Fast context adaptation via meta-learning. In: International Conference on Machine Learning. pp. 7693–7702 (2019)