A Novel Few-Shot Learning with Meta-Gradient Memory

Lin Hui[1]*, Yi-Cheng Chen[2], Huang-Wen Huang[3], Pin-Chen Tseng[4]

[1,3,4] Department of Computer Science and Information Engineering, Tamkang University, Taiwan
[2]Department of Information Management, National Central University, Taiwan
[1] amar0627@gmail.com, [2]ycchen@mgt.ncu.edu.tw, [3]hhw402@mail.tku.edu.tw
[4]610420019@gms.tku.edu.tw

Abstract. The concept of few-shot learning allows users to train models with a limited amount of data while still achieving a high level of generalization. Impressive techniques like meta-learning and continual learning models have shown remarkable performance in model development. However, there are still two crucial challenges to overcome: unstable performance and catastrophic forgetting, especially when dealing with new tasks while retaining knowledge of previous tasks. To tackle these issues, a new approach called Enhanced Model-Agnostic Meta-learning (EN-MAML) has been proposed. It combines the adaptable adaptation characteristics of meta-learning with the stable performance of continual learning. By employing this method, users can efficiently and effectively train their models even with limited data, ensuring stability throughout the process. Experimental results demonstrate that EN-MAML outperforms other state-of-the-art models across multiple real datasets, achieving superior accuracy, exhibiting more stable performance, and converging faster.

Keywords: Machine Learning, Deep Learning, Meta-learning, Continual Learning

1 Introduction

Model-agnostic meta-learning (MAML) [1] is noteworthy not only for its simplicity but also for its effectiveness. On the other hand, continual learning excels in few-shot learning tasks by learning new tasks without forgetting previous ones. Gradient episodic memory for continual learning (GEM) [2] addresses the issue of catastrophic forgetting through quadratic programming and even achieves positive forward transfer (FWT), where the model learns new tasks better by leveraging previous task knowledge, as well as positive backward transfer (BWT), where the learning of current tasks benefits previous ones.

Figure 1 illustrates the training accuracy of MAML with three different seeds, revealing its unstable performance during training. To tackle this instability problem identified by Antoniou et al. [3], we combine MAML with GEM. By employing a GEM quadratic program originally used to prevent catastrophic forgetting in Lopez-

Paz and Ranzato [2], we store meta-gradients to adjust the updating direction, facilitating the learning of previous episodes. Consequently, our method, EN-MAML, aims to achieve effects similar to FWT and BWT while accelerating model convergence.

Furthermore, we address the stability-plasticity dilemma [4] and reconsider the features of meta-learning and continual learning in the context of few-shot learning. As a result, we design our approach to effectively combine the learning processes for current tasks and previous tasks. This allows our model to possess both adaptability to new tasks and improved stability.

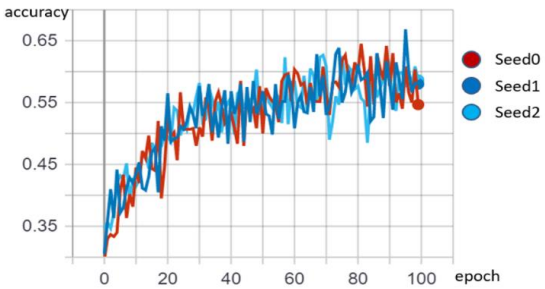


Fig. 1. Unstable performance of MAML.

2 Related Work

Recent studies based on this concept have become more complicated and delicate in the embedding process [8, 14] and even utilize data-dependent initializations to adapt well in a low-dimensional latent space. By designing an object detection network with a weight generator based on an attention mechanism, the method of Gidaris and Komodakis [5] also uses a representation space to acquire different task knowledge.

The method proposed in Finn et al. [1] can be applied to different kinds of model structures to promote their generalization ability. To explore a set of appropriate initial parameters, the approach proposed in Nichol and Schulman [6] reduces the cost of calculating in the differentiating process. Moreover, Antoniou et al. [3], presents various modifications of Finn et al. [1] and analyzes the framework of MAML. For few-shot image classification, Chen et al. [7] proposed a meta-learning system to achieve time and resource efficiency and to generalize unknown feedback datasets. Kuo et al. [8] alleviated catastrophic forgetting, prevented base learners from inducing overfitting, and achieved strong robustness.

It is difficult to train a model to generalize well with little data, incremental learning aims to gradually learn via continuous training. Incremental learning is divided into task-based incremental learning [9] and class-based incremental learning [10]. Hu et al. [10] found that data replay is a reliable technology. Using the causal effect of introducing old data in an end-to-end manner, old data can be stored in a

CIL network to prevent forgetting without actually storing them.

Many continual learning approaches use extra memory to store data for the purpose of alleviating catastrophic forgetting [11]. In addition to storing data to ensure that networks remember these previous tasks, there is a network designed to generate data to review the knowledge that has been learned previously [12]. Rather than determining the important parameters, Lopez-Paz and Ranzato [2] focused on modifying the angle of the model's gradient and even proposed the metrics of *forward transfer* and *backward transfer* to evaluate the performance of continual learning approaches.

From our observation, and as noted in Riemer et al. [9], the classic stability-plasticity dilemma [4] concept seems to match the characteristics of meta-learning and continual learning. The main concept of Riemer et al. [9] is easing the interference between transfer and retention with gradient alignment, which was proposed in Lopez-Paz and Ranzato [2]. The stability-plasticity dilemma mentioned in Abraham and Robins [4] means that there is a regulated balance between synaptic stability and synaptic plasticity. Meta-learning presents great adaptation to nonstationary task distribution. However, the problems of training instability and overfitting occurred in Finn et al. [1]. Most continual learning approaches enhance network stability, such as by adding more constraints when updating parameters [2] [18] or using a buffer to store data [11]. The features of these approaches allow continual learning to correspond to the stability of the stability-plasticity dilemma. To maintain the information of previous tasks, Lopez-Paz and Ranzato [2] compares the gradients of different tasks to confirm that updating the direction will not lead to serious forgetting.

Several approaches have crossed the border between meta-learning and continual learning and leveraged the advantages of each to overcome their shortcomings. Gai et al. [13] uses meta continual learning to mitigate forgetting with gradient episodic memory. De Lange et al. [14] compared 11 state-of-the-art continual learning methods and 4 baselines. Riemer et al. [9] combines meta-learning with GEM [2] so that networks become generable based on past and future task distributions. Our approach focuses on addressing the problem of MAML pointed out by Antoniou et al. [3] with GEM. In our work, we migrate the GEM quadratic program into the MAML framework to make MAML more stable and to fit it with other gradient-based meta-learning approaches to enhance their performance.

3 Proposed Model : EN-MAML

From Fig. 2, the entire EN-MAML framework can be segmented into two parts. On the left side, EN-MAML produces fast weights to adapt to a new batch of tasks and produces a meta-gradient for the current batch. We see this process as "learning" because EN-MAML acquires novel knowledge from new tasks that consist of unseen categories of images. When EN-MAML completes the learning process, it produces the meta-gradient according to the loss from the current batch. On the left side, EN-

MAML computes the batch of tasks stored in the meta-gradient buffer to generate the meta-gradient for the previous batch. We see this process as "reviewing" because the model performs previous tasks again with its current parameter state. In the next step, the meta-gradient from the current batch will be modified by the process of continual learning, which integrates the gradient from the previous batch in computation to migrate the knowledge the model learned previously. Finally, EN-MAML can reduce the conflict updating caused by the nonstationary environment and update its parameters in a more stable way.

In the original MAML, the outer-loop updating generates the gradient, which comes from the loss of an entire batch. These gradients contain information about cross-task knowledge, which is the key to allowing networks to acquire the ability to continue promoting adaptation to different tasks. In our work, we called this kind of gradient a "meta-gradient", and it has a significant impact on the directing network in exploring more adaptive initialization parameters.

Therefore, networks trained on few-shot learning usually have to face difficulty in dealing with conflicts between gradients' directions. Accordingly, we assume that this issue causes the problem of MAML training instability. In our observation of other learning methods for handling this issue, we found that GEM, an approach proposed for continual learning, focuses on adjusting the updating gradient angle to make the network learn the new task without forgetting previous tasks, and [9] also proved its effectiveness. Conflicts between gradients will occur in the following situation:

$$\frac{\partial L(f_\theta(x_i),y_i)}{\partial \theta} \cdot \frac{\partial L(f_\theta(x_j),y_j)}{\partial \theta} < 0, \quad (1)$$

$(X_i,\ y_i)$ and $(X_j,\ y_j)$ are different sampled data points from different tasks. When the inner product of the gradients is negative, it means that the network loses knowledge of the previous task if the parameters are updated for the current task. To avoid forgetting, GEM updates the parameter only if the following constraint is satisfied:

$$\frac{\partial L(f_\theta(x),y)}{\partial \theta} \cdot \frac{\partial L(f_\theta(x_M),y_M)}{\partial \theta} \geq 0, \quad (2)$$

M is the buffer used to store the data from the observed task. $x_M,\ y_M$ indicate the images and labels stored in *M*. GEM uses a quadratic program to modify the updating gradients that originally violated this constraint. Most recent approaches utilize task memory buffers to store task-level data or gradients. However, we pay attention to the meta-gradient produced after the network learns all batch tasks. In other words, the information of the meta-gradient is at the batch level, which contains more varied and general task knowledge, and this property could be more likely to make MAML avoid overfitting. This is the reason why we store the meta-gradient in the buffer instead of the task-level gradient. In addition, the buffer replaces the oldest meta-gradient with the newest one. Thus, the network can prevent overfitting on certain tasks and can learn from the distribution.

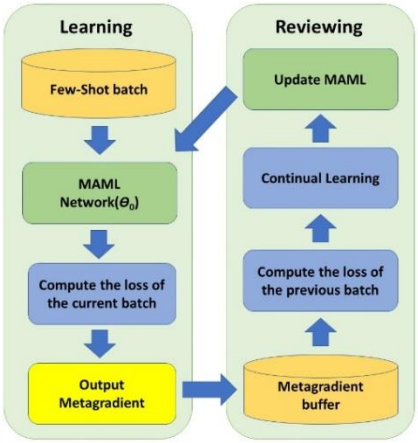


Fig. 2. The architecture of EN-MAML.

We assume that there are *n* batches of tasks sampled from task distribution $p(T)$ in a training epoch. EN-MAML learns a new batch by performing inner updates and computes the meta-gradient of the current batch g_c. The current meta-gradient is modified by using (5) and the GEM quadratic program to compare it with other meta-gradients from the previous batch. After the gradient tuning process, we acquire the modified meta-gradient g_c' that EN-MAML applies to outer-loop updates.

To enhance training stability, EN-MAML calculates the loss not only from the tasks of the current batch but also from the tasks of the previous batch stored in the buffer. With the loss from learning and reviewing, we design EN-MAML to automatically decide how important the parts are, so there are trainable weights before the two losses. Therefore, EN-MAML can balance the stability-plasticity dilemma in different learning environments and training stages because it can adjust the attention that it gives to learning and reviewing. We use the cross-entropy loss function to calculate the loss of image classification, which is expressed by (6). The loss function of EN-MAML is expressed by (7).

$$l_c\big(f_\theta(x,y)\big) = \sum_{x,y\sim T} y\, log f_\theta(x), \tag{3}$$

$$L_{total} = w_c \sum_{t=1}^{T} l_c\big(f_\theta(x,y)\big) + w_p \sum_{t=1}^{T} l_p(f_\theta(x_M, y_M)), \tag{4}$$

w_c is the weight used to represent how important EN-MAML considers the current batch of tasks to be, and w_p is the weight that represents how much attention EN-MAML gives to reviewing the previous batch of tasks. l_c is the loss from a task in the current batch, and l_p is the loss from a task in the previous batch. Mem is the meta-gradient memory buffer, where we store previous data to compute the previous meta-gradient.

4 Performance Evaluation

During our experimental evaluations, we utilize the well-established benchmarks in the field of few-shot learning: Omniglot [15] and Mini-ImageNet [16]. The Omniglot dataset comprises 1,623 handwritten characters, classified into 50 distinct letters. Each letter category contains 20 instances of handwritten symbols. In our experiments using Torchmeta, we divide the Omniglot dataset into three subsets: a training set with 1,028 classes, a validation set with 172 classes, and a testing set with 423 classes. Traditionally, most few-shot learning methods utilize the first 1,200 classes from the Omniglot dataset for training [3]. However, it has been acknowledged in previous research that preserving a few classes for validation purposes is crucial [3]. Therefore, we also allocate a portion of the dataset for validation to conduct our experiments effectively. Moving on to the Mini-ImageNet dataset in Torchmeta, each class consists of 600 instances, and the dataset encompasses 64 classes for training, 16 classes for validation, and 20 classes for testing. To enhance the datasets, we apply augmentation techniques such as rotating the images by 90 degrees and resizing them to 28 × 28 for Omniglot and 84 × 84 for Mini-ImageNet.

4.1 Performance Comparison

We compare the performance of different few-shot learning models under N-way K-shot experiments, which means that a task has images from N kinds of classes and that each class has K examples. As well as MAML, we demonstrate the performance of EN-MAML with other famous few-shot learning models proposed in recent years: *Siamese Nets* [17], *Matching Nets* [16], *Neural Statistician*: [18], *Memory Mod*: [19]. *Meta networks*: [20], and *Reptile*: Reptile [6].

Table 1. Accuracy of Omniglot for 5-way classification

Omniglot 5-way few-shot classification		
Model	**Accuracy**	
	1-SHOT	5-SHOT
Siamese Nets	97.3%	98.4%
Matching Nets	98.1%	98.9%
Neural Statistician	98.1%	99.5%
Memory Mod.	98.4%	99.6%
MAML	98.25%	98.85%
Reptile	95.30%	98.80%
EN-MAML	**98.77%**	**99.67%**

Evaluate EN-MAML by performing N-way K-shot experiments on the Omniglot and Mini-ImageNet datasets. First, the results of 5-way few-shot classification on Omniglot show that EN-MAML reaches state-of-the-art performance and improves accuracy compared to MAML, as shown in Table 1. Compared to the performance of MAML, EN-MAML improves the accuracy by approximately 0.52%, as shown in Table 1.

For the Mini-ImageNet datasets, EN-MAML also demonstrated dramatically higher performance on 5-way classification experiments, as shown in Table 2. EN-MAML improves the accuracy by approximately 5.17% compared to the MAML performance from our replication in terms of accuracy in the Mini-ImageNet 5-way 1-shot setting. For the Mini-ImageNet 5-way 5-shot setting, EN-MAML is approximately 5.45% more accurate than MAML.

Table 2. Accuracy of Mini-ImageNet 5-way classification

Mini-ImageNet 5-way few-shot classification		
Model	**Accuracy**	
	1-SHOT	5-SHOT
Siamese Nets	47.8%	63.66%
Matching Nets	43.56%	55.31%
Neural Statistician	48.60%	63.09%
Memory Mod.	49.21%	65.42%
MAML	49.38%	66.55%
Reptile	46.81%	62.37%
EN-MAML	**54.55%**	**72%**

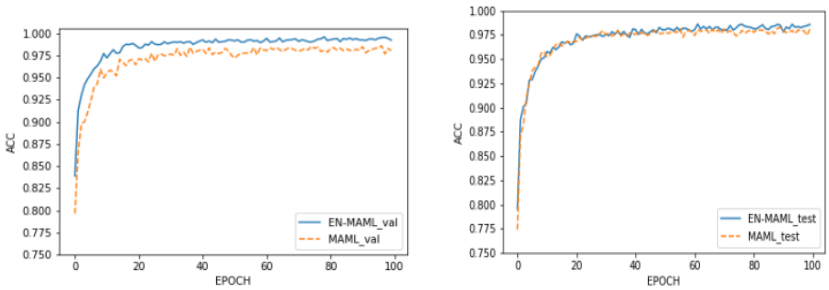
4.2 Stability and Accuracy Comparison with MAML

To fully compare and analyze the performance of EN-MAML and MAML, we demonstrate how the models' testing performance improves as the number of epochs increases. We show all performance curves from the experiments mentioned in the above sections. First, we perform 5-way and 20-way classification, both with 1 shot and 5 shots in the Omniglot dataset. Additionally, we perform 5-way classification with 1 shot and 5 shots in Mini-ImageNet. Moreover, we reproduce MAML with the above experimental protocol setting. Second, we perform model training stability experiments to examine whether our method alleviates the unstable training problem proposed in Antoniou et al. [3].

We can observe that the testing accuracy of EN-MAML starts to surpass that of MAML when the model has been trained for approximately 40 epochs, as shown in

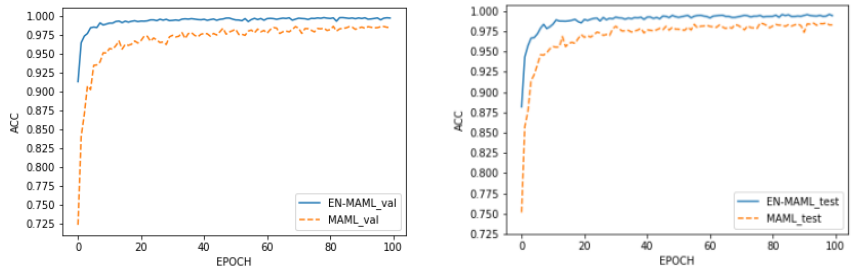
Fig. 3(a). Our method provides more stable validation accuracy, which is one of our method's objectives. In Fig. 3(b), we can see that EN-MAML maintains higher validation accuracy at all times. Therefore, the combination of meta-learning and continual learning is actually positive in terms of enhancing the stability of MAML. Fig. 4 shows that EN-MAML can not only improve the accuracy of the original MAML but also enhance the training stability. EN-MAML obtains higher accuracy from earlier epochs to the end of the testing experiment in Fig. 4(a), and this result can also be observed in the validation experiment in Fig. 4(b).

For the Mini-ImageNet experiments, we can observe that the performances of both EN-MAML and MAML become more unstable than in the tests on the Omniglot dataset. Both testing accuracy and validation accuracy fluctuate dramatically because the difficulty of the dataset and the few-shot setting makes the models unable to capture general features easily.



(a) **(b)**

Fig. 3. Comparison of EN-MAML and MAML in the 5-way 1-shot setting on the Omniglot dataset: (a) Validation accuracy; (b) Testing accuracy.

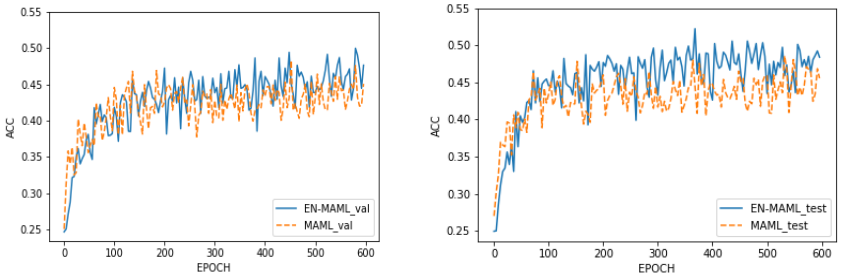


(a) **(b)**

Fig. 4. Comparison of EN-MAML and MAML in the 5-way 5-shot setting on the Omniglot dataset: (a) Validation accuracy; (b) Testing accuracy.

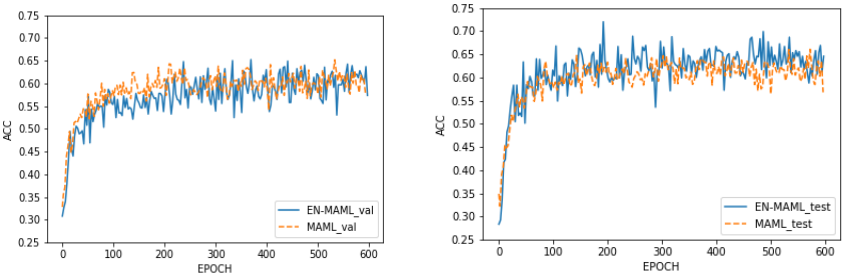
However, EN-MAML still reaches the highest accuracy in the 5-way 1-shot setting in Fig. 5(a) and maintains an equivalent level of validation accuracy in Fig5(b). Additionally, EN-MAML outperforms MAML most of the time in the 5-way 5-shot setting on Mini-ImageNet in Fig. 6(a). EN-MAML starts to surpass it and obtains higher accuracy in the middle epochs. In contrast, MAML shows more stable

performance in validation accuracy in this setting. We analyzed the results, and we will discuss this phenomenon in the next section. To summarize all the experimental results on Omniglot and Mini-ImageNet, our observation is that EN-MAML either improves the testing accuracy or promotes validation accuracy. EN-MAML progresses on at least one metric and keeps the other metric at an equivalent level. On the Omniglot dataset, EN-MAML demonstrates dramatic improvement in validation accuracy. In contrast, EN-MAML shows greater enhancement in testing accuracy in all Mini-ImageNet experiments.



(a) **(b)**

Fig. 5. Comparison of EN-MAML and MAML in the 5-way 1-shot setting on the Mini-ImageNet dataset: (a) Validation accuracy; (b) Testing accuracy.



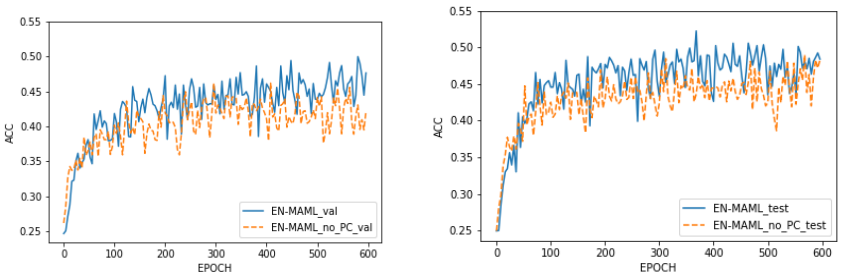
(a) **(b)**

Fig. 6. Comparison of EN-MAML and MAML in the 5-way 5-shot setting on the Mini-ImageNet dataset: (a) Validation accuracy; (b) Testing accuracy.

4.3 The Effectiveness of Combining Meta-learning with Continual Learning

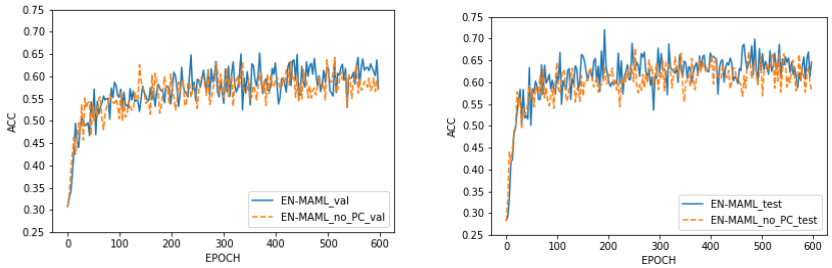
As shown in Figs. 3 and 4, EN-MAML can truly improve the stability of the validation accuracy, which means a more reliable and stable training process in all of the Omniglot experimental settings. However, the positive effect of combining meta-

learning with continual learning is not only stability promotion but also enhancement of the model in terms of reaching higher testing accuracy, which is shown more clearly in Figs. 4, 5 and 6. From our experimental observation, the modified meta-gradient, which is generated from quadratic programming to maintain the meta-gradient information from previous batches, can have approximately the same effect as the FWT proposed in Lopez-Paz and Ranzato [2].



(a) **(b)**

Fig. 7. Comparison of EN-MAML and EN-MAML without the previous-current vector in the 5-way 1-shot setting on the Mini-ImageNet dataset: (a) Validation accuracy; (b) Testing accuracy.



(**a**) (**b**)

Fig. 8. Comparison of EN-MAML and EN-MAML without the previous-current vector in the 5-way 5-shot setting on the Mini-ImageNet dataset: (a) Validation accuracy; (b) Testing accuracy.

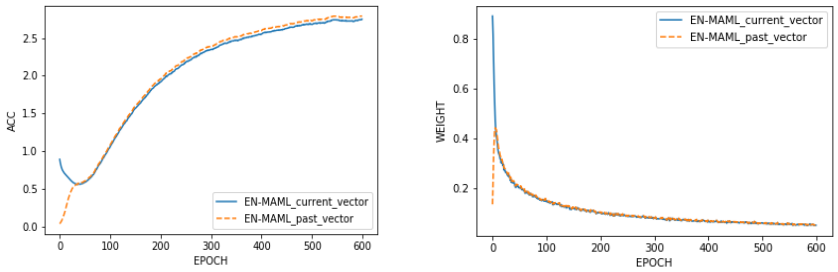
From Figs. 7 and 8, we can observe that there is an obvious performance gap between the original EN-MAML and EN-MAML without considering dynamic weights. Particularly in the 5-way 1-shot setting, the original EN-MAML can improve both its testing accuracy and validation accuracy with fewer data provided. With the mechanism of weighted current knowledge and weighted knowledge, we find a possible solution to overcome the stability-plasticity dilemma.

In our experiment, we set the initial values of the importance of current tasks and previous tasks to 0.9 and 0.1, respectively. As the epoch increases, we find that the network gradually pays more attention to both the current task and previous tasks

under the 5-way 1-shot setting on Mini-ImageNet in Fig. 9(a). In contrast, the network pays less attention to both the current task and previous tasks under the 5-way 5-shot setting on Mini-ImageNet in Fig. 9(b).

EN-MAML can increase the attention to previous and current knowledge to overcome the limitation of the few training data. And EN-MAML can determine the current learning problem that is the most influential in the experimental setting and dynamically adjust the importance of different kinds of learning knowledge.

In addition, different classes in our experiment can be sampled repeatedly, so a gradually better-trained EN-MAML can learn the seen classes better after it acquires metalevel knowledge from other batches of tasks. We take this effect to be nearly the same as that of the BWT proposed in Lopez-Paz and Ranzato [2] EN-MAML absorbs new metalevel knowledge with the MAML framework, digests new information with the FWT effect, and then acquires a better understanding of previously learned knowledge. This is the reason why EN-MAML can concurrently promote stability and testing accuracy, which is demonstrated in most of our experimental settings.



(a) **(b)**

Fig. 9. The weight change in the current-past vector: (a) Current and past vector change under the 5-way 1-shot setting on Mini-ImageNet; (b) Current and past vector change under the 5-way 5-shot setting on Mini-ImageNet.

Notably, even under Mini-ImageNet, a more difficult dataset, and a smaller batch size setting, which means the meta-gradient will be generated from fewer tasks, EN-MAML can outperform MAML in most of the experiments in Figs. 5 and 6. We also observe that EN-MAML still shows FWT and BWT effects, even under a harder learning environment. As the epoch grows, higher performance also appears more frequently. As we mentioned in the above paragraphs, EN-MAML needs time to accumulate powerful meta-gradient memory, and the phenomenon illustrated in Figs. 5 and 6 is demonstrated more clearly. In these figures, we find that EN-MAML outperforms MAML more dramatically and reaches the highest accuracy in later epochs.

5 Conclusions

Presenting a groundbreaking technique called EN-MAML, our approach merges meta-learning and continual learning by utilizing the meta-gradient property alongside quadratic programming. This innovative method offers enhanced stability during model training, outperforming MAML in terms of testing accuracy across various experimental scenarios. Our experimental results highlight the potential of combining meta-learning and continual learning to achieve simultaneous improvements in flexibility and stability. Looking ahead, the field of few-shot learning can further investigate additional strategies to leverage the unique characteristics of meta-learning and continual learning, effectively addressing the stability-plasticity dilemma and pushing the boundaries of research in this domain.

References

[1] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proceedings of the 34th International Conference on Machine Learning*, Sydney, Australia, 2017, pp. 1126-1135.

[2] D. Lopez-Paz and M. A. Ranzato, "Gradient episodic memory for continual learning," 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 2017.

[3] A. Antoniou, H. Edwards, and A. Storkey, "How to train your MAML," *arXiv preprint arXiv:1810.09502*, pp. 1-11, 2019.

[4] W. C. Abraham and A. Robins, "Memory retention--the synaptic stability versus plasticity dilemma," *Trends in Neuroscience*, vol. 28, no. 2, pp. 73-78, 2005.

[5] S. Gidaris and N. Komodakis, "Dynamic few-shot visual learning without forgetting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, 2018, pp. 4367-4375.

[6] A. Nichol and J. Schulman, "Reptile: a scalable metalearning algorithm," *arXiv preprint arXiv:1803.02999*, 2018.

[7] Y. Chen, C. Guan, Z. Wei, X. Wang, and W. Zhu, "MetaDelta: a meta-learning system for few-shot image classification," *arXiv preprint arXiv:2102.10744*, 2021.

[8] N. I. Kuo, M. Harandi, N. Fourrier, C. Walder, G. Ferraro, and H. Suominen, "Learning to continually learn rapidly from few and noisy data," *arXiv preprint arXiv:2103.04066*, 2021.

[9] M. Riemer, I. Cases, R. Ajemian, M. Liu, I. Rish, Y. Tu, and G. Tesauro, "Learning to learn without forgetting by maximizing transfer and minimizing interference," *arXiv preprint arXiv:1810.11910*, 2019.

[10] X. Hu, K. Tang, C. Miao, X. S. Hua, and H. Zhang, "Distilling causal effect of data in class-incremental learning," in *Proceedings of the IEEE/CVF Conference*

on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, 2021, pp. 3957-3966.

[11] S. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert, "iCaRL: incremental classifier and representation learning," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, 2017, pp. 5533-5542.

[12] K. Shaheen, M.A. Hanif, O. Hasan et al. “Continual Learning for Real-World Autonomous Systems: Algorithms, Challenges and Frameworks”. J Intell Robot Syst 105, 9 (2022).

[13] S. Gai, Z. Chen and D. Wang, "Multi-Modal Meta Continual Learning,"*2021 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2021, pp. 1-8.

[14] M. De Lange, R. Aljundi, M. Masana, S. Parisot, X. Jia, A. Leonardis, G. Slabaugh,, T. Tuytelaars,, “A Continual Learning Survey: Defying Forgetting in Classification Tasks,” IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2022, pp. 3366-3385.

[15] B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum, "Human-level concept learning through probabilistic program induction," *Science*, vol. 350, no. 6266, pp. 1332-1338, 2015.

[16] O. Vinyals, C. Blundell, T. Lillicrap, and D. Wierstra, "Matching networks for one shot learning," *Advances in Neural Information Processing Systems*, vol. 29, pp. 3630-3638, 2016.

[17] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *Proceedings of the 32 nd International Conference on Machine Learning*, Lille, France, 2015, p. 8.

[18] H. Edwards and A. Storkey, "Towards a neural statistician," in *5th International Conference on Learning Representations*, Toulon, France, 2017, pp. 1-13.

[19] Ł. Kaiser, O. Nachum, A. Roy, and S. Bengio, "Learning to remember rare events," *arXiv preprint arXiv:1703.03129*, 2017.

[20] T. Munkhdalai and H. Yu, "Meta networks," in *Proceedings of the 34th International Conference on Machine Learning*, Sydney, Australia, 2017, pp. 2554-2563.