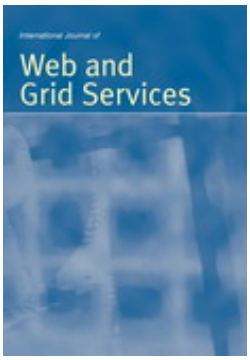
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Web and Grid Services

ISSN online: 1741-1114 - ISSN print: 1741-1106
https://www.inderscience.com/ijwgs

Effectively learn how to learn: a novel few-shot learning with meta-gradient memory

Lin Hui, Yi-Cheng Chen

DOI: 10.1504/IJWGS.2024.10060211

Article History:

Received:	13 April 2023
Last revised:	05 July 2023
Accepted:	23 September 2023
Published online:	25 March 2024

Effectively learn how to learn: a novel few-shot learning with meta-gradient memory

Lin Hui*

Department of Computer Science and Information Engineering,
Tamkang University, Taiwan
Email: 121678@mail.tku.edu.tw
*Corresponding author

Yi-Cheng Chen

Department of Information Management,
National Central University, Taiwan
Email: ycchen@mgt.ncu.edu.tw

Abstract: Recently, the importance of few-shot learning has tremendously grown due to its widespread applicability. Via few-shot learning, users can train their models with few data and maintain high generalisation ability. Meta-learning and continual learning models have demonstrated elegant performance in model development. However, unstable performance and catastrophic forgetting are still two fatal issues with regard to retaining the memory of knowledge about previous tasks when facing new tasks. In this paper, a novel method, enhanced model-agnostic meta-learning (EN-MAML), is proposed for blending the flexible adaptation characteristics of meta-learning and the stable performance of continual learning to tackle the above problems. Based on the proposed learning method, users can efficiently and effectively train the model in a stable manner with few data. Experiments show that when following the N-way K-shot experimental protocol, EN-MAML has higher accuracy, more stable performance and faster convergence than other state-of-the-art models on several real datasets.

Keywords: machine learning; deep learning; meta-learning; continual learning.

Reference to this paper should be made as follows: Hui, L. and Chen, Y-C. (2024) 'Effectively learn how to learn: a novel few-shot learning with meta-gradient memory', *Int. J. Web and Grid Services*, Vol. 20, No. 1, pp.3–24.

Biographical notes: Lin Hui received her PhD in Computer Science and Information Engineering from Tamkang University (TKU), Taiwan in 2006. She is currently an Associate Professor with the Department of Computer Science and Information Engineering, Tamkang University (TKU). Her research interests include operation research, data mining, machine learning, multimedia applications, and mobile information systems. She has published some journal articles, book chapters, and conference papers related to these research fields. She had served as journal guest editor/reviewer, and program co-chair/chair for many international conferences and workshops.

Yi-Cheng Chen received his PhD degree in the Department of Computer Science at National Chiao Tung University (NCTU), Taiwan in 2012. Currently, he is an Associate Professor in the Department of Information Management in National Central University (NCU), Taiwan. He has been

active in international academic activities, as conference organiser, journal editor/reviewer. He published some papers in several prestigious conferences and journal and also had the best paper awards in several conferences. His research interests include machine learning, data mining, social network analysis and cloud computing.

1 Introduction

Although modern deep neural networks present outstanding performance in different applications and domains, such as social network, computer vision, speech, Industrial 4.0 and natural language processing, to name a few. However, most of them require a large amount of data, a long training time and many computing resources to achieve a state-of-the-art level. For example, Om et al. (2020) show both LSTM and RNN can achieve high accuracy for large datasets in social network derived from e-mail. This also means that current deep neural networks have difficulty learning a new concept quickly with only a few available samples. In contrast, human intelligence is able to recognise new objects or learn new tasks well with only a small amount of data and practice to become familiar with novel concepts.

Neural networks usually do not perform as well as expected in training networks with very few samples, not to mention generalisation to novel tasks. Transfer learning (Yosinski et al., 2014), which takes models that were previously trained on one or more tasks and uses them as starting points in training a model on a similar target task, was expected to use previous task knowledge to learn new tasks quickly with few data. Nevertheless, this approach also does not perform well when the target task distribution is not similar to the distribution of the training task, and it has the risk of overfitting on the target task.

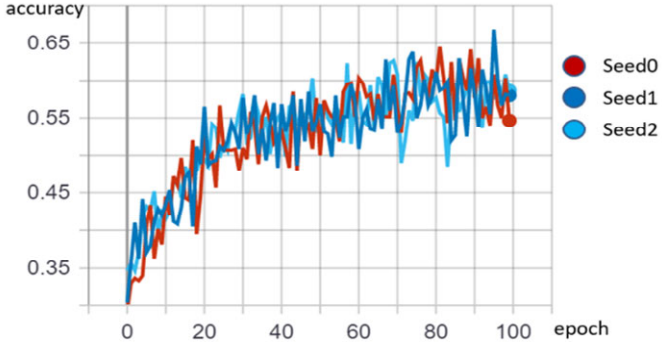
To meet the challenge of training neural networks with few data and adapting to new tasks well, growing importance is being placed on research in few-shot learning, which pursues the goal of training neural networks to acquire new concepts with only a few sampled data points. Usually, few-shot learning approaches have a specific dataset format that consists of a variety of tasks, and each of them is divided into a support set and query set that contain several samples from different classes. Under this dataset structure, few-shot learning networks were expected to learn a new task quickly with few samples in the support set and then precisely recognise the other samples in the query set.

Among the various learning approaches designed to solve few-shot learning problems, meta-learning and continual learning both demonstrate outstanding performance in different benchmark datasets. In recent years, meta-learning has become well known for its powerful generalisation ability when performing few-shot learning tasks. One of the approaches is model-agnostic meta-learning (MAML) (Finn et al., 2017), which is surprising not only in its simplicity but also in its effectiveness. On the other hand, continual learning also performs well on few-shot learning tasks with its principle of learning new tasks without forgetting previous tasks. Gradient episodic memory (GEM) for continual learning (Lopez-Paz and Ranzato, 2017) utilises quadratic programming to address the issue of catastrophic forgetting and even achieves positive forward transfer (FWT), meaning that the model learns new tasks better by using

previous task knowledge, and positive backward transfer (BWT), meaning that the learning of current tasks benefits previous tasks.

Even though MAML manifests outstanding progress on few-shot learning and generalisation, it also suffers from being prone to overfitting (Rusu et al., 2019) and having unstable training performance (Antoniou et al., 2018) during the training process, as shown in Figure 1. From our viewpoint, high learning plasticity is the critical factor in acquiring excellent generalisation ability, but it also means that neural networks may bear the risk of generating unstable performance. In addition, we noted the importance of the gradient produced during the MAML outer loop, which we call the 'meta-gradient' in this paper. Therefore, we propose our enhanced model-agnostic meta-learning (EN-MAML) approach to enhance the stability of MAML with the characteristics of meta-gradients and the features of continual learning. In this paper, our method combines MAML with GEM to resolve the instability problem of MAML mentioned in Antoniou et al. (2018). We use a GEM quadratic program, which was originally used to avoid catastrophic forgetting in Lopez-Paz and Ranzato (2017), to memorise the meta-gradient and adjust the updating direction to promote the learning of previous episodes. Thus, EN-MAML could be expected to achieve a similar effect to that of FWT and BWT and make the models converge faster.

Figure 1 Unstable performance of MAML (see online version for colours)



Notes: This figure illustrates the training accuracy of MAML with three different seeds. The accuracy curves illustrate the unstable performance of MAML during the training time.

Furthermore, we take the stability-plasticity dilemma (Abraham and Robins, 2005) into consideration and rethink the features of meta-learning (Plaat, 2022) and continual learning in the few-shot learning field. Therefore, we design our method to appropriately combine the learning process for current tasks and previous tasks. Consequently, our model can both acquire good adaptability to novel tasks and become more stable.

The main contributions of this paper are listed as follows:

- A novel learning model, EN-MAML, is proposed to cross the border between meta-learning and continual learning to overcome their shortcomings by combining their advantages, providing a possible way to combine these two powerful learning methods.

- We propose a method not only to address the instability of MAML but also to achieve faster convergence on the Omniglot dataset and higher accuracy on the Mini-ImageNet dataset.
- Our method could work on other gradient-based meta-learning algorithms with the same manipulation of the meta-gradient in EN-MAML.
- We also analyse the effect of the meta-gradient buffer setting, such as the mechanism of meta-gradient memory replacement and the size of the meta-gradient buffer, which affect whether EN-MAML achieves a balance in the stability-plasticity dilemma.

The organisation of the paper is as follows. Section 2 discusses the literature review and Section 3 presents the proposed EN-MAML in details. We provide the experimental results in a performance study in Section 4, and conclude the paper in Section 5.

2 Literature review

Our work combines both meta-learning and continual learning, which have become increasingly important in the few-shot learning field in recent years. In this section, we introduce our related works in three parts: meta-learning, continual learning and hybrid methods.

2.1 Meta-learning

Meta-learning mainly focuses on solving the problem of few-shot learning and has also been proven to be effective in the field. By computing the distance of a prototype representation, the model can classify different classes with a small amount of data (Snell et al., 2017). Recent studies based on this concept have become more complicated and delicate in the embedding process (Gidaris and Komodakis, 2018; Dhillon et al., 2020) and even utilise data-dependent initialisations 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 (2018) also uses a representation space to acquire different task knowledge. With a metric space that is based on metric scaling and metric task conditioning, the model can learn novel concepts well under task-dependent scaled metrics (Oreshkin et al., 2018).

Several approaches present powerful generalisation ability in adapting to unseen tasks. Most of these approaches design networks to acquire a learning ability to replace some artificial neural network settings. Using a long short-term memory (LSTM)-based high-level model to learn how to update the base-level model (Ravi and Larochelle, 2016), the method can automatically produce different optimisation algorithms (Finn et al., 2017; Kuo et al., 2021) and train the network to determine appropriate initialisation parameters that perform well on different task distributions. The method proposed in Finn et al. (2017) can be applied to different kinds of model structures to promote their generalisation ability. To explore a set of appropriate initial parameters, the approach proposed in Nichol and Schulman (2018) reduces the cost of calculating in the differentiating process. Moreover, Antoniou et al. (2018), presents various modifications of Finn et al. (2017) and analyses the framework of MAML. It also notes potential

problems in training MAML. Recently, Dhillon et al. (2020), proposed a metric for different few-shot benchmark datasets to evaluate their hardness so that different meta-learning models could compare their performances in a more convincing way. For few-shot image classification, Chen et al. (2021) proposed a meta-learning system to achieve time and resource efficiency and to generalise unknown feedback datasets. Kuo et al. (2021) alleviated catastrophic forgetting, prevented base learners from inducing overfitting, and achieved strong robustness.

2.2 Continual learning

It is difficult to train a model to generalise well with little data, Aljundi et al. (2017) focused on how to make networks absorb new knowledge without forgetting the knowledge from previous tasks that the networks have learned.

Incremental learning aims to gradually learn via continuous training. Incremental learning is divided into task-based incremental learning (Davidson and Mozer, 2020; Riemer et al., 2019; Zhao et al., 2020) and class-based incremental learning (Hu et al., 2021; Liu et al., 2020; Zhang et al., 2021).

Regarding class-based incremental learning, called class-incremental learning (CIL), Liu et al. (2021) proposed a novel network architecture to solve the stability and plasticity dilemma between the old and new classes of learning. This approach can adjust the specific level and weight of a specific stage in an existing CIL method to improve its performance. Hu et al. (2021) 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. The authors showed that the proposed causal effect distillation technique could greatly improve the state-of-the-art CIL methods.

Many continual learning approaches use extra memory to store data for the purpose of alleviating catastrophic forgetting (Rebuffi et al., 2017). 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 (Shaheen et al., 2022). Kirkpatrick et al. (2017) utilised the importance of each model parameter to avoid tuning the sensible weights that are more likely to lead to catastrophic forgetting. Rather than determining the important parameters, Lopez-Paz and Ranzato (2017) focused on modifying the angle of the model's gradient and even proposed the metrics of FWT and BWT to evaluate the performance of continual learning approaches.

From our observation, and as noted in Riemer et al. (2019), the classic stability-plasticity dilemma (Abraham and Robins, 2005) concept seems to match the characteristics of meta-learning and continual learning. The main concept of Riemer et al. (2019) is easing the interference between transfer and retention with gradient alignment, which was proposed in Lopez-Paz and Ranzato (2017). The stability-plasticity dilemma mentioned in Abraham and Robins (2005) means that there is a regulated balance between synaptic stability and synaptic plasticity. Meta-learning presents great adaptation to non-stationary task distribution. However, the problems of training instability and overfitting occurred in Finn et al. (2017). In the research of Rusu et al. (2019) mentioned that it is difficult for gradient-based meta-learning methods to perform few-shot learning in a high-dimensional latent environment. This evidence suggests that meta-learning might implicitly correspond to the plasticity of the stability-plasticity dilemma. On the other hand, most continual learning approaches enhance network stability, such as by

adding more constraints when updating parameters (Lopez-Paz and Ranzato, 2017; Kirkpatrick et al., 2017) or using a buffer to store data (Rebuffi et al., 2017). For the parameters that are influential on previous tasks, the approach presented in Kirkpatrick et al. (2017) updates them carefully and slowly. 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 (2017) compare the gradients of different tasks to confirm that updating the direction will not lead to serious forgetting.

2.3 Hybrid methods

In recent years, some approaches have crossed the border between meta-learning and continual learning and leveraged the advantages of each to overcome their shortcomings. Gai et al. (2021) use meta continual learning to mitigate forgetting with GEM. De Lange et al. (2022) contribute comprehensive experimental comparison of 11 state-of-the-art continual learning methods and four baselines. Riemer et al. (2019) combines meta-learning with GEM (Lopez-Paz and Ranzato, 2017) 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. (2018) 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 Methodology

First, we introduce the framework of MAML, which achieves state-of-the-art few-shot learning by training the network to determine a set of adaptive initialisation parameters that can generalise to various tasks composed of different classes in the dataset. With the specific initialisation parameters, the network can adapt well to different tasks through only a few update steps.

According to the concept of meta-learning, there are two kinds of knowledge networks acquired during the training phase. As a result, we define the *base learner* as a network f_θ with task-level knowledge θ. When the network encounters the support set S_t from a task t, it will update only a few times to adapt to new tasks with step size α. This process is inner-loop updating; i is the time step of the updates, and there are I updates in total. The process can be expressed as:

$$\theta_i^t = \theta_{i-1}^t - \alpha \nabla_\theta L_{S_t}\left(f_{\theta_{i-1}^t}\right), \tag{1}$$

Normally, a batch in a few-shot learning setting contains many tasks that are made according to the N-way K-shot setting. We assume the batch size is T, and then the performance of the initialisation parameters θ_0 can be evaluated by the following:

$$L\left(f_{\theta_0}\right) = \sum_{t=1}^{T} L_{Q_t}\left(f_{\theta_I^t}\right) \tag{2}$$

The loss accumulates as the base learners learn the Q_t query sets of each task from a whole batch, and the total loss is used as a criterion to measure how adaptive the set of initialisation parameters would be. With that, the network cannot only explore the direction in which to adjust θ_0 but also promotes the model's generalisation ability on the next batch of tasks. The process by which the network updates its θ_0 to acquire cross-task-level knowledge to fit better on the next batch is called outer-loop updating, which can be expressed as:

$$\theta_0' = \theta_0 - \beta\nabla\sum_{t=1}^{T} L_{S_t}\left(f_{S_t^N}(\theta)\right), \tag{3}$$

As a result, MAML repeats the process of the inner loop and outer loop and determines an excellent set of initialisation parameters. Hence, the problem could be defined as follows,

(Problem definition) For a batch of tasks $B \in p(T)$, the support set of a batch $S_t = \langle (X_{S_{t1}}, Y_{S_{t1}}), (X_{S_{t2}}, Y_{S_{t2}}), ..., (X_{S_{tn}}, Y_{S_{tn}})\rangle$ is utilised to produce fast weights and to adapt to a new task; the query set of a batch $Q_t = \langle (X_{Q_{t1}}, Y_{Q_{t1}}), (X_{Q_{t2}}, Y_{Q_{t2}}), ..., (X_{Q_{tn}}, Y_{Q_{tn}})\rangle$ is given to evaluate the generalisation of networks to tasks.

Although MAML has shown powerful adaptation ability in the field of few-shot learning, it also has some potential problems, for example, training instability and being prone to overfitting. In our study, we proposed our EN-MAML method to overcome these problems and to improve the performance of the original MAML.

3.1 EN-MAML architecture

From Figure 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 non-stationary 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 initialisation parameters.

However, the task distribution of few-shot learning can be seen as a non-stationary environment (Yosinski et al., 2014). 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 Riemer et al. (2019) also proved its effectiveness. Conflicts between gradients will occur in the following situation:

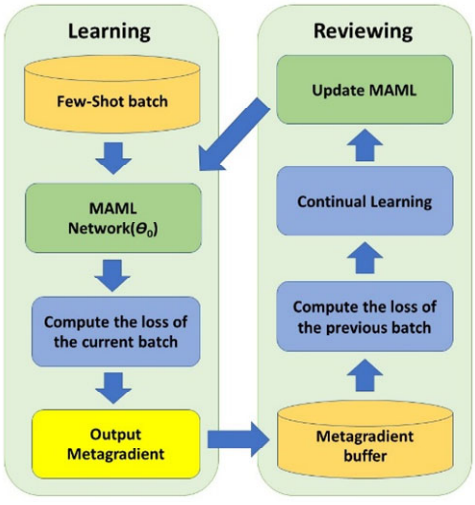
$$\frac{\partial L\left(f_\theta\left(x_i\right), y_i\right)}{\partial \theta} \cdot \frac{\partial L\left(f_\theta\left(x_j\right), y_j\right)}{\partial \theta} < 0, \tag{4}$$

(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:

$$\frac{\partial L\left(f_\theta(x), y\right)}{\partial \theta} \cdot \frac{\partial L\left(f_\theta\left(x_M\right), y_M\right)}{\partial \theta} \geq 0, \tag{5}$$

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 utilise 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.

Figure 2 The architecture of EN-MAML (see online version for colours)



3.2 Meta-learning with the meta-gradient algorithm

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. More details of EN-MAML are described in the following algorithm:

Algorithm EN-MAML for supervised few-shot learning

Require $p(T)$, distribution of tasks

Procedure (f_θ, α, β, M)

while not done:

Sample a batch of tasks $B_0, B_1 \ldots B_n$ from $p(T)$

for each $B_0, B_1 \ldots B_n$ **do**

if M **not** full:

Store B_1 to M

else:

Replace the oldest B with the next B

for each $t_0, t_1 \ldots t_n$ in the current B **do**

for each step from i to I:

$$\theta_i^t = \theta_{i-1}^t - \alpha \nabla_\theta l_{S_t}\left(f_{\theta_{i-1}^t}\right)$$

compute $L_{Q_t}\left(f_{\theta_I^t}\right)$

end For

end For

$$L\left(\theta_0^{B_i}\right) = \sum_{t=1}^{T} L_{Q_t}\left(f_{\theta_I^t}\right)$$

Compute the gradient of the current batch g_c

Compute g_{B_n} for all B_n in M

Get g_c' by quadratic programming

$$\theta_0^{B_i} = \theta_0^{B_{i-1}} - \beta g_c'$$

end For

3.3 The loss function of EN-MAML

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).

$$l_c = \left(f_\theta(x, y)\right) = \sum_{x,y\sim T} y \log f_\theta(x), \tag{6}$$

$$L_{total} = w_c \sum_{t=1}^{T} l_c \left(f_\theta(x, y)\right) + w_p \sum_{t=1}^{T} l_p \left(f_\theta\left(x_M, y_M\right)\right), \tag{7}$$

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

In this section, we follow the experimental protocol of classification in the MAML paper (Antoniou et al., 2018) and evaluate the performance of EN-MAML. To fairly compare the performance, we also compare with MAML. In addition, we use Torchmeta (Deleu et al., 2019), a powerful tool package built by PyTorch, to reproduce and design our model architecture. Although it is difficult to replicate the same results as in the MAML paper (Antoniou et al., 2018), we use the relative performance from our implementation to compare the two models.

Table 1 Dataset description

	Omniglot	*Mini-ImageNet*
Number of classes	1,623	100
Number of images	32,460	60,000
Training classes	1,028	64
Validation classes	172	16
Testing classes	423	20

In our experiments, the datasets that we use to evaluate our model are Omniglot (Lake et al., 2015) and Mini-ImageNet (Vinyals et al., 2016), which are the benchmarks in the few-shot learning field. There are 1,623 handwritten characters classified as 50 different letters in the Omniglot dataset. Each of the classes contains 20 instances of handwritten symbols. In Torchmeta, the Omniglot dataset is split into a training set that contains 1,028 classes, a validation set that contains 172 classes and a testing set that contains 423 classes. The majority of few-shot learning methods use the first 1,200 classes in Omniglot for training (Antoniou et al., 2018). Other research has mentioned that preserving a few classes to perform validation is important (Antoniou et al., 2018). Therefore, we also use validation to perform our experiments. In Torchmeta, the Mini-ImageNet dataset, which contains 600 instances in each class, consists of 64 classes in the training set, 16 classes in the validation set and 20 classes in the testing set. For

both datasets, we augment them by rotating the images 90 degrees and decreasing the image sizes to 28 × 28 in Omniglot and 84 × 84 in Mini-ImageNet. Dataset information is shown in Table 1.

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* (Koch et al., 2015): Siamese nets utilise the similarity between the inputs to effectively perform image classification.
- *Matching nets* (Vinyals et al., 2016): Based on the metric learning concept, matching nets design their model with external memory and an attention mechanism to make the model learn the important image feature representation rapidly.
- *Neural statistician* (Edwards and Storkey, 2017): The model is efficiently trained from a statistical viewpoint. By observing the statistics of a dataset, the model can use parameters and data to perform few-shot learning.
- *Memory mod* (Kaiser et al., 2017): With a lifelong memory module and fast nearest-neighbor algorithm, this method enables the network to perform lifelong few-shot learning
- *MAML* (Finn et al., 2017; Antoniou et al., 2018): This approach presents fast adaptation with few-shot learning to learn metalevel knowledge by modifying the inductive bias.
- *Reptile* (Nichol and Schulman, 2018): Reptile is a gradient-based meta-learning method that trains the model to find a set of excellent initialisation parameters through an efficient updating process.

We 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 2. Second, we perform 20-way few-shot classification on Omniglot. EN-MAML can also achieve better performance than state-of-the-art models, as shown in Table 3. Compared to the performance of MAML, EN-MAML improves the accuracy by approximately 0.12%, as shown in Table 3. We can also enhance the accuracy by approximately 0.52% in the Omniglot 5-way 1-shot setting, as shown in Table 2.

For the Omniglot 5-way 5-shot setting, EN-MAML also improves accuracy by approximately 6.09% compared to our MAML replication in the Omniglot 20-way 1-shot setting. For the Omniglot 20-way 5-shot setting, EN-MAML improves accuracy by approximately 0.07% compared to MAML, as shown in Table 3. For the Mini-ImageNet datasets, EN-MAML also demonstrated dramatically higher performance on 5-way classification experiments, as shown in Table 4. 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 Omniglot for 5-way classification

Model	*Accuracy*	
	1-SHOT	*5-SHOT*
Siamese nets (Koch et al., 2015)	97.3%	98.4%
Matching nets (Vinyals et al., 2016)	98.1%	98.9%
Neural statistician (Edwards and Storkey, 2017)	98.1%	99.5%
Memory mod. (Kaiser et al., 2017)	98.4%	99.6%
MAML (Finn et al., 2017; Antoniou et al., 2018)	98.25%	98.85%
Reptile (Nichol and Schulman, 2018)	95.30%	98.80%
EN-MAML	*98.77%*	*99.67%*

Table 3 Accuracy of Omniglot for 20-way classification

Model	*Accuracy*	
	1-SHOT	*5-SHOT*
Siamese nets (Koch et al., 2015)	88.2%	97.0%
Matching nets (Vinyals et al., 2016)	93.8%	98.5%
Neural statistician (Edwards and Storkey, 2017)	93.2%	98.1%
Memory mod. (Kaiser et al., 2017)	95.0%	98.6%
MAML (Finn et al., 2017; Antoniou et al., 2018)	93.58%	97.81%
Reptile (Nichol and Schulman, 2018)	87.99%	96.32%
EN-MAML	*93.70%*	*97.88%*

Table 4 Accuracy of Mini-ImageNet 5-way classification

Model	*Accuracy*	
	1-SHOT	*5-SHOT*
Siamese nets (Koch et al., 2015)	47.8%	63.66%
Matching nets (Vinyals et al., 2016)	43.56%	55.31%
Neural statistician (Edwards and Storkey, 2017)	48.60%	63.09%
Memory mod. (Kaiser et al., 2017)	49.21%	65.42%
MAML (Finn et al., 2017; Antoniou et al., 2018)	49.38%	66.55%
Reptile (Nichol and Schulman, 2018)	46.81%	62.37%
EN-MAML	*54.55%*	*72%*

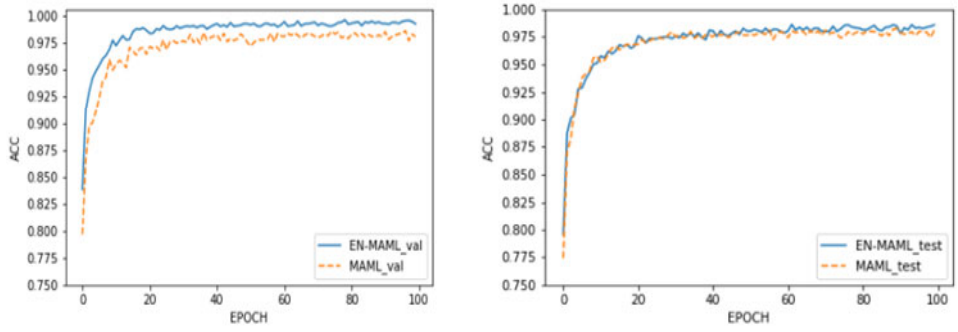
4.2 Stability and accuracy comparison with MAML

To fully compare and analyse 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. (2018).

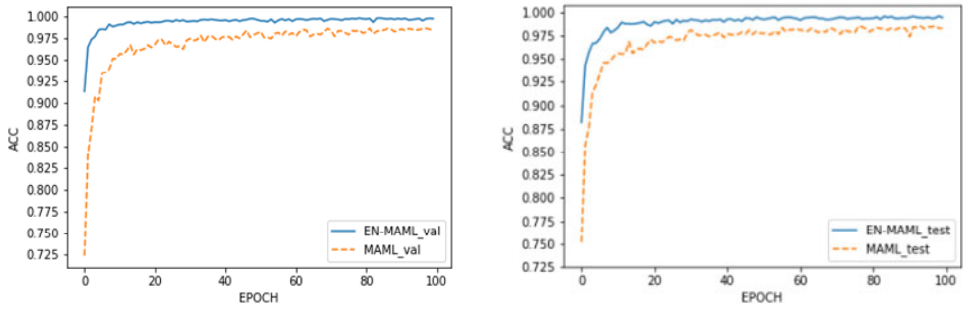
We can observe that the testing accuracy of EN-MAML starts to surpass that of MAML when the model has been trained for approximately 40 epochs, as shown in Figure 3(a). Moreover, the gap in testing accuracy between EN-MAML and MAML gradually increases as the number of epochs increases. Additionally, our method provides more stable validation accuracy, which is one of our method's objectives. In Figure 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. Figure 4 shows that EN-MAML cannot 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 Figure 4(a), and this result can also be observed in the validation experiment in Figure 4(b).

Figure 3 Comparison of EN-MAML and MAML in the 5-way 1-shot setting on the Omniglot dataset, (a) validation accuracy (b) testing accuracy (see online version for colours)



(a) (b)

Figure 4 Comparison of EN-MAML and MAML in the 5-way 5-shot setting on the Omniglot dataset, (a) validation accuracy (b) testing accuracy (see online version for colours)

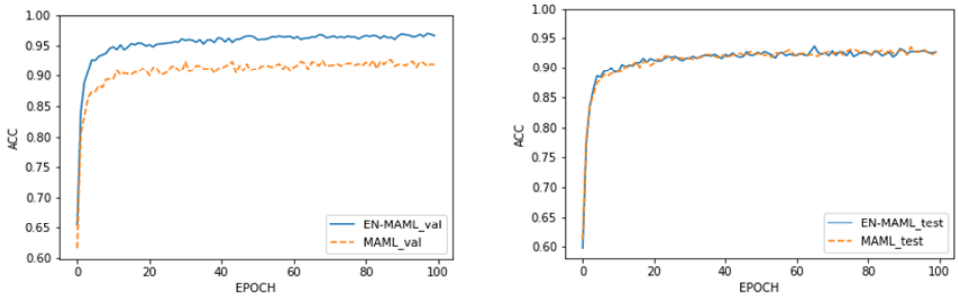


(a) (b)

In the Omniglot 20-way 1-shot experimental setting, EN-MAML obtains the highest testing accuracy in Figure 5(a), and EN-MAML obviously outperforms MAML in the

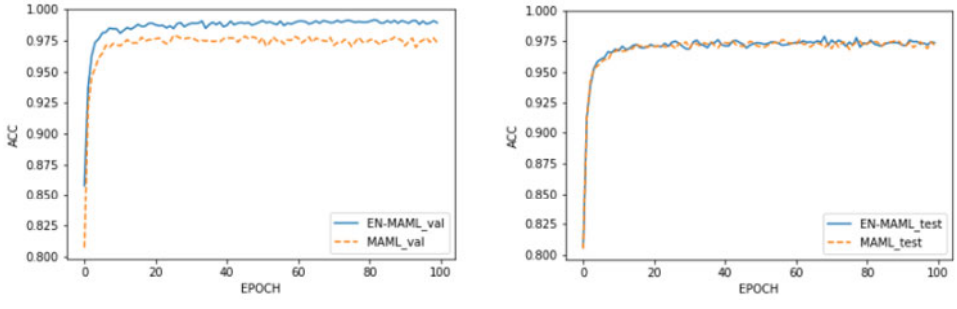
validation accuracy experiment in Figure 5(b). Although the testing accuracy improves slightly, EN-MAML proves that it can effectively promote training stability. In Figure 6(a), EN-MAML shows little improvement in testing accuracy in the 20-way 5-shot setting on Omniglot. However, EN-MAML has quite a large improvement in the validation accuracy performance. We also note that EN-MAML has greater stability improvement in the 20-way experimental setting than in the 5-way experimental setting when both EN-MAML and MAML have converged.

Figure 5 Comparison of EN-MAML and MAML in the 20-way 1-shot setting on the Omniglot dataset, (a) validation accuracy (b) testing accuracy (see online version for colours)



(a) (b)

Figure 6 Comparison of EN-MAML and MAML in the 20-way 5-shot setting on the Omniglot dataset, (a) validation accuracy (b) testing accuracy (see online version for colours)



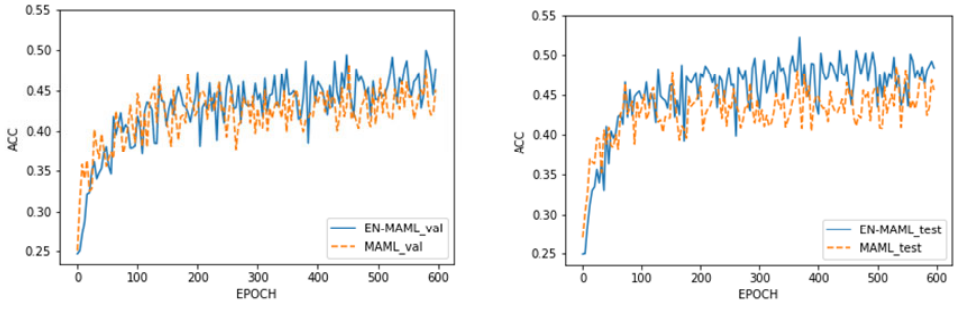
(a) (b)

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.

However, EN-MAML still reaches the highest accuracy in the 5-way 1-shot setting in Figure 7(a) and maintains an equivalent level of validation accuracy in Figure 7(b). Additionally, EN-MAML outperforms MAML most of the time in the 5-way 5-shot setting on Mini-ImageNet in Figure 8(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 analysed the results, and we will

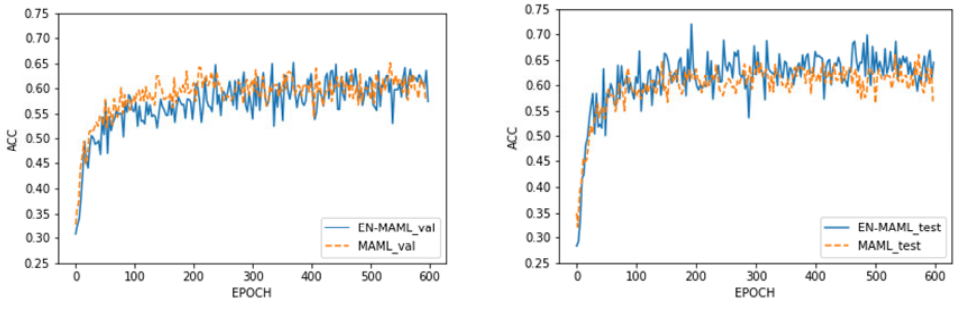
discuss this phenomenon in the next section. To summarise all the experimental results on Omniglot and Mini-ImageNet, our observation is that EN-MAML either improves the testing accuracy or promotes validation accuracy. EN-MAML progresses on at least one metric and keeps the other metric at an equivalent 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.

Figure 7 Comparison of EN-MAML and MAML in the 5-way 1-shot setting on the Mini-ImageNet dataset, (a) validation accuracy (b) testing accuracy (see online version for colours)



(a) (b)

Figure 8 Comparison of EN-MAML and MAML in the 5-way 5-shot setting on the Mini-ImageNet dataset, (a) validation accuracy (b) testing accuracy (see online version for colours)



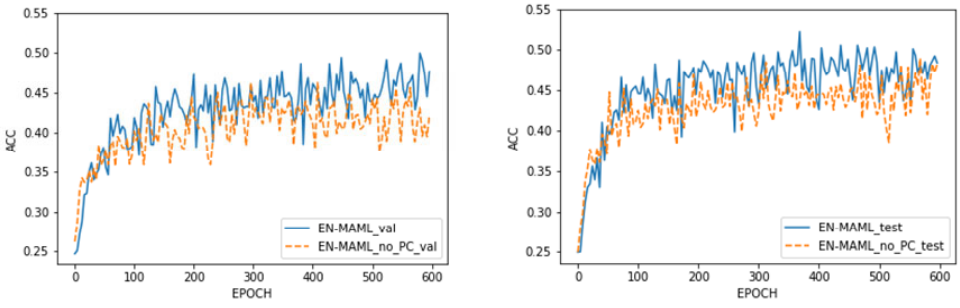
(a) (b)

4.3 The effectiveness of combining meta-learning with continual learning

Initially, we expected the addition of quadratic programming, which is the mechanism used in GEM to alleviate catastrophic forgetting, to only promote the stability of MAML. As per our expectation, 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, as shown in Figures 3–6. 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

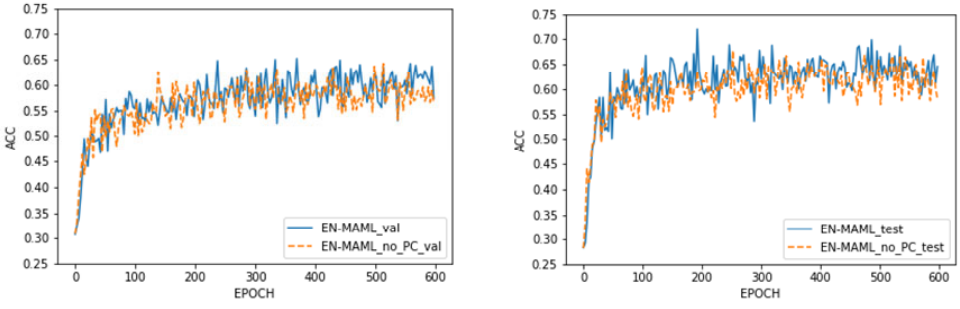
more clearly in Figures 4, 7 and 8. From our experimental observation, the modified meta-gradient, which is generated from quadratic programming to maintain the meta-gradient information from previous batches, can have approximately the same effect as the FWT proposed in Lopez-Paz and Ranzato (2017). Operating with our algorithm, the meta-gradient stored in the buffer will migrate new metalevel information from different batches of tasks. As the training continues, the meta-gradient in our buffer accumulates more sufficient information about the distribution of the experimental dataset, and then the model can update its parameters in a more correct and stable direction. Therefore, our learning algorithm can be separated into learning from current knowledge and learning from previous knowledge. Continual learning excels in maintaining previous task knowledge, so it can demonstrate excellent stability. MAML has flexible learning ability, but it is difficult to promote stability and plasticity concurrently, which is a difficult issue to overcome.

Figure 9 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 (see online version for colours)



(a) (b)

Figure 10 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 (see online version for colours)



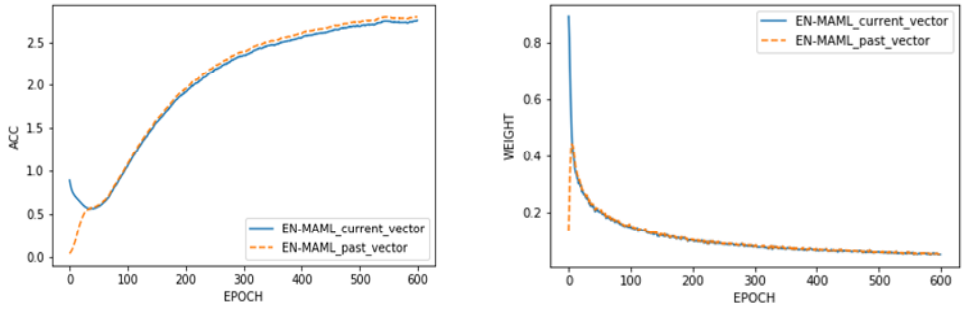
(a) (b)

To evaluate the effectiveness of weighted current knowledge and previous knowledge, we test the performance of the original EN-MAML and EN-MAML without considering dynamic weights. From Figures 9 and 10, 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.

Originally, we expected that the model would focus on current task knowledge at first and gradually shift the importance from current tasks to previous tasks. However, we observe that not only the time but also different experimental settings influence the importance of the two kinds of knowledge in Figure 11. 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 can see 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 Figure 11(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 Figure 11(b).

Figure 11 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 (see online version for colours)



(a) (b)

From our observation and analysis, the reason why EN-MAML can maintain higher stability and accuracy concurrently is that it increases the attention to previous and current knowledge to overcome the limitation of the few training data. On the other hand, when EN-MAML learns the task knowledge with relatively more data for training, it gradually decreases the attention on the coming tasks to overcome the unstable environment of few-shot learning. In other words, 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 (2017) 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.

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 Figures 7 and 8. 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 Figures 7 and 8 is demonstrated more clearly. In these figures, we find that EN-MAML outperforms MAML more dramatically and reaches the highest accuracy in later epochs.

Finally, we discuss the design of our meta-gradient buffer. First, the size of the gradient buffer is vital to EN-MAML. An oversized meta-gradient buffer will lead to learning much more slowly or even crashing. The appropriate batch size can be determined by the difficulty of the dataset and the batch size of the tasks. In a relatively easy experimental environment, we suggest that the model should be trained for small meta-gradient batch sizes. Conversely, it should be increased in a more difficult setting. In our implementation, the meta-gradient batch size is 1 in Omniglot and 2 in Mini-ImageNet. Compared to other few-shot learning methods with external memory, our buffer size can be adjusted with different settings and environments. Moreover, we use a relatively small memory size to reduce the computation and storage space because the meta-gradient stored in our buffer can be continually updated to become more adaptive. This is why we only need a little memory for storage. Regarding the computation in migrating each meta-gradient, we also consider applying weights to each meta-gradient before performing quadratic programming. Additionally, how frequently the old memory is replaced is an important setting. In our method, we replace old memory when the new meta-gradient is generated.

4.4 Hyperparameter settings

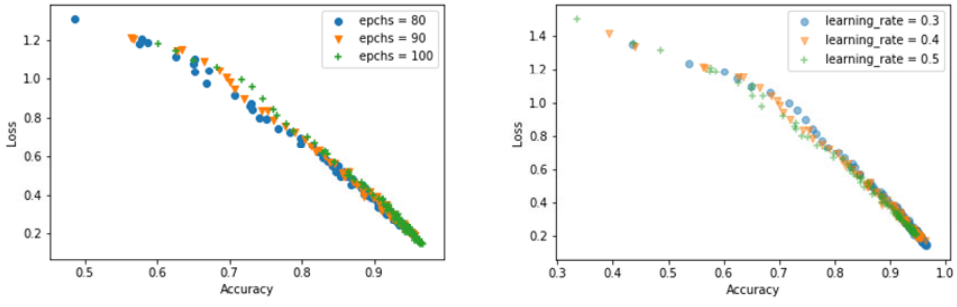
To evaluate our model, we followed the experimental protocol proposed in Riemer et al. (2019), which is also followed by MAML. The N-way K-shot protocol is often used to evaluate a model's classification ability in few-shot learning. In the N-way K-shot experimental protocol, N classes are randomly selected, and each of them has K samples.

To determine more appropriate hyperparameter settings, we not only study the experimental results from previous meta-learning methods but also train EN-MAML with different epochs and learning rates to search for better hyperparameter settings. From the experimental results in Figure 12, we find that EN-MAML demonstrates better performance under 100 epochs and a 0.3 learning rate. Additionally, we train EN-MAML with 600 epochs and a 0.4 learning rate according to the experimental results demonstrated in Figure 12.

According to the experimental results in Figures 12 and 13, we train the models for 100 epochs, and each epoch contains 100 iterations with a step size of 0.4 in all Omniglot experiments. For all Mini-ImageNet experiments, we train the models for 600 epochs, and each epoch consists of 100 iterations with a step size of 0.1. Each classification experiment trains for 600 epochs, and each epoch consists of 100 iterations. Every classification task in a batch is randomly generated and shuffled. In the stability experiment, we show the validation accuracy curve tendency in the first 100 epochs. In both the 5-way and 20-way Omniglot experiments, we set the meta-batch size to 32 tasks. For Mini-ImageNet, we train our model with a step size of 0.01. We set the meta-batch

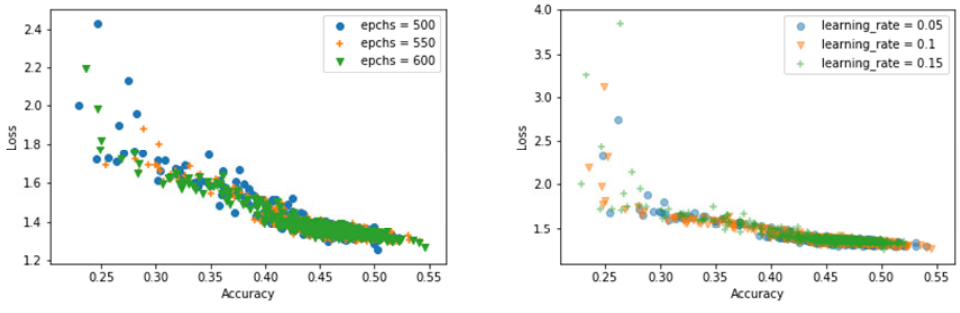
size to 2 tasks in the 5-shot experiment and set the meta-batch size to 2 tasks in the 1-shot experiment. In addition, we train the models with one gradient step and test the models with 5 gradient steps in all experiments.

Figure 12 The performance of EN-MAML under different hyperparameter settings on the Omniglot dataset, (a) different numbers of epoch settings under the 5-way 1-shot setting (b) different learning rate settings under the 5-way 1-shot setting (see online version for colours)



(a) (b)

Figure 13 The performance of EN-MAML under different hyperparameter settings on the Mini-ImageNet dataset, (a) different numbers of epoch settings under the 5-way 1-shot setting (b) different learning rate settings under the 5-way 1-shot setting (see online version for colours)



(a) (b)

The performance of EN-MAML is analysed above, proving that it prevails over the other frequent-used few-shot learning models, such as MAML, Siamese nets, matching nets, neural statistician, memory mod, and reptile, in terms of stability and accuracy. In addition, the meta-gradient stored in the buffer can be continuously updated, making the proposed approach more adaptive than other models. Model training stability experiments are carried to examine our method, showing that it can alleviate the unstable training condition problem proposed in Antoniou et al. (2018).

5 Conclusions

In this paper, we introduce a novel method, EN-MAML, which combines meta-learning with continual learning by leveraging the meta-gradient property with quadratic programming. We provide higher stability in model training and better testing accuracy than MAML in most experimental settings. The experimental results indicate that the combination of meta-learning and continual learning can have the potential to increase flexibility and stability concurrently. In the future, few-shot learning research can explore more possible ways to leverage the features of meta-learning and continual learning to overcome the shortcomings of both to address the stability-plasticity dilemma.

References

Abraham, W.C. and Robins, A. (2005) 'Memory retention – the synaptic stability versus plasticity dilemma', *Trends in Neuroscience*, Vol. 28, No. 2, pp.73–78.

Aljundi, R., Chakravarty, P. and Tuytelaars, T. (2017) 'Expert gate: lifelong learning with a network of experts', *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.3366–3375.

Antoniou, A., Edwards, H. and Storkey, A. (2018) 'How to train your MAML', in *Proceedings of the International Conference on Learning Representations*.

Chen, Y., Guan, C., Wei, Z., Wang, X. and Zhu, W. (2021) 'MetaDelta: a meta-learning system for few-shot image classification', in *Proceedings of Machine Learning Research, AAAI Workshop on Meta-Learning and MetaDL Challenge*, Vol. 140, pp.17–28.

Davidson, G. and Mozer, M. (2020) 'Sequential mastery of multiple visual tasks: networks naturally learn to learn and forget to forget', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, pp.9282–9293.

De Lange, M., Aljundi, R., Masana, M., Parisot, S., Jia, X., Leonardis, A., Slabaugh, G. and Tuytelaars, T. (2022) 'A continual learning survey: defying forgetting in classification tasks', *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, pp.3366–3385.

Deleu, T., Würfl, T., Samiei, M., Cohen, J.P. and Bengio, Y. (2019) *Torchmeta: A Meta-Learning Library for Pytorch*, arXiv preprint arXiv:1909.06576.

Dhillon, G.S., Chaudhari, P., Ravichandran, A. and Soatto, S. (2020) 'A baseline for few-shot image classification', in *International Conference on Learning Representations (ICLR)*.

Edwards, H. and Storkey, A. (2017) 'Towards a neural statistician', in *Proceedings of the 5th International Conference on Learning Representations*, Toulon, France, pp.1–13.

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

Gai, S., Chen, Z. and Wang, D. (2021) 'Multi-modal meta continual learning', *2021 International Joint Conference on Neural Networks (IJCNN)*, IEEE, pp.1–8.

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

Hu, X., Tang, K., Miao, C., Hua, X.S. and Zhang, H. (2021) '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, pp.3957–3966.

Kaiser, Ł., Nachum, O., Roy, A. and Bengio, S. (2017) 'Learning to remember rare events', in *Proceedings of the Seventh International Conference on Learning Representations (ICLR)*.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D. and Hadsell, R. (2017) 'Overcoming catastrophic forgetting in neural networks', *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 114, No. 13, pp.3521–3526.

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

Kuo, N.I., Harandi, M., Fourrier, N., Walder, C., Ferraro, G. and Suominen, H. (2021) 'Learning to continually learn rapidly from few and noisy data', *AAAI Workshop on Meta-Learning and MetaDL Challenge, PMLR*, Vol. 140, pp.65–76.

Lake, B.M., Salakhutdinov, R. and Tenenbaum, J.B. (2015) 'Human-level concept learning through probabilistic program induction', *Science*, Vol. 350, No. 6266, pp.1332–1338.

Liu, Y., Schiele, B. and Sun, Q. (2021) 'Adaptive aggregation networks for class-incremental learning', in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, pp.2544–2553.

Liu, Y., Su, Y., Liu, A.A., Schiele, B. and Sun, Q. (2020) 'Mnemonics training: multi-class incremental learning without forgetting', in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, pp.12245–12254.

Lopez-Paz, D. and Ranzato, M.A. (2017) 'Gradient episodic memory for continual learning', in *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA.

Nichol, A. and Schulman, J. (2018) *Reptile: A Scalable Metalearning Algorithm*, arXiv preprint arXiv:1803.02999.

Om, K., Boukoros, S., Nugaliyadde, A. et al. (2020) 'Modelling email traffic workloads with RNN and LSTM models', *Hum. Cent. Comput. Inf. Sci*., Vol. 10, No. 1.

Oreshkin, B.N., Rodriguez, P. and Lacoste, A. (2018) 'Tadam: task dependent adaptive metric for improved few-shot learning', in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., Red Hook, NY, pp.721–731.

Plaat, A. (2022) 'Meta-Learning', in *Deep Reinforcement Learning*, Springer, Singapore, https://doi.org/10.1007/978-981-19-0638-1_9.

Ravi, S. and Larochelle, H. (2016) 'Optimization as a model for few-shot learning', in *International Conference on Learning Representations (ICLR)*, Toulon, France, pp.281–288.

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

Riemer, M., Cases, I., Ajemian, R., Liu, M., Rish, I., Tu, Y. and Tesauro, G. (2019) 'Learning to learn without forgetting by maximizing transfer and minimizing interference', in *Proceedings of the Seventh International Conference on Learning Representations (ICLR)*.

Rusu, A.A., Rao, D., Sygnowski, J., Vinyals, O., Pascanu, R., Osindero, S. and Hadsell, R. (2019) 'Meta-learning with latent embedding optimization', in *Proceedings of the Seventh International Conference on Learning Representations (ICLR)*.

Shaheen, K., Hanif, M.A., Hasan, O. et al. (2022) 'Continual learning for real-world autonomous systems: algorithms, challenges and frameworks', *J. Intell. Robot Syst*., Vol. 105, No. 1.

Snell, J., Swersky, K. and Zemel, R. (2017) 'Prototypical networks for few-shot learning', in *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, pp.4077–4087.

Vinyals, O., Blundell, C., Lillicrap, T. and Wierstra, D. (2016) 'Matching networks for one shot learning', *Advances in Neural Information Processing Systems*, Vol. 29, pp.3637–3645.

Yosinski, J., Clune, J., Bengio, Y. and Lipson, H. (2014) 'How transferable are features in deep neural networks?', in *Proceedings of the 27th International Conference on Neural Information Processing Systems*, Montreal, Canada, pp.3320–3328.

Zhang, C., Song, N., Lin, G., Zheng, Y., Pan, P. and Xu, Y. (2021) 'Few-shot incremental learning with continually evolved classifiers', in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, pp.12455–12464.

Zhao, B., Xiao, X., Gan, G., Zhang, B. and Xia, S.T. (2020) 'Maintaining discrimination and fairness in class incremental learning', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, pp.13208–13217.