Self-Attention Meta-Learner for Continual Learning

Ghada Sokar
Eindhoven University of Technology
Eindhoven, The Netherlands
g.a.z.n.sokar@tue.nl

Decebal Constantin Mocanu
University of Twente
Eindhoven University of Technology
The Netherlands
d.c.mocanu@utwente.nl

Mykola Pechenizkiy
Eindhoven University of Technology
Eindhoven, The Netherlands
m.pechenizkiy@tue.nl

Abstract
Continual learning aims to provide intelligent agents capable of learning multiple tasks sequentially with neural networks. In most settings of the current approaches, the agent starts from randomly initialized parameters and is optimized to master the current task regardless of the usefulness of the learned representation for future tasks. Moreover, each of the future tasks uses all the previously learned knowledge although parts of this knowledge might not be helpful for its learning. These cause interference among tasks, especially when the data of previous tasks is not accessible. In this paper, we propose a new method, named Self-Attention Meta-Learner (SAM), which learns a prior knowledge for continual learning that permits learning a sequence of tasks, while avoiding catastrophic forgetting. SAM incorporates an attention mechanism that learns to select the particular relevant representation for each future task. We empirically show that we can achieve a better performance than several state-of-the-art methods for continual learning by building on top of the selected representation learned by SAM. We also show the role of the meta-attention mechanism in boosting informative features corresponding to the input task and identifying the correct target in the task agnostic inference. Finally, we demonstrate that popular existing continual learning methods gain a performance boost when they adopt SAM as a starting point.

Keywords
Continual Learning; Prior Knowledge; Self-Attention; Task Agnostic Inference

ACM Reference Format:

1 INTRODUCTION

Deep neural networks have achieved outstanding performance in different areas such as visual recognition, natural language processing, and speech recognition [1, 6, 10, 11, 24]. However, the performance degrades when the model interacts with a dynamic open environment and operates on non-stationary data, a phenomenon known as catastrophic forgetting [15]. Continual learning (CL) addresses this problem and aims to provide neural networks with lifelong learning capability. Many works have been proposed to address the CL paradigm [7, 12, 16, 19–23]. This paradigm includes many desiderata, other than mitigating forgetting, such as allowing forward transfer and dealing with the inaccessibility of previous tasks data (see [2, 8, 18] for the complete list).

In this work, we shed the light on two other desiderata that are not widely addressed in the state-of-the-art and illustrate their effective role in boosting the performance and satisfying other CL desiderata. First, the necessity of having a good quantity of prior knowledge to help new tasks to learn in the continual learning paradigm. Second, selecting the useful and relevant parts only from the previous knowledge to learn each of the future tasks instead of using the whole knowledge.

Our contributions in this paper can be summarized as follows:
First, we propose a Self-Attention Meta learner (SAM) that builds a prior knowledge that permits learning a continual sequence of tasks. In addition, SAM learns to pick the relevant representation for each task. Second, we address the challenging and realistic scenario where the task identity is not available during inference (task agnostic). We also assume that the data of previous tasks is not accessible. Third, we achieved a better performance than the state-of-the-art methods by building on top of the learned representation by SAM. Finally, we show that SAM significantly improves the performance of popular existing continual learning strategies.

2 SELF-ATTENTION META-LEARNER (SAM)

Figure 1 shows an overview of our proposed approach, SAM. Herein, we briefly discuss SAM, while its complete description can be found in the full version1. We can divide our approach into two main...
phases: prior knowledge construction and the continuous learning of tasks. In particular, the model consists of two parts. The first part represents the prior knowledge parameterized by the shared learned meta-parameters \( \theta \). This prior knowledge should be characterized by good generality that enables out-of-domain tasks to learn on top of it. To satisfy this objective, we train the shared parameters \( \theta \) using the optimization-based meta-learning algorithm MAML [3] which proves its ability to generalize to out-of-distribution tasks [4]. The second part contains a specific branch for each task \( t_i \) parameterized by \( \phi_i \) to capture the specific discriminative representation. Each branch consists of a few layers that are added on the top of the shared sub-network when the model faces a new task. Unlike previous methods in which all the previous knowledge is used in learning each task, in this work, we select the relevant sparse representation and each task builds a specific representation branch on the top of the selected knowledge. To address this goal, we incorporate the self-attention mechanism proposed by [5] in our meta-learner. An attention module follows each layer in the shared sub-network which learns to pick the relevant features from that layer. Rather than the standard training of the self-attention mechanism as in [5], we exploit meta-learning to allow the network to learn to boost the informative features corresponding to the incoming data. Our analysis shows that the meta-attention mechanism plays an effective role in boosting the performance of the CL setting and mitigating forgetting.

3 EXPERIMENTS AND RESULTS

We compare SAM with the state-of-the-art approaches in the regularization and architectural strategies on the commonly used benchmarks for CL: split CIFAR-10/100 and split MNIST [23]. In addition, we provide an extensive analysis of the role of each of the proposed desiderata in the performance of the CL paradigm. Herein, we briefly discuss the results on the more complex inputs (split CIFAR-10/100) and illustrate how SAM enhances the performance of popular CL methods. For the interested reader, please see the full version of this paper for the full experiments and analysis.

For the split CIFAR-10/100 benchmark, the shared sub-network of SAM is trained on MiniImagenet [17] to construct the prior knowledge and learn the meta-attention mechanism. Then, the model faces each of the 6 tasks of split CIFAR-10/100 sequentially. Figure 2 shows the accuracy of each task after training all tasks along with the average accuracy across all tasks. The regularization methods (EWC [7], LWF [9], and SI [23]) suffer from forgetting previous tasks while having a good performance on the last trained task. On the other hand, the performance achieved by SAM on each of the tasks is close to each other, outperforming the regularization methods by a big margin. In addition, the results show that the learned representation by SAM generalizes better than the counterpart CWR method [13], outperforming it by around 24.5%. It worths to be highlighted that SAM achieves a performance that is better than optimizing a separate network for each task from scratch.

To investigate further the importance of our proposed desiderata, we analyze the performance of popular CL methods when they are combined with SAM as well as their original form. In particular, we allow for accumulating the knowledge from each CL task in the shared sub-network of SAM while the catastrophic forgetting is addressed using each of the studied methods. We also add the fine-tuning method as another baseline, where new tasks are trained continuously without any mechanism to avoid forgetting. Interestingly, SAM always improves the performance as shown in Table 1. Although the regularization methods have low performance in the task agnostic scenario as shown before, combining SAM with the SI method leads to a significant improvement: around 47% and 29% on the split MNIST and split CIFAR-10/100 benchmarks respectively. SAM enhances the performance of the optimization-based meta-learning method (MER) by 17.5% on split MNIST. Moreover, the combination of SAM with the fine-tuning baseline increases its performance despite that there is no forgetting avoidance strategy. SAM reduces the forgetting by allowing an adaptive update for the weights. The update of the weights becomes a function of the recalibrated activations by SAM. Therefore, the knowledge accumulated by the new tasks affects a subset of the previously learned representation which mitigates forgetting.

4 CONCLUSION

In this paper, we propose SAM, a self-attention meta-learner for the continual learning paradigm. SAM learns a prior knowledge that can generalize to new distributions and learns to boost the relevant features to the input task. These two desiderata are largely overlooked in the state-of-the-art; however, our empirical evaluation and analysis show their effective role in improving the performance of the CL paradigm. Finally, we demonstrate that combining SAM with the existing continual learning methods boosts their performance. Our results show the potential of the proposed method in the CL setting and open the path for several new research directions.

<table>
<thead>
<tr>
<th>Method</th>
<th>Split MNIST</th>
<th>Split CIFAR-10/100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fine-tuning</td>
<td>19.86 ± 0.04</td>
<td>53.87 ± 1.73</td>
</tr>
<tr>
<td>SI</td>
<td>19.99 ± 0.06</td>
<td>67.32 ± 0.43</td>
</tr>
<tr>
<td>MER</td>
<td>32.66 ± 2.33</td>
<td>50.04 ± 1.85</td>
</tr>
</tbody>
</table>

Table 1: Enhancing existing CL strategies by SAM. “Standard” represents the original form of the methods.
REFERENCES


