Dark Experience for General Continual Learning: a Strong, Simple Baseline

Summary and Notes

Qilin Ye
Department of Computer Science
University of Southern California
Los Angeles, CA 90007
yeqilin@usc.edu

1 Summary

This paper presents two novel General Continual Learning (GCL) models, Dark Experience Replay (DER) and DER++, which can learn incrementally without knowing the task boundaries or identities. The models use a memory buffer of sample images and logits, and upon receiving a new task, they regularize the loss by requiring the newly logins to be consistent with the buffered ones. The authors evaluate their models on several benchmark datasets and find that DER and DER++ outperform other traditional CL methods under most settings. The authors also propose a new protocol for CL evaluation, MNIST-360, and demonstrate DER and DER++'s strong performance on this benchmark. Finally, the authors examine some intriguing properties of their novel models, including convergence to flatter minima and more calibrated networks, and they hypothesize that these properties are linked to DER and DER++'s strong performance.[†]

1.1 Background

Continual Learning (CL) aims to develop models that can maintain their learned knowledge as new tasks arrive. This is a response to the challenge of catastrophic forgetting, a phenomenon where models may abruptly forget previously learned knowledge when handed new tasks. There are three main categories of CL, which are Task-IL, Domain-IL, and Class-IL.

Currently, common strategies to tackle catastrophic forgetting include rehearsal-based methods (keeping a memory buffer and mixing old examples with recent ones), knowledge distillation (appointing a past state as "teacher" and transfer knowledge to the "student"), regularization (additional loss to penalize change of weight parameters), and so on.

The authors of this paper argue that existing CL models fail to reflect real-world scenarios. To address this, they further introduce three restrictions on CL: (i) no tasks (input is not partitioned into tasks), (ii) no test-time oracle (task have no identities), and (iii) constant memory (an upper bound on memory training). These requirements define the General Continual Learning (GCL). Among existing baselines, Experienced Replay (ER) is an efficient model fully compliant with the GCL criterion. To this end, the authors construct their models based on ER.

[†] I am unclear on the 2-page requirement, whether it only applies to the Summary or includes later sections as well. If it only covers the Summary, great! My Section 1 is precisely two pages long. If the requirement applies to all sections, please proceed to Section 2 directly and skip all subsections under Section 1.

Preprint. Under review.

1.2 DER and DER++ Objectives, Oversimplified

The Dark Experience Replay (DER) and DER++ models employ a buffer $\mathcal{M}$, built using reservoir sampling, to store data points and logits. DER's objective includes the lost of current task as well as the L^2 distances between previous and current outputs for data points stored in the buffer. DER++, in addition, contains a regularization term penalizing predictions that deviate significantly from ground truth. To sum up, the following depicts the objective function of DER++. In particular, if $\beta = 0$, the objective reduces to that of DER.

$$\text{DER++ Objective} = \underbrace{\mathcal{L}}_{\substack{\text{loss on}\\\text{current task}}} + \alpha \cdot \underbrace{\mathbb{E}_{(x',z')\sim\mathcal{M}}[\|z' - h_\theta(x')\|_2^2]}_{\text{consistency with previous output}} + \beta \underbrace{\mathbb{E}_{(x'',y'',z'')\sim\mathcal{M}}[\ell(y'', f_\theta(x''))]}_{\text{approximation of ground truth}} .$$

1.3 Evaluations Settings

The authors of the paper evaluate their novel General Continual Learning (GCL) models, Dark Experience Replay (DER) and DER++, using benchmarks from three subcategories of Continual Learning: Task-IL, Domain-IL, and Class-IL. They choose CIFAR-10 and Tiny ImageNet for both Task-IL and Class-IL, and they choose Permuted MNIST and Rotated MNIST for Domain-IL.

To ensure fair comparison, the authors take extra steps to standardize evaluation settings. They provide necessary information, such as task boundaries, for non-GCL-compliant methods, train their models using the same architecture as their competitor methods, set (mini)batch size constant across all methods under the same setting, and apply Stochastic Gradient Descent for all optimizations.

In addition, the authors propose a new protocol, MNIST-360, which they claim to be the first GCL-compliant protocol that includes both discrete (change in digit class) and continuous (rotation) changes in data distribution. They evaluate several selected models on MNIST-360 to showcase their models' effectiveness.

1.4 Findings

Performance. The results of the comparison suggest that both DER and DER++ exhibit strong performance across most settings. The authors find DER and DER++ to outperform traditional regularization-based methods and other non-rehearsal-based methods, including Knowledge Distillation and architectural. When compared to rehearsal-based methods, DER and DER++ still show leading performance, with the exception being Task-IL, where DER is on par with ER (the method on which DER is based). In the newly proposed MNIST-360 protocol, DER and DER++ maintain their dominant performance.

Convergence behavior. In terms of convergence behavior, the authors empirically show that DER converges to flatter minima compared to similar models by inspecting the empirical Fisher Information Matrix. This means that DER and DER++ have more freedom to explore neighboring regions and potentially reach a new minimum without drastically increasing the loss during the process. This hypothesis is subsequently confirmed by the authors' introduction of independent Gaussian noise, which show that DER and DER++ are less prone to perturbations compared to similar methods.

Calibration. The authors also find that DER and DER++ have lower calibration errors, which indicates that their methods do not suffer from the common issue of trained models being overconfident, making them more suitable for real-life applications.

Buffer informativeness. The authors note that the buffers created by DER and DER++ contain a more informative summary of the task, in the sense that DER and DER++ attain highest accuracy when one trains a new learner based solely on the buffered information.

Training time. Finally, the authors run different methods under the same hardware conditions and find the amount of time elapsed to train DER and DER++ is on par with similar methods.

2 Comments

2.1 Technical Strengths

(1) First and foremost, like the title suggests, DER and DER++ are simple to implement and strong in performance, even against other highly optimized existing methods.

(2) The authors provide thoughtful analysis regarding the efficiency of their methods, and their models' observed mathematical properties suggest new research opportunities.

(3) At the end of their abstract, the authors links a repository that stores dozens of Continual Learning models, making it convenient for readers to inspect the implementation of a particular model if they have any questions.

(4) Additionally, the proposed methods, supplemented by clear pseudocode and implementations, along with a thorough mathematical explanation, are easily reproducible and may inspire further modifications.

(5) The authors demonstrate strong writing skills, with a natural flow of writing, appropriate use of technical terms, and a well-structured and self-contained paper.

2.2 Technical Weaknesses

(1) The categorization of the three Continual Learning settings is first mentioned in Section 1 (Introduction), but it is not explained until later in the paper (Section 4, Experiments). This may cause confusion for general ML audiences unfamiliar with specific CL tasks.

(2) The authors state that they use reservoir sampling for memory buffers, referencing J.S. Vitter's paper, but it's not clear which of the four reservoir sampling algorithms Vitter discussed is used by DER. The answer remains unclear until one dives into the code repository and find that `utils.buffer.reservoir` uses the naïve implementation.

(3) While the authors have conducted extensive comparisons among various models, it is not clear — and labor intensive to verify — whether the implementations of all methods are correct and properly optimized. To address this, a comparison between the original performance of other models and the reproduced performance listed in Table 2 could be added.

(4) The MNIST-360 dataset provides essential features for GCL, but its practicality is limited as real-world digit streams don't usually come in the format of consecutive digits being consecutive numbers, handwritten digits can contain extra or missing strokes, and so on. MNIST-360 does not address these issues. That being said, MNIST-360 nevertheless will have its own research potential.

(5) The authors do not explicitly state the assumptions on which DER and DER++ apply, nor do they discuss the statistical properties of their methods. (I come from a background of reading theorems or statements that are rigorously preceded by explicitly stated conditions, so naturally when reading this paper, I looked for similar statements too. But perhaps this is not what experiment-oriented ML papers focus on.)

3 Additional Questions

Q: What can you see yourself contributing to a paper like this?

Personally, I find this question equivalent to "what aspects of this paper are you *unable* to contribute to?" One definite answer for the latter question is to come up with DER in the first place, or more generally, having the acuteness to identify potential research topics by reading existing literature. This unfortunately intimately ties with one's accumulation of research experience, and I am afraid there is no easy shortcut. On a lighter note, I also struggle with summarizing, writing abstracts, conclusions, and coming up with a suitable title, so please, don't leave those to me!

Once a solid research idea has been proposed, I see myself as a good fit in writing any specific sections of the paper, especially Section 3, where I can describe the mathematical model in detail, and Section 5, where I can analyze the properties of DER, as my strengths lie in mathematical reasoning. With sufficient preparation, I am confident that I will also be capable of writing clear and comprehensive Background and Experimental Setting paragraphs.

Q: Can you see yourself working on a project that would result in publishing a similar paper?

Yes, absolutely. I am planning on pursuing a Ph.D. after finishing my undergraduate studies, so with probability $(1 - \epsilon)$ I will (have to) end up producing a similar paper in finite time. This potential opportunity to work as a research assistant now provides me with an unparalleled chance to gain valuable experience in Ph.D.-level research years ahead of my schedule. If I am lucky enough to be selected, I am committed to making the most of this opportunity, fully investing myself and contributing to the group to the best of my abilities.