Adaptive Memory Replay for Continual Learning
-Supplementary Materials (Appendix)-

James Seale Smith[*] [1,2] **Lazar Valkov**[1] **Shaunak Halbe**[2] **Vyshnavi Gutta**[2]
Rogerio Feris[1] **Zolt Kira**[2] **Leonid Karlinsky**[1]
[1]MIT-IBM Watson AI Lab [2]Georgia Institute of Technology

A. Method

This section shows how to express the CL objective (Eq. 1) in terms of the amount of forgetting. To start off, for task T, we denote the optimal parameters found on the previous task as θ^*_{T-1}. Then, we define the forgetting for some parameter on some example to be positive if the loss on that example has increased: $\mathcal{F}(x;\theta) = \mathcal{L}(x;\theta) - \mathcal{L}(x;\theta^*_{T-1})$. Starting from our objective in Eq. 1, we write:

$$
\begin{aligned}
&\min_\theta \left[\sum_{x \in X_T} \frac{L(x;\theta)}{|X_T|} + \sum_{t=1}^{T-1} \sum_{x \in X_t} \frac{L(x;\theta)}{|X_t|} \right] \\
&= \min_\theta \left[\sum_{x \in X_T} \frac{L(x;\theta)}{|X_T|} \right. \\
&\qquad \left. + \sum_{t=1}^{T-1} \sum_{x \in X_t} \frac{L(x;\theta) - L(x;\theta^*_{T-1}) + L(x;\theta^*_{T-1})}{|X_t|} \right] \\
&= \min_\theta \left[\sum_{x \in X_T} \frac{L(x;\theta)}{|X_T|} \right. \\
&\qquad + \sum_{t=1}^{T-1} \sum_{x \in X_t} \frac{L(x;\theta) - L(x;\theta^*_{T-1})}{|X_t|} \\
&\qquad \left. + \sum_{t=1}^{T-1} \sum_{x \in X_t} L(x;\theta^*_{T-1}) \right] \\
&= \min_\theta \left[\sum_{x \in X_T} \frac{L(x;\theta)}{|X_T|} + \sum_{t=1}^{T-1} \sum_{x \in X_t} \frac{\mathcal{F}(x;\theta)}{|X_t|} + C \right]
\end{aligned}
$$

Finally, we note that when minimizing the forgetting $\mathcal{F}(x;\theta) = \mathcal{L}(x;\theta) - \mathcal{L}(x;\theta^*_{T-1})$, only only needs to compute and minimize the loss on the new task $\mathcal{L}(x;\theta)$, since $\mathcal{L}(x;\theta^*_{T-1})$ is a fixed value. Therefore, we can optimize $\mathcal{F}$ without introducing extra computational demands to our training process.

[*]Work done during internship at MIT-IBM Watson AI Lab.

B. On Regularization Losses

In our approach, we prioritize computational efficiency and focus on methods that do not incur additional computational costs. This decision is informed by the findings of Ghunaim *et al.* [4], who demonstrate that both simple and advanced regularization-based continual learning techniques struggle to perform effectively under computational budget constraints. Moreover, their research suggests that simple experience replay is a more effective strategy in such scenarios. Thus, when extending such computational considerations to the setting of extended continual pre-training, we focus on *outperforming iid experience replay without introducing any additional computational costs.* Furthermore, we consider gains of our approach to be orthogonal to the realms of non-replay regularization-based continual learning methods, and thus our method could potentially be integrated with these regularization techniques to enhance overall performance, offering a synergistic effect.

C. Expanded Implementation Details

We use A100 GPUs to generate all results. The hyperparameters for our experiments were meticulously chosen based on a series of small task experiments in which we use only used half of the number of tasks. We update our model on $10,000$ new data examples per task. In the interest of computational resources for the larger Llama model, we approximate the training of all the model parameters with LoRA finetuning [5] in the language modeling experiments. In our experience, conclusions attained for LoRA finetuning reflect the same in full model training. We use a learning rate of $2e-5$ for full model fine-tuning and $2e-4$ for LoRA-based fine-tuning. For LoRA-based fine-tuning, we use a rank of 8 for the Llama model experiments. For our proposed adaptive memory replay bandit scheme, we found that a temperature of $t = 0.1$ and forgetting mean update ratio of $\beta = 0.01$ performed best. We compose our replay batches for both iid replay and our adaptive memory replay with a 1:1 ratio of replay data to new task training data. We conducted evaluations on a hold-out test dataset comprising

500 samples per dataset. We used a batch size of 128 and 16 for the Masked Autoencoder and Llama models, respectively, which was chosen based on GPU memory. For the Llama experiments, we leveraged low-precision training.

D. Expanded Benchmark Details

In our main text, we evaluated the Masked Autoencoder model for three vision datasets. The first dataset is the DomainNet [7] dataset, containing 6 different domains of common objects. The next is the Medical MNIST dataset [11], from which we sampled 5 standardized biomedical image datasets containing the highest number of samples. Finally, we use 4 attribute splits from the Synthetic Visual Concepts (SyViC) dataset [3].

For the Llama model, we benchmarked on a 5-dataset sequence using datasets from Huggingface [10]. The datasets involved in this sequence were *banking77* [2], *wikicat-sum/animal* [8], *bigbio/hallmarks-of-cancer* [1], *bigpatent* [9], and *wikitext* [6].

References

[1] Simon Baker, Ilona Silins, Yufan Guo, Imran Ali, Johan H"ogberg, Ulla Stenius, and Anna Korhonen. Automatic semantic classification of scientific literature according to the hallmarks of cancer. *Bioinform.*, 32(3):432–440, 2016. 2

[2] Iñigo Casanueva, Tadas Temcinas, Daniela Gerz, Matthew Henderson, and Ivan Vulic. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on NLP for ConvAI - ACL 2020*, 2020. Data available at https://github.com/PolyAI-LDN/task-specific-datasets. 2

[3] Paola Cascante-Bonilla, Khaled Shehada, James Seale Smith, Sivan Doveh, Donghyun Kim, Rameswar Panda, Gul Varol, Aude Oliva, Vicente Ordonez, Rogerio Feris, et al. Going beyond nouns with vision & language models using synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20155–20165, 2023. 2

[4] Yasir Ghunaim, Adel Bibi, Kumail Alhamoud, Motasem Alfarra, Hasan Abed Al Kader Hammoud, Ameya Prabhu, Philip HS Torr, and Bernard Ghanem. Real-time evaluation in online continual learning: A new hope. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11888–11897, 2023. 1

[5] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 1

[6] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models, 2016. 2

[7] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1406–1415, 2019. 2

[8] Laura Perez-Beltrachini, Yang Liu, and Mirella Lapata. Generating summaries with topic templates and structured convolutional decoders. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, 2019. Association for Computational Linguistics. 2

[9] Eva Sharma, Chen Li, and Lu Wang. BIGPATENT: A large-scale dataset for abstractive and coherent summarization. *CoRR*, abs/1906.03741, 2019. 2

[10] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019. 2

[11] Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023. 2