ConSlide: Asynchronous Hierarchical Interaction Transformer with Breakup-Reorganize Rehearsal for Continual Whole Slide Image Analysis ———Supplementary Material———

Yanyan Huang[1,2*], Weiqin Zhao[1*], Shujun Wang[3], Yu Fu[2], Yuming Jiang[4], Lequan Yu[1†]
[1]The University of Hong Kong
[2]Zhejiang University
[3]The Hong Kong Polytechnic University
[4]Stanford University
{yanyanh, wqzhao98}@connect.hku.hk, shu-jun.wang@polyu.edu.hk,
yufu1994@zju.edu.cn, ymjiang2@stanford.edu, lqyu@hku.hk

Overview. In the supplementary material, we elaborate on the details of metrics and experimental settings. Specifically, we demonstrate the detailed definitions of AUC, ACC, Masked ACC, BWT, and Forgetting. Then, we provide more experimental details about comparison with other WSI analysis approaches and comparison with other continual learning approaches.

1. More Information about Metrics

In this section, we provide more details about the metrics we used in our experiments.

AUC. The Area Under the receiver operating characteristic Curve (AUC) is an important metric to evaluate the performance of medical image analysis and WSI analysis models. In our WSI continual learning setting, we evaluate the performance after the final task conducted. Specifically, we compute the AUC of each class against the rest [13, 7], with OvR (*i.e.*, One vs Rest) strategy.

ACC. The Accuracy (ACC) metric we used in this project is traditional multi-class ACC metric, and we also evaluate ACC after the final task is conducted.

Masked ACC. The above AUC and ACC metrics only evaluate the classification performance around all classes, and cannot measure the performance in single task. Therefore, we also applied Masked ACC as a evaluation metric in our work. The Masked ACC is calculated by masking task-irrelevant categories from different datasets, which can also reflect the performance of continual learning on the task-incremental scenario.

BWT. Backward transfer (BWT) is the influence of learning a new task t has on a previous task k ($k < t$) [11]. After the model finishes learning about the task t, we evaluate its test performance on all T tasks, and we can construct the matrix $R \in \mathbb{R}^{T\times T}$, where $R_{i,j}$ is the test classification accuracy of the model on task j after observing the last sample from task i. The BWT can be calculated from the constructed $R \in \mathbb{R}^{T\times T}$ as:

$$BWT = \frac{1}{T-1}\sum_{i=1}^{T-1} R_{T,i} - R_{i,i} \tag{1}$$

Forgetting. Forgetting measures how much an algorithm forgets what is learned in the past [3]. And it is defined as the difference between the maximum knowledge gained about the task throughout the learning process in the past and the knowledge the model currently has about it:

$$Forgetting = \frac{1}{T-1}\sum_{i=1}^{T-1} \max_{t\in\{1,\cdot,T-1\}} R_{t,i} - R_{T-1,i} \tag{2}$$

Larger AUC, ACC, Masked ACC, and BWT, and lower Forgetting indicate better performance of the model.

2. Additional Experimental Details

In this section, we demonstrate more detailed settings in our experiments.

Details of Comparison with Other WSI Analysis Approaches. As introduced in the paper, we first evaluate the performance of the proposed HIT model by comparing with other start-of-the-art WSI analysis models. We set up four tasks/datasets for this project, and each task/dataset contains two classes. We merge these four datasets and conduct a eight-class classification task in this comparison scenario.

*These authors contributed equally to this work.
†Corresponding Author.

CL Type	Method	Buffer size	Avg. AUC (↑)	Avg. ACC (↑)	Avg. Masked ACC (↑)
Baselines	JointTrain (Upper)	-	-	-	-
	Finetune (Lower)		0.786 ± 0.011	0.485 ± 0.018	0.824 ± 0.020
Regularization based	LwF [10]	-	0.801 ± 0.019	0.461 ± 0.009	0.836 ± 0.017
	EWC [9]		0.802 ± 0.012	0.480 ± 0.013	0.809 ± 0.017
Rehearsal based	GDumb [12]	5 WSIs	-	-	-
	ER-ACE [2]		0.843 ± 0.019	0.523 ± 0.032	0.805 ± 0.023
	A-GEM [4]		0.850 ± 0.025	0.529 ± 0.034	0.810 ± 0.040
	DER++ [1]		0.852 ± 0.027	0.540 ± 0.041	0.814 ± 0.028
	ConSlide w/o BuRo		0.883 ± 0.009	0.569 ± 0.020	0.864 ± 0.017
	ConSlide	1100 regions (≈ 5 WSIs)	**0.926 ± 0.019**	**0.677 ± 0.026**	**0.865 ± 0.026**
	GDumb [12]	10 WSIs	-	-	-
	ER-ACE [2]		0.866 ± 0.009	0.581 ± 0.029	0.817 ± 0.026
	A-GEM [4]		0.879 ± 0.012	0.562 ± 0.028	0.853 ± 0.011
	DER++ [1]		0.874 ± 0.012	0.606 ± 0.033	0.856 ± 0.021
	ConSlide w/o BuRo		0.901 ± 0.009	0.626 ± 0.019	0.861 ± 0.019
	ConSlide	2200 regions (≈ 10 WSIs)	**0.938 ± 0.010**	**0.711 ± 0.023**	**0.867 ± 0.015**
	GDumb [12]	30 WSIs	-	-	-
	ER-ACE [2]		0.910 ± 0.007	0.686 ± 0.024	0.832 ± 0.017
	A-GEM [4]		0.913 ± 0.009	0.616 ± 0.056	0.863 ± 0.019
	DER++ [1]		0.922 ± 0.017	0.715 ± 0.029	0.873 ± 0.022
	ConSlide w/o BuRo		0.942 ± 0.007	**0.732 ± 0.030**	**0.878 ± 0.019**
	ConSlide	6600 regions (≈ 30 WSIs)	**0.942 ± 0.011**	0.729 ± 0.018	0.871 ± 0.014

Table 1. Comparison of average results among different continual learning methods. The best performances are shown in **bold**.

Besides, the original HIPT model [5] are conducted under the minimal patch size set as 16×16, and it has a three-layer hierarchical structure. To make a fair comparison with proposed HIT, we re-implement HIPT with the minimal patch size of 512×512 (extracted features with pre-trained CNN feature extractor) and with a two-layer hierarchical structure (*i.e.*, patch- and region-level respectively) similar to HIT.

Details of Comparison with Other Continual Learning Approaches. In the paper, we compare our proposed ConSlide with several state-of-the-art continual learning approaches. However, previous continual learning methods are conducted on natural images by using CNN encoder (*e.g.*, ResNet [8]), and directly applying them on WSI data is not feasible. Therefore, we reproduce several powerful baselines with the proposed HIT as backbone to realize a fair comparison. Besides, for the rehearsal-based methods, we save the patch- and region-level features to the buffer respectively, for the ease of WSI replay. The reproduced approaches include: regularization-based methods LwF [10] and EWC [9], and rehearsal-based methods GDumb [12], ER-ACE [2], A-GEM [4], and DER++ [1].

3. Additional Experimental Results

In the paper, we report the comparison results with other continual learning approaches in Table 5, and the results are calculated on all datasets after the final task is conducted (*i.e.*, evaluate eight-class classification performance). We notice that in some works [6, 14, 15], the average results of models at different time steps are also reported (*i.e.*, evaluate the performances of two-, four-, six-, and eight-class classification tasks respectively and calculate the mean value of them). So we also follow these works and report the average results in Table 1.

As the JointTrain and GDumb are conducted once in the whole process, we only report the last results in the paper and there are no averaged results. It is observed that the regularization-based methods still perform worse than rehearsal-based methods, and only have a little improvement over the lower-bound baseline in average AUC metric. Compared with DER++ under different buffer size, the proposed ConSlide w/o BuRo can consistently achieve 3.0%, 2.7%, and 2.0% improvements in Avg. AUC, 2.9%, 2.0%, and 1.7% improvements in Avg. ACC, and 5.0%, 0.5%, and 0.5% improvements in Avg. Masked ACC. By incorporating the BuRo module, ConSlide can further boost the average performance, especially under small buffer sizes.

References

[1] Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline. *Advances in neural information processing systems*, 33:15920–15930, 2020. 2

[2] Lucas Caccia, Rahaf Aljundi, Nader Asadi, Tinne Tuytelaars, Joelle Pineau, and Eugene Belilovsky. New insights

on reducing abrupt representation change in online continual learning. In *International Conference on Learning Representations*, 2022. 2

[3] Arslan Chaudhry, Puneet K Dokania, Thalaiyasingam Ajanthan, and Philip HS Torr. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 532–547, 2018. 1

[4] Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with a-gem. In *International Conference on Learning Representations*, 2019. 2

[5] Richard J Chen, Chengkuan Chen, Yicong Li, Tiffany Y Chen, Andrew D Trister, Rahul G Krishnan, and Faisal Mahmood. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16144–16155, 2022. 2

[6] Arthur Douillard, Alexandre Ramé, Guillaume Couairon, and Matthieu Cord. Dytox: Transformers for continual learning with dynamic token expansion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9285–9295, 2022. 2

[7] Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006. 1

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[9] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017. 2

[10] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017. 2

[11] David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30, 2017. 1

[12] Ameya Prabhu, Philip HS Torr, and Puneet K Dokania. Gdumb: A simple approach that questions our progress in continual learning. In *European conference on computer vision*, pages 524–540. Springer, 2020. 2

[13] Foster Provost and Pedro Domingos. Well-trained pets: Improving probability estimation trees. *Raport instytutowy IS-00-04, Stern School of Business, New York University*, 2000. 1

[14] Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, et al. Dualprompt: Complementary prompting for rehearsal-free continual learning. *arXiv preprint arXiv:2204.04799*, 2022. 2

[15] Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 139–149, 2022. 2