ConStruct-VL: Data-Free Continual Structured VL Concepts Learning -Supplementary Materials (Appendix)-

James Seale Smith*[1,2] Paola Cascante-Bonilla[1,3] Assaf Arbelle[4]
Donghyun Kim[1,4] Rameswar Panda[1,4] David Cox[1,4] Diyi Yang[5]
Zsolt Kira[2] Rogerio Feris[1,4] Leonid Karlinsky*[1,4]
[1]MIT-IBM Watson AI Lab [2]Georgia Institute of Technology [3]Rice University
[4]IBM Research [5]Stanford University

A. Discussion on Choice of BLIP

We chose to base our method on BLIP as it had highest out-of-the-box performance (as a pre-trained model) on the ConStruct-VL tasks (Table 1) compared to numerous VL models including the very recent CyCLIP [4], thus making it a good representative source model for CL on ConStruct-VL. We also evaluated the out-of-the-box performance on ConStruct-VL tasks using METER [3], X-VLM [7], VLMO [1], and FIBER [2], and observed an average performance of 56.8%, 58.9%, 54.6%, and 73.9% respectively (21.0%, 18.9%, 23.2, and 3.9% below out-of-the-box BLIP), which further demonstrates the difficulty of VL models to understand VL concepts. As our approach is orthogonal to continued improvements in VL, we note that a great future direction is to explore future improved VL models with our approach on ConStruct-VL.

B. Details on Prompting Baselines

In Section 4 (Experiments), we discuss our PyTorch implementations of the very recent and influential L2P [6] and DualPrompt [5] works which are state-of-the-art (SOTA) data-free visual continual learning (CL) prompting-based methods. In our PyTorch implementation, we rigorously followed the description and the JAX code of L2P and DualPrompt. Furthermore, we tuned their hyperparameters for ConStruct-VL by maximizing their performance on the same 3 task sequence of ConStruct-VL as for all of the compared methods, including all the baselines and our own approach (Sec. 4). In this section, we provide additional details on these L2P and DualPrompt baselines.

L2P and DualPrompt work by learning a key-value paired prompt pool based on an instance-wise query mechanism. For L2P, we use a prompt size of 4, prompt pool size of 50, and choose the 5 closest prompts from the pool at a time. For DualPrompt, we use a prompt length of 20 for the 'expert' prompts, and a prompt length of 6 for the 'general' prompts. Importantly, these hyperparameters were tuned in the same manner as for all other compared methods in our paper by maximizing performance on the same 3 tasks sequence of ConStruct-VL (starting from the hyperparameters recommended in the original papers [5, 6]). We also searched *where* to insert prompts. Whereas originally L2P has prompting in layer 1 only, and DualPrompt has 'general' prompts in layers 1,2 and 'expert prompts' in layers 3,4,5; through tuning L2P and DualPrompt on ConStruct-VL, we found that adding prompts in every layer of the model for both methods (i.e., layers 1-12 for L2P and layers 3-12 for DualPrompt 'expert' prompts) maximizes their ConStruct-VL performance.

We note that the under-performance of these methods on the proposed ConStruct-VL benchmark (Tab. 1, Tab. 2a, Tab. 2b). is likely an indication that the proposed problem of multi-modal continual learning of SVLCs in ConStruct-VL is challenging to the vision-only CL SOTA and is thus an exciting new CL goal, which we just started to explore in the current work.

References

[1] Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. VLMo: Unified vision-language pre-training with mixture-of-modality-experts. 2022. 1

[2] Zi-Yi* Dou, Aishwarya* Kamath, Zhe* Gan, Pengchuan Zhang, Jianfeng Wang, Linjie Li, Zicheng Liu, Ce Liu, Yann LeCun, Nanyun Peng, Jianfeng Gao, and Lijuan Wang. Coarse-to-fine vision-language pre-training with fusion in the backbone. In *NeurIPS*, 2022. 1

[3] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, Zicheng Liu, and Michael Zeng. An empirical study of training end-to-end vision-and-language transformers. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1

*Equal contribution

[4] Shashank Goel, Hritik Bansal, Sumit Bhatia, Ryan A Rossi, Vishwa Vinay, and Aditya Grover. Cyclip: Cyclic contrastive language-image pretraining. *arXiv preprint arXiv:2205.14459*, 2022. 1

[5] 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. 1

[6] 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. 1

[7] Yan Zeng, Xinsong Zhang, and Hang Li. Multi-grained vision language pre-training: Aligning texts with visual concepts. *arXiv preprint arXiv:2111.08276*, 2021. 1