Architectural Analysis of Vision Transformers in Continual Learning

August 4, 2024

Deep neural networks (DNN) deployed in the real world are frequently exposed to non-stationary data distributions and required to sequentially learn multiple tasks. This requires that DNNs acquire new knowledge while retaining previously obtained knowledge and this is imperative in applications like autonomous driving and robotics. However, continual learning in DNNs, in which networks are trained in a sequence of tasks, results in catastrophic forgetting of previously learned information. Therefore, to combat this, a variety of approaches have been proposed for convolutional neural networks (CNN) [ASZ22, BZA22, SAZ22].

On the other hand, the recent breakthrough of vision transformers (VTs) and their compelling performance in different vision tasks present them as an alternative architectural paradigm. Due to their global receptive field, self-attention-based transformer architectures have a unique advantage over CNNs in terms of robustness and generalizability [JKV+22]. However, VTs struggle in the low training data regime [TCD+21]. Thus, various architectural modifications have been proposed to incorporate convolutional biases and increase data efficiency in VTs [WXZ+22, WXC+21, dTL+21, YCW+21].

Most of the research has focused on developing effective training methodologies to combat catastrophic forgetting in CNN and VT. However, only a few works have investigated the effect of architectural choices on lifelong learning. Moreover, these studies only investigate the effect of CNN architecture choices, such as batchnorm layer and network depth, on catastrophic forgetting [MCY+22, PLH22]. However, works exploring the effect of architectural choices of VTs in continual learning performance is almost non-existent. VTs architectures are results of different architectural design choices including attention mechanism, positional embedding, token embedding, and class token. Therefore, the objective of this study would be to conduct a comprehensive study on the impact of architectural choices in VTs on various aspects of continual learning, including catastrophic forgetting, plasticity to learn new tasks, and task-recency bias, etc.

References

[ASZ22] Elahe Arani, Fahad Sarfraz, and Bahram Zonooz. Learning fast, learning slow: A general continual learning method based on complementary learning system. *arXiv preprint arXiv:2201.12604*, 2022.

[BZA22] Prashant Bhat, Bahram Zonooz, and Elahe Arani. Task agnostic representation consolidation: a self-supervised based continual learning approach. *arXiv preprint arXiv:2207.06267*, 2022.

[dTL+21] Stéphane d'Ascoli, Hugo Touvron, Matthew L Leavitt, Ari S Morcos, Giulio Biroli, and Levent Sagun. Convit: Improving vision transformers with soft convolutional inductive biases. In *International Conference on Machine Learning*, pages 2286–2296. PMLR, 2021.

[JKV+22] Kishaan Jeeveswaran., Senthilkumar Kathiresan., Arnav Varma., Omar Magdy., Bahram Zonooz., and Elahe Arani. A comprehensive study of vision transformers on dense prediction tasks. In *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 4: VISAPP,*, pages 213–223. INSTICC, SciTePress, 2022.

[MCY+22] Seyed Iman Mirzadeh, Arslan Chaudhry, Dong Yin, Huiyi Hu, Razvan Pascanu, Dilan Gorur, and Mehrdad Farajtabar. Wide neural networks forget less catastrophically. In *International Conference on Machine Learning*, pages 15699–15717. PMLR, 2022.

[PLH22] Quang Pham, Chenghao Liu, and Steven Hoi. Continual normalization: Rethinking batch normalization for online continual learning. *arXiv preprint arXiv:2203.16102*, 2022.

[SAZ22] Fahad Sarfraz, Elahe Arani, and Bahram Zonooz. Synergy between synaptic consolidation and experience replay for general continual learning. *arXiv preprint arXiv:2206.04016*, 2022.

[TCD+21] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021.

[WXC+21] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22–31, 2021.

[WXZ+22] Cong Wang, Hongmin Xu, Xiong Zhang, Li Wang, Zhitong Zheng, and Haifeng Liu. Convolutional embedding makes hierarchical vision transformer stronger. *arXiv preprint arXiv:2207.13317*, 2022.

[YCW+21] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 558–567, 2021.