APP: Anytime Progressive Pruning

Diganta Misra, Bharat Runwal,
Tianlong Chen, Zhangyang Wang,
Irina Rish
{diganta, bharat}@landskape.ai

Motivation and Problem Statement

Although many methods have been investigated for optimal learning settings in scenarios where the data stream is continuous over time, sparse networks training in such settings have often been overlooked. In this paper, we aim to answer the following question:

"Given a dense neural network and a target sparsity, what should be the optimal way of pruning the model in ALMA setting?"

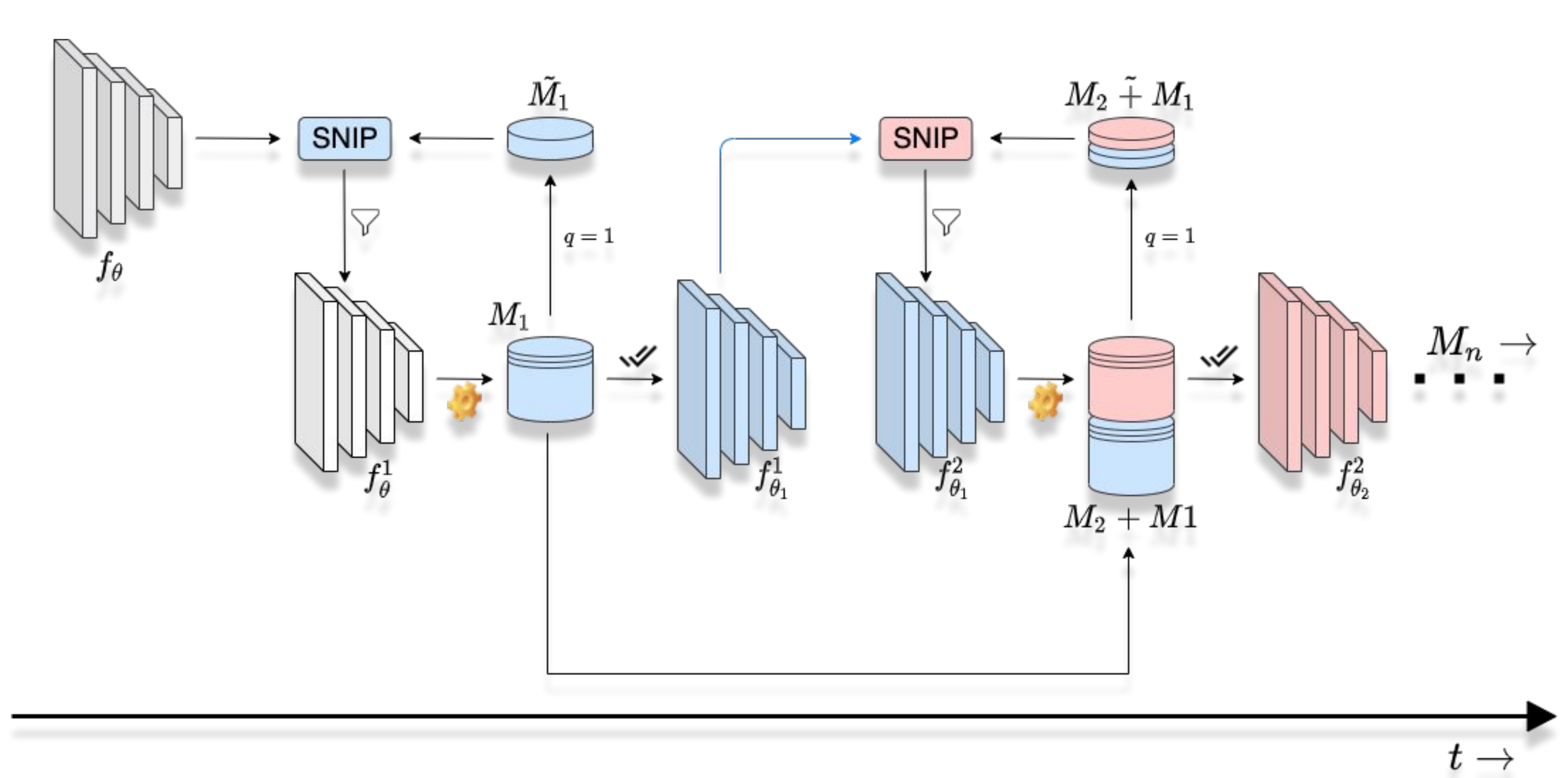
Solution: **Anytime Progressive Pruning (APP)**

Our Contributions:

- We provide the first comprehensive study into pruning of deep neural networks in an ALMA setting.
- We therefore propose a novel approach of progressively pruning dense neural networks in the ALMA paradigm, which we term Anytime Progressive Pruning(APP).
- We further investigate the training dynamics of APP as compared to the baselines in ALMA

Anytime Progressive Pruning: Algorithm

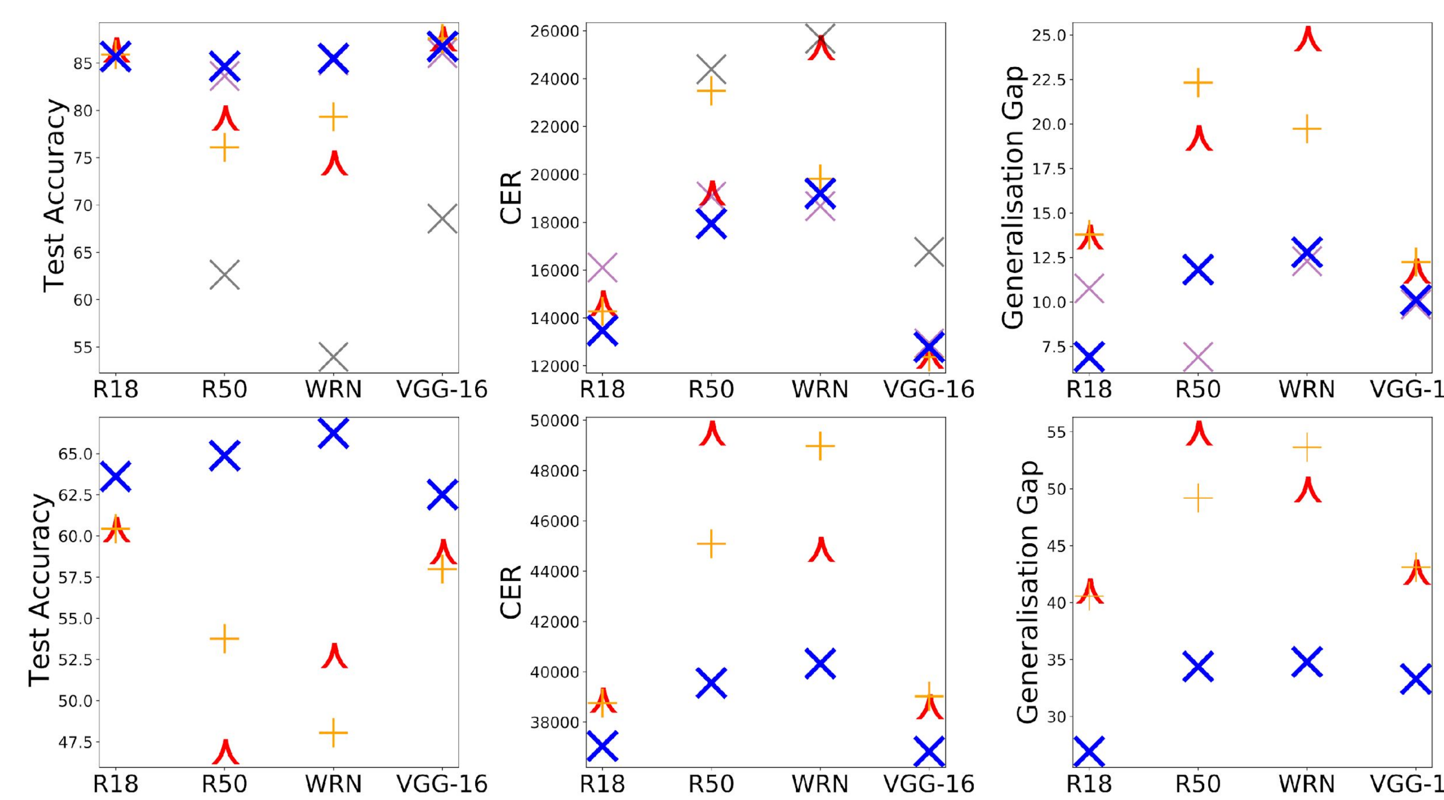
For each megabatch $M_t \in S_B$, we construct the replay inclusive megabatch M_t by taking the union of all previous megabatches along with the current megabatch and then create a small sample set π_t of size $0.2 * |M_t|$ to be used to prune the model to $0.8 * \delta_t \times 100\%$ sparsity. Here, δ_t is obtained from a predetermined list δ of uniformly spaced values that denote the target sparsity levels for each megabatch in the stream S_B. After pruning the model, we train it on the Mt megabatch and evaluate it on a holdout test set.



Framework design of APP

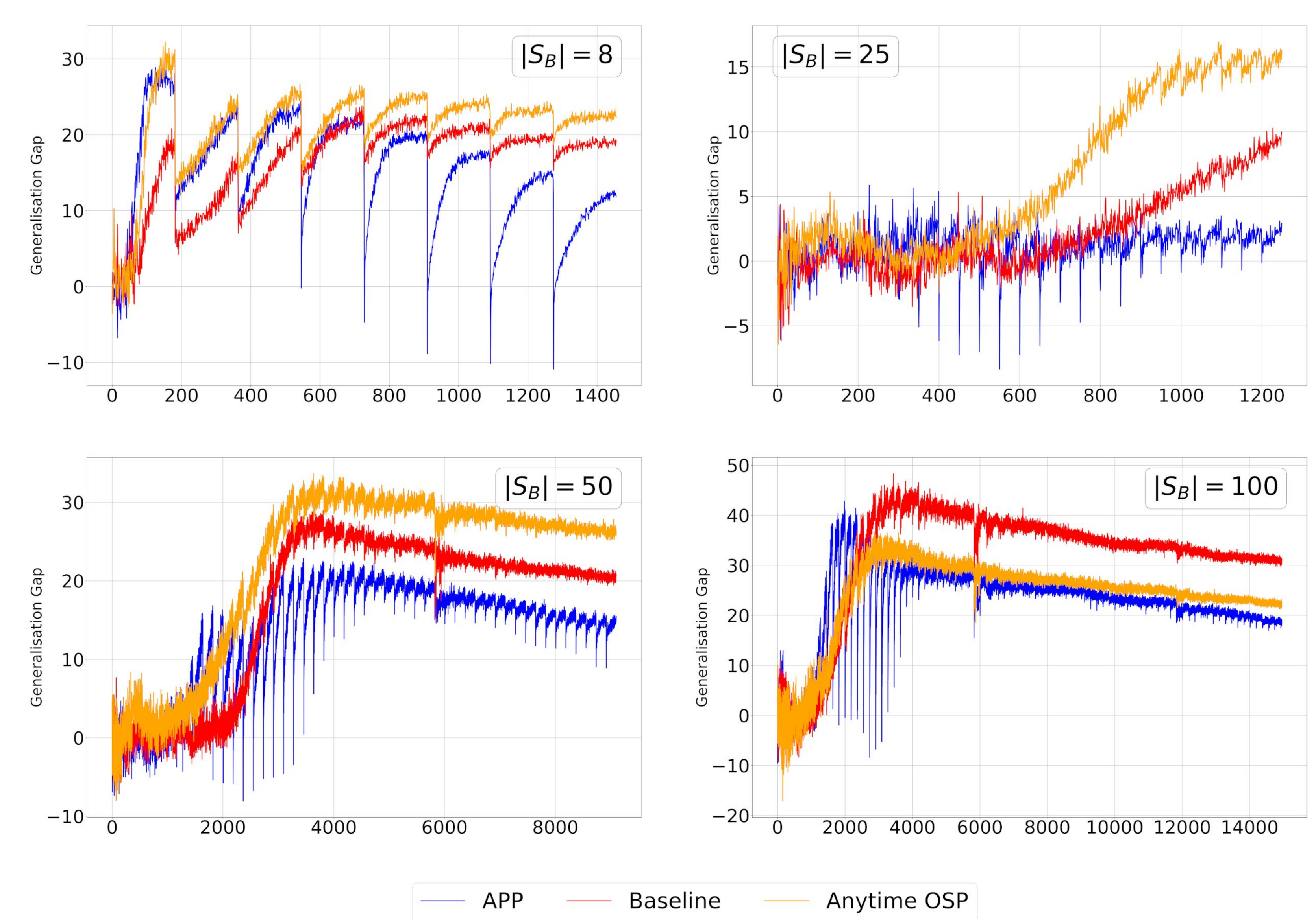
Experimental Analysis

Results on short sequence ALMA for CIFAR-10 & 100:

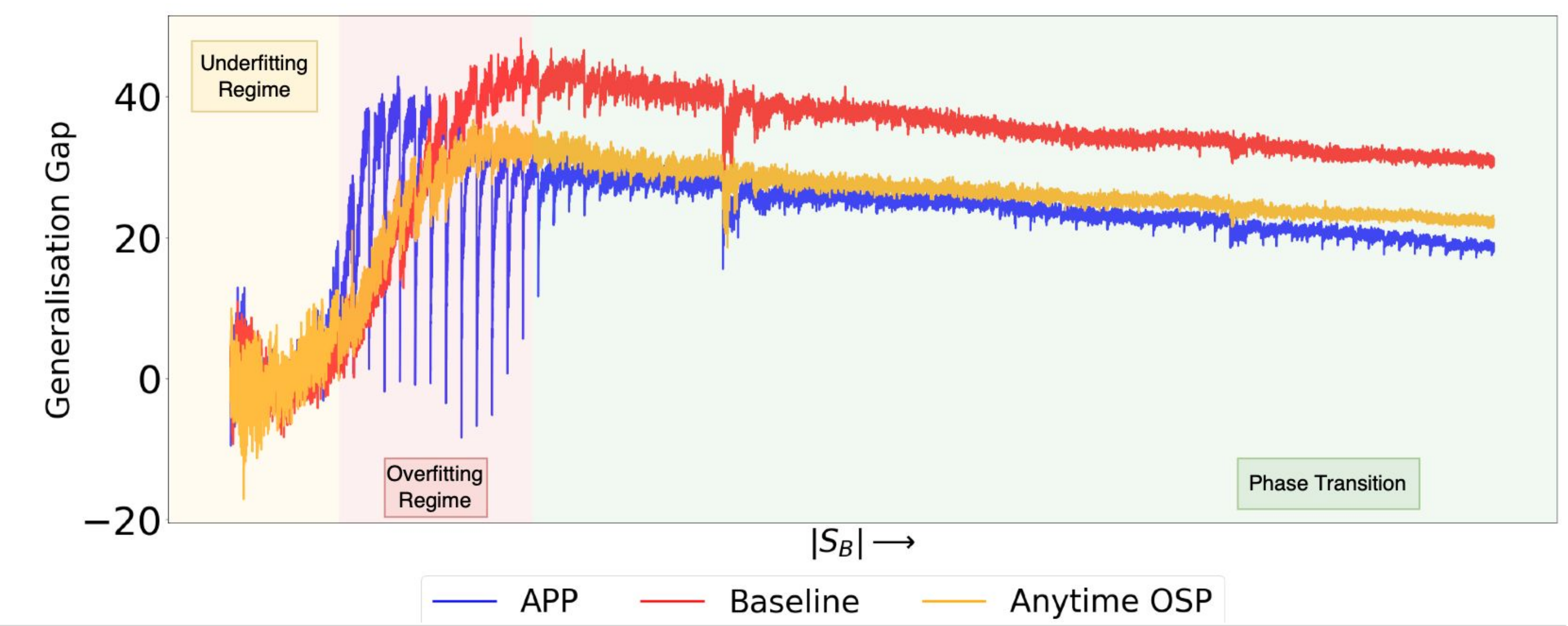


With ResNet-50 (R50) in C-100, APP improved the test accuracy by 17.97% and 11.12%, reduced CER by 9927 and 5533 and decreased the generalisation gap by 20.49% and 14.79% compared to baseline and Anytime OSP

Training dynamics analysis



Profiling non-monotonic generalisation gap curves under long sequence ALMA

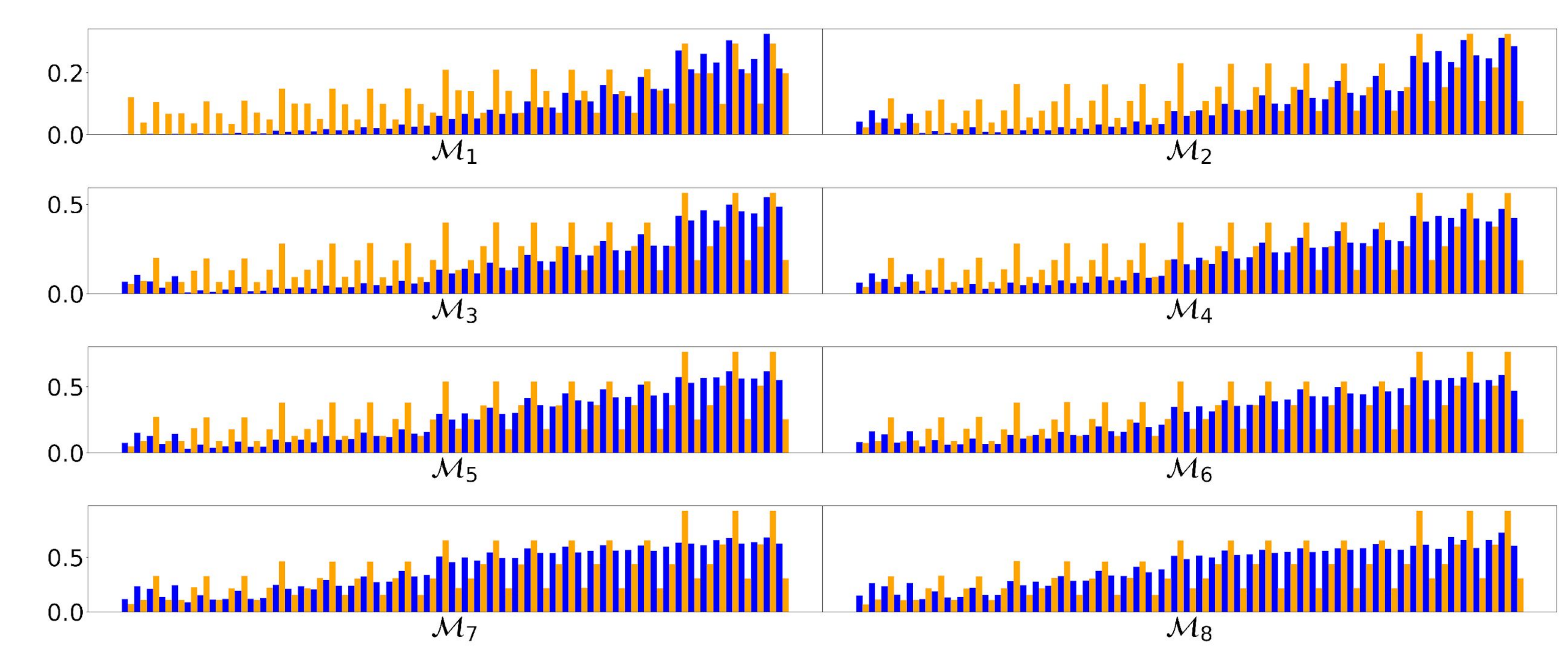


Generalisation gap as a function of number of megabatches (100) for CIFAR-10 using ResNet-50 architecture.

Ablation and future prospects

Analysis into effect of pruners

We analysed the effect of SNIP vs Magnitude Pruning for APP under 8 megabatch setting for CIFAR-10 and observed that magnitude pruning aggressively prunes earlier layers at initial megabatches as compared to SNIP.



References

- Lucas Caccia, Jing Xu, Myle Ott, Marc'Aurelio Ranzato, and Ludovic Denoyer. On anytime learning at macroscale. arXiv preprint arXiv:2106.09563, 2021.
- Tianlong Chen, Zhenyu Zhang, Sijia Liu, Shiyu Chang, and Zhangyang Wang. Long live the lottery: The existence of winning tickets in lifelong learning. In International Conference on Learning Representations, 2020.
- Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. arXiv preprint arXiv:1803.03635, 2018.