A weight-averaging approach to speeding up model training on resource-constrained devices
Samarth Tripathi, Jiayi (Jason) Liu, **Unmesh Kurup**, Mohak Shah

Training machine learning models on edge devices has definite advantages for security, privacy and latency. However, techniques such as Deep Neural Networks (DNNs) are unsuitable given the resource constraints on such devices. Optimizing DNNs is especially challenging due to the nonconvex nature of their loss function. While gradient-based methods that use back-propagation have been crucial to neural network adoption, optimal convergence of the loss function is still time-consuming, volatile, and needs many finely tuned hyperparameters. One key hyperparameter is the learning rate. A high learning rate can produce fast results faster but at the increased risk of the model never converging.

In this paper we show that by manipulating the model weights directly using their distributions over batch-wise updates, we can achieve significant intermediate improvements in training convergence, and add more robustness to the optimization process with negligible cost of additional training time. More importantly, this approach allows deep neural networks to be trained at higher than usual learning rates resulting in fewer epochs which reduces resource use and allows for lower total training time.

We introduce a trio of techniques (PSWA, PWALKS, and PSWM) centered around periodic sampling of model weights that provide consistent and more robust convergence on gradient update methods (vanilla SGD, Momentum, Adam) for a variety of vision problems (classification, detection, segmentation). Our techniques use existing optimal training policies but converge in a less volatile fashion with performance improvements that are approximately monotonic. We conduct a variety of experiments to quantify these improvements and identify scenarios where each of these techniques could be more useful.