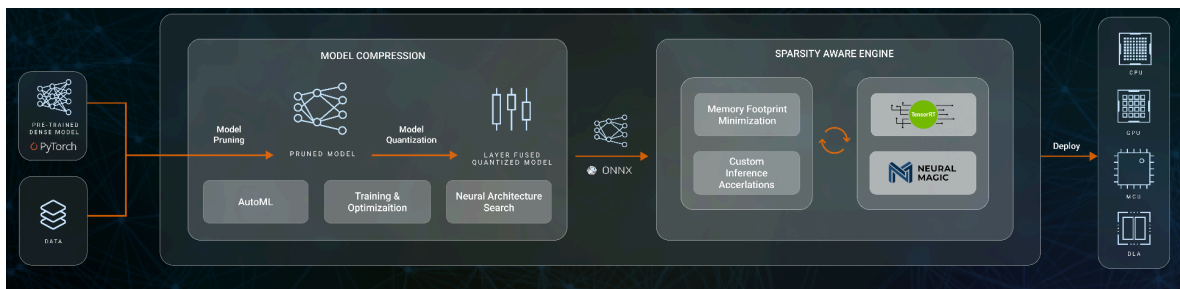


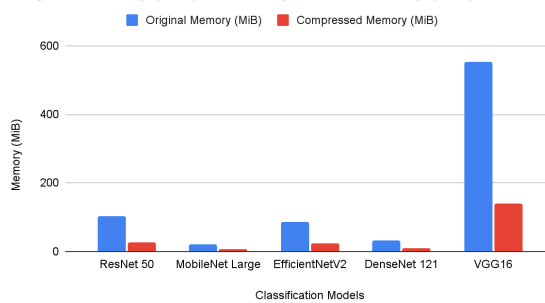
Sconce: An End-to-End AutoML Framework for Model Compression through Multi-Stage Pruning and Quantization

Kien Pham, Adam Chestovaliev, Ryan Albright, Sathyaprakash Narayanan, Golam Md Muktadir

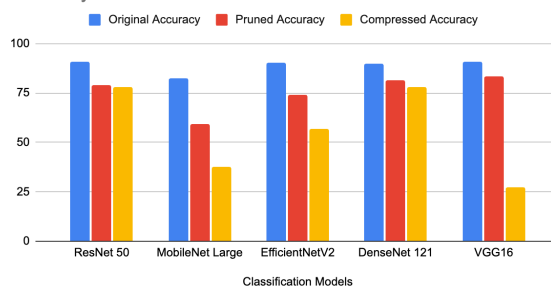
The rapid advancement of deep learning has been primarily driven by the availability of large datasets and the computational power to train large models. The size grows exponentially with each passing year, and it is becoming increasingly difficult to train and deploy these models. To address this bottleneck, we introduce Sconce, an AutoML package designed to compress a model end-to-end with minimal human intervention. Sconce employs a multi-stage compression pipeline integrating channel-wise pruning, granular magnitude pruning, quantization, and sparsity engines to eliminate redundant parameters while preserving model accuracy. Across six convoluted neural networks, Sconce reduced memory by 72-75% and parameters by 84-94%, with only a marginal 5-10% accuracy degradation post-compression. Furthermore, fine-tuning the compressed models using NVIDIA CUDA GPU and Inference Acceleration yielded a 2.8x reduction in inference latency. Sconce thus provides an efficient, hardware-friendly solution for deploying large deep learning models without substantial accuracy trade-offs.



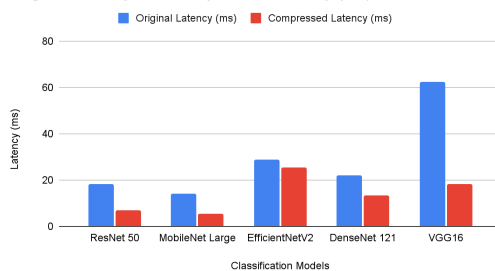
Original Memory (MiB) and Compressed Memory (MiB)



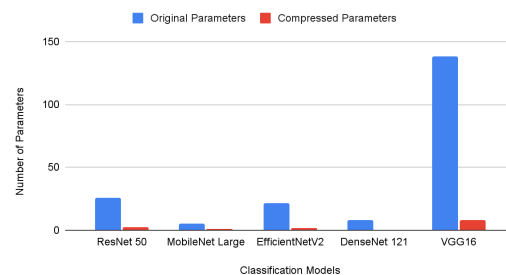
Original Accuracy, Pruned Accuracy and Compressed Accuracy



Original Latency and Compressed Latency (ms)



Number of Original Parameters and Compressed Parameters



Sconce Package Python Code, Implementation into Classification/Large Language Models, and other UCSC SIP-related research/results can be found [here](#).