ABSTRACT

Over the last decade, we have witnessed immense growth in the usage of IoT devices. These devices have limited storage and computational capacities and are heavily dependent on cloud platforms. The ever increasing size of deep neural networks makes them infeasible to run on these IoT devices. Efficient compression of machine learning models helps improve the latency of these IoT devices by reducing their dependence on the cloud. This has led to an increasing interest in the research of model compression techniques. We first review a few existing compression techniques which aim to reduce the space requirements of convolutional layers and the fully connected layers. We also propose a clustering-based approach to prune the fully connected layers of a pre-trained neural network. Our approach is data-independent and solely depends on the weights of the trained model. We evaluate our algorithm on a LeNet architecture [1] using the MNIST dataset [2]. Our algorithm performs better by a small margin in most cases compared to the existing approaches. We observe a maximum additive benefit of 2.5% and an average additive benefit of 0.5% in the accuracy drop.