

In this paper we present a novel and general method to accelerate convolutional neural network (CNN) inference by taking advantage of feature map sparsity. V