This work studies the model compression for deep convolutional neural networks (CNNs) via filter pruning. The workflow of a traditional pruning consists of thre