Structured pruning of filters or neurons has received increased focus for compressing convolutional neural networks. Most existing methods rely on multi-stage