Abstract

Computer vision was researched since the late 1960's but image classification is still a challenging task. After Geoffrey Hinton won the ImageNet Large Scale Vision Recognition Challenge(ILSVRC)[1] as known as ImageNet with AlexNet[2] on 2012, people started researching about Convolutional Neural Networks(CNN) for image classification. Modern neural networks[3, 4] achieved nearly 90% accuracy on the ImageNet dataset, however the number of parameters are tremendously large. The most popularly used CNN models are VGG16[5], InceptionV3[6], and ResNet18[7], which have 138M, 24M, 11M parameters respectively. Involution[8] successfully reduced the number of parameters of CNNs by replacing all the 3×3 convolution kernels with involution kernels, which use 1×1 convolution for the convolution layer's kernel generation. In result, involution networks achieved similar performance with 34.1% fewer parameters. Although the models introduced above have great performance on image tasks, they still require high computational cost therefore applying deep neural networks on mobile devices remain challenging.

There were several approaches for reducing the number of parameters trying not to lose the performance. Knowledge distillation[9] is a method which use a dense network as a teacher model, and a sparse network as a student model. The term distillation means that the student model learns the soft label of the teacher model. By learning both hard label which is the loss function of the prediction and the ground truth, and the soft label which is the loss between the prediction of the student and the teacher model, the student model can mimic the output of the teacher model. If properly trained, the student model will act similarly with the teacher model with fewer parameters. Raphael Gontijo Lopes et al. proposed a knowledge distillation algorithm which does not require data when training the student model by reconstructing the input image using the activation statistics and layer gradients. Gongfan Fang et al. improved the data free distillation algorithm by training an image generator for data reconstruction.

Filter pruning[10] is a method for reducing the number of filters in CNN. When pruning the filters we sort by the sum of weights for each filter for every layer. Since filters that have weights that are close to 0 will not affect much of the performance of the model, therefore when given a proper threshold we can get a sparse model with a similar performance.

However, it is hard to apply the conventional pruning method for involution since it requires sorting the filters. Involution has reshaping layers therefore if the filters are sorted, they lose the spatial information. To overcome this problem we need to rewrite the code for the model which is hard to implement and time consuming. In this research we propose a pruning method called the model diet which is easy to implement, and effective for CNN models including involution. Instead of sorting the filters for each layer, we reduce a certain portion of the filters by grouping the kernel weights therefore for involution, the spatial information is not lost. Since the model depth is maintained but the filters are reduced, we call this pruning method a model diet, and we will show that diet models have faster convergence compared with randomly initialized models.

The model diet is consisted of 2 stages, the kernel grouping stage and the group selection stage. kernel grouping is an algorithm that splits the kernel weights into groups. The kernel weights are split in order therefore when applied to involution, the involution kernel does not lose the spatial information. Once the kernel weights are split into groups, we take the sum of the weights for each group. Then we use the group that has the biggest sum ans we call this operation group selection.

Deep learning frameworks such as Tensorflow or Pytorch save the model's weights as matrices. For involution the kernel weights are saved as a vector. When the weights are loaded, the vector reshapes itself into the corresponding shape. Therefore the element of the vector indicates a certain location in an image. If we apply conventional pruning algorithms, the weights of the involution kernel will be sorted also, resulting a loss of spatial information. However the kernel grouping keeps the order of the weights, therefore when applied to involution the loss of the spatial information does not happen. Also the computational complexity of the model diet is $\mathcal{O}(n)$ where the computational complexity of the conventional pruning is $\mathcal{O}(n \log n)$. Since model diet has the same computational complexity with selecting the maximum element in a vector where conventional methods have the same computational complexity with sorting.

In this research we show the effectiveness of the model diet in 3 vision tasks, image classification, image segmentation, and depth estimation. We test the performance of the diet model and the random initialized model. The diet model showed faster convergence and performance compared with the random initialized model. For image segmentation the dataset was easy to generalize and lacked difficulty therefore the diet model and the random initialized model showed equal performance but the diet model had faster convergence. For depth estimation both the diet model and the random initialized model showed poor performance even though the loss converged. Also the difference between groups was studied. We split the full model into 2 groups and compared the performance. Both the group with the bigger sum and smaller sum showed equal performance and speed of convergence. Since pruning can be regarded as weight initialization, we hypothesize that both the group with bigger sum and smaller sum started from a different location, but shared the same local optimum point.