摘要

近年来，以深度神经网络为代表的人工智能浪潮极大地推动了计算机视觉领域的迅猛发展。然而，为了获得更好的性能，人们越来越倾向于收集更多的数据并且设计更为复杂的网络结构。随之而来的问题是模型复杂度急剧上升，直观的表现为参数越来越多、计算量越来越大。这使得深度神经网络的存储和计算成本变得更高，在手机、机器人、车载等移动端设备上的应用受到了极大的限制。模型压缩是一种有效降低深度神经网络存储和计算量的方案，通过修剪网络中不重要的参数或者使用更少的比特表示等方法，减少模型参数的存储并且降低网络计算带来的能耗。本文围绕着挖掘神经网络压缩潜能的目标，系统地进行了深度神经网络压缩方法的研究，从层内自适应、层间自适应、样本自适应三个角度进行探索，减少神经网络的冗余，在此基础上对神经网络模型进行压缩和加速。本文的主要创新点包括：

第一，提出了一种层内自适应的深度神经网络频域动态剪枝方法，利用权重参数的频域稀疏性减少深度神经网络的冗余。该方法首先使用离散余弦变换将深度神经网络的卷积核变换到频域，根据离散余弦变换的线性变换性质将卷积操作在频域中进行了等效实现。其次，基于不同频带系数对于神经网络的贡献不同，该方法提出对不同频带系数设置不同的剪枝率。在此基础上，提出了一种可以在剪枝过程中将神经元重新连接回来的动态算法，以防止剪枝过程中的误删操作。在国际基准数据集 CIFAR-10 和 ImageNet 上的实验结果表明，该方法在模型精度相当的情况下压缩率优于 DNS 等国际前沿方法。

第二，提出了一种层间自适应的深度神经网络混合比特量化方法，通过对不同网络层设置不同的量化比特宽度降低深度神经网络的冗余。首先，该方法对不同网络层的权重值和激活值分别进行编码，然后将不同编码组成的比特串（即不同量化方案的量化模型）视为一个个体，所有的个体会构成一个种群。这样，该方法将混合比特量化任务建模成一个遗传问题，并通过进化算法来得到最优的量化模型。为了提升进化算法的效率，该方法设计了针对量化模型的个体适应度度量函数，平衡量化模型的精度和压缩率。在国际基准数据集 ImageNet 上的实验结果表明，该方法在相同的压缩率下精度优于 HAQ、HAWQ 等国际前沿方法。

第三，提出了一种层间自适应的 Transformer 后训练量化方法，不需要重训练就可以对深度神经网络进行压缩。该方法将后训练量化任务建模成寻找权重值和激活值的最优量化步长问题，通过迭代搜索的方式获取最优的量化模型。优化目标是最大化量化前后的中间特征相似度以及最小化注意力层的排序损失，其中排序损失函数的引入

是为了更好地保持 Transformer 模型中注意力层的功能性。在此基础上，该方法会根据注意力层特征和多层感知机层特征的核范数来决定相应网络层的量化比特宽度。实验结果表明，该方法 8 比特量化模型的精度损失仅为 0.1%，优于 EasyQuant、BitSplit 等国际前沿方法。

第四，提出了一种样本自适应的深度神经网络量化方法，通过对不同样本分配不同的计算量来进一步挖掘神经网络压缩的潜能。给定一个网络结构，该方法会生成大量不同量化方案的隐藏子网络。在推理过程中，难以被准确识别的样本会被分配到一个较高计算资源的子网络，反之亦然。为了不增加额外的计算成本，该方法设计了一个轻量级的比特控制器来预测输入样本的难易程度，并据此分配各个网络层权重值和激活值的量化比特宽度。在国际基准数据集 ImageNet 上的实验结果表明，该方法在模型精度相当的条件下比 PACT 等国际前沿方法可以节省 30% 以上的计算量。

综上所述，本文系统地研究了深度神经网络压缩方法，通过对网络层内的不同频带、不同网络层、不同样本进行自适应压缩，可以更大程度地降低神经网络的冗余。本文所提的方法可以在图像分类、目标检测等任务上进行应用，为深度神经网络在移动端和嵌入式等设备上的应用奠定了基础。

关键词：深度神经网络压缩，频域动态剪枝，混合比特量化，后训练量化，样本自适应量化