Motivation:

随着深度学习技术的快速发展，深度学习模型在不断刷新任务精度极限的同时，其深度和尺寸也在成倍增长。目前，最先进的深度学习模型通常具有数百万个需要存储的参数，而设备上的内存是有限的。但是当这些模型开始由研究步入到具体场景应用时，一方面为了保护数据数据安全保留用户隐私，另一方面，为减少用户感知的查询时间同时更加充分的利用移动端和边缘设备的计算能力，需要将深度神经网络提供的智能服务迁移到移动边缘计算设备是非常有必要的。而面对规模巨大的模型直接移植到移动端和嵌入式芯片当中使用会对设备有很高的性能要求，大尺寸的模型也对设备功耗和运行速度带来了巨大的挑战，较高的带宽占用也让很多用户望而生畏。为了更好的将模型部署到移动端和边缘设备，模型压缩是一种必要的手段。我们通过对经典的深度神经网络模型进行压缩实践，分析不同方法对于不同模型减小模型尺寸大小和推理时间的效果，进一步分析不同模型在不同方法下的性能表现。我们主要使用了量化和剪裁对模型进行压缩，在量化中，我们测试了10个模型,包括了mobilenet，vggnet, inception net, resnet四种不同结构的网络，通过量化，将模型大小平均压缩为原来的25.5%左右，推理时间普遍增加，但不同模型增加的幅度不同。进一步可以证明量化方法对于不同的模型有不同的效果，其中inception的推理时间增加最多，结果最不好。Vggnet时间增加最少，表现最优。在剪裁中，我们共剪裁三种模型包括Vgg16，inception v3 ,Resnet50。通过剪裁，vgg16剪裁后大小减少8%,准确率下降3%，推理时间减少36%。Resnet50通过剪裁后，大小减少18%，准确率下降4%，推理时间减少20%。可以看到Vgg16的时间减少最明显，但大小变化较小；Resnet50的时间变化和参数量的变化几乎一致；inception v3的准确率损失过大，剪裁效果不好。