1 Background:

剪裁：

剪裁的主要思想是：通过对神经网络的节点重要性进行评估，然后进行移除剪枝。其核心步骤是首先进行神经元重要性评估，确定可以移除的对象；然后进行移除操作；最后对剪枝后的神经网络进行平滑操作的从而剪枝过程损失的精度。对神经网络权重进行剪枝的方法定义权重连接重要与否是通过检查权重的绝对值大小实现。当权重绝对值大于设定的阈值，即认为重要的连接不可以移除;权重绝对值小于设定的阈值是，即认为是不重要的连接，可以移除。目前这种方法的基本流程是：训练-剪枝-重训练。

rank the neurons in the network according to how much they contribute, you could then remove the low ranking neurons from the network, resulting in a smaller and faster network.

The idea of pruning is that among the many parameters in the network, some are redundant and don’t contribute a lot to the output.

The idea of pruning is to evaluate the importance of each parameter in the network, and remove the ones that do not contribute a lot to the output.

The idea of pruning is that some parameters in the network are redundant and can be removed, resulting in a smaller and faster network. The pruning is a three stage iterative process: Prune / Train / Repeat. Firstly, evaluating the importance of each neuron and ranks the neurons according to how much they contribute, and removing the least importance one; after the pruning, the accuracy will drop, so the pruned network need to be retrained to regain accuracy; the pruned network could be damaged if too many parameters are pruned at once, so we pruned a few neurons and repeat the this process.

Pruning is a three-stage approach, evaluating the importance of each neuron and ranks the neurons according to how much they contribute; removing the least importance one that do not contribute a lot to the output; retraining the pruned networks to regain accuracy. Furthermore, the pruning is an iterative process: Prune / Train / Repeat, because it could be damaged if too many parameter are pruned at once.

The core of pruning is that removing the low ranking neurons from the network, resulting in a smaller and faster network.

thdetermines the ones that can be removed; performing the re

The main idea of tailoring is to evaluate the importance of the nodes of the neural network and then remove the pruning. The core steps are to first evaluate the importance of neurons, determine the objects that can be removed, and then perform the removal operation; finally, the smoothing operation of the pruned neural network and the accuracy of the pruning process loss. The method of pruning the weight of the neural network defines whether the weight connection is important or not by checking the absolute value of the weight. When the absolute value of the weight is greater than the set threshold, the connection considered to be important cannot be removed; the absolute value of the weight is less than the set threshold, that is, the connection considered to be unimportant can be removed. The basic process of this method is currently: training - pruning - heavy training.

量化：

量化主要是用更少的数据位对神经网络存储和计算，主要的优点是减少模型存储空间。

其主要的思想是：原始用4个字节来表示一个浮点数，量化后使用1个字节来定点化原始数据，即将一个浮点数据转换为0-255的整型数据。这是因为由于神经元只抽取必要信息并给出响应，这样整个网络是抗噪和鲁棒的。训练完成后进行推断时，减低精度的定点化量化也可以考虑为噪声，因此不会对模型精度产生大影响。网络参数是按层组织，每层数值都在同一数量级，有大量论文研究表明确认值最大和最小后每层数据使用8bit定点化量化已可以很好满足推断计算。对量化的实现是通过把常见操作转换为等价的八位达到的，涉及的操作包括卷积，矩阵乘法，激活函数，池化操作，以及拼接等。

2 movtivation：

Setup

最后我们分别使用量化和剪裁两种方法对模型进行压缩。

量化模型时我们采用的是tensorflow开源的量化工具，量化步骤：首先在源码下编译tensorflow的量化工具；然后将ckpt模型进行成pb，从graph中读取模型的输入输出节点；最后对模型使用工具进行8bit量化。在使用量化工具时需要根据要求输入模型的具体参数，例如输入输出节点，图片的尺寸等参数。

剪裁我们采用的主要方法是根据权值绝对值的大小评估神经元的重要性，然后进行剪裁，具体的步骤：首先计算每个神经元权值绝对值的大小；然后设定百分比移除权值的百分比，将权值绝对值小的神经元进行移除；最后重新训练剪裁后的模型恢复精度。在剪裁CNN模型时，需要对每一层模型进行上述的步骤，然后根据每一层的精度损失确定剪裁的阈值，最终完成整个模型的剪裁。

Result:

补充解释量化时间增大的原因：

根据实验结果可以看到：量化的平均精度损失在2%左右，这主要是由于神经网络的健壮性，同时可以减少模型的存储空间75%，但是推理时间增大。通过使用tensorboard查看量化前后的pb发现，为了保证输出层的输入数据的准确性，量化后的graph里加入了含有转换函数的子图，把8位转回32位作为输出层的输入，导致模型的推理时间增大。