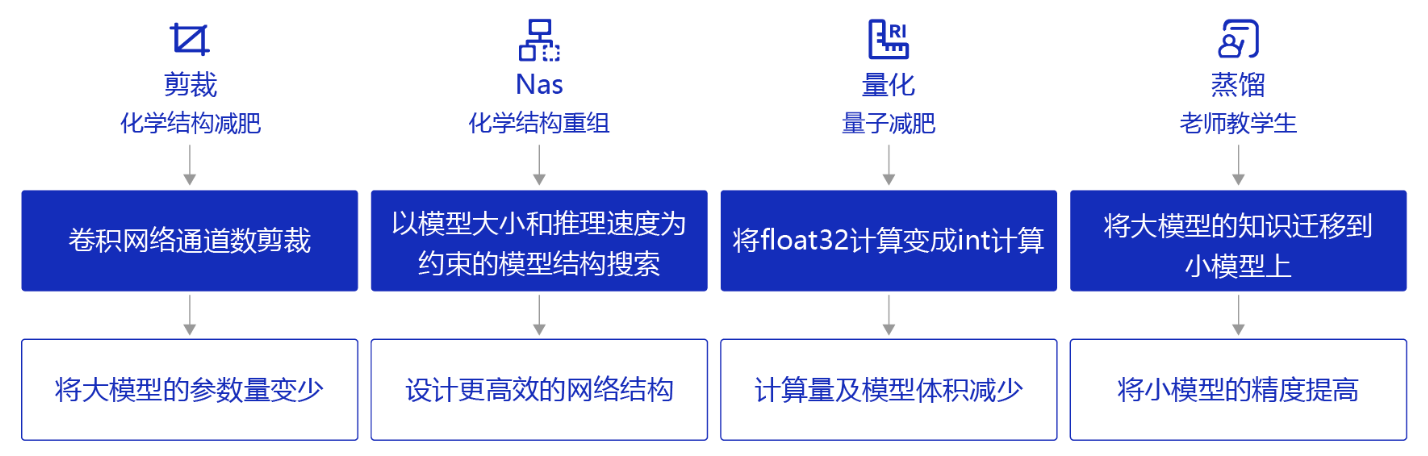
模型压缩报告



主要的模型压缩方法有四种：

1. 剪枝（Parameter Pruning）：

1. 方法描述：这种方法通过移除模型中不重要的参数（例如权重接近零的连接）来减小模型大小。剪枝后的模型需要更少的计算，需要重新训练以恢复性能。

2. 代码量：相对较少

3. 训练时间：相对较短

4. 涉及的网络层：卷积层和全连接层

二、网络结构搜索（Neural Architecture Search, NAS）：

1. 方法描述：这个方法使用机器学习寻找最优的模型结构，帮助设计更小、更快和/或更有效的模型。

2. 代码量：相对较大

3. 训练时间：相对较长

4. 涉及的网络层：卷积层/卷积模块

三、量化（Quantization）：

1. 方法描述：这是一种将浮点数权重转换为低精度表示（例如二进制或三进制）的方法。这可以显著减少模型大小、加速推理，但可能会牺牲一些预测精度。

2. 代码量：相对较少。主要是对模型参数进行位数截断或编码压缩等操作。

3. 训练时间：相对较短

4. 涉及的网络层：卷积、全连接、激活、BN层等。

四、知识蒸馏（Knowledge Distillation）：

1. 方法描述：属于迁移学习的一种，主要思想是将学习能力强的复杂教师模型中的“知识”迁移到简单的学生模型中。

2. 代码量：相对较大。需要构建教师模型和学生模型，并定义相应的损失函数。

3. 训练时间：相对较长，需要训练教师模型和学生模型。

4. 涉及的网络层：卷积层和全连接层。

实验：在MNIST数据集上使用量化方法对MobileNet进行压缩

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Mode | Iteration | Top 1 | Top 5 | Loss |
| Train MobileNet | 900 | 0.969 | 1.0 | 1.494 |
| Test MobileNet |  | 0.926 | 0.994 |  |
| Train quant | 800 | 0.984 | 1.0 | 1.488 |
| Test quant |  | 0.957 | 0.997 |  |

原模型在900 iter时达到了top1=0.926的高精准度，

在量化重训的开始阶段，模型acc下降至top1=0.875, top5=1.0，

在重新训练到800 iter后，模型acc回到top1=0.921875, top5=1.0，

测试集acc回到top1=0.957, top5=0.997。

可见对于基于KL散度方法的INT8量化策略来说，重新训练是很重要的，能够有效恢复压缩后模型的精度。