# 压缩模型的方式

xbZhong

2025-06-26

## Contents



本页 PDF

**模型压缩三部分优化：**

1. 减少内存密集的范围量
2. 提高获取模型参数时间
3. 加速模型推理时间

## 模型剪枝（Pruning）

研究模型权重的冗余，尝试**删除/修改**冗余或者非关键权重，会**改变模型参数量**

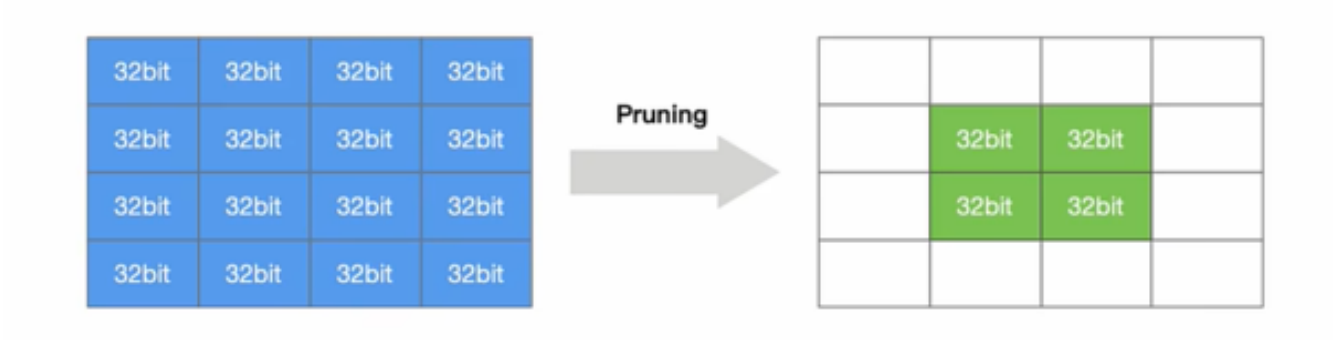


Figure 1: image-20250703160937958

**剪枝算法分类**

- **非结构化剪枝**：剪枝算法简单，模型压缩比高，**权重矩阵会稀疏**
- **结构化剪枝**：在 **channel 和 layer** 上进行剪枝，保留原始卷积结构，但算法相对复杂

**模型剪枝流程**

**常见三种方法**

1. 训练一个模型-> 对模型进行剪枝-> 对剪枝后的模型进行微调
2. 在模型训练过程中进行剪枝-> 对剪枝后的模型进行微调
3. 进行剪枝-> 从头训练剪枝后的模型

## 模型量化（Quantization）

**减少**权重表示或激活所需的**比特数**来压缩模型，也就是降低模型参数的精度，**是不改变模型参数量的**


Figure 2: image-20250703160948495

## 模型蒸馏

核心思想是通过让小型学生模型（**Student Model**）模仿大型教师模型（**Teacher Model**）的行为或知识，从而在保持较高性能的同时大幅减少模型的计算量和参数量

我第一段实习的时候做的是**知识蒸馏**，教师模型仅作**推理任务**，直接生成数据给小模型训练

#### 知识蒸馏

**教师模型指导学生模型训练**，通过**蒸馏**的方式让学生模型学习到教师模型的认识