Day1

1.引言

一、模型压缩的背景

- 随着神经网络模型越来越大，硬件负担加重，尤其是巨型模型（几十亿甚至上千亿参数）的推理和训练。
- **模型压缩**技术让模型“瘦身”，在有限硬件上高效运行。主要方法有剪枝、量化、蒸馏和架构搜索。

二、核心方法

1. 剪枝

- **目标**：找到模型中不太重要的部分，减去它们。
- **实现**：剪去冗余的神经元或连接，减少计算量，同时保持准确度。
- **优缺点**：简化模型，但可能导致架构不规则，难加速。

2. 量化

- **目标**：降低计算精度以减少模型大小。
- **实现**：用低位数（如8位）表示权重，适合内存小、计算能力低的设备。
- **优缺点**：压缩率高，但可能损失精度。

3. 知识蒸馏

- **目标**：让小模型学习大模型的“经验”，适用于小设备部署。
- **实现**：教师模型指导学生模型，以保留性能。
- **优缺点**：保持高准确率，但训练过程较复杂。

4. 神经网络架构搜索

- **目标**：通过算法寻找最优结构，代替手动设计。
- **实现**：定义搜索空间和策略，自动生成高效架构。
- **优缺点**：效果好但计算资源需求大。

三、常见的评估指标

- **准确率**：压缩后模型的任务表现是否基本保持。
- **参数量**：模型大小的直接体现，越少越节省资源。

- **模型大小**：最终文件的大小，直接影响部署。
- **推理时间**：模型处理数据所需时间，反映实时处理能力。

四、总结

模型压缩为小设备、边缘设备提供了可能性，有效平衡了模型性能与硬件资源需求。

2.CNN基础

一、CNN简介

- **CNN（卷积神经网络）**主要用于处理图像、视频等网格结构的数据，自动提取特征，特别适合图像分类、目标检测等任务。
- 核心特点：**局部连接**、**参数共享**，以及**平移不变性**（对图像中的小移动不敏感）。

二、核心组成

1. **卷积层**：
   - 通过卷积核滑动提取局部特征，生成特征图（Feature Map）。
   - 多个卷积核可以提取不同类型的特征，比如边缘、纹理。
2. **激活层**：
   - 引入非线性，常用激活函数如ReLU。
   - 使模型能学习复杂特征。
3. **池化层**：
   - 通过降低特征图的尺寸，减少参数，控制计算量。
   - 提高对特征的稳定性（如最大池化）。
4. **全连接层**：
   - 将特征图转换为输出，最终进行分类或回归。

三、常见术语

- **通道（Channel）**：图像的颜色维度（如RGB有3个通道）。
- **卷积核（Kernel）**：用于提取特征的小矩阵，滑动计算。
- **滤波器（Filter）**：由多个卷积核组成，每个通道对应一个卷积核。
- **特征图（Feature Map）**：通过滤波器提取的特征，生成二维数据。

四、总结

CNN通过层层提取特征并组合，用较少参数高效处理图像任务。在网络深层，特征越抽象，帮助模型在分类、检测中准确识别对象。