深度学习模型压缩方法及产品研究

汇报人：

2024-01-17

目录

contents

01

引言

研究背景与意义

深度学习模型压缩的重要性

随着深度学习技术的不断发展，模型规模不断扩大，导致存储和计算资源需求急剧增加。深度学习模型压缩技术的出现，对于降低模型复杂度、减少资源消耗和提高模型推理速度具有重要意义。

推动相关产品和应用的发展

深度学习模型压缩技术的研究和应用，有助于推动一系列相关产品和应用的发展，如智能手机、嵌入式设备、自动驾驶等领域的深度学习应用。这些应用对于模型的实时性、资源占用等方面有较高要求，深度学习模型压缩技术能够为其提供更好的支持。

国内外研究现状及发展趋势

国内外研究现状及发展趋势

发展趋势

未来，深度学习模型压缩技术的发展将呈现以下趋势

多方法融合

将不同压缩方法进行融合，形成优势互补，进一步提高模型压缩效果。

自适应压缩

根据模型的特点和应用需求，自适应地选择最合适的压缩方法和参数设置。

国内外研究现状及发展趋势

硬件加速优化

针对特定的硬件平台，设计专门的模型压缩算法和加速策略，以充分利用硬件资源，提高推理速度。

模型可解释性与鲁棒性增强

在模型压缩过程中，注重提高模型的可解释性和鲁棒性，以增强其在实际应用中的可靠性和安全性。

02

深度学习模型压缩方法概述

模型压缩基本概念

模型压缩定义

模型压缩是指通过一系列技术和方法，减小深度学习模型的存储空间和计算资源消耗，同时保持或尽可能减少模型性能损失的过程。

压缩对象

主要包括神经网络的权重、激活值、梯度等。

压缩目标

减小模型大小、提高计算效率、降低能耗等。

常见模型压缩方法

- 剪枝（Pruning）：通过去除神经网络中的一部分连接或神经元，减小模型大小和计算量。剪枝可分为结构化剪枝和非结构化剪枝，前者去除整个滤波器或通道，后者去除单个连接。
- 量化（Quantization）：通过降低神经网络中权重和激活值的精度（如使用8位整数代替32位浮点数），减少存储空间和计算资源消耗。量化可分为静态量化和动态量化。
- 知识蒸馏（Knowledge Distillation）：利用一个较大、性能较好的教师模型（Teacher Model）来指导一个较小、性能较差的学生模型（Student Model）的训练，使得学生模型能够学习到教师模型的知识和经验，提高性能。
- 神经架构搜索（Neural Architecture Search, NAS）：通过自动搜索神经网络的最佳结构和参数配置，找到性能优异且资源消耗较少的模型。NAS可分为基于强化学习、进化算法和梯度下降等方法。

模型压缩效果评估指标

模型大小

压缩后的模型所占用的存储空间大小，通常以MB或GB为单位。

计算量

压缩后的模型进行前向推理所需的计算资源，通常以FLOPs（浮点运算次数）或MACs（乘加运算次数）为衡量标准。

推理速度

压缩后的模型在特定硬件平台上的推理速度，通常以每秒推理的图片数量（FPS）或推理延迟（Latency）为衡量标准。

性能损失

压缩后的模型相对于原始模型的性能损失程度，通常以准确率（Accuracy）、召回率（Recall）、F1分数等指标进行评估。

03

基于剪枝的深度学习模型压缩

剪枝算法原理及分类

剪枝算法原理

通过移除神经网络中的一部分连接或神经元，减小模型大小和计算复杂度，同时尽可能地保持模型的性能。

剪枝算法分类

根据剪枝粒度可分为连接剪枝、神经元剪枝和层剪枝；根据剪枝方式可分为结构化剪枝和非结构化剪枝；根据剪枝策略可分为重要性剪枝、随机剪枝和正则化剪枝等。

经典剪枝算法介绍

Optimal Brain Damage

一种基于损失函数对模型参数进行二阶泰勒展开，通过最小化损失函数来选择需要剪掉的连接或神经元的方法。

Deep Compression

一种结合权重剪枝、量化和霍夫曼编码的深度学习模型压缩方法，可以显著减小模型大小和计算复杂度。

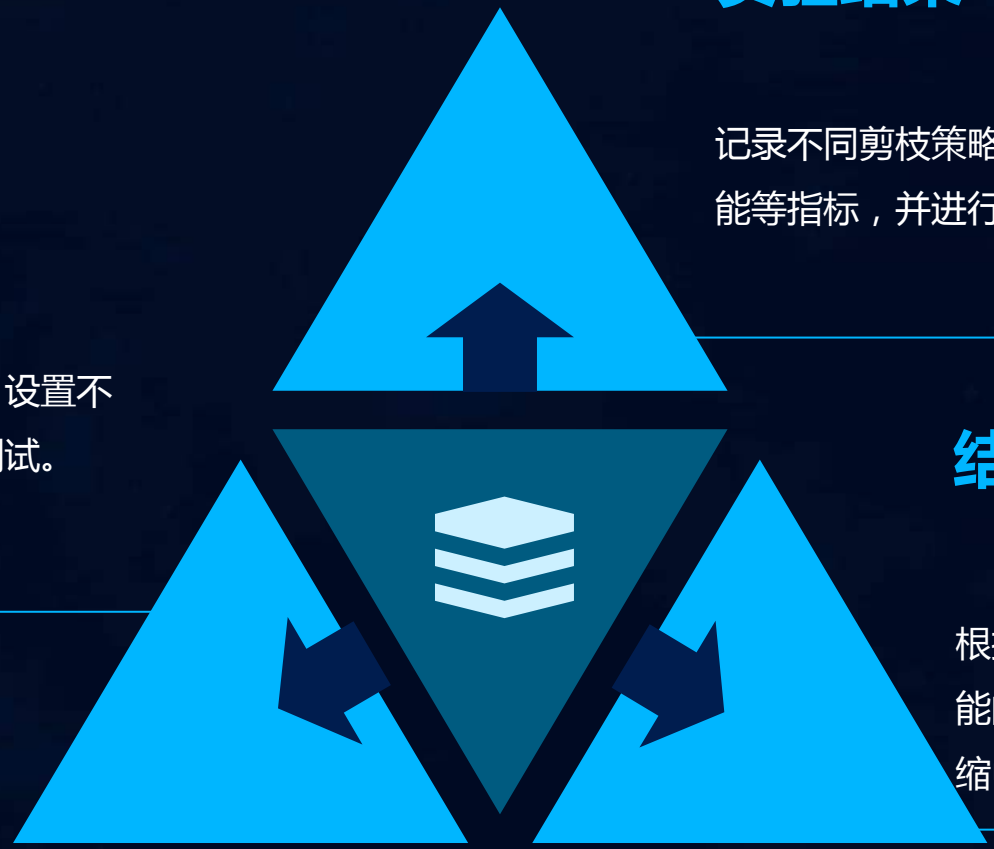
Lottery Ticket Hypothesis

一种基于迭代式剪枝和重置的方法，通过多次迭代找到一组稀疏的子网络，这些子网络在训练后可以达到与原始网络相当的性能。

基于剪枝的深度学习模型压缩实验

实验设置

选择适当的深度学习模型和数据集，设置不同的剪枝策略和参数，进行训练和测试。



实验结果

记录不同剪枝策略下的模型大小、计算复杂度和性能等指标，并进行比较和分析。

结果分析

根据实验结果，分析不同剪枝策略对模型性能的影响，探讨剪枝算法在深度学习模型压缩中的应用前景和改进方向。