

# 动态稀疏神经网络在计算机视觉中的应用

何斌<sup>1</sup>, 王欣<sup>2</sup>

(1. 中国电子科技集团公司第十五研究所, 北京 100083; 2. 武汉工程大学 邮电与信息工程学院, 武汉 430073)

**摘要:** 面向现代复杂任务场景, 深度神经网络庞大的参数总量和计算开销给模型部署带来严峻挑战。模型剪枝作为神经网络轻量化的关键技术, 现有的  $N:M$  算法仅解决了如何以统一的方式训练  $N:M$  稀疏神经网络的挑战, 但在高稀疏性条件下, 其精度明显下降。为突破传统  $N:M$  算法的性能局限, 提出了一种新的动态分层稀疏神经网络技术。该技术基于预先训练的稠密深度神经网络, 通过网络层次化处理, 每一层根据数据特征优化调整可训练的阈值向量, 并确定网络的最佳稀疏度, 从而寻找动态混合  $N:M$  稀疏性方案, 最终实现细粒度精准剪枝, 保证最大化压缩模型冗余参数的同时, 有效保留关键层级的特征提取能力。为验证方法的有效性, 在大规模 ImageNet 图像分类数据集上开展对比实验, 选取 ResNet18 与 ResNet50 两种经典计算机视觉模型作为基准, 以 top-1 分类精度为核心评价指标, 与传统剪枝方法进行性能对比。实验结果表明, 在相同高稀疏性条件下, 本文所提细粒度动态分层稀疏方案对应的基准模型 top-1 精度均显著优于传统方法, 充分验证了该方法的优越性与有效性, 在计算机视觉领域具有广阔的优势和应用前景。

**关键词:** 深度神经网络; 细粒度结构化稀疏; 计算机视觉; 模型压缩; 卷积; 剪枝

中图分类号: TP18; TP391.4

文献标志码: A

文章编号: 1003-7241(2026)10-0006-05

## Application of dynamic neural networks in computer vision

He Bin<sup>1</sup>, Wang Xin<sup>2</sup>

(1. The 15th Research Institute of China Electronics Technology Group Corporation, Beijing 100083, China;

2. College of Post and Telecommunication, Wuhan Institute of Technology, Wuhan 430073, China)

**Abstract:** In the context of modern complex task scenarios, the large parameter volume and high computational cost of deep neural networks pose significant challenges to model deployment. As a key technique for neural network lightweighting, existing  $N:M$  sparsity algorithms primarily adopt a uniform approach to training sparse networks. However, under high sparsity conditions, these methods often suffer from substantial accuracy degradation. To overcome the performance limitations of conventional  $N:M$  approaches, this paper proposes a novel dynamic hierarchical sparsity framework for neural networks. The method is built upon a pre-trained dense deep neural network and employs hierarchical optimization, whereby each layer adaptively learns and adjusts a trainable threshold vector based on data-specific features. This enables the determination of optimal, layer-wise sparsity levels and facilitates the discovery of a dynamic hybrid  $N:M$  sparsity pattern. As a result, the approach achieves fine-grained, precision-driven pruning, maximizing the compression of redundant parameters while preserving the feature extraction capabilities of critical layers. To evaluate the effectiveness of the proposed method, comparative experiments are conducted on the large-scale ImageNet image classification benchmark. Two widely used computer vision architectures, ResNet18 and ResNet50, are selected as baseline models, with top-1 classification accuracy serving as the primary evaluation metric, and performance is compared against traditional pruning techniques. Experimental results demonstrate that, under identical high-sparsity conditions, the top-1 accuracy of the models using the proposed fine-grained dynamic hierarchical sparsity scheme significantly exceeds that of conventional methods. These findings clearly validate the superiority and efficacy of the proposed approach, indicating its strong potential and broad applicability in the field of computer vision.

**Key words:** deep neural networks; fine-grained structural sparse; computer vision; model compression; convolution; pruning

深度神经网络 (deep neural network, DNN) 在计算机视觉<sup>[1]</sup>、认知科学<sup>[2]</sup>、军事智能博弈<sup>[3-5]</sup>等人工智能领域取得很大成功, 得益于更深、更宽的网络结构, 模型通过参数化来实现先进的泛化, 但参数众多的神经网络会带来计算冗余, 需要消耗巨大的计算、内存和能量资源。为了解决这些挑战, 研究人员提出了许多神经网络模型压缩技术

(network compression, NC) 来优化 DNN 模型。

最新的研究表明, 相比于知识蒸馏<sup>[6]</sup>、参数量化<sup>[7-8]</sup>、架构设计、正交量化<sup>[9]</sup>和网络结构搜索<sup>[10-12]</sup>等技术, 神经网络剪枝<sup>[13-14]</sup>是一种很有前途的模型压缩手段<sup>[15-16]</sup>。本文提出一种新的细粒度  $N:M$  稀疏技术, 使用幅值标准修剪选择阈值, 结合权重惩罚指导结果向阈值变化, 同时使

收稿日期: 2025-10-27; 录用日期: 2025-12-25

基金项目: 鄂教科函[2019]2号/2018519

作者简介: 何斌(1991—), 男, 高级工程师, 研究方向为军事智能化。

通信作者: 王欣(1985—), 女, 副教授, 研究方向为智能控制。

引用本文: 何斌, 王欣. 动态稀疏神经网络在计算机视觉中的应用[J]. 自动化技术与应用, 2026, 45(10): 6-10. (He Bin, Wang Xin. Application of dynamic neural networks in computer vision[J]. Techniques of Automation and Applications, 2026, 45(10): 6-10.)

用分层的惩罚因子来平衡参数的启发式分层冗余和分层计算复杂性,最终通过不断迭代从预训练的稠密权重中找出分层的细粒度稀疏方案。

## 1 动态稀疏神经网络

目前,硬件架构<sup>[17-18]</sup>和  $N:M$  细粒度稀疏 DNNs 剪枝算法<sup>[19]</sup>可以在保持规则稀疏结构的同时实现高稀疏比。硬件架构主要是通过对 DNN 进行特定体系结构设计提升神经网络模型的运算效率和速度,比如英伟达 Ampere 架构的 A100 Tensor Core GPU<sup>[20-21]</sup>,能够实现神经网络两倍的运算加速。但是随着深度学习模型不断增大<sup>[22-23]</sup>,在资源受限的场景中,由于能源功耗的约束<sup>[24]</sup>,无法部署足够的硬件资源,这种背景下,模型稀疏成为了神经网络压缩的重要研究方向<sup>[25]</sup>。

模型稀疏从粒度可以分为单个权重剪枝<sup>[26]</sup>、核内权重剪枝、卷积核剪枝<sup>[27]</sup>、通道剪枝<sup>[28]</sup>和隐层剪枝。

非结构化剪枝主要是对单个权重剪枝,结构化剪枝是对整个卷积核或者通道进行剪枝,不会破坏原有的卷积结构,更适合在通用的硬件设备上使用。

由于 DNN 往往具有许多网络层,冗余分布非常不均匀,基于传统的非结构化稀疏性虽然可以达到较高的稀疏比,但不能保持规则的稀疏结构;而基于传统的结构稀疏性可以达到规则的稀疏结构,但不能达到高压缩比,使得模型运行加速变得困难。细粒度结构化稀疏架构能有效的缓解这一矛盾,通过在模型权重中引入了约束获得规整的结构,最大限度地利用加速器的多核并行运算能力,是一种高模型准确率且硬件友好的平衡稀疏模式,是神经网络轻量化的重要手段。

如下图所示,分别为传统的非结构化修剪(图 1),结构化修剪(图 2)和  $N:M$  细粒度结构化修剪(图 3)对比结果(规定组大小  $M=4$ )。

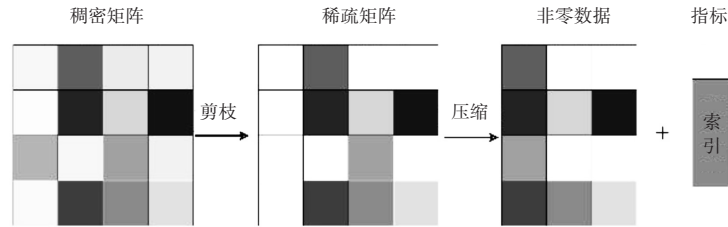


图 1 非结构化修剪

Fig. 1 Unstructured pruning

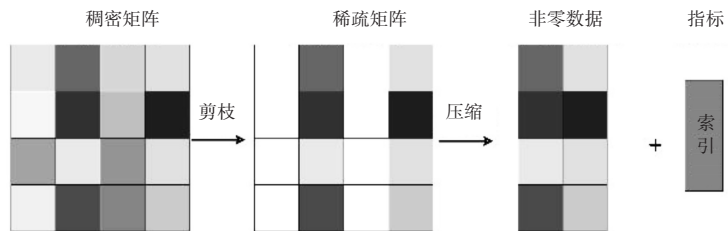


图 2 结构化修剪

Fig. 2 Structured pruning

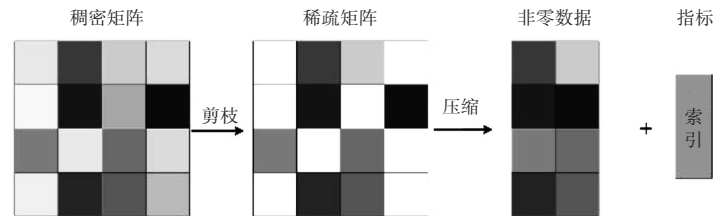


图 3  $N:M$  细粒度结构化修剪

Fig. 3  $N:M$  fine-grained structured pruning

现有的研究主要侧重于非结构化修剪,仅单独处理一层内的权重,并没有细粒度位置约束,不能用于细粒度  $N:M$  稀疏。因此,需要一种新的方法来分层细粒度  $N:M$  稀疏。

## 2 逐层稀疏的神经网络训练

### 2.1 问题定义

给定预先训练的含权向量  $\mathbf{w}^1, \dots, \mathbf{w}^i, \dots, \mathbf{w}^L$  的稠密模

型(卷积层和全连接层共  $L$  层)、模型复杂度约束  $S_{pt}$  (如模型大小或模型计算量)、稀疏方案的候选池  $N_p$  (如  $[1:8, 2:8, 4:8, 8:8]$ )。上述数学问题可以表述为

$$\begin{aligned} & \max \text{Accuracy}(S(\mathbf{w}, N)) \\ & \text{s. t. } \text{Complexity}(S, N) \leq S_{pt} \end{aligned} \quad (1)$$

$$N^i \in N_p, i = 1, 2, \dots, L$$

其中,  $S(\cdot)$  表示稀疏模型,可通过稠密权值  $\mathbf{w}$  和稀疏比  $N$

得到;  $N^i$  表示第  $i$  层稀疏配置; 复杂度  $(\cdot)$  表示模型约束;  $t$  表示阶段;  $p$  表示复杂度类型, 可以是模型大小、模型计算量、真实延迟, 甚至能源消耗, 这取决于应用实例和设计目标。最终输出是为整个神经网络找到最优的分层方案, 由于巨大的离散设计空间和评估设计点的训练/微调成本 (即恢复稀疏模型的准确性), 通常不可能人为设计最优解或穷举搜索。

基于混合位搜索方法的研究成果, 将混合位策略搜索和模型精度恢复解耦, 本文采用两步法进行求解: 1) 从候选池  $N_p$  中找到分层稀疏方案。2) 利用搜索的稀疏性 ( $N_i$ ) 对稀疏权值进行微调/再训练, 以恢复精度。

## 2.2 细粒度结构化方法流程

### 2.2.1 组内阈值和稀疏性

细粒度的  $N:M$  稀疏方法从组的角度处理权重, 如图3所示。本文定义了组阈值向量  $T^i$  和组稀疏性向量  $N_g^i$ , 其中  $i$  为层索引,  $g$  为组索引。这两个向量的长度是  $K = \frac{|w^i|}{M}$ , 其中,  $|\cdot|$  表示基数, 即参数的数量;  $M$  表示组的尺度大小;  $K$  表示组的数量;  $|w^i|$  应该确保被  $M$  整除。

### 2.2.2 基于组的阈值选择

本文使用一个阈值向量  $T^i$  来表示每个组的权重分布, 为了有效的表示组的群体特征, 同时降低离群值的影响, 选择基于绝对值最小  $\frac{M}{2}$  元素的平均值作为阈值。一方面, 使用  $\frac{M}{2}$  元素可以提供相对足够的搜索空间和代表性, 使用更多的元素, 虽然可以提供更好的代表性, 但是由于包含的元素更多, 阈值越大, 起始  $N_g^i$  越小, 搜索空间变小。如图所示, 为按组选择阈值的示例图。

$$\begin{matrix} w^i & T^i & N_g^i \\ \begin{bmatrix} 0.012 & 0.021 & 0.003 & 0.004 \\ 0.030 & 0.068 & 0.031 & 0.098 \\ 0.079 & 0.192 & 0.031 & 0.030 \\ 0.086 & 0.017 & 0.000 & 0.019 \end{bmatrix} & \rightarrow \begin{bmatrix} 0.003 \\ 0.027 \\ 0.030 \\ 0.008 \end{bmatrix} & \rightarrow \begin{bmatrix} 3 \\ 3 \\ 3 \\ 3 \end{bmatrix} \end{matrix}$$

图4 稠密权矩阵和相应阈值向量的示例

Fig. 4 Example of the dense weight matrix and the corresponding threshold vector

另一方面, 分组阈值将导致分组稀疏性, 列向量  $N_g^i$  编码每组适配权重的个数, 据此可以推断出这一层的稀疏性, 即  $N:M = 3:4$ 。在实际应用过程中, 可以通过使用绝对多数投票法来放宽这一要求, 只要  $N_g^i$  中的稀疏比例  $x$  的个数比率达到预定义的投票比率  $v_r$ , 层稀疏性  $N^i$  就可以推断得出为  $x$ 。所以, 权值和阈值可以表示为  $w_k^i$  和  $T_k^i$ , 其中  $k$  是组别。

### 2.3 自定义逐层惩罚方法

本文使用标准的带随机梯度下降 (stochastic gradient descent, SGD) 更新规则的 DNNs 训练过程。

$$w(t+1) := w(t) - \xi_t g(w(t)) \quad (2)$$

其中,  $\xi_t$  表示第  $t$  阶段权重  $w$  的学习速率;  $g(w)$  表示权

重  $w$  的梯度。

为了获得稀疏的模型, 对于  $w(t)$  应用掩码  $m \in \{0, 1\}$ , 推导的稀疏模型为  $\tilde{w}(t) := w(t) \odot m$ , 其中符号  $\odot$  表示矩阵运算。在模型训练迭代  $t$  过程中, 掩码依赖于权值  $w(t)$  可以动态变化。基于分组和分层维数, 掩码  $m$  表示为

$$m_k^i(t) := w_k^i(t) > T_k^i \quad (3)$$

其中,  $i$  表示层索引;  $k$  表示组别;  $w_k^i(t)$  表示第  $t$  阶段第  $i$  层第  $k$  组内权重聚合值;  $T_k^i$  表示第  $i$  层第  $k$  组引入的组阈值; 掩码  $m_k^i(t)$  由布尔逻辑算子 ( $>$ ) 推出, 其中  $m = 1$  表示保留,  $m = 0$  表示修剪。

一般情况下, 应用额外的正则化惩罚来引入更多的稀疏性, 包括  $L_1$  范数和  $L_2$  范数。本文研究周期性地使用  $L_2$  范数, 不仅可以实现分层稀疏性, 还可以保持组的特征。其更新规则表示为

$$w_k^i(t+1) := w_k^i(t) - \varepsilon_i (g(\tilde{w}_k^i(t)) + p(t) \cdot \eta_i(t) \cdot \tilde{w}_k^i(t)) \quad (4)$$

$$p(t) = \begin{cases} 1, & t \% k_p = 0 \\ 0, & \text{其他} \end{cases} \quad (5)$$

其中,  $i$  表示层索引;  $\eta_i(t)$  是每一层控制惩罚强度的分层惩罚因子;  $k_p$  表示预设正则周期。本文使用了一个周期性的权值惩罚, 逐步驱动适配权值逐渐趋向于阈值, 并留下  $k_p$  迭代作为缓冲时间, 以适应应用的惩罚; 在搜索阶段, 分层稀疏性  $N^i$  和模型复杂度降低; 当模型复杂度达到搜索约束  $S_{pt}$  时, 终止搜索过程, 并基于搜索的分层细粒度稀疏方案对权值进行微调。

## 3 实验结果与分析

### 3.1 实验设计

ImageNet 数据集是用于计算机视觉识别项目的大型可视化数据库, 被广泛地用于世界大型顶级视觉识别挑战项目中, 具有较强的代表性和权威性。残差网络是在残差快之间使用了跳跃链接的卷积神经网络, 方便优化, 可有效缓解深度神经网络中增加深度带来的梯度消失问题。ResNet18 是由 17 个卷积层 (convolutional layer, Conv) 和 1 个全连接层 (fully connected layer, FC) 所搭建的 18 层残差网络结构, 图像数据集经过网络的预处理模块后被 4 个残差网络块进一步处理, 每 1 个残差网络块包含 2 个基本块和 2 个残差, 最后输入全连接层。Resnet50 网络中包含了 49 个卷积层, 1 个全连接层, 比 ResNet18 规模更大。

本文基于 ResNet18 和 Resnet50 在大型 ImageNet 数据集上验证了方法的有效性, 展示了分层细粒度  $N:M$  稀疏性的优势。在动态地训练模型和进行模型结构剪枝过程中, 为了避免直接通过稀疏权重的梯度更新稠密权重, 造成被剪枝的权重梯度更新误差增大, 降低网络性能的问题, 通过引入惩罚项, 减少“已剪掉权重”的梯度, 缩小更新速度, 最终稀疏网络训练过程会趋于稳定。实验首先从

预先训练的稠密模型中找到分层稀疏方案,并使用搜索的分层方案对预先训练的稠密权值进行微调/再训练稀疏权值,以恢复精度,最终分别以模型大小和模型计算量为约束来计算模型结果的精度,并和当前主流的方法比较,验证方法的合理性和可行性。

3.2 实验结果

1) ResNet18。表 1 和表 2 显示了在 ResNet18 上的结果。根据实验,动态分层稀疏 ResNet18 在等效模型大小情况下,其精度比统一基线提高 2.1%,结果如表 1 所示。动态分层稀疏 ResNet18 在等效模型计算量的情况下,其

精度比统一基线提高 1.3%,结果如表 2 所示。本文的方法与稠密稀疏模型所达到的 69.8%的 top-1 精度具有竞争性,后者参数量和模型计算量较大。

2) ResNet50。表 3 和表 4 显示了在 ResNet50 上的结果。根据实验,动态分层稀疏 ResNet50 在等效模型大小情况下,其精度比统一基线提高 1.3%,结果如表 3 所示。动态分层稀疏 ResNet50 在等效模型计算量的情况下,其精度比统一基线提高 1.0%,结果如表 4 所示。同样,本文的方法与参数量和模型计算量较大的稠密稀疏模型所达到的 76.2%的 top-1 精度具有一定的竞争性。

表 1 分层稀疏 ResNet18 与具有等效模型大小的统一基线对比

Tab. 1 Hierarchical sparse ResNet18 compared with a unified baseline of equivalent model size

| 方法                   | 参数量/ $10^6$ | 模型计算量/ $10^6$ | Top1 精度    | 结构          |
|----------------------|-------------|---------------|------------|-------------|
| ResNet18             | 11.7        | 1814          | 69.8       | 稠密          |
| 均匀剪枝 <sup>[19]</sup> | 1.46        | 227           | 66.6       | 2:16        |
| 动态稀疏                 | 1.46        | 329           | 68.7(+2.1) | 动态分层 $N:16$ |

表 2 分层稀疏 ResNet18 与具有等效模型计算量的统一基线对比

Tab. 2 Comparison of hierarchical sparse ResNet18 with a unified baseline of equivalent model computational capacity

| 方法                   | 参数量/ $10^6$ | 模型计算量/ $10^6$ | Top1 精度    | 结构          |
|----------------------|-------------|---------------|------------|-------------|
| ResNet18             | 11.7        | 1814          | 69.8       | 稠密          |
| 均匀剪枝 <sup>[19]</sup> | 1.46        | 227           | 66.6       | 2:16        |
| 动态稀疏                 | 1.29        | 227           | 67.9(+1.3) | 动态分层 $N:16$ |

表 3 分层稀疏 ResNet50 与具有等效模型大小的统一基线对比

Tab. 3 Hierarchical sparse ResNet50 compared with a unified baseline of equivalent model size

| 方法                   | 参数量/ $10^6$ | 模型计算量/ $10^6$ | Top1 精度    | 结构          |
|----------------------|-------------|---------------|------------|-------------|
| ResNet50             | 25.6        | 4 090         | 76.2       | 稠密          |
| 均匀剪枝 <sup>[19]</sup> | 3.2         | 511           | 74.4       | 2:16        |
| 动态稀疏                 | 3.2         | 731           | 75.7(+1.3) | 动态分层 $N:16$ |

表 4 分层稀疏 ResNet50 与具有等效模型计算量的统一基线对比

Tab. 4 Comparison of hierarchical sparse ResNet50 with a unified baseline of equivalent model computational capacity

| 方法                   | 参数量/ $10^6$ | 模型计算量/ $10^6$ | Top1 精度    | 结构          |
|----------------------|-------------|---------------|------------|-------------|
| ResNet50             | 25.6        | 4 090         | 76.2       | 稠密          |
| 均匀剪枝 <sup>[19]</sup> | 3.2         | 511           | 74.4       | 2:16        |
| 动态稀疏                 | 3.3         | 511           | 75.4(+1.0) | 动态分层 $N:16$ |

通过对比,在精度-稀疏性和精度-模型计算量权衡方面,动态分层细粒度稀疏神经网络始终优于其统一基线的比较对象;和稠密网络效果相比,动态分层细粒度稀疏神经网络同时保持稀疏性和模型计算量的优势,未来能够更灵活地支持模型和硬件设备之间的综合系统设计。

5 结论

本文提出了一种新颖高效的神经网络剪枝框架,从稠密权值中找到分层细粒度的  $N:M$  稀疏性,与等效模型大小或模型计算量的统一对等模型相比,该框架可以提高模型精度。基于 ResNet18 和 ResNet50 在 ImageNet 大规模

数据集上的实验,在同等的网络模型大小和模型计算量的情况下,本文的动态分层  $N:M$  稀疏性相比当前均匀剪枝方法精度更高;模型取得的精度和稠密网络的结果相当,在稀疏性、模型计算量和精度权衡方面,本文的方法更具有竞争性。实验结果表明,细粒度  $N:M$  结构化稀疏神经网络在计算机视觉领域具有广阔的优势和应用价值。

参考文献

[1] 晓勇,林璐,谭贵蓉,等.一种基于卷积神经网络的高影响天气识别方法[J].自动化技术与应用,2024,43(10):212-216.  
[2] 鲍泓,郑颖,梁天骄.认知机器的模型与结构研究进展[J].中国图象图形学报,2025,30(4):895-921.

- [3] 姜嘉成, 贾政轩, 徐钊, 等. 基于博弈对抗复杂系统的决策建模与求解[J]. 系统仿真学报, 2025, 37(1): 66-78.
- [4] 黄迎春, 邢秀祺. 融合 CNN-GRU 和 Transformer 的网络入侵检测方法[J]. 火力与指挥控制, 2025, 50(6): 21-27.
- [5] 王浩乾, 刘局, 李腾, 等. 基于多尺度特征融合的激光干扰图像复原性能提升[J]. 激光与光电子学进展, 2025, 62(4): 377-385.
- [6] Wang L, Yoon K J. Knowledge distillation and student-teacher learning for visual intelligence: a review and new outlooks [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(6): 3048-3068.
- [7] Shen J, Duan Y, Shen H W. Surroflow: a flow-based surrogate model for parameter space exploration and uncertainty quantification [J]. Visualization and Computer Graphics, IEEE Transactions on, 2025, 31(1): 635-644.
- [8] 高晗, 田育龙, 许封元, 等. 深度学习模型压缩与加速综述[J]. 软件学报, 2021, 32(1): 68-92.
- [9] Gholami A, Kim S, Dong Z, et al. A survey of quantization methods for efficient neural network inference [PP/OL]. arXiv (2021-03-25) [2024-05-11]. <https://doi.org/10.48550/arXiv.2103.13630>.
- [10] Zhou H, Yang M, Wang J, et al. Bayesnas: A bayesian approach for neural architecture search [C]// Proceedings of the 36th International Conference on Machine Learning. Long Beach: PMLR, 2019: 7603-7613.
- [11] Liu F, Zhang T, Dai W, et al. Few-shot adaptation of multi-modal foundation models: a survey [J]. Artificial Intelligence Review, 2024, 57(10): 1-30.
- [12] Krestinskaya O, Fouda M E, Benmezziane H, et al. Neural architecture search for in-memory computing-based deep learning accelerators [J]. Nature Reviews Electrical Engineering, 2024, 1(6): 374-390.
- [13] 宋清昆, 郝敏. 基于改进相关性剪枝算法的 BP 神经网络的结构优化[J]. 自动化技术与应用, 2006, 25(12): 4-6.
- [14] 闫涵, 卢伟, 吴玉虎. 基于轻量化卷积神经网络的金属断口图像识别[J]. 控制与决策, 2024, 39(9): 2858-2866.
- [15] 宁欣, 赵文尧, 宗易昕, 等. 神经网络压缩联合优化方法的研究综述[J]. 智能系统学报, 2024, 19(1): 36-57.
- [16] 程点, 郑海斌, 陈晋音. 基于相似度感知的深度卷积神经网络剪枝方法[J]. 小型微型计算机系统, 2024, 45(11): 2656-2662.
- [17] 王鹏, 张嘉诚, 范毓洋. 适应于硬件部署的神经网络剪枝量化算法[J]. 计算机工程与科学, 2024, 46(9): 1547-1553.
- [18] Anzt H, Tsai Y M, Abdelfattah A, et al. Evaluating the performance of NVIDIA's A100 ampere GPU for sparse and batched computations [C]//2020 IEEE/ACM Performance Modeling, Benchmarking and Simulation of High Performance Computer Systems (PMBS), Atlanta: IEEE, 2020: 26-38.
- [19] Zhou A J, Ma Y K, Zhu J N, et al. Learning N: M fine-grained structured sparse neural networks from scratch [C]//9th International Conference on Learning Representations (ICLR 2021). Virtual Event, Austria: ICLR, 2021.
- [20] Zhang L, Wei H, Chen M. Efficient CNN model optimization via structured pruning and sparse tensor core acceleration on NVIDIA A100 GPUS: a hardware-aware approach with fine-tuning and sparse matrix computation techniques [J]. International Journal of Electronic Devices and Networking, 2024, 5(2): 12-17.
- [21] Ruda D, Turek S, Ribbrock D. Fastsemi-iterative finite element poisson solvers for tensor core GPUs based on prehandling [C]//European Conference on Numerical Mathematics and Advanced Applications. Cham: Springer, 2025: 320-330.
- [22] Shen L, Sun Y, Yu Z, et al. On efficient training of large-scale deep learning models [J]. ACM Computing Surveys, 2024, 57(3): 1-36.
- [23] Mei T, Zi Y, Cheng X, et al. Efficiency optimization of large-scale language models based on deep learning in natural language processing tasks [C]//2024 IEEE 2nd International Conference on Sensors, Electronics and Computer Engineering (ICSECE). Jinzhou: IEEE, 2024: 1231-1237.
- [24] Lazzaro D, CinÀ A E, Pintor M, et al. Minimizing energy consumption of deep learning models by energy-aware training [C]//International Conference on Image Analysis and Processing. Cham: Springer Nature Switzerland, 2023: 515-526.
- [25] Jia J, Liu J, Ram P, et al. Model sparsity can simplify machine unlearning [J]. Advances in Neural Information Processing Systems, 2023, 36: 51584-51605.
- [26] Fang G, Ma X, Song M, et al. Depgraph: towards any structural pruning [C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Piscataway: IEEE, 2023: 16091-16101.
- [27] Koo K, Kim H. V-skip: vectorized kernel-based structured kernel pruning for accelerating deep convolutional neural networks [J]. IEEE Access, 2023(11): 118547-118557.
- [28] Zhang Z, Cao C, Zheng F, et al. GCPNet: gradient-aware channel pruning network with bilateral coupled sampling strategy [J]. Expert Systems with Applications, 2025, 266: 126104.