摘要

近年来，深度神经网络研究有力推动了计算机视觉和人工智能技术的发展。随着性能的提升，深度神经网络的复杂度持续增加，阻碍了其在轻量化设备上的应用。如何在保持深度神经网络性能的前提下，提升其运行效率是计算机视觉领域中长期存在的基础难题之一。

深度神经网络压缩的目标是利用参数量化和剪枝等方法，降低模型复杂度。近几年虽然取得了较大的进展，但是深度神经网络压缩研究仍然面临一些亟待解决的难题。首先，网络不同层的结构、作用、参数量、计算量各异，导致单一的压缩策略难以适应不同网络层。其次，参数量化会引入量化误差，而传统网络量化方法难以自适应优化特征残差，导致特征残差逐层累积，使量化网络与浮点值网络性能差异增大。第三，参数离散性使低比特量化网络对数据域变化的自适应能力变差，严重制约了其在域外（Out-Of-Domain, OOD）数据集上的性能。本文围绕深度神经网络压缩主题，针对以上三个问题开展如下研究：

（1）针对单一压缩策略难以适应不同网络结构问题，提出了一种网络结构自适应剪枝和量化方法。基于元学习框架，本文针对网络不同层自适应地寻找剪枝与量化参数的最优组合。此外，本文提出了一种正则化损失函数优化方法，可以统一约束每一层的压缩参数，使压缩网络能够满足给定的计算量上限约束。在多个视觉任务上的实验结果表明，该方法能够在不同网络结构下取得比同期方法更高的压缩率和准确率。例如，该方法所生成网络的紧凑性和 ImageNet 分类准确率均超越了同期工作 DJPQ。在图片超分辨任务上，该方法达到与 DHP 相近的峰值信噪比，并节省了近一半的计算量。此外，该方法还能迁移到视觉 Transformer 压缩中。在 ImageNet 数据集上，该方法能将模型 DeiT-Small 参数量压缩至原来的 5.3%，并取得了 78.61% 的准确率，超过了同期工作 PTQ4ViT。

（2）针对量化网络结构难以自适应优化特征残差问题，提出了一种量化网络特征残差自适应优化方法。首先，本文提出了逐层特征蒸馏方法优化每个量化卷积层，最小化量化网络和浮点值网络中间特征图间的特征残差。其次，为提升量化网络的视觉信息建模能力，本文在每个量化卷积层内引入了轻量化的残差短接分支。本文提出了一种 SI 结构来实现该分支。此方法只需较少（如 10%）的额外计算开销，就能显著提升量化网络性能。本文在数据集 ImageNet 上进行了对比实验，实验结果表明，所提方法显著提升了网络分类准确率。例如，基于网络结构 ResNet-18，该方法训练的 1 比特量化网络在 ImageNet 上达到了 60.45% 的准确率，优于许多同期方法。

（3）针对低比特量化网络难以适应数据域变化问题，提出了一种量化网络跨域优化方法。该方法通过优化低比特量化网络的权重分布和激活值分布来提高低比特量化网络的跨域鲁棒性。量化网络需要存储一套浮点值隐参数，用于辅助量化权重的更新。所提方法通过优化浮点值隐参数为低比特量化网络找到参数空间中的平坦极小值。此外，该方法还通过将网络激活值分布优化为平均分布、减小量化误差，来提升跨域鲁棒性。在四个数据集上的实验结果表明，所提方法能有效提高低比特量化网络对域外数据的泛化能力，并对不同网络结构表现出了良好的兼容性。与同期方法相比，该方法训练的低比特量化网络具有更高的跨域泛化能力，并在同域图像分类数据集上取得了更高的准确率。

（4）基于所提网络压缩方法，本文应用行人重识别技术构建了密接人员高效识别系统。疫情防治工作需要运用行人重识别技术在海量视频数据中查找特定人员、分析人员轨迹、识别密接人员。本文通过对深度神经网络模型进行压缩和加速，构建了一个高效的行人重识别系统，并在真实场景数据集 FSID 上验证了所提方法的有效性。所提方法在保持行人特征提取模块性能的前提下，实现约 11 倍的加速。该系统在某市部分社区部署，并为防疫人员找到了 10 多位密接人员。

综上所述，本文分别针对单一压缩策略难以适应不同网络结构、量化网络结构难以自适应优化特征残差、低比特网络难以适应数据域变化三个难点问题开展研究。提出了网络结构自适应剪枝和量化方法、量化网络特征残差自适应优化方法、量化网络跨域优化方法。基于 ResNet、VGG、视觉 Transformer 等多种网络结构，所提方法在图像分类、图像分割、图像超分辨等任务中进行了验证。此外，本文构建了基于紧凑网络的密接人员高效识别系统，并在公开数据集 Market-1501 和真实场景数据集 FSID 中进行了验证和应用。实验结果显示，所提方法能有效降低深度神经网络计算量，显著优化了紧凑网络在多种视觉任务中的准确率和泛化能力。本文的成果为深度神经网络在手持、车载、机载设备等移动端中高效部署奠定了基础。

关键词：网络压缩，网络剪枝，二值网络，混合比特量化

Adaptive Compression Methods for Deep Neural Networks

Jianming Ye (Technology of Computer Application)
Directed by: Prof. Shiliang Zhang

ABSTRACT

In recent years, research on deep neural networks has significantly advanced the development of computer vision and artificial intelligence technology. As performance continues to improve, the complexity of deep neural networks has also increased, hindering their application on lightweight devices. One of the long-standing fundamental problems in the field of computer vision is how to improve the efficiency of deep neural networks while maintaining their performance.

The goal of deep neural network compression is to reduce the complexity of models through quantization and reducing network parameters. Although significant progress has been made in recent years, there are still several challenges that need to be addressed in deep neural network compression research. First, a single compression strategy is difficult to adapt to different network structures. Networks have varying layer structures, functions, parameter amounts, and computational loads, making it difficult for a single compression strategy to adapt to different layers. Second, the quantized network structure is difficult to adaptively optimize the feature residual. The quantization process introduces quantization errors. Traditional network quantization methods are unable to adaptively optimize feature residuals, leading to the accumulation of feature residuals layer by layer and increasing the performance gap between compressed and floating-point networks. Third, low-bit networks are difficult to adapt to changes in data domain. Due to the discreteness of parameters, low-bit quantized networks have poor adaptive ability to changes in the data domain. Once transferred to a complex and diverse out-of-domain dataset, the accuracy will decrease significantly. This dissertation focuses on the theme of deep neural network compression and conducts research on the above three issues.

(1) To address the problem that a single compression strategy is difficult to adapt to different network structures, a network structure adaptive pruning and quantization method is proposed. Based on the meta-learning framework, this dissertation adaptively searches for the optimal combination of pruning and quantization parameters for different layers of the network.

In addition, a regularization loss function optimization method is proposed to uniformly constrain the compression parameters of each layer, allowing the compressed network to meet the given computational constraint. Experimental results on multiple visual tasks show that this method can achieve higher compression rates and accuracy than the state-of-the-art methods for different network structures. For example, the compactness and ImageNet classification accuracy of the generated network using this method both surpass the same period method DJPQ. In the image super-resolution task, this method achieves similar accuracy to DHP while saving nearly half of the computational cost. Moreover, this method can also be applied to compressing visual Transformer (Vision Transformer). On the ImageNet dataset, this method can compress the parameters of the DeiT-Small model to 5.3% of the original size and achieve an accuracy of 78.61%, outperforming the same period method PTQ4ViT.

(2) To address the problem that quantized network structures are difficult to adaptively optimize feature residuals, an adaptive optimization method for feature residuals in quantized networks is proposed. First, the dissertation proposes a layer-by-layer feature distillation method to optimize each quantized convolutional layer, minimizing the feature residual between the feature maps of the quantized network and the floating-point convolutional neural network. Second, to improve the modeling ability of the quantized network, this dissertation introduces a lightweight residual shortcut branch inside each quantized convolutional layer to assist in optimizing each layer's distillation loss. A novel Squeeze-and-Interaction (SI) structure is proposed to implement this branch. This method can significantly improve the performance of the quantized network with only a small additional computation cost (e.g., 10%). The proposed method was validated on ImageNet dataset. The experiments results show that the proposed method significantly improves the network's classification accuracy. For example, based on the ResNet-18 network structure, the 1-bit quantized network trained with this method achieves an accuracy of 60.45% on ImageNet, surpassing many contemporary methods.

(3) To address the problem that low-bit quantized neural networks are difficult to adapt changes in data domain, a cross-domain optimization method is proposed. The dissertation finds that low-bit quantization compression significantly reduces the network's cross-domain generalization ability through experiments. To improve the network's cross-domain robustness, the dissertation proposes to optimize the weight and activation value distributions of the low-bit quantized network. A set of floating-point hidden parameters is required to assist the update of the quantized weight parameters. The proposed method helps the low-bit quantized network find a flat minimum by optimizing the floating-point hidden parameters, and reduces

quantization errors by optimizing the quantized network activation values to approach the distribution average. The proposed method was validated respectively on four datasets. The experiments results show that the proposed method effectively improves the quantized network's generalization ability to out-of-domain data and shows good compatibility with different network structures. Compared with concurrent methods, the low-bit quantized network trained by this method has higher cross-domain generalization ability and achieves higher accuracy on the same-domain image classification dataset.

(4) This dissertation applies and validates the proposed methods by building an efficient identification system for close contacts of COVID-19 cases based on a compact neural network. In the prevention and control of epidemics, person re-identification technology is needed to search for specific individuals in massive video data. By compressing deep neural network models, this dissertation constructs an efficient person re-identification system. The system uses the network compression algorithm proposed in this dissertation to quantize the model, achieving efficient feature extraction and retrieval. The experiment results on the large-scale real-world scenario dataset FSID demonstrate the effectiveness of the proposed methods. For example, the compression method proposed in this dissertation can accelerate the feature extraction module by about 11 times while maintaining the performance. The system helped epidemic prevention personnel identify more than 10 close contacts of COVID-19 cases.

In summary, this dissertation presents a research on model adaptive compression methods for deep neural networks, focusing on the challenges of a single compression strategy is difficult to adapt to different network structures, quantized network structures are difficult to adaptively optimize feature residuals, and low-bit networks are difficult to adapt to changes in data domain. To address these challenges, the dissertation proposes 1) network structure adaptive pruning and quantization methods, 2) an adaptive optimization method for feature residuals in quantized networks, and 3) a domain generalization capability enhancement method for low-bit quantized neural networks. The proposed methods are validated on tasks such as image classification, image segmentation, and image super-resolution, using network structures such as ResNet, VGG, and visual transformers. Additionally, the dissertation presents a compact network-based efficient personnel recognition system, which validates and deploys the proposed algorithms. Experimental results show that the proposed methods can effectively reduce the computational complexity of deep neural networks, improve computation speed, and significantly optimize the accuracy and generalization ability of the network in various visual tasks. The achievements of this dissertation lay the foundation for the efficient deployment of deep neural networks in

mobile devices such as handheld, vehicle-mounted, and airborne equipment.

KEY WORDS: neural network compression, network pruning, binary network compression, mixed-precision quantization