紧致化神经网络的鲁棒性分析

陈光耀¹, 彭佩玺^{1,2}, 田永鸿^{1,2*}

1. 北京大学计算机科学技术系, 北京 100871;

2. 鹏城实验室, 深圳 518055

* E-mail: yhtian@pku.edu.cn

收稿日期: 2021-06-07; 接受日期: 2021-08-27; 网络版发表日期: 2022-03-17

广东省重点领域研发计划资助项目(编号: 2019B010153002)

摘要 深度神经网络(deep neural network, DNN)目前已经在很多视觉识别任务中达到了超越人类性能的表现. 然而, 将深度神经网络部署到真实场景的终端设备面临着两个挑战: 首先, 目前的深度神经网络模型非常耗费计算资源和内存, 很难直接应用在资源受限的终端设备上; 其次, 真实场景的数据往往受到各种各样噪声的影响, 因此部署在真实场景的模型应该拥有良好的鲁棒性(robustness). 近年来出现了很多对深度神经网络进行压缩和加速的方法以适应真实场景的资源受限终端, 而对这些压缩模型的鲁棒性分析也得到了越来越多的关注. 本文首次对神经网络剪枝、量化及知识蒸馏等常用的压缩算法进行对抗鲁棒性及噪声鲁棒性分析, 总结了不同压缩算法对神经网络鲁棒性的影响. 并对目前同时解决神经网络鲁棒性及模型大小问题的一些方案进行了整理与分析, 展望了该领域未来研究的技术挑战与可能的发展方向.

关键词 神经网络, 量化与剪枝, 知识蒸馏, 鲁棒性分析

1 引言

在深度学习的推动下, 深度神经网络在很多实际任务^[1-6]中取得了巨大的进步. 然而如今的神经网络模型在真实场景部署中仍然面临着两个巨大的挑战. 首先, 现有的神经网络模型大多数参数量巨大, 这导致其运行与部署需要强大的算力支持和巨大的存储开销^[7,8]. 这些巨大的开销阻碍了神经网络部署到一些资源受限设备上运行, 例如嵌入式设备、手机、机器人, 及物联网设备^[8-10]. 另一个主要的问题就是深度神经网络的安全性问题. 深度神经网络虽然取得了极佳的性能, 但是如果通过特定的策略给神经网络的输入添

加微小的扰动, 其分类或者检测结果就会发生明显错误^[11-14]. 而真实场景往往容易受到各种各样噪声的影响, 因此越来越多的研究^[15-17]关注在如何使得神经网络对抗这些未知扰动. 然而, 现有改善神经网络的模型大小或计算复杂度的研究与提升神经网络对抗稳定性的研究都是孤立发展的. 如今深度神经网络在实际部署中往往需要进行模型压缩后再部署, 因此我们认为这两个挑战必须结合起来考虑, 从而使得神经网络的部署具有高效性和安全性. 那么如今的神经网络压缩方法是否能够实现紧致化与鲁棒性的统一呢?

目前关于神经网络压缩的方法大致包含神经网络参数的剪枝(parameter pruning)^[18-21]和量化(quan-

引用格式: 陈光耀, 彭佩玺, 田永鸿. 紧致化神经网络的鲁棒性分析. 中国科学: 技术科学, 2022, 52: 689-703

Chen G Y, Peng P X, Tian Y H. Robustness analysis for compact neural networks (in Chinese). Sci Sin Tech, 2022, 52: 689-703, doi: 10.1360/SST-2021-0233

size)^[22~24]、知识蒸馏(knowledge distillation, KD)^[25~28]、低秩分解(low-rank factorization)^[29,30]、神经网络架构搜索(neural architecture search, NAS)^[31,32]等方法。其中神经网络剪枝、量化和知识蒸馏是目前广泛使用的大规模神经网络模型压缩方法。剪枝和量化都是为了减少神经网络参数的冗余。其中剪枝主要通过筛选出神经网络中不重要的权重参数,用于降低网络的复杂度或解决过拟合问题,而量化主要通过减少表示每个权重所需的比特数来压缩原始网络。知识蒸馏则是通过原始的大型网络的输出,训练适应不同任务的小型网络,用于模仿复杂模型能够实现的功能。以基于神经网络参数的压缩方法为例,目前的大部分方法都采用了相似的框架进行压缩:首先在预训练好的大规模模型上进行剪枝或量化,然后通过原始数据重训练以恢复性能,并不断迭代上述过程。对于剪枝来说,一种常用权重筛选策略^[18~21]便是通过权重值的大小来判断权重的重要性,值越小的权重就会被判断为不重要的权重被剪去。这种方法已经在多种任务上证明能够实现高压缩率的同时保持高准确度的性能。但是这些压缩方式真的可以减小模型大小的同时保持模型的鲁棒性吗?

如图1所示,我们展示了不同神经网络模型的权重值分布情况。图1(a)和(b)展示了自然训练的模型和通过对抗训练得到的鲁棒性模型的权值分布。可以看到,原始自然图片训练得到的神经网络模型相对于通过鲁棒性训练得到的神经网络更加稀疏,其分布在零值附近的权重更多。这也意味着鲁棒的模型往往需要更多的参数量。图1(c)展示了不同大小的神经网络模型权值分布情况。这里ResNet18^[33]比VGG16^[34]拥有更好的分类性能和更少的网络参数量。图1(c)可以明显地看到ResNet18的权重分布更加紧凑。因此可以推断出鲁棒性好的神经网络模型应该拥有更紧凑化的参数分布。那么通过压缩算法紧致化神经网络是否可以提高神经网络模型的鲁棒性呢?或者这些压缩算法是否会影响神经网络模型的鲁棒性?

目前已经有一部分针对紧致化神经网络对抗鲁棒性增强的研究工作^[35~38],而其中大部分研究主要集中在剪枝算法。虽然一些合理的剪枝算法能够实现神经网络鲁棒性能的提升,但其压缩比例与现有压缩算法存在一定的差距。还有一些方法通过在重训练阶段添加对抗样本攻击或者添加稀疏正则化来提高神经网络

图1 不同神经网络模型的权重值分布

Figure 1 Weight distributions of different neural network models. (a) ResNet18; (b) VGG16; (c) ResNet18 and VGG16.

的对抗鲁棒性。但是目前大部分方法仅针对模型的对抗鲁棒性进行评估,仍然没有针对各种压缩算法的对抗鲁棒性和噪声鲁棒性的影响进行整体性的评估和分析。

本文首次针对剪枝、量化和知识蒸馏三类神经网络压缩算法进行了整体性回顾和分类,通过定量化的实验分析了三类方法对神经网络对抗鲁棒性及噪声鲁

棒性的影响, 并给出参考性的结论分析. 发现了目前的常用的压缩策略都会对其鲁棒性能带来一定的影响, 而这种影响存在负面影响, 也存在积极影响方面. 最后, 针对这些结论对目前针对紧致化神经网络鲁棒性的算法进行整体性回顾, 并进一步展望了该领域未来研究的重大挑战与可能发展方向.

2 深度神经网络模型压缩

2.1 神经网络剪枝

如今的神经网络模型的层次结构越来越复杂, 参数数目巨大, 其中冗余也越来越大. 有研究表明, 神经网络中激活值中只有44%的值为零, 然而零值对于神经网络来说没有任何的表达作用, 所以大部分剪枝方法将其看作为无效连接. 同时在神经网络前向和反向运算过程中, 仍然要为这些冗余的神经元浪费大量的运算成本. 神经网络剪枝方法则是通过一些策略或者算法来判断不同权重或神经元的重要程度, 将网络中不重要的神经元或通道移除, 从而减小神经网络的参数规模, 并提升模型的运算效率.

而大部分神经网络剪枝方法的基本流程保持一致, 即首先对神经网络的参数重要性进行判定, 然后进行网络剪枝, 最后对网络重训练, 并不断的迭代这个过程. 当然也有一些算法可以只通过参数的重要性进行剪枝而无需网络的重训练过程, 但相对来说这样的准确率会带来较大的损失.

针对剪枝的策略我们又可以将剪枝算法分为结构化剪枝和非结构化剪枝. 其中非结构化剪枝不考虑模型的结构信息, 例如通道、秩等. 而针对深度神经网络的结构化剪枝大多数直接通过启发式的方法对网络的权重进行稀疏化, 例如我们前面所提到的权重值大小. 但是非结构剪枝由于其不规则性, 往往很难针对其做硬件上训练或者推理的加速. 目前有很多的研究工作, 包括使用梯度^[39,40]、动量和基于Hessian的启发式方法^[18,41~44], 以及具有权值大小的剪枝方法^[8,21,45~53]. 同时非结构化稀疏也可以通过引入 L_0 或 L_1 正则化^[54]和变分丢失(variational dropout)实现^[55].

结构化剪枝则是考虑到神经网络模型的结构化信息. 而这种方法可以使得神经网络模型在现有的硬件系统上具有良好的伸缩性, 可用于硬件的加速和优化. 现在的结构化剪枝的方法利用低秩权重张量^[56~58]、

通道修剪、通道过滤和诱导块稀疏性的方法. 虽然结构化剪枝可以提供帮助硬件实现加速有优化, 但是针对更高层次的剪枝只能通过非结构剪枝实现^[59~63].

2.2 神经网络量化

神经网络量化算法的核心思想是将高精度的权重映射到低精度的集合中, 即通过低精度参数, 例如16位精度浮点或8位定点整数, 来代替高精度的32位全精度浮点参数, 理想情况下能够有效减小模型大小, 同时不会丢失太多精确度. 二值网络(binanzed neural networks)^[64]和XNOR-Net^[65]是最早使用1-bit表示神经网络的工作. 但是大多数情况下只使用1-bit性能会明显的下降. 另外一些研究提出了低比特的方法^[66]来表示任意比特宽度的权重、激活和梯度, 这样的方法也可以适当地实现前向传播和反向传播的速度. 此后的研究关注在如何将全精度的参数量化到低比特. 而这些量化策略可以分为均匀量化^[66~69]、对数量化^[70,71]和自适应量化^[72,73]. 进一步, 混合精度量化^[74~76]则是通过适当的量化比特的组合来匹配深度神经网络各层的灵敏度, 这类方法在相同的约束条件下往往可以获得更好的性能. 在实际的应用中, 我们也可以根据是否需要重训练来选择这些量化方法. 根据重训练的需求, 又可将量化分为训练后量化和训练量化, 其中训练后量化方法(post-training quantization, PTQ)^[77,78]适用于训练环境或者数据集完全不可用的场景, 而训练量化方法(quantization aware training, QAT)^[79,80]则需要数据来训练出更有效的量化神经网络模型.

2.3 知识蒸馏

神经网络的训练和推理在很多场景都是两个不同任务. 在训练阶段模型大多数情况下不需要面临计算资源的限制, 而实际部署推理则需要对延迟和资源消耗进行关注. 因此训练模型通常比理想的推理模型要大的多, 而知识蒸馏便是将大模型中有效内容提取到小规模模型的方法. 知识蒸馏最早由Hinton等人^[27]在2015年提出, 其认为对于不同的预测标签, 虽然都是Softmax函数分配概率, 但是仍有数量级上的差异, 其中可能包含的信息被称作“暗知识”, 因此通过教师学生训练模型, 让一个规模小的学生网络, 同时学习图片本身的标签和规模较大的高精度教师网络的输出, 从而将知识从大模型转移到小模型. 随后的研究专注

于如何改变学习目标从而促进知识的转移. 一些工作通过中间特征^[81]或特征转换的变体^[82~85]. 还有一些工作^[86~90]将一元匹配的概念扩展到二元关系模型. Da-bouei等人^[91]研究了数据增强对知识蒸馏的影响, 证明了一些数据增强方式能够可以传递额外的知识. 还有一些工作^[92~94]研究了知识蒸馏中的数据效率问题, 其目的是在数据有限的环境中获得高的学生准确性.

2.4 低秩分解

低秩分解利用矩阵或张量分解来估计神经网络中卷积层的权重参数. 卷积核可以看作是一个三维张量, 而低秩分解则是通过张量分解的思想来消除三维张量中存在的冗余. 低秩分解来解决卷积加速问题已经有了很多研究, 例如最早的高维DCT(discrete cosine transform)和由一维DCT变换和一维小波构造的小波系统. Denton等人^[95]提出了卷积核的低秩近似和聚类方法, 实现单个卷积层2倍加速. Li等人^[96]提出用于寻找训练CNN的最优低秩近似, 即基于约束优化的低秩近似, 从而限制累加操作和内存占比. 除此之外, 仍然有很多基于低秩分解的加速方法^[29,30,97~99], 虽然低秩分解对于模型加速和压缩简单有效, 但其实现由于涉及分解操作, 不仅带来额外的计算压力, 同时还需要大量的模型重训练以实现收敛. 并且不适用不包含卷积操作的神经网络, 本文中不针对低秩分解方法进行特定分析.

2.5 神经网络架构搜索

目前的压缩和加速方法主要针对特定的神经网络结构, 但当我们考虑它们的组合时, 可以探索的架构策略就会极具增加. 神经网络架构搜索^[31,32]就是为了对定义的神经网络不同组件的一组决策的搜索. 是一种系统化、自动化的学习最佳神经网络架构的方法. 根据Elsken等人^[100]的总结, 神经网络架构搜索方法可分为搜索空间(search space)、搜索策略(search strategy)、性能评估策略(performance estimation strategy)等. 针对压缩预训练好的模型场景, 神经网络架构搜索方法可以与剪枝、量化等压缩方法进行结合, He等人^[101]提出AutoML for model compression (AMC), 即通过一种强化学习策略来逐层对模型进行修剪, 通过性能评估策略, 可以省去对剪枝后模型的重训练. Cao等人^[102]使用贝叶斯优化搜索最佳压缩架构, 根据压缩

性能及精度性能选择集合中最佳模型.

3 神经网络鲁棒性

3.1 对抗鲁棒性

目前已经有很多的工作^[11~14]证明了对抗样本成功的实验, 即在推理阶段将人类不可察觉的扰动人为地添加到神经网络的输入, 从而愚弄神经网络. 在不同约束条件下, 多次对抗攻击可以产生让神经网络分类或检测错误的成功的扰动. 而这些攻击方法^[11~14]往往需要神经网络的梯度, 针对性地在原始数据上添加噪声.

为了应对这些攻击方法, 对抗鲁棒性(adversarial robustness)训练被提出, 在尽量保持模型原始的分类或检测能力的同时, 优化抵御对抗攻击的鲁棒性. 目前大部分研究工作都是通过解决一个稳健的优化问题^[12,15,103,104]来实现的. 在优化过程中, 自然输入和带有扰动的输入同时被优化目标损失最小, 即在自然训练中加入对抗攻击样本. 而目前的研究也发现, 在训练过程中使用更强的攻击方式得到的对抗样本, 往往可以更好地提升神经网络模型抵御对抗攻击^[12,15,105]的鲁棒性. 然而这种鲁棒性只适用于基于梯度的攻击方法, 因此也被称为经验鲁棒性(empirical robustness). 另外一种鲁棒性训练方向被称为可验证的鲁棒性训练(verifiable robustness)^[106~108], 其目的是使得神经网络模型具有可验证的鲁棒性. 可验证的鲁棒性训练方法使用一种声音验证技术(sound verification techniques)^[109]将对抗样本的输出过度逼近神经网络自然图片输出, 并优化具有可证明鲁棒性的目标损失. 但是这种方法很难扩展到大型的网络模型和数据集.

3.2 噪声鲁棒性

目前的大部分数据集在收集时, 并没有考虑实际应用案例中遇到的各种干扰. 实际应用中, 传感器特性、照明条件或运动会引入各种噪声变换, 图像的处理或数据的传输会导致意外的失真, 这是广泛存在的现象. 因此基于这些干净数据集训练的神经网络往往也会受一些自然场景变化^[110,111]和未知场景^[112,113]的影响, 产生错误的预测和行为, 如雨、雾等自然场景的变化. 针对各种噪声鲁棒性问题, 目前提出了各种数据增强策略^[114~117]来提高神经网络的泛化性, 从而抵抗各

种噪声带来的潜在影响。

4 模型压缩算法鲁棒性分析

4.1 神经网络剪枝模型鲁棒性分析

本小节探究了不同剪枝算法对神经网络模型鲁棒性的影响。对于剪枝模型在不同稀疏率下对自然图片性能影响我们采用了和大多数压缩模型同样的评价策略, 即Top1准确率。对于对抗鲁棒性分析, 我们采用FGSM(fast gradient sign method)^[11]白盒攻击方法生成对抗样本:

$$\hat{x} = x + \epsilon \text{sign}(\nabla_x J(\theta, x, y)), \quad (1)$$

其中 x 为原始样本, $J(\theta, x, y)$ 是训练神经网络的损失, $\epsilon=8/255$ 。即通过损失梯度不断地对原始图像添加微小噪声, 直到神经网络判断错误。对于噪声鲁棒性, 采用基于常见腐败和扰动构建的数据集^[96]ImageNet-C, CIFAR10-C或CIFAR100-C测试模型在各种噪声下的泛化能力。

这里我们采用了两种目前为止性能优越的剪枝方法, 一种是不需要进行重训练的软阈值权重参数化非结构化剪枝方法(soft threshold weight reparameterization, STR)^[118]和需要重训练的自动化结构搜索人工蜂群的结构化剪枝策略(artificial bee colony algorithm, ABC)^[119]。在不同剪枝率下ResNet50^[33]在ImageNet^[120]上性能及对抗鲁棒性及噪声鲁棒性表现如表1和2所示。

首先, 对于不需要重训练的非结构化剪枝方法STR, 我们可以从表1中看到, 当稀疏率是81.27%时, 虽然自然图片的性能只下降了1.1%。但是其对抗攻击防御性能却下降了7.3%, 远远的超过了自然图像的性能的损失。相反, 对于噪声鲁棒性, 除98.98%稀疏率之外, 噪声图像上的性能下降基本与自然图像一致, 但过于稀疏时, 其也表现出噪声鲁棒性损失远远大于自然损失的现象。因此, 目前大部分方法所采用的只通过自然图像来评判剪枝模型的性能损失往往是片面的。其整体的泛化性及鲁棒性能要比目前评价策略所表现的严重的多。

我们再关注一下使用结构化搜索并重训练的剪枝策略ABC剪枝。可以观察到即使在稀疏度只有29.5%的情况下, 其自然图像性能下降了1.1%, 但对抗鲁棒

表 1 基于不需要重训练的非结构化剪枝策略的ResNet50在不同稀疏率下ImageNet上性能及鲁棒性损失

Table 1 Performance and robustness loss of ResNet50 on ImageNet at different sparsity rates based on an unstructured pruning strategy without retraining

稀疏率 (%)	自然图像 Top1 (%)	对抗图像 Top1 (%)	噪声图像 Top1 (%)
0	77.1	22.9	39.0
81.27	76.0 (-1.1)	15.6 (-7.3)	38.3 (-0.7)
90.23	74.2 (-2.9)	10.5 (-12.4)	35.9 (-3.1)
95.15	70.2 (-6.9)	5.7 (-17.2)	31.0 (-8.0)
96.11	67.8 (-9.3)	3.6 (-19.3)	28.2 (-10.8)
98.05	61.4 (-15.7)	1.6 (-21.3)	23.0 (-16.0)
98.98	51.7 (-25.4)	0.4 (-22.5)	16.8 (-32.2)

表 2 基于重训练下结构化剪枝策略的ResNet50在不同剪枝率下ImageNet性能及鲁棒性损失

Table 2 Performance and robustness loss of ResNet50 on ImageNet under different pruning rates based on structured pruning strategy with retraining

稀疏率 (%)	自然图像 Top1 (%)	对抗图像 Top1 (%)	噪声图像 Top1 (%)
0	76.0	7.9	38.1
29.50	74.9 (-1.1)	6.1 (-1.8)	36.4 (-1.7)
54.02	73.8 (-2.2)	5.5 (-2.4)	35.3 (-2.7)
56.01	73.4 (-2.6)	5.0 (-2.9)	34.8 (-3.3)
64.38	72.5 (-3.5)	4.1 (-3.8)	32.5 (-5.6)
71.24	72.5 (-5.8)	2.7 (-5.2)	29.9 (-8.2)

性能仅下降1.8%, 噪声鲁棒性下降了1.7%。这看起来要比不需要重训练的非结构剪枝策略下降的要平缓一些, 但是也可以看到大部分稀疏率下其对抗鲁棒性及噪声鲁棒性的损失要比自然图像上的性能损失大的多, 尤其是噪声鲁棒性。因此可以推断出重训练策略可以在一定程度上缓解剪枝带来的鲁棒性损失, 并带来更多的可能性。

整体上可以看到剪枝方法的鲁棒性损失要比其在自然图像上表现的损失要严重。但基于重训练的结构化剪枝方法也许可以对解决这样的损失带来更多的可能性。

4.2 神经网络量化模型鲁棒性分析

本小节探究在不同比特位数下的量化模型的性能

及鲁棒性能损失. 针对对抗鲁棒性以及噪声鲁棒性的评估策略和4.1小节一致. 我们针对任意精度的量化策略^[121]得到的不同比特位数的ResNet50在CIFAR10^[122]和ImageNet上进行了评估, 实验结果如表3所示.

可以看到在CIFAR10数据集上, 不同比特位数的自然图像性能影响不大, 甚至在4比特量化模型上还有性能的提升. 但是可以看到其抵御对抗攻击的损失相比于自然图像要大的多. 这样的表现在ImageNet上表现的更为突出. 由于不同算法采用的预训练模型不一样, 其性能存在一定的差异. 但是鲁棒性损失要大于自然图像的损失的现象仍然可以观察到. 对于噪声图像, 高比特条件下其性能变化并不是十分明显, 但当比特数较低时, 我们惊奇的发现其噪声鲁棒性在CIFAR10上有一定的性能提升, 但在ImageNet上有较大损失. 这可能与量化权重后, 低比特权重的有限表示能够消除部分自然小噪声以及扰动的变化影响. 但当换到分辨率更大的图像上时, 这种噪声往往就会被放大, 超过量化模型所能抵御噪声的范围. 同时, 该结论并不适用于对抗鲁棒性, 因为对抗攻击可以获得网络错误预测的梯度, 进行针对性的扰动干预.

表 3 不同量化比特下ResNet50性能及鲁棒性损失

Table 3 Performance and robustness loss of ResNet50 under different quantization bits

比特	CIFAR10			ImageNet		
	自然图像	对抗图像	噪声图像	自然图像	对抗图像	噪声图像
	Top1 (%)	Top1 (%)	Top1 (%)	Top1 (%)	Top1 (%)	Top1 (%)
32	93.6	21.2	77.2	75.9	8.2	36.2
8	93.4 (-0.2)	20.1 (-1.1)	77.1 (-0.1)	75.0 (-0.9)	6.9 (-1.3)	33.9 (-2.3)
4	93.6 (+0.0)	17.8 (-3.4)	77.3 (+0.1)	74.1 (-1.8)	6.1 (-2.1)	31.3 (-4.9)
2	93.3 (-0.3)	18.6 (-2.6)	78.4 (+1.2)	68.6 (-7.3)	2.5 (-5.7)	21.6 (-14.6)
1	91.5 (-2.1)	17.6 (-3.6)	78.5 (+1.3)	58.0 (-17.9)	0.9 (-7.3)	13.9 (-23.3)

表 4 不同蒸馏模型的CIFAR100性能及鲁棒性损失

Table 4 Performance and robustness loss of different distillation models on CIFAR100

老师模型	老师模型性能			学生模型	学生模型性能		
	自然图像	对抗图像	噪声图像		自然图像	对抗图像	噪声图像
	Top1 (%)	Top1 (%)	Top1 (%)		Top1 (%)	Top1 (%)	Top1 (%)
ResNet50	78.7	11.0	50.3	ShuffleNetV1	75.2 (-3.5)	10.2 (-0.8)	50.3 (-0.0)
WRN-40-2	76.7	11.9	47.9	ShuffleNetV1	75.9 (-0.8)	12.5 (+0.6)	50.1 (+2.2)
ResNet50	78.7	11.0	50.3	VGG8	73.5 (-5.2)	4.8 (-6.2)	48.7 (-1.6)
VGG16	75.3	19.5	49.2	VGG8	72.5 (-2.8)	3.5 (-16.0)	47.8 (-1.4)

在对抗鲁棒性方面, 整体上量化和剪枝算法表现类似, 其针对特定扰动的损失要大于自然图像损失. 但是相对于剪枝算法来说, 量化保持了神经网络的整体结构, 其对抗鲁棒性的性能损失要比剪枝算法小一些. 对于噪声鲁棒性, 低比特的量化可以消除部分随机扰动对网络输出带来的影响. 相比于剪枝, 量化策略能够有效抵抗随机噪声扰动带来的影响.

4.3 知识蒸馏模型鲁棒性分析

最后, 我们探究知识蒸馏的教师和学生模型的性能及鲁棒性能损失. 鲁棒性评估策略与前两节一致. 我们针对知识蒸馏(KD)^[27]最初的方法进行评估, 即让一个规模小的学生网络, 同时学习图片本身的标签和规模较大的高精度教师网络的输出, 从而将知识从大模型转移到小模型. 我们这里分别采用ResNet50, VGG16^[34], WRN-40-2^[123]作为教师模型, 采用ShuffleNetV1^[124], VGG8作为学生模型.

表4展示了不同教师和学生模型下的性能, 以及对鲁棒性和噪声鲁棒性损失. 首先, 对于ShuffleNetV1的学生模型, 其在自然图像上相对于教师模型要有一

定的损失. 但是其对抗鲁棒性和噪声鲁棒性相对于自然图像损失较小, 甚至在ResNet50作为教师网络下还有提升. 其原因可能与图1(c)所示, 用更小的模型容纳相同知识, 能带来更好的鲁棒性. 但是VGG8上学生模型的对抗鲁棒性损失相对于自然图像损失就大很多, 而其噪声鲁棒性损失是其中最小的, 而这有可能和VGG8网络结构有关. 因此学生模型的性能及鲁棒性损失不仅和其模型容量有关, 也有可能与其自身的网络结构存在关联.

5 紧致化鲁棒神经网络模型

本节中, 结合第4节中定量化实验结果, 我们对目前神经网络模型压缩及鲁棒性统一优化的工作进行回顾与总结, 并进一步展望了该领域未来研究的重大挑战与可能发展方向. 如表5, 我们按照是否针对是否考虑模型鲁棒性, 对目前的模型压缩算法进行了重新划分.

5.1 剪枝模型及其鲁棒性

目前已经有很多的研究工作^[125~127]都对稀疏化神经网络的对抗鲁棒性进行分析, 发现当剪枝率比较低时(如5%~10%), 神经网络抵御对抗攻击的鲁棒性有所提高. 但这些方法大部分都是在MNIST和CIFAR10等小规模数据集上进行的测试, 在大规模的ImageNet上反映情况更加严重一些. 适当地对网络进行稀疏化的确可以提高网络的鲁棒性, 这也是 L_1 和 L_2 正则化^[54]所证明的. 但目前的剪枝方法大部分要求80%以上的剪枝效果, 这样高的剪枝率所带来的过度稀疏化反过来会严重影响神经网络的鲁棒性能.

Schwag等人^[37]首次系统性地对剪枝模型和对抗鲁棒性进行实验, 其发现和我们的实验结果基本一致, 即不进行重训练的剪枝往往会影响对抗鲁棒性, 而进

行重训练后的模型的对抗鲁棒性损失得到缓解, 但是可解释的对抗鲁棒性却出现了显著的下降, 这可能是由于重训练和初始训练的目标不一致导致. 同时指出了适当的剪枝并重训练的策略往往比从头训练的方法能更好地维护对抗鲁棒性能, 但是仍然会损失部分自然图像的性能.

进一步Schwag等人分析了结构化剪枝和非结构化剪枝对自然图像性能 and 对抗鲁棒性的影响. 非结构化剪枝能够更好地维护自然图像性能, 但是其对抗鲁棒性会出现很大的损失. 而结构化剪枝能够比非结构化剪枝维护较好的对抗鲁棒性, 但是其自然图像性能往往会受到很大的影响. 这和我们实验结果的表1和2表现是一致的.

Ye等人^[36]在剪枝的重训练过程中加入了对抗训练的优化目标, 成功提升了剪枝后模型的对抗鲁棒性能. 而Schwag等人^[38]在最新的工作中, 将对抗鲁棒性的优化问题加入到剪枝的策略中, 从而提升剪枝后模型的对抗鲁棒性. 目前这些方法的剪枝效率还不能和传统的剪枝方法相对比, 但传统方法带来的更多是虚假的性能, 其损失比表现出的要严重的多.

对于剪枝模型的噪声鲁棒性, 除本文外并没有系统性分析的文章. 但从第4节中实验结果可以看到, 剪枝模型的噪声鲁棒性变化与对抗鲁棒性变化趋势基本一致. 通过重训练, 并引入针对对抗鲁棒性以及噪声鲁棒性优化目标, 可以有效地提升剪枝模型的鲁棒性. 但如何进一步缩短剪枝后模型与原始模型之间的鲁棒性差异, 仍然值得进一步的探索. 同时关于结构化鲁棒性一体化的剪枝方法仍然是一个值得更多探索的领域.

5.2 量化模型及其鲁棒性

神经网络量化方法在训练过程中往往涉及到对量化过程的模拟, 使得训练后的模型高度依赖于目标的比特宽度和精确的量化方式. Lin等人^[128]首次对量化和对抗鲁棒性的关系进行了分析, 其指出量化虽然在自然图像上带来微小的性能损失, 但是其鲁棒性损失很大, 这与表3中的实验现象一致. 同时他们也发现对于微小的人工噪声, 量化后的模型要比原始模型拥有更好的鲁棒性, 但是随着噪声的增大, 这样的鲁棒性会消失并带来损失. 这其实与我们所分析的噪声鲁棒性结果相对应, 即低比特量化模型的有限表达空间,

表 5 模型压缩算法分类

Table 5 Classification of model compression algorithms

模型压缩算法分类	模型紧致化方法	模型鲁棒性压缩方法
剪枝	非结构化剪枝	[8,18,21,39~53] [37,38,125~127]
	结构化剪枝	[56~63] [37]
量化	[64~80]	[35,128~131]
知识蒸馏	[81~94]	[132]

可以有效抵御部分随机自然小噪声的影响. 这与量化本身的特点有关, 其能够通过量化中心操作消除微小噪声. 同时, 量化模型能够抵御随机小噪声的现象, 值得进一步的理论探索.

还有一些研究工作^[35,128-131]对权重的量化的对抗鲁棒性维护提出了解决方案. Lin等人^[128]在训练过程中加入权重正则化项使得训练后的量化模型具有更好的对抗鲁棒性. Shkolnik等人^[35]发现均匀量化要比正态量化方式具有更好抵御噪声波动的能力. 目前大部分方法都是在训练过程中加入一定的优化策略以维护模型的对抗鲁棒性. 因此, 针对需要训练的量化方法来说, 更有效的对抗训练方法是未来主要的探索方向. 同时, 如何维护不需要重训练量化模型的对抗鲁棒性, 仍然值得进一步探索.

另一方面, 我们根据第4节的结果也可以分析出对抗鲁棒性与噪声鲁棒性是相互独立的, 这与Laugros等人^[133]的结论是一致的. 当神经网络模型的梯度暴露, 往往可以对模型中的弱点进行针对性的扰动.

5.3 鲁棒模型与知识蒸馏

知识蒸馏的研究主要关注在如何有效地将大模型的教师模型的知识传递到小模型的学生模型. 其已经被广泛用于大规模神经网络模型压缩, 例如GTP-3^[134], BERT^[135]等大规模模型. 其是目前将大模型应用到下游任务中常用的模型压缩方式. 目前没有针对模型对抗鲁棒性以及噪声鲁棒性与知识蒸馏方法的系统性分析, 本文表4中的实验是较为完整的学生模型鲁棒性分析. 可以看到知识蒸馏方式相对于剪枝和量化来说, 没有特别统一的结论, 其鲁棒性不仅与模型容量大小有关, 也有可能与学生模型的结构存在关联. 但其对抗鲁棒性以及噪声鲁棒性表现较为一致. 因此更鲁棒性的网络结构也许是知识蒸馏鲁棒性压缩算法未来值得探究的方向. 另一方面, 由于知识蒸馏需要对学生模型进行训练, 因此在训练阶段可以针对不同任务对学生模型进行针对性的调整. Dabouei等人^[91]研究了数据增强对知识蒸馏的影响, 证明了一些数据增强方式能够可以传递额外的知识. Goldblum等人^[132]进一步在训练学生模型阶段加入对抗样本, 从而提升了学生模型的鲁棒性能. 因此, 可以在学生模型训练阶段引入更多能够提升模型鲁棒性的策略, 从而改善现有的知识蒸馏框架.

6 总结与展望

本文针对神经网络模型中常用的剪枝、量化和知识蒸馏三类模型压缩方法进行了定量化实验的对抗鲁棒性及噪声鲁棒性分析. 并对目前针对紧致化神经网络鲁棒性的算法进行回顾与总结, 实验结论与最新的研究进展相一致. 最后我们从鲁棒性分析、鲁棒性提升解决方案, 以及未来可能的研究方法对三类压缩方法进行总结.

(1) 剪枝模型

鲁棒性分析. 适度的增加神经网络模型的稀疏率往往可以提升模型的鲁棒性. 但是目前的高剪枝率的剪枝算法往往会带来较大的鲁棒性损失. 其中非结构化剪枝能够更好地维护自然图像性能, 但是其鲁棒性会出现很大的损失; 而结构化剪枝能够比非结构化剪枝维护较好的鲁棒性, 但是其自然图像性能往往会受到很大的影响.

解决方案. 通过重训练, 并引入针对对抗鲁棒性以及噪声鲁棒性优化目标, 可以有效地提升剪枝模型的鲁棒性.

未来可能的研究方向. 不需要重训练的情况下, 如何进一步缩短剪枝后模型与原始模型之间的鲁棒性差异, 仍然值得进一步的探索.

(2) 量化模型

鲁棒性分析. 量化模型和剪枝模型的对抗鲁棒性变化类似, 自然图像上针对性的微小变化往往会引起对抗鲁棒性的巨大损失. 但低比特量化模型的有限表达空间, 可以有效抵御微小人工噪声和部分随机自然小噪声的影响. 同时也证明了对抗鲁棒性与噪声鲁棒性之间的独立性.

解决方案. 加入对抗样本的重训练方法是能够提升量化模型对抗鲁棒性的有效方法.

未来可能研究方向. 如何维护不需要重训练量化模型的对抗鲁棒性, 仍然值得进一步探索. 同时, 量化模型能够抵御随机小噪声的现象, 值得进一步的理论探索.

(3) 知识蒸馏

鲁棒性分析. 知识蒸馏中的学生模型的对抗鲁棒性以及噪声鲁棒性表现较为一致, 其鲁棒性不仅与模型容量大小有关, 也有可能与学生模型的结构存在关联.

解决方案. 在知识蒸馏中加入额外的对抗鲁棒性以及噪声鲁棒性的优化目标能够提升学生模型鲁棒性.

未来可能研究方向. 如今常用的大规模预训练模型是否能够有效提升学生模型的对抗鲁棒性以及噪声鲁棒性值得进一步探索.

整体上传统压缩方法带来的性能损失远比我们想象大的多, 因此对于未来紧致化鲁棒模型的研究仍然

有很多内容需要我们去探索. 首先, 模型压缩算法的评估需要补充对模型对抗鲁棒性以及噪声鲁棒性评估. 同时, 目前针对这三类压缩算法鲁棒性提升的策略, 几乎都是在训练阶段增加关于鲁棒性的约束项或对抗样本. 虽然这样可以高效提升模型的鲁棒性, 但我们认为更多针对鲁棒性优化的策略需要添加到剪枝、量化或者知识蒸馏的压缩策略中, 探索一种维护模型整体性能的可解释压缩手段.

参考文献

- 1 Carrère V, Conel J E. Recovery of atmospheric water vapor total column abundance from imaging spectrometer data around 940 nm—Sensitivity analysis and application to airborne visible/infrared imaging spectrometer (AVIRIS) data. *Remote Sens Environ*, 1993, 44: 179–204
- 2 Esteva A, Robicquet A, Ramsundar B, et al. A guide to deep learning in healthcare. *Nat Med*, 2019, 25: 24–29
- 3 Hinton G, Deng L, Yu D, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Process Mag*, 2012, 29: 82–97
- 4 Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Boston, 2015. 3431–3440
- 5 Bojarski M, Del Testa D, Dworakowski D, et al. End to end learning for self-driving cars. 2016, arXiv: [1604.07316](#)
- 6 Taigman Y, Yang M, Ranzato M, et al. DeepFace: Closing the gap to human-level performance in face verification. In: *Proceedings of the IEEE Computer Vision and Pattern Recognition*. Columbus, 2014
- 7 Canziani A, Paszke A, Culurciello E. An analysis of deep neural network models for practical applications. 2016, ArXiv: [1605.07678](#)
- 8 Han S, Pool J, Tran J, et al. Learning both weights and connections for efficient neural networks. 2015, ArXiv: [1506.02626](#)
- 9 Parashar A, Rhu M, Mukkara A, et al. SCNN: An accelerator for compressed-sparse convolutional neural networks. *SIGARCH Comput Archit News*, 2017, 45: 27–40
- 10 Yao S, Zhao Y, Zhang A, et al. Deepiot: Compressing deep neural network structures for sensing systems with a compressor-critic framework. In: *Proceedings of the 15th ACM Conference on Embedded Network Sensor Systems*. Delft, 2017. 1–14
- 11 Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. In: *Proceedings of the International Conference on Learning Representations (ICLR)*. San Diego, 2015
- 12 Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks. In: *Proceedings of the International Conference on Learning Representations (ICLR)*. Vancouver, 2018
- 13 Carlini N, Wagner D. Towards evaluating the robustness of neural networks. In: *Proceedings of the IEEE Symposium on Security and Privacy*. San Jose, 2017. 39–57
- 14 Athalye A, Carlini N, Wagner D. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In: *Proceedings of the International Conference on Machine Learning*. Stockholm, 2018. 274–283
- 15 Tramèr F, Boneh D, Kurakin A, et al. Ensemble adversarial training: Attacks and defenses. In: *Proceedings of the International Conference on Learning Representations (ICLR)*. Vancouver, 2018
- 16 Song C, He K, Wang L, et al. Improving the generalization of adversarial training with domain adaptation. In: *Proceedings of the International Conference on Learning Representations (ICLR)*. Vancouver, 2018
- 17 Wang S, Pei K, Whitehouse J, et al. Efficient formal safety analysis of neural networks. In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. Montréal, 2018. 6369–6379
- 18 LeCun Y, Denker J, Solla S. Optimal brain damage. In: *Proceedings of the Advances in Neural Information Processing Systems*. Denver, 1989. 598–605
- 19 Li H, Kadav A, Durdanovic I, et al. Pruning filters for efficient convnets. In: *Proceedings of the International Conference on Learning Representations (ICLR)*. San Juan, 2016

- 20 Han S, Mao H, Dally W J. Compressing deep neural networks with pruning, trained quantization and Huffman coding. In: Proceedings of the International Conference on Learning Representations (ICLR). San Diego, 2015
- 21 Guo Y, Yao A, Chen Y. Dynamic network surgery for efficient DNNs. In: Proceedings of the Advances in Neural Information Processing Systems. Barcelona, 2016. 1379–1387
- 22 Gong Y, Liu L, Yang M, et al. Compressing deep convolutional networks using vector quantization. 2014, ArXiv: [1412.6115](#)
- 23 Wu J, Leng C, Wang Y, et al. Quantized convolutional neural networks for mobile devices. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, 2016. 4820–4828
- 24 Han S, Mao H, Dally W J. Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding. In: Proceedings of the International Conference on Learning Representations (ICLR). San Juan, 2016
- 25 Zhang L, Chen G, Shi Y, et al. Anonymous model pruning for compressing deep neural networks. In: Proceedings of the IEEE Conference on Multimedia Information Processing and Retrieval (MIPR). Shenzhen, 2020. 157–160
- 26 Ba L J, Caruana R. Do deep nets really need to be deep? In: Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, 2014. 2654–2662
- 27 Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. 2015, ArXiv: [1503.02531](#)
- 28 Korattikara A, Rathod V, Murphy K, et al. Bayesian dark knowledge. In: Proceedings of the 28th International Conference on Neural Information Processing Systems. Bali, 2015. 3438–3446
- 29 Rigamonti R, Sironi A, Lepetit V, et al. Learning separable filters. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Portland, 2013. 2754–2761
- 30 Denton E, Zaremba W, Bruna J, et al. Exploiting linear structure within convolutional networks for efficient evaluation. In: Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, 2014. 1269–1277
- 31 Zoph B, Le Q V. Neural architecture search with reinforcement learning. 2016, ArXiv: [1611.01578](#)
- 32 Real E, Moore S, Selle A, et al. Large-scale evolution of image classifiers. In: Proceedings of the International Conference on Machine Learning. Sydney, 2017. 2902–2911
- 33 He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, 2016. 770–778
- 34 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: Proceedings of the International Conference on Learning Representations (ICLR). San Diego, 2015
- 35 Shkolnik M, Chmiel B, Banner R, et al. Robust quantization: One model to rule them all. 2020, ArXiv: [2002.07686](#)
- 36 Ye S, Xu K, Liu S, et al. Adversarial robustness vs. model compression, or both? In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Long Beach, 2019. 111–120
- 37 Schwag V, Wang S, Mittal P, et al. Towards compact and robust deep neural networks. 2019, ArXiv: [1906.06110](#)
- 38 Schwag V, Wang S, Mittal P, et al. Hydra: Pruning adversarially robust neural networks. In: Proceedings of the Advances in Neural Information Processing Systems (NeurIPS). Virtual-only, 2020
- 39 Chen Y, Yu Z, Fang W, et al. Pruning of deep spiking neural networks through gradient rewiring. 2021, ArXiv: [2105.04916](#)
- 40 Kusupati A, Ramanujan V, Somani R, et al. Soft threshold weight reparameterization for learnable sparsity. In: Proceedings of the International Conference on Machine Learning. Virtual-only, 2020. 5544–5555
- 41 Evci U, Gale T, Menick J, et al. Rigging the lottery: Making all tickets winners. In: Proceedings of the International Conference on Machine Learning. Virtual-only, 2020. 2943–2952
- 42 Lee N, Ajanthan T, Torr P H S. SNIP: Single-shot network pruning based on connection sensitivity. In: Proceedings of the International Conference on Learning Representations (ICLR). Vancouver, 2018
- 43 Hassibi B, Stork D G. Second order derivatives for network pruning: Optimal brain surgeon. In: Proceedings of the 5th International Conference on Neural Information Processing Systems. San Francisco, 1992. 164–171
- 44 Dettmers T, Zettlemoyer L. Sparse networks from scratch: Faster training without losing performance. 2019, ArXiv: [1907.04840](#)
- 45 Zhu M, Gupta S. To prune, or not to prune: Exploring the efficacy of pruning for model compression. 2017, ArXiv: [1710.01878](#)
- 46 Frankle J, Carbin M. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In: Proceedings of the International Conference on Learning Representations (ICLR). Vancouver, 2018

- 47 Gale T, Elsen E, Hooker S. The state of sparsity in deep neural networks. 2019, ArXiv: [1902.09574](#)
- 48 Mostafa H, Wang X. Parameter efficient training of deep convolutional neural networks by dynamic sparse reparameterization. In: Proceedings of the International Conference on Machine Learning. Long Beach, 2019. 4646–4655
- 49 Bellec G, Kappel D, Maass W, et al. Deep rewiring: Training very sparse deep networks. In: Proceedings of the International Conference on Learning Representations (ICLR). Toulon, 2017
- 50 Mocanu D C, Mocanu E, Stone P, et al. Scalable training of artificial neural networks with adaptive sparse connectivity inspired by network science. *Nat Commun*, 2018, 9: 2383
- 51 Narang S, Elsen E, Diamos G, et al. Exploring sparsity in recurrent neural networks. In: Proceedings of the International Conference on Learning Representations (ICLR). Toulon, 2017
- 52 Kusupati A, Singh M, Bhatia K, et al. FastGRNN: A fast, accurate, stable and tiny kilobyte sized gated recurrent neural network. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montréal, 2018. 9031–9042
- 53 Wortsman M, Farhadi A, Rastegari M. Discovering neural wirings. In: Proceedings of the Advances in Neural Information Processing Systems. Vancouver, 2019. 2684–2694
- 54 Louizos C, Welling M, Kingma D P. Learning sparse neural networks through L_0 regularization. In: Proceedings of the International Conference on Learning Representations (ICLR). Vancouver, 2018
- 55 Molchanov D, Ashukha A, Vetrov D. Variational dropout sparsifies deep neural networks. In: Proceedings of the International Conference on Machine Learning. Sydney, 2017. 2498–2507
- 56 Jaderberg M, Vedaldi A, Zisserman A. Speeding up convolutional neural networks with low rank expansions. In: Proceedings of the British Machine Vision Conference. Nottingham, 2014
- 57 Prabhu A, Farhadi A, Rastegari M. Butterfly transform: An efficient FFT based neural architecture design. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual-only, 2020. 12024–12033
- 58 Lu Z, Sindhwani V, Sainath T N. Learning compact recurrent neural networks. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Shanghai, 2016. 5960–5964
- 59 Liu Z, Sun M, Zhou T, et al. Rethinking the value of network pruning. In: Proceedings of the International Conference on Learning Representations (ICLR). Vancouver, 2018
- 60 Wen W, Wu C, Wang Y, et al. Learning structured sparsity in deep neural networks. In: Proceedings of the 30th International Conference on Neural Information Processing Systems. Barcelona, 2016. 2082–2090
- 61 Li H, Kadav A, Durdanovic I, et al. Pruning filters for efficient convnets. In: Proceedings of the International Conference on Learning Representations (ICLR). San Juan, 2016
- 62 Luo J H, Wu J, Lin W. ThiNet: A filter level pruning method for deep neural network compression. In: Proceedings of the IEEE International Conference on Computer Vision. Venice, 2017. 5058–5066
- 63 Gordon A, Eban E, Nachum O, et al. MorphNet: Fast & simple resource-constrained structure learning of deep networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake, 2018. 1586–1595
- 64 Courbariaux M, Hubara I, Soudry D, et al. Binarized neural networks: Training deep neural networks with weights and activations constrained to +1 or -1. 2016, ArXiv: [1602.02830](#)
- 65 Rastegari M, Ordonez V, Redmon J, et al. XNOR-Net: ImageNet classification using binary convolutional neural networks. In: Proceedings of the European Conference on Computer Vision. Amsterdam, 2016. 525–542
- 66 Zhou S, Wu Y, Ni Z, et al. DoReFa-Net: Training low bitwidth convolutional neural networks with low bitwidth gradients. 2016, ArXiv: [1606.06160](#)
- 67 Choi J, Wang Z, Venkataramani S, et al. Pact: Parameterized clipping activation for quantized neural networks. 2018, ArXiv: [1805.06085](#)
- 68 Jung S, Son C, Lee S, et al. Learning to quantize deep networks by optimizing quantization intervals with task loss. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, 2019. 4350–4359
- 69 Sun Q, Shang F, Yang K, et al. Multi-precision quantized neural networks via encoding decomposition of $\{-1, +1\}$. In: Proceedings of the AAAI Conference on Artificial Intelligence. Honolulu, 2019. 5024–5032
- 70 Miyashita D, Lee E H, Murmann B. Convolutional neural networks using logarithmic data representation. 2016, ArXiv: [1603.01025](#)
- 71 Zhou A, Yao A, Guo Y, et al. Incremental network quantization: Towards lossless CNNs with low-precision weights. In: Proceedings of the

- International Conference on Learning Representations (ICLR). Toulon, 2017
- 72 Jin Q, Yang L, Liao Z. Adabits: Neural network quantization with adaptive bit-widths. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual-only, 2020. 2146–2156
- 73 Li Y, Dong X, Wang W. Additive powers-of-two quantization: An efficient non-uniform discretization for neural networks. In: Proceedings of the International Conference on Learning Representations (ICLR). New Orleans, 2019
- 74 Cai Z, Vasconcelos N. Rethinking differentiable search for mixed-precision neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual-only, 2020. 2349–2358
- 75 Uhlich S, Lukas M, Fabien C, et al. Mixed precision DNNs: All you need is a good parametrization. In: Proceedings of the International Conference on Learning Representations (ICLR). Virtual-only, 2020
- 76 Sun Q, Ren Y, Jiao L, et al. MWQ: Multiscale wavelet quantized neural networks. 2021, ArXiv: [2103.05363](#)
- 77 Banner R, Nahshan Y, Soudry D. Post training 4-bit quantization of convolutional networks for rapid-deployment. In: Proceedings of the Advances in Neural Information Processing Systems. Vancouver, 2019. 7950–7958
- 78 Nahshan Y, Chmiel B, Baskin C, et al. Loss aware post-training quantization. 2019, ArXiv: [1911.07190](#)
- 79 Esser S K, McKinstry J L, Bablani D, et al. Learned step size quantization. In: Proceedings of the International Conference on Learning Representations (ICLR). New Orleans, 2019
- 80 Choi J, Wang Z, Venkataramani S, et al. Pact: Parameterized clipping activation for quantized neural networks. 2018, ArXiv: [1805.06085](#)
- 81 Romero A, Ballas N, Kahou S E, et al. Fitnets: Hints for thin deep nets. In: Proceedings of the International Conference on Learning Representations (ICLR). Banff, 2014
- 82 Heo B, Lee M, Yun S, et al. Knowledge transfer via distillation of activation boundaries formed by hidden neurons. In: Proceedings of the AAAI Conference on Artificial Intelligence. Honolulu, 2019. 3779–3787
- 83 Lee S H, Kim D H, Song B C. Self-supervised knowledge distillation using singular value decomposition. In: Proceedings of the European Conference on Computer Vision (ECCV). Munich, 2018. 335–350
- 84 Kim J, Park S, Kwak N. Paraphrasing complex network: Network compression via factor transfer. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montréal, 2018. 2765–2774
- 85 Yang J, Martinez B, Bulat A, et al. Knowledge distillation via adaptive instance normalization. 2020, ArXiv: [2003.04289](#)
- 86 Ahn S, Hu S X, Damianou A, et al. Variational information distillation for knowledge transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, 2019. 9163–9171
- 87 Liu Y, Cao J, Li B, et al. Knowledge distillation via instance relationship graph. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, 2019. 7096–7104
- 88 Park W, Kim D, Lu Y, et al. Relational knowledge distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, 2019. 3967–3976
- 89 Tian Y, Krishnan D, Isola P. Contrastive representation distillation. In: Proceedings of the International Conference on Learning Representations (ICLR). New Orleans, 2019
- 90 Xu G, Liu Z, Li X, et al. Knowledge distillation meets self-supervision. In: Proceedings of the European Conference on Computer Vision. Virtual-only, 2020. 588–604
- 91 Dabouei A, Soleymani S, Taherkhani F, et al. Exploiting joint robustness to adversarial perturbations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual-only, 2020. 1122–1131
- 92 Kimura A, Ghahramani Z, Takeuchi K, et al. Few-shot learning of neural networks from scratch by pseudo example optimization. In: Proceedings of the British Machine Vision Conference. Newcastle, 2018
- 93 Li T, Li J, Liu Z, et al. Few sample knowledge distillation for efficient network compression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual-only, 2020. 14639–14647
- 94 Nayak G K, Mopuri K R, Shaj V, et al. Zero-shot knowledge distillation in deep networks. In: Proceedings of the International Conference on Machine Learning. Long Beach, 2019. 4743–4751
- 95 Denton E L, Zaremba W, Bruna J, et al. Exploiting linear structure within convolutional networks for efficient evaluation. In: Proceedings of the Advances in Neural Information Processing Systems. Montréal, 2014. 1269–1277
- 96 Li C, Shi C J. Constrained optimization based low-rank approximation of deep neural networks. In: Proceedings of the European Conference on

- Computer Vision (ECCV). Munich, 2018. 732–747
- 97 Lu Y, Kumar A, Zhai S, et al. Fully-adaptive feature sharing in multi-task networks with applications in person attribute classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, 2017. 5334–5343
 - 98 Denil M, Shakibi B, Dinh L, et al. Predicting parameters in deep learning. In: Proceedings of the Advances in Neural Information Processing Systems. Lake Tahoe, 2013
 - 99 Sainath T N, Kingsbury B, Sindhvani V, et al. Low-rank matrix factorization for deep neural network training with high-dimensional output targets. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Vancouver, 2013. 6655–6659
 - 100 Elsken T, Metzen J H, Hutter F. Neural architecture search: A survey. *J Mach Learn Res*, 2019, 20: 1997–2017
 - 101 He Y, Lin J, Liu Z, et al. AMC: Automl for model compression and acceleration on mobile devices. In: Proceedings of the European Conference on Computer Vision (ECCV). Munich, 2018. 784–800
 - 102 Cao S, Wang X, Kitani K M. Learnable embedding space for efficient neural architecture compression. In: Proceedings of the International Conference on Learning Representations (ICLR). New Orleans, 2019
 - 103 Liu Y, Chen X, Liu C, et al. Delving into transferable adversarial examples and black-box attacks. In: Proceedings of the International Conference on Learning Representations (ICLR). San Juan, 2016
 - 104 Zhang H, Yu Y, Jiao J, et al. Theoretically principled trade-off between robustness and accuracy. In: Proceedings of the International Conference on Machine Learning. Long Beach, 2019. 7472–7482
 - 105 Shafahi A, Najibi M, Ghiasi A, et al. Adversarial training for free! In: Proceedings of the Advances in Neural Information Processing Systems 32. Vancouver, 2019. 3358–3369
 - 106 Wong E, Schmidt F R, Metzen J H, et al. Scaling provable adversarial defenses. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montréal, 2018. 8410–8419
 - 107 Mirman M, Gehr T, Vechev M. Differentiable abstract interpretation for provably robust neural networks. In: Proceedings of the International Conference on Machine Learning. Stockholm, 2018. 3578–3586
 - 108 Gowal S, Dvijotham K, Stanforth R, et al. On the effectiveness of interval bound propagation for training verifiably robust models. 2018, ArXiv: [1810.12715](https://arxiv.org/abs/1810.12715)
 - 109 Wang S, Pei K, Whitehouse J, et al. Formal security analysis of neural networks using symbolic intervals. In: Proceedings of the 27th USENIX Security Symposium. Baltimore, 2018. 1599–1614
 - 110 Hendrycks D, Dietterich T. Benchmarking neural network robustness to common corruptions and perturbations. In: Proceedings of the International Conference on Learning Representations (ICLR). Vancouver, 2018
 - 111 Hendrycks D, Gimpel K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In: Proceedings of the International Conference on Learning Representations (ICLR). San Juan, 2016
 - 112 Chen G, Qiao L, Shi Y, et al. Learning open set network with discriminative reciprocal points. In: Proceedings of the 16th European Conference on Computer Vision. Glasgow, 2020. 507–522
 - 113 Chen G, Peng P, Wang X, et al. Adversarial reciprocal points learning for open set recognition. 2021, ArXiv: [2103.00953](https://arxiv.org/abs/2103.00953)
 - 114 Kim I, Han S, Baek J W, et al. Quality-agnostic image recognition via invertible decoder. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021. 12257–12266
 - 115 Hendrycks D, Basart S, Mu N, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. 2020, ArXiv: [2006.16241](https://arxiv.org/abs/2006.16241)
 - 116 Hendrycks D, Mu N, Cubuk E D, et al. Augmix: A simple data processing method to improve robustness and uncertainty. In: Proceedings of the International Conference on Learning Representations (ICLR). New Orleans, 2019
 - 117 Chen G, Peng P, Ma L, et al. Amplitude-phase recombination: Rethinking robustness of convolutional neural networks in frequency domain. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Virtual-only, 2021: 458–467
 - 118 Kusupati A, Ramanujan V, Somani R, et al. Soft threshold weight reparameterization for learnable sparsity. In: Proceedings of the International Conference on Machine Learning. Virtual-only, 2020. 5544–5555
 - 119 Lin M, Ji R, Zhang Y, et al. Channel pruning via automatic structure search. In: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI). Virtual-only, 2020
 - 120 Russakovsky O, Deng J, Su H, et al. Imagenet large scale visual recognition challenge. *Int J Comput Vis*, 2015, 115: 211–252

- 121 Yu H, Li H, Shi H, et al. Any-precision deep neural networks. In: Proceedings of the AAAI Conference on Artificial Intelligence. Virtual-only, 2021. 10763–10771
- 122 Krizhevsky A. Learning multiple layers of features from tiny images. Dissertation for Master's Degree. Toronto: University of Toronto, 2009
- 123 Zagoruyko S, Komodakis N. Wide residual networks. In: Proceedings of the British Machine Vision Conference. York, 2016
- 124 Zhang X, Zhou X, Lin M, et al. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake, 2018. 6848–6856
- 125 Wang S, Wang X, Ye S, et al. Defending DNN adversarial attacks with pruning and logits augmentation. In: Proceedings of the IEEE Global Conference on Signal and Information Processing (GlobalSIP). Anaheim, 2018. 1144–1148
- 126 Guo Y, Zhang C, Zhang C, et al. Sparse DNNs with improved adversarial robustness. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montréal, 2018. 240–249
- 127 Dhillon G S, Azizzadenesheli K, Lipton Z C, et al. Stochastic activation pruning for robust adversarial defense. In: Proceedings of the International Conference on Learning Representations (ICLR). Vancouver, 2018
- 128 Lin J, Gan C, Han S. Defensive quantization: When efficiency meets robustness. In: Proceedings of the International Conference on Learning Representations (ICLR). Vancouver, 2018
- 129 Gui S, Wang H N, Yang H, et al. Model compression with adversarial robustness: A unified optimization framework. In: Proceedings of the Advances in Neural Information Processing Systems. Vancouver, 2019. 1285–1296
- 130 Wijayanto A W, Choong J J, Madhawa K, et al. Towards robust compressed convolutional neural networks. In: Proceedings of the IEEE International Conference on Big Data and Smart Computing (BigComp). Kyoto, 2019. 1–8
- 131 Alizadeh M, Behboodi A, van Baalen M, et al. Gradient 1 regularization for quantization robustness. 2020, ArXiv: [2002.07520](https://arxiv.org/abs/2002.07520)
- 132 Goldblum M, Fowl L, Feizi S, et al. Adversarially robust distillation. In: Proceedings of the AAAI Conference on Artificial Intelligence. Virtual-only, 2020. 3996–4003
- 133 Laugros A, Caplier A, Ospici M. Are adversarial robustness and common perturbation robustness independent attributes? In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. Seoul, 2019
- 134 Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, 2019. 4171–4186
- 135 Brown T B, Mann B, Ryder N, et al. Language models are few-shot learners. 2020, ArXiv: [2005.14165](https://arxiv.org/abs/2005.14165)

Robustness analysis for compact neural networks

CHEN GuangYao¹, PENG PeiXi^{1,2} & TIAN YongHong^{1,2}

¹ Department of Computer Science and Technology, Peking University, Beijing 100871, China;

² Peng Cheng Laboratory, Shenzhen 5180550, China

Deep neural networks (DNNs) have achieved comparable performance to humans on many tasks. However, there are two problems in deploying a DDN to terminal devices of a real scene. First, DNNs consume huge computing resources and memory; therefore, it is difficult to directly apply DNNs to resource-constrained terminal devices. Second, real-world data are often affected by noise; accordingly, the model deployed in the real scene needs to have good robustness. Recently, many compression methods for DNNs have been adapted to resource-constrained terminal devices, and robustness analyses of these compression models have increasingly received attention. In this paper, the adversarial and corruption robustness of several compression algorithms, such as pruning, quantization, and knowledge distillation, is first analyzed. Then, several studies of compact networks with improved robustness are summarized. Finally, several challenges are discussed, and possible research directions for compact robustness networks are proposed.

neural networks, quantization and pruning, knowledge distillation, robustness analysis

doi: [10.1360/SST-2021-0233](https://doi.org/10.1360/SST-2021-0233)