资源有限条件下的高效训练

Mahdi Zolnouri
华为诺亚方舟实验室，蒙特利尔′，魁北克省，加拿大
mahdi.zolnouri@huawei.com

Dounia Lakhmiri
GERAD 和蒙特利尔理工学院，蒙特利尔，魁北克省，加拿大
dounia.lakhmiri@polymtl.ca

Christophe Tribes
GERAD 和蒙特利尔理工学院，蒙特利尔，魁北克省，加拿大 christophe.tribes@polymtl.ca

Eyy¨ub Sari
华为诺亚方舟实验室，蒙特利尔，魁北克省，加拿大
eyyub.sari@huawei.com

塞巴斯蒂安·勒·迪加贝尔 GERAD and Polytech
蒙特利尔大学′蒙特利尔′，加拿大魁北克省 sebastien.le.digabel@gerad.ca

arXiv:2301.09264v1 [cs.LG] 23 Jan 2023

摘要

训练时间预算和数据集规模是影响深度神经网络（DNN）性能的因素之一。本文表明，在这两个因素受限的情况下，神经架构搜索（NAS）、超参数优化（HPO）和数据增强有助于显著提升 DNN 的表现。然而，在资源有限的情况下，要找到最优架构、最佳超参数值以及良好的数据增强技术组合，需要进行大量实验。我们提出了分三步实现这一目标的方法：在保持模型性能与原始模型相当的同时，通过压缩模型缩短训练轮次时间；在数据集较小时防止模型过拟合；以及进行超参数调优。我们使用了 NOMAD，这是一款基于无导数算法的黑盒优化软件，用于执行神经架构搜索和超参数优化。在 ICLR 2021 硬件感知高效训练（HAET）挑战赛中，我们在 Mini-ImageNet（Vinyals 等，2016）的一个极小子集上取得了 86.0% 的准确率，并获得比赛第二名。比赛结果可见 haet2021.github.io/challenge，源代码可见 github.com/DouniaLakhmiri/ICLR HAET2021。

1 引言

深度神经网络的压缩主要针对推理阶段，但近期研究也开始尝试压缩训练过程。不过，这仍然颇具挑战。ICLR 2021 HAET 挑战赛是一项年度竞赛，旨在评估分类模型在训练时间和数据有限的情况下的性能。训练时间预算为 10 分钟，使用配备 32 GB 内存的英伟达 GPU V100，并运行在配备 12 GB RAM 的 Intel(R) Xeon(R) Gold 6230 CPU @ 2.10GHz 处理器上。数据集是一个小型未知数据集，包含 10 个类别，开发阶段提供 5K 张训练图像和 1K 张测试图像。模型输入为 32×32 RGB 图像，类似 CIFAR-10。评估阶段使用包含 80×80 张图像的 Mini-ImageNet 子集来评估候选模型的性能。参赛者可以使用自己的优化器、训练循环和数据增强流程。在训练过程中引入这些限制会带来两项主要挑战。首先，由于参数数量庞大，大多数分类器需要远超 10 分钟才能收敛。深度神经网络领域的大多数最先进方法普遍倾向于加深、加宽网络，以提升模型容量（Simonyan 和 Zisserman，2014；Szegedy 等，2016；2017；He 等，2016）。更大的网络需要更多时间和计算资源才能达到最佳准确率。本文通过应用神经架构搜索，使用压缩模型来减少模型的计算操作。其次，小型数据集会导致

模型过拟合。在这种情况下，模型会过度拟合训练图像，在训练数据集上表现良好，但在真实世界图像上的表现不佳。由于数据增强（DA）是一种生成额外训练数据的有效技术（Taylor 和 Nitschke，2017），我们在流水线中加入了多种数据增强策略的组合，以解决过拟合问题。

2 相关工作

自 AlexNet（Krizhevsky 等人，2012）通过堆叠卷积层赢得 2012 年 ImageNet ILSVRC-2012 竞赛以来，深度神经网络通过增加参数和隐藏层的数量，提高了准确率。2014 年，拥有 19 层的 VGGNet（Simonyan 和 Zisserman，2014）以及拥有 22 层的 GoogleNet（Szegedy 等人，2015）分别在 ILSVRC 竞赛的定位和分类任务中取得了最低错误率。2015 年，ILSVRC 冠军 ResNet 提出了跳跃连接，以解决深度神经网络中的梯度消失问题，并通过堆叠层将网络深度扩展至 152 层。SENet（Hu 等人，2018）赢得了 2017 年 ImageNet 竞赛。它拥有 1.45 亿个参数，规模几乎是拥有 680 万个参数的 GoogleNet 的 20 倍。最近，NFNet-F4+ （ Brock 等人，2021 ）以 5.27 亿个参数，在 ImageNet 数据集上取得了 89.2% 的 top-1 准确率，创下新的最先进水平。另一方面，一些网络尝试在模型大小和准确率之间取得良好的平衡。例如，MobileNet（Howard 等人，2017）为模型的深度、宽度和分辨率引入了三个乘数。研究表明，用户可以通过牺牲准确率来缩小模型。EfficientNet（Tan 和 Le，2019）使用神经架构搜索设计了一个新的基线模型，并引入了一种新的缩放方法，以构建一个达到最先进水平的新网络系列。这些工作主要关注推理方面，所提出的模型在训练过程中不受数据和资源方面的限制。

3 提出的方法

3.1 模拟挑战赛评估

作为初始步骤，我们首先实现了一个训练流程。我们以挑战赛网站提供的代码为基础，并添加了数据预处理、模型构建和训练器等新模块。在预处理模块中，我们添加了一个数据采样器，用于构建符合竞赛要求的数据集：训练集包含5千张图像，验证集包含1千张图像。我们选择CIFAR-10作为代理数据集，因为它是分类任务中最常见的十类别数据集。其规模使我们能够通过打乱选定图像的索引后再对数据进行子采样，构建各种子集，并帮助候选模型获得更好的泛化能力。模型构建模块提供了一个经过设计的模型，我们将其作为已知模型的神经架构搜索变体进行预览。训练器模块运行训练循环，时长恰好为十分钟，以模拟竞赛环境。

3.2 搜索基准网络

搜索空间越大，神经架构搜索所需的时间和成本就越高。高效利用资源是本次挑战赛的首要任务。因此，我们没有从头开始进行架构设计，而是将方法的第二步设为寻找一个基准网络候选模型，并以此为起点开展搜索。我们列出了一系列知名分类器，主要包括ImageNet ILSVRC挑战赛的获胜模型。我们改编了其中大多数分类器的实现，来源为Kuang (2021)。随后，我们使用从CIFAR-10中抽取的不同子集，对每个分类器训练十次。多次训练网络十分重要，这有助于平滑小规模训练集所引起的方差影响。我们使用自己的流水线训练所有分类器，训练时间预算为10分钟，且使用相同的数据集。其他超参数则与各模型对应论文中的设置相近。最后，我们选出了平均准确率最高的前两个模型。表1显示，SENet和ResNet-18的验证准确率分别为77.0%和75.5%，位列前两名。

3.3 神经网络架构搜索

为了缩短训练轮次时间，我们使用公式 1 对 SENet 和 ResNet-18 进行神经架构搜索。第三步的目标是找到一个规模更小、性能不低于基线模型的变体

模型。我们使用 NOMAD（Le Digabel，2011）黑盒优化软件，该软件实现了 MADS 算法（Audet & Dennis, Jr.，2006）。NOMAD 可以在满足特定约束条件的情况下优化目标函数，而无需了解函数的内部机制。我们将 $g(\phi)$ 定义为表示资源消耗的函数，例如乘加（MAC）运算次数等显式函数。我们还将模型性能限制为不低于基线模型。我们参考 Howard 等人（2017），通过为模型的三个维度——深度、宽度和输入分辨率——定义三个乘数，对基线模型进行缩放。深度乘数决定模型的模块数量。宽度乘数的作用是通过统一减少每层的输入和输出通道来缩窄网络。输入分辨率乘数会缩放输入图像，并相应地缩放每一层的内部表示。对于每次神经架构搜索试验，我们都会向 NOMAD 返回根据模型新维度计算出的 MAC 运算次数以及模型的准确率。

$$\min_{\phi \in \Phi} \quad g(\phi) \;\; \text{subject to} \;\; f(\mathbf{w}, \phi \mid \mathbf{x}, \mathbf{y}) \geq f(\mathbf{w}_0, \phi_0 \mid \mathbf{x}, \mathbf{y}) \tag{1}$$

其中，w 为神经网络的权重，Φ 表示神经网络的超参数空间，$\phi \in \Phi$ 表示用于定义网络架构的深度、宽度和分辨率值向量，而 (x, y) 表示数据特征和标签。我们将网络训练后的验证准确率记为 $f(\phi, \mathrm{w} \mid \mathrm{x}, \mathrm{y})$；将反映乘加运算次数的函数记为 $g(\phi)$。

3.4 数据增强方法

鉴于数据集规模较小，我们发现候选模型容易过拟合。由于数据集非常小，使用基本的图像增强技术无法让模型在训练过程中获得足够的性能提升。我们采用了 AutoAugment（Cubuk 等人，2018）方法，以自动选择最优的数据增强策略。此外，我们还加入了 CUTOUT（DeVries 和 Taylor，2017）技术，该技术会在训练期间随机遮蔽输入图像的部分区域。

3.5 超参数优化

最后，我们使用 NOMAD 对候选模型进行超参数优化，以找到超参数的最佳取值。公式 2 展示了我们提出的目标函数。

$$\max_{\lambda \in \Lambda} \quad f(\lambda, \mathbf{w} \mid \mathbf{x}, \mathbf{y}) \tag{2}$$

其中，λ 是来自 Λ 的超参数空间的取值向量。$f(\lambda, \mathrm{w} \mid \mathrm{x}, \mathrm{y})$ 表示神经网络的验证准确率。我们选择了三种已知会显著影响验证集表现的标准超参数：初始学习率、权重衰减和优化器类型。由于数据集很小，我们还选择调整批量大小。由于不同优化器对应的学习率取值范围不同，我们在训练期间根据每种优化器预期的学习率调整学习率。例如，随机梯度下降通常使用 0.1，而 ADAM 使用 0.001。最终，超参数优化实验的搜索空间如下：

- 均匀类型的学习率，范围为 [0.6;1E-3]
- 离散型权重衰减，取值为 $\{0, 0.00005、0.0005、0.005、0.05、0.5\}$
- 离散型优化器，取值为 {“Adadelta”(1)、“Adagrad” (0.01)、“随机梯度下降” (0.1)、“Adam” (0.01)、“AdamW” (0.01)、“Adamax” (0.002)、“随机梯度下降”}
- 离散型批量大小，取值为 $\{128, 256, 512\}$

超参数优化使我们的模型准确率提高了 1.1%。

4 实验与结果

在本节中，我们通过实验探究所提出方法在多种模型架构和数据集上的有效性。

4.1 CIFAR-10

CIFAR-10 是一个由 6 万张 32×32 彩色图像组成的数据集，划分如下：5 万张用于训练集，1 万张用于验证集。该数据集包含十个类别。由于它与竞赛评估数据集的描述十分接近，且规模足够大，可以划分出多个不同的子样本，因此我们决定在开发阶段将其用作代理数据集。在所有实验中，数据集规模如下：训练集为 5 千张图像，验证集为 1 千张图像。表1展示了我们寻找基线模型过程中的中间结果。所有分类器均在 CIFAR-10 的一个子集上训练十分钟。SE Net-18 的准确率为 77.0%，ResNet-18 的准确率为 75.5%，均优于其他分类器。

准确率 77.0 75.5 74.6 74.2 74.0 70.4 70.0 69.8 69.6 68.7 67.6 67.2 64.8 64.7 62.0 58.3 52.7 41.9

表1：搜索基准网络。最终验证准确率（%）：十次运行的平均值 s.

通过在 SENet-18 和 ResNet-18 上应用神经架构搜索，我们为每个模型找到了一个更小的变体，其表现与对应的基线模型相近。NOMAD-NAS-SENet-18 的深度和输入分辨率与 SENet-18 相同，但网络宽度仅为原模型的 67%。NOMAD-NAS-ResNet-18 的深度与 ResNet-18 相同，但其宽度和输入分辨率分别为原模型的 73% 和 118%。尽管输入分辨率比原模型高出 18%，NOMAD-NAS-ResNet-18 的乘加运算减少了 22%，参数数量减少了 47%。我们发现这两个候选模型都容易过拟合。我们在数据变换列表中加入 AutoAugment 和 CUTOUT 技术，以改进数据变换。表2展示了在应用神经架构搜索和数据增强后 SENet-18 与 ResNet-18 的表现。完成这一步后，我们继续对 NOMAD-NAS-SENet-18 进行超参数优化。表3展示了对候选模型 NOMAD-NAS-SENet-18 应用超参数优化后的验证准确率。我们使用 NOMAD 为模型寻找最佳超参数值。为此，我们使用 MADS 算法，对最终候选模型的多个参数进行了黑盒优化。如 3.5 节所述，我们得到了一组经过调优的超参数值：学习率 =0.042、权重衰减 = 0.005、优化器 = 随机梯度下降、批量大小 = 512。我们进行了多项测试，以确保这些最优值能够达到最佳准确率。我们观察到验证准确率从 86.7% 提升至 87.8%。

表2：在 SENet-18 和 ResNet-18 上应用神经架构搜索和数据增强。最终验证准确率：十次运行的平均值。

模型	验证准确率（%）
SENet-18	**86.7**
ResNet-18	84.8

表3：在 NOMAD-NAS-SENet-18 上应用超参数优化（HPO）。最终验证准确率：十次运行的平均值。

模型	验证准确率（%）
NOMAD-NAS-SENet-18	**87.8**

4.2 MINI-IMAGENET

比赛截止日期之后，ICLR HAET 挑战赛委员会公布了结果和评估数据集的名称。表 4 展示了我们的模型在 Mini-ImageNet 子集上的最终结果。我们在 CIFAR-10 上的评估结果与委员会在 Mini-ImageNet 上的评估结果之间存在显著差距，达 1.8%。硬件和训练循环等其他潜在差异，也可能造成这一差距。然而，影响模型性能的主要因素是数据变换中有一项额外操作，而挑战赛说明中并未提及。Mini-ImageNet 中的图像为 80×80。委员会在数据变换列表中添加了缩放操作，以便将图像转换为初始公告中所述的 32×32。我们采用余弦退火作为学习率衰减调度器，最大训练轮数设为 240，并在十分钟内完成训练。这项额外的数据变换开销较大，导致训练过程延误了几个训练轮次，进而降低了模型性能。我们认为这一结果是公平的，因为这项额外开销也影响了所有其他提交作品。

表 4：NOMAD-NAS-SENet-18 在 Mini-ImageNet 上的表现。最终验证准确率：五次运行的平均值。数据集是 Mini-ImageNet 的一个子集：训练集包含 5K 张图像，验证集包含 1K 张图像

模型	验证准确率（%）
NOMAD-NAS-SENet-18	86.0

5 结论

ICLR 2021 HAET 挑战赛聚焦于如何在资源有限的条件下训练神经网络，更确切地说，是在数据量和训练时间受限的情况下进行训练。模型压缩领域的大多数研究都致力于优化深度神经网络的推理部分。为了高效训练模型，我们提出了一种方法，能够在极小的数据集和有限的训练时间预算下找到准确率高的分类器。我们使用了 NOMAD，它为神经架构搜索和超参数优化实现了一种无导数优化算法。此外，我们还使用数据增强技术来提升所提出模型的性能。我们的模型 NOMAD-NAS-SENet-18 在 Mini-ImageNet 的一个子集上用时不到 10 分钟，取得了 86% 的成绩，并在比赛中获得二等奖。

参考文献

C. Audet 和 J.E. Dennis, Jr.《用于约束优化的网格自适应直接搜索算法》。《SIAM 优化杂志》, 17(1):188–217, 2006。doi: 10.1137/040603371。Andrew Brock、Soham De、Samuel L Smith 和 Karen Simonyan。《无需归一化的高性能大规模图像识别》。arXiv 预印本 arXiv:2102.06171，2021。Ekin D Cubuk、Barret Zoph、Dandelion Mane、Vijay Vasudevan 和 Quoc V Le。《Autoaugment：从数据中学习增强策略》。arXiv 预印本 arXiv:1805.09501，2018。Terrance DeVries 和 Graham W Taylor。《通过 CUTOUT 改进卷积神经网络的正则化》。arXiv 预印本 arXiv:1708.04552，2017。何恺明、张祥雨、任少卿和孙剑。《用于图像识别的深度残差学习》。载于 IEEE 计算机视觉与模式识别会议论文集，第 770–778 页，2016。Andrew G Howard、Menglong Zhu、Bo Chen、Dmitry Kalenichenko、Weijun Wang、Tobias Weyand、Marco Andreetto 和 Hartwig Adam。《MobileNets：面向移动视觉应用的高效卷积神经网络》。arXiv 预印本 arXiv:1704.04861，2017。Jie Hu、Li Shen 和 Gang Sun。《挤压与激励网络》。载于 IEEE 计算机视觉与模式识别会议论文集，第 7132–7141 页，2018。Alex Krizhevsky、Ilya Sutskever 和 Geoffrey E Hinton。《使用深度卷积神经网络进行 ImageNet 分类》。神经信息处理系统进展，25:1097–1105，2012。Liu Kuang。《Liu/pytorch-cifar》。https://github.com/kuangliu/pytorch-cifar，2021。S. Le Digabel。《算法 909：NOMAD：使用 MADS 算法进行非线性优化》。《ACM 数学软件汇刊》，37(4):44:1–44:15，2011。doi: 10.1145/1916461.1916468。URL http://dx.doi.org/10.1145/1916461.1916468。Karen Simonyan 和 Andrew Zisserman。《用于大规模图像识别的超深卷积网络》。arXiv 预印本 arXiv:1409.1556，2014。Christian Szegedy、Wei Liu、Yangqing Jia、Pierre Sermanet、Scott Reed、Dragomir Anguelov、Dumitru Erhan、Vincent Vanhoucke 和 Andrew Rabinovich。《深入卷积》。载于 IEEE 计算机视觉与模式识别会议论文集，第 1–9 页，2015。Christian Szegedy、Vincent Vanhoucke、Sergey Ioffe、Jon Shlens 和 Zbigniew Wojna。《重新思考面向计算机视觉的 Inception 架构》。载于 IEEE 计算机视觉与模式识别会议论文集，第 2818–2826 页，2016。Christian Szegedy、Sergey Ioffe、Vincent Vanhoucke 和 Alexander Alemi。《Inception-v4、Inception-ResNet 以及残差连接对学习的影响》。载于 AAAI 人工智能大会论文集，第 31 卷，2017。Mingxing Tan 和 Quoc Le。《EfficientNet：重新思考卷积神经网络的模型缩放》。载于国际机器学习大会，第 6105–6114 页。PMLR，2019。Luke Taylor 和 Geoff Nitschke。《利用通用数据增强改进深度学习》。arXiv 预印本 arXiv:1708.06020，2017。Oriol Vinyals、Charles Blundell、Timothy Lillicrap、Koray Kavukcuoglu 和 Daan Wierstra。《用于单样本学习的匹配网络》。载于第 30 届神经信息处理系统国际会议论文集，第 3637–3645 页，2016。