arXiv:2301.05744v1 [cs.LG] 13 Jan 2023

使用残差拟合的自适应神经网络

Noah Ford (Noah.Ford@jhuapl.edu) [1]、John Winder [1], 和 Joshua McClellan [1]

[1] 约翰斯·霍普金斯大学应用物理实验室，，美国马里兰州劳雷尔市约翰斯·霍普金斯路 11100 号，邮编 20723,

2023年1月18日

1 摘要

当前，针对给定问题类别估算所需神经网络规模的方法，主要集中于神经架构搜索和剪枝等计算密集型方法。相比之下，按需增加神经网络容量的方法或许能达到与架构搜索和剪枝相近的效果，同时无需耗费那么多计算资源来确定合适的网络规模。本文提出一种网络增长方法，通过搜索网络残差中的可解释误差，并在检测到足够大的误差时扩展网络。我们通过分类、模仿学习和强化学习任务中的示例展示了该方法。在这些任务中，与不会增长的小型网络相比，增长型网络往往能取得更好的性能；与初始规模大得多的网络相比，其性能则相近。

2 引言

在对一组数据训练神经网络之前，很难估计出能够拟合数据复杂度所需网络的大小或形状。网络拟合数据集的能力还取决于网络的初始化方式，这进一步增加了寻找“合适”网络规模的难度。对于大多数问题而言，如果我们知道如何初始化权重 [1]，那么所需网络规模本可以小于通常采用的规模。由于我们事先对性能最佳网络的结构知之甚少，因此我们可以训练一个远大于所需规模的网络，或者使用神经架构搜索 [2], 网络剪枝[3], 和网络增长等技术，搜索一个合适的较小网络。在本文中，我们提出并测试了一种新的网络增长技术：在较大网络的残差上训练一个小型神经网络。该方法既可以作为神经架构搜索、网络剪枝及其他增长方法的补充，也可以替代这些方法。

NAVAIR 公共发布 2022-360。分发声明 A – “已批准公开发布；分发不受限制”

通用架构搜索可以为具有特定复杂度的一组问题找到合适的网络大小和形状 [4]。这些方法会同时或依次训练多个神经网络，以发现适用于特定问题的正确架构。例如， [5] 使用神经架构搜索，为翻译任务寻找一种效率更高、性能更优的类 Transformer 网络。尽管神经架构搜索能够生成效率更高的网络，但搜索过程本身计算成本很高，因为它需要训练大量网络。

网络剪枝是寻找合适网络结构的另一种方法，但其计算成本通常低于架构搜索。剪枝尤其常用于寻找小型且性能出色的网络，以实现快速推理。已有研究提出了多种标准，用于选择要剪枝的节点及剪枝时机（[6], [3]）。这些方法有助于找到性能出色的小型网络，但通常需要训练大得多的网络，才能找到性能最佳的子网络（[1]）。此外，剪枝方法可能会产生在特定问题上表现良好但无法很好适应数据变化的小型神经网络。

与架构搜索或剪枝相比，网络增长方法受到的研究较少，但网络增长具有一些优势，例如允许初始网络更小，并在训练期间通过额外增长来适应数据中的分布偏移。要增长神经网络，我们必须决定：1. 何时增长；2. 如何增长。为帮助确定何时增长神经网络， [7] 中提出的方法会在损失函数进入平台期时增长神经网络，以摆脱梯度下降减速。为确定如何增长， [8] 中的方法提出了梯度最大化增长，通过最大化新权重的梯度来加速训练。针对非随机问题， [9] 和 [10] 中提出的方法根据网络误差求解神经网络的增长时机和方式，但不易扩展到损失无法降至零的问题。与本文工作最为相似的是 AdaNet [11] 方法，它会在训练前迭代增长神经网络。该方法使用拉德马赫复杂度来估算神经网络的泛化误差，并训练预测泛化误差最低的网络。

所提出的方法与 AdaNet 具有类似的增长机制，但此处的方法通过关注网络的均方误差（MSE），在训练过程中增大网络规模。与 AdaNet 中近似泛化误差的做法不同，我们训练一个残差网络，使其近似基础神经网络的损失。如果残差网络能够预测基础网络损失中的结构，算法便会推断基础神经网络缺少一定程度的网络复杂度，于是将两个网络合并为一个并继续训练。随后，算法会初始化一个新的残差网络，再次拟合基础网络的误差。

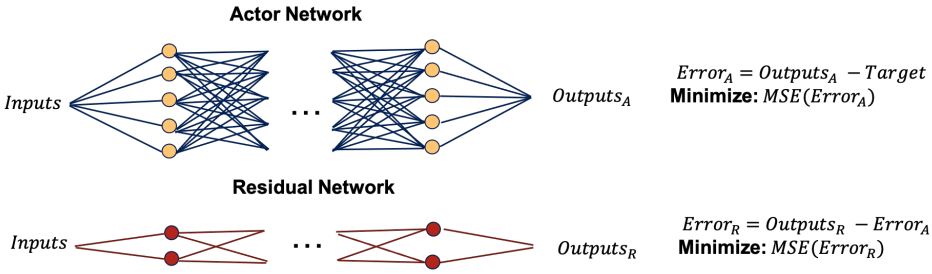


图1：基础网络和残差网络示意图

3 方法

训练一个具有 n 个隐藏层的全连接多层感知机时，我们会初始化一个同样具有 n 个隐藏层、但规模更窄的神经网络。我们称这个网络为残差神经网络。每经过一个训练轮次，我们就训练该残差网络，使其拟合较大网络当前的残差（图 1）。

如果将预测的残差加到较大网络的预测结果中，能使均方误差（MSE）的改善幅度超过用户设定的阈值，我们就通过将残差网络与基础神经网络融合，增大神经网络的潜在维度。合并后的两个网络构成一个新的全连接神经网络，其隐藏层宽度等于基础网络隐藏层的节点数与残差网络隐藏层的节点数之和。

第一层和输出层的权重完全由基础神经网络和残差网络决定。然而，内部隐藏层中，基础网络节点与残差网络节点之间新增了连接。这些权重需要初始化。将这些权重初始化为零，可以使网络扩展后的均方误差（MSE）最小，但反向传播算法会导致这些权重始终彼此相等。这里，我们将新权重随机初始化，使其幅度为残差网络中现有权重的十分之一。

将各个网络融合后，我们继续训练更大的网络。此外，我们重置残差神经网络的权重，并继续使残差网络拟合更大网络的残差。该算法详见算法 1。

由于这项技术着重拟合网络残差中可预测的误差，因此容易出现过拟合。减少过拟合及神经网络后续增长的一种方法是在训练基础神经网络和残差网络时使用 Dropout层。在一些实验中，使用丢弃法的网络表现不如不使用的网络。因此，为避免过度的网络增长，我们不必依赖丢弃法，而是采用一项增长标准：当前均方误差必须比网络上次增长时的均方误差低出增长阈值。例如，如果网络上次增长时的均方误差为

NAVAIR 公共发布 2022-360。分发声明 A –“批准公开发布；不限制分发”

10，且增长阈值为 10%，那么基础网络的均方误差必须达到 9，之后才能再次增长。

算法 1 SANN算法

训练数据： $\{X_i\}$, $\{Y_i\}$，适应阈值 γ
$f(x)$ 神经网络
$g(x)$ 残差网络
$\alpha_{prev} \leftarrow 1 \times 10^6$ ▷ 将 α_{prev} 设为某个较大的数
对于 i 个训练轮次 **执行**
  将 $f(x)$ 拟合到训练数据 $\{X_i\}$, $\{Y_i\}$。收集残差 $\{r_i\}$
  将 $g(x)$ 拟合到 $\{X_i\}$, $\{r_i\}$
  $\alpha \leftarrow$ 均方误差($f(x)$)
  $\beta \leftarrow$ 均方误差($f(x) + g(x)$)
  如果 $\beta/\alpha < 1 - \gamma$ 并且 $\alpha/\alpha_{prev} < 1 - \gamma$ **则**
    通过组合 $f(x)$ 和 $g(x)$ 的权重来扩展网络
    重置 $g(x)$ 的权重
    $\alpha_{prev} = \alpha$
  结束 if
结束 for

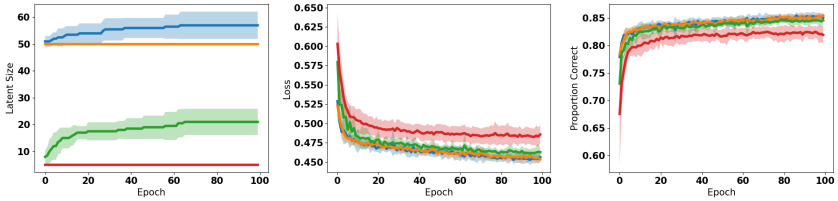
4 结果

4.1 CIFAR

继 AdaNet [11], 之后，我们使用 CIFAR 图像的直方图数据，对这些网络在 CIFAR-10 分类任务上的性能进行比较。图像分类的最先进方法通常包括卷积神经网络，但遵循 [11], ，我们使用图像数据的直方图来检验这种方法的可行性。在这些 CIFAR-10 实验中，我们仅使用每种颜色分布的直方图，每种颜色有 40 个分箱，总计 120 个分箱。我们在留出集上评估这种方法。我们测试了小型固定网络、小型增长网络、大型固定网络和大型增长网络的性能。我们测试大型增长网络的性能，以确保网络规模并非总会增长。对于每种网络类型的每项成对分类实验，我们进行了 10 次训练。

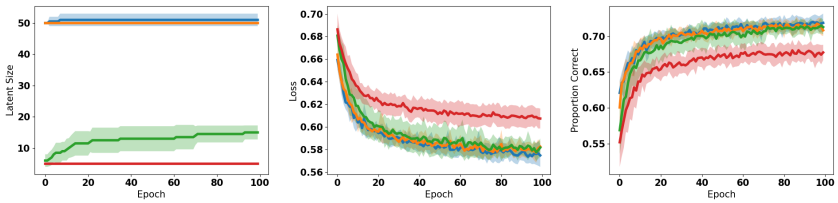
对于鹿与卡车的分类，增长型网络的表现优于小型固定网络，与大型网络相当（图2）。大型增长型网络的规模略有增长，但其均方误差与大型固定网络相近。

在测试网络在鸟与青蛙分类问题上的性能时，我们发现，训练并未像鹿与卡车问题中那样促使网络大幅增长（图3）。与之前一样，增长的网络表现优于小型固定网络，与较大型网络的表现相近。



(a) 潜在维度 (b) 均方误差 (c) 正确率

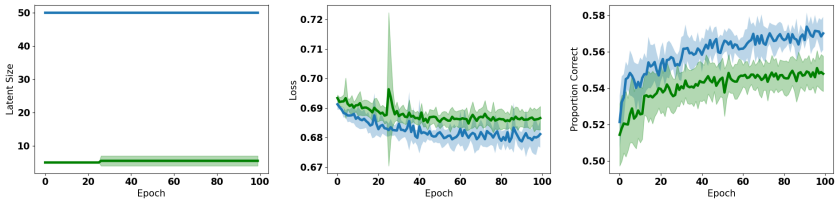
图2：鹿与卡车



(a) 潜在维度 (b) 均方误差 (c) 正确率

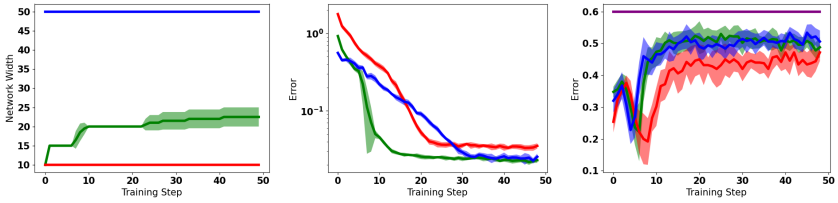
图3：鸟与青蛙

对于猫狗分类问题，小型网络和大型网络通常都不会增长（图4）。这种缺乏增长的现象意味着，残差网络通常不会发现基础网络未能拟合的较大损失。小型网络的表现不如大型网络，但差异相对较小（平均损失小于0。01）。猫狗分类所用的颜色直方图可能没有包含足够的信息，无法让神经网络做出复杂的分类决策。在这种情况下，小型和大型网络中的神经网络都不会增长。



(a) 潜在维度 (b) 均方误差 (c) 正确率

图4：猫与狗



(a) 潜在维度 (b) 均方误差 (c) 得分

图5：使用 DAgger 的 PyUXV

4.2 模仿学习

我们将模仿学习作为自适应神经网络方法的另一项测试任务。在 DAgger算法 [12] 中，我们比较了不同规模网络的性能。DAgger 是一种常用的模仿学习方法，适用于训练期间可以访问环境和专家本身的情况。我们运行 DAgger算法，在 PyUXV 环境中模仿一位专家；该环境也用于 [13]。在此环境中，智能体必须绕开障碍物以到达目标。在这里，我们模仿避开障碍物所需的指令。

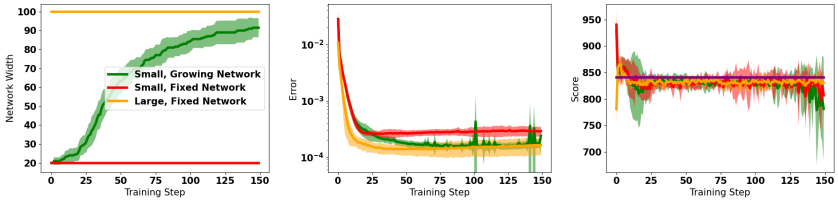
在每种条件下运行 10 次后，我们发现，增长型网络的训练速度快于大型和小型固定网络（图5）。此外，增长型网络取得了与大型网络相近的得分，高于小型网络的得分。

我们使用 MuJoCo 作为基准，测试自适应网络在行为克隆中的性能。我们考虑两种环境：蚂蚁和 Half Cheetah，并在这两种环境中分别进行 10 次迭代。我们使用 10 条专家轨迹进行训练，然后使用另外 10 条未见过的轨迹进行验证。专家轨迹使用 [14] 生成。在这些实验中，我们不使用丢弃法，因为我们发现丢弃法会降低网络的性能。相反，我们通过阻止网络再次增长来避免失控增长，直到网络达到其均方误差改进阈值。

在蚂蚁和 Half Cheetah 中，增长型网络的均方误差会在训练期间的某些时刻突然上升。这些跳升发生在网络增长之后，因为评估是在网络中新随机初始化的权重完成训练之前进行的。

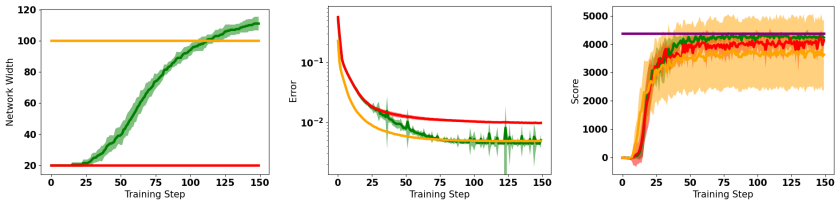
对于蚂蚁环境，增长型网络在均方误差方面优于小型固定网络，并且与大型增长型网络的均方误差相近（图6）。三个网络在该环境中的得分相近，但大型增长型网络的误差在得分上的方差更小。

在 Half Cheetah 环境中，网络平均会增长到比大型固定网络更大的规模。不过，在达到更大规模之前，增长中的网络就能取得与大型网络相近的均方误差（不计网络增长并导致均方误差激增的步骤）。采用增长型网络的模仿器似乎比大型固定



（a）潜在维度　（b）均方误差　（c）得分

图6：蚂蚁上的行为克隆



（a）潜在维度　（b）均方误差　(c) 得分

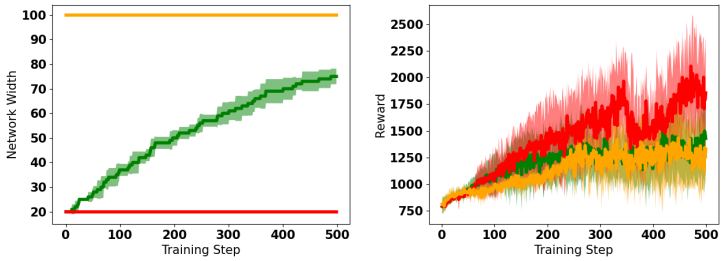
图7：Half Cheetah 上的行为克隆

网络取得了更低的均方误差（图7）。此外，采用小型增长型网络的模仿器获得的平均环境奖励最高。

5 强化学习中的自适应网络

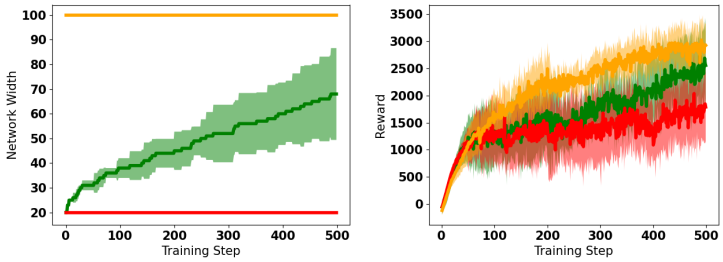
我们将相同的自适应网络技术应用于强化学习算法近端策略优化（[15]）中的价值函数。如果通过添加残差网络可以改善拟合效果，我们就会扩大价值函数网络的规模。在所有实验条件下，策略网络均固定为两个隐藏层，每层宽度为 64 个节点；而价值函数网络则包含两个宽度可变的隐藏层。这些实验中我们不使用丢弃法；不使用丢弃法时，即使是大型网络也可能持续增长。

我们在 MuJoCo 蚂蚁和 Half Cheetah 环境中，使用近端策略优化（PPO）训练不同的价值网络。我们在每种环境的每种条件下进行了五次运行。在蚂蚁环境中，我们发现，渐增网络的表现与大型网络相近（图 8）。然而，使用较小价值网络的近端策略优化运行表现优于使用较大和渐增价值网络的运行。在 Half Cheetah 环境中，使用渐增网络的近端策略优化运行初期表现与小型网络相近，但随着价值



(a) 潜在维度　　(b) 得分

图8：蚂蚁上的强化学习



(a) 潜在维度　　(b) 得分

图9：猎豹上的强化学习

网络不断增大，这些运行的得分开始接近使用大型价值网络的近端策略优化运行的得分（图 9）。

结果表明，采用不断扩大的价值网络的近端策略优化（PPO）运行，其表现可能与采用更大型网络的近端策略优化运行相近，但达到类似熟练程度可能需要更多时间。还需要进一步研究这类不断扩大的价值网络的有效性，并减少网络不必要的扩张。

6 结论

本文研究了一种方法：将一个小型神经网络拟合到另一个神经网络的残差上，并在残差网络发现基础网络的拟合存在显著且可预测的误差时，将两个网络组合起来。通过分类、模仿学习和强化学习中的测试案例，我们发现，即使事先不了解问题的复杂程度，这种方法也能构建出有效的网络。

在许多情况下，这种拟合残差的方法能够达到

与大得多的多层感知机相近的评估损失，适用于多种类型的问题。在拟合 CIFAR 数据时，我们发现，针对不同问题，网络会增长到不同的规模，这或许反映了直方图数据中可用的信息量。在模仿学习中，我们发现，用于行为克隆以及 DAgger算法中的自适应网络，能够达到与起初具有更宽隐藏层的固定规模网络相近的效果。使用自适应价值网络的强化学习训练似乎比使用较大固定网络的训练速度更慢，但也能取得相近的结果。

未来的工作将着重于改进网络组合过程中新权重的初始化，例如采用 [8] 等方法。我们还希望像 AdaNet ([11]) 那样，实现网络深度可变的增长。

致谢

本研究部分由测试资源管理中心 (TRMC) 和测试评估/科学与技术 (T&E/S&T) 计划资助，合同编号为 W900KK-19-C-004。

本材料中表达的任何意见、研究结果、结论或建议均属于作者所有，并不一定反映测试资源管理中心 (TRMC) 和测试评估/科学与技术 (T&E/S&T) 计划的观点。

分发声明 A - 已获准公开发布。

参考文献

[1] J. Frankle 和 M. Carbin，彩票假设：寻找稀疏可训练神经网络（2019），1803.03635。[2] T. Elsken、J. H. Metzen 和 F. Hutter，神经架构搜索：综述（2019），1808.05377。[3] D. Blalock、J. J. G. Ortiz、J. Frankle 和 J. Guttag，神经网络剪枝的现状如何？（2020），2003.03033。[4] Y. Alparslan、E. J. Moyer、I. M. Isozaki、D. Schwartz、A. Dunlop、S. Dave 和 E. Kim，探索二元分类问题中的高效且准确的神经网络架构（2021），2101.06511。[5] D. R. So、C. Liang 和 Q. V. Le，进化型*Transformer*（2019），1901.11117。[6] P. Molchanov、S. Tyree、T. Karras、T. Aila 和 J. Kautz，为资源高效推理剪枝卷积神经网络（2017），1611.06440。

[7]*Y. Kilcher*、*G. B´ecigneul* 和 *T. Hofmann*，通过保持函数不变的结构网络修改逃离平坦区域（*2019*），网址 *https: //openreview.net/forum?id= H1eadi0cFQ*。[8] *U. Evci*、*M. Vladymyrov*、*T. Unterthiner*、*B. van Merri¨enboer* 和 *F. Pe-dregosa*，*Gradmax*：使用梯度信息扩展神经网络（*2022*），*2201.05125*。[9] *M. Liu*、*Z. Cai* 和 *J. Chen*，自适应双层 *ReLU* 神经网络：*I.* 最佳最小二乘近似（*2022*），*2107.08935*。[10] *Z. Cai*、*J. Chen* 和 *M. Liu*，自适应深度神经网络：函数和偏微分方程的数值近似（*2022*），*2109.02839*。[11] *C. Cortes*、*X. Gonzalvo*、*V. Kuznetsov*、*M. Mohri* 和 *S. Yang*，*AdaNet*：人工神经网络的自适应结构学习（*2017*），*1607. 01097*。[12] *S. Ross*、*G. J. Gordon* 和 *J. A. Bagnell*，将模仿学习和结构化预测归约为无悔在线学习（*2011*），*1011.0686*。[13] *C. A. Lowman*、*J. S. McClellan* 和 *G. E. Mullins*，载于《*2021*年 *IEEE/RSJ* 国际智能机器人与系统会议》（*I ROS*）（*2021*），第 *8921–8927* 页。[14]*J. Ho* 和 *S. Ermon*，生成对抗模仿学习（*2016*），*1606.03476*。[15]*J. Schulman*、*F. Wolski*、*P. Dhariwal*、*A. Radford* 和 *O. Klimov*，*CoRR abs/1707.06347*（*2017*），*1707.06347*，网址 *http://arxiv.org/abs/1707.06347*。