网络压缩探幽 (三)

陈超
南京大学

前言

本周实现了一个简易的 ENAS(Efficient NAS), 并在 cifar10 上进行测试.

此外，还阅读了一些 NAS 相关的文章. 之前在周报 (一) 中参考两年前的综述论文[1] 陈列了几个 NAS 未来的发展方向, 现在重新审视一下它们.

发展方向

更优的搜索空间

搜索空间目前朝着变大和变小的两个方向发展. 为了加速训练，B Zoph[8] 等人提出了基本搜索单元 normal cell 和 reduction cell. JD Dong[10] 等人只搜索卷积层和 BN 层的组合方式. C Liu[9] 等人使用相加 (add) 而非拼接 (concatenate) 来结合两个张量，并且让 reduction 操作的通道数固定为 norml 操作的一半，这样就少搜索了两个维度. 前人已经积累了许多设计网络的技巧，如果引入这些先验知识可以极大减小搜索空间，更方便落地. 随之而来的弊端是 NAS 难以搜索到超越人类知识的结构.

为了挖掘 NAS 的潜力，Jieru Mei[7] 将搜索单元划分得更细，搜索空间涵盖了通道数和操作的组合.

更好的搜索策略

尽管搜索策略不一而足，Yu K 等人[12] 对现有的方法产生质疑，他们发现随机搜索的网络性能甚至更好. 个人认为出现这个尴尬的主要原因是搜索空间小，随机搜索更可能遇到最优的网络结构之一.

统一的基准

Dong X[5] 等人为研究者们提供了 15625 个网络结构的训练结果，既有利于评估搜索策略又节省了单纯的模型训练的时间.

可解释性

Yao Shu[6] 等发现一些 NAS 方法搜索得到的 cell 呈现出特殊的结构，可能是因为这种结构具有更快的收敛性.

更多的应用

除了最常实验的图像分类和句法标记，NAS 还被应用于 GAN[2]/语义分割[3]/目标检测[11]/视频[4] 等领域.

实验：ENAS

实验代码见`github.com/passerer/NetworkCompression/blob/master/ENAS.ipynb`

本实验目的是在 cifar-10 上搜寻一个较优的网络结构，采用的算法为 ENAS[13]. ENAS 是一种加速的 NAS, 简单来说就是 RL+one-shot，其加速的动力主要来自于权重共享. 实验流程可总结如下：

1. 构建一个 CNN 超图 (supergraph). 见图 1.
2. 控制器从超图中采样一个子图. 控制器为 lstm.
3. 在训练集上训练子图. 子图权与超图权重共享.
4. 训练控制器. 采用策略梯度方法进行优化，奖励为子图在验证集上的准确率. 为了使控制器的输出更稳定，奖励再减去控制器的熵.
5. 重复 2 至 4 步 n 次.
6. 从训练好的控制器中采样 m 个子图，选择验证准确率最高的一个作为最终模型.
7. 在训练集上微调最终模型

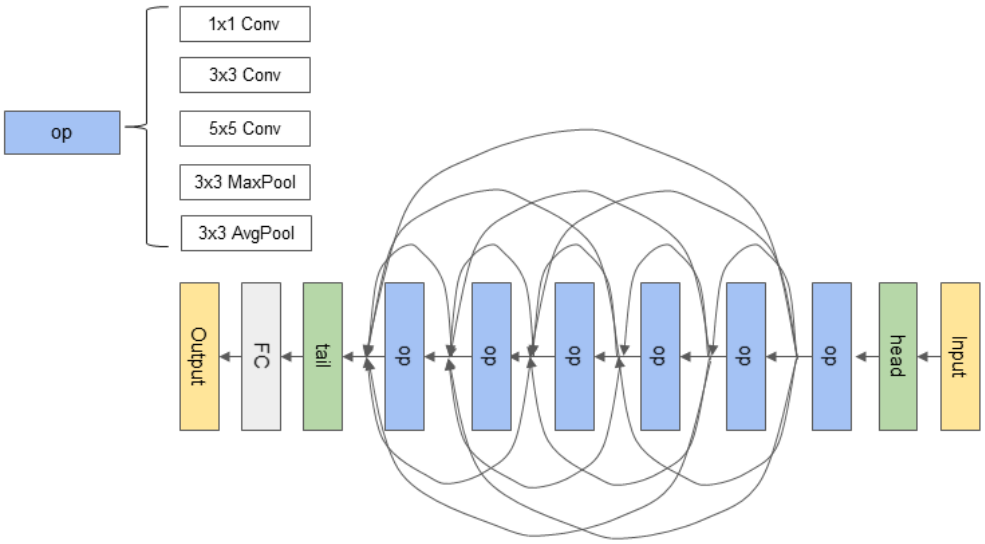

Figure 1: 超图. 黑色的箭头表示数据流动方向. 矩形表示输入输出或操作，其中 op 表示五种操作中的一种，head 和 tail 为辅助层.

为了加速训练，实验采用了如下几个设计：

- 基于链式结构而非 cell. 如果像原文那样基于 cell 搜索，则要分别训练微控制器和宏控制器，较为繁琐. 这里使用链式结构，而且操作只包括 1x1 卷积/3x3 卷积/5x5 卷积/最大池化/平均池化 (池化步长为 1). 控制器除了要决定 5 种操作中选择哪一种，还决定各层之间是否有跳跃连接.
- 固定的通道/激活函数/步长. 搜索空间大为减小，省事很多.
- 小批量验证. 训练控制器时要计算验证集上的准确率. 遍历整个验证集比较耗时，替代的做法是从中随机抽出一小部分.

如图 1所示，网络中间设计了六个节点，每个节点可表示五种操作之一. 搜索空间大小为 $5^6 \times 2^{15} \approx 5 \times 10^8$.

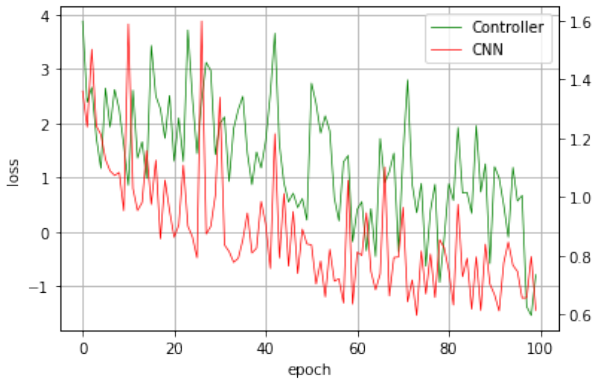

Figure 2: 控制器和 CNN 的损失曲线

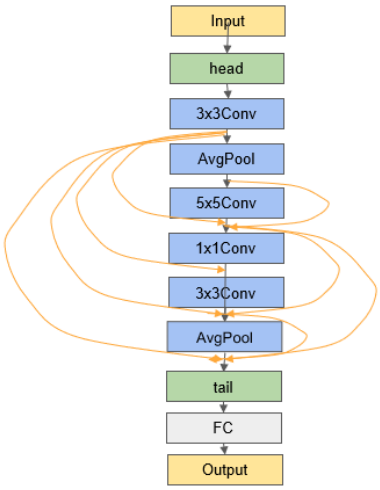

Figure 3: 搜索出的网络结构

图 2展示了 CNN 和控制器 100 轮训练的损失曲线. 因为每轮采样的子图基本上与之前不同的, 所以 CNN 的损失有明显的波动. 控制器的震荡较大, 可能是因为小批量验证, 也可能是因为子图不能适应超图的权重, 二者都会使实验中的准确率不能代表真实准确率.

图 3展示了搜寻到的最优网络结构.

总结

NAS 的可挖的坑感觉蛮多的, 比如搜索空间往大了设计, 可以衍生出一系列新的搜索策略. 如果足够大了, 可能就不叫 " 搜索 " 而叫 " 生成 " .

下次的安排是总结网络压缩的部分论文.

References

[1] Elsken T, Metzen J H, Hutter F. Neural architecture search: A survey[J]. arXiv preprint arXiv:1808.05377, 2018.

[2] Gong X, Chang S, Jiang Y, et al. Autogan: Neural architecture search for generative adversarial networks[C]//Proceedings of the IEEE International Conference on Computer Vision. 2019: 3224-3234.

[3] Chen L C, Collins M, Zhu Y, et al. Searching for efficient multi-scale architectures for dense image prediction[C]//Advances in neural information processing systems. 2018: 8699-8710.

[4] Piergiovanni A J, Angelova A, Toshev A, et al. Evolving space-time neural architectures for videos[C]//Proceedings of the IEEE international conference on computer vision. 2019: 1793-1802.

[5] Dong X, Yang Y. Nas-bench-102: Extending the scope of reproducible neural architecture search[J]. arXiv preprint arXiv:2001.00326, 2020.

[6] Shu Y, Wang W, Cai S. Understanding Architectures Learnt by Cell-based Neural Architecture Search[C]//International Conference on Learning Representations. 2019.

[7] Mei J, Li Y, Lian X, et al. Atomnas: Fine-grained end-to-end neural architecture search[J]. arXiv preprint arXiv:1912.09640, 2019.

[8] Zoph B, Vasudevan V, Shlens J, et al. Learning transferable architectures for scalable image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 8697-8710.

[9] Liu C, Zoph B, Neumann M, et al. Progressive neural architecture search[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 19-34.

[10] Dong J D, Cheng A C, Juan D C, et al. Dpp-net: Device-aware progressive search for pareto-optimal neural architectures[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 517-531.

[11] Chen Y, Yang T, Zhang X, et al. DetNAS: Backbone search for object detection[C]//Advances in Neural Information Processing Systems. 2019: 6642-6652.

[12] Yu K, Sciuto C, Jaggi M, et al. Evaluating the Search Phase of Neural Architecture Search[J]. arXiv preprint arXiv:1902.08142, 2019.

[13] Pham H, Guan M Y, Zoph B, et al. Efficient neural architecture search via parameter sharing[J]. arXiv preprint arXiv:1802.03268, 2018.