网络压缩探幽（四）

陈超
南京大学

前言

本周主要阅读了最新的一些论文，做些记录.

剪枝

Y H[1] 等人根据卷积核的相似性去除冗余的卷积核. 若几何中心接近则可认为相似.

J Chen[2] 等人构造了一个轻量网络来预测每个通道的重要性. 预测网络实际上就是一种 SE 模块[3]. 重要性较低的通道将被抛弃.

NAS

A Bulat[4] 等人将 NAS 与 BNN 结合. 为了避免特征严重退化，BNN 的搜索空间不能用 1x1 卷积和深度可分离卷积, 为了使搜索策略作出更的确定的决定，损失函数引入了 T（参考知识蒸馏的软目标）.

J Yu[5] 等人将 NAS 与知识蒸馏结合. 与 ENAS 同样构建一个超图. 不同地，每次同时训练一个最大模型、一个最小模型、几个随机挑出的模型. 最大模型往往收敛最快，用它的软目标当作其他模型的标签. 训练到最后，超图中任何一个子图都能有很高的准确率，而不需要单独微调。最后根据内存/推理时间要求从中挑出最合适的模型.

C Liu[6] 等人将 NAS 与自监督结合. 在搜索阶段，用自监督的方法训练采样的网络. 在评估阶段，用有监督的方法训练并评估网络。结论是自监督 NAS 可以媲美以前的有监督 NAS.

X Dong[7] 等人为了进一步减少 NAS 的搜索时间，搜索空间仅限于网络的宽度 (通道数) 和深度 (层数).

References

[1] He Y, Liu P, Wang Z, et al. Filter pruning via geometric median for deep convolutional neural networks acceleration[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 4340-4349.

[2] Chen J, Zhu Z, Li C, et al. Self-adaptive network pruning[C]//International Conference on Neural Information Processing. Springer, Cham, 2019: 175-186.

[3] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7132-7141.

[4] Bulat A, Martinez B, Tzimiropoulos G. BATS: Binary ArchitecTure Search[J]. arXiv preprint arXiv:2003.01711, 2020.

[5] Yu J, Jin P, Liu H, et al. Bignas: Scaling up neural architecture search with big single-stage models[J]. arXiv preprint arXiv:2003.11142, 2020.

[6] Liu C, Dollár P, He K, et al. Are Labels Necessary for Neural Architecture Search?[J]. arXiv preprint arXiv:2003.12056, 2020.

[7] Dong X, Yang Y. Network pruning via transformable architecture search[C]//Advances in Neural Information Processing Systems. 2019: 760-771.