智源学者成果展示——机器学习方向

作者　叶杰平（滴滴出行）

2020 年 6 月

项目负责人：叶杰平

AutoCompress: An Automatic DNN Structured Pruning Framework for Ultra-High Compression Rates. AAAI 2020

近年来，随着深度神经网络模型性能不断刷新，模型的骨干网络参数量愈发庞大，存储和计算代价不断提高，从而导致难以部署在资源受限的嵌入式平台上。我们提出了一种基于 AutoML 思想的自动结构化剪枝的算法框 AutoCompress，能自动化的去寻找深度模型剪枝中的超参数，去除模型中不同层的参数冗余，替代人工设计的过程并实现了超高的压缩倍率。从而满足嵌入式端上运行深度模型的实时性能需求。相较之前方法的局限性，该方法提出三点创新性设计：

（1）提出混合型的结构化剪枝维度；

（2）采用高效强大的神经网络剪枝算法 ADMM（交替乘子优化算法）对训练过程中的正则项进行动态更新；

（3）利用了增强型引导启发式搜索的方式进行行为抽样。在 CIFAR 和 ImageNet 数据集的大量测试表明 AutoCompress 的效果显著超过各种神经网络压缩方法与框架。在相同准确率下，实际参数量的压缩相对之前方法最大可以提高超 120 倍。

该工作发表在机器学习顶级会议 AAAI 2020.

Beijing Academy of Artificial Intelligence

微信关注
北京智源人工智能研究院