Knowledge Transfer Office
香港城市大學
City University of Hong Kong

用于自适应压缩和量化的全嵌套神经网络

信息和通信

建筑和施工技术

计算机/人工智能/数据处理和信息技术

机会

这一发明对于基于或辅助神经网络的产品意义重大，例如自动驾驶汽车、视频监控、智能物联网设备和智能家居业务。全球自动驾驶汽车和卡车市场规模预计在2020年将达到约6700台，并预计在2021年至2030年期间以63.1%的复合年增长率持续增长。视频分析与智能视频监控市场在2018年达到了281.3亿美元，预计到2027年底将达到1038.3亿美元，全球复合年增长率为15.14%。智能物联网的市场规模在2018年为1900亿美元，智能家居在2018年的市场规模为799亿美元，预计到2026年将达到6225.9亿美元，预测期内的复合年增长率为29.3%。

技术

神经网络的压缩和量化是将最先进的模型适配到移动设备和嵌入式硬件的计算、内存和功率限制中的重要任务。最近的模型压缩/量化方法基于强化学习或搜索方法，将神经网络量化为特定的硬件平台。然而，这些方法需要多次运行以将同一基础神经网络压缩/量化到不同的硬件设置中。在这一发明中，我们提出了一种完全嵌套神经网络（FN3），只需运行一次即可构建嵌套的压缩/量化模型集，这些模型在不同资源约束下是最优的。具体而言，我们利用神经网络中不同层次构建块的累加特性，并提出了一种有序丢弃（ODO）操作来对构建块进行排序。给定一个训练好的FN3，离线运行一个快速启发式搜索算法，找到在不同约束条件下最大化精度的组件移除方案。实证结果验证了所提出方法的强大实用性能。

优势

- 适用于广泛的神经网络组件
- 更好的预测精度
- 部署神经网络的更大灵活性

应用

- 自动驾驶车辆
- 视频分析
- 智能物联网
- 智能家居

IP状态
专利已存档

技术成熟度等级 (TRL) 4

发明人

Prof. CHAN Antoni Bert

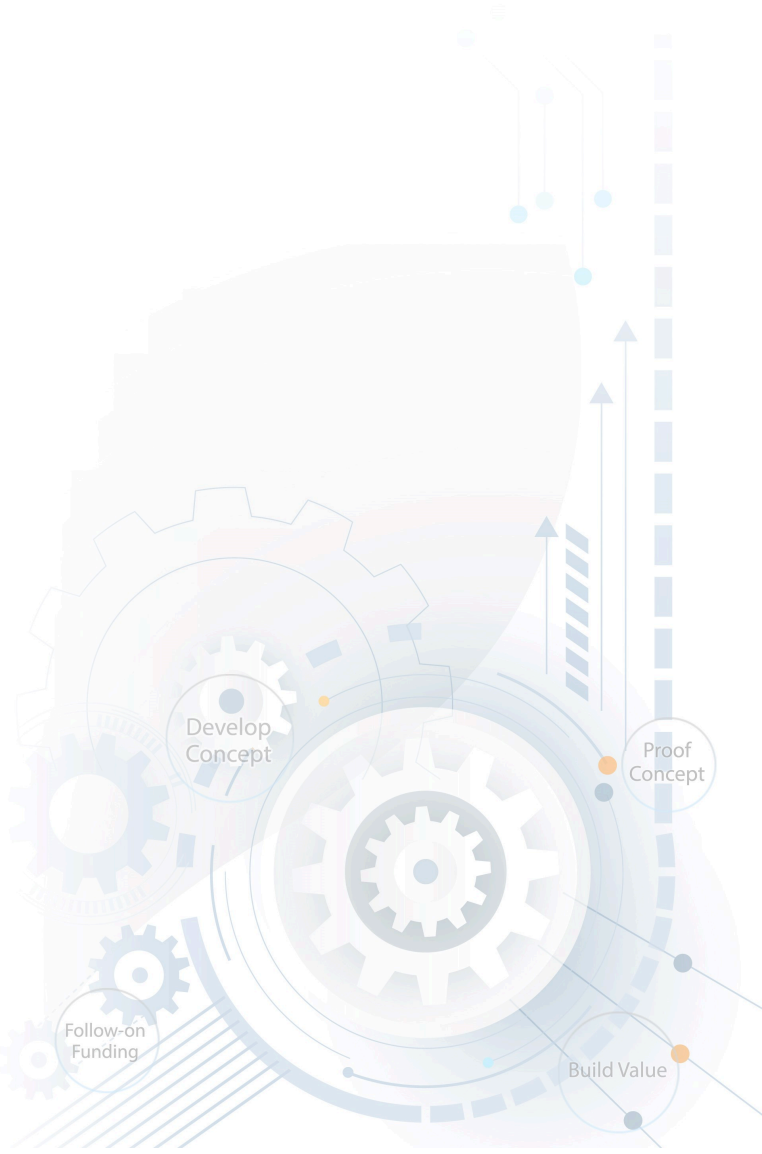
崔宇飞

李乔

刘子泉

姚武冠楠

询问: kto@cityu.edu.hk


Develop
Concept
Proof
Concept
Follow-on
Funding
Build Value