Neural Network Optimization for Machine Vision Edge Devices with Model Distillation Techniques

Robert W. Bickley[1,2], in conjunction with the FLIR IIS Research Division, based on work by Han Cai[3] et al.
[1] University of Victoria, [2] NTCO, [3] Massachusetts Institute of Technology

Background / Motivation

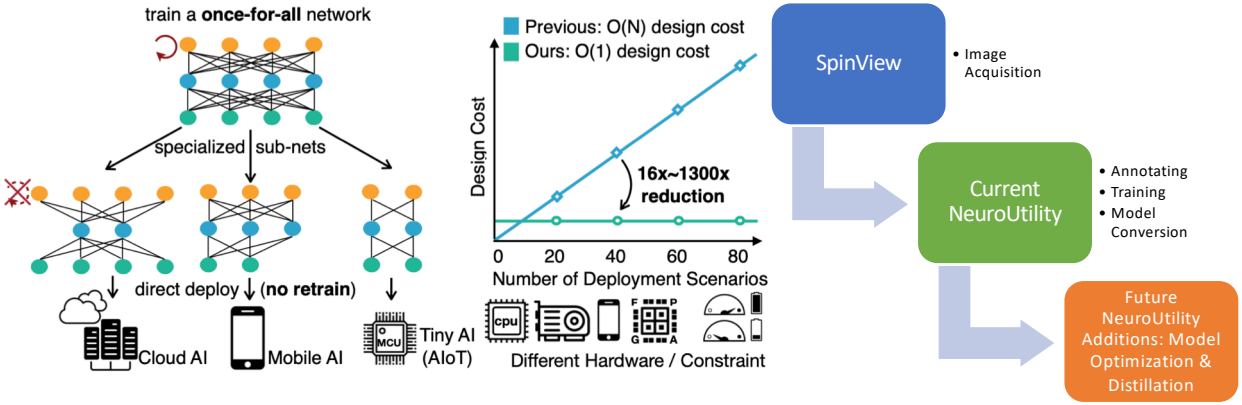
- Machine learning, deep learning, and Convolutional Neural Networks (CNNs) / Deep Neural Networks (DNNs) have gained favor in the last decade for industrial and academic image classification and detection problems.
- DNN accuracy is such that the number of deployment scenarios (i.e. specific integrated workflows) has skyrocketed in recent years.
- Unique hardware systems with a range of computational abilities previously necessitated the costly construction and training of multiple highly specified networks.
- Promising new techniques have allowed for singly-trained networks to be adjusted on the fly to meet the needs of a range of hardware systems and latency constraints.

Challenge: Applications to the Firefly Camera

- FLIR's deep learning Firefly camera and software allow for user-end training and deployment of specialized neural networks for image classification or object detection.
- The camera is paired with a suite of software for data collection and model preparation.
- A range of customers stand to benefit from adjustable latency parameters.
- Model distillation techniques like Once-For-All (Cai et al. 2019) pave the way for customers to specify deployment scenarios to fit individual computational constraints without sacrificing accuracy.

Proposed Solution: Once-For-All

- In their 2019 publication, Cai et al. at the Massachusetts Institute of Technology detailed the Once-For-All model distillation technique.
- Once-For-All uses a progressive shrinking algorithm in order to identify countless sub-networks within a larger pre-trained network that retain much of the parent model's performance, without retraining.
- In a Firefly-based use case, the Once-For-All technique will be used to tune the camera's performance and latency with careful consideration of accuracy retention.
- Transfer learning on the fully-connected layers will allow users to further customize the camera's performance for specific applications.



The FLIR Firefly DL camera, and FLIR Spinnaker software for user-end data collection (reproduced from flir.com). Customers can gather images for their specific training scenario following instructions in Spinnaker. FLIR's NeuroUtility software tool is used for conversion between neural network formats and uploading to the Firefly DL camera, and will be leveraged for distillation.

Reproduced from Cai et al. 2019: the Once-For-All workflow from bulk training, to sub-network identification via progressive shrinking, and finally deployment without retraining on a diverse suite of cloud and edge devices (left), and the design cost versus number of deployment scenarios for traditional deployment architectures (blue) versus once-for-all (green) (center). Right: AI Development Workflow, outlining the present and future roles of available software tools.

Acknowledgements:

- FLIR IIS Research Group, including Stephen Se, Di Xu, Douglas Chong, and others
- NTCO-CREATE
- University of Victoria Faculty of Graduate Studies
- University of Victoria Physics & Astronomy Department

Future Work

- Investigating the potential value of incorporating Once-For-All, as well as other novel model distillation techniques, into the existing framework for Firefly camera model preparation
- Preserving excellent performance while offering lighter-weight networks for edge deployment
- Providing an opportunity for customers to manually shrink the computational expense and predicted inference latency of a pre-trained model, a novel offering in the intersection between edge devices and machine vision.