Differentiable Network Pruning to Enable Smart Applications

EDGAR LIBERIS, University of Cambridge, United Kingdom
NICHOLAS D. LANE, University of Cambridge Samsung AI Centre Cambridge, United Kingdom

Wearable, embedded, and IoT devices are a centrepiece of many ubiquitous computing applications, such as fitness tracking, health monitoring, home security and voice assistants. By gathering user data through a variety of sensors and leveraging machine learning (ML), applications can adapt their behaviour: in other words, devices become "smart". Such devices are typically powered by microcontroller units (MCUs). As MCUs continue to improve, smart devices become capable of performing a non-trivial amount of sensing and data processing, including machine learning inference, which results in a greater degree of user data privacy and autonomy, compared to offloading the execution of ML models to another device.

Advanced predictive capabilities across many tasks make neural networks an attractive ML model for ubiquitous computing applications; however, on-device inference on MCUs remains extremely challenging. Orders of magnitude less storage, memory and computational ability, compared to what is typically required to execute neural networks, impose strict structural constraints on the network architecture and call for specialist model compression methodology. In this work, we present a differentiable structured pruning method for convolutional neural networks, which integrates a model's MCU-specific resource usage and parameter importance feedback to obtain highly compressed yet accurate models. Compared to related network pruning work, compressed models are more accurate due to better use of MCU resource budget, and compared to MCU specialist work, compressed models are produced faster. The user only needs to specify the amount of available computational resources and the pruning algorithm will automatically compress the network during training to satisfy them.

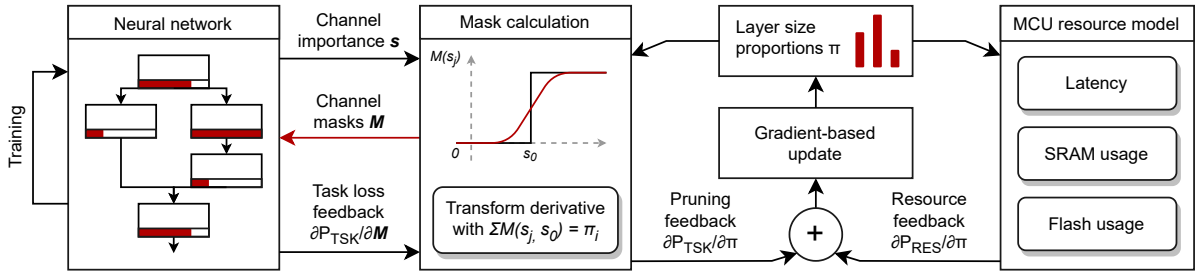

Fig. 1. A diagrammatic summary of the implementation of microcontroller-aware differentiable pruning. Pruning provides channel masks $\mathbf{M}$ to the neural network, and in return receives channel importance measures s, and task loss feedback $\frac{dPtsk}{dM}$. The latter is transformed into feedback with respect to layer size proportions π. That, and the feedback from MCU resource model, are combined to produce gradient-based updates to π.

We evaluate our methodology using benchmark image and audio classification tasks and find that it (a) improves key resource usage of neural networks up to 80×; (b) has little to no overhead or even improves model training time; (c) produces compressed models with matching or improved resource usage up to 1.7× in less time compared to prior MCU-specific model compression methods.

Acknowledgements. This work was supported by Samsung AI and by the UK's Engineering and Physical Sciences Research Council (EPSRC) with grants EPM50659X1 and EPS0015301 (the MOA project) and the European Research Council via the REDIAL project (Grant Agreement ID: 805194).