SpArSe: Sparse Architecture Search for CNNs on Resource-Constrained Microcontrollers

Igor Fedorov (Arm ML Research), Ryan P. Adams (Princeton University), Matthew Mattina (Arm ML Research), Paul N. Whatmough (Arm ML Research)

The vast majority of processors in the world are actually microcontroller units (MCUs), which find widespread use performing simple control tasks in applications ranging from automobiles to medical devices and office equipment. The Internet of Things (IoT) promises to inject machine learning into many of these every-day objects via tiny, cheap MCUs. However, these resource-impoverished hardware platforms severely limit the complexity of machine learning models that can be deployed. For example, although convolutional neural networks (CNNs) achieve state-of-the-art results on many visual recognition tasks, CNN inference on MCUs is challenging due to severe memory limitations. To circumvent the memory challenge associated with CNNs, various alternatives have been proposed that do fit within the memory budget of an MCU, albeit at the cost of prediction accuracy. This work challenges the idea that CNNs are not suitable for deployment on MCUs. We demonstrate that it is possible to automatically design CNNs which generalize well, while also being small enough to fit onto memory-limited MCUs. Our Sparse Architecture Search method combines neural architecture search with pruning in a single, unified approach, which learns superior models on four popular IoT datasets. The CNNs we find are more accurate and up to 7.4x smaller than previous approaches, while meeting the strict MCU working memory constraint.