Resource Efficient ML in a few KBs of RAM

Prateek Jain, Senior Principal Researcher, Microsoft Research India

Several critical applications require ML inference on resource-constrained devices, especially in the domain of Internet of Things like smartcity, smarthouse etc. Furthermore, many of these problems reduce to time-series classification. Unfortunately, existing techniques for time-series classification like recurrent neural networks are very difficult to deploy on the tiny devices due to computation and memory bottleneck. In this talk, we will discuss two new methods FastGRNN and SRNN that can enable time-series inference on devices as small as Arduino Uno that have 2KB of RAM. Our methods can provide as much as 70x speed-up and compression over state-of-the-art methods like LSTM, GRU, while also providing strong theoretical guarantees.