Demo Title: Online Hand Gesture Recognition with Temporal Shift Module (TSM)

Presenter: Song Han, Assistant Professor, EECS, MIT

Type of items to be demonstrated:
We are going to present an online hand gesture recognition demo on an input video stream from the camera. Recognizing hand gestures requires not only spatial features but also temporal cues in a video (e.g., swipe left or swipe right). Conventional 2D CNNs are computationally cheap but cannot capture temporal relationships; 3D CNN based methods can achieve good performance but are computationally intensive, making it expensive to deploy. To this end, we propose an efficient video understanding model called Temporal Shift Module (TSM). TSM shifts parts of the channels along the temporal dimension to enable temporal information exchange among neighboring frames, which comes at zero computation and zero parameters. In this way, we can achieve the performance of 3D convolutions using just 2D convolution. Our model is also hardware-friendly due to the highly optimized 2D convolution, enabling real-time low-latency online video recognition. Our work is featured by MIT News, MIT Tech Review, WIRED, Engadget, and NVIDIA Jetson forum.

Relevance to tinyML:
Video recognition models usually consume a lot of computation and parameters, making it hard to deploy on edge devices. Making deep learning models tiny has drawn people's attention in image recognition models, but little has been paid to video understanding models. With the explosive growth of video data, we believe it is important to give a solution to the problem. Our proposed TSM helps to achieve real-time, low-latency, low-power, online video recognition on an edge device Jetson Nano, using only 5 watts and running at 70 FPS. Our model is among the first tiny-scale models in the video recognition domain, which suits the tinyML topic.