Fresh Air

On Efficient Machine Learning

IoT + AI = Future ?

Microsoft to launch its biggest AI, IoT lab in Shanghai

JAN 17, 2019 | IN HEAVY HITTERS, WITH CHINESE CHARACTERISTICS | BY JILL SHEN

1 min read

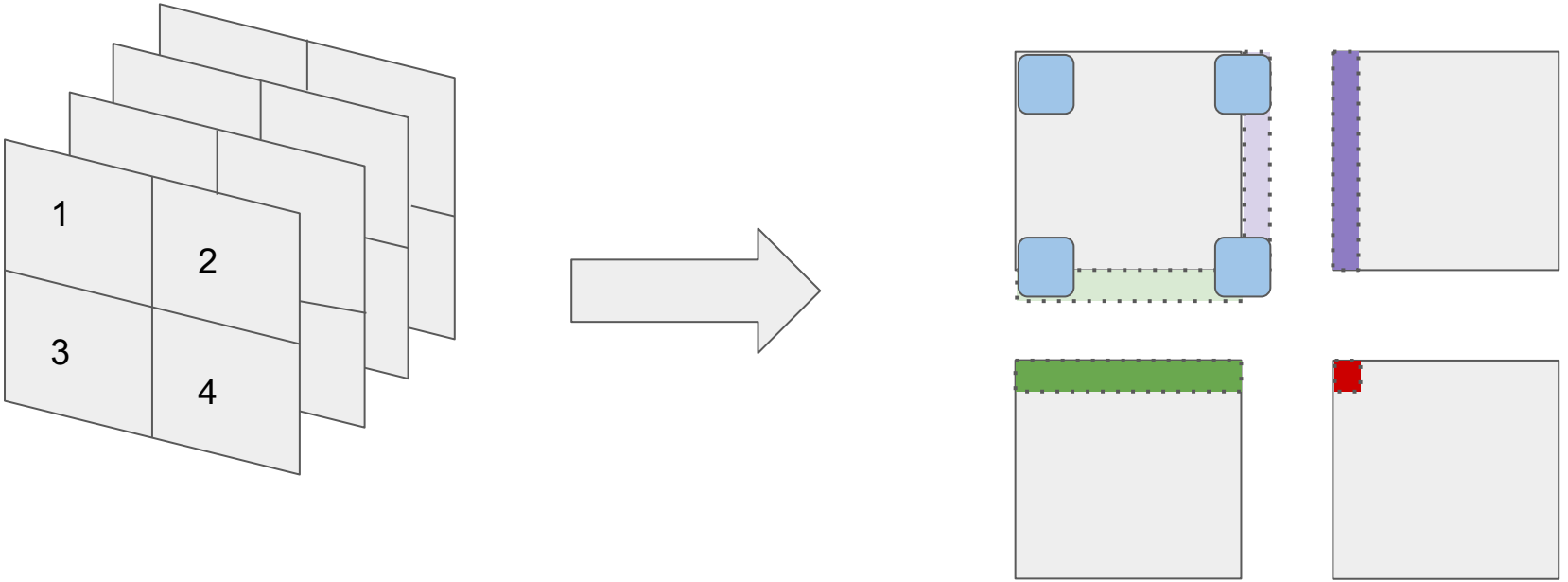
ML + IoT

- Embedded devices:
  - Number is large
  - Individual device is weak
- End-devices + Cloud
  - Rely on capability of networking
  - Transmission is power consuming

Can we split a neural network on **several** devices ?

Level of Partition

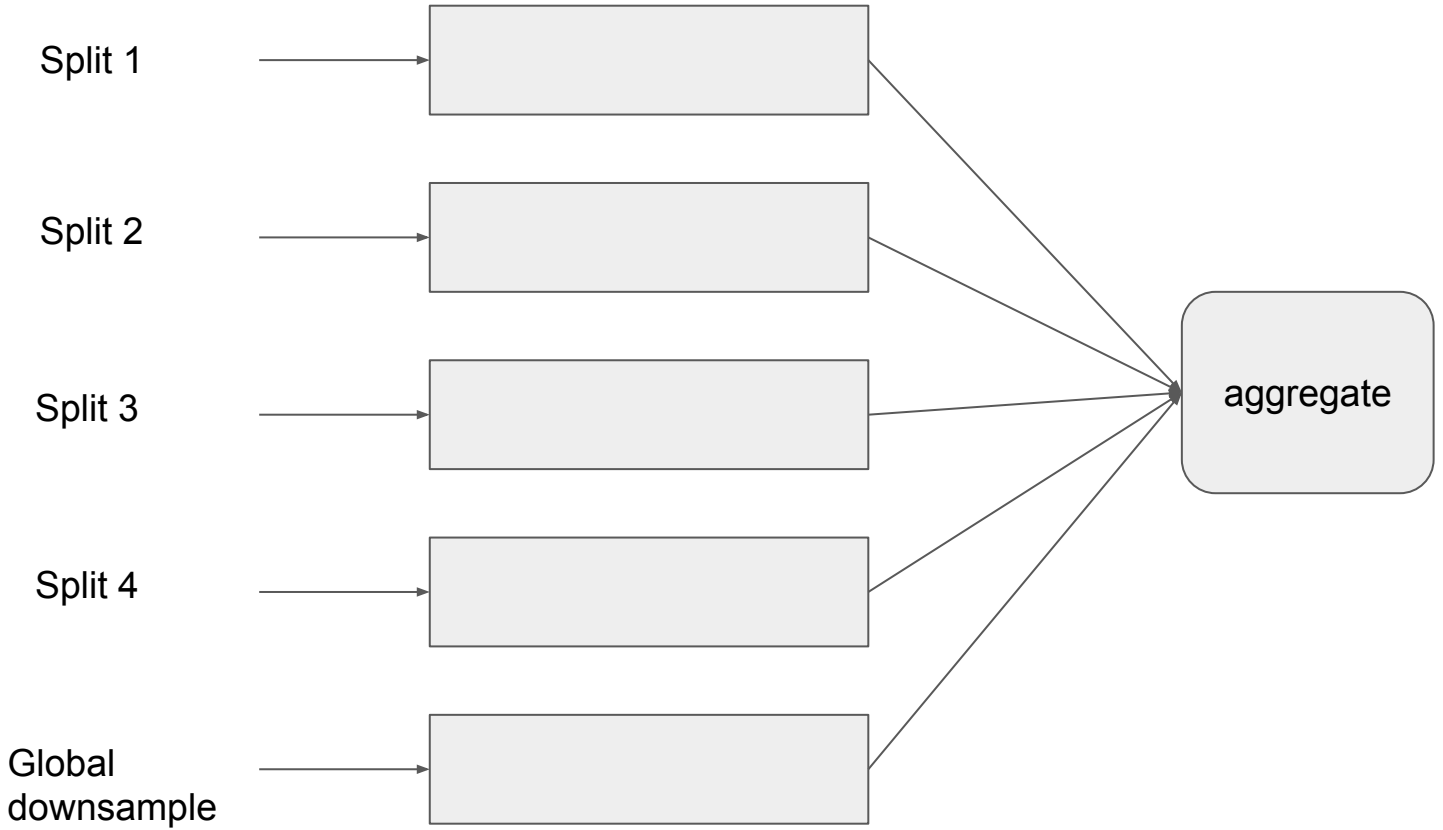
Level	Method	Transmission	Feature
Layer	Different layers on different devices	Large, have to transfer all feature map	High latency, but “early exist”[1][2] can be utilized to reduce latency.
Crop	Crop the input image directly or downsample image [3]	Zero	Accuracy is lower
Channel	Different channels on different devices	Large, have to transfer feature map for every layer	High latency. Transmission introduce loss.
Spatial	Split feature map	Low	Low latency, and high accuracy


1
2
3
4

Reduce transmission

1. Introduce more sparsity with ReLU(x-threshold) for only edge pixels
2. Quantize edge pixels to reduce bits
3. Introduce transmission loss/noise into training

Future Work



Federated Learning

- Data is local
- Training is local
- Good for privacy
- Fast iteration for produce