MLPerf Edge Inference

- The MLPerf Inference is still in discussion about:
  - Metrics to be decided:
    - Power, cost, accuracy, latency etc.
    - The accuracy/throughput/latency tradeoffs are different in batch and online inference with totally different models
  - Framework must be available and widely-used
    - PyTorch/Caffe2, TensorFlow, ONNX-compatible frameworks, etc.
  - Task/Scenario
    - Each task is assigned to a certain model using distinct framework and model
    - ONNX currently used for image classification/segmentation
  - Boundary for Edge device
    - For example for an edge video monitor , how the test dataset is read: thru file system or lens? Should its boundary include the CMOS sensor, auto-focus lens for performance /power?

HUAWEI

ONNX Edge & MLPerf Edge Inference

- The MLPerf community right now is discussing on the categories of each scenario for the **Inference Rule**
- Metrics including accuracy, size, speed, power are still under discussion. No agreement for accuracy while huge disagreement for speed and power.
- Model size is not talked about yet, and all the ONNX model for MLPerf tasks are non-quantized version. MLPerf community considers quantization for mobilenet and ssd specifically while other models will be measured based on FP32

HUAWEI

Future Plans

ONNX Edge

- Scenario coverage
- Low-precision op, data type and quantized models for edge
- Meet the requirements for each metric
- Cooperating with other WGs (Quantization etc.)
- Verification with MLPerf Edge Inference

MLPerf

- Power measurement tools for edge inference
- More edge benchmarking metrics and DUTs definition

HUAWEI

Thank you