Efficientnet b0 Model Card

Model Card

Model Details

- The EfficientNet-B0 INT8 model is a quantized deep learning model optimized for efficient image classification, designed to run with Intel's Deep Learning Streamer (DL Streamer) for high-performance inference on Intel hardware. It is based on the EfficientNet-B0 architecture, which balances accuracy and computational efficiency. It takes a 224x224 RGB image as input and outputs a softmax probability distribution over 1,000 ImageNet classes, predicting the most likely category for the given image.

Intended Use

- Our application uses this model for AI inferencing on input video and we collect metrics while the pipeline is running
- The model is primarily used for general-purpose image classification tasks, making it suitable for applications like automated tagging, visual search, and edge AI deployment.
- The INT8 quantization enables lower latency and reduced computational cost, making it ideal for real-time inference in resource-constrained environments.

Training and validation data

- We are not training or validating this model in our reference implementation

Ethical Considerations

- We are using person-bicycle-car-detection.mp4 from https://github.com/intel-iot-devkit/sample-videos as input video to test this application tool.
- We are not storing any person or user related personal information.

Caveats and Considerations

- The model's accuracy may vary depending on the quality and resolution of the input images. Ensure that the images used are of sufficient quality for reliable detection.
- Preprocess images to normalize lighting conditions and remove noise.

Quantitative Analysis

- We are not doing quantitative analysis in this application tool but we do display metrics mentioned below to the user.

Factors

- We are also not evaluating this model in this reference implementation

Metrics

- We are displaying metrics including throughout (FPS) and system level metrics: CPU/GPU utilization, memory utilization, CPU/GPU frequency, CPU/system temp, GPU power, GPU engine, and package power. In this application these metrics are collected and displayed to the user via gauges.