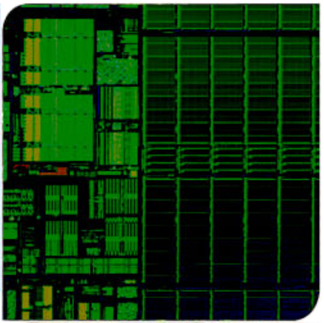
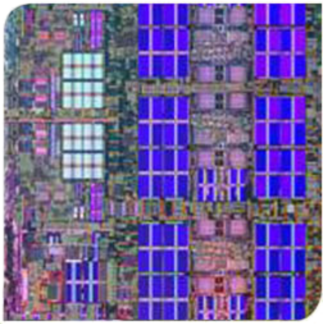
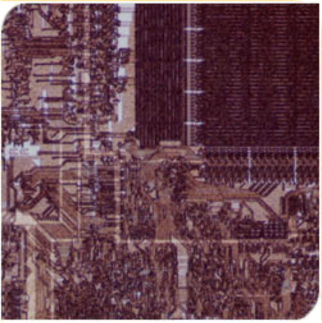




Work-In-Progress: Video Analytics From Edge To Server



Jiashen Cao, Ramyad Hadidi, Joy Arulraj, Hyesoon Kim

Motivation

- Camera systems generate massive amount of data nowadays.
  - According to Lucid Motors, 6 – 12 cameras are able to produce 60 – 400 MB data per second.
  - It is no longer possible to analyze large-scale data by hands.
- The advancements in **deep neural networks** encourage engineers to use it to understand data without manual efforts.
- In a system, more devices (cameras, sensors) are deployed on the edge.
  - More computation resources are available on the edge.
  - Edge devices are usually under-utilized in the system.

Motivation

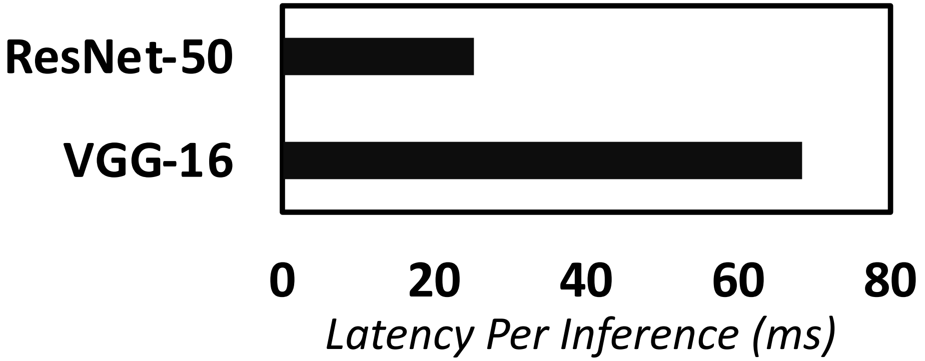
- Camera systems generate massive amount of data nowadays.
  - According to Lucid Motors, 6 – 12 cameras are able to produce 60 – 400 MB data per second.
  - It is no longer possible to analyze large-scale data by hands.
- The a ge engin l effort

Deep Neural Networks based System Processes Real-Time Inference On The Edge

- In a system, more devices (cameras, sensors) are deployed on the edge.
  - More computation resources are available on the edge.
  - Edge devices are usually under-utilized in the system.

Challenges

- Deep neural networks inferences are compute intensive.
  - VGG-16 model has 16 GFLOPs.
- Each edge device has limited computation resource.
  - A Nvidia TX2 development board.
    - □ 2 GHz ARM CPU processor and a low end GPU.
- As results,
  - Limited computation resource causes longer latency.

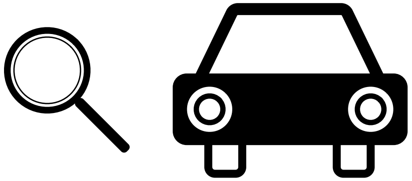


Observation

- In video analytics system, not all requests have the same accuracy requirements.
  - To identify the license plate number of a vehicle, the system needs to run deep neural network prediction with high accuracy.
  - To estimate number of cars passing a traffic intersection, the system requires lower accuracy support.

Scenario 1

Scenario 2

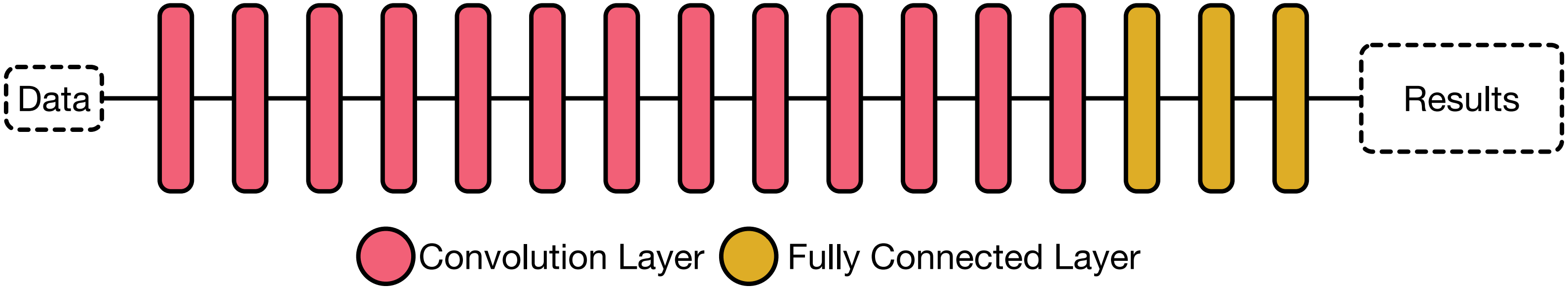


- Opportunities exist to leverage accuracy and improve the performance.

Our Approach

- A Multi-Stage Neural Network.
  - Support multiple accuracy requirements in a single model.
  - Stop in the middle of inference if accuracy requirements are met.
- Conduct case study on VGG-16.

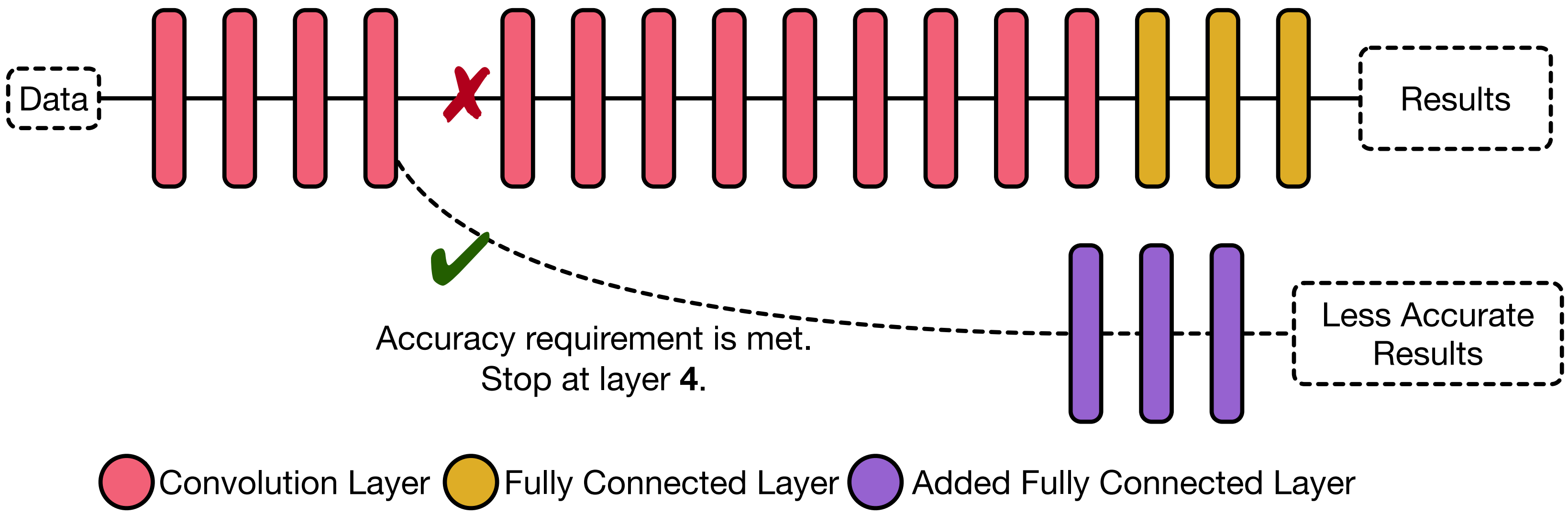
Original VGG-16



Our Approach (cont'd)

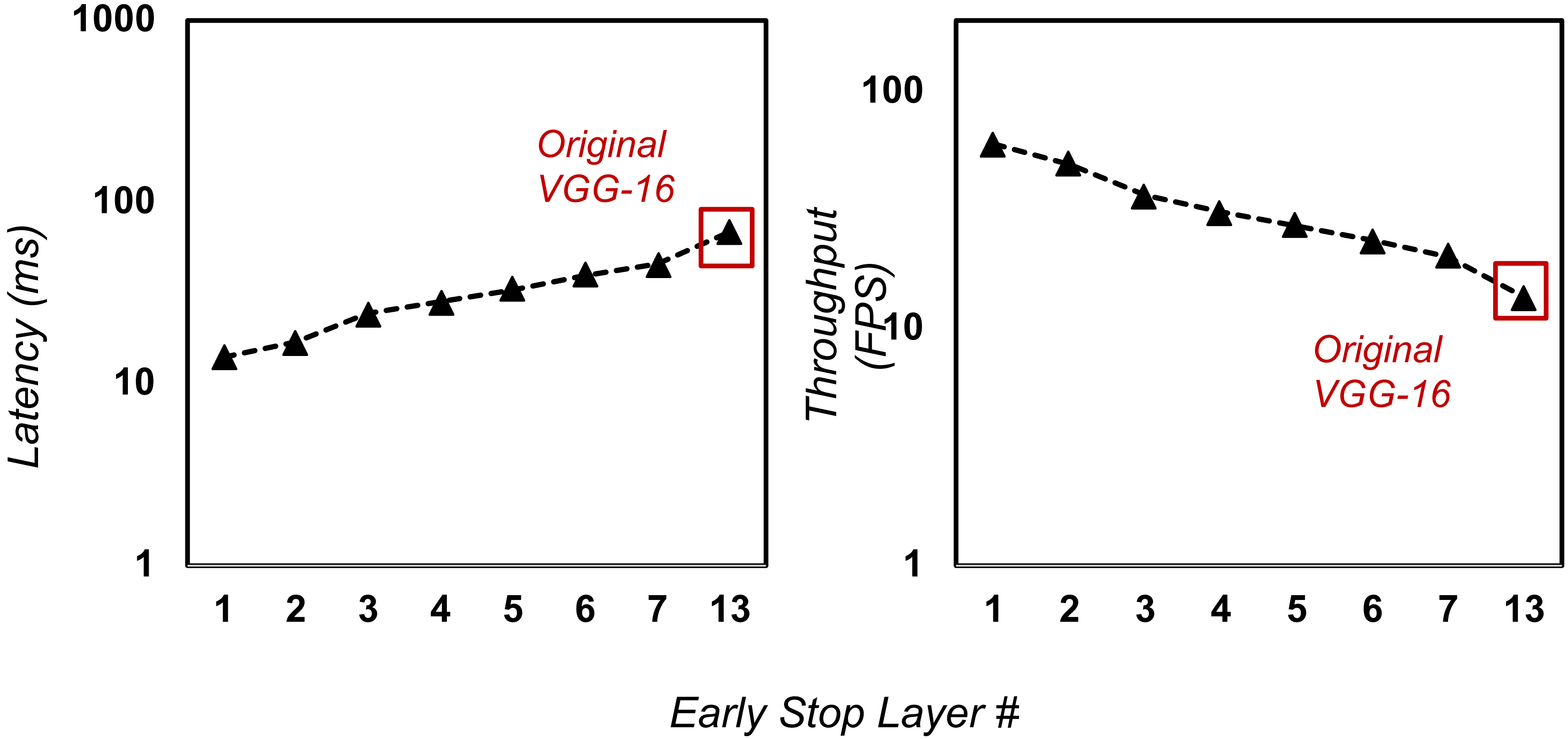
- Multi-Stage VGG-16 properties.
  - Add customized fully connected layers to shallow convolution layers.
  - Inferences stop early if accuracy requirements are met.

An Example of Multi-Stage VGG-16



Preliminary Results

- Profile performances on Nvidia TX2.


Latency (ms)
Original
VGG-16
Throughput
(FPS)
Original
VGG-16
Early Stop Layer #