Computation Offloading Scheduling for Deep Neural Network Inference in Mobile Computing

Yubin Duan and Jie Wu
Dept. of Computer and Information Sciences
Temple University, USA

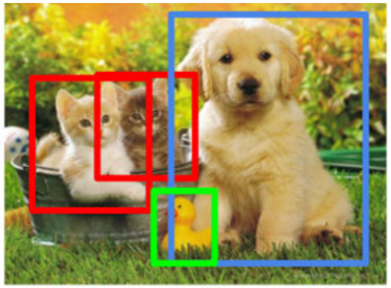
IWQoS 2021

Outline

1. Introduction
2. Model
3. Tree-structure DNN Scheduling
   - Path-wise scheduling
   - Layer-wise scheduling
4. Scheduling for DAG-Style DNNs
5. Experiment
6. Conclusions

1. Introduction

- DNN inference in mobile applications
  - Image classification
  - Object detection

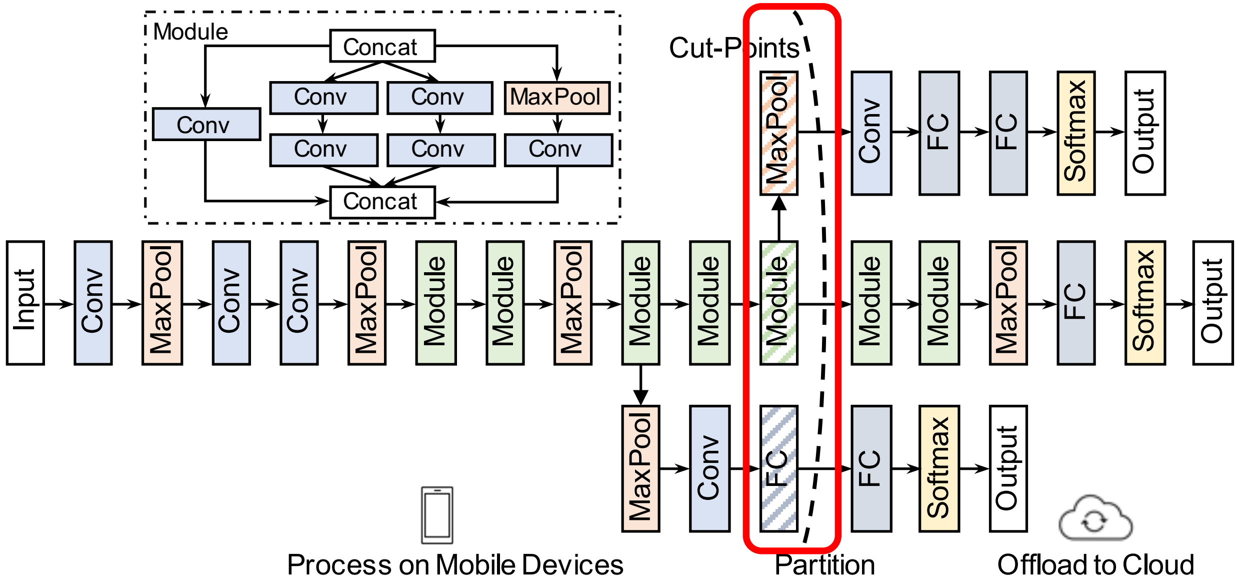


CAT, DOG, DUCK

- QoS measurement
  - Inference latency
- Cooperative DNN inference
  - Computation offloading

Motivation

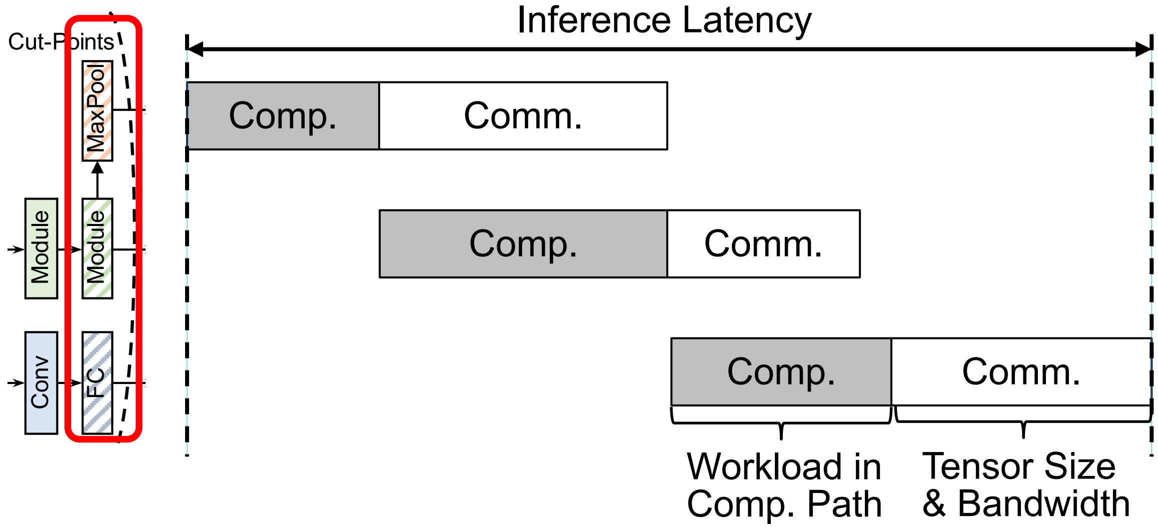
- Offloading pipeline
  - May have multiple offloading subtasks



- Scheduling problem
  - Computation and communication priorities

2. Model

- Two-stage offloading pipeline



  - Comp.: process DNN layers from input to cut-points
  - Comm.: upload intermediate results to cloud servers
  - Cloud processing time is negligible

Problem Formulation

- Objective
  - Minimize inference latency for a given DNN

$$\min_{\sigma} \quad \tau = t_{|S|} - t_0$$

  - Recursive calculation of the completion time

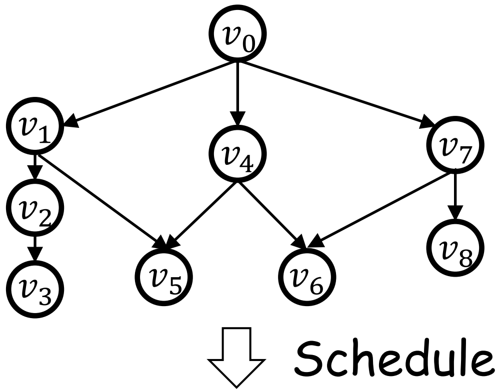
$$t_i = \max\{t_0 + \sum_{k=1}^{i} f(x_i), t_{i-1}\} + g(x_i), \forall x_i \in \sigma$$

- Constraints
  - Precedence constraint

$$i \leq j, \forall x_i \prec x_j, \forall x_j \in \sigma$$

  - Permutation constraint

$$\cup_{x_i \in \sigma} x_i = S, |\sigma| = |S|$$

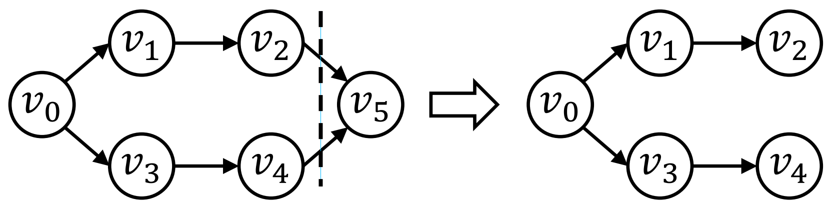


$\sigma = [x_0, x_1, \ldots, x_8]$

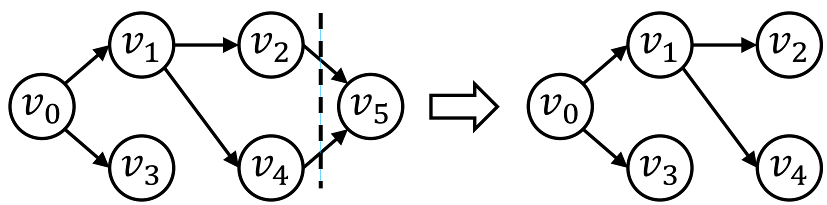
A permutation of $\{v_0, \ldots, v_8\}$

DNN Structures after Partition

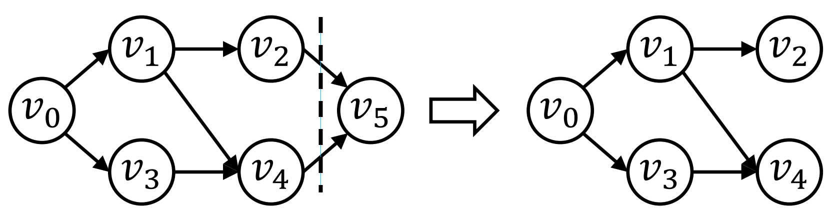
- Tree-structure
  - Multi-path tree



  - General tree

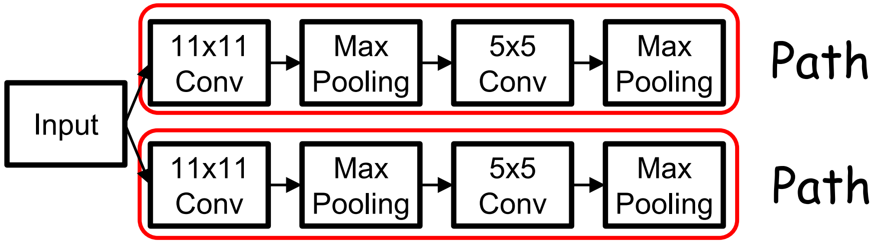


- DAG

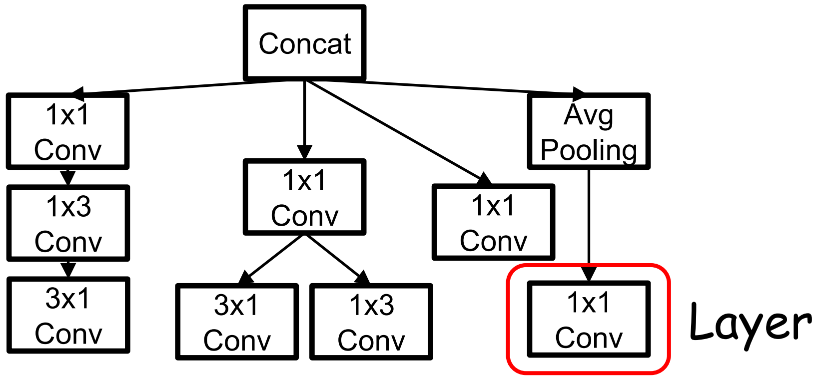


3. Tree-structure DNN Scheduling

- Scheduling granularities
  - Path-wise scheduling
    - Can be optimally solve by applying Johnson's rule

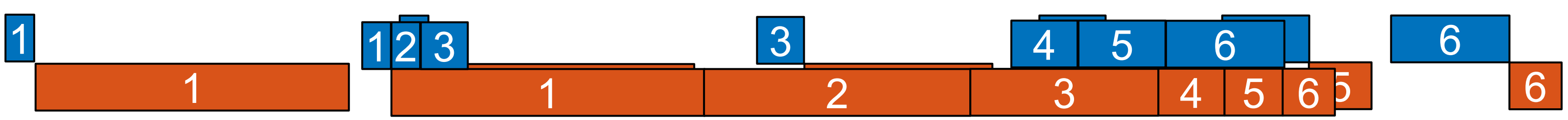


  - Layer-wise scheduling
    - Extend Johnson's rule for the optimal solution



Path-wise Scheduling

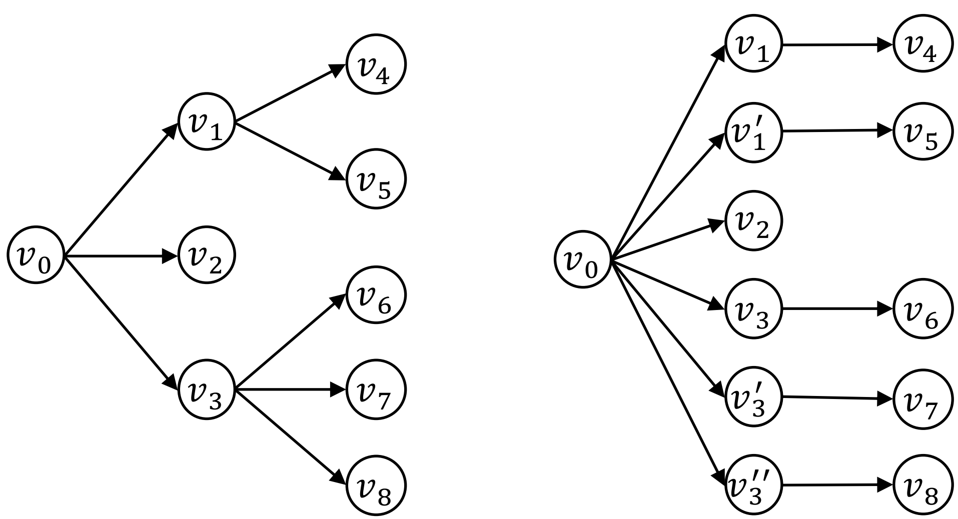
- Optimal scheduling with Johnson's rule
  - Each path is a task with two operations
  - Split tasks into comm./comp.-domination groups
  - H = {1, 2, 3}, increasing order of comp.
  - L = {4, 5, 6}, decreasing order of comm.



- Can extend to any structures with conversion

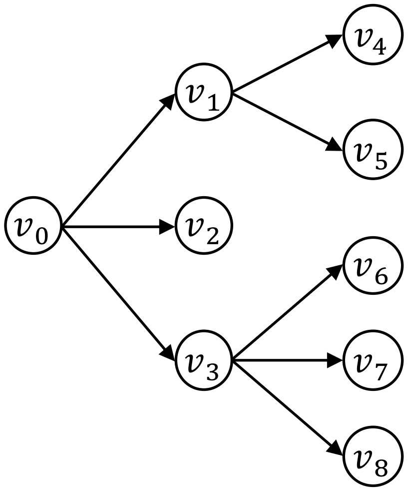
DAG Conversion

- Apply path-wise scheduling on arbitrary DNNs
  - Breadth-first search on graph
  - Duplicate each internal nodes
  - Avoid re-processing duplicated layers in inference



Layer-wise Scheduling

- For arbitrary tree-structure DNNs
  - Johnson's rule + conversion is suboptimal
  - Challenge: precedence constraints
- Recursively merge schedules of subtrees
  - Schedule of a subtree:
    - list covering all its node
  - At internal nodes:
    - Merge lists of children nodes
    - Group it with head of the merged list
  - Johnson's rule for comparisons



Layer-wise Scheduling

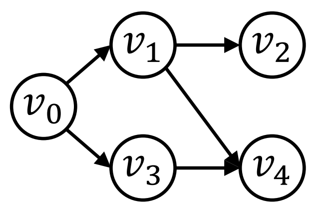
- Property

> Theorem 1: The schedule generated by the recursive merging approach is optimal for tree-structure DAGs.

- Proof
  - Sketch : mathematical induction
    - Merging and grouping will not lose optimal schedule
  - Insights:
    - Merging preserve the precedence constraints
    - For nodes without precedence constraints, Johnson's rule finds their optimal schedule

4. Scheduling for DAG-Style DNNs

- More complex precedence constraints
  - DAG scheduling is NP-hard
  - Inspired by topological sort
    - Iteratively sort nodes with no successors with Johnson's rule
    - Scheduled nodes are removed from the DAG



Schedule

$$\sigma = [v_0, v_1, v_3, v_4, v_2]$$

5. Experiment

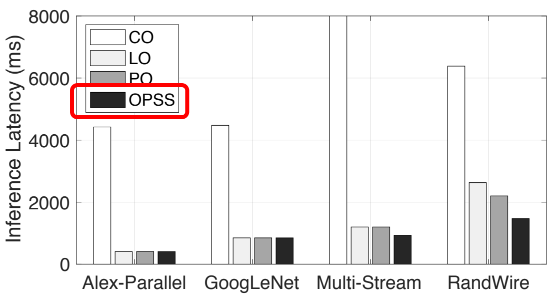
- Prototype implemented with PyTorch
  - gRPC is used for offloading
  - PyTorch Profiler is used to measure comp. time
- DNNs used in evaluation
  - Alex-Parallel[1]: multi-path tree
  - GoogleNet[2]: tree
  - Multi-Stream Network[3]: tree
  - RandWire[4]: DAG

[1].K. He et. al., "Deep residual learning for image recognition," in IEEE CVPR, 2016, pp. 770–778
[2].C. Szegedy et. al., "Going deeper with convolutions," in IEEE CVPR, 2015, pp. 1-9.
[3].Y.-W. Chao, et. al., "Learning to detect human-object interactions," in IEEE WACV, 2018, pp. 381-389.
[4].S. Xie, et. al., "Exploring randomly wired neural networks for image recognition," in IEEE ICCV, 2019, pp. 1284–1293.

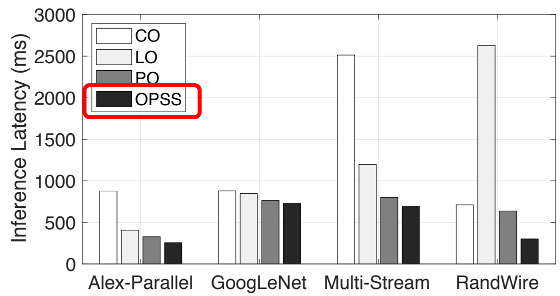
Experiment Results

- Latency on different network environment

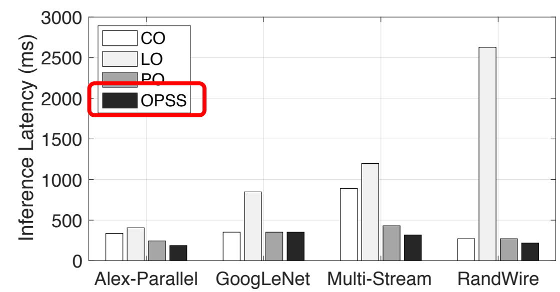
  (CO: Cloud-Only, LO: Local-Only, PO: Partition-Only)



3G

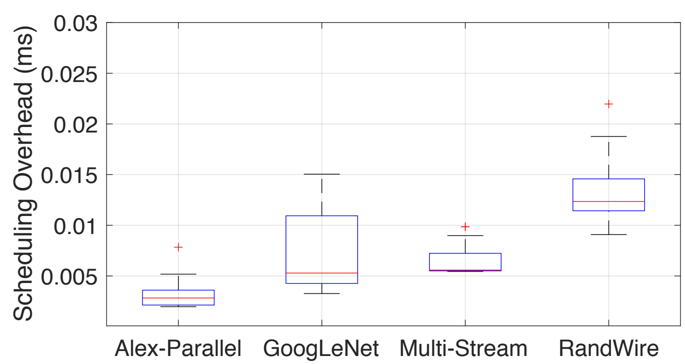


4G



Wi-Fi

- Scheduling overhead



6. Conclusion

- Proposed an offloading pipeline
  - Hide comm. time behind comp.
- Optimal path-wise scheduling
  - Intend for trees with multi-paths
  - Can apply to arbitrary DNNs with conversion
- Optimal layer-wise scheduling
  - Can apply to arbitrary tree-structure DNNs
  - Recursively merge schedule lists
- Evaluation on a prototype system

Questions

yubin.duan@temple.edu