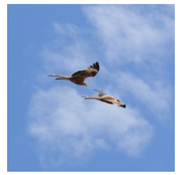
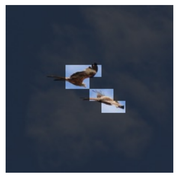
Latency-aware Spatial-wise Dynamic Networks

Yizeng Han[1]* , Zhihang Yuan[2]*, Yifan Pu[1]*, Chenhao Xue[2] , Shiji Song[1] , Guangyu Sun[2] , and Gao Huang[1]
[1]Tsinghua University, Beijing, China; [2]Peking University, , Beijing, China
*Equal contribution. Corresponding author.

paper/code

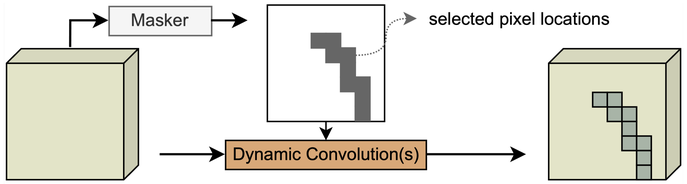
Background

Spatial Redundancy

- Some image regions are less important.
- Treat them equally cause computation redundancy.

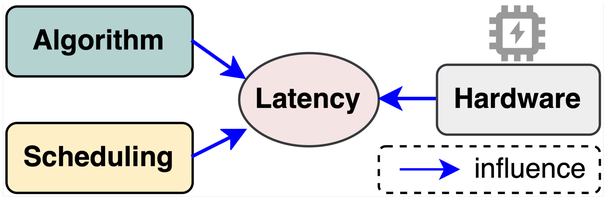
Spatial-wise Dynamic Computation



- Adaptive allocating more computation to more informative regions (e.g. foreground).

Observation & Motivation

Realistic speedup is hardly attained 😥

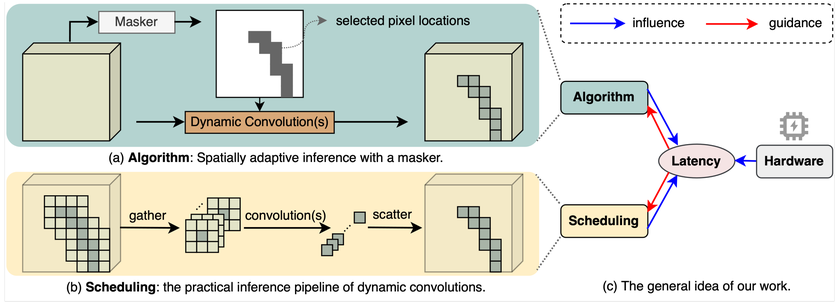


- Real latency is affected by three key factors
  - Algorithm design
  - Scheduling Strategy
  - Hardware Property
- Most existing works only consider part of them

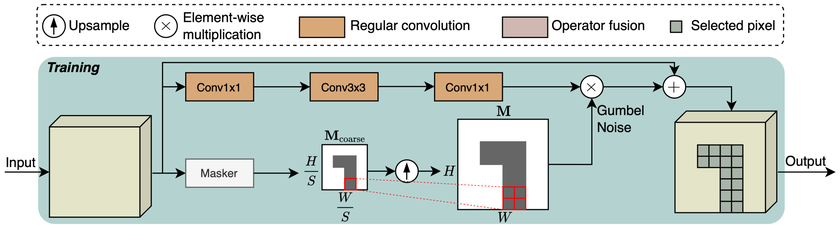
We propose a co-design framework! 😃

Method

Co-design Framework

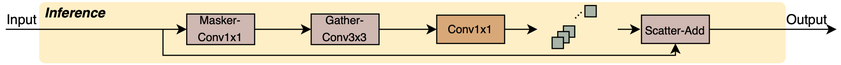


- **Algorithm design:** *Coarse-grained spatially adaptive*

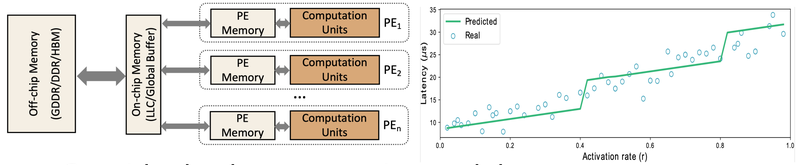


Larger S: more contiguous memory access & less flexibility

- **Scheduling strategy:** *Operator fusion*



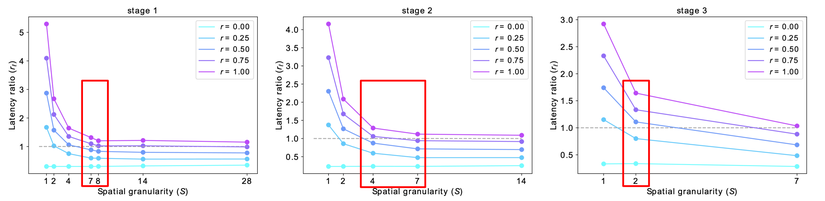
- **Hardware awareness:** *Latency prediction model*



Consider both computation and data movement.
Guide the choice of S of each model block.

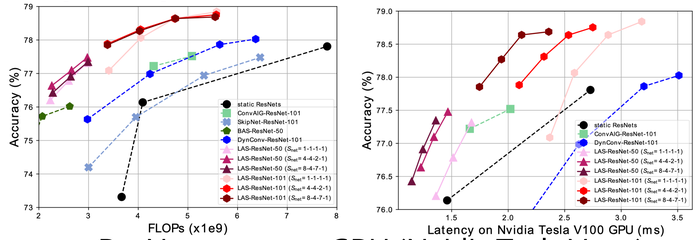
Experiments

Latency prediction model guide S design



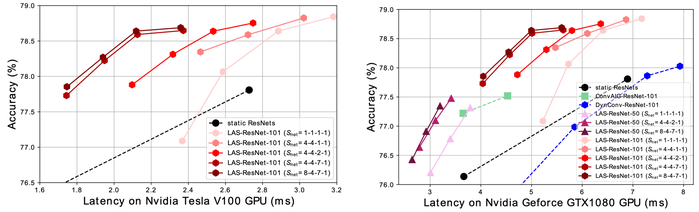
Latency prediction results on different stages

Real speedup is achieved on GPU



ResNet on server GPU (Nvidia Tesla V100)

Ablation study on granularity (S)



Visualization on computed regions