Dynamic Multi-path Neural Network

Yichao Wu

Sensetime Group Limited

Overview

- Overwhelming burden on computation of deeper neural networks
- Dynamic inference mechanism
  - Change the inference path for different samples at runtime
  - Existing methods only reduce the depth by skipping an entire specific layer
- Dynamic Multi-path Neural Network
  - Provide more topology choices in terms of both width and depth

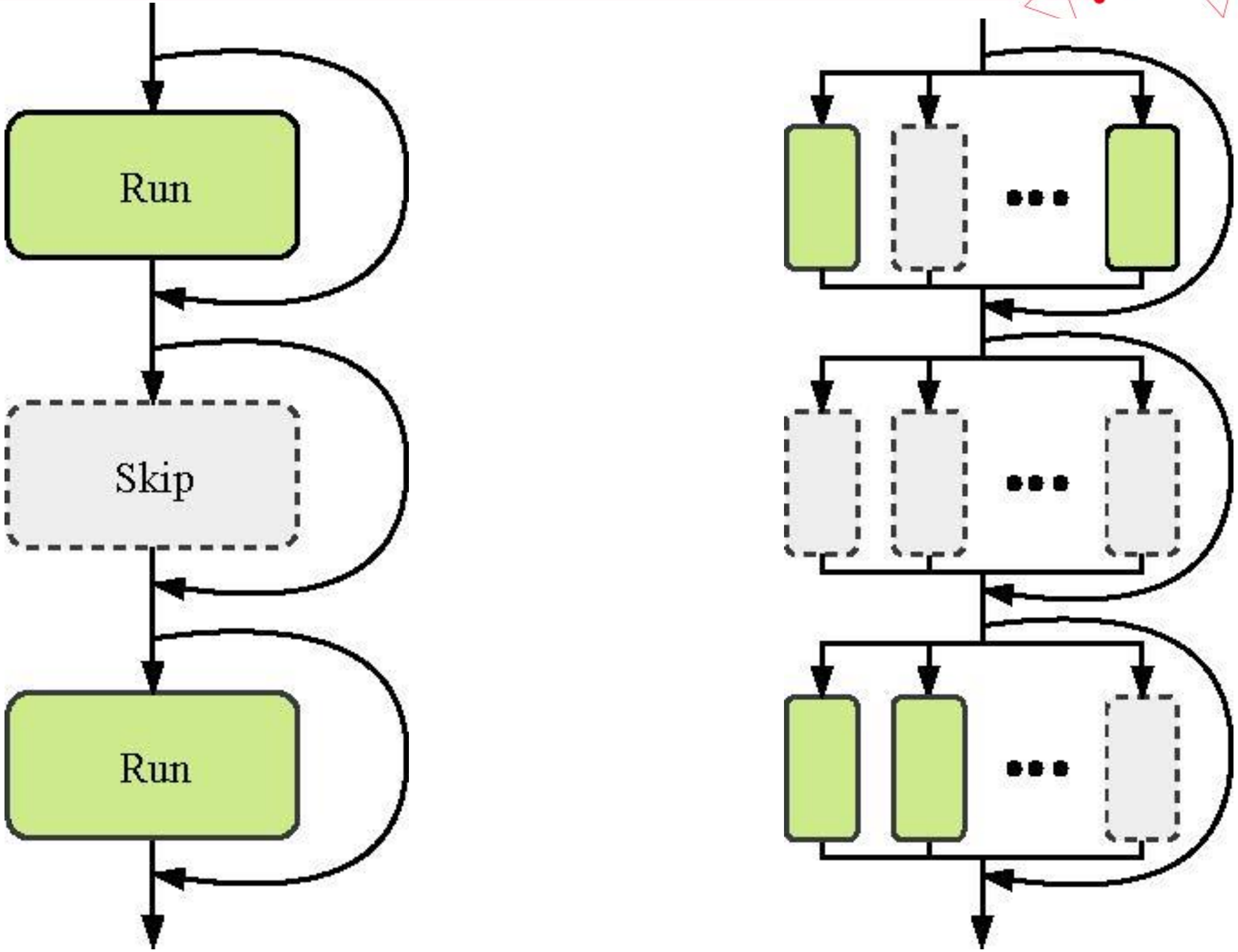
Outline

- Introduction
- Proposed Approach
- Experiments

Introduction

- Dynamic inference mechanism
  - Elegant solution to lightweight deployment
  - Prevalent dynamic inference techniques are mostly layer-wise
- We aim to improve the conventional dynamic inference scheme in terms of both network width and depth.
- Challengings: efficiency and effectiveness
  - Block split
  - Gate controller
- Experimental results demonstrate the superiority of our method.

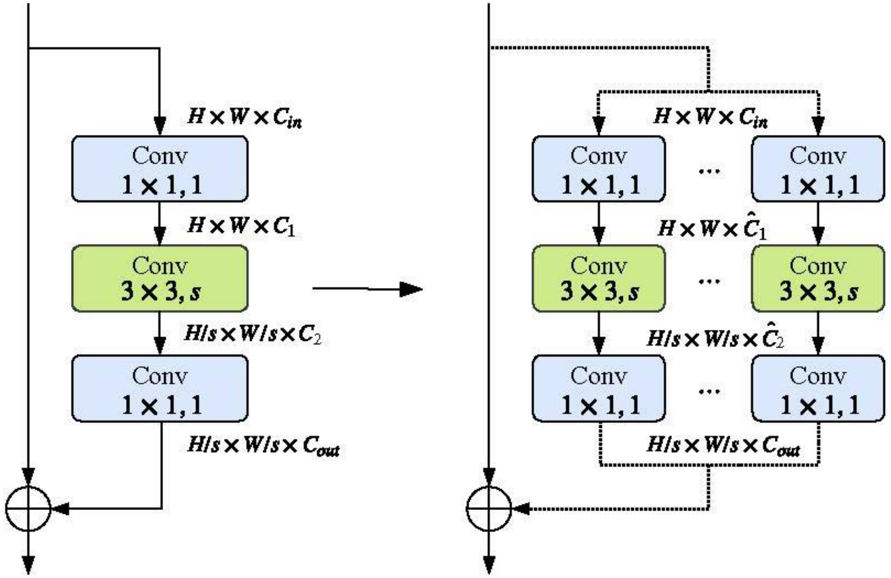
Overview of DMNN

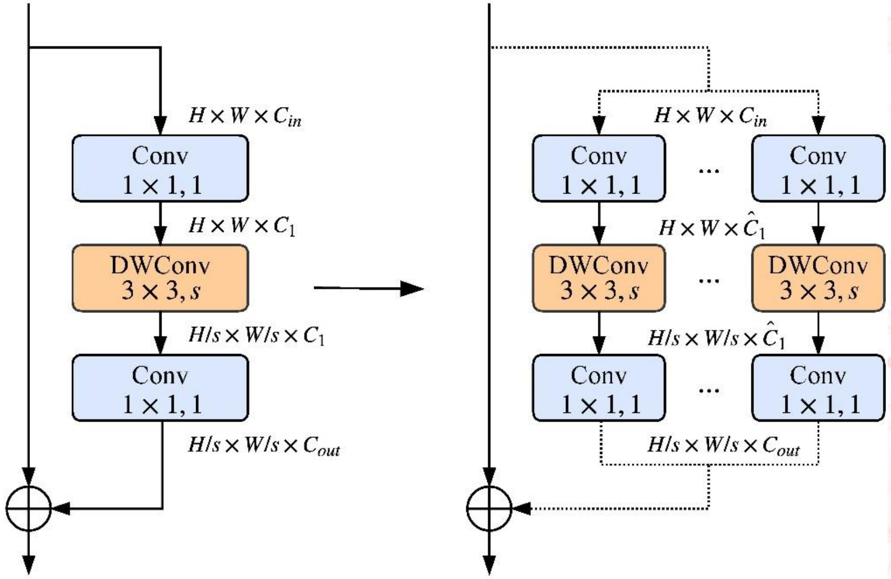


Different from altering on depth by skipping an entire specific layer, DMNN alters on both width and depth.

Block Subdivision

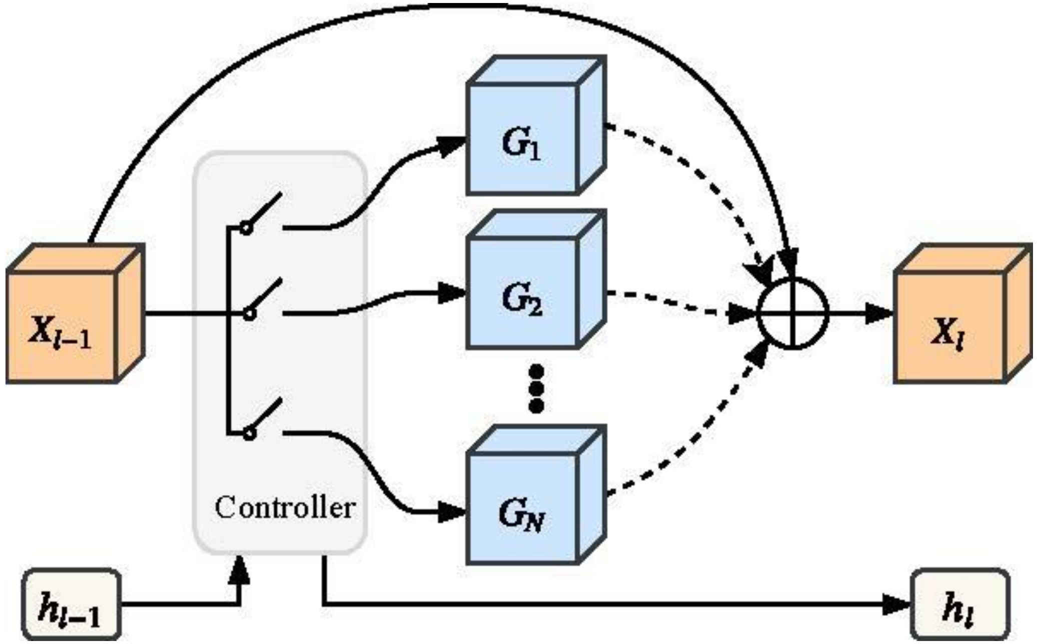
- Efficiency: impractical to control each channel
- Procedure
  - Divide the origin block of the network into several sub-blocks
  - Each sub-block has its switch to decide whether to execute or not



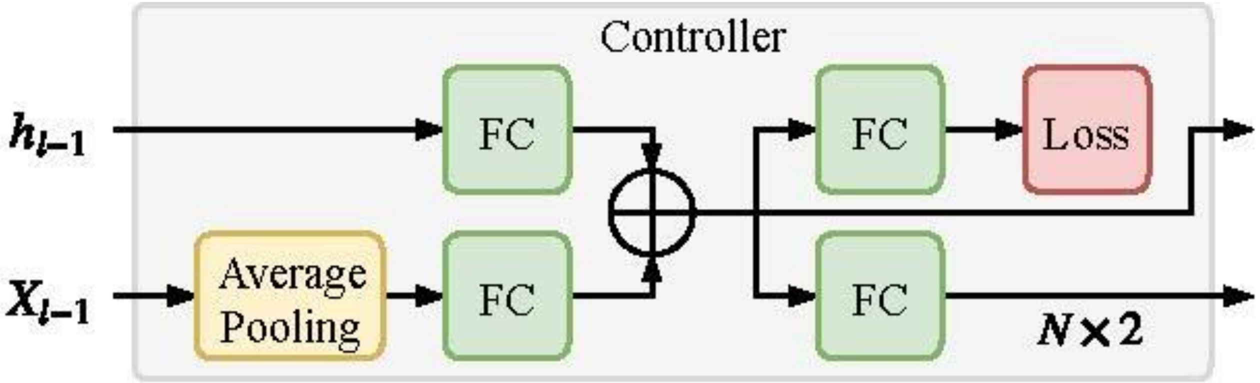


Design of Controller

- Overview of gate controller
  - Predict the status of each sub-block (on/off)

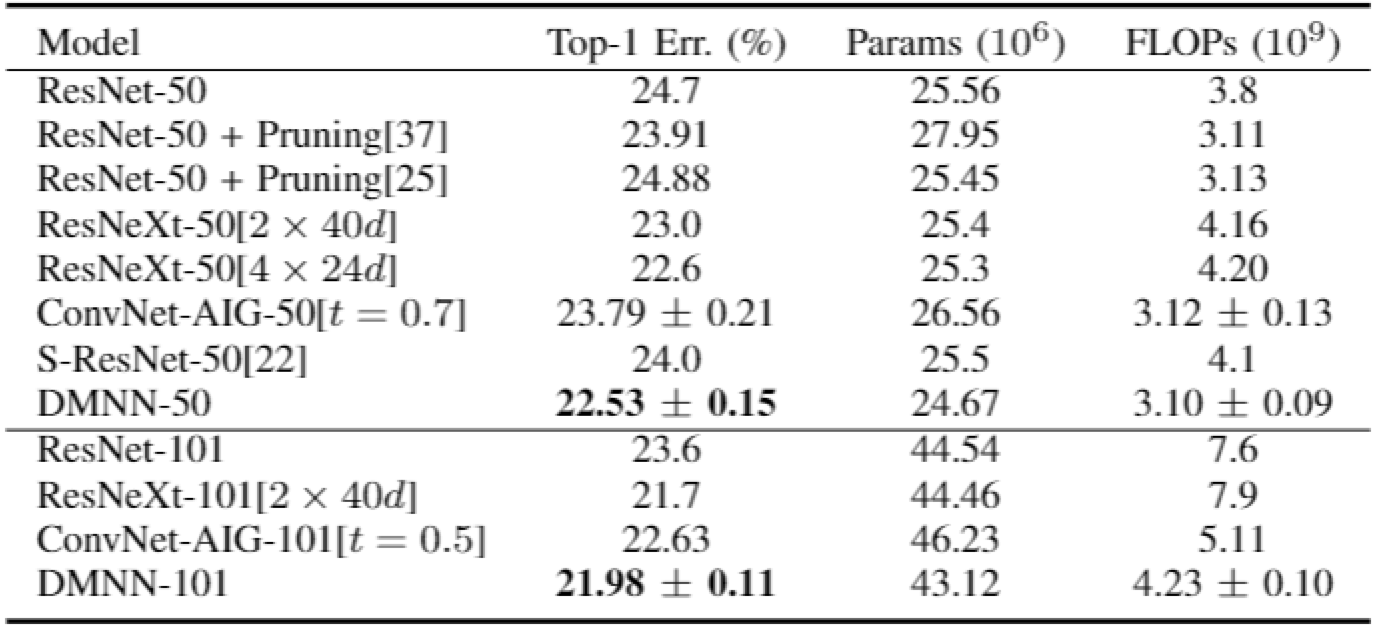


- Design
  - Previous state information embedding
  - Auxiliary classification task



Experiment

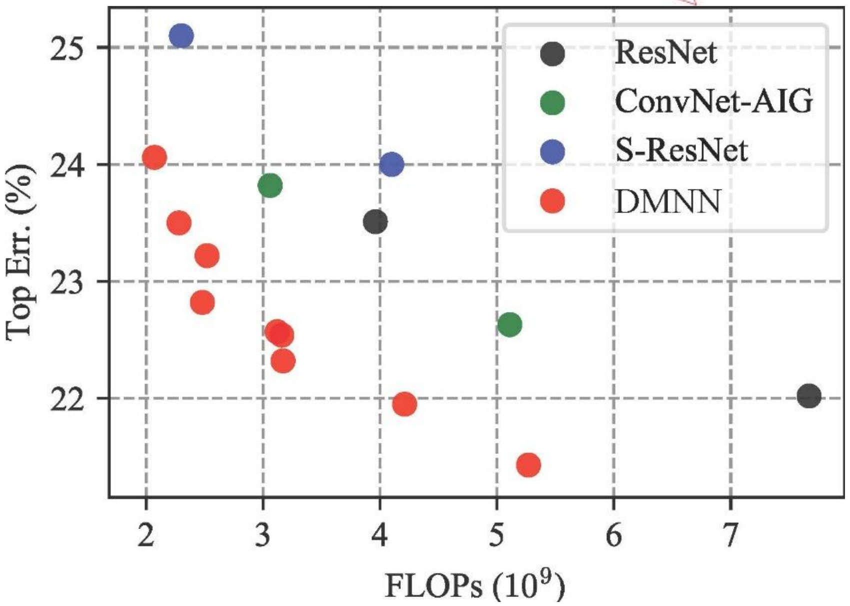
- Benchmark: ImageNet
- Training setup
  - Standard practice
- Performance analysis



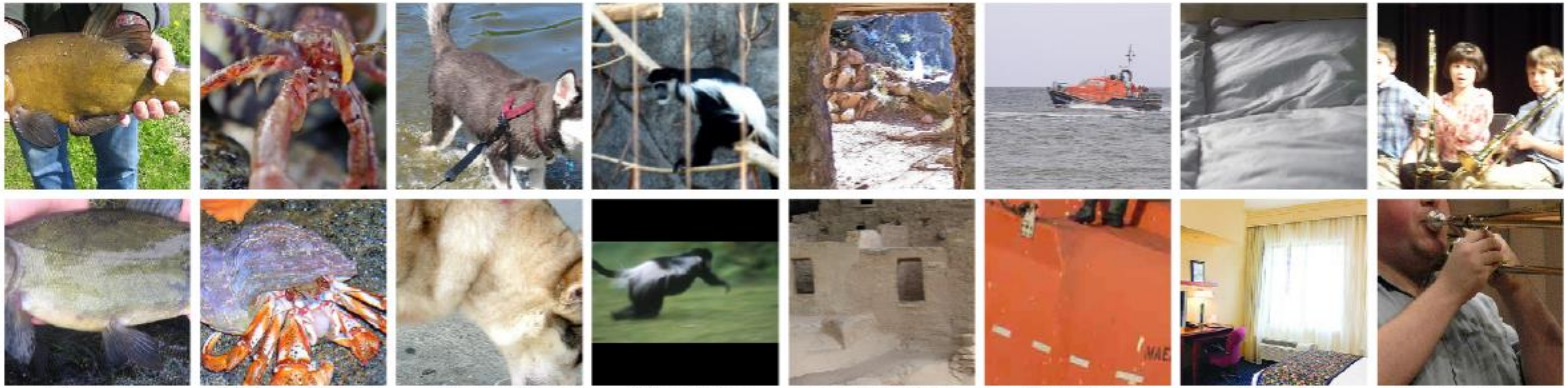
Model	Top-1 Err. (%)	Params (10^6)	FLOPs (10^9)
ResNet-50	24.7	25.56	3.8
ResNet-50 + Pruning[37]	23.91	27.95	3.11
ResNet-50 + Pruning[25]	24.88	25.45	3.13
ResNeXt-50[$2 \times 40d$]	23.0	25.4	4.16
ResNeXt-50[$4 \times 24d$]	22.6	25.3	4.20
ConvNet-AIG-50[$t = 0.7$]	23.79 ± 0.21	26.56	3.12 ± 0.13
S-ResNet-50[22]	24.0	25.5	4.1
DMNN-50	**22.53 ± 0.15**	24.67	3.10 ± 0.09
ResNet-101	23.6	44.54	7.6
ResNeXt-101[$2 \times 40d$]	21.7	44.46	7.9
ConvNet-AIG-101[$t = 0.5$]	22.63	46.23	5.11
DMNN-101	**21.98 ± 0.11**	43.12	4.23 ± 0.10

Further Analysis

- **Top-1 error vs FLOPs**



- **Visualization of “easy” and “hard” samples**



(a) tench (b) hermit crab (c) malamute (d) colobus (e) dwelling (f) lifeboat (g) quilt (h) trombone

Conclusion

- DMNN
  - Provide more path selection choices in terms of network width and depth during inference
- Experimental results
  - Superior performance in terms of efficiency and accuracy
- Future work
  - Apply the framework to practical systems

Thanks && Questions ?