Improved Techniques for Training Adaptive Deep Networks

Hao Li[1]*, Hong Zhang[2]*, Xiaojuan Qi[3], Ruigang Yang[2], Gao Huang[1]

[1] Tsinghua University [2] Baidu Inc. [3] University of Oxford

*** Equal Contribution**

ICCV 2019 Seoul, Korea

code link

arxiv link

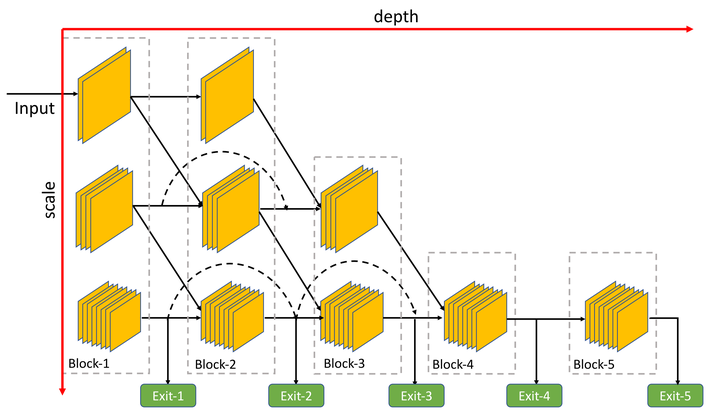
Motivation

Adaptive Inference

- Adjust the network structure dynamically based on inputs
- Improve computational efficiency at test time
- Use small models for “easy” inputs while big models for “hard” inputs

“Easy” Dogs

“Hard” Dogs



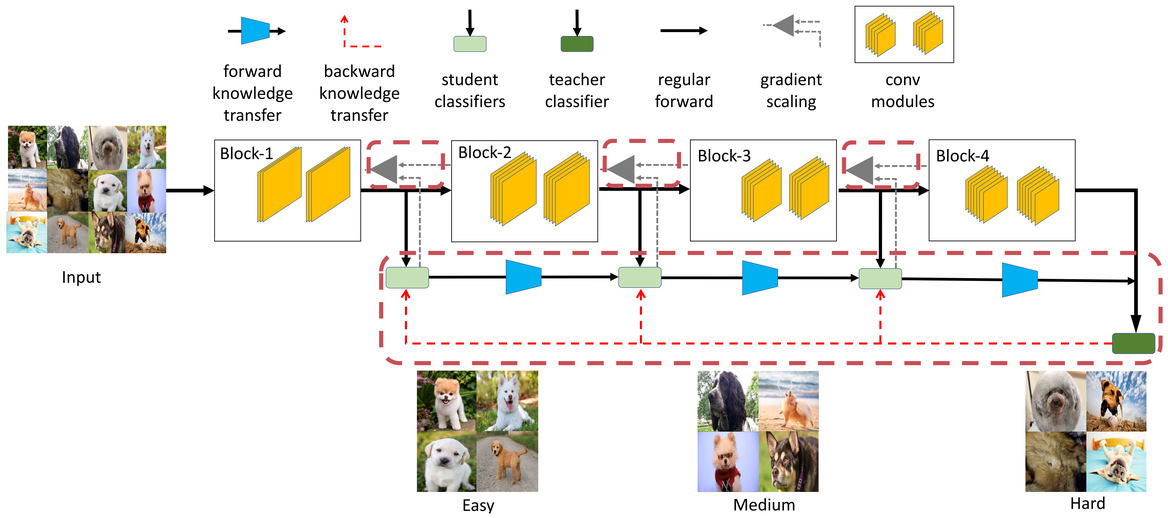
MSDNet[1]

Training adaptive inference networks effectively is difficult:

- How to resolve the conflicts among classifiers
- How to encourage the collaboration of classifiers

[1] Huang *et al.*, Multi-scale dense networks for resource efficient image classification. ICLR 2018

Method



Resolve gradient conflicts among classifiers

Gradient Equilibrium (GE)

- Rescale the magnitude of gradients along its backward propagation path.

$$R(x;s) = x;\ \nabla_x R(x;s) = s$$

Encourage collaboration of classifiers

Inline Subnetwork Collaboration (ISC)

- Prediction of previous stage serves as a prior to facilitate learning of classifiers.

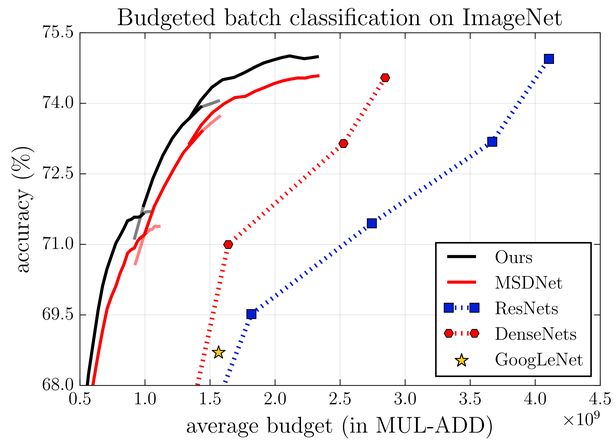
One-for-all Knowledge Distillation (OFA)

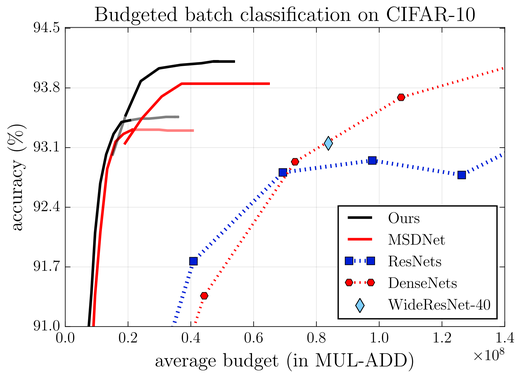
- The last classifier serves as a teacher model whose knowledge could be distilled into earlier exits.

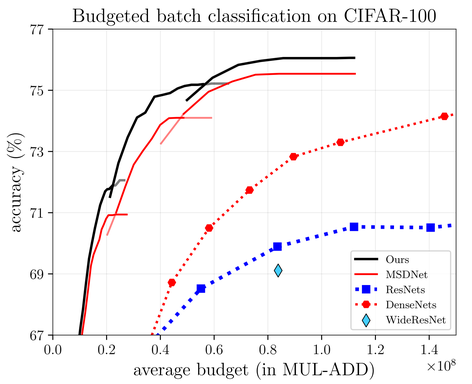
$$L_i = \alpha \mathrm{CE}_i + (1-\alpha)\mathrm{KLD}_i$$

$$\mathrm{KLD}_i = \sum_{c \in Y} p_k(c \mid x;\theta,T) \log \frac{p_i(c \mid x;\theta,T)}{p_k(c \mid x;\theta,T)}$$

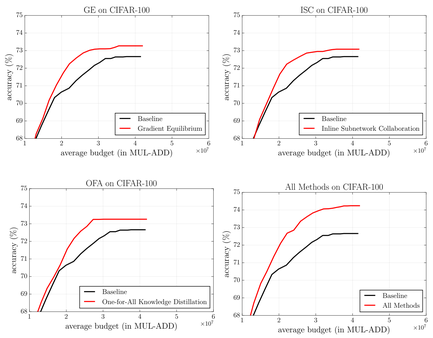
Results







Ablation Studies



Method			Accuracy @TOP1				
GE	ISC	OFA	E-1	E-2	E-3	E-4	E-5
-	-	-	60.09	63.73	67.89	70.48	71.81
✓			60.35	64.38	68.72	70.65	71.94
	✓		60.19	64.72	68.07	70.94	73.28
		✓	60.39	64.20	68.10	70.65	71.85
✓	✓	✓	**60.78**	**65.54**	**69.98**	**72.27**	**73.45**

Links

PyTorch Implementation: https://github.com/kalviny/IMTA