CVPR
JUNE 17-21, 2024
SEATTLE, WA

AdaShift: Learning Discriminative Self-Gated Neural Feature Activation With an Adaptive Shift Factor

Sudong Cai

Graduate School of Informatics, Kyoto University

京都大学

KYOTO UNIVERSITY

Introduction

- Preliminaries

$$\rightarrow \text{Layer-}1 \rightarrow \text{Layer-}2 \rightarrow \cdots \rightarrow \text{Layer-}\tau \rightarrow \cdots \rightarrow \text{Layer-}T$$

Comprehensive scoring
(relaxed similarity)

$$\tilde{x} = \langle \boldsymbol{w}, \boldsymbol{x} \rangle$$

$$\phi\left(\tilde{x}\right) = \rho\left(\tilde{x}\right) \cdot \tilde{x}$$

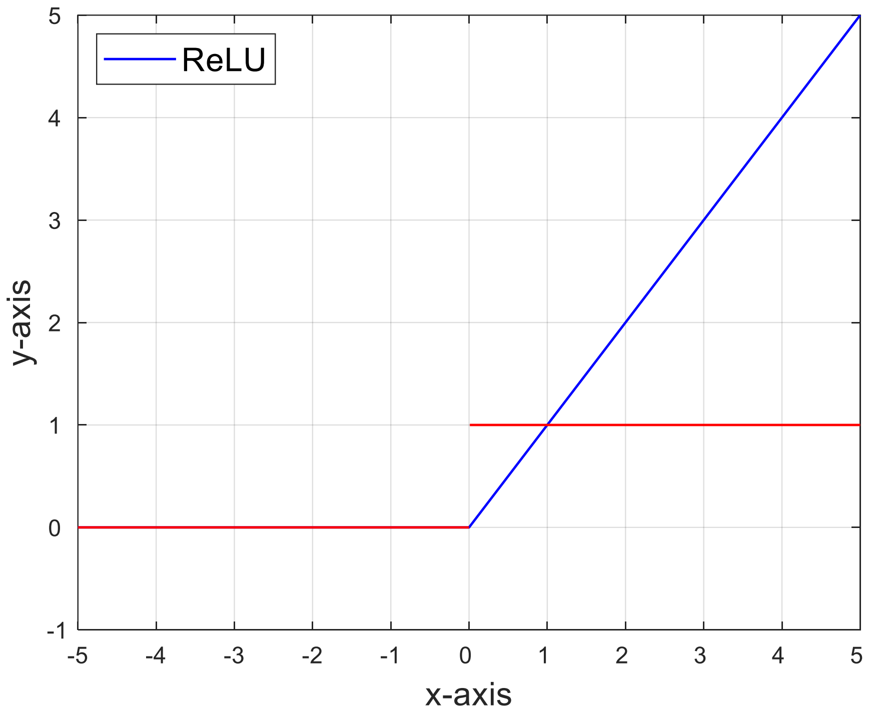
Neural Activation:
Feature selection by weights

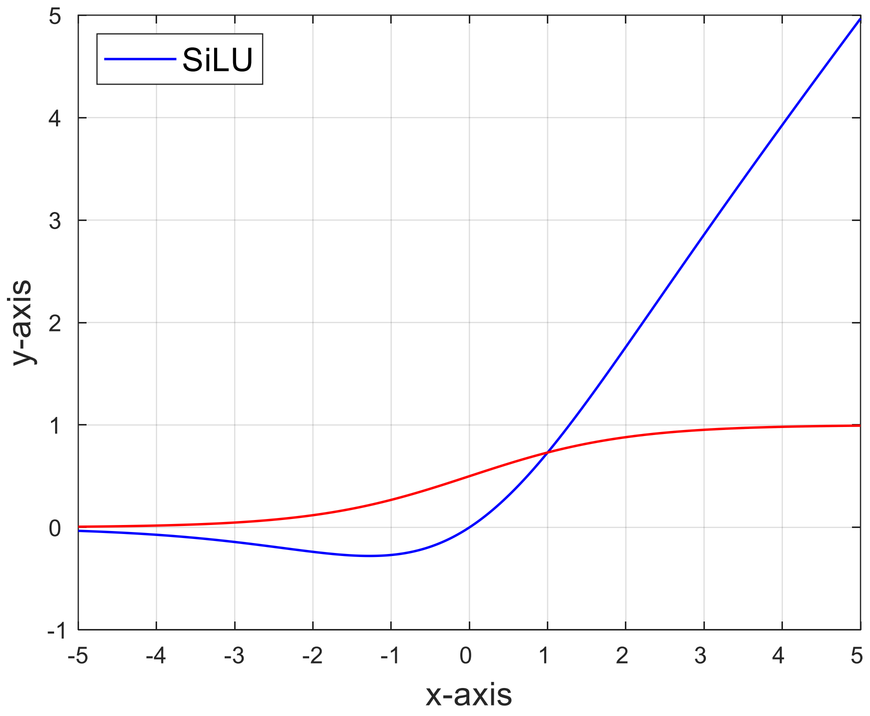
Importance score (weight)

Introduction

- Self-Gated Activation Functions

$$\rho\left(\tilde{x}\right) = \begin{cases} 1, & \tilde{x} > 0\,; \\ 0, & \tilde{x} \leq 0\,. \end{cases} \xrightarrow{\text{Smoothing}} \rho\left(\tilde{x}\right) = \mathrm{sigmoid}\left(\tilde{x}\right)\;.$$

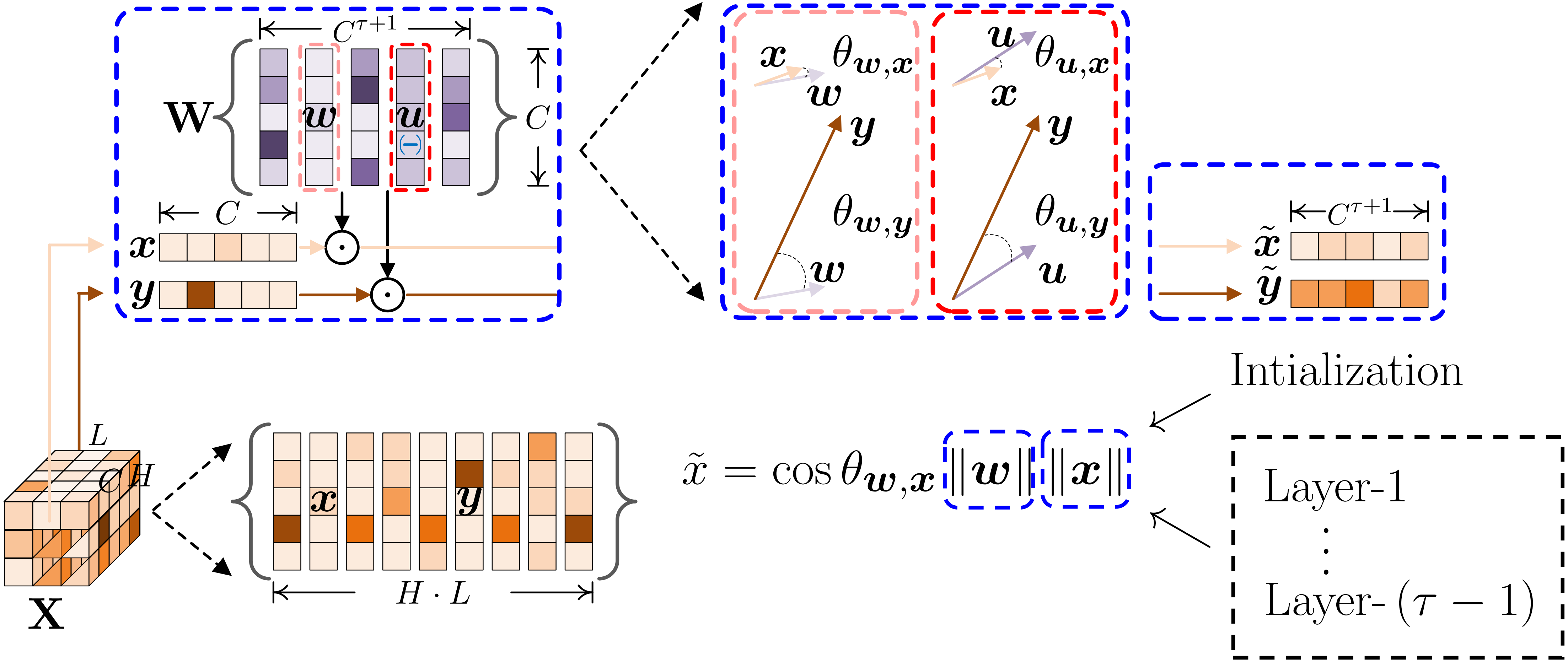




京都大学

Intuition

- **A critical problem:** *Mismatched Feature Scoring (MFS)*[1]



1. Sudong Cai. IIEU: Rethinking Neural Feature Activation from Decision-Making. In *Proc. ICCV*, pages 5796-5806, 2023.

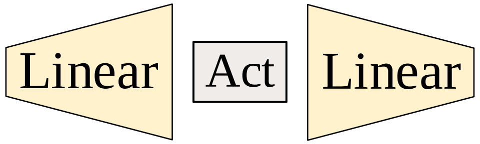
Intuition

- Why AdaShift – Remaining problems in the modeling of IIEU

1. Efficiency of training

$$\phi\left(\tilde{x}\right)=\varsigma\left(\boxed{\frac{\tilde{x}}{\left\|\boldsymbol{x}\right\|\left\|\boldsymbol{w}\right\|+\epsilon}}+\nu\right)\cdot\tilde{x}$$

$$\boxed{\nabla_{\boldsymbol{w}}s\left(\boldsymbol{w}\right)}=\frac{\left\|\boldsymbol{w}\right\|^{2}\boldsymbol{x}-\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}}{\left\|\boldsymbol{x}\right\|\left\|\boldsymbol{w}\right\|^{3}}$$



Expensive FFN with high expansion ratio

2. Extensibility

$$\phi\left(\tilde{x}\right)=\varsigma\left(\frac{\tilde{x}}{\boxed{\left\|\boldsymbol{x}\right\|}\boxed{\left\|\boldsymbol{w}\right\|}+\epsilon}+\nu\right)\cdot\tilde{x}$$

$$\tilde{x}:=\tilde{x}+\tilde{y}$$

How about if we also consider activating an un-projected feature (unit)?

Can feature and filter norms play a useful role in feature re-weighting?

Intuition of AdaShift

- Rethinking the meaning of feature and filter norms from a Softmax-based classification
  - **(Intuition 5.1)** Feature and filter norms present local and dataset-level non-local cues, respectively

Pre-conditions: $\boldsymbol{w}(i)$: The representative filter of the class-i — N classes in total

$\xrightarrow{\text{Let}}$ $\boldsymbol{w}(j)$: The representative filter of an arbitrarily class-j — $\boldsymbol{x}$ is classified as the class-i

$\boldsymbol{x}$: The representative feature of the input (image or pixel) — $\forall \boldsymbol{w}(i) \neq 0, \forall \boldsymbol{x} \neq 0$

$$\xrightarrow{\text{For } i \neq j} \quad \frac{e^{\tilde{x}_i + \boxed{b_i}}}{\sum_{c=1}^{C} e^{\tilde{x}_c + b_c}} > \frac{e^{\tilde{x}_j + \boxed{b_j}}}{\sum_{c=1}^{C} e^{\tilde{x}_c + b_c}} \iff e^{\tilde{x}_i + \boxed{b_i}} > e^{\tilde{x}_j + \boxed{b_j}}$$

$$\longrightarrow \quad \|\boldsymbol{w}(i)\| \, \|\boldsymbol{x}\| \cos \theta_{\boldsymbol{w}(i), \boldsymbol{x}} + b_i > \|\boldsymbol{w}(j)\| \, \|\boldsymbol{x}\| \cos \theta_{\boldsymbol{w}(j), \boldsymbol{x}} + b_j$$

Intuition of AdaShift

$$\|\boldsymbol{w}(i)\| \, \|\boldsymbol{x}\| \cos\theta_{\boldsymbol{w}(i),\boldsymbol{x}} + b_i \; > \; \|\boldsymbol{w}(j)\| \, \|\boldsymbol{x}\| \cos\theta_{\boldsymbol{w}(j),\boldsymbol{x}} + b_j$$

$\xrightarrow{\text{Let}}$ $\alpha = b_j - b_i$

$\longrightarrow$ $\|\boldsymbol{w}(i)\| \cos\theta_{\boldsymbol{w}(i),\boldsymbol{x}} - \|\boldsymbol{w}(j)\| \cos\theta_{\boldsymbol{w}(j),\boldsymbol{x}} > \frac{\alpha}{\|\boldsymbol{x}\|}$

Case 1: $\|\boldsymbol{x}\| \gg |\alpha|$

$\longrightarrow$ Filter norms are influential & Feature norms are not decisive

Case 2: $\|\boldsymbol{x}\| \ll |\alpha|$

$\longrightarrow$ Filter norms are influential & Feature norms are influential

Case 3: Others

$\longrightarrow$ Filter norms are influential & Feature norms matters

Complementary information : tensor-level non-local cues

AdaShift Prototype

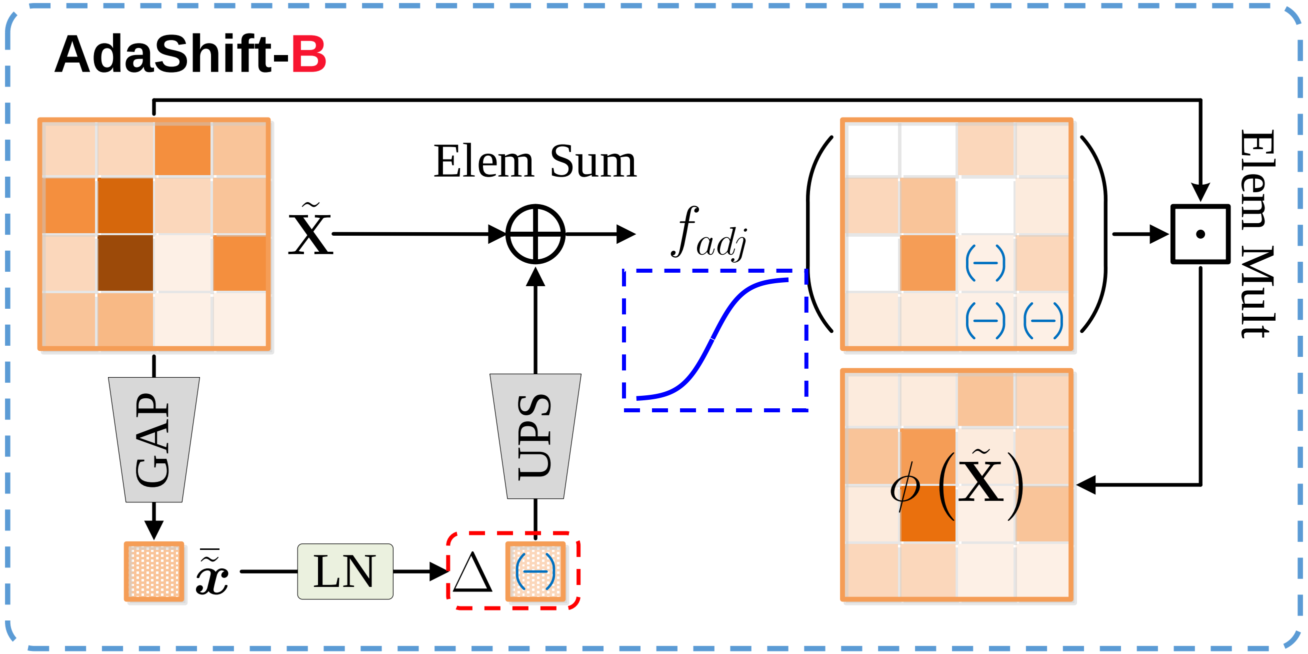
$\longrightarrow$ $\phi(\tilde{x}) = \varsigma\left(\tilde{x} + \boxed{\Delta}\right)\tilde{x}$

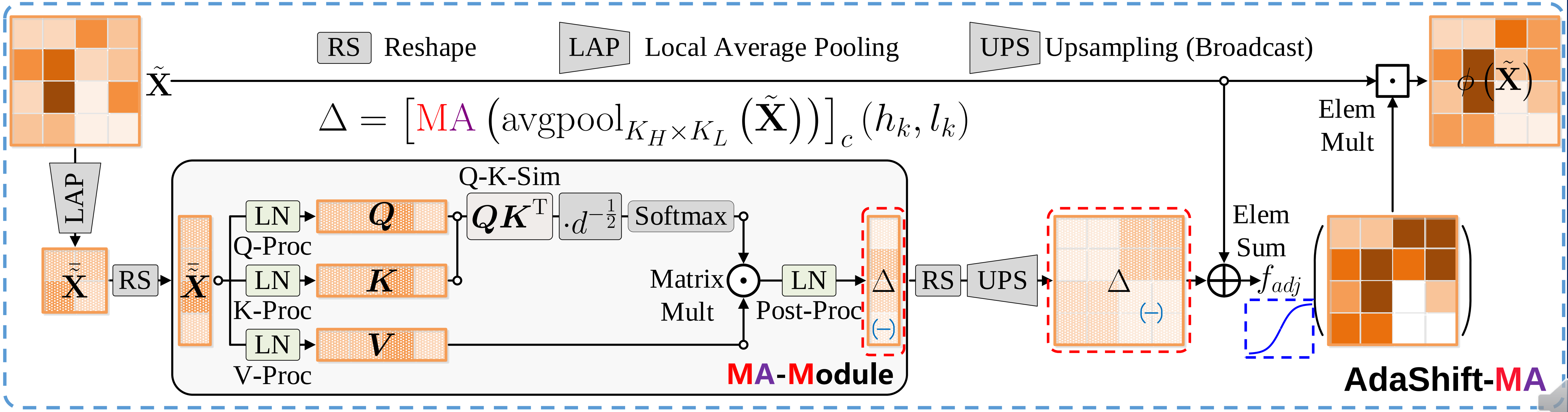
Method

AdaShift Prototype

$$\phi\left(\tilde{x}\right) = \boxed{f_{\text{adj}}}\left(\tilde{x} + \boxed{\Delta}\right) \cdot \tilde{x}$$

Complementary information:
tensor-level non-local cues





Experiment

- Dataset
  - ImageNet [90]
    - The most popular large-scale visual (image) recognition benchmark dataset
  - CIFAR(-100) [27]
    - A popular image recognition benchmark dataset of small-size images
  - COCO [21]
    - A popular large-scale object detection benchmark dataset
  - KITTI-Materials [23]
    - The benchmark dataset of RGB RMS

[90] J. Deng et al., Imagenet: A largescale hierarchical image database, *CVPR*, 2009

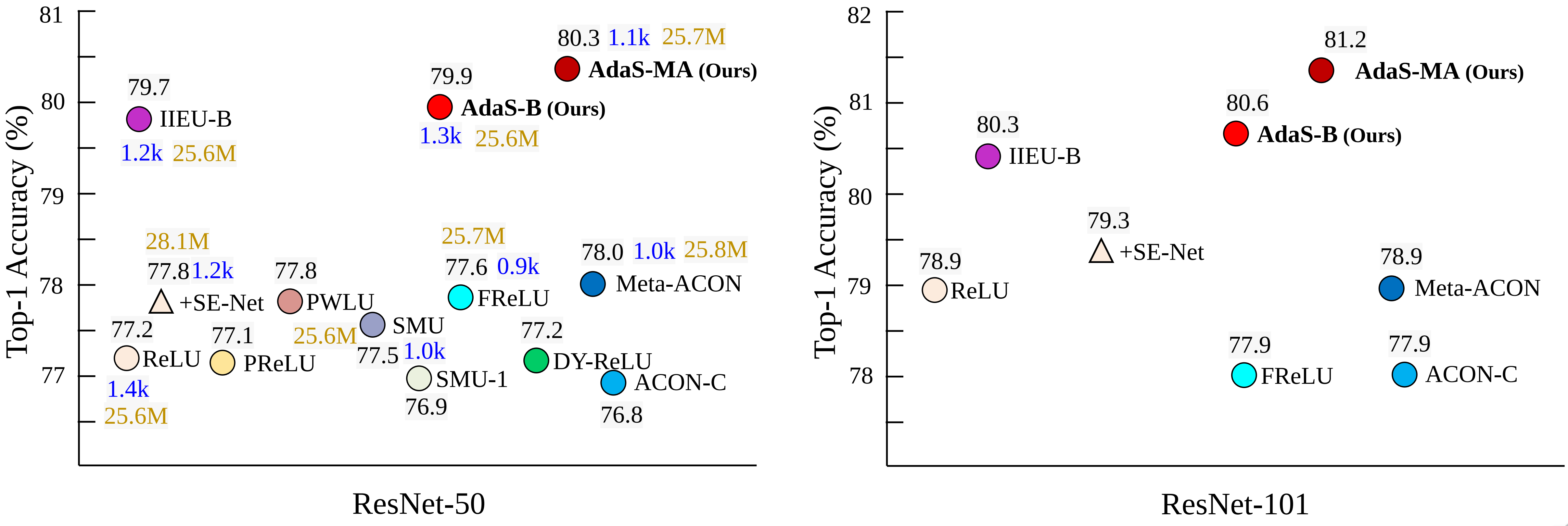
[27] A. Krizhevsky et al., Learning multiple layers of features from tiny images, *Master's thesis,* University of Toronto, 2009

[21] T.-Y. Lin et al., Microsoft coco: Common objects in context, *ECCV*, 2014

[23] **Sudong Cai** et al., RGB Road Scene Material Segmentation, *ACCV*, 2022

Results

Comparison of AdaShift-B and -MA with prevailing and (other) SOTA activation models on ImageNet using ResNet-50 and ResNet-101 backbones. In the comparison using the ResNet-50 backbone, the *number of parameters* and *throughput* of each model are indicated in Purple and Blue colors, respectively



京都大学

Results

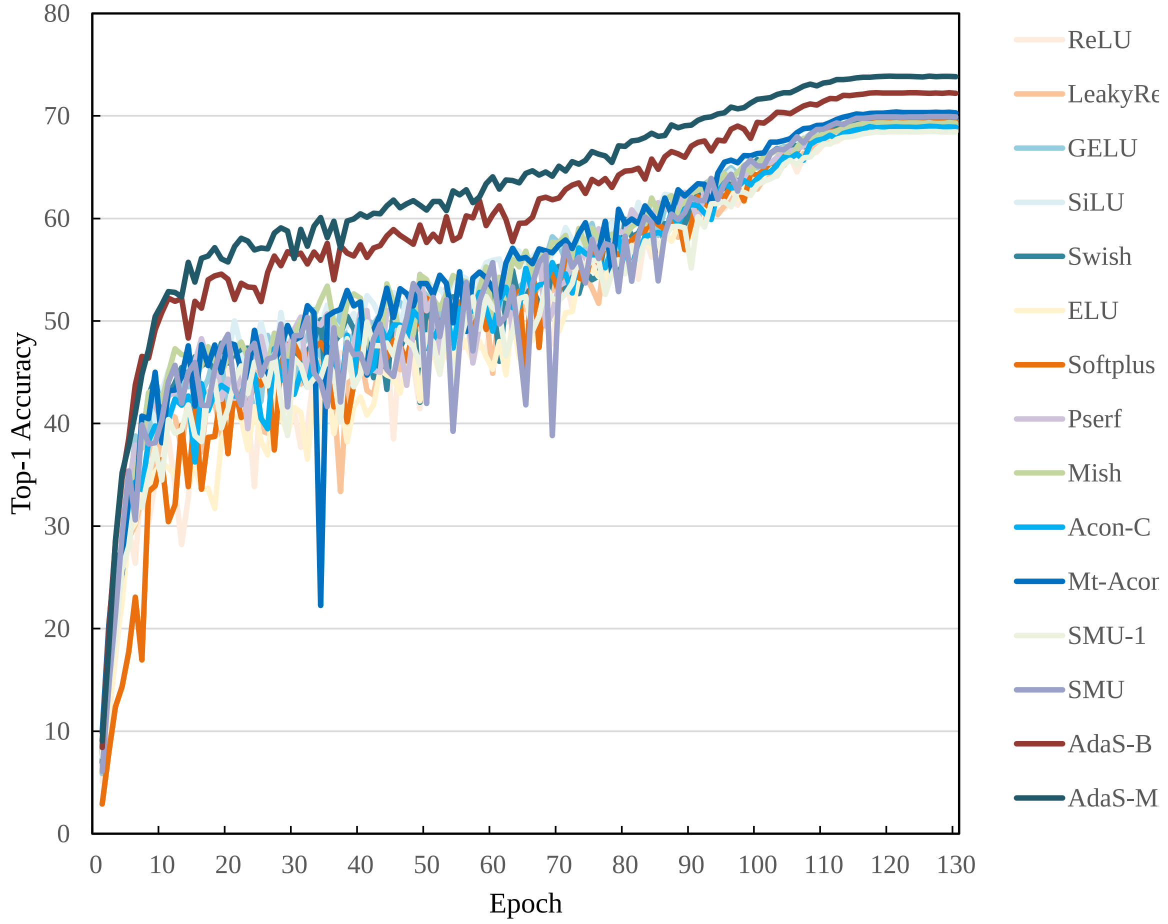
Comparison of AdaShift-enhanced ResNet-50(s) to representative vision Transformer counterparts. "$\ast$" denotes the improved ViT trained with an extra regularization. Details can be found in Appendix G.2. (Training with Advanced Recipe)

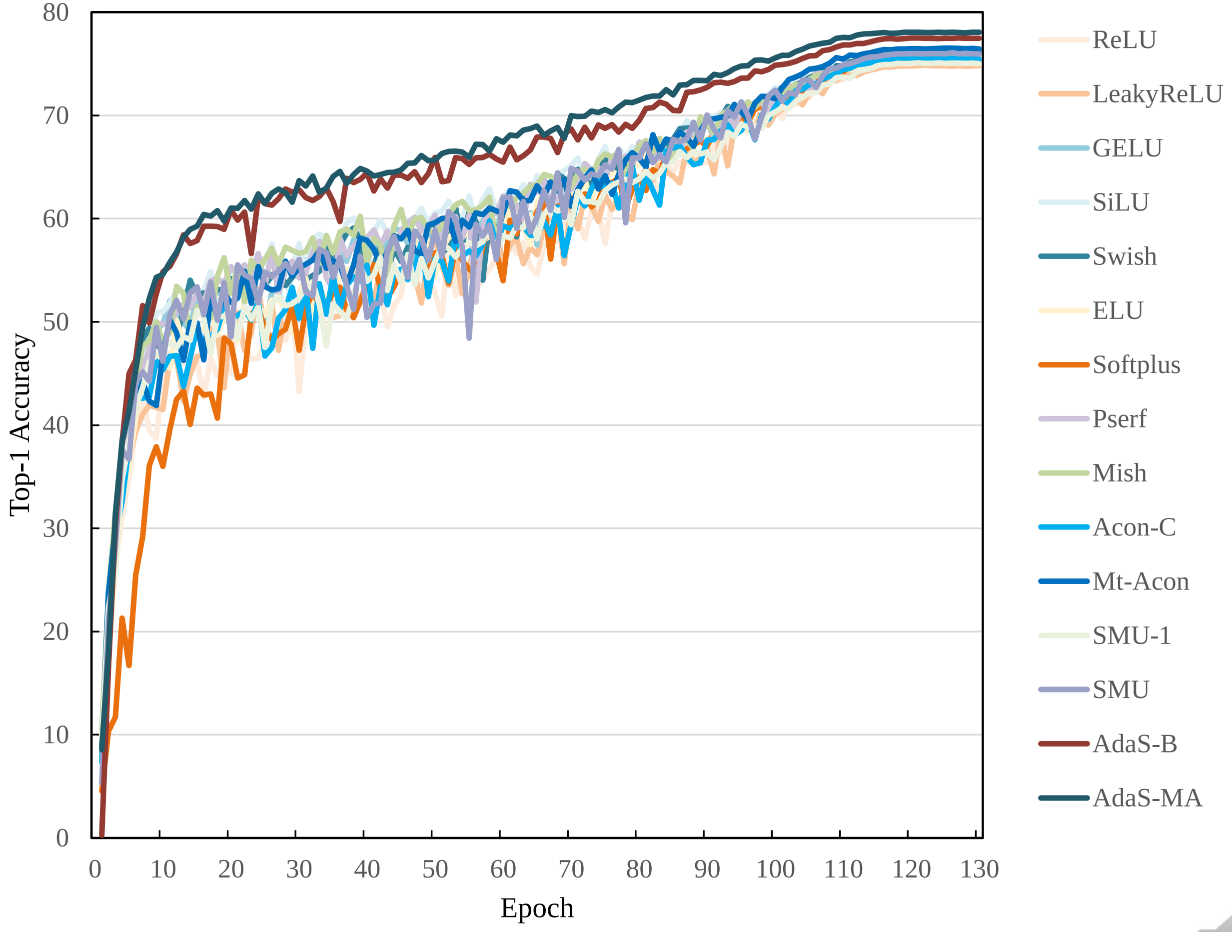
Network	Activation	Resolution	Train. Epoch	#Params.	FLOPs	Throughput	Top-1(%)↑
ViT-B/16*	GELU	224×224	300	86.6M	16.9G	775.6	79.7
PoolFormer-S24	GELU	224×224	300	21.4M	3.5G	1144.6	80.3
Swin-T	GELU	224×224	300	28.3M	4.5G	1052.2	**81.3**
ResNet-50	**AdaS-B (ours)**	224×224	300	25.6M	4.1G	**1352.8**	**80.8**
ResNet-50	**AdaS-Hyb (ours)**	224×224	300	28.2M	4.2G	**1201.6**	**81.7**

京都大学

Results

The Acc. curves (left) and loss curves (right) of ResNet-14 (Top) and ResNet-26 (Bottom) backbones with different activation models





Gratitude for your time and patience!

AdaShift: Learning Discriminative Self-Gated Neural Feature Activation With an Adaptive Shift Factor

Sudong Cai

Graduate School of Informatics, Kyoto University

京都大学