Jiajia JIAO, Ran WEN, Hong YANG, 2025. An end-to-end automatic methodology to accelerate the accuracy evaluation of deep neural networks under hardware transient faults. *Frontiers of Information Technology & Electronic Engineering*, 26(7):1099-1114. https://doi.org/10.1631/FITEE.2400547

An end-to-end automatic methodology to accelerate the accuracy evaluation of deep neural networks under hardware transient faults

Key words: Analytical model; Deep neural networks; Hardware transient faults; Fast evaluation; Automatic evaluation tool

Corresponding author: Jiajia JIAO
E-mail: jiaojiajia@shmtu.edu.cn
ORCID: https://orcid.org/0000-0003-3680-787X

Motivation

Hardware transient faults are proven to have a significant impact on deep neural networks (DNNs), whose safety-critical misclassification (SCM) in autonomous vehicles, healthcare, and space applications is increased up to four times. However, the inaccuracy evaluation using accurate fault injection is time-consuming. To accelerate the evaluation of hardware transient faults on DNNs, a unified and end-to-end automatic methodology, A-Mean, is proposed to estimate the general classification metric accuracy and application-specific metric SCM.

Main idea

A unified two-level rapid and systematic assessment method, A-Mean, is proposed to evaluate the impact of transient faults on DNNs. Our approach can take advantages of one-time fault-free dynamic information, a static two-level mean calculation model, and a worst-case policy to effectively capture the inner-layer (data type and operator) and inter-layer (input feature map, depth, and topology) fault impacts.

Main idea

We conduct some experiments and compare our A-Mean with single-convolution silent data corruption (SDC) and no max-policy. As we implement three groups of SDC_{Conv} to replace other SDC_{op}, the different operators remain highly effective and consistently outperform the alternatives. The max-policy and no max-policy reveal distinct characteristics in topology, such as branches without hidden layers behaving like a sequential structure, while branches with hidden layers require consideration of both individual branches and their interactions.

Main idea

We develop an automated tool to accelerate the evaluation of accuracy and SCM in DNNs under transient faults. This tool directly uses the model's printed structure and parameters, including operators and input feature maps, to automatically compute the SDC rate for each model. It is then combined with the accuracy and SCM of a no-fault case to estimate the corresponding metrics under transient faults.

Method

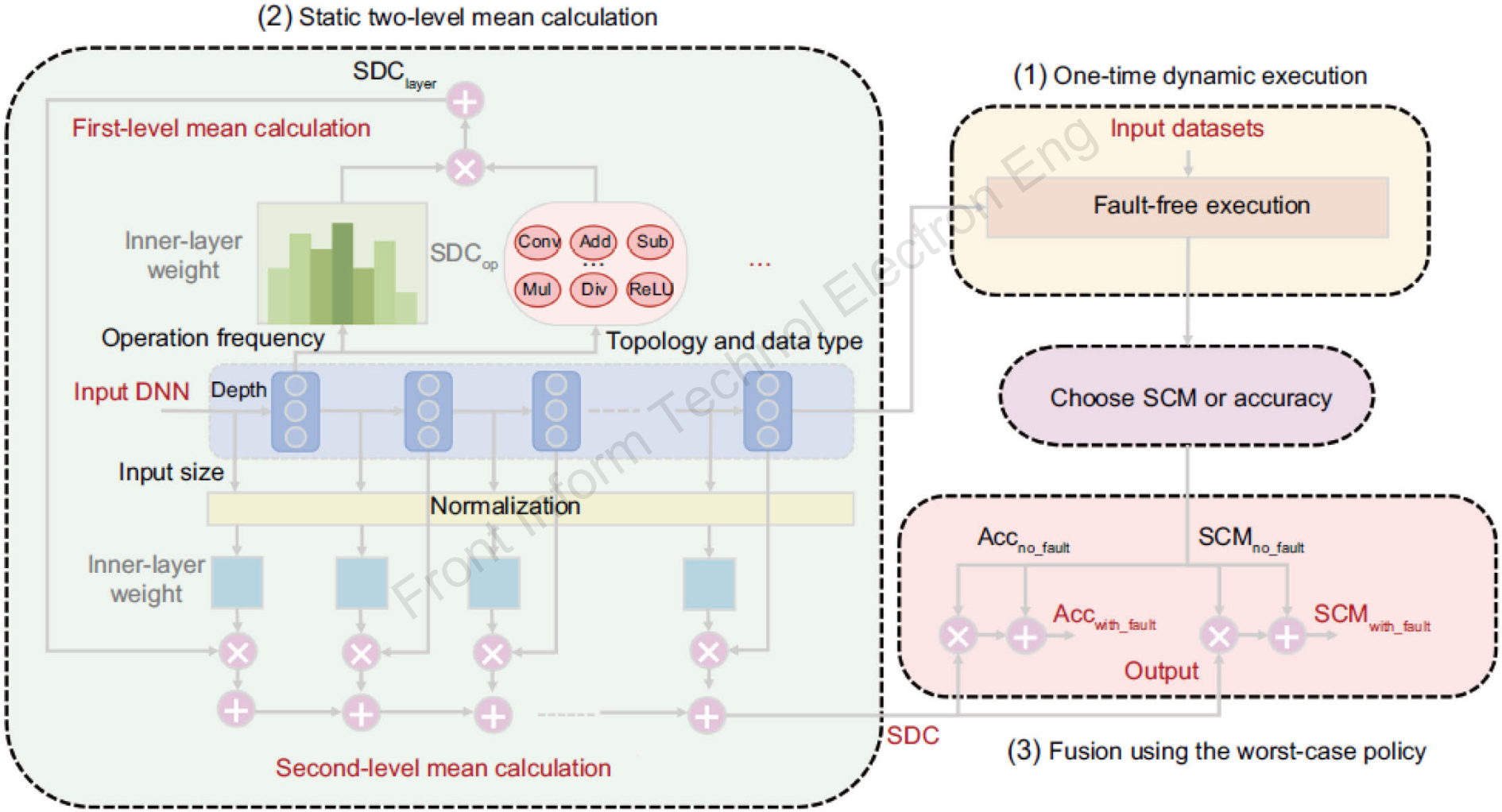


Fig. 3 Comparison of different methods in terms of outlier detection accuracy

Method

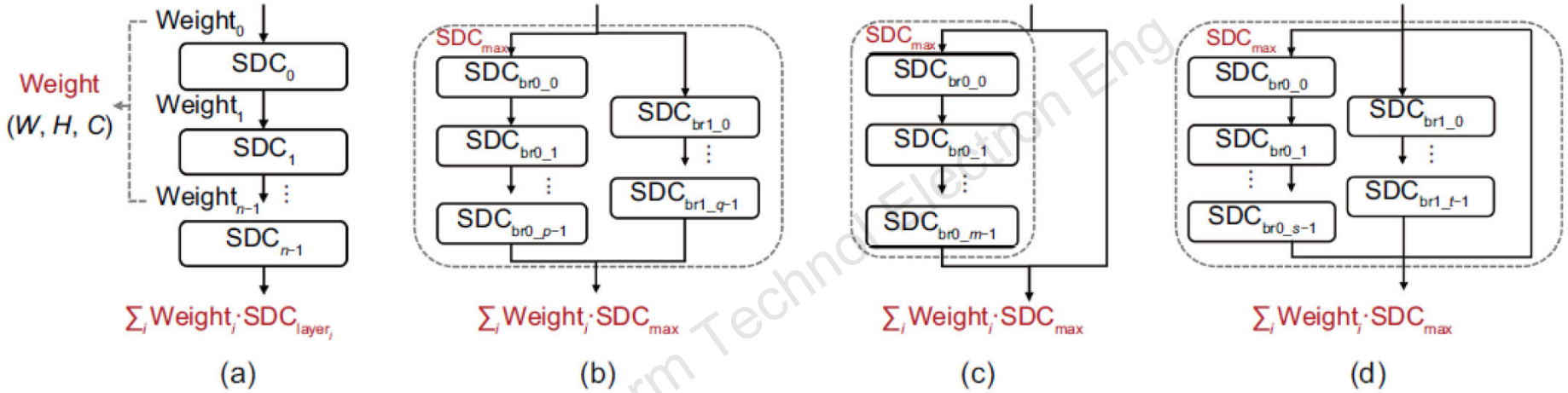


Fig. 4 Various DNN topologies: (a) sequential structure; (b) branch topology 1 (both branches have hidden layers); (c) branch topology 2 (one branch contains hidden layers); (d) branch topology 3 (some branches have hidden layers and some do not)

Method

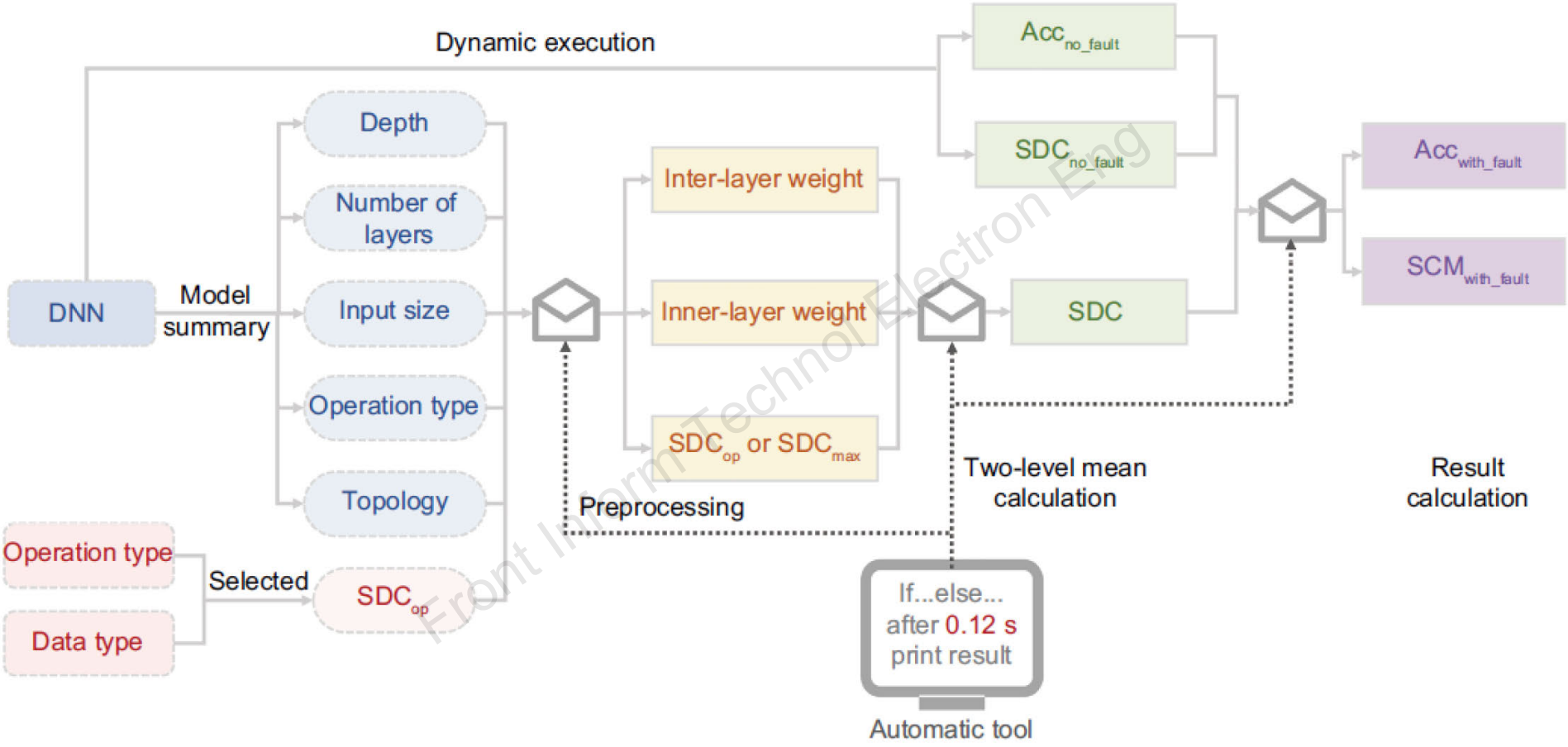


Fig. 5 End-to-end automatic tool

Results

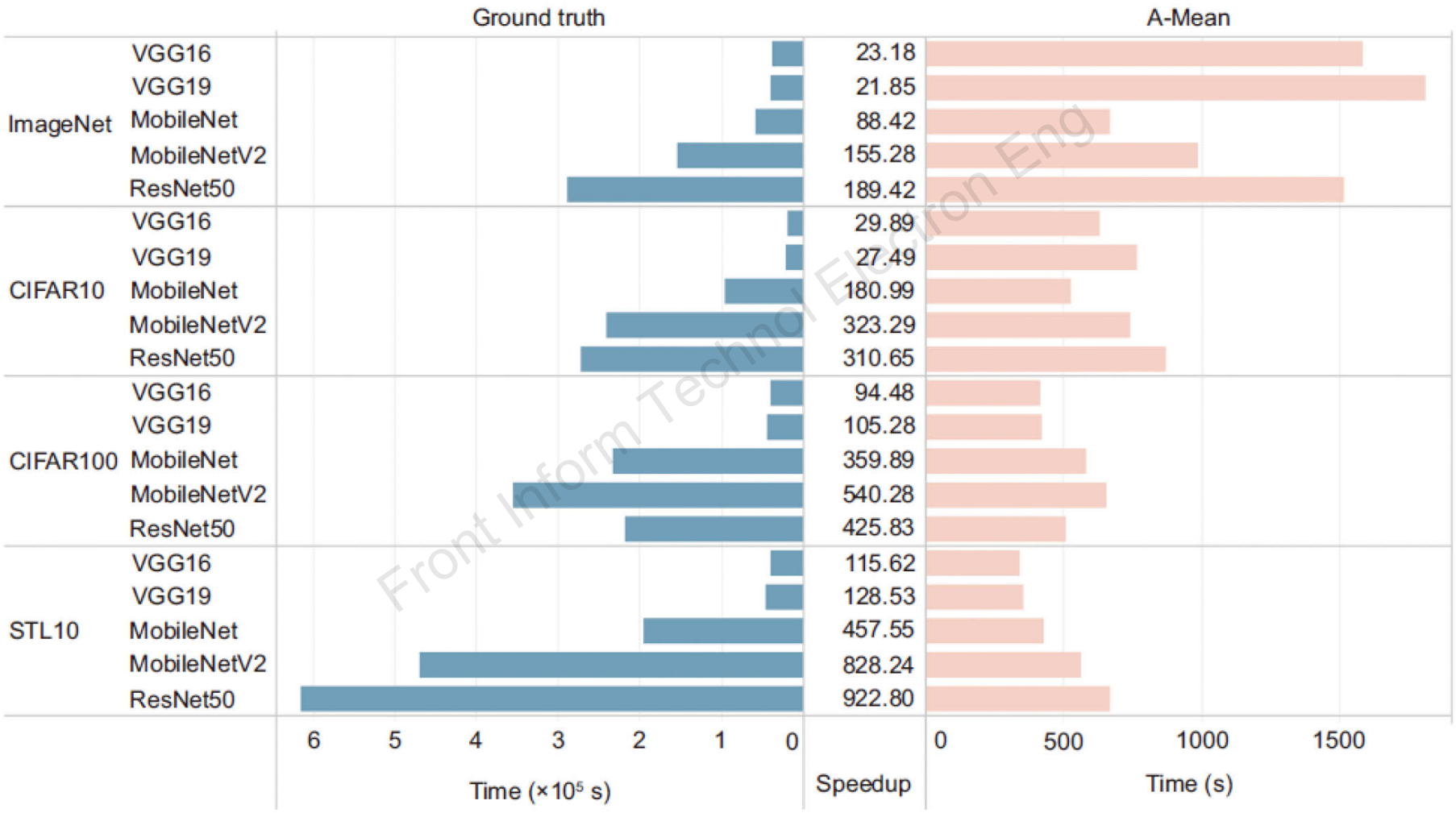


Fig. 6 Speedup between A-Mean and TensorFI+ (ground truth)

Results

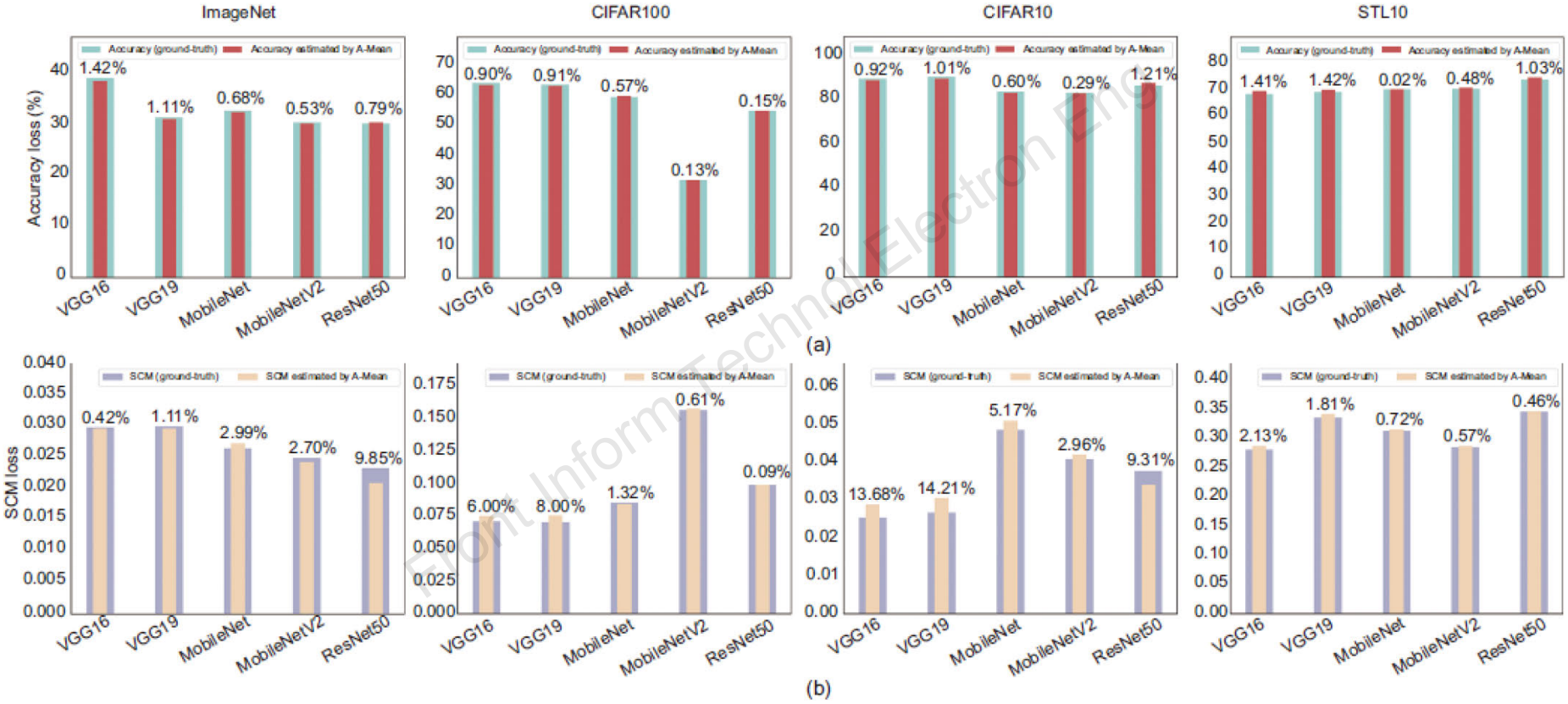


Fig. 7 Accuracy loss (a) and SCM loss (b) on four datasets

Results

Table 2 Different metrics with and without faults

Dataset	DNN	Acc_{no_fault} (%)	SCM_{no_fault} (%)	Fault sensitivity with faults	
				Ground truth	A-Mean
ImageNet	VGG16	70.43	2.11	0.3625	0.5949
	VGG19	71.18	2.04	0.3765	0.5881
	MobileNet	70.28	2.43	0.4231	0.5942
	MobileNetV2	70.71	2.21	0.4085	0.5915
	ResNet50	74.58	1.86	0.3950	0.5580
CIFAR100	VGG16	64.16	6.81	0.5270	0.6262
	VGG19	63.82	6.81	0.3810	0.6301
	MobileNet	59.60	8.28	0.5263	0.6695
	MobileNetV2	31.79	15.79	0.5238	1.2245
	ResNet50	54.72	9.84	0.5500	0.7238
CIFAR10	VGG16	90.28	2.07	0.4563	0.4424
	VGG19	91.20	2.18	0.4505	0.4363
	MobileNet	83.45	4.83	0.3000	0.4702
	MobileNetV2	83.05	4.00	0.4500	0.4780
	ResNet50	87.65	3.19	0.3742	0.4523
STL10	VGG16	39.05	27.90	0.2500	0.8232
	VGG19	31.48	33.35	0.1786	0.9588
	MobileNet	32.43	31.14	0	0.9619
	MobileNetV2	30.21	28.33	0.2000	0.7834
	ResNet50	30.35	34.38	0.3030	0.9811

Results

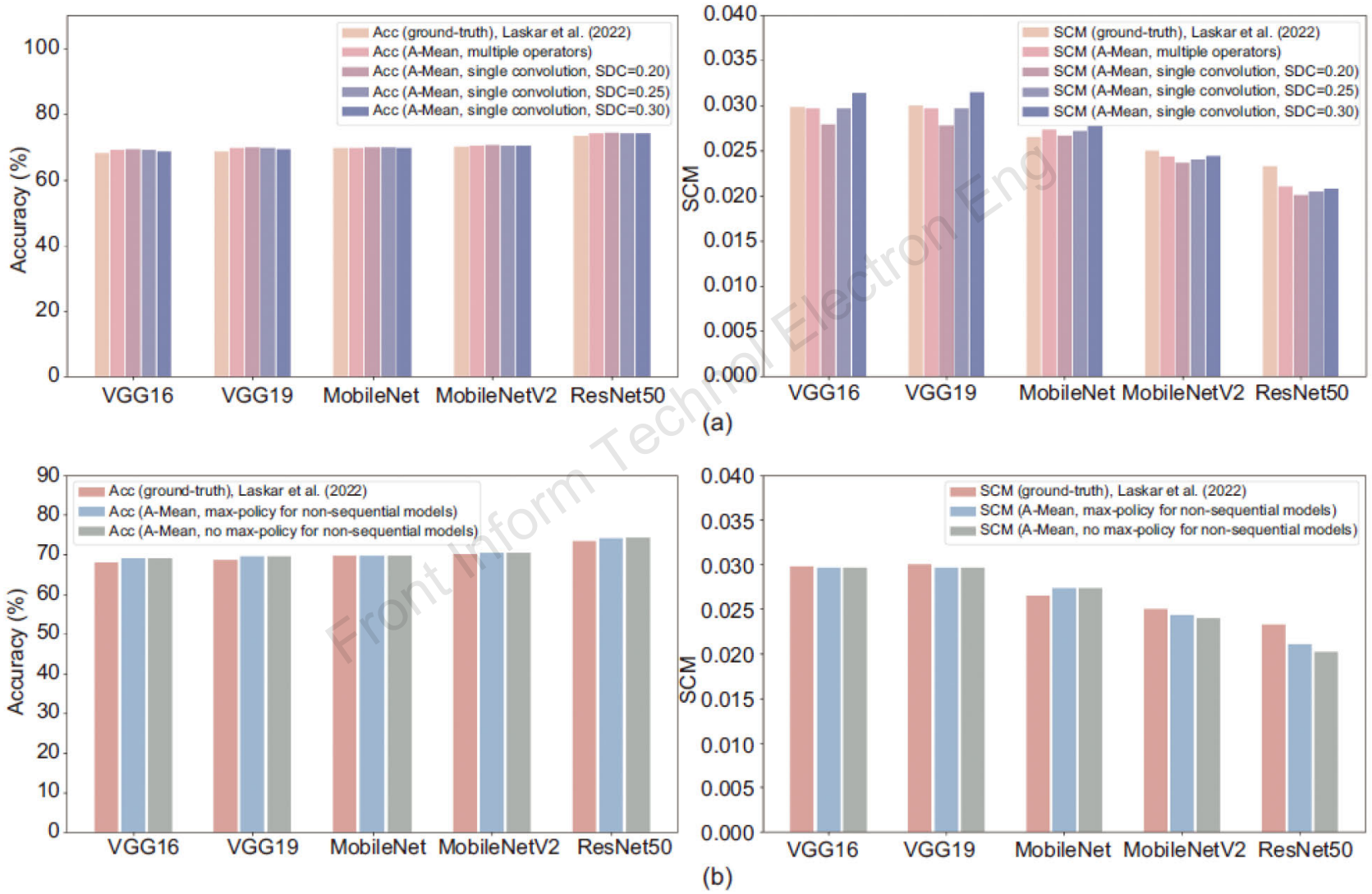
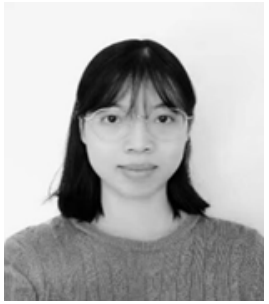


Fig. 8 SCM and accuracy improvement using max-policy on different strategies: (a) different topologies; (b) different operations

Conclusions

This paper presents a novel static two-level mean calculation model, A-Mean. This proposed approach can take advantages of one-time dynamic execution and static analysis for accurate, fast, and automatic estimation using the worst-case policy. The static two-level mean calculation refers to inner-layer and inter-layer mean calculations. More importantly, the sequential and non-sequential DNN topologies are considered, and a max-policy is used to estimate the upper bound of multiple non-sequential cases.



Jiajia JIAO is currently an Associate Professor with Shanghai Maritime University. She received her Ph.D. degree in computer science engineering from Shanghai Jiao Tong University in 2014. From 2013 to 2014, she was a visiting Ph.D. student at Carnegie Mellon University; from 2019 to 2020, she was a visiting scholar at Cornell University. Her research interests include hardware reliability and security in processors and machine learning-assisted processor design.

Ran WEN received a B.S. degree in Internet of Things engineering from Jiangsu University of Science and Technology in 2022. She is currently working toward an M.S. degree in computer science and technology with the College of Information Engineering, Shanghai Maritime University.

Hong YANG received a B.S. degree in computer science and technology from Shanghai Maritime University in 2022. He is currently working toward an M.S. degree in computer science and technology with the College of Information Engineering, Shanghai Maritime University.