CICC: Channel Pruning via the Concentration of Information and Contributions of Channels

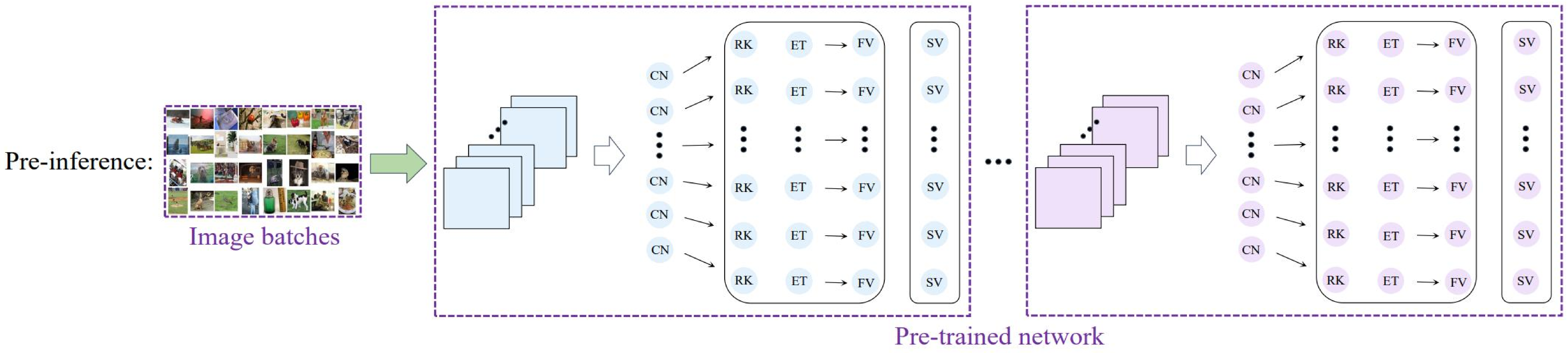
Yihao Chen[1] yihaochen@zju.edu.cn
Zhishan Li[2] zhishanli@zju.edu.cn
Yingqing Yang[1] yingqingyang@zju.edu.cn
Lei Xie[2] leix@iipc.zju.edu.cn
Yong Liu[3] yongliu@iipc.zju.edu.cn
Longhua Ma[4] lhma_zju@zju.edu.cn
Shanqi Liu[2] shanqiliu@zju.edu.cn
Guanzhong Tian*[1] gztian@zju.edu.cn

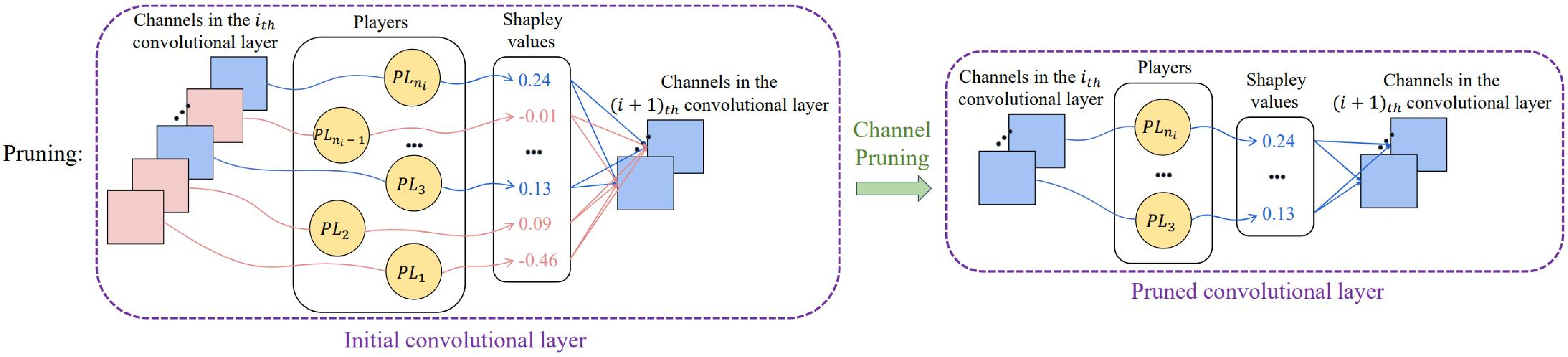
1 Ningbo Research Institute, Zhejiang University, Ningbo, China
2 College of Control Science and Engineering, Zhejiang University, Hangzhou, China
3 Zhejiang - Singapore Innovation and AI Joint Research Lab
4 College of Information Science and Engineering, NingboTech University, Ningbo, China

Abstract

Channel pruning provides a promising prospect to compress and accelerate convolutional neural networks. However, existing pruning methods neglect the compression sensitivity of different layers and adjust the pruning rate through engineering tuning. To address this problem, we propose to assign the layer-wise pruning ratio via the concentration of information for the convolutional layers. Specifically, we introduce the rank and entropy of convolutional layers as indicators of the redundancy and amount of information, respectively. After that, we define a fusion function, which compromises these two indicators, to represent the concentration of information for the convolutional layers. Additionally, for pruning filters with interpretability and intuition, we propose to evaluate the importance of channels by leveraging Shapley values, which fairly distribute the *average marginal contributions* among them. Extensive experiments on various architectures and benchmarks demonstrate the promising performance of our proposed method (CICC). For example, CICC achieves an accuracy increase of 0.21% with FLOPs and parameters reductions of 45.5% and 40.3% on CIFAR-10. Besides, CICC obtains Top-1/Top-5 accuracy of 0.43%/0.11% with FLOPs and parameters reductions of 41.6% and 35.0% on ImageNet. It is worth noting that our method can still achieve excellent accuracy under high acceleration rates for pruning ResNet-110 on CIFAR-10.

Methodology




Pruning:
Channels in the i_{th} convolutional layer
Players
Shapley values
PL_{n_i}
PL_{n_i-1}
PL_3
PL_2
PL_1
0.24
-0.01
0.13
0.09
-0.46
Channels in the $(i+1)_{th}$ convolutional layer
Initial convolutional layer
Channel Pruning
Pruned convolutional layer

The framework of our method is divided into two phases. (1) Pre-inference: We feed randomly sampled image batches to obtain the rank and entropy (denoted by "RK" and "ET"), the corresponding fusion values (denoted by "FV") and the Shapley values (denoted by "SV") of each channel (denoted by "CN") in the convolutional layers. (2) Pruning: The channels in a layer are regarded as players (denoted by "PL"), and a negative Shapley value indicates that the player poses an adverse contribution to the cooperation. In each layer, the channels with the smallest Shapley values (pink squares) are discarded.

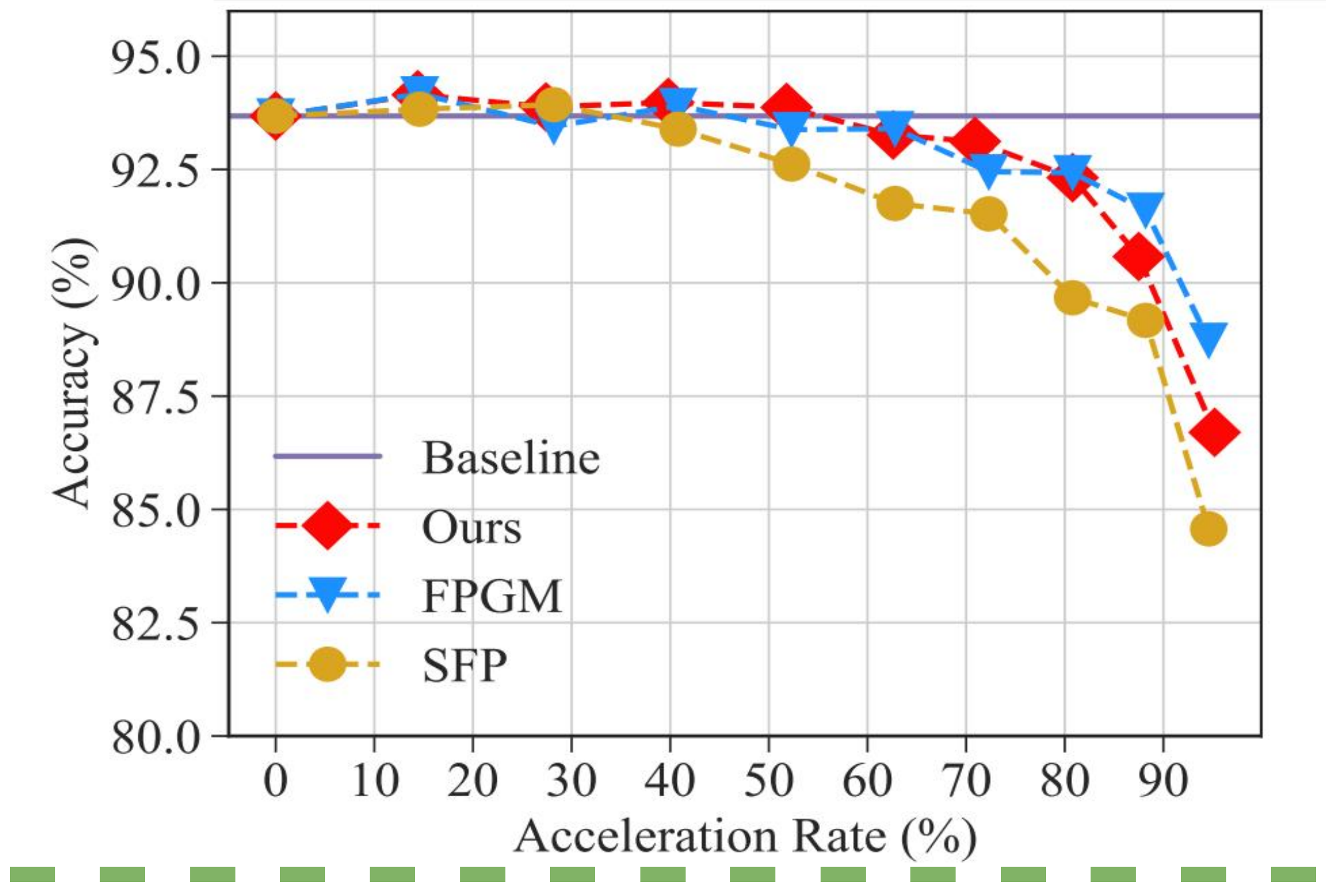
Results

CIFAR-10

Model	Method	Base. Acc. (%)	Accl. Acc. (%)	Acc. ↓ (%)	FLOPs ↓ (%)	Params ↓ (%)
VGG-16	SSS [15]	93.96	93.02	0.94	41.6	**73.8**
	CP [12]	93.26	90.80	2.46	50.6	–
	CICC	93.91	93.17	**0.74**	**52.3**	45.7
	CICC	93.91	93.38	**0.53**	61.0	50.7
	HRank [23]	93.96	92.34	1.62	**65.3**	**82.1**
ResNet-56	GAL [24]	93.33	92.98	0.35	37.6	11.8
	ACTD [50]	93.69	93.76	-0.07	40.0	**50.0**
	CICC	93.39	93.60	**-0.21**	45.5	40.3
	AMC [13]	92.80	91.90	0.90	**50.0**	–
	FPGM [11]	93.59	93.26	0.33	52.6	–
	DBP [49]	93.69	93.27	0.42	52.0	40.0
	CICC	93.39	93.11	0.28	58.1	**43.9**
	Graph [31]	93.27	93.38	**-0.11**	**60.3**	43.0
ResNet-110	SFP [10]	93.68	93.86	-0.18	40.8	–
	HRank [23]	93.50	94.23	-0.73	41.2	39.4
	CICC	93.68	94.56	**-0.88**	**45.6**	**40.4**
	GAL [24]	93.50	92.74	0.76	48.5	**44.8**
	FPGM [11]	93.68	93.74	-0.16	52.3	–
	CICC	93.68	94.16	**-0.48**	**58.1**	44.0
DenseNet-40	CC [20]	94.81	94.67	**0.14**	**47.0**	51.9
	CICC	94.22	93.56	0.66	44.4	**60.8**
	HRank [23]	94.81	93.53	**1.28**	54.7	56.7
	CICC	94.22	92.54	1.68	**59.6**	**68.6**

ImageNet

Model	Method	Base. Top-1 Acc. (%)	Accl. Top-1 Acc. (%)	Base. Top-5 Acc. (%)	Accl. Top-5 Acc. (%)	Top-1 Acc. ↓ (%)	Top-5 Acc. ↓ (%)	FLOPs ↓ (%)	Params ↓ (%)
ResNet-50	DSA [36]	76.02	75.10	92.86	92.45	0.92	0.41	40.0	-
	CICC	76.13	75.70	92.86	92.75	**0.43**	**0.11**	41.6	35.0
	SFP [10]	76.15	74.61	92.87	92.06	1.54	0.81	41.8	–
	DECORE [1]	76.15	74.58	92.87	92.18	1.57	0.69	**44.7**	**42.3**
	DSA [36]	76.02	74.69	92.86	92.06	1.33	0.80	50.0	–
	TPP [47]	76.13	75.60	–	–	0.53	–	–	–
	CICC	76.13	75.29	92.86	92.47	0.84	**0.39**	**50.4**	**44.2**
	Fisher [27]	76.79	76.42	–	–	**0.37**	–	**50.4**	–
ResNet-101	FPGM [11]	77.37	77.32	93.56	93.56	0.05	0.00	42.2	–
	CICC	77.37	77.35	93.55	93.59	**0.02**	**-0.04**	**43.7**	**42.6**
	Rethinking [29]	77.37	75.27	–	–	2.10	–	47.0	–
	CICC	77.37	76.10	93.55	92.94	**1.27**	**0.61**	**54.4**	**54.0**



Compared with SFP and FPGM for pruning ResNet-110 on CIFAR-10 *w.r.t* the acceleration rates, our method achieves higher accuracy than the baseline model (93.68%) when the acceleration ratio is not more than 51.8%, while the performance of FPGM only exceeds the baseline model under 14.6% and 40.8% FLOPs reductions.