RUTGERS
THE STATE UNIVERSITY
OF NEW JERSEY

CHIP: CHannel Independence-based Pruning for Compact Neural Networks

Yang Sui, Miao Yin, Yi Xie, Huy Phan,
Saman Zonouz, Bo Yuan

Outline

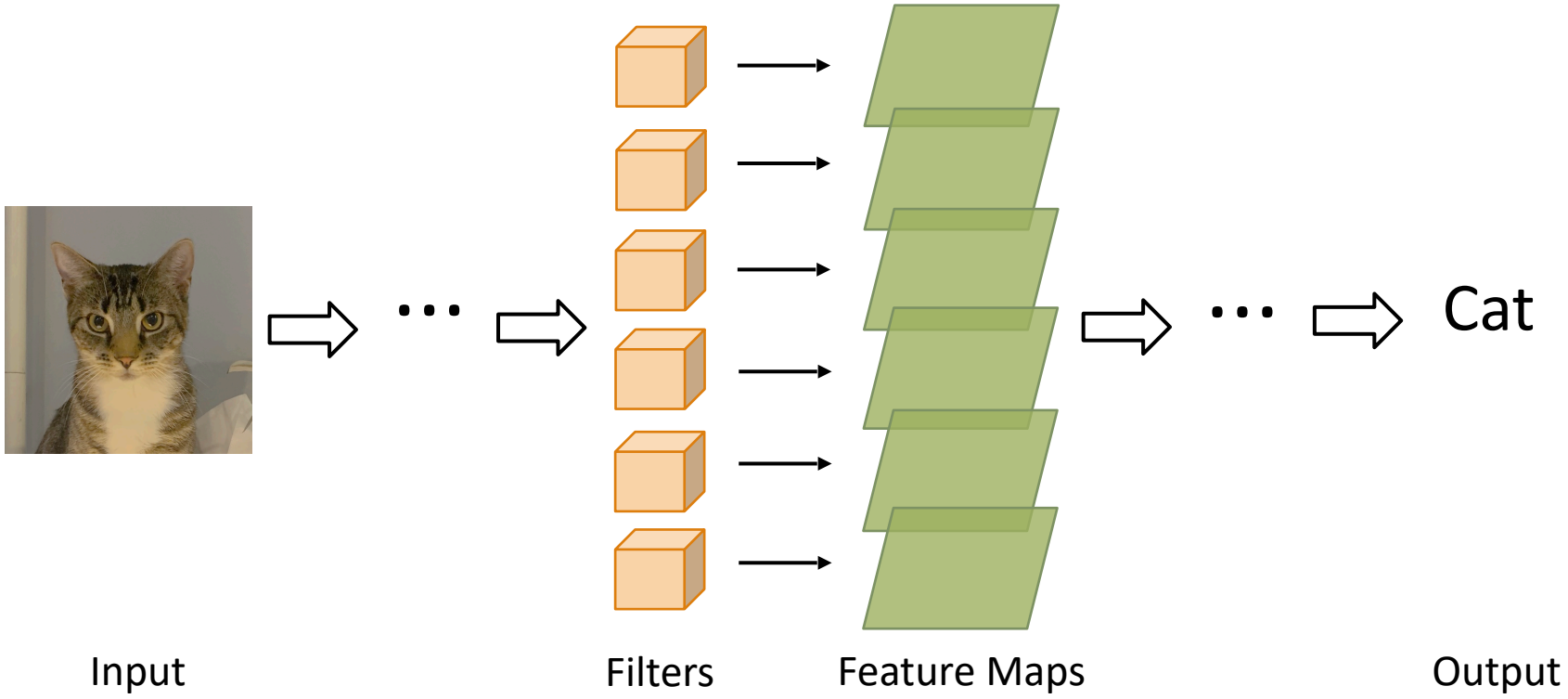
- Background
  - What is the model pruning in CNN?
- Motivation
  - Why do we propose CHIP?
- Approach
  - How does CHIP work?
- Results

BACKGROUND

- What is channel pruning in CNN?

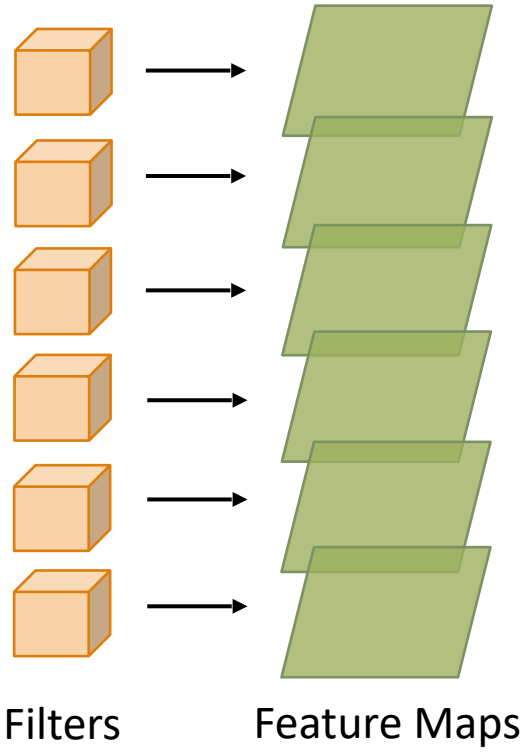
What is the channel pruning in CNN?

- Remove unimportant channels(filters) in convolutional layers
- Make model smaller and run faster



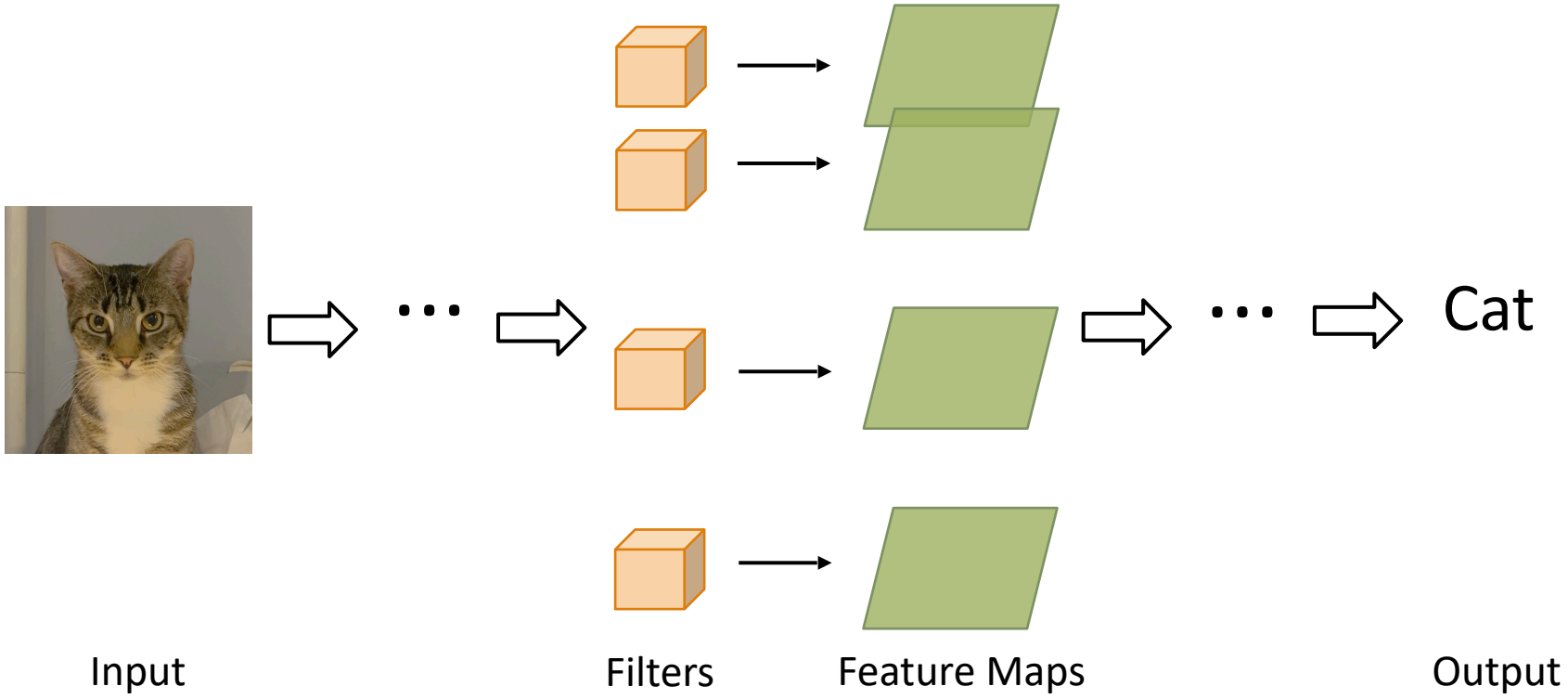
What is the channel pruning in CNN?

- Remove unimportant channels(filters) in convolutional layers
- Make model smaller and run faster



What is the channel pruning in CNN?

- Remove unimportant channels(filters) in convolutional layers
- Make model smaller and run faster



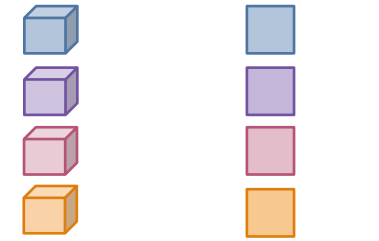
MOTIVATION

- Why do we propose CHIP?

Which filter is unimportant?

- From the perspective of Linear Independence

Given



If  are linearly dependent,

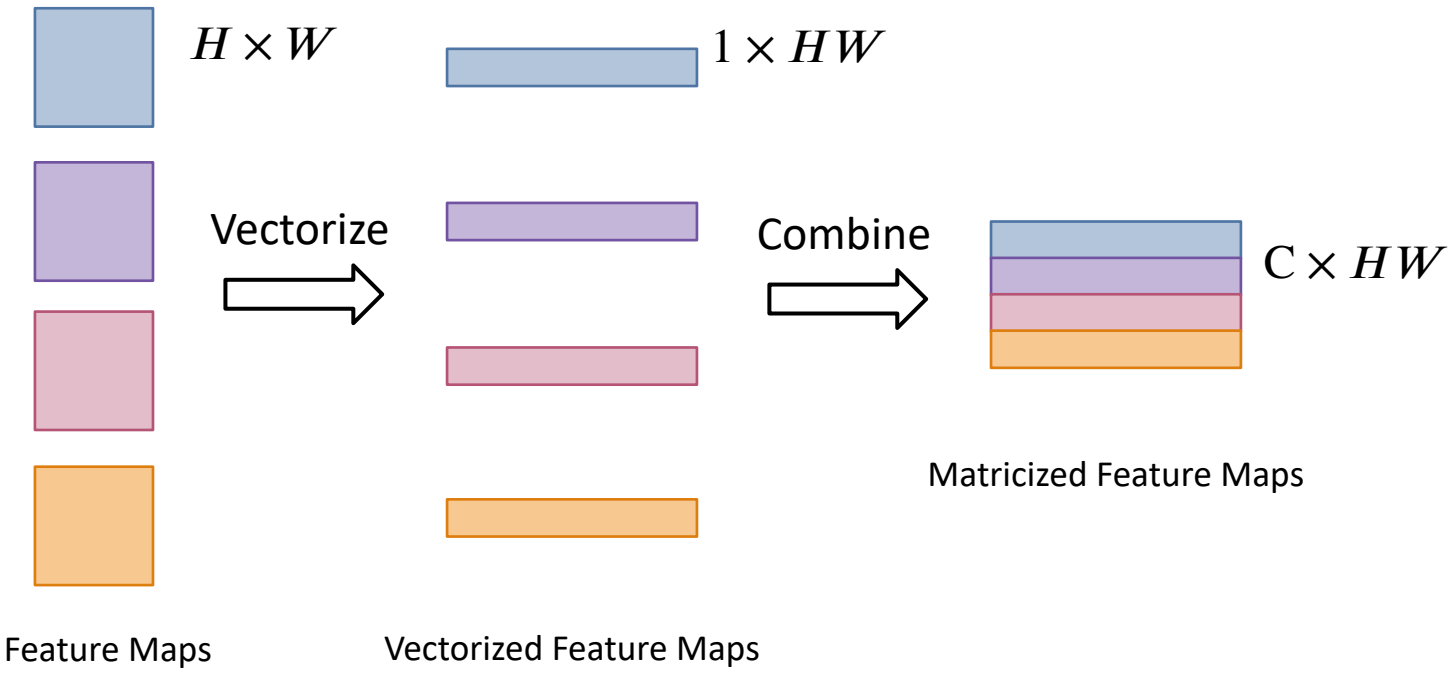
then, the channel from one of  can be removed.

Because  can be linearly represented by  and  .

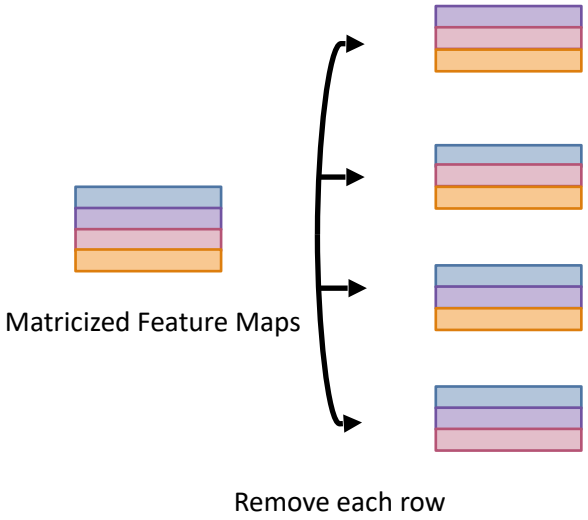
APPROACH

- From linear independence, how does CHIP work?

Define: Matricized Feature Maps



Measure Linear Independence of each channel



Compute the Rank

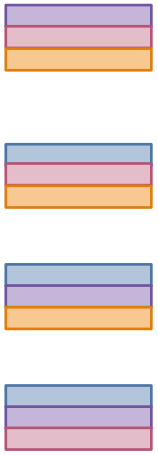
Measure Linear Independence of each channel

IF  $= 1.2 \times$  $+ \ 0.7 \times$ 

Contribution to Rank

Matricized Feature Maps

Rank = 3

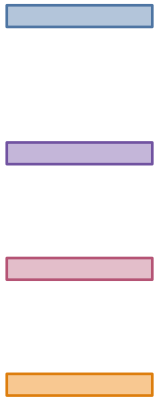


Rank = 3

Rank = 3

Rank = 3

Rank = 2



= 0

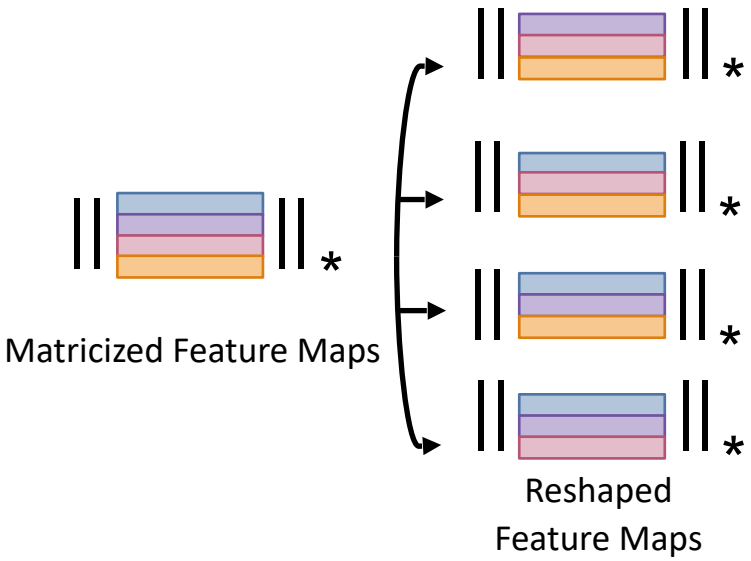
= 0

= 0

= 1 Important

Q:Which channel is least unimportant among  ?

Use Nuclear Norm instead of Rank

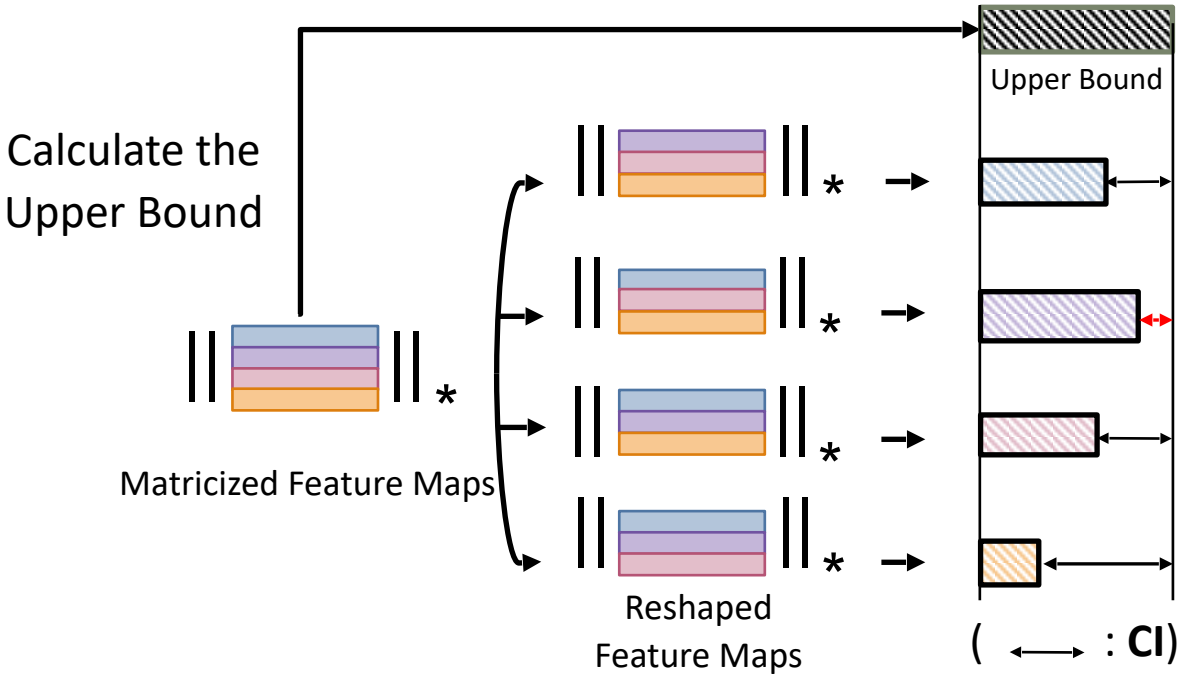


Pruning Criterion: CI(Channel Independence)

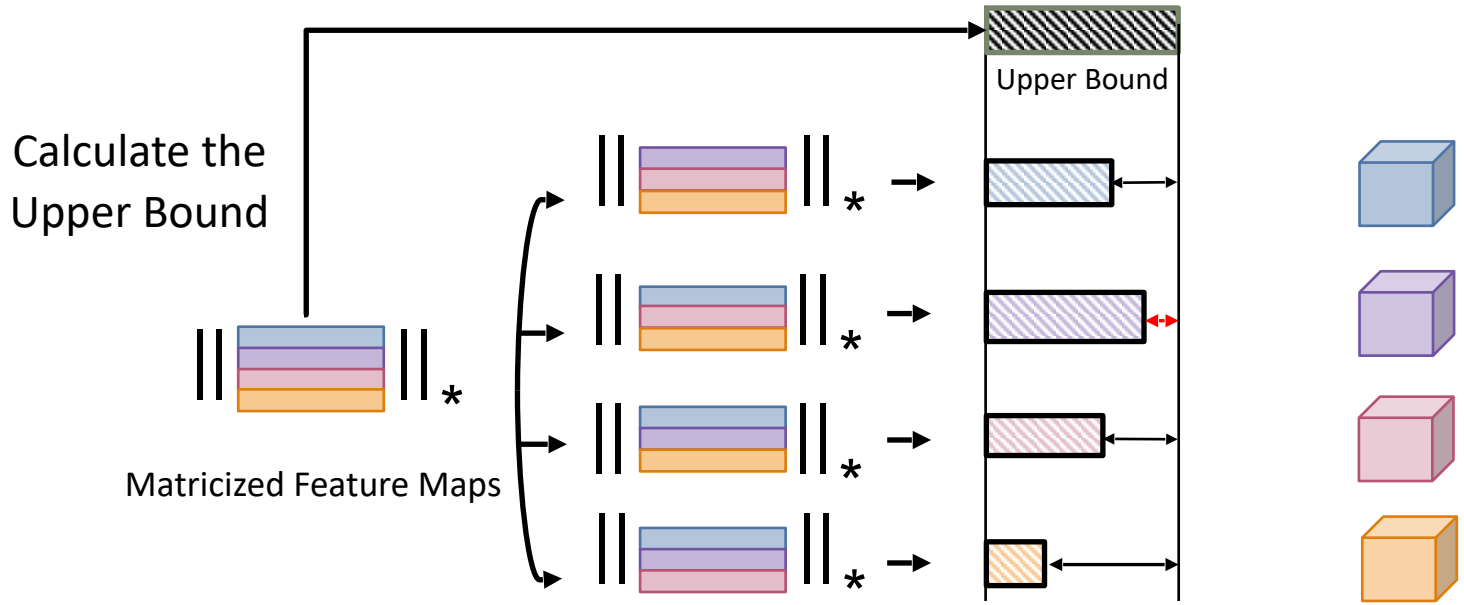
Definition 1 (Channel independence of single feature map) *For the i-th layer with output feature maps $\mathcal{A}^l = \{A_1^l, A_2^l, \cdots, A_{c^l}^l\} \in \mathbb{R}^{c^l \times h \times w}$, the Channel Independence (CI) of one feature map $A_i^l \in \mathbb{R}^{h \times w}$ in the i-th channel is defined and calculated as:*

$$CI(A_i^l) \triangleq \|A^l\|_* - \|M_i^l \odot A^l\|_*, \tag{3}$$

where $A^l \in \mathbb{R}^{c^l \times hw}$ is the matricized $\mathcal{A}^l$, $\|\cdot\|_*$ is the nuclear norm, $\odot$ is the Hardmard product, and $M_i^l \in \mathbb{R}^{c^l \times hw}$ is the *row mask matrix* whose i-th row entries are zeros and other entries are ones.



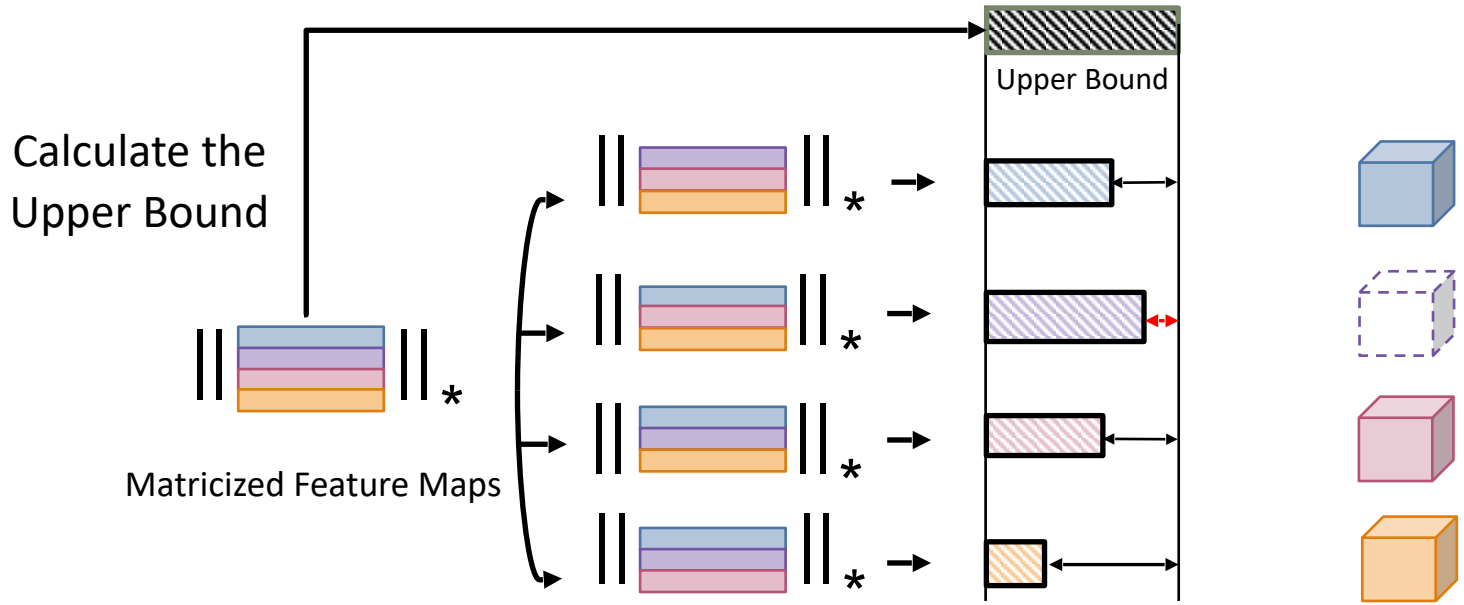
Remove the channel with the least CI



Q:Which channel is least unimportant among  ?

A: 

Remove the channel with the least CI



Q:Which channel is least unimportant among  ?

A: 

RESULTS

Table 2: Experimental results on ImageNet dataset.

Method	Top-1 Accuracy (%)			Top-5 Accuracy (%)			Params. ↓(%)	FLOPs ↓(%)
	Baseline	Pruned	Δ	Baseline	Pruned	Δ		
ResNet-50								
ThiNet (2017) [22]	72.98	72.04	-0.94	91.14	90.67	-0.47	33.72	36.8
SFP (2018)[9]	76.15	74.61	-1.54	92.87	92.06	-0.81	N/A	41.8
Autopruner (2020) [21]	76.15	74.76	-1.39	92.87	92.15	-0.72	N/A	48.7
FPGM (2019) [10]	76.15	75.59	-0.56	92.87	92.27	-0.60	37.5	42.2
Taylor (2019) [23]	76.18	74.50	-1.68	N/A	N/A	N/A	44.5	44.9
C-SGD (2019) [2]	75.33	74.93	-0.40	92.56	92.27	-0.29	N/A	46.2
GAL (2019) [19]	76.15	71.95	-4.20	92.87	90.94	-1.93	16.9	43
RRBP (2019) [36]	76.10	73.00	-3.10	92.9	91.00	-1.90	N/A	54.5
PFP (2019) [17]	76.13	75.91	-0.22	92.87	92.81	-0.06	18.1	10.8
HRank (2020) [18]	76.15	74.98	-1.17	92.87	92.33	-0.54	36.6	43.7
SCOP (2020) [30]	76.15	75.95	-0.20	92.87	92.79	-0.08	42.8	45.3
CHIP (Ours)	76.15	**76.30**	**+0.15**	92.87	**93.02**	**+0.15**	40.8	44.8
CHIP (Ours)	76.15	76.15	0.00	92.87	92.91	+0.04	**44.2**	**48.7**
PFP (2019) [17]	76.13	75.21	-0.92	92.87	92.43	-0.44	30.1	44
SCOP (2020) [30]	76.15	75.26	-0.89	92.87	92.53	-0.34	51.8	54.6
CHIP (Ours)	76.15	**75.26**	**-0.89**	92.87	**92.53**	**-0.34**	**56.7**	**62.8**
HRank (2020) [18]	76.15	71.98	-4.17	92.87	91.01	-1.86	46.0	62.1
HRank (2020) [18]	76.15	69.10	-7.05	92.87	89.58	-3.29	67.5	76.0
CHIP (Ours)	76.15	**73.30**	**-2.85**	92.87	**91.48**	**-1.39**	**68.6**	**76.7**