Constraint-Aware Importance Estimation for Global Filter Pruning under Multiple Resource Constraints

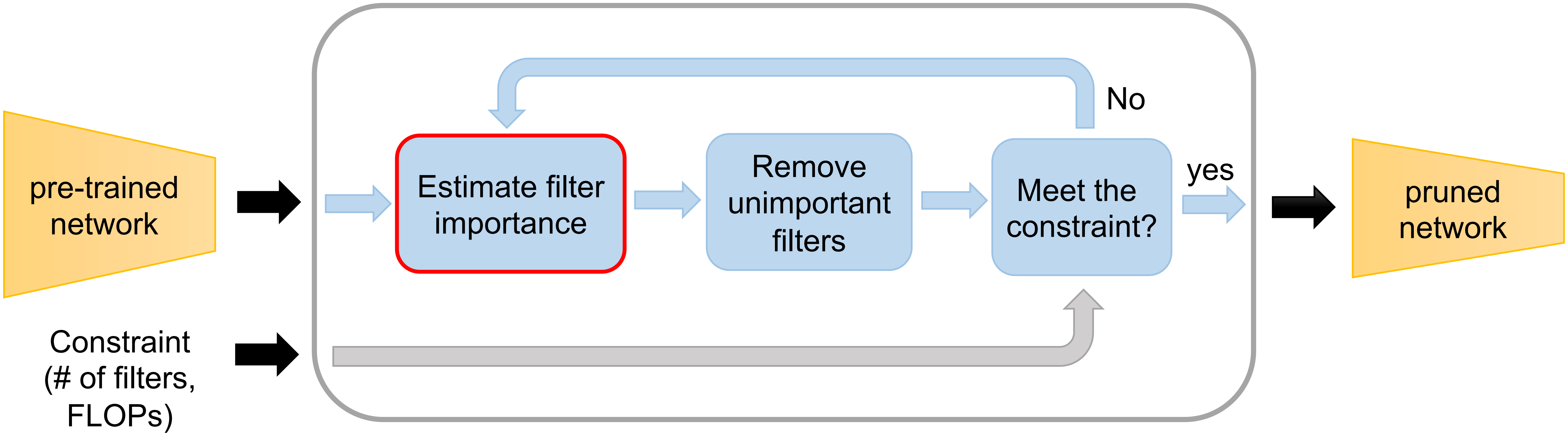
Yu-Cheng Wu, Chih-Ting Liu, Bo-Ying Chen, Shao-Yi Chien

NTU IoX Center, National Taiwan University

Graduate Institute of Electronic Engineering, National Taiwan University

Github: https://github.com/mediaic/CAIE-Filter-Pruning

Preliminary -- Filter Pruning



Proposed Method -- CAIE

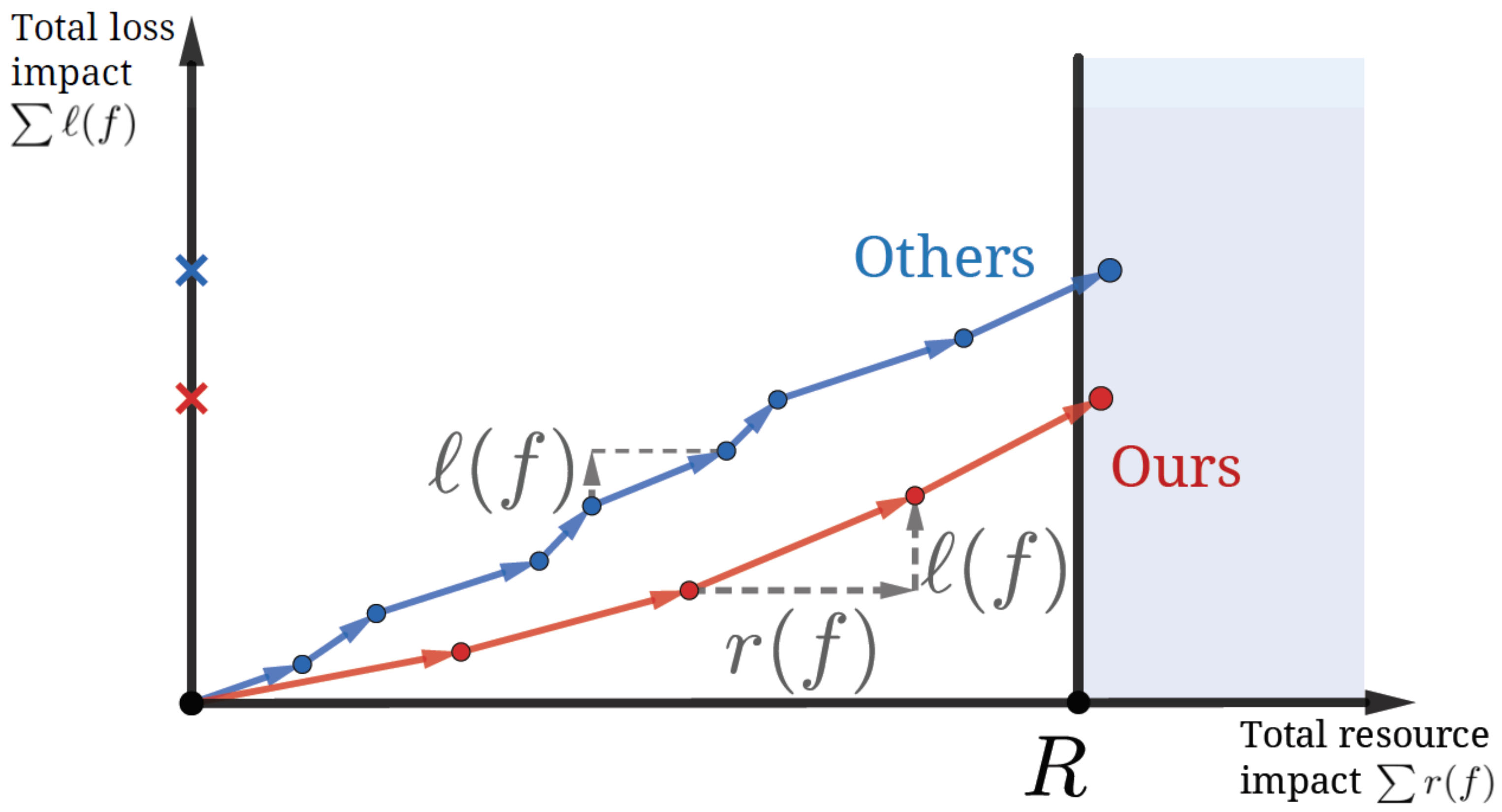
- Problem in previous methods
  - Information of the **constraint is not considered** during importance estimation
  - Under **multiple constraints**, they can only keep pruning until the network **separately** matching all constraints.

- Solution: our **Constraint-Aware Importance Estimation (CAIE)**
  - **Integrating constraint information** in the phase of importance estimation
  - Can be generalized to the problem of **multiple-constraint** pruning

Keywords and Notation

- Loss impact $\ell(f)$
  - The change in the loss induced by removing the filter *f*
- Resource impact $r(f)$
  - The proportion of reduction in the concerned resource induced by removing the filter *f*
- Pruning objective *R*
  - The minimum proportion of total reduction in the resource

CAIE in Single-constraint Pruning



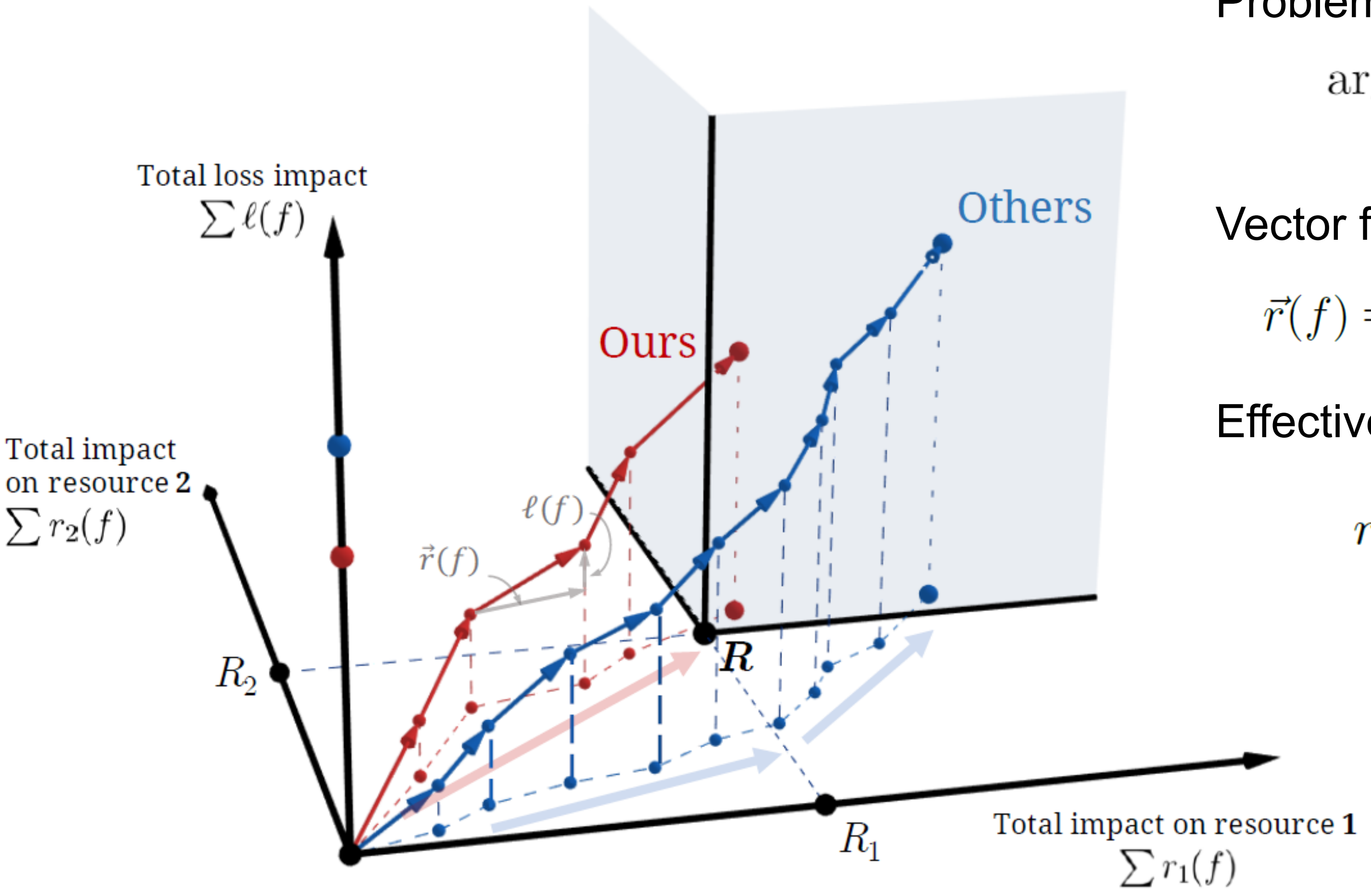
Problem formulation:

$$\underset{F}{\operatorname{argmin}} \sum_{f \in F} \ell(f) \quad s.t. \sum_{f \in F} r(f) \geq R$$

CAIE :

$$\boxed{\mathcal{I}_{sing}(f) = \frac{\ell(f)}{r(f)}}$$

CAIE in Multiple-constraint Pruning



Problem formulation:

$$\underset{F}{\operatorname{argmin}} \sum_{f \in F} \ell(f) \quad s.t. \sum_{f \in F} r_i(f) \geq R_i, \ \forall i \leq k$$

Vector form:

$$\vec{r}(f) = \langle r_1(f), r_2(f), \ldots, r_k(f) \rangle \qquad \vec{R} = \langle R_1, R_2, \ldots, R_k \rangle$$

Effective resource impact :

$$r_e(f) = \vec{r}(f) \cdot \frac{\vec{R}}{|\vec{R}|} = \frac{\sum_i r_i(f) R_i}{\sqrt{\sum_i R_i^2}}$$

CAIE :

$$\boxed{\mathcal{I}_{mul}(f) = \frac{\ell(f)}{r_e(f)}}$$

Effectiveness of CAIE

Model	Constraints	w/ CAIE	FLOPs left (%)	Param. left (%)	P. Top-1 (%)	Top-1↓ (%)	w/ − w/o CAIE (%)
ImageNet [19]							
ResNet-50 (orig. top-1 : 76.13%)	$f_{.33}$, $p_{.31}$	✗	**32.83**	**25.94**	71.57	4.56	-
	$f_{.33}$	✓	32.95	49.40	73.90	2.23	2.33
	$p_{.26}$	✓	46.64	25.80	71.96	4.17	0.39
	$f_{.33}$, $p_{.31}$	✓	32.90	30.76	72.39	3.74	0.82
	$f_{.33}$, $p_{.26}$	✓	**32.47**	**25.89**	71.92	4.22	**0.34**
ResNet-50 (orig. top-1 : 76.13%)	$f_{.65}$, $p_{.70}$	✗	**64.83**	**64.27**	75.59	0.54	-
	$f_{.65}$	✓	64.58	85.72	76.02	0.11	0.43
	$p_{.65}$	✓	79.80	64.70	75.80	0.33	0.21
	$f_{.65}$, $p_{.70}$	✓	64.95	69.88	75.83	0.30	0.24
	$f_{.65}$, $p_{.65}$	✓	**64.81**	**64.61**	75.69	0.44	**0.10**
ResNet-34 (orig. top-1 : 73.31%)	$f_{.78}$, $p_{.79}$	✗	**77.55**	**71.43**	72.67	0.64	-
	$f_{.78}$	✓	77.47	90.43	73.15	0.16	0.48
	$p_{.72}$	✓	85.89	71.29	72.72	0.59	0.05
	$f_{.78}$, $p_{.79}$	✓	77.43	78.94	72.91	0.40	0.24
	$f_{.78}$, $p_{.72}$	✓	**77.72**	**71.32**	72.73	0.58	**0.06**
CIFAR-10 [11]							
VGG16-BN (orig. top-1 : 93.34%)	$f_{.44}$, $p_{.20}$	✗	**43.32**	**9.93**	92.94	0.40	-
	$f_{.44}$	✓	44.00	12.55	93.06	0.28	0.12
	$p_{.10}$	✓	42.90	9.69	93.02	0.32	0.08
	$f_{.44}$, $p_{.20}$	✓	43.07	12.19	93.11	0.23	0.17
	$f_{.44}$, $p_{.10}$	✓	**42.43**	**9.89**	92.98	0.36	**0.04**
ResNet-34 (orig. top-1 : 94.13%)	$f_{.40}$, $p_{.15}$	✗	**29.90**	**14.48**	93.34	0.79	-
	$f_{.30}$	✓	29.82	19.95	93.48	0.65	0.14
	$p_{.15}$	✓	35.69	14.79	93.46	0.67	0.12
	$f_{.40}$, $p_{.15}$	✓	35.10	14.88	93.50	0.63	0.16
	$f_{.30}$, $p_{.15}$	✓	**29.64**	**14.79**	93.40	0.73	**0.06**

Comparison to state-of-the-arts (ImageNet)

Model	Orig. Top-1 (%)	Method	FLOPs left (%)	Param. left (%)	P. Top-1 (%)	Top-1↓ (%)
ResNet-50	76.18	Taylor-FO-BN-56% [16]	32.76	30.86	71.69	4.49
	76.13	**Ours** ($f_{.33}$, $p_{.31}$)	32.90	30.76	**72.39**	**3.74**
ResNet-50	72.88	Thinet-30 [15]	34.66	28.49	68.42	4.46
	76.13	**Ours** ($f_{.33}$, $p_{.26}$)	32.47	25.89	**71.92**	**4.22**
ResNet-50	76.15	FPGM-only 30% [8]	58.80	-	75.59	0.56
	76.13	**Ours** ($f_{.55}$)	54.77	77.35	**75.62**	**0.53**
ResNet-50	76.18	Taylor-FO-BN-81% [16]	65.03	69.92	75.48	0.70
	76.13	**Ours** ($f_{.65}$, $p_{.70}$)	64.95	69.88	**75.83**	**0.30**
ResNet-50	-	NISP-50-B [24]	55.99	56.18	-	0.89
	76.13	**Ours** ($f_{.56}$, $f_{.56}$)	55.89	55.84	**75.25**	**0.88**
ResNet-34	73.31	Taylor-FO-BN-82% [16]	77.74	78.90	72.83	0.48
	73.23	Li *et al.* [12]	75.80	89.20	72.17	1.04
	73.31	**Ours** ($f_{.78}$, $p_{.79}$)	77.43	78.94	**72.91**	**0.40**