LoFT: Finding Lottery Tickets through Filter-wise Training

Qihan Wang*, Chen Dun*, Fangshuo Liao*, Chris Jermaine, Anastasios Kyrillidis

*Equal Contribution

New Metric for Filter Distance

Let $\mathcal{R}$ and $\hat{\mathcal{R}}$ be two rankings of the filters. Let $\sigma : \mathcal{R} \to \hat{\mathcal{R}}$ such that $\sigma(\mathcal{R}_i) = \hat{\mathcal{R}}_i$. We introduce a new metric to measure filter similarity based on Spearman's footrule

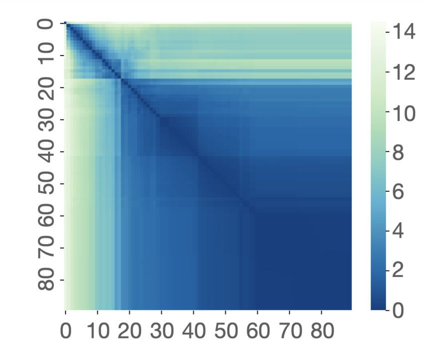
$$F_{\text{filter}}(\sigma) = \sum_i \frac{1}{i} \cdot |\ln(i) - \ln(\sigma(i))|$$

Properties of the Metric

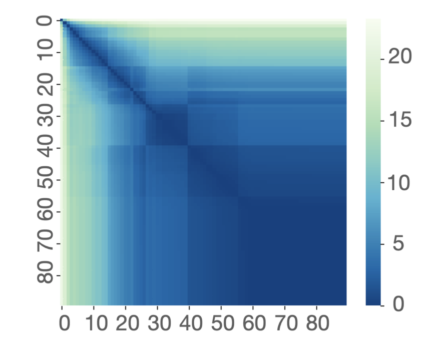
- $\ln(\cdot)$ is used to approximate the summation.
- $|\ln(i) - \ln(\sigma(i))|$ is larger if i is significantly different from $\sigma(i)$
- $\frac{1}{i}$ puts larger weight on filters with higher ranking in $\mathcal{R}$

Winning tickets appears before loss converges (darker=smaller distance)

- See figures on the right.

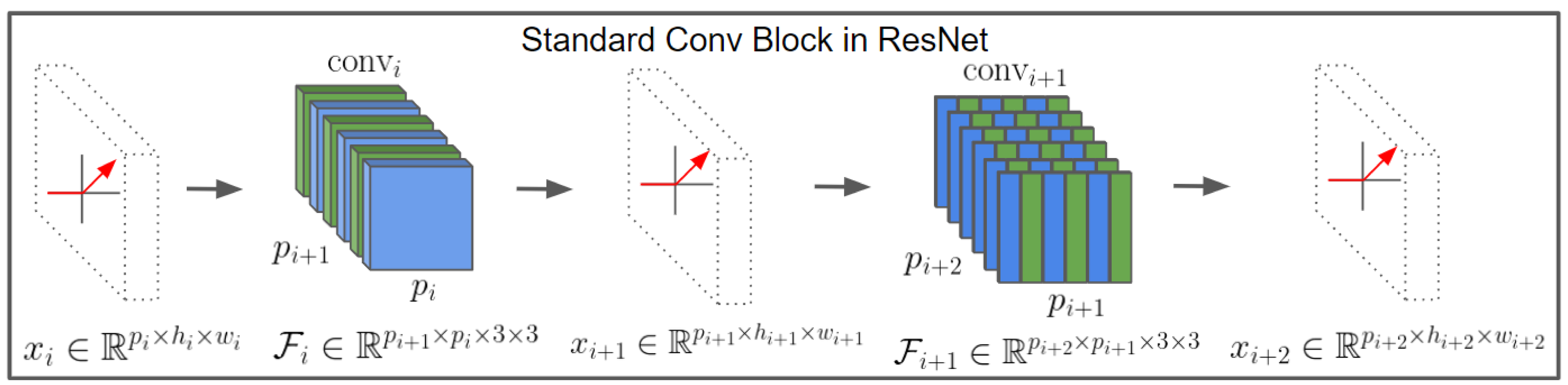


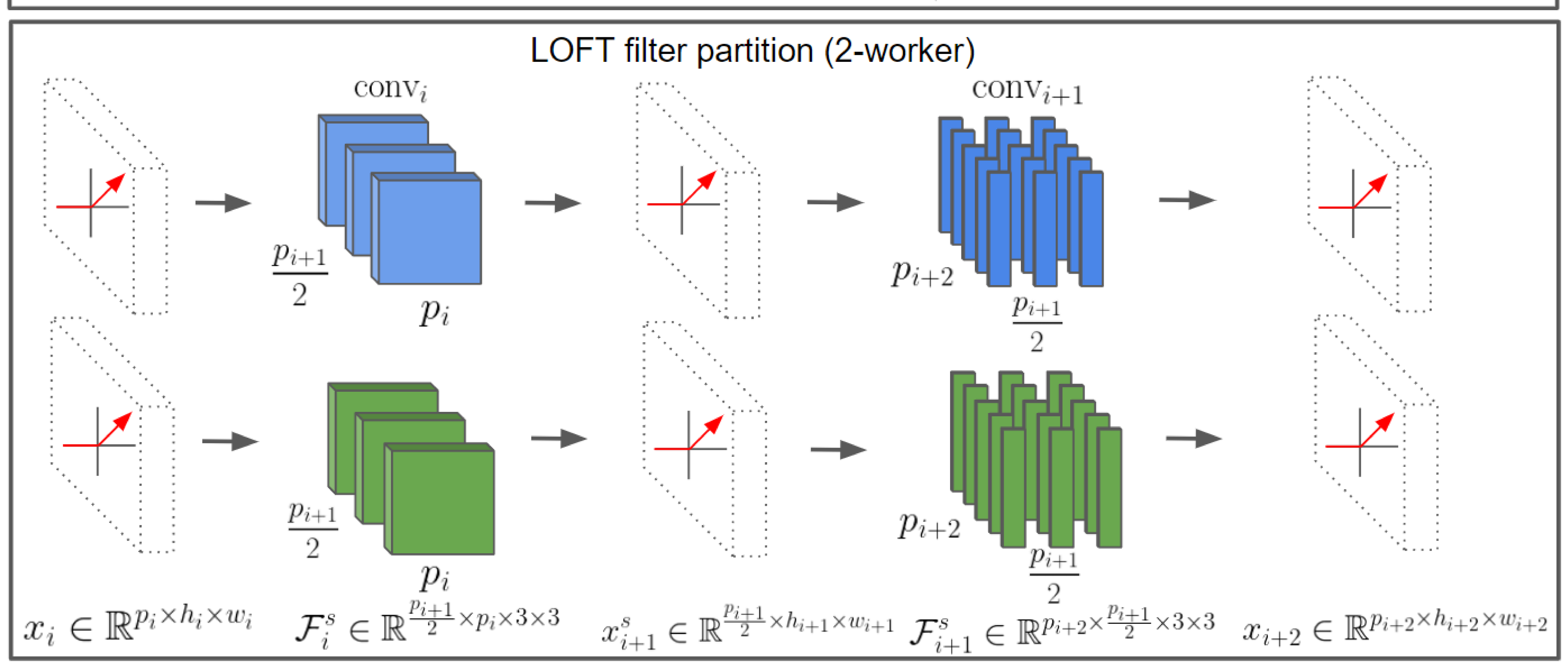
(a) conv-layer2



(b) conv-layer4

LoFT: A Filter-wise Partitioning Approach





LoFT Achieves Lower Communication Cost

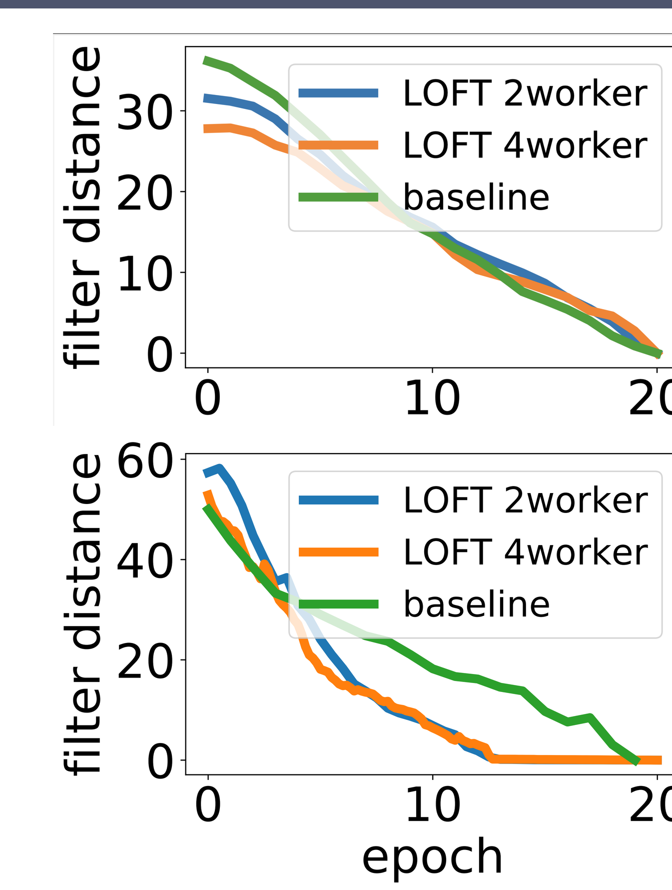
Setting	No-Prune	Methods	Pruning Ratio 80%	50%	30%	Comm cost	Improv.
PreResNet-34 CIFAR-10	93.51	GPipe-2	93.93	94.38		131.88G	
		LoFT-2	93.25	93.43		104.59G	1.26×
		GPipe-4	93.93	94.38		461.60G	
		LoFT-4	93.89	94.02		144.27G	3.20×
ResNet-34 CIFAR-10	93.22	GPipe-2	93.69	93.81		131.88G	
		LoFT-2	93.38	93.41		104.60G	1.26×
		GPipe-4	93.69	93.81		461.60G	
		LoFT-4	93.41	93.60		144.29G	3.20×
PreResNet-34 CIFAR-100	76.57	GPipe-2	76.72	77.09		131.88G	
		LoFT-2	75.93	77.27		104.77G	1.26×
		GPipe-4	76.72	77.09		461.60G	
		LoFT-4	75.77	76.79		144.64G	3.19×
ResNet34 CIFAR-100	75.93	GPipe-2	75.51	76.00		131.88G	
		LoFT-2	76.11	77.07		104.78G	1.26×
		GPipe-4	75.51	76.00		461.60G	
		LoFT-4	75.05	76.51		144.66G	3.19×
PreResNet-18 ImageNet	70.71	GPipe-2	66.71	69.14	70.29	20954.24G	
		LoFT-2	65.41	69.12	69.64	791.09G	21.60×
		GPipe-4	66.71	69.14	70.29	52385.59G	
		LoFT-4	65.60	68.93	69.77	1284.84G	40.77×

Finding Winning Tickets Faster

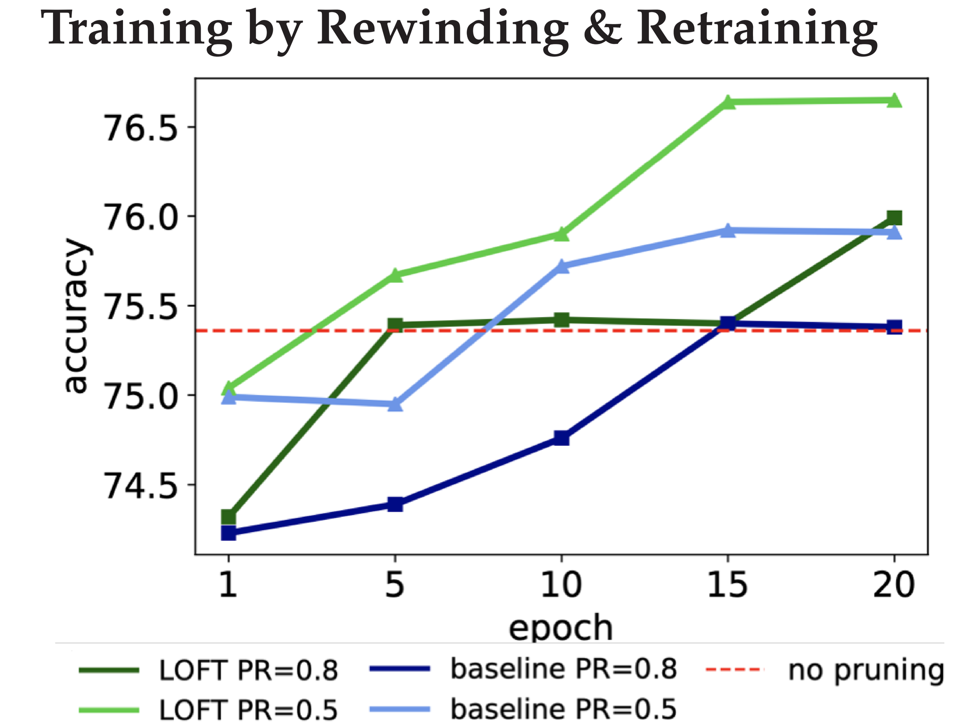
Filter distance between filter during training and the winning filter (See figures on the right.)

Top Figure: Results on CIFAR-100

Bottom Figure: Results on ImageNet



Evaluate Snapshots of Tickets during Training by Rewinding & Retraining



Theoretical Result: LoFT trajectory stays near GD trajectorye

Let $\mathbf{X} \in \mathbb{R}^{n \times d \times p}$ be the input data and $\mathbf{y} \in \mathbb{R}^n$ be the labels. Let f be a one-hidden-layer CNN with only the first layer filters $\mathbf{W}$ trainable. Let $\{\mathbf{W}_t\}_{t=0}^T$ and $\{\hat{\mathbf{W}}_t\}_{t=0}^T$ be the weights in the trajectory of LoFT and GD. Let S be the number of workers.

Theorem 1. *Assume the number of hidden filters satisfies* $m = \Omega\left(\frac{n^4T^2}{\lambda_0^4\delta^2}\max\{n,d\}\right)$ *and the step size satisfies* $\eta = O\left(\frac{\lambda_0}{n^2}\right)$*. Then, with probability at least* $1 - O(\delta)$ *we have:*

$$\mathbb{E}_{[\mathbf{M}_T]}\left[\left\|\mathbf{W}_T - \hat{\mathbf{W}}_T\right\|_F^2\right] + \eta\sum_{t=0}^{T-1}\mathbb{E}_{[\mathbf{M}_T]}\left[\left\|f(\mathbf{X},\mathbf{W}_t) - f\left(\mathbf{X},\hat{\mathbf{W}}_t\right)\right\|_2^2\right] \le O\left(\frac{n^2\sqrt{d}}{\lambda_0^2\kappa m^{\frac{1}{4}}\sqrt{\delta}} + \frac{2\eta^2T\theta^2(1-\xi)\lambda_0}{S}\right).$$

Reference

[1] Binhang Yuan, Anastasios Kyrillidis, and Christopher M. Jermaine. Distributed Learning of Deep Neural Networks using Independent Subnet Training. *arXiv e-prints*, page arXiv:1910.02120, 2019.

[2] Jonathan Frankle and Michael Carbin. The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks. *arXiv e-prints*, page arXiv:1803.03635, 2018.