Few Shot Network Compression via Cross Distillation

Haoli Bai[1], Jiaxiang Wu[2], Irwin King[1], Michael Lyu[1]

*{**hlbai**, king, lyu}@cse.cuhk.edu.hk, jonathanwu@tencent.com*

[1]The Chinese University of Hong Kong, [2]Tencent AI Lab

Looking for internship!

Introduction

Most prevalent network compression methods require fine-tuning with sufficient training data to ensure accuracy, which could be challenged by privacy and security issues. Network compression with few shot training instances is a new direction for research.

Our Work: We propose cross distillation, a novel network compression approach specialized for few shot training samples. The proposed method offers a general framework compatible with pruning or quantization.

Methods

Overall Framework

❑ Task Definition

Given an over-parameterized teacher network $\mathcal{F}^T$, our goal is to learn a compact student network $\mathcal{F}^S$. We proceed in a layer-wise manner:

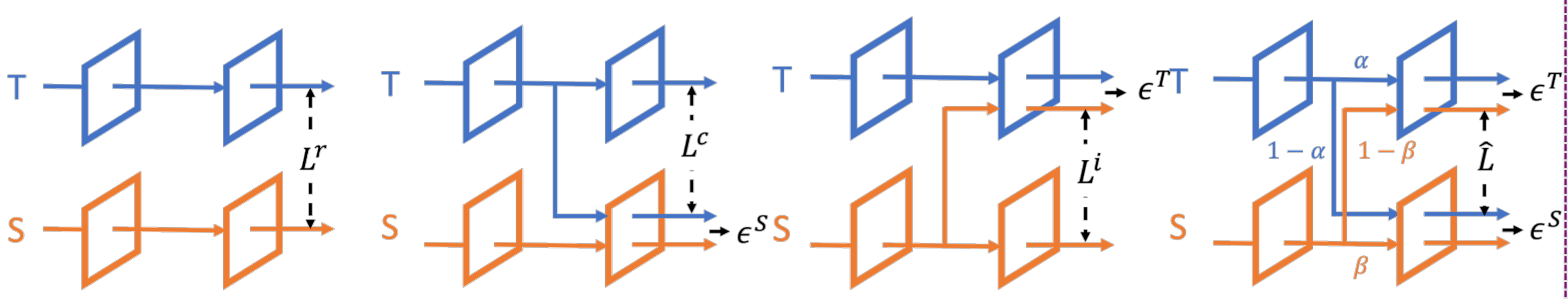
$$\mathbf{W}_*^S = \arg\min_{\mathbf{W}^S} \frac{1}{N}\mathcal{L}^r(\mathbf{W}^S) + \lambda\mathcal{R}(\mathbf{W}^S),$$

where $\mathcal{L}^r(\mathbf{W}^S) = \|\sigma(\mathbf{W}^T * \mathbf{h}^T) - \sigma(\mathbf{W}^S * \mathbf{h}^S)\|_F^2$ is the **estimation error**, and $\mathcal{R}(\mathbf{W}^S)$ is some regularization for sparsity/quantization.

⚠ Fewer training samples → larger estimation errors;

⚠ Estimation errors accumulate and propagate layer-wisely.

❑ Cross Distillation



(a) Layer-wise distillation (b) Correction (c) Imitation (d) Soft cross distillation

(b) Correction loss – reduce the historically accumulated errors ✓

$$\mathcal{L}^c(\mathbf{W}^S) = \|\sigma(\mathbf{W}^T * \mathbf{h}^T) - \sigma(\mathbf{W}^S * \mathbf{h}^T)\|_F^2$$

(c) Imitation loss – teacher-aware accumulated errors on student ✓

$$\mathcal{L}^i(\mathbf{W}^S) = \|\sigma(\mathbf{W}^T * \mathbf{h}^S) - \sigma(\mathbf{W}^S * \mathbf{h}^S)\|_F^2$$

(d) Soft cross distillation – a trade-off

$$\hat{\mathcal{L}}(\mathbf{W}^S) = \|\sigma(\mathbf{W}^T * \hat{\mathbf{h}}^T) - \sigma(\mathbf{W}^S * \hat{\mathbf{h}}^S)\|_F^2$$

(*) Convex combination between (b) and (c) – another trade-off

$$\tilde{\mathcal{L}} = \mu\mathcal{L}^c + (1-\mu)\mathcal{L}^i, \quad \mu \in [0,1]$$

Combination with Pruning/Quantization

We adopt proximal mapping update for different $\mathcal{R}(\mathbf{W}^S)$

$$\mathbf{W}_{t+1}^S = \mathrm{Prox}_{\lambda\mathcal{R}}(\mathbf{W}_t^S - \eta\nabla\tilde{\mathcal{L}}(\mathbf{W}_t^S)) \ \in R^{c_o\times c_i\times k\times k}$$

❑ Structured Pruning

$$\mathcal{R}(\mathbf{W}^S) = \|\mathbf{W}^S\|_{2,1} = \textstyle\sum_i \|\mathbf{W}_i^S\|_2$$

$$\mathrm{Prox}_{\lambda\|\cdot\|_2}(\mathbf{W}_i^S) = \max(1 - \frac{\lambda}{\|\mathbf{W}_i^S\|_2}, 0)\cdot\mathbf{W}_i^S$$

❑ Unstructured Pruning

$$\mathcal{R}(\mathbf{W}^S) = \|\mathbf{W}^S\|_1 = \textstyle\sum_{i,j,h,w}|W_{ijhw}^S|$$

$$\mathrm{Prox}_{\lambda\|\cdot\|_1}(W_{ijhw}^S) = \begin{cases} W_{ijhw}^S - \lambda & W_{ijhw}^S > \lambda \\ 0 & |W_{ijhw}^S| \le \lambda \\ W_{ijhw}^S + \lambda & W_{ijhw}^S < -\lambda \end{cases}$$

❑ Quantization

$$\mathcal{R}(\mathbf{W}^S) = \Pi_Q(g(\mathbf{W}^S)) \quad \mathcal{Q} = \{0, \tfrac{\pm1}{2^{B-1}-1}, \tfrac{\pm2}{2^{B-1}-1}, ..., \pm1\}$$

$$\mathrm{Prox}_{\Pi_Q}(W_{ijhw}^S) = 0 \text{ if } W_{ijhw}^S \in Q \text{ else } \infty$$

Theoretical Analysis

Theorem 1. *Suppose both $\mathcal{F}^T$ and $\mathcal{F}^S$ are L-layer convolutional neural networks followed by the un-pruned softmax fully-connected layer. If the activation functions $\sigma(\cdot)$ are Lipchitz-continuous such as $\mathrm{ReLU}()$, the gap of softmax cross entropy $\mathcal{L}^{ce}$ between the network logits $\mathbf{o}^T = \mathcal{F}^T(\mathbf{x})$ and $\mathbf{o}^S = \mathcal{F}^S(\mathbf{x})$ can be bounded by*

$$|\mathcal{L}^{ce}(\mathbf{o}^T;\mathbf{y}) - \mathcal{L}^{ce}(\mathbf{o}^S;\mathbf{y})| \le C\tilde{\mathcal{L}}_L + \sum_{l=1}^{L-1}\prod_{k=l}^{L} C_k'(\mu)\tilde{\mathcal{L}}_l, \quad (6)$$

where C and $C'(\mu)$ are constants and $C'(\mu)$ is linear in μ.

Experiments

Main Results

❑ Structured Pruning with ResNet-34 on ImageNet

Methods	50	100	500	1	2	3
L1-norm	$72.94_{\pm0.00}$	$72.94_{\pm0.00}$	$72.94_{\pm0.00}$	$72.94_{\pm0.00}$	$72.94_{\pm0.00}$	$72.94_{\pm0.00}$
BP	$83.18_{\pm1.86}$	$84.32_{\pm1.29}$	$85.34_{\pm0.89}$	$85.76_{\pm0.73}$	$86.05_{\pm0.51}$	$86.29_{\pm0.56}$
FSKD	$82.53_{\pm1.52}$	$84.58_{\pm1.13}$	$86.67_{\pm0.78}$	$87.08_{\pm0.76}$	$87.23_{\pm0.52}$	$87.20_{\pm0.43}$
FitNet	$86.86_{\pm1.81}$	$87.12_{\pm1.63}$	$87.73_{\pm0.96}$	$87.66_{\pm0.84}$	$88.61_{\pm0.76}$	$\mathbf{89.32}_{\pm0.78}$
ThiNet	$85.67_{\pm1.57}$	$85.54_{\pm1.39}$	$86.97_{\pm0.89}$	$87.42_{\pm0.76}$	$87.52_{\pm0.68}$	$87.53_{\pm0.50}$
CP	$86.34_{\pm1.24}$	$86.38_{\pm1.37}$	$87.41_{\pm0.80}$	$88.03_{\pm0.66}$	$87.98_{\pm0.49}$	$88.21_{\pm0.37}$
Ours-NC	$86.51_{\pm1.71}$	$86.61_{\pm1.20}$	$87.92_{\pm0.75}$	$87.98_{\pm0.60}$	$88.63_{\pm0.49}$	$88.82_{\pm0.38}$
Ours	$86.95_{\pm1.59}$	$87.60_{\pm1.13}$	$88.34_{\pm0.69}$	$88.17_{\pm0.73}$	$88.57_{\pm0.40}$	$88.59_{\pm0.41}$
Ours-S	$\mathbf{87.42}_{\pm1.69}$	$\mathbf{87.73}_{\pm1.17}$	$\mathbf{88.60}_{\pm0.82}$	$\mathbf{88.40}_{\pm0.61}$	$\mathbf{88.84}_{\pm0.48}$	$88.87_{\pm0.35}$

❑ Unstructured Pruning with VGG-16 on ImageNet

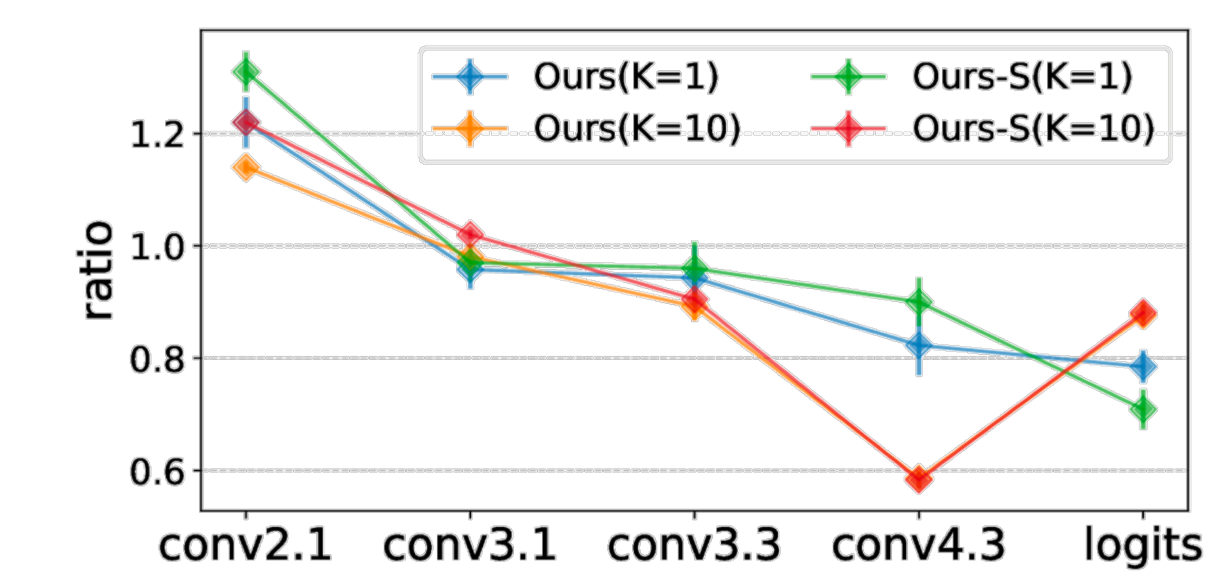
Methods	50	100	500	1	2	3
L1-norm	$0.5_{\pm0.00}$	$0.5_{\pm0.00}$	$0.5_{\pm0.00}$	$0.5_{\pm0.00}$	$0.5_{\pm0.00}$	$0.5_{\pm0.00}$
BP	$42.87_{\pm2.07}$	$48.78_{\pm1.43}$	$65.47_{\pm1.15}$	$71.25_{\pm0.97}$	$74.85_{\pm0.71}$	$76.04_{\pm0.48}$
FitNet	$52.66_{\pm2.93}$	$57.09_{\pm2.14}$	$76.59_{\pm1.45}$	$80.14_{\pm1.23}$	$82.27_{\pm0.70}$	$83.14_{\pm0.51}$
Ours-NC	$78.73_{\pm1.78}$	$83.29_{\pm1.12}$	$85.04_{\pm0.93}$	$85.36_{\pm0.61}$	$85.21_{\pm0.41}$	$85.49_{\pm0.46}$
Ours	$\mathbf{83.81}_{\pm1.49}$	$86.21_{\pm1.09}$	$87.19_{\pm0.96}$	$87.61_{\pm0.82}$	$87.78_{\pm0.45}$	$87.86_{\pm0.39}$
Ours-S	$83.67_{\pm1.52}$	$\mathbf{86.72}_{\pm1.23}$	$\mathbf{87.82}_{\pm1.04}$	$\mathbf{88.14}_{\pm0.74}$	$\mathbf{88.23}_{\pm0.61}$	$\mathbf{88.38}_{\pm0.43}$

❑ Weight Quantization with ResNet-56 on Cifar-10

	K=1		K=5	
	W2A32	W4A32	W2A32	W4A32
Ours-NC	$72.48_{\pm1.94}$	$85.75_{\pm0.96}$	$84.67_{\pm1.89}$	$91.09_{\pm0.37}$
Ours	$\mathbf{80.92}_{\pm2.23}$	$\mathbf{90.42}_{\pm0.53}$	$\mathbf{86.11}_{\pm1.97}$	$\mathbf{91.23}_{\pm0.45}$

Further Analysis

❑ Generalization Ability



❑ Sensitivity Analysis