A Discriminant Information Approach to Deep Neural Network Pruning

Zejiang Hou and Sun-Yuan Kung

Department of electrical engineering, Princeton University

{zejiangh, kung}@princeton.edu

Overview

DNN acceleration

- Motivation for feature-map pruning
- Taxonomy & challenges

Proposed Discriminant Information Feature-map pruning

- Quantifying feature-maps discriminant power
- Differential discriminant for channel pruning
- Intra-layer mixed precision quantization
- Performance evaluation and inference speedups

Summary

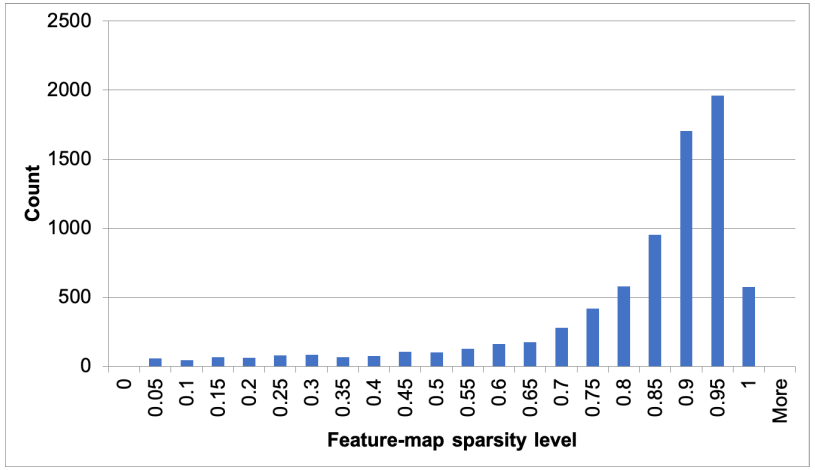
Feature-map Pruning: One of Many Efficient DNN Methods

Optimization goals for network inference

- Reduce network model size
- Speedup network execution time
- Maintain model accuracy

Observation that inspires feature-map pruning

- Biology: not all neurons are activated to solve a task
- Neural network:
  - Trained model has many redundant feature-maps, pruning which causes little degradation
  - Feature-maps can have (nearly) all 0s due to ReLU



Most of feature-maps have (nearly) all 0s (ResNet50 on ImageNet)

Taxonomy & Challenges

- Core of feature-map pruning – determine which feature-maps to prune based on a importance characterization metric
- Existing pruning methods
  - Importance metric based on filter weights information, e.g. filter norm, filter geometric median
  - Current discriminant feature-map pruning requires auxiliary cross-entropy losses for measuring and selecting feature-maps – retraining step is heavy in both computation time and human labor.
- *Many importance metric performs no better than random selection*

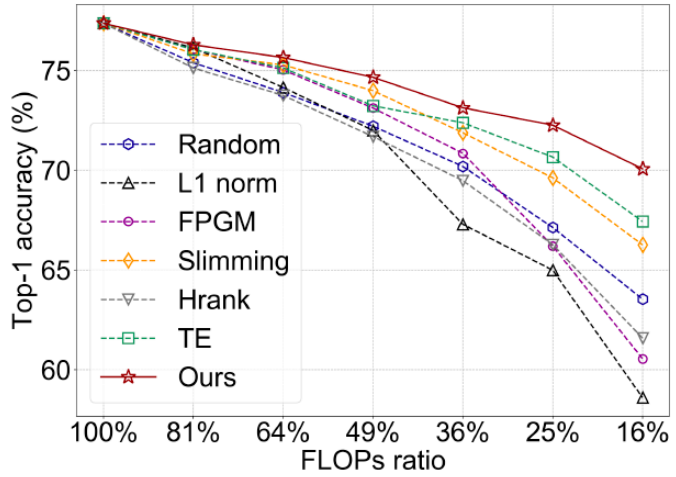


Fig. 1. Comparison of channel selection criteria in terms of testing accuracy. We prune VGG16 on CUB-200 with different FLOPs reductions.

Proposed Feature-map Pruning Method

- Quantifying feature-map discriminant power
  - Multi-class discriminant analysis $\underset{\mathbf{F}:\mathbf{F}^T(\bar{\mathbf{K}}^l+\rho\mathbf{I})\mathbf{F}=\mathbf{I}}{\text{maximize}} \ \text{tr}(\mathbf{F}^T\mathbf{K}_B^l\mathbf{F}) \Rightarrow DI = \text{tr}((\bar{\mathbf{K}}^l+\rho\mathbf{I})^{-1}\mathbf{K}_B^l)$
  - Relation to predictor learning $MRLSE = -\text{tr}((\bar{\mathbf{K}}^l+\rho\mathbf{I})^{-1}\mathbf{K}_B^l) + \|\mathbf{YC}\|_F^2 = -DI + \|\mathbf{YC}\|_F^2$
- Differential discriminant for feature-map pruning
  - Given a desired channel sparsity κ^l, we aim to find $\kappa^l C^l$ channels that maximally preserve the discriminant power
  - Channel importance by measuring influence on the DI value, i.e. the difference of DI value when $m_j^l = 1$ (channel j is present) and $m_j^l = 0$ (channel j is pruned)

$$\phi_j^l = \text{tr}\{[\text{diag}(\mathbf{1})\bar{\mathbf{K}}^l\,\text{diag}(\mathbf{1})+\rho\mathbf{I}]^{-1}\,\text{diag}(\mathbf{1})\mathbf{K}_B^l\,\text{diag}(\mathbf{1})\}- \quad (5)$$
$$\text{tr}\{[\text{diag}(\mathbf{1}-\mathbf{e}_j)\bar{\mathbf{K}}^l\,\text{diag}(\mathbf{1}-\mathbf{e}_j)+\rho\mathbf{I}]^{-1}\,\text{diag}(\mathbf{1}-\mathbf{e}_j)\mathbf{K}_B^l\,\text{diag}(\mathbf{1}-\mathbf{e}_j)\}$$

  - By relaxing the binary constraint on the indicator vector $\boldsymbol{m}^l$, ϕ_j^l can be approximated by the derivative of DI with respect to m_j^l (*differential discriminant*)

$$\phi_j^l \approx \frac{\partial\,\text{tr}\{[\text{diag}(\mathbf{m}^l)\bar{\mathbf{K}}^l\,\text{diag}(\mathbf{m}^l)+\rho\mathbf{I}]^{-1}\,\text{diag}(\mathbf{m}^l)\mathbf{K}_B^l\,\text{diag}(\mathbf{m}^l)\}}{\partial m_j^l}\Big|_{\mathbf{m}^l=\mathbf{1}} \quad (6)$$
$$= 2\rho([\bar{\mathbf{K}}^l]^{-1}\mathbf{K}_B^l[\bar{\mathbf{K}}^l]^{-1})_{jj}$$

Performance Evaluation

- Pruning ResNet-18/50 and MobileNetV2 on ImageNet

TABLE III

PRUNING RESULTS ON IMAGENET. "ΔTOP-1": TOP-1 ACCURACY DIFFERENCE DUE TO PRUNING, CALCULATED AS *pruned top-1 − pre-trained top-1*. "FLOPS": FLOPS(PRUNING RATIO). "PARAMS.": PARAMETERS(PRUNING RATIO). "-": RESULTS NOT REPORTED BY CORRESPONDING METHOD.

Model	Method	Top-1 (%)	ΔTop-1 (%)	FLOPs	Params.
ResNet18	TAS [49]	70.65 → 69.15	-1.50	1.21E9 (33.3%)	-
	FPGM [12]	70.28 → 68.41	-1.87	1.05E9 (41.8%)	-
	DI-greedy(Ours)	**69.76 → 68.91**	**-0.85**	**1.04E9 (42.5%)**	**7.82E6 (33.1%)**
	Sampling [42]	69.76 → 67.38	-2.38	1.28E9 (29.3%)	6.57E6 (43.8%)
	DCP [10]	69.76 → 67.35	-2.41	0.98E9 (46.1%)	6.19E6 (47.1%)
	DI-unif(Ours)	**69.76 → 68.15**	**-1.61**	**0.98E9 (46.1%)**	**6.19E9 (47.1%)**
ResNet50	SSS-32 [22]	76.10 → 74.18	-1.92	2.82E9 (31.1%)	18.60E9 (27.3%)
	Taylor-81 [19]	76.18 → 75.48	-0.70	2.66E9 (34.9%)	17.90E6 (30.1%)
	SFP [14]	76.15 → 74.61	-1.54	2.38E9 (41.8%)	-
	FPGM-30 [12]	76.15 → 75.59	-0.56	2.36E9 (42.2%)	-
	GAL-0.5 [23]	76.15 → 71.95	-4.20	2.33E9 (43.1%)	21.20E6(17.2%)
	LeGR [13]	76.10 → 75.70	-0.40	2.37E9 (42.0%)	-
	DI-SD(Ours)	**76.10 → 76.10**	**0**	**2.31E9 (43.5%)**	**16.70E6 (34.8%)**
	Sampling [42]	76.13 → 75.21	-0.92	2.86E9 (30.1%)	14.33E6 (44.0%)
	Hrank [11]	76.15 → 74.98	-1.17	2.30E9 (43.8%)	16.15E6 (36.9%)
	Taylor-72 [19]	76.18 → 74.50	-1.68	2.25E9 (45.0%)	14.20E6 (44.5%)
	FPGM-40 [12]	76.15 → 74.83	-1.32	1.90E9 (53.5%)	-
	C-SGD-50 [50]	75.34 → 74.54	-0.80	1.82E9 (55.8%)	-
	ThiNet-50 [27]	72.88 → 71.01	-1.87	1.82E9 (55.8%)	12.40E6(51.6%)
	DCP [10]	76.10 → 74.95	-1.15	1.82E9(55.8%)	12.40M (51.6%)
	DI-unif(Ours)	**76.10 → 75.50**	**-0.60**	**1.77E9 (56.7%)**	**12.10E6 (52.7%)**
MobileNetV2	DCP [10]	70.11% → 64.22%	-5.89%	1.65E8 (45%)	2.57E6 (25.9%)
	CPLI [29]	72.19% → 67.35%	-4.84%	1.65E8 (45%)	2.57E6 (25.9%)
	DI-unif(Ours)	71.80% → **69.33%**	**-2.47%**	**1.50E8 (50%)**	**1.92E6 (44.7%)**

Performance Evaluation (Cont.)

- Quantization results of compression ResNet50 on ImageNet

TABLE IV
COMPARISON OF DIFFERENT QUANTIZATION METHODS FOR COMPRESSING RESNET50 ON IMAGENET.

Method	Precision	Model Size	Top-1	Top-5
ResNet50	float 32-bit	97.49 MB	76.10%	92.93%
DC [52]	fixed 2-bit	6.32 MB	69.85%	88.68%
FSNet-WQ [53]	fixed 8-bit	≈8.37 MB	69.87%	89.61%
HAQ [4]	inter-layer mixed	6.30 MB	70.63%	89.93%
Ours	intra-layer mixed	6.09 MB	**73.21%**	**91.27%**

- Practical latency speedup

Table 11: Actual inference time speedup of pruned models on ImageNet. We measure on PyTorch platform with single NVIDIA P100 GPU using batch-size 64.

Model	Method	Top-1	Runtime	Latency↓
ResNet18	Baseline	69.76%	14.41ms	-
	DI-unif	68.15%	10.52ms	27%
ResNet50	Baseline	76.10%	49.97ms	-
	DI-unif	75.50%	30.98ms	38%
MobileNetV2	Baseline	71.80%	30.50ms	-
	DI-unif	69.33%	16.86ms	45%

Summary

- A feature-map discriminant perspective for feature-map pruning in deep neural networks
- Theoretical guidelines to effectively quantify the feature-map discriminant power
- An intra-layer mixed precision quantization scheme to further compress the network based on the same metric
- DI-based greedy pruning algorithm to automatically decides the target pruned architecture
- Experiments on various CNN architectures and benchmarks validates the effectiveness of our method