PDP: Parameter-free Differentiable Pruning is All You Need

Minsik Cho, Saurabh Adya, Devang Naik (minsik@apple.com)
NeurIPS 2023 · Apple Inc.

Methods

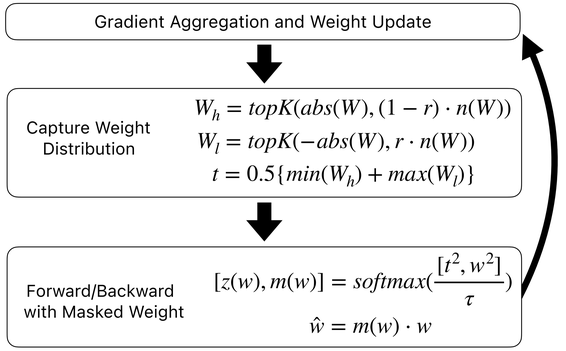
Differentiable pruning without extra-parameters

- Learning masks is expensive
  - Extra-parameter overheads
  - Complex training recipes and flows
- Dynamically inferring the soft masks
  - Capture the weight distributions using a threshold
  - Use probabilistic masks for differentiability
  - Second chances for the weights on the boundary
  - Easily expanded to structured pruning

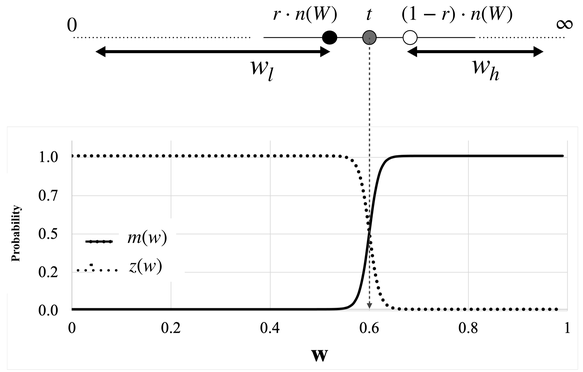
GPT2+OpenWebText

Method	Perplexity	Model #param	Extra #param	GPU cost($)
Dense	22.4	163M	0	
GMP [58]	37.7	163M	0	6997
OptG	33.7	163M	124M	11210
PDP	33.7	163M	0	7499

PDP training flow



Probability plot by w



Masked weight

$$\hat{w} = m(w)\cdot w = \frac{e^{\frac{w^2}{\tau}}}{e^{\frac{w^2}{\tau}} + e^{\frac{t^2}{\tau}}}\cdot w$$

Weight gradient

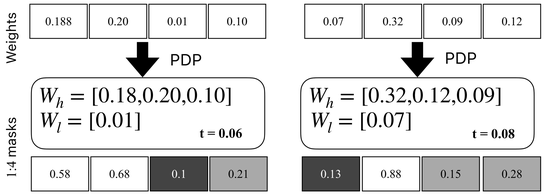
$$\Delta w = m(w)\Delta\hat{w} + 2\frac{w^2}{\tau}m(w)\{1-m(w)\}\Delta\hat{w}$$

Conventional gradient from masked weight

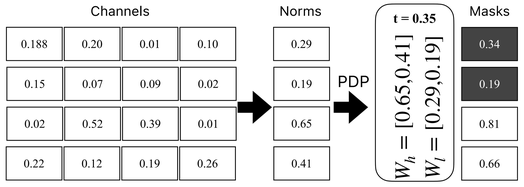
Always Positive
Maximized when m(w)=0.5

- Accelerate the learning of weights on the pruning boundary
- Additional gradient is zero when m(w) is 0 or 1
- τ is an inverse scaling factor

Extension



Application to N:M Pruning



Application to Channel Pruning

Results

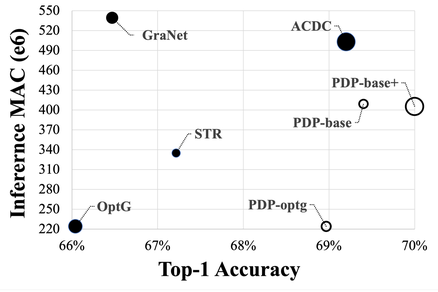
BERT + GLUE:MNLI

Validation dataset	Sparsity	Methods						
		Dense	STR*	OptG	GMP	MVP	POFA	PDP
matched	90	84.5	75.8	78.5	n/a	81.2	81.5	**83.1**
	94		74.4	76.9	74.8	80.7	n/a	**82.0**
mismatched	90	84.9	76.3	78.3	n/a	81.8	82.4	**83.0**
	94		74.1	76.5	75.6	81.2	n/a	**82.4**

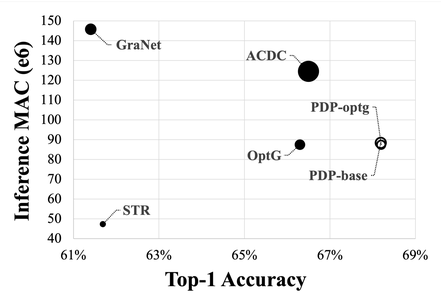
ResNet18/50 Structured N:M pruning + ImageNet1k

Network	Method	Batch size	#epochs	N:M				avg GPU cost ($)
				2:4	4:8	1:4	2:8	
ResNet18	LNM	256	120	69.6	**70.2**	65.1	68.4	395
	PDP	1024	100	**70.2**	70.1	**68.7**	**69.1**	275
ResNet50	LNM	256	120	74.6	75.1	74.1	75.0	812
	PDP	1024	100	**75.9**	**75.8**	**75.0**	**75.3**	380

ResNet50 + ImageNet1k



MobileNetV1 + ImageNet1k



ResNet50 Channel pruning + ImageNet1k

Method	Batch size	#epochs	Top-1 (%)	MAC drop (%)
NISP	?	90	75.3	44.0
DCP	256	60	75.0	55.0
SCP	256	100	75.3	54.3
PDP	1024	100	**75.9**	54.9

Easy to expand to structured pruning
Low cost training due to being parameter-free
SOTA results on ResNets+ImageNet1k

STR: soft-threshold weight reparameterization for learnable sparsity [ICML20]
GradNet: Sparsity training via boosting pruning plasticity with neuroregeneration [NeurIPS21]
OptG: Optimizing gradeint-driven criteria in network sparsity [CoRR22]
ACDC: Alternating compressed/decompressed training of deep neural networks [NeurIPS21]

GMP: To prune or not to prune: Exploring the efficacy of pruning for model compression [ICLR18]
DNW: Discovering neural wings [NeurIPS19]
MVP: Movement pruning: Adapative sparsity by fine-tuning [NeurIPS20]
POFA: Prune once for all: Sparse pre-trained language models [NeurIPS21]

LNM: Learning N:M fine-grained structured sparse neural networks from scratch [ICLR21]
NISP: pruning networks using neuron importance score propagation [CVPR18]
DCP: Discrimination-aware channel pruning for deep neural networks [NeurIPS19]
SCP: Operation-aware soft channel pruning using differentiable masks [ICML20]