sconce v0.99

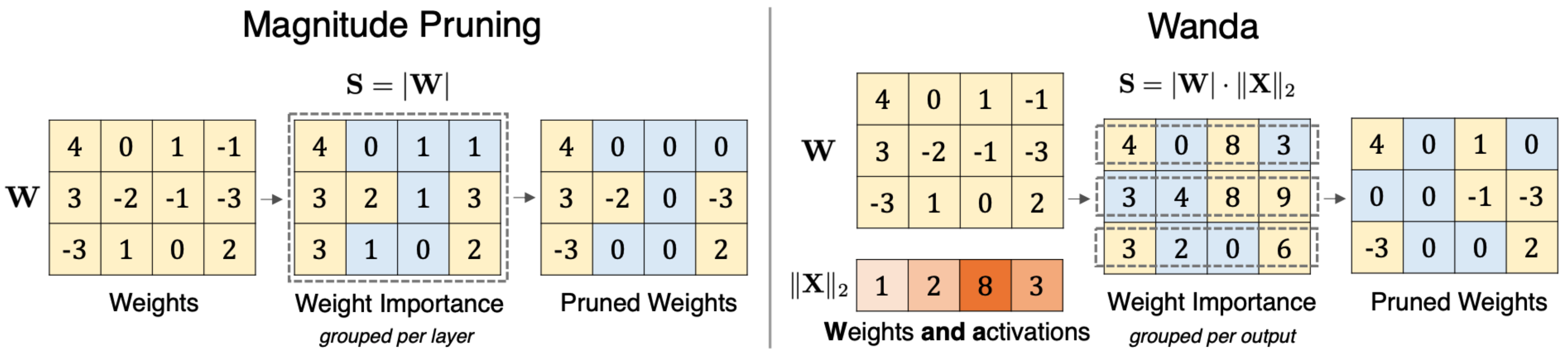
- Auto Sensitivity Scan for Pruning -> Finds Best Sparsity Ratio for Pruning [Least Performance Degradation and Max Performance]
- Supports CWP, GMP Pruning. Room for WANDA, GPTQ, etc..
- QAT
- Auto-Layer Fusion

sconce v1.1

Altruism is all you need !!! Don’t just be self attentive

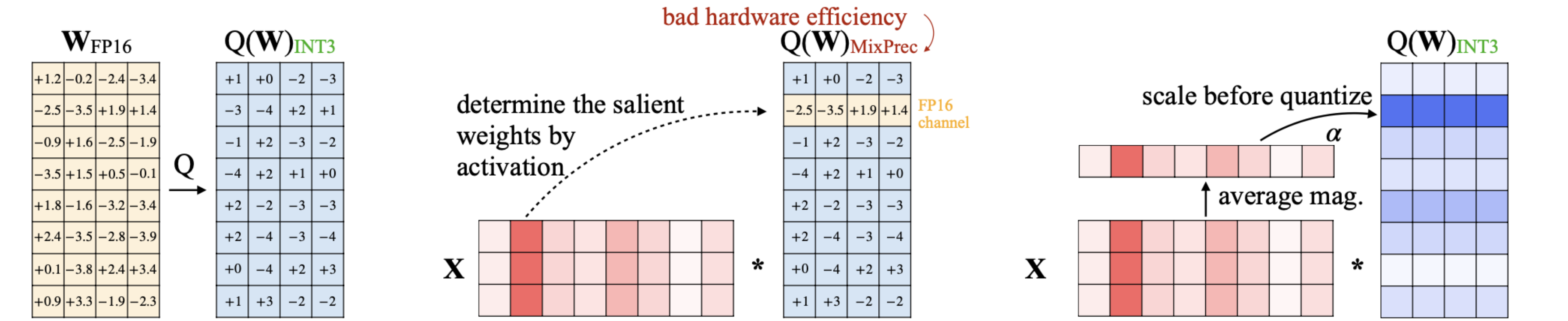
A SIMPLE AND EFFECTIVE PRUNING APPROACH FOR LARGE LANGUAGE MODELS

Mingjie Sun[1*] Zhuang Liu[2*] Anna Bair[1] J. Zico Kolter[1,3]
[1]Carnegie Mellon University [2]Meta AI Research [3]Bosch Center for AI



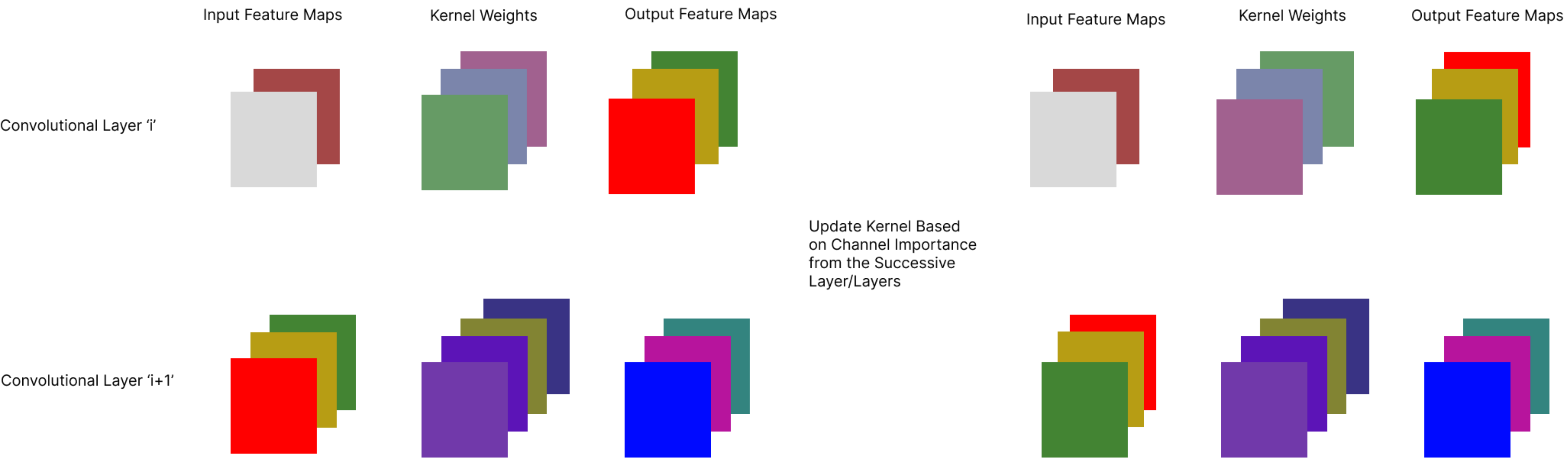
AWQ: Activation-aware Weight Quantization for LLM Compression and Acceleration

Ji Lin[1*] Jiaming Tang[1,2*] Haotian Tang[1] Shang Yang[1] Xingyu Dang[3] Chuang Gan[1] Song Han[1]
[1]MIT [2]SJTU [3]Tsinghua University
https://github.com/mit-han-lab/llm-awq



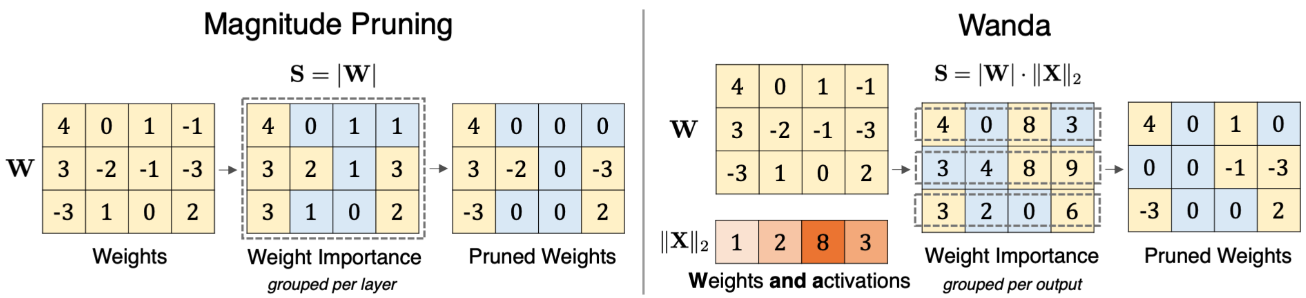
(a) RTN quantization (**PPL** 43.2) (b) Keep 1% salient weights in FP16 (**PPL** 13.0) (c) Scale the weights before quantization (**PPL** 13.0)

Make Kernels Aware of the Future Kernel Spaces



Code: https://github.com/satabios/sconce/blob/main/tutorials/Pruning.ipynb
Citations:
EIE:
Multi-scale channel importance sorting and spatial attention mechanism for retinal vessels segmentation
EACP: An effective automatic channel pruning for neural networks

Channel-Based Activation Aware Pruning



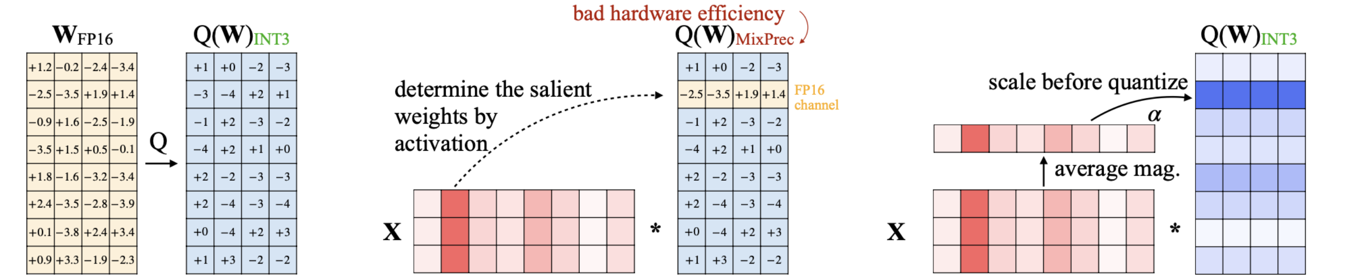
But Channel Wise!!

Method	Weight Update	Calibration Data	Pruning Metric $\mathbf{S}_{ij}$	Complexity
Magnitude	✗	✗	$\lvert\mathbf{W}_{ij}\rvert$	$O(1)$
SparseGPT	✓	✓	$\left[\lvert\mathbf{W}\rvert^2/\mathrm{diag}\left[(\mathbf{X}\mathbf{X}^T+\lambda\mathbf{I})^{-1}\right]\right]_{ij}$	$O(d^3_{\mathtt{hidden}})$
Wanda	✗	✓	$\lvert\mathbf{W}_{ij}\rvert \cdot \lVert\mathbf{X}_j\rVert_2$	$O(d^2_{\mathtt{hidden}})$

			LLaMA				LLaMA-2		
Method	Weight Update	Sparsity	7B	13B	30B	65B	7B	13B	70B
Dense	-	0%	59.99	62.59	65.38	66.97	59.71	63.03	67.08
Magnitude	✗	50%	46.94	47.61	53.83	62.74	51.14	52.85	60.93
SparseGPT	✓	50%	**54.94**	58.61	63.09	66.30	**56.24**	60.72	**67.28**
Wanda	✗	50%	54.21	**59.33**	**63.60**	**66.67**	**56.24**	**60.83**	67.03
Magnitude	✗	4:8	46.03	50.53	53.53	62.17	50.64	52.81	60.28
SparseGPT	✓	4:8	**52.80**	55.99	60.79	64.87	**53.80**	**59.15**	65.84
Wanda	✗	4:8	52.76	**56.09**	**61.00**	**64.97**	52.49	58.75	**66.06**
Magnitude	✗	2:4	44.73	48.00	53.16	61.28	45.58	49.89	59.95
SparseGPT	✓	2:4	**50.60**	**53.22**	58.91	62.57	**50.94**	54.86	63.89
Wanda	✗	2:4	48.53	52.30	**59.21**	**62.84**	48.75	**55.03**	**64.14**

- **Register hooks to fetch O/P Feature Maps**
- **Run through a Calibration Dataset**
- **Prune Channels(Kernel Weights) Based on Activations**

Activation Aware - QAT

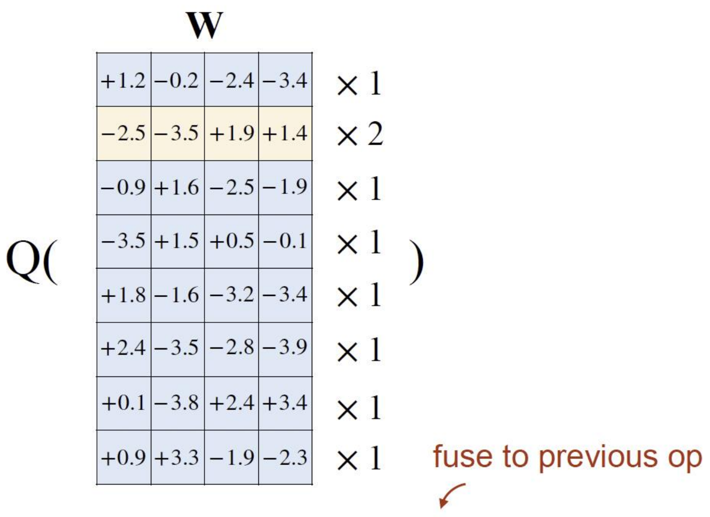


(a) RTN quantization (**PPL 43.2**) (b) Keep 1% salient weights in FP16 (**PPL 13.0**) (c) Scale the weights before quantization (**PPL 13.0**)

$$\mathbf{s}^* = \arg\min_{\mathbf{s}} \mathcal{L}(\mathbf{s}), \quad \mathcal{L}(\mathbf{s}) = \|Q(\mathbf{W}\cdot\mathbf{s})(\mathbf{s}^{-1}\cdot\mathbf{X}) - \mathbf{W}\mathbf{X}\| \tag{3}$$

But Channel Wise!!

- **Apply Scaling on Feature Maps based on Activation Saliency**
- **QAT**



$$\mathbf{WX} \longrightarrow Q(\mathbf{W}\cdot\mathbf{s})(\mathbf{s}^{-1}\cdot\mathbf{X})$$

PPL↓		Llama-2			LLaMA			
		7B	13B	70B	7B	13B	30B	65B
FP16	-	5.47	4.88	3.32	5.68	5.09	4.10	3.53
INT3 g128	RTN	6.66	5.52	3.98	7.01	5.88	4.88	4.24
	GPTQ	6.43	5.48	3.88	8.81	5.66	4.88	4.17
	GPTQ-R	6.42	5.41	3.86	6.53	5.64	4.74	4.21
	AWQ	**6.24**	**5.32**	**3.74**	**6.35**	**5.52**	**4.61**	**3.95**
INT4 g128	RTN	5.73	4.98	3.46	5.96	5.25	4.23	3.67
	GPTQ	5.69	4.98	3.42	6.22	5.23	4.24	3.66
	GPTQ-R	5.63	4.99	3.43	5.83	5.20	4.22	3.66
	AWQ	**5.60**	**4.97**	**3.41**	**5.78**	**5.19**	**4.21**	**3.62**

Complete Flow

- Sort Channels Based on Successive Channels
- Activation Aware Pruning (WANDA- like)
- Activation Aware Quantization (AWQ- like)

Possible Additions

- **Layer-Wise Neural Network Compression via Layer Fusion: https://proceedings.mlr.press/v157/o-neill21a/o-neill21a.pdf**
- **Layer-Selective Rank Reduction:** https://github.com/pratyushasharma/laser