SPP: Sparsity-Preserved Parameter-Efficient Fine-Tuning for Large Language Models

吴雨欣

Xudong Lu * 1 Aojun Zhou * 1 Yuhui Xu * 2 Renrui Zhang 1 3 Peng Gao 3 Hongsheng Li 1 4

1Multimedia Laboratory (MMLab), The Chinese University of Hong Kong
2Salesforce AI Research
3Shanghai Artificial Intelligence Laboratory
4CPII under InnoHK

Background

- LLMs' impressive capabilities、large number of parameters、fine-tuning->cumbersome、difficult to deploy

Background

- LLMs' impressive capabilities、large number of parameters、fine-tuning->cumbersome、difficult to deploy
- Many post training pruning methods have emerged, such as SparseGPT and Wanda, which have improved the sparsity rate of the model

Background

- LLMs' impressive capabilities、large number of parameters、fine-tuning->cumbersome、difficult to deploy
- Many post training pruning methods have emerged, such as SparseGPT and Wanda, which have improved the sparsity rate of the model
- Direct pruning -> information loss、in medium and high sparsity -> difficult to maintain performance

Background

- Restore the model performance through retraining

Background

- Restore the model performance through retraining
- Traditional retraining methods -> Full parameter backpropagation -> High costs

Background

- Restore the model performance through retraining
- Traditional retraining methods -> Full parameter backpropagation -> High costs
- Parameter Efficient Fine Tuning(PEFT)

Background

- Restore the model performance through retraining
- Traditional retraining methods -> Full parameter backpropagation -> High costs
- Parameter Efficient Fine Tuning(PEFT)
- Current PEFT -> Cause the sparse model to revert back to a dense model

Meaning

target:

- Sparse LLMs after pruning

characteristic:

Meaning

target:

- Sparse LLMs after pruning

characteristic:

- Fine tune the model during the retraining phase without changing its sparsity

Meaning

target:

- Sparse LLMs after pruning

characteristic:

- Fine tune the model during the retraining phase without changing its sparsity
- Restore the performance degradation caused by pruning without compromising the pruning effect

Meaning

target:

- Sparse LLMs after pruning

characteristic:

- Fine tune the model during the retraining phase without changing its sparsity
- Restore the performance degradation caused by pruning without compromising the pruning effect
- Modular approach, targets some layer of LLMs

Meaning

target:

- Sparse LLMs after pruning

characteristic:

- Fine tune the model during the retraining phase without changing its sparsity
- Restore the performance degradation caused by pruning without compromising the pruning effect
- Modular approach, targets some layer of LLMs
- Residual connection

Impression

Achieved good results in both structured and unstructured pruning, compared to the DSnoT method and LoRA.

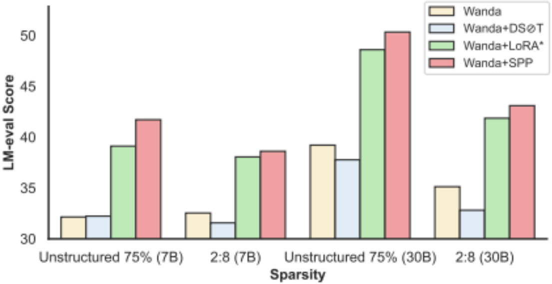

图 1: Experiment Results

The First Step

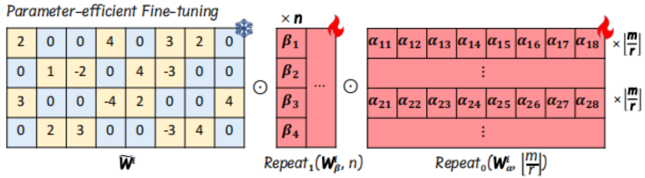

图 2: SPP

- Freeze the original pruned sparse linear matrix $\widetilde{\mathbf{W}}^{\mathbf{i}} \in \mathbb{R}^{m\times n}$, Insert two learnable matrices and adjust only these two matrices： $\mathbf{W}^{\mathbf{i}}_{\alpha} \in \mathbb{R}^{r\times n}$ 和 $\mathbf{W}^{\mathbf{i}}_{\beta} \in \mathbb{R}^{m\times 1}$

The First Step

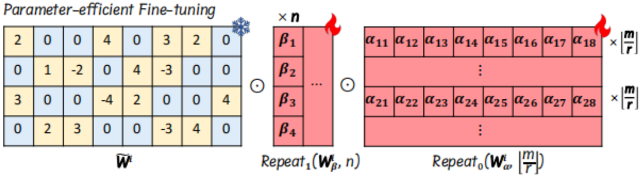

图 2: SPP

- Freeze the original pruned sparse linear matrix $\widetilde{\mathbf{W}}^{\mathbf{i}} \in \mathbb{R}^{m \times n}$, Insert two learnable matrices and adjust only these two matrices： $\mathbf{W}^{\mathbf{i}}_{\alpha} \in \mathbb{R}^{r \times n}$ 和 $\mathbf{W}^{\mathbf{i}}_{\beta} \in \mathbb{R}^{m \times 1}$
- Only modify these $m + rn$ additional parameters

The First Step

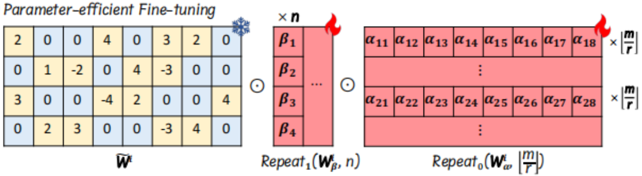

图 2: SPP

- Freeze the original pruned sparse linear matrix $\widetilde{\mathbf{W}}^{\mathbf{i}} \in \mathbb{R}^{m\times n}$, Insert two learnable matrices and adjust only these two matrices： $\mathbf{W}^{\mathbf{i}}_{\alpha} \in \mathbb{R}^{r\times n}$ 和 $\mathbf{W}^{\mathbf{i}}_{\beta} \in \mathbb{R}^{m\times 1}$
- Only modify these $m + rn$ additional parameters
- r is a hyperparameter that m can be divided by r

The Second Step

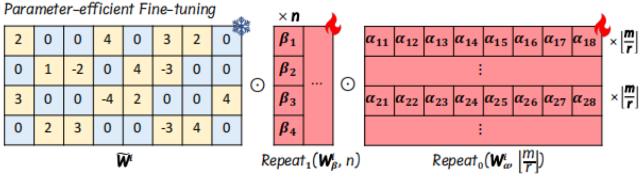

图 3: SPP

- Scale $\mathbf{W}^{\mathbf{i}}_{\alpha}$ and $\mathbf{W}^{\mathbf{i}}_{\beta}$ to the same size as $\widetilde{\mathbf{W}}^{\mathbf{i}}$:

$$\widetilde{W}^{i'} = \widetilde{W}^{i} \odot \text{Repeat}_0(W^i_\alpha, \left\lfloor \frac{m}{r} \right\rfloor) \odot \text{Repeat}_1(W^i_\beta, n)$$

The Second Step

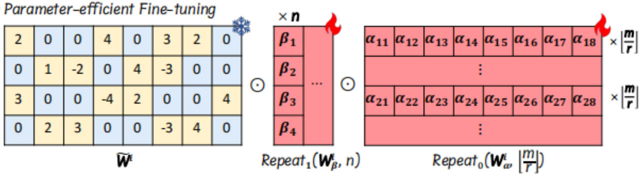

图 3: SPP

- Scale $\mathbf{W^i_\alpha}$ and $\mathbf{W^i_\beta}$ to the same size as $\widetilde{\mathbf{W}}^{\mathbf{i}}$:

$$\widetilde{W}^{i'} = \widetilde{W}^i \odot \text{Repeat}_0(W^i_\alpha, \left\lfloor \frac{m}{r} \right\rfloor) \odot \text{Repeat}_1(W^i_\beta, n)$$

- where $\odot$ represents element wise multiplication operation , Repeat_0 and Repeat_1 represents scaling operations on $\mathbf{W^i_\alpha}$ and $\mathbf{W^i_\beta}$, to ensure they match the size of $\widetilde{W}^i$

The Third Step

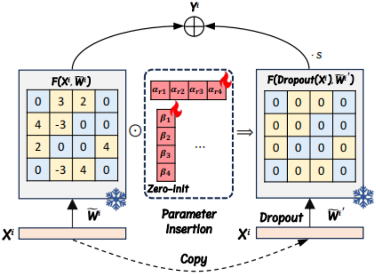

图 4: SPP framework

- The i Layer of the Model:

$$\mathbf{Y}^i = \mathbf{F}(\mathbf{X}^i, \widetilde{W}^i) + s \cdot \mathbf{F}(\mathrm{Dropout}(\mathbf{X}^i), \widetilde{W}^{i'})$$

The Third Step

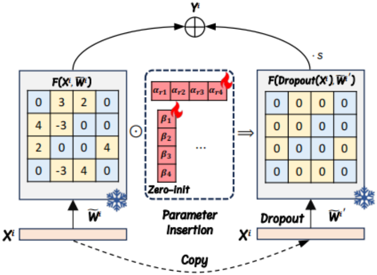

图 4: SPP framework

- The i Layer of the Model:

$$\mathbf{Y}^i = \mathbf{F}(\mathbf{X}^i, \widetilde{W}^i) + s \cdot \mathbf{F}(\text{Dropout}(\mathbf{X}^i), \widetilde{W}^{i\prime})$$

- Initialization: set $\mathbf{W}^{\mathbf{i}}_{\beta}$ to All Zeros, and randomly initialize $\mathbf{W}^{\mathbf{i}}_{\alpha}$

Why SPP does not destroy the sparse structure?

- Element wise multiplication (Hadamard product)(X*0=0)

Why SPP does not destroy the sparse structure?

- Element wise multiplication (Hadamard product)(X*0=0)
- Freeze the original pruning weights

Why SPP does not destroy the sparse structure?

- Element wise multiplication (Hadamard product)(X*0=0)
- Freeze the original pruning weights
- Maintain sparsity during the weight merging process(0+0=0)

1. Motivation
2. Method
3. **Experiments**
4. Conclusion

Experiment Settings

- Instruction fine-tuning dataset: Stanford Alpaca

Experiment Settings

- Instruction fine-tuning dataset: Stanford Alpaca
- Model: LLaMA 7B/13B/30B/65B, LLaMA-2 7B/13B/70B

Experiment Settings

- Instruction fine-tuning dataset: Stanford Alpaca
- Model: LLaMA 7B/13B/30B/65B, LLaMA-2 7B/13B/70B
- Hardware: 8 * NVIDIA A100-80GB GPU

Experiment Settings

- Instruction fine-tuning dataset: Stanford Alpaca
- Model: LLaMA 7B/13B/30B/65B, LLaMA-2 7B/13B/70B
- Hardware: 8 * NVIDIA A100-80GB GPU
- Evaluation metrics: LM Eval, Perplexity, MMLU

Experiment Details

- Add learnable parameters on linear layers such as'q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj, score'

Experiment Details

- Add learnable parameters on linear layers such as'q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj, score'
- Set r to 16

Experiment Details

- Add learnable parameters on linear layers such as'q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj, score'
- Set r to 16
- For the training of the 7B/13B/30B/65B/70B models, learning rates of 4e-3/2e-3/4e-3/5e-4/5e-4 were used, with batch sizes set to 8/4/16/8/8

Experiment Details

- Add learnable parameters on linear layers such as'q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj, score'
- Set r to 16
- For the training of the 7B/13B/30B/65B/70B models, learning rates of 4e-3/2e-3/4e-3/5e-4/5e-4 were used, with batch sizes set to 8/4/16/8/8
- Set a 0.03 warm-up ratio、the AdamW optimizer、a 0.001 weight decay

Experiment Details

- Add learnable parameters on linear layers such as'q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj, score'
- Set r to 16
- For the training of the 7B/13B/30B/65B/70B models, learning rates of 4e-3/2e-3/4e-3/5e-4/5e-4 were used, with batch sizes set to 8/4/16/8/8
- Set a 0.03 warm-up ratio、the AdamW optimizer、a 0.001 weight decay
- Fine-tune 7B/13B/30B models by 3 epochs, and 65B/70B models by 1 epoch

Comparison of the number of Trainable Parameter

	LLaMA				LLaMA-2		
	7B	13B	30B	65B	7B	13B	70B
Trainable Parameters	2.0×10^7	3.1×10^7	6.0×10^7	9.8×10^7	2.0×10^7	3.1×10^7	1.1×10^8
All Parameters	6.8×10^9	1.3×10^{10}	3.3×10^{10}	6.5×10^{10}	6.8×10^9	1.3×10^{10}	6.9×10^{10}
Per mille (‰)	**2.90**	**2.35**	**1.83**	**1.50**	**2.90**	**2.35**	**1.54**

- the number of Trainable Parameter: m*n -> m+r*n

50%Sparisty,LLaMA,Zero-shot

LLaMA	Method	Sparsity	BoolQ	RTE	HellaSwag	WinoGrande	ARC-e	ARC-c	OBQA	Average
7B	None	Dense	75.11	66.43	56.96	69.85	75.25	41.89	34.40	59.98
	SparseGPT	Unstructured 50%	73.36	57.76	51.44	68.03	70.45	36.35	28.40	55.11
	SparseGPT+**SPP**	Unstructured 50%	72.84	65.70	56.40	67.88	72.35	41.04	32.80	**58.43**
	SparseGPT	2:4	70.09	57.76	43.37	63.46	61.62	29.27	22.60	49.74
	SparseGPT+**SPP**	2:4	72.39	59.57	53.33	64.17	68.39	37.54	26.80	**54.60**
	Wanda	Unstructured 50%	71.01	55.23	51.90	66.22	69.36	36.95	28.60	54.18
	Wanda+**SPP**	Unstructured 50%	70.86	66.06	55.92	67.64	72.81	41.64	32.00	**58.13**
	Wanda	2:4	69.27	51.26	42.07	62.67	60.52	27.99	24.60	48.34
	Wanda+**SPP**	2:4	71.19	63.90	52.77	64.88	68.18	37.03	30.00	**55.42**
13B	None	Dense	77.98	70.40	59.92	72.61	77.36	46.50	33.20	62.57
	SparseGPT	Unstructured 50%	76.54	62.09	54.94	71.59	72.35	41.64	32.20	58.76
	SparseGPT+**SPP**	Unstructured 50%	79.20	64.62	59.27	70.32	74.83	46.59	34.60	**61.35**
	SparseGPT	2:4	70.80	56.68	48.09	69.22	66.88	36.26	26.20	53.45
	SparseGPT+**SPP**	2:4	77.65	63.54	56.55	69.69	71.21	40.96	32.60	**58.89**
	Wanda	Unstructured 50%	76.27	62.82	55.78	71.98	73.32	43.77	31.80	59.39
	Wanda+**SPP**	Unstructured 50%	78.29	66.43	58.88	70.32	75.59	46.93	34.40	**61.55**
	Wanda	2:4	70.21	53.79	46.78	68.82	65.74	33.70	26.20	52.18
	Wanda+**SPP**	2:4	75.99	58.12	56.07	68.90	70.37	40.53	32.40	**57.48**
30B	None	Dense	82.63	66.79	63.36	75.85	80.39	52.82	36.00	65.41
	SparseGPT	Unstructured 50%	82.63	58.84	59.20	73.48	78.79	49.15	33.20	62.18
	SparseGPT+**SPP**	Unstructured 50%	84.43	68.23	63.18	73.56	81.57	52.56	37.00	**65.79**
	SparseGPT	2:4	76.57	61.01	53.52	72.30	74.66	42.06	31.60	58.82
	SparseGPT+**SPP**	2:4	81.65	66.43	60.46	72.45	78.75	50.17	36.20	**63.73**
	Wanda	Unstructured 50%	81.93	64.98	60.95	73.64	79.38	50.17	34.80	63.69
	Wanda+**SPP**	Unstructured 50%	84.19	66.79	62.52	71.59	77.10	51.79	34.80	**64.11**
	Wanda	2:4	75.14	63.54	54.53	72.45	74.24	41.89	31.80	59.08
	Wanda+**SPP**	2:4	81.38	69.68	59.99	71.59	76.73	48.63	34.60	**63.23**
65B	None	Dense	84.55	69.68	65.40	77.35	52.82	81.00	38.00	66.97
	SparseGPT	Unstructured 50%	84.90	70.04	63.95	77.27	79.65	50.17	37.40	66.20
	SparseGPT+**SPP**	Unstructured 50%	84.95	70.04	64.25	77.19	79.85	50.94	37.80	**66.43**
	SparseGPT	2:4	84.55	69.31	57.95	76.95	78.00	45.39	31.20	63.34
	SparseGPT+**SPP**	2:4	84.25	68.23	58.40	76.87	78.10	45.99	31.40	63.32
	Wanda	Unstructured 50%	85.05	71.84	64.60	77.35	79.65	50.26	38.40	66.74
	Wanda+**SPP**	Unstructured 50%	85.25	71.84	65.30	77.19	79.95	51.11	38.60	**67.03**
	Wanda	2:4	83.40	61.01	58.55	75.22	76.60	45.56	33.20	61.93
	Wanda+**SPP**	2:4	83.30	61.37	61.85	76.16	78.60	47.70	36.20	**63.60**

75%Sparisty,LLaMA,Zero-shot

LLaMA	Method	Sparsity	LM-eval	PPL (↓)
7B	Wanda	Unstructured 75%	32.14	1285.24
	Wanda+DS∅T	Unstructured 75%	32.23	646.40
	Wanda+**SPP**	Unstructured 75%	**41.71**	**21.80**
	Wanda	2:8	32.53	3284.43
	Wanda+DS∅T	2:8	31.57	2742.98
	Wanda+**SPP**	2:8	**38.61**	**42.07**
30B	Wanda	Unstructured 75%	39.21	149.63
	Wanda+DS∅T	Unstructured 75%	37.77	184.51
	Wanda+**SPP**	Unstructured 75%	**50.33**	**10.89**
	Wanda	2:8	35.12	1057.58
	Wanda+DS∅T	2:8	32.81	903.17
	Wanda+**SPP**	2:8	**43.09**	**19.83**

- 7B LLaMA LM-Eval: 59.98

75%Sparisty,LLaMA,Zero-shot

LLaMA	Method	Sparsity	LM-eval	PPL (↓)
7B	Wanda	Unstructured 75%	32.14	1285.24
	Wanda+DS∅T	Unstructured 75%	32.23	646.40
	Wanda+**SPP**	Unstructured 75%	**41.71**	**21.80**
	Wanda	2:8	32.53	3284.43
	Wanda+DS∅T	2:8	31.57	2742.98
	Wanda+**SPP**	2:8	**38.61**	**42.07**
30B	Wanda	Unstructured 75%	39.21	149.63
	Wanda+DS∅T	Unstructured 75%	37.77	184.51
	Wanda+**SPP**	Unstructured 75%	**50.33**	**10.89**
	Wanda	2:8	35.12	1057.58
	Wanda+DS∅T	2:8	32.81	903.17
	Wanda+**SPP**	2:8	**43.09**	**19.83**

- 7B LLaMA LM-Eval: 59.98
- 30B LLaMA LM-Eval: 65.41

75%Sparisty,LLaMA,Zero-shot

LLaMA	Method	Sparsity	LM-eval	PPL (↓)
7B	Wanda	Unstructured 75%	32.14	1285.24
	Wanda+DS∅T	Unstructured 75%	32.23	646.40
	Wanda+**SPP**	Unstructured 75%	**41.71**	**21.80**
	Wanda	2:8	32.53	3284.43
	Wanda+DS∅T	2:8	31.57	2742.98
	Wanda+**SPP**	2:8	**38.61**	**42.07**
30B	Wanda	Unstructured 75%	39.21	149.63
	Wanda+DS∅T	Unstructured 75%	37.77	184.51
	Wanda+**SPP**	Unstructured 75%	**50.33**	**10.89**
	Wanda	2:8	35.12	1057.58
	Wanda+DS∅T	2:8	32.81	903.17
	Wanda+**SPP**	2:8	**43.09**	**19.83**

- 7B LLaMA LM-Eval: 59.98
- 30B LLaMA LM-Eval: 65.41
- The results of high sparsity rate is not ideal

50%Sparisty,MMLU,5-shot

		LLaMA				LLaMA-2		
Method	**Sparsity**	7B	13B	30B	65B	7B	13B	70B
None	Dense	35.64	47.63	58.58	63.78	46.56	55.30	69.56
SparseGPT	Unstructured 50%	32.19	40.44	52.62	59.37	36.41	47.47	65.57
SparseGPT+**SPP**	Unstructured 50%	30.77	43.91	54.73	59.38	39.78	48.31	65.60
SparseGPT	2:4	28.24	32.31	43.79	49.79	29.16	38.41	57.66
SparseGPT+**SPP**	2:4	27.81	37.55	49.01	49.50	33.28	45.63	57.85
Wanda	Unstructured 50%	31.50	39.43	52.84	58.75	34.20	47.78	64.45
Wanda+**SPP**	Unstructured 50%	31.74	43.34	53.89	59.02	38.08	48.97	64.39
Wanda	2:4	27.14	31.26	41.36	45.68	28.33	35.16	56.86
Wanda+**SPP**	2:4	28.56	35.73	46.19	47.67	30.47	42.79	57.98

- Certain gap on difficult problems

50%Sparisty,MMLU,5-shot

		LLaMA				LLaMA-2		
Method	**Sparsity**	7B	13B	30B	65B	7B	13B	70B
None	Dense	35.64	47.63	58.58	63.78	46.56	55.30	69.56
SparseGPT	Unstructured 50%	32.19	40.44	52.62	59.37	36.41	47.47	65.57
SparseGPT+**SPP**	Unstructured 50%	30.77	43.91	54.73	59.38	39.78	48.31	65.60
SparseGPT	2:4	28.24	32.31	43.79	49.79	29.16	38.41	57.66
SparseGPT+**SPP**	2:4	27.81	37.55	49.01	49.50	33.28	45.63	57.85
Wanda	Unstructured 50%	31.50	39.43	52.84	58.75	34.20	47.78	64.45
Wanda+**SPP**	Unstructured 50%	31.74	43.34	53.89	59.02	38.08	48.97	64.39
Wanda	2:4	27.14	31.26	41.36	45.68	28.33	35.16	56.86
Wanda+**SPP**	2:4	28.56	35.73	46.19	47.67	30.47	42.79	57.98

- Certain gap on difficult problems
- The author: due to the small size of the dataset

Ablation Study

Method	Sparsity	Zero-init	$\mathbf{W}_\beta$	r	LM-eval
Wanda+**SPP**	2:4	✓	✓	4	54.04
		✓	✓	8	54.87
		✓		16	54.52
			✓	16	53.52
		✓	✓	16	**55.42**
	Unstructured 50%	✓	✓	4	57.86
		✓	✓	8	56.39
		✓		16	57.81
			✓	16	57.59
		✓	✓	16	**58.13**
SparseGPT+**SPP**	2:4	✓	✓	4	**54.82**
		✓	✓	8	54.24
		✓		16	54.62
			✓	16	54.01
		✓	✓	16	54.60
	Unstructured 50%	✓	✓	4	57.58
		✓	✓	8	57.32
		✓		16	57.66
			✓	16	57.12
		✓	✓	16	**58.43**

- $\mathbf{W}_\beta$, r,the initialization of $\mathbf{W}_\beta$

Ablation Study

Method	Sparsity	Zero-init	$\mathbf{W}_\beta$	r	LM-eval
Wanda+**SPP**	2:4	✓	✓	4	54.04
		✓	✓	8	54.87
		✓		16	54.52
			✓	16	53.52
		✓	✓	16	**55.42**
	Unstructured 50%	✓	✓	4	57.86
		✓	✓	8	56.39
		✓		16	57.81
			✓	16	57.59
		✓	✓	16	**58.13**
SparseGPT+**SPP**	2:4	✓	✓	4	**54.82**
		✓	✓	8	54.24
		✓		16	54.62
			✓	16	54.01
		✓	✓	16	54.60
	Unstructured 50%	✓	✓	4	57.58
		✓	✓	8	57.32
		✓		16	57.66
			✓	16	57.12
		✓	✓	16	**58.43**

- $\mathbf{W}_\beta$, r,the initialization of $\mathbf{W}_\beta$
- r=32?

Comparison with LoRA*

LLaMA	Method	Sparsity	BoolQ	RTE	HellaSwag	WinoGrande	ARC-e	ARC-c	OBQA	Average
7B	Wanda+LoRA*	Unstructured 75%	62.39	53.07	31.81	52.64	38.22	21.08	14.60	39.11
	Wanda+**SPP**	Unstructured 75%	60.67	56.32	35.06	52.64	47.05	22.44	17.80	**41.71**
	Wanda+LoRA*	2:8	61.47	53.07	29.18	53.12	34.68	20.22	14.60	38.05
	Wanda+**SPP**	2:8	54.50	59.21	31.29	52.09	37.46	19.11	16.60	**38.61**
30B	Wanda+LoRA*	Unstructured 75%	65.08	55.60	44.35	62.27	60.69	29.18	23.00	48.60
	Wanda+**SPP**	Unstructured 75%	67.95	54.15	47.28	62.51	63.68	30.72	26.00	**50.33**
	Wanda+LoRA*	2:8	62.17	52.71	35.96	54.30	48.48	23.21	16.20	41.86
	Wanda+**SPP**	2:8	62.05	54.87	38.17	55.09	49.62	23.63	18.20	**43.09**

- LoRA's parameter number: m*r+r*n

Comparison with LoRA*

LLaMA	Method	Sparsity	BoolQ	RTE	HellaSwag	WinoGrande	ARC-e	ARC-c	OBQA	Average
7B	Wanda+LoRA*	Unstructured 75%	62.39	53.07	31.81	52.64	38.22	21.08	14.60	39.11
	Wanda+**SPP**	Unstructured 75%	60.67	56.32	35.06	52.64	47.05	22.44	17.80	**41.71**
	Wanda+LoRA*	2:8	61.47	53.07	29.18	53.12	34.68	20.22	14.60	38.05
	Wanda+**SPP**	2:8	54.50	59.21	31.29	52.09	37.46	19.11	16.60	**38.61**
30B	Wanda+LoRA*	Unstructured 75%	65.08	55.60	44.35	62.27	60.69	29.18	23.00	48.60
	Wanda+**SPP**	Unstructured 75%	67.95	54.15	47.28	62.51	63.68	30.72	26.00	**50.33**
	Wanda+LoRA*	2:8	62.17	52.71	35.96	54.30	48.48	23.21	16.20	41.86
	Wanda+**SPP**	2:8	62.05	54.87	38.17	55.09	49.62	23.63	18.20	**43.09**

- LoRA's parameter number: m*r+r*n
- In LoRA, r=8 is used to ensure that the parameter count is similar to that of SPP

Comparison with LoRA*

LLaMA	Method	Sparsity	BoolQ	RTE	HellaSwag	WinoGrande	ARC-e	ARC-c	OBQA	Average
7B	Wanda+LoRA*	Unstructured 75%	62.39	53.07	31.81	52.64	38.22	21.08	14.60	39.11
	Wanda+**SPP**	Unstructured 75%	60.67	56.32	35.06	52.64	47.05	22.44	17.80	**41.71**
	Wanda+LoRA*	2:8	61.47	53.07	29.18	53.12	34.68	20.22	14.60	38.05
	Wanda+**SPP**	2:8	54.50	59.21	31.29	52.09	37.46	19.11	16.60	**38.61**
30B	Wanda+LoRA*	Unstructured 75%	65.08	55.60	44.35	62.27	60.69	29.18	23.00	48.60
	Wanda+**SPP**	Unstructured 75%	67.95	54.15	47.28	62.51	63.68	30.72	26.00	**50.33**
	Wanda+LoRA*	2:8	62.17	52.71	35.96	54.30	48.48	23.21	16.20	41.86
	Wanda+**SPP**	2:8	62.05	54.87	38.17	55.09	49.62	23.63	18.20	**43.09**

- LoRA's parameter number: m*r+r*n
- In LoRA, r=8 is used to ensure that the parameter count is similar to that of SPP
- Is it more meaningful?

Conclusion

- a novel Sparsity-Preserved Parameter-efficient fine-tuning (SPP) method

Conclusion

- a novel Sparsity-Preserved Parameter-efficient fine-tuning (SPP) method
- to tackle the challenge of restoring the performance of LLMs after pruning

Conclusion

- a novel Sparsity-Preserved Parameter-efficient fine-tuning (SPP) method
- to tackle the challenge of restoring the performance of LLMs after pruning
- PEFT without changing its sparsity