Maintaining Structural Integrity in Parameter Efficient Fine-Tuning

Chongjie Si[1], Xuehui Wang[1], Xue Yang[2], Zhengqin Xu[1], Qingyun Li[2], Jifeng Dai[2], Yu Qiao[2], Xiaokang Yang[1], Wei Shen[1]
[1]Shanghai Jiao Tong University [2]Shanghai AI Laboratory

Observation

Task:

Parameter Efficient Fine-Tuning (PEFT), minimal trainable parameters, comparable or even superior performance.

Problem:

- Almost all of existing methods focus on linear weights, neglecting other higher dimensions spaces like convolution.
- Some methods adapt high-dimensional parameter spaces by compressing changes into a two-dimensional representation and applying low-rank matrix adaptations, yet this compression undermines the inherent structural integrity of the original space.

An Ideal Method Should Satisfy:

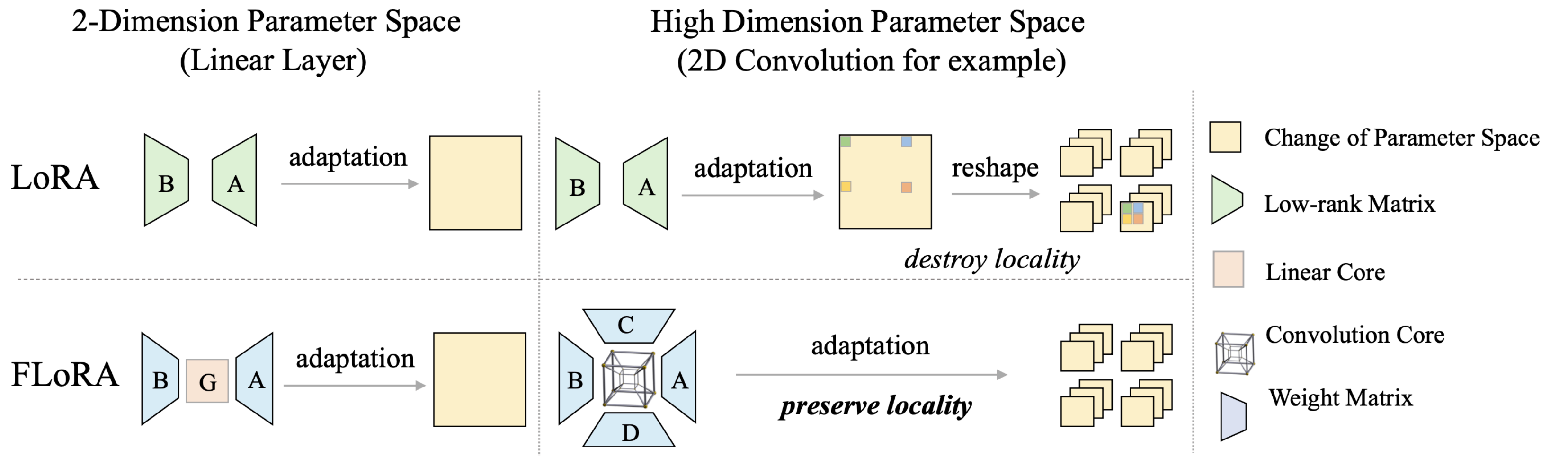
- Appropriate low-rank representation for the changes in various dimensional parameter spaces, without destructing the structural integrity of the original parameter spaces.
- Maintain a consistent formulation across various dimensional parameter spaces.

Contributions:

- We propose a novel PEFT method, FLoRA. To the best of our knowledge, it is the first time that a PEFT method has been designed for different dimensional parameter spaces, aiming to preserve their topological structure while seeking low-rank representations.
- Extensive experiments on different tasks, include computer vision, natural language processing and multi-modal tasks, demonstrates that FLoRA significantly surpasses other baselines, validating the effectiveness of FLoRA.

Method

Framework:



- LoRA employs low-rank matrix adaptation for each dimensional parameter space. However, for higher dimensional parameter space, the reshaping operation causes adjacent elements within the kernel to be separated in the matrix, disrupting the spatial locality inherent in the original convolutional space.
- FLoRA asserts that the alternations of each dimensional parameter space has a low-rank core space with the consistent topological structure. This enables FLoRA to effectively preserve the structural integrity of the original parameter space.

For N-dimension:

$$\mathcal{W}_0 \to \mathcal{W}_0 + s * \Delta\mathcal{W} = \mathcal{W}_0 + s * \mathcal{G} \times_1 \mathbf{A}^{(1)} \times_2 \mathbf{A}^{(2)} \times \cdots \times_N \mathbf{A}^{(N)}$$

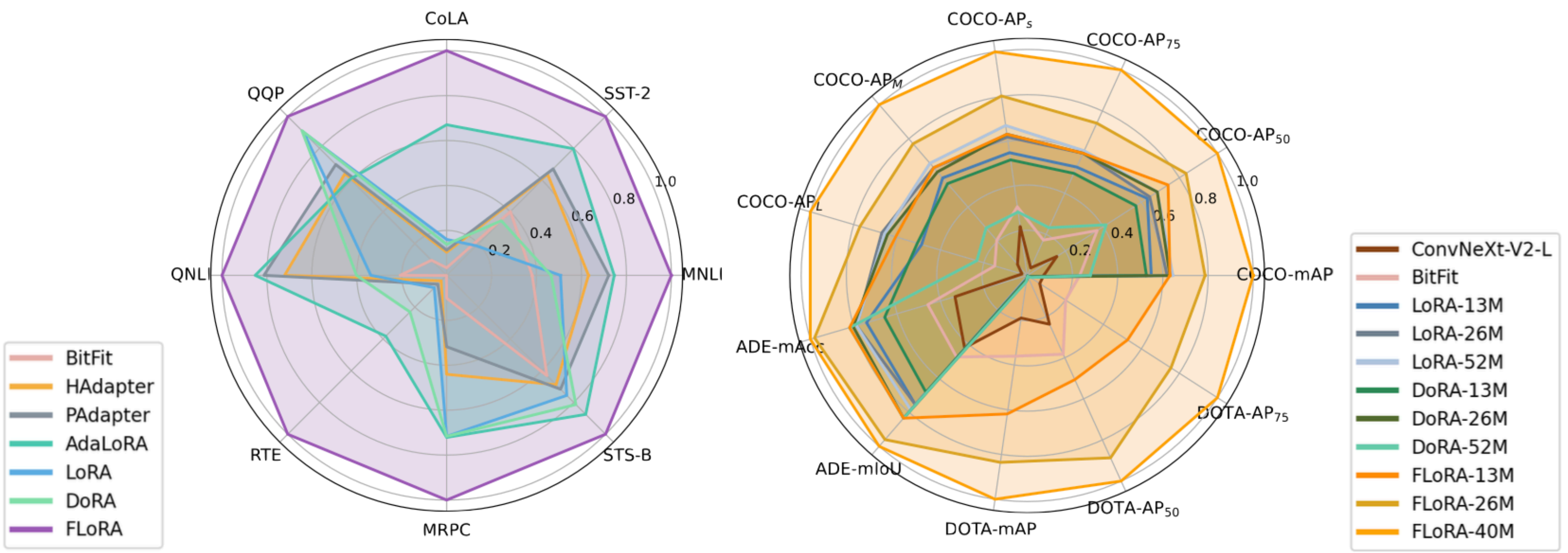
For 4-dimension:

$$\Delta\mathcal{W} = \mathcal{G} \times_1 \mathbf{A} \times_2 \mathbf{B} \times_3 \mathbf{C} \times_4 \mathbf{D}$$

For 2-dimension:

$$\Delta\mathbf{W} = \mathcal{G} \times_1 \mathbf{A} \times_2 \mathbf{B} = \mathbf{AGB}^{\mathsf{T}}$$

Experiment



Method	# Params	COCO						ADE20K		DOTA			All
		mAP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	mAcc	mIoU	mAP	AP_{50}	AP_{75}	Avg.
Base	-	37.3	63.3	39.7	27.8	41.0	46.2	59.6	48.5	31.4	61.8	27.3	44.0
Fully FT	196M	52.7	74.3	58.7	38.3	56.9	67.3	64.7	53.1	33.9	59.9	34.0	54.0
BitFit	0.2M	43.1	67.6	47.4	29.5	46.7	55.3	61.2	49.1	34.6	64.2	32.9	48.3
LoRA	12.94M	47.4	70.3	53.0	32.4	51.8	61.2	63.6	51.4	18.3	36.2	16.3	45.6
DoRA	13.07M	47.2	69.8	52.7	32.1	51.5	61.4	63.0	50.9	19.6	37.9	17.2	45.8
FLoRA	12.77M	**48.1**	**71.1**	**53.6**	**33.1**	**52.3**	**62.3**	**64.1**	**51.9**	**37.3**	**65.6**	**37.7**	**52.5**
LoRA	25.89M	48.0	70.4	53.6	33.0	52.3	62.8	63.9	51.4	20.0	38.3	18.3	46.5
DoRA	26.16M	48.1	70.7	53.6	33.1	52.1	62.6	64.0	51.9	21.1	39.7	19.1	46.9
FLoRA	25.65M	**49.2**	**71.7**	**54.7**	**34.3**	**53.3**	**63.5**	**65.0**	**52.6**	**38.8**	**68.4**	**39.5**	**53.7**
LoRA	51.78M	48.2	70.7	53.7	33.4	52.5	62.7	63.9	51.6	20.4	39.4	19.1	46.9
DoRA	51.95M	44.0	68.2	48.9	29.1	48.2	57.5	64.1	51.9	21.7	40.9	20.3	45.0
FLoRA	40.49M	**50.4**	**72.6**	**56.2**	**35.4**	**54.6**	**64.8**	**65.1**	**52.8**	**39.7**	**69.0**	**40.9**	**54.7**

a mouse *a dog* *a cat* *an elephant*