LEARNING IN THE FISHER SUBSPACE: A GUIDED INITIALIZATION FOR LORA FINE-TUNING

Presenter: Zhi-Quan Feng

National Cheng Kung University, CSIE

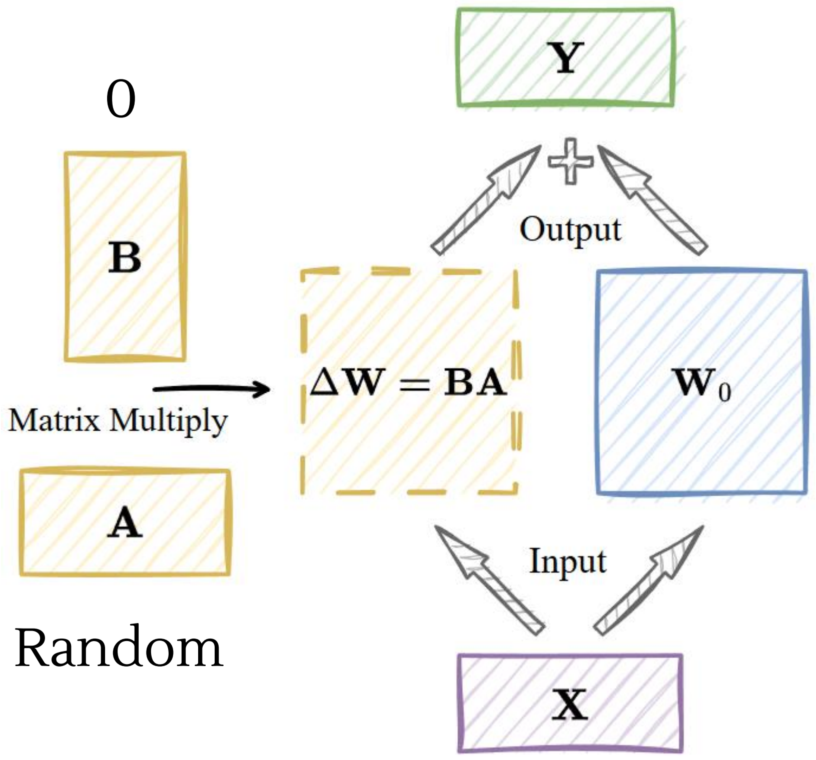
National Tsing Hua University, CS

LORA INITIALIZATION

LoRA Initialization determines the starting point of finetuning.

- The original initialization method is:
    - Matrix A: Random initialization
    - Matrix B: 0

Under random initialization, the LoRA weights contain no prior information and may converge to suboptimal local minima during fine-tuning.



LORA INITIALIZATION

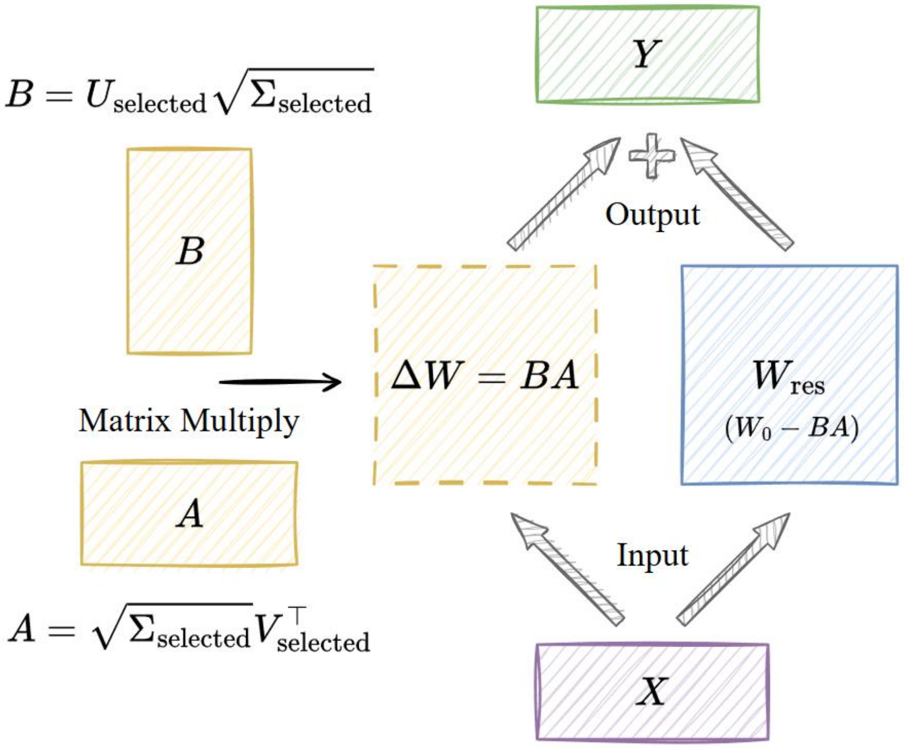
A common approach is to decompose a high-dimensional linear transformation into a set of one-dimensional directions (e.g., SVD, QR, etc.).

SVD:

$$W_0 = U_0\Sigma_0V_0^\top + U_1\Sigma_1V_1^\top + \ldots + U_d\Sigma_dV_d^\top$$

$$\Delta W = U_{\text{selected}}\Sigma_{\text{selected}}V_{\text{selected}}^\top \qquad W_{\text{res}} = W_0 - \Delta W$$

$$B = U_{\text{selected}}\sqrt{\Sigma_{\text{selected}}} \qquad A = \sqrt{\Sigma_{\text{selected}}}V_{\text{selected}}^\top$$



A subset of these directions is used to initialize the LoRA modules.

METHOD

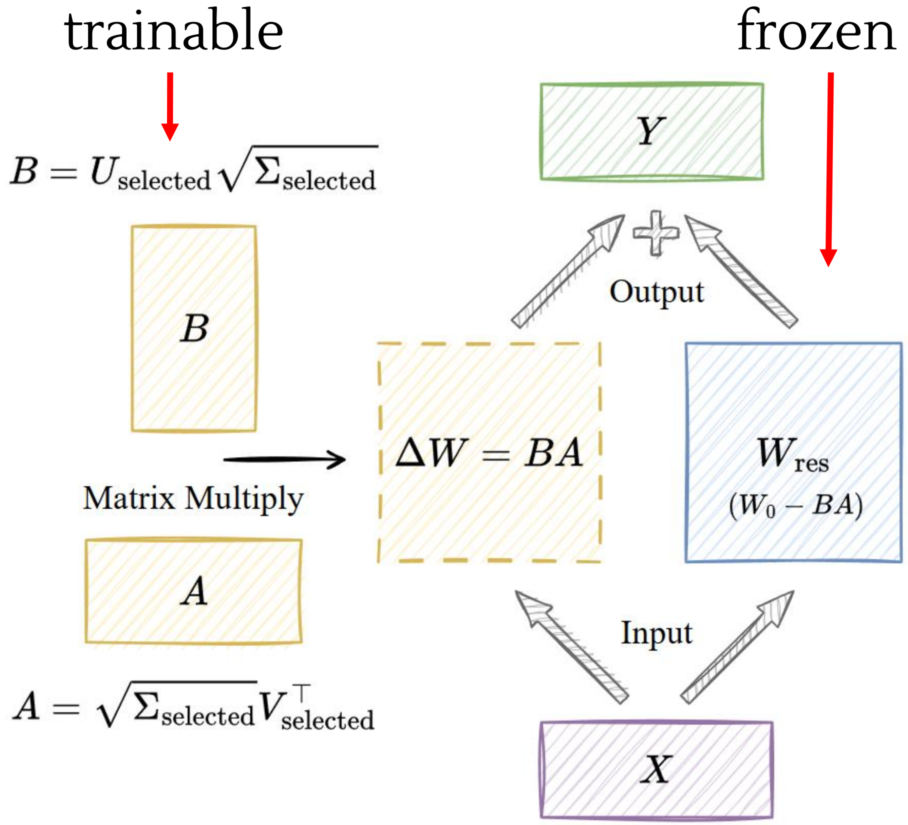
PROBLEM DEFINITION

LoRA initialization devides the pretrained weight into two parts, one is used to initialize LoRA while the other is frozen.

- As a result, our goal is to :
  - preserve the most task-relevent information in the residual weight

    and
  - task-irrelevent information in the LoRA module.

We use "sensitivity" to mesure "task-relevance"



METHOD

FILET

We employ the Fisher Information to characterize the sensitivity associated with each direction.

$$S_W = \mathbb{E}\left[\operatorname{vec}\left(\nabla_W \mathcal{L}\right) \operatorname{vec}\left(\nabla_W \mathcal{L}\right)^\top\right]$$

We then define the Fisher Energy and seek to initialize LoRA with directions that minimize this quantity.

$$\begin{aligned}\mathcal{E}(Z) :=& \operatorname{vec}(Z)^\top S_W \operatorname{vec}(Z) \\ =& \operatorname{vec}(uv^\top)^\top S_W \operatorname{vec}(uv^\top)\end{aligned}$$

We propose "Fisher-guided Initialization for LoRA Finetuning (FILet)", which use Fisher Energy to select candidate directions.

METHOD

FILET

Fisher Energy:

$$\mathcal{E}(Z) = \operatorname{vec}(uv^\top)^\top S_W \operatorname{vec}(uv^\top)$$

The Fisher information matrix can be factorized using K-FAC (ICML 2015):

$$S_W = \mathbb{E}\Big[\operatorname{vec}(\nabla_W \mathcal{L})\ \operatorname{vec}(\nabla_W \mathcal{L})^\top\Big]$$

$$\nabla_W \mathcal{L} = (\nabla_Y \mathcal{L}) X^\top.$$

The Fisher Energy could be factorized as:

$$\begin{aligned}\mathcal{E}(Z) &= \operatorname{vec}(Z)^\top (S_X \otimes S_Y) \operatorname{vec}(Z) \\ &= \left(v^\top S_X v\right)\left(u^\top S_Y u\right).\end{aligned}$$

FILET

Obtain the candidate directions:

$$\hat{V} = \text{normalize}(W_0^\top W_0),$$
$$\hat{U} = \text{normalize}(W_0 \hat{V}).$$

We compute the Fisher Energy of all candidate directions $-\Sigma_{\text{fisher}}$;

Select directions with lowest Fisher Energies:

$$\mathcal{I} = \text{TopK}(-\Sigma_{\text{fisher}}, r),$$
$$\hat{V}_{\text{sel}} = \hat{V}[:, \mathcal{I}],$$
$$\hat{U}_{\text{sel}} = \hat{U}[:, \mathcal{I}],$$
$$\Sigma_{\text{sel}} = \Sigma_{\text{fisher}}[\mathcal{I}, \mathcal{I}].$$

Then initialize LoRA modules

$$A = \sqrt{\Sigma_{\text{sel}}}\, \hat{V}_{\text{sel}}^\top,$$
$$B = \hat{U}_{\text{sel}} \sqrt{\Sigma_{\text{sel}}},$$

The pre-trained weight is decomposed as: $W_0 = W_{\text{res}} + \alpha(BA),$

EXPERIMENTS

REASONING EXPERIMENTS

LoRA rank = 32

Model	PEFT	**BoolQ**	**PIQA**	**SIQA**	**HellaS.**	**WinoG.**	**ARC-e**	**ARC-c**	**OBQA**	**Avg.**
ChatGPT†	-	73.1	85.4	68.5	78.5	66.1	89.8	79.9	74.8	77.0
Llama2-7b	LoRA†	69.8	79.9	79.5	83.6	82.6	79.8	64.7	81.0	77.6
	DoRA†	71.8	83.7	76.0	89.1	82.6	83.7	68.2	82.4	79.7
	PiSSA†	67.6	78.1	78.4	76.6	78.0	75.8	60.2	75.6	73.8
	MiLoRA†	67.6	83.8	80.1	88.2	82.0	82.8	68.8	80.6	79.2
	NEAT†	71.9	84.0	80.4	88.9	84.6	86.5	**71.6**	**83.0**	81.4
	KaSA	$74.3_{0.71}$	$85.2_{0.66}$	$81.9_{0.32}$	$93.4_{0.23}$	$85.3_{0.59}$	$86.3_{0.62}$	$67.4_{0.87}$	$80.2_{0.77}$	$81.7_{0.22}$
	LoRA-Dash	$73.9_{0.58}$	$85.0_{0.79}$	$82.0_{0.38}$	$\mathbf{94.2}_{0.19}$	$84.3_{0.90}$	$85.7_{0.49}$	$68.6_{0.92}$	$79.4_{0.79}$	$81.7_{0.38}$
	FILet	$\mathbf{75.0}_{0.40}$	$\mathbf{86.6}_{0.50}$	$\mathbf{82.8}_{0.50}$	$\mathbf{94.2}_{0.22}$	$\mathbf{85.7}_{0.53}$	$\mathbf{87.1}_{0.58}$	$70.5_{0.45}$	$82.6_{0.52}$	$\mathbf{83.1}_{0.26}$
Llama3-8b	LoRA†	70.8	85.2	79.9	91.7	84.3	84.2	71.2	79.0	80.8
	DoRA†	71.8	83.7	76.0	89.1	82.6	83.7	68.2	82.4	79.7
	PiSSA†	67.1	81.1	77.2	83.6	78.9	77.7	63.2	74.6	75.4
	MiLoRA†	68.8	86.7	77.2	92.9	85.6	86.8	75.5	81.8	81.9
	NEAT†	72.1	87.0	80.9	94.3	86.7	91.4	78.9	84.8	84.5
	KaSA	$74.3_{0.71}$	$87.8_{0.70}$	$82.2_{0.55}$	$95.2_{0.38}$	$87.5_{0.54}$	$91.6_{0.57}$	$\mathbf{80.3}_{0.63}$	$87.0_{0.62}$	$85.7_{0.23}$
	LoRA-Dash	$74.7_{0.67}$	$87.4_{0.53}$	$82.0_{0.42}$	$95.8_{0.30}$	$87.2_{0.55}$	$92.0_{0.43}$	$80.0_{0.66}$	$86.0_{1.06}$	$85.6_{0.29}$
	FILet	$\mathbf{75.4}_{0.55}$	$\mathbf{90.6}_{0.57}$	$\mathbf{82.8}_{0.52}$	$\mathbf{96.1}_{0.28}$	$\mathbf{89.0}_{0.45}$	$\mathbf{92.2}_{0.58}$	$79.9_{0.39}$	$\mathbf{87.4}_{0.61}$	$\mathbf{86.7}_{0.13}$

NLG EXPERIMENTS

LoRA rank = 8

Model	PEFT	GSM8k	MBPP	E2E					MT-bench
				BLEU	NIST	METEOR	Rouge-L	CIDEr	
Llama2-7b	LoRA	32.9	15.7	39.8	5.50	65.6	54.4	12.9	4.26
	KaSA	33.2	15.6	39.3	5.43	65.4	54.5	12.5	4.12
	LoRA-Dash	34.1	15.8	40.1	5.55	65.8	54.6	12.8	4.29
	FILet	**34.5**	**16.6**	**40.7**	**5.57**	**66.5**	**54.9**	**13.3**	**4.39**
Llama3-8b	LoRA	55.6	22.2	39.3	5.47	65.6	54.9	12.4	4.91
	KaSA	53.7	23.4	40.4	5.48	66.0	55.0	12.9	4.75
	LoRA-Dash	55.8	22.8	40.5	**5.60**	**66.3**	54.8	**13.2**	4.95
	FILet	**58.8**	**27.1**	**40.7**	5.56	**66.3**	**55.1**	13.1	**5.01**
Gemma-7b	LoRA	54.1	29.4	39.7	5.49	65.6	54.0	12.6	4.32
	KaSA	54.5	29.9	38.9	5.41	**65.8**	54.4	12.8	4.37
	LoRA-Dash	58.7	29.5	39.3	5.46	65.0	54.5	12.9	4.43
	FILet	**63.5**	**30.8**	**39.9**	**5.48**	65.4	**54.7**	**13.0**	**4.56**
Qwen2.5-7b	LoRA	77.6	39.3	42.2	5.64	66.6	55.7	13.5	4.99
	KaSA	76.4	40.2	41.8	5.63	66.7	55.6	13.3	5.05
	LoRA-Dash	75.9	39.2	42.3	5.65	**67.4**	56.1	13.5	5.03
	FILet	**78.8**	**42.5**	**42.4**	**5.68**	66.9	**56.3**	**13.7**	**5.11**

CONCLUSION

We presented FILet, a data-aware LoRA initialization method that leverages Fisher Energy to identify and scale directions

FILet captures the local parameter sensitivity induced by the target data distribution and achieved better performance .