Low-Rank and Sparsity Are All You Need: Exploring Robust Hierarchical Latent Subspaces for Transferable Adversarial Attack

Shuangshuang Pu[1], Wen Yang[1], Min Li[1], Guodong Liu[2], Chris Ding[3], Di Ming[1 4] ✉
[1] School of Computer Science and Engineering, Chongqing University of Technology, Chongqing, China
[2] Eli Lilly & Company, Indiana, USA
[3] School of Data Science, The Chinese University of Hong Kong, Shenzhen, China
[4] Chongqing Key Laboratory of Trusted Human-Machine Interaction and Security/Safety for Intelligent Connected Vehicles, Chongqing, China

Introduction

➢ Background

❖ Adversarial Example: created by introducing subtle perturbations into a benign sample.

❖ Low-Rank Adversarial Attacks: imposing structural constraints on intermediate representations to guide perturbations toward dominant semantic subspaces.

$$\max_{x_{adv}} \mathcal{L}(\beta \cdot \mathcal{F}(x_{adv}) + (1-\beta)\cdot \widetilde{\mathcal{F}}(x_{adv}), y),$$
$$\text{s.t.} \quad \|x_{adv} - x\|_\infty \le \epsilon.$$

where $\beta \in [0,1]$ is a hyperparameter that controls the relative contribution of the dense and low-rank components.

➢ Contribution

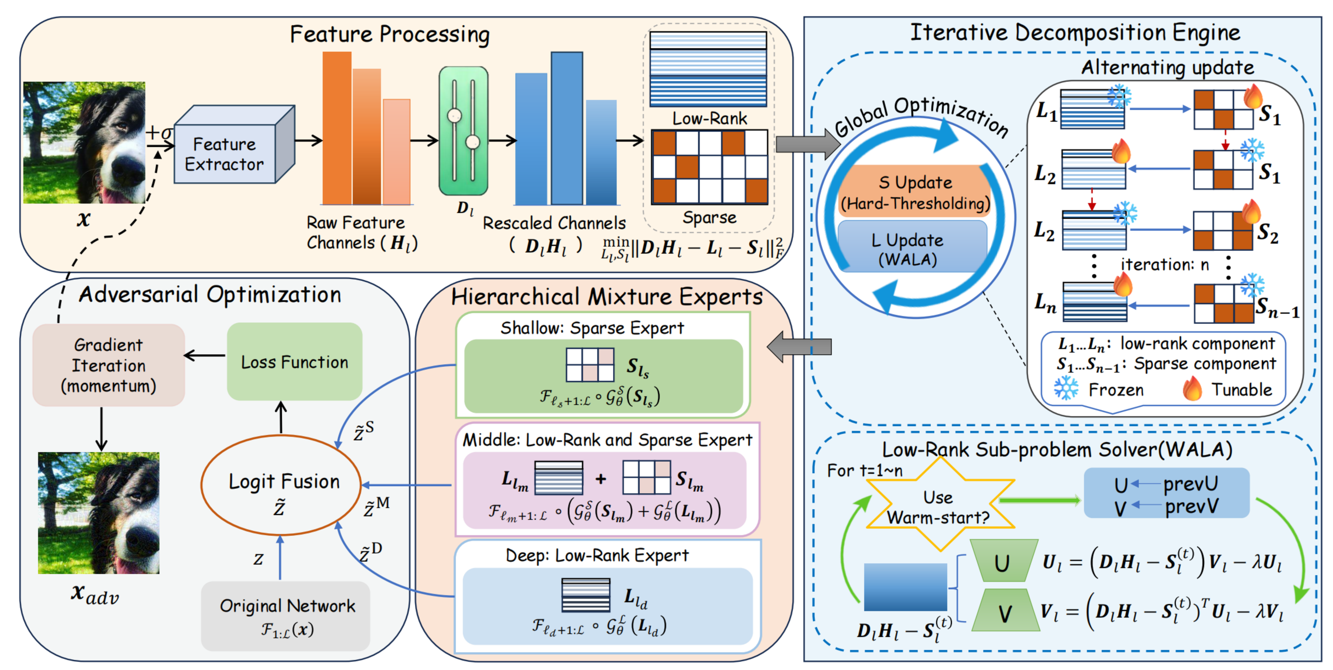
❖ We propose a low-rank and sparse feature-space attack that explicitly decomposes intermediate representations into global semantic components and localized discriminative perturbations, enabling structured intervention beyond conventional dense attacks.

❖ We introduce an efficient Warm-started Alternating Low-rank Approximation (WALA) solver to extract low-rank features, achieving a balance between attack effectiveness and computational efficiency.

❖ We design a hierarchical mixture of robust experts, applying sparse, balanced, and low-rank constraints to shallow, middle, and deep layers respectively, which significantly improves cross-layer consistency and adversarial transferability.

Methodology

➢ The Overall Framework of Our LRS-Attack Method



➢ Learning Robust Latent Subspace with Low-Rank and Sparse Decomposition

❖ At first, we formulate the proposed rescaled low-rank and sparse decomposition as follows:

$$\min_{L_l, S_l} \|D_l H_l - L_l - S_l\|_F^2,$$
$$\text{s.t.} \quad \text{rank}(L_l) \le r, \ \|S_l\|_0 \le k,$$

➢ Alternating Updates with Hard-Thresholding Operator and Regularized Low-Rank Approximation

❖ (i) Sparse component Update:

$$S_l^{(t)} = \arg\min_{\|S_l\|_0 \le k} \|(D_l H_l - L_l^{(t-1)}) - S_l\|_F^2.$$

❖ Hard-thresholding:

$$S_l^{(t)} = M \odot (D_l H_l - L_l^{(t-1)}),$$

❖ (ii) Low-rank Update:

$$L_l^{(t)} = \arg\min_{\text{rank}(L_l)\le r} \|(D_l H_l - S_l^{(t)}) - L_l\|_F^2.$$

❖ (ii) Reformulate the low-rank component subproblem as a regularized low-rank factorization problem:

$$\min_{U_l, V_l} \|(D_l H_l - S_l^{(t)}) - U_l V_l^\top\|_F^2 + \lambda\|U_l\|_F^2 + \lambda\|V_l\|_F^2,$$

➢ Warm-started Alternating Low-rank Approximation

❖ The alternative update:

$$U_l = (D_l H_l - S_l^{(t)})V_l - \lambda U_l,$$
$$V_l = (D_l H_l - S_l^{(t)})^\top U_l - \lambda V_l,$$

❖ the low-rank term is efficiently approximated by :

$$L_l^{(t)} = U_l^* \Sigma_l (V_l^*)^\top$$

❖ the final rescaled de-composition is obtained :

$$\widetilde{H}_l = \mathcal{G}(H_l) = D_l^{-1}(L_l^* + S_l^*),$$

➢ Hierarchical Mixture of Robust Low-Rank and Sparse Experts.

❖ Sparse Expert in Shallow Layers：

structured representation: $\widetilde{H}_l = \mathcal{G}_\theta^S(H_l) = D_l^{-1} S_l$

The final logits: $\tilde{z}^S = \widetilde{\mathcal{F}}^S(x; l, \theta) = \mathcal{F}_{l+1:L} \circ \mathcal{G}_\theta^S \circ \mathcal{F}_{1:l}(x)$.

❖ Low-Rank and Sparse Expert in Middle Layers：

structured representation: $\widetilde{H}_l = \mathcal{G}_\theta^S(H_l) + \mathcal{G}_\theta^L(H_l) = D_l^{-1} S_l + D_l^{-1} L_l$.

the final logits: $\tilde{z}^M = \widetilde{\mathcal{F}}^M(x; l, \theta) = \mathcal{F}_{l+1:L} \circ (\mathcal{G}_\theta^S + \mathcal{G}_\theta^L) \circ \mathcal{F}_{1:l}(x)$.

❖ Low-Rank Expert in Deep Layers：

structured representation: $\widetilde{H}_l = \mathcal{G}_\theta^L(H_l) = D_l^{-1} L_l$.

the final logits: $\tilde{z}^D = \widetilde{\mathcal{F}}^D(x; l, \theta) = \mathcal{F}_{l+1:L} \circ \mathcal{G}_\theta^L \circ \mathcal{F}_{1:l}(x)$.

➢ Hierarchical Mixture of Robust Experts

❖ Logits Fusion:

$$\tilde{z} = w_O \cdot z + w_S \cdot \tilde{z}_S + w_M \cdot \tilde{z}_M + w_D \cdot \tilde{z}_D$$

❖ The hierarchical low-rank and sparse decomposition-based attack is formulated as:

$$\max_{x_{adv}} \mathcal{L}(\mathcal{F}(x_{adv}) + \widetilde{\mathcal{F}}^S(x_{adv}) + \widetilde{\mathcal{F}}^M(x_{adv}) + \widetilde{\mathcal{F}}^D(x_{adv}), y)$$
$$\text{s.t.} \quad \|x_{adv} - x\|_\infty \le \epsilon.$$

Experiment

➢ Comparison with State-of-the-Art Methods

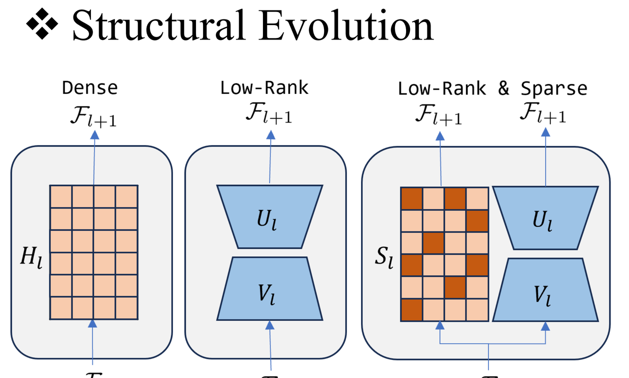
❖ Surrogate Model: Inc-v3, Vgg-19, Densenet-121, ViT-B, DeiT-B, CaiT-S.

Model	Attack	Method	Avg^{CNN}	$\text{Avg}^{\text{Def-CNN}}$	Avg^{ViT}	$\text{Avg}^{\text{Def-ViT}}$	Avg^{Def}
Inc-v3	MI-FGSM	W/O	51.7	20.6	21.0	34.7	32.0
		SVD	55.5	22.3	21.3	34.9	33.5
		LRS	**61.8**	**23.2**	**24.3**	**35.3**	**36.1**
	DI-FGSM	W/O	42.1	14.1	14.7	32.0	25.7
		SVD	46.3	14.4	15.0	31.9	26.9
		LRS	**49.6**	**15.3**	**16.4**	**32.5**	**28.4**
	TI-FGSM	W/O	29.8	11.7	12.6	31.9	21.5
		SVD	31.4	11.8	12.8	**32.3**	22.1
		LRS	**35.2**	**13.3**	**14.1**	**32.3**	**23.7**
	PI-FGSM	W/O	63.4	42.4	20.7	45.4	43.0
		SVD	65.5	44.8	21.0	46.3	44.4
		LRS	**67.7**	**45.6**	**21.6**	**47.6**	**45.6**
	SI-NI-FGSM	W/O	74.4	36.5	33.8	**37.5**	45.5
		SVD	77.7	37.2	35.0	37.1	46.8
		LRS	**80.4**	**39.2**	**37.0**	37.3	**48.5**
	VMI-FGSM	W/O	68.0	36.0	37.5	36.1	44.4
		SVD	74.0	41.6	39.8	36.3	47.9
		LRS	**76.6**	**41.7**	**41.1**	**36.8**	**49.0**
	GI-FGSM	W/O	61.2	21.8	24.1	35.9	35.8
		SVD	67.5	22.5	25.9	35.9	38.0
		LRS	**73.5**	**24.8**	**28.3**	**36.4**	**40.8**
Vgg-19	MI-FGSM	W/O	73.1	37.8	37.5	46.7	48.8
		SVD	74.1	37.7	35.2	46.8	48.5
		LRS	**77.4**	**40.5**	**40.0**	**47.0**	**51.2**
	DI-FGSM	W/O	64.6	25.1	25.9	40.4	39.0
		SVD	63.3	24.6	24.4	40.5	38.2
		LRS	**68.1**	**27.8**	**27.7**	**40.6**	**41.0**
	TI-FGSM	W/O	56.5	26.7	23.2	40.5	36.7
		SVD	54.0	25.3	21.6	**40.8**	35.4
		LRS	**61.4**	**28.8**	**25.8**	40.6	**39.1**
	PI-FGSM	W/O	69.3	47.4	23.2	52.8	48.2
		SVD	69.1	47.1	23.1	53.5	48.2
		LRS	**71.1**	**48.6**	**24.0**	**54.2**	**49.4**
	SI-NI-FGSM	W/O	86.5	54.7	47.9	48.1	59.3
		SVD	87.2	53.7	47.3	**48.3**	59.1
		LRS	**89.6**	**58.3**	**51.0**	**48.3**	**61.8**
	VMI-FGSM	W/O	83.6	52.5	50.6	48.0	58.7
		SVD	84.7	52.6	49.5	47.6	58.6
		LRS	**85.4**	**54.3**	**52.5**	**48.1**	**60.1**
	GI-FGSM	W/O	81.2	41.4	39.1	44.3	51.5
		SVD	82.6	41.7	39.1	44.5	51.9
		LRS	**83.1**	**42.9**	**40.9**	**44.6**	**52.9**
Densenet-121	MI-FGSM	W/O	76.3	47.5	40.7	38.8	50.8
		SVD	77.0	45.7	37.8	39.3	49.9
		LRS	**83.5**	**54.4**	**46.1**	**39.5**	**55.9**
	DI-FGSM	W/O	73.0	40.4	33.4	**41.3**	47.0
		SVD	72.7	38.3	30.2	41.1	45.6
		LRS	**79.1**	**43.9**	**36.5**	**41.3**	**50.2**
	TI-FGSM	W/O	56.6	33.6	26.0	41.1	39.3
		SVD	54.7	32.2	23.2	40.7	37.7
		LRS	**63.7**	**39.1**	**29.1**	**41.4**	**43.3**
	PI-FGSM	W/O	74.1	61.8	26.4	55.7	54.5
		SVD	74.8	61.7	25.2	55.2	54.2
		LRS	**76.9**	**65.1**	**27.3**	**56.0**	**56.3**
	SI-NI-FGSM	W/O	88.4	68.3	56.4	49.2	65.6
		SVD	88.8	67.1	53.1	49.0	64.5
		LRS	**91.5**	**73.1**	**60.4**	**49.8**	**68.7**
	VMI-FGSM	W/O	88.0	69.1	61.4	49.4	67.0
		SVD	89.4	70.5	60.6	49.2	67.4
		LRS	**91.4**	**73.2**	**64.7**	**50.2**	**69.9**
	GI-FGSM	W/O	86.6	58.5	49.7	45.6	60.1
		SVD	87.7	57.1	47.6	45.7	59.5
		LRS	**91.5**	**65.3**	**55.6**	**46.0**	**64.6**

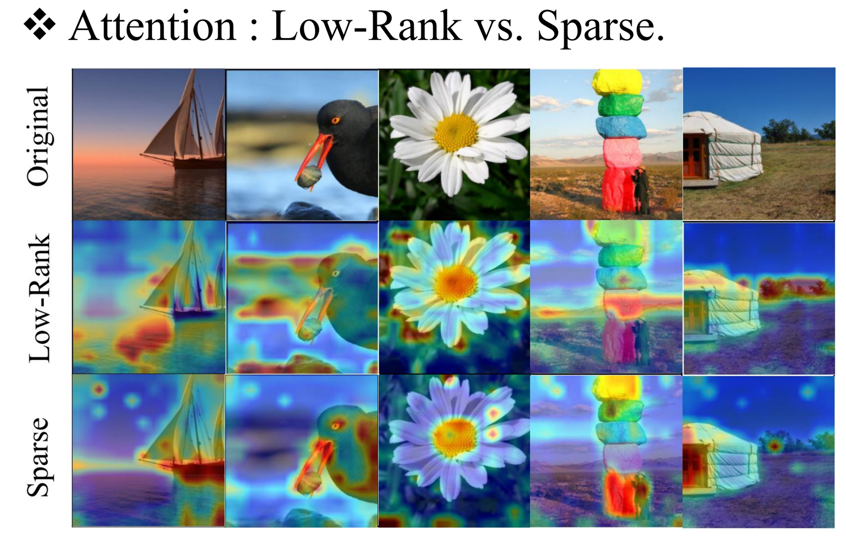
Model	Attack	Method	Avg^{CNN}	$\text{Avg}^{\text{Def-CNN}}$	Avg^{ViT}	$\text{Avg}^{\text{Def-ViT}}$	Avg^{Def}
ViT-B	MIM	W/O	51.1	34.4	64.2	48.0	49.4
		SVD	53.6	35.8	68.2	48.5	51.5
		LRS	**57.7**	**38.6**	**69.8**	**49.4**	**53.9**
	PNA	W/O	52.7	33.5	69.0	45.8	50.2
		SVD	53.6	35.0	71.2	46.2	51.5
		LRS	**59.8**	**38.3**	**71.8**	**47.2**	**54.3**
	TGR	W/O	59.5	40.5	68.9	51.1	55.0
		SVD	60.1	40.4	69.3	51.7	55.4
		LRS	**61.1**	**41.3**	**69.7**	**52.7**	**56.2**
	ATT	W/O	65.0	**46.2**	77.0	51.8	60.0
		SVD	65.8	45.6	76.5	52.3	60.1
		LRS	**66.3**	45.4	**77.2**	**53.1**	**60.5**
DeiT-B	MIM	W/O	60.9	40.7	78.0	48.8	57.1
		SVD	62.8	43.0	81.2	49.4	59.1
		LRS	**68.1**	**48.2**	**86.6**	**50.7**	**63.4**
	PNA	W/O	63.9	42.8	98.7	98.5	76.0
		SVD	66.8	45.8	98.9	**98.8**	77.6
		LRS	**71.2**	**50.1**	**99.4**	**98.8**	**79.9**
	TGR	W/O	78.8	58.8	92.7	54.7	71.2
		SVD	79.6	60.4	93.1	55.0	72.1
		LRS	**82.2**	**63.1**	**94.1**	**56.7**	**74.0**
	ATT	W/O	81.4	63.7	95.2	54.6	73.7
		SVD	81.1	62.4	95.2	54.1	73.2
		LRS	**83.4**	**65.2**	**95.8**	**56.2**	**75.2**
CaiT-S	MIM	W/O	68.5	47.0	83.3	49.4	62.1
		SVD	70.6	49.3	85.9	50.4	64.0
		LRS	**74.6**	**56.0**	**88.4**	**52.4**	**67.8**
	PNA	W/O	65.1	43.1	81.4	45.5	58.8
		SVD	68.3	47.7	85.5	46.2	61.9
		LRS	**73.7**	**53.9**	**87.4**	**49.1**	**66.0**
	TGR	W/O	80.4	60.9	90.5	54.1	71.5
		SVD	80.3	62.2	90.8	55.4	72.2
		LRS	**83.1**	**65.7**	**91.7**	**58.0**	**74.6**
	ATT	W/O	81.3	66.2	91.3	53.8	73.2
		SVD	82.0	66.1	91.8	55.1	73.8
		LRS	**83.0**	**66.8**	**92.0**	**56.9**	**74.7**

➢ Visualization & Analysis

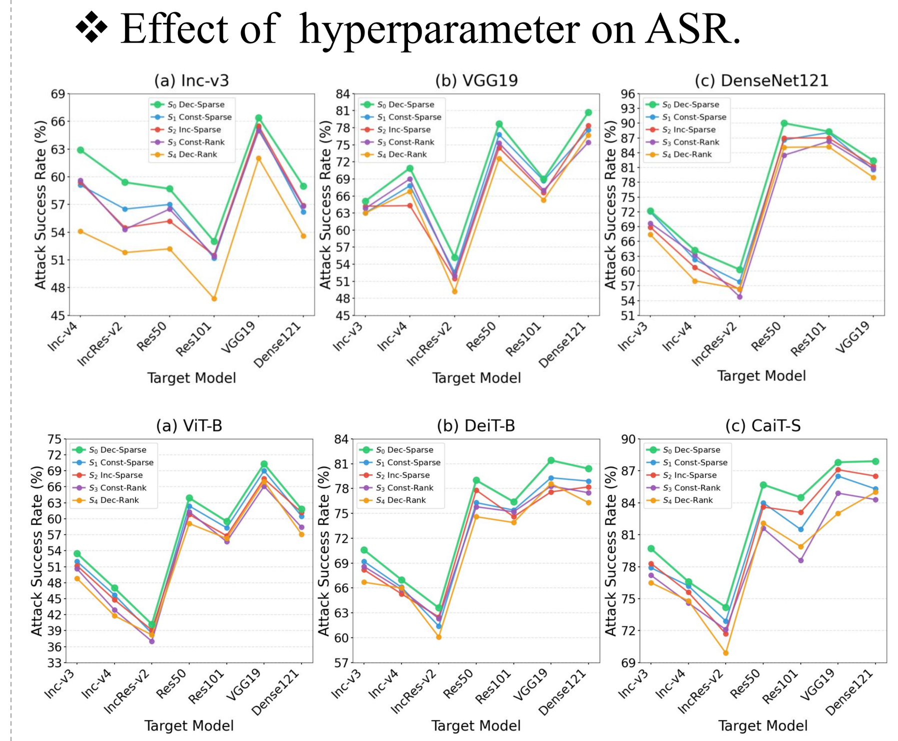
❖ Structural Evolution



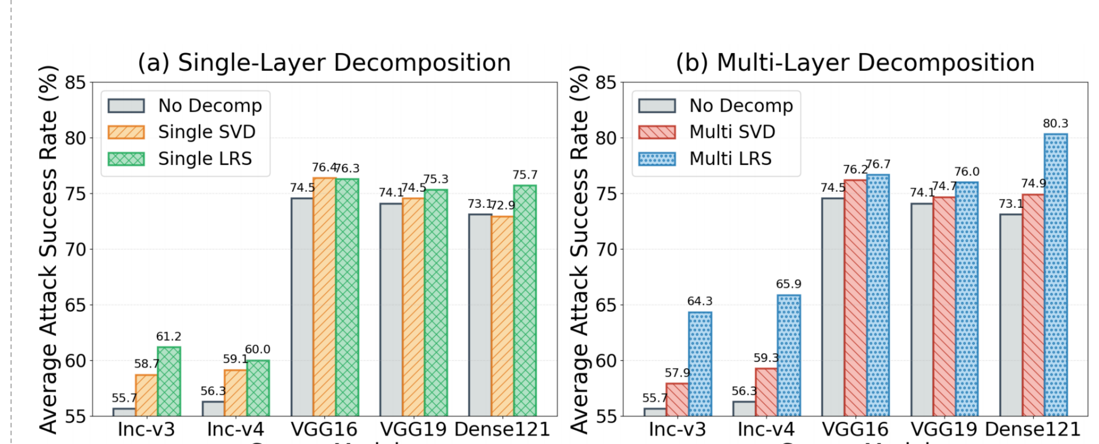
❖ Attention : Low-Rank vs. Sparse.



❖ Effect of hyperparameter on ASR.



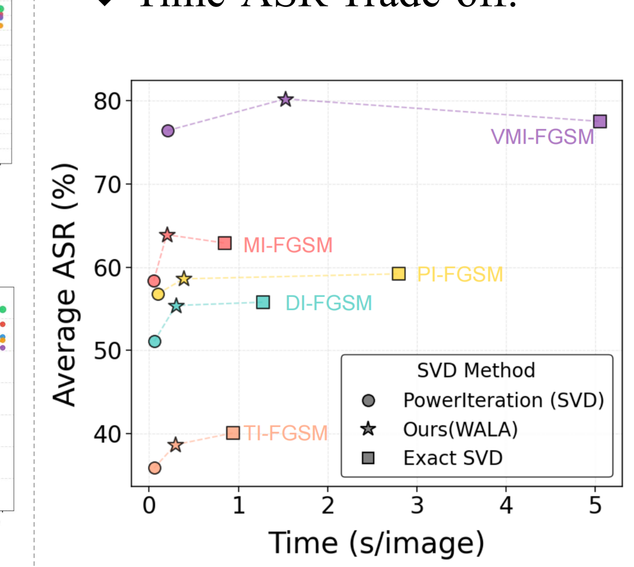
❖ Effect of Layer Configuration on ASR.



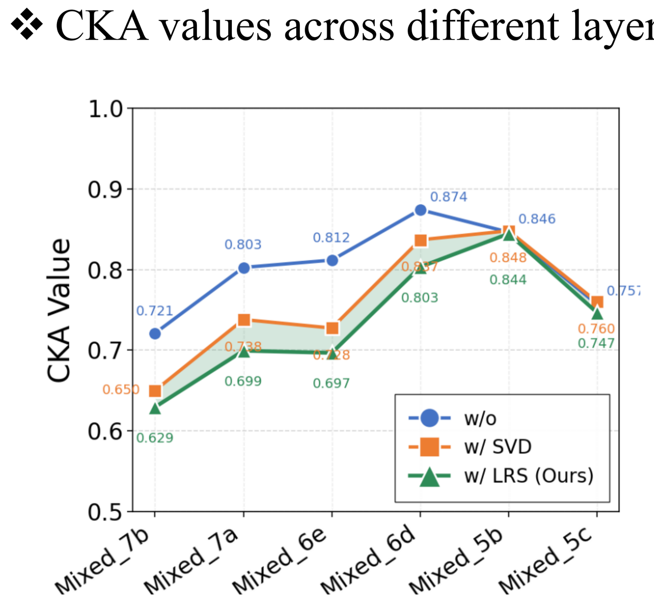
❖ Exploration of Expert Assignment Strategies.

Source	Strategy	Expert			Inc-v3	Vgg19	Dense121	CaiT-S	DeiT-B	ViT-B	Avg
		Shallow	Middle	Deep							
Inc-v3	w/o	-	-	-	**100.0**	59.9	47.4	15.0	17.1	25.0	44.1
	All-Spa.	Spa.	Spa.	Spa.	**100.0**	65.5	57.5	18.9	19.0	25.9	47.8
	All-Lrk.	Lrk.	Lrk.	Lrk.	99.9	64.8	56.2	19.1	19.6	26.2	47.6
	L-M-S	Lrk.	Mix.	Spa.	**100.0**	61.4	51.1	17.1	18.0	24.5	45.4
	L-S-M	Lrk.	Spa.	Mix.	**100.0**	64.2	57.0	17.8	20.3	26.5	47.6
	S-L-M	Spa.	Lrk.	Mix.	**100.0**	65.1	58.4	18.1	19.2	26.1	47.8
	S-M-L	**Spa.**	**Mix.**	**Lrk.**	**100.0**	**66.5**	**60.4**	**19.6**	**21.0**	**26.5**	**49.0**
ViT-B	w/o	-	-	-	39.0	57.3	44.8	44.4	49.0	99.9	55.7
	All-Spa.	Spa.	Spa.	Spa.	52.7	69.6	59.7	75.2	79.1	**100.0**	72.7
	All-Lrk.	Lrk.	Lrk.	Lrk.	52.6	69.3	60.9	76.0	78.6	**100.0**	72.9
	L-M-S	Lrk.	Mix.	Spa.	53.0	70.0	60.2	76.2	77.4	**100.0**	72.8
	L-S-M	Lrk.	Spa.	Mix.	52.2	68.0	60.3	76.6	78.1	**100.0**	72.5
	S-L-M	Spa.	Lrk.	Mix.	52.7	68.3	60.5	76.5	76.3	**100.0**	72.4
	S-M-L	**Spa.**	**Mix.**	**Lrk.**	**53.5**	**70.3**	**61.8**	**77.3**	**79.9**	**100.0**	**73.8**

❖ Time-ASR Trade-off.



❖ CKA values across different layers.



❖ CKA across different models.

Transfer	Layer	Clean	SVD	LRS
Inc-v3 → Inc-v4	Pool	0.7986	0.4233	0.4813
	FC	0.5626	0.1891	0.2086
Inc-v3 → IncRes-v2	Pool	0.7612	0.3927	0.4495
	FC	0.5854	0.1873	0.2135
Inc-v3 → Vgg-16	Pool	0.5065	0.4512	0.4887
	FC	0.2944	0.1341	0.1532
Inc-v3 → Vgg-19	Pool	0.5079	0.4373	0.4767
	FC	0.2986	0.1310	0.1541

Links & Contacts

GitHub Project

WeChat