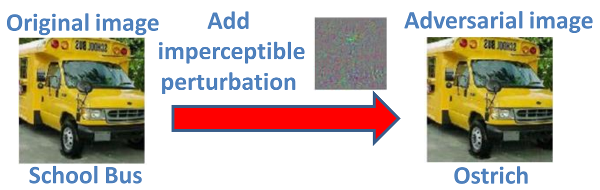
RBFormer: Improve Adversarial Robustness of Transformer by Robust Bias

Hao Cheng[1], Jinhao Duan[2], Hui Li[3], Jiahang Cao[1], Ping Wang[4], Lyutianyang Zhang[5], Jize Zhang[6], Kaidi Xu[2], Renjing Xu[1]

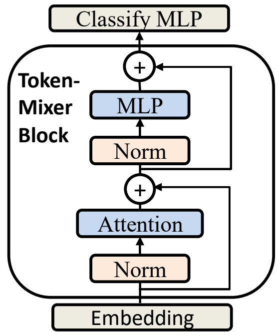
[1]The Hong Kong University of Science and Technology (Guangzhou), [2]Drexel University, [3]Samsung R&D Institute China Xi'an, [4]Xi'an Jiaotong University, [5]University of Washington, [6]The Hong Kong University of Science and Technology.

Background: Adversarial Robustness:



1. Adversarial examples are obtained by adding imperceptible perturbations to a correctly classified input image;
2. Adversarial training uses the min-max optimization to improve the adversarial robustness of corresponding models

Background: Transformer Structure



Vision Transformer (ViT) and Vision MLP (VMLP) Components:

- **Essential components(Robust Bias — Adding convolutional operation):**

1) Embedding, 2) Token-Mixer (TM) Block; 3) Classifying MLP (CMLP)

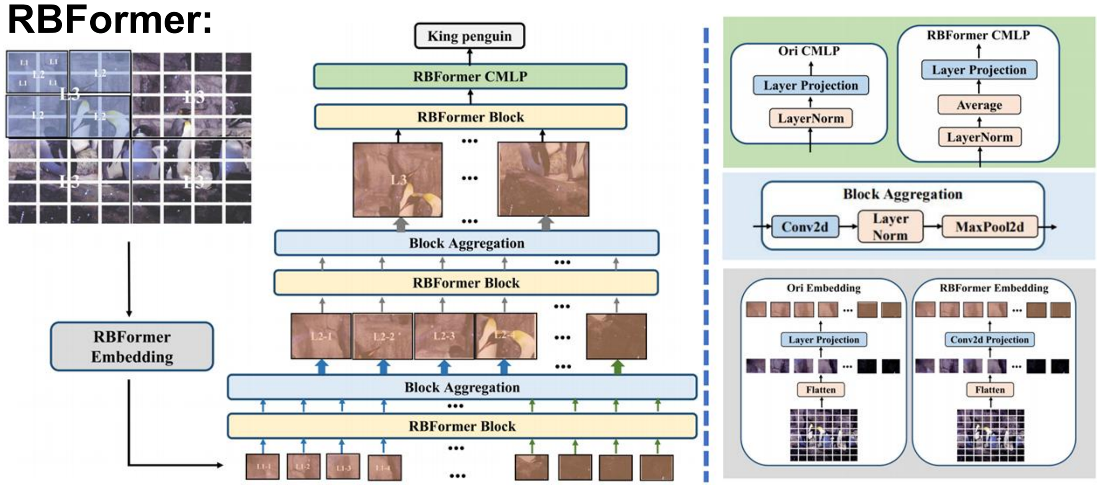
- **Multi-hierarchy layer Stacking (Robust Bias– Trying various strategies):**

1) OriViT, 2) CNN-based, 3) Swin, 4) Image Pyramid (ImagePy) Structures

- **Training facilitation techniques:**

1) Norm, 2) Skip-connection.

RBFormer:



CMLP removes CLS token and adopting an average pooling layer to process patches.

Block Aggregation is adopted to aggregate different token vectors hierarchically. Adding convolutional operation through replacing the Linear Layer with Conv Layer

Embedding removes the CLS token and directly averages all tokens. The original embedding dimension modification process is directly changed to convolution operation.

ImagePy Structure first splits and then aggregates non-overlap image patches in a hierarchy way.

TM Block keeps the original Multi-head Self Attention (MSA) and MLP structures in ViT and VMLP. The Norm and Skip-connection are both retained. About introducing convolution, all Linear Layers are replaced into Conv Layer according to the modification of other components and ensure dimensional transformation.

Experiments:

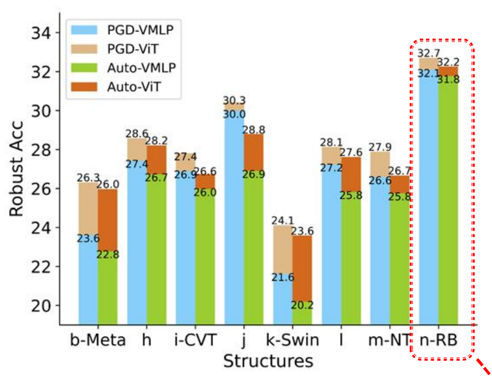


Figure 1: ImageNet-1k robust accuracy for various structures under PDG and Auto Attack

ViT/VMLP						Clean Accuracy	PGD (8/255)	Auto-Attack (8/255)	Lipschitz Constant
Components Combine	Embedding	TM	CMLP	Norm	Stacking Structure				
(a)	Ori	Ori	Ori	None	oriViT	79.88/71.06	52.66/45.56	51.12/44.37	159.2/163.2
(b)-Ori	-	-	-	LN	-	79.93/66.38	52.70/44.06	51.45/43.10	157.7/164.7
(c)	-	-	CONV	None	-	82.81/78.83	54.79/54.24	53.83/53.69	151.3/152.3
(d)	-	-	-	LN	-	81.66/77.58	54.69/51.00	53.85/50.88	152.7/157.5
(e)	CONV	-	Ori	None	-	82.77/77.86	55.85/53.22	54.98/52.89	151.4/155.8
(f)	-	-	-	LN	-	80.50/75.92	54.40/50.89	53.69/48.99	153.1/162.3
(g)	-	-	CONV	None	-	80.57/79.25	55.63/53.81	54.23/52.45	146.3/148.5
(h)	-	-	-	LN	-	**82.35/81.42**	**56.41/56.89**	**56.12/57.02**	**140.3/141.3**
(i)-CVT	PCONV	CONV	Ori	-	CNN-based	79.62/77.62	53.15/52.12	52.11/50.21	143.2/146.9
(j)	-	-	CONV	-	-	80.98/79.64	57.67/56.83	57.34/57.06	136.9/138.1
(k)-Swin	Ori	WBM+SWBM	Ori	-	Swin-based	76.39/75.23	48.93/46.34	47.64/45.21	152.0/154.2
(l)	PCONV	-	CONV	-	-	80.08/78.34	52.10/50.92	50.48/50.22	146.3/145.2
(m)-NT	Ori	Ori	Ori	-	ImagePy	76.22/75.94	52.45/51.82	51.28/49.14	158.2/167.9
(n)-RB	PCONV	CONV	CONV	-	-	**83.74/82.19**	**60.91/59.88**	**59.69/59.22**	**89.1/98.7**

Table 1: The robust performance in CIFAR-10 of ViT/VMLP accuracy (%). 1) Structure (a)-(h):adding convolution operation to different components; 2) Structure (h)-(n): adopting various multi-hierarchy layer stacking strategies. (b) is the original ViT/VMLP (Ori), (i) is corresponding to CVT, (k) Swin, (m) is the NesT . (n) is our final RBViT/RBMLP (RB).

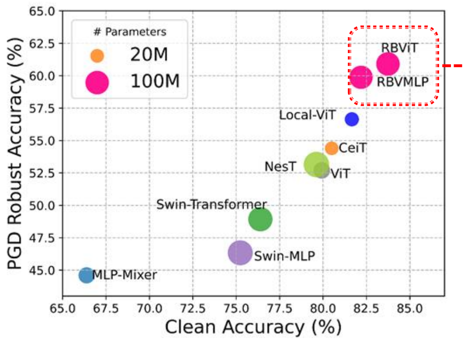


Figure 2:Comparison Results of RBFormer (RBViT/RBVMLP) with current SOTA in clean/robust accuracy and model size.

Tab. 1 and Fig.1: According to the results of adding convolution operation (structures (a) to (h)) and multi-hierarchy layer stacking strategies (structures (h) to (n)) under CIFAR-10 and ImageNet-1k, we conclude: 1) In each kind of layer stacking strategy, adding convolution operation to any components could generate a positive effect on improving robustness; 2) Not any layer stacking strategy could successfully introduce robust bias to boost robustness, like Swin; 3) ImagePy structure would be the best choice and make the final structure attain the best adversarial robustness.

Fig. 2: Under comparing RBViT/RBVMLP (n) with some popularly used ViT/Mixer-MLP (b), CeiT and Local-ViT CVT/CVT-based VMLP(i), Swin ViT/VMLP (k), NesT/NesT-based VMLP. Our RBFormer could attain the best robust and clean accuracy with a relatively small model size.