Boosting Adversarial Transferability with Spatial Adversarial Alignment

Zhaoyu Chen[1,*], Haijing Guo[2,*], Kaixun Jiang[1], Jiyuan Fu[2], Xinyu Zhou[2], Dingkang Yang[1], Hao Tang[3], Bo Li[4], Wenqiang Zhang[1,2,†]
[1]College of Intelligent Robotics and Advanced Manufacturing, Fudan University
[2]College of Computer Science and Artificial Intelligence, Fudan University
[3] School of Computer Science, Peking University [4]vivo Mobile Communication Co., Ltd

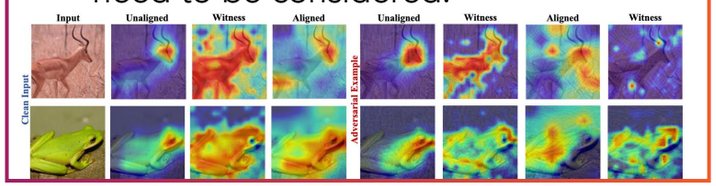
Motivation

Cross-model transferability has been extensively studied for CNNs. However, few works explore adversarial transferability on ViT and the performance of existing work extending CNN to ViT is poor due to significant structural differences.

We argue that **unique structural features** are critical to cross-architecture adversarial transferability. If we can obtain a surrogate model whose features are similar to those of models with different architectures, then the resulting adversarial perturbation can be transferable across different models.

However, directly applying this idea to black-box attacks may lead to the degradation of cross-architecture transferability. The main reasons are:

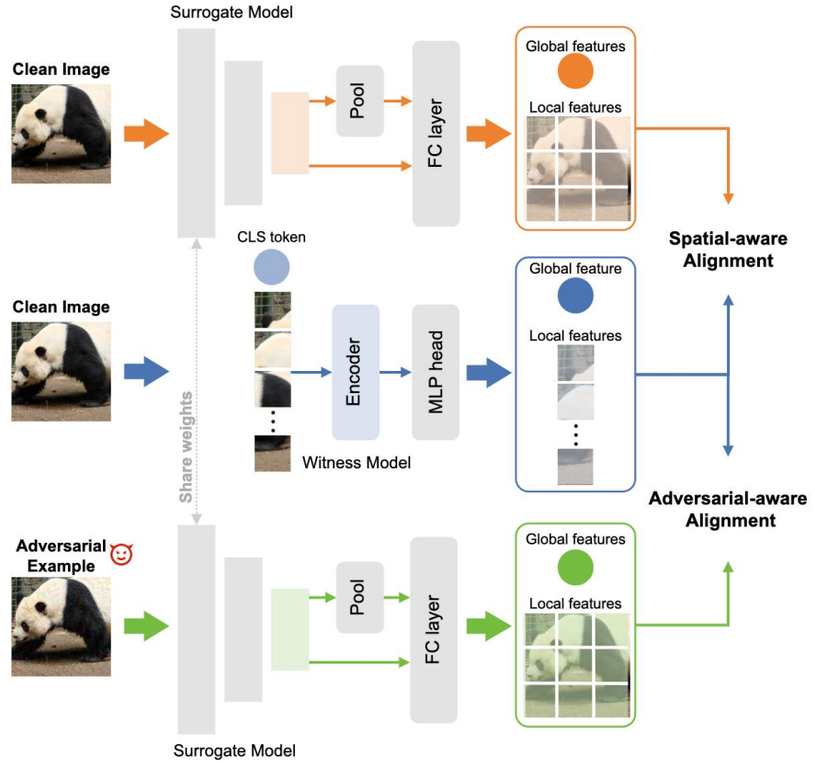
1. Features are not aligned in space. It is hard to directly constrain the similarity of features only by the final logits.
2. Features are not aligned from the perspective of adversarial features. The features of adversarial examples also have similarities across models and need to be considered.



Method

Spatial Adversarial Alignment consists of two key parts:

1. In **spatial-aware alignment**, in addition to aligning on the final global features, we also focus on the features of local regions.
2. In **adversarial-aware alignment**, we introduce a self-adversarial strategy, which constructs adversarial examples so that the model can learn the differences between different architectures in adversarial features, thereby enabling the model to further capture more common features.



Experiments

SAA has stronger adversarial transferability with transfer attacks.

Attack	Target Model: CNNs Res18	Res50	Res101	VGG19	DN121	Inc-v3	ViTs ViT-B	Swin-B	PVT-v2	MobViT	Avg. ASR (%)
MI	57.7	**99.9**	58.1	54.2	55.1	39.0	9.4	33.0	38.0	35.7	42.2
MI-SAA	**84.1**	99.6	**74.7**	**80.3**	**81.8**	**65.7**	**24.3**	**48.7**	**52.9**	**62.5**	**63.9**
NI	58.9	**100.0**	63.2	59.3	61.4	40.0	9.6	37.4	41.8	38.1	45.5
NI-SAA	**86.1**	**99.9**	**76.3**	**82.2**	**83.7**	**67.6**	**24.0**	**50.6**	**55.7**	**64.8**	**65.7**
GI	57.3	**100.0**	62.3	60.5	60.5	40.7	12.5	36.8	40.7	39.6	45.7
GI-SAA	**86.5**	99.7	**78.8**	**83.9**	**84.8**	**70.4**	**27.6**	**52.7**	**55.8**	**66.3**	**67.4**
DI	44.1	**95.8**	41.7	56.1	44.2	26.1	5.6	30.9	36.7	35.1	35.6
DI-SAA	**74.6**	94.4	**61.1**	**81.1**	**73.2**	**53.1**	**10.6**	**44.0**	**50.8**	**63.6**	**56.9**
TI	38.5	**99.9**	**37.8**	33.9	36.1	24.2	5.4	21.0	29.0	21.1	27.4
TI-SAA	**59.7**	94.9	35.2	**50.6**	**54.5**	**40.6**	**9.5**	**22.8**	**29.3**	**33.6**	**37.3**
SSA	75.8	**99.9**	**78.6**	76.0	77.8	57.0	16.5	**48.5**	55.0	50.5	59.5
SSA-SAA	**91.5**	99.5	77.8	**85.7**	**88.4**	**74.9**	**23.4**	46.7	**57.1**	**66.0**	**67.9**
DI-MI	65.5	97.0	65.0	74.7	65.7	49.1	16.4	49.0	54.9	57.9	55.4
DI-MI-SAA	**91.9**	**98.7**	**84.9**	**94.1**	**90.5**	**78.3**	**34.1**	**69.8**	**76.1**	**86.8**	**78.5**
TI-MI	61.4	**99.9**	60.5	60.9	60.9	44.3	15.2	37.4	42.3	41.8	47.2
TI-MI-SAA	**84.8**	99.3	**71.9**	**79.0**	**81.8**	**69.1**	**27.0**	**45.0**	**52.8**	**62.4**	**63.8**
SSA-MI	89.6	**99.9**	92.2	89.5	91.0	77.6	39.2	**74.4**	76.4	76.3	78.5
SSA-MI-SAA	**96.3**	99.8	**95.6**	**96.5**	**97.2**	**91.5**	**46.3**	74.1	**80.4**	**88.4**	**85.1**
SSA-DI-TI-MI	93.5	98.5	92.3	95.0	93.7	85.5	**55.9**	**83.2**	87.1	89.8	86.2
SSA-DI-TI-MI-SAA	**97.5**	**98.8**	**93.8**	**97.6**	**96.7**	**94.3**	53.6	81.6	**84.4**	**94.7**	**88.2**

SAA further improves the adversarial transferability on ViTs.

Attack	Target Model: CNNs Res18	Res50	Res101	VGG19	DN121	Inc-v3	ViTs ViT-B	Swin-B	PVT-v2	MobViT	Avg. ASR (%)
SGM	82.9	67.6	59.4	81.2	75.4	71.3	**99.7**	83.3	72.7	78.8	78.3
SGM-SAA	**91.1**	**79.8**	**73.3**	**87.5**	**87.3**	**80.9**	99.5	**90.5**	**82.6**	**86.3**	**86.6**
PatchOut	45.6	27.4	20.3	45.5	36.1	33.9	93.0	41.0	34.2	40.5	43.3
PatchOut-SAA	**76.5**	**72.4**	**70.3**	**79.4**	**78.1**	**71.3**	**94.7**	**83.6**	**77.2**	**76.8**	**78.7**
PNA	61.2	45.0	38.1	60.8	54.8	49.0	**99.6**	66.3	55.8	56.8	60.3
PNA-SAA	**82.7**	**78.1**	**73.4**	**85.6**	**84.0**	**75.1**	97.4	**89.3**	**80.3**	**82.1**	**83.3**
TGR	74.0	55.6	48.4	73.2	66.6	59.0	**99.3**	74.5	61.6	69.6	69.6
TGR-SAA	**85.9**	**78.1**	**71.5**	**87.4**	**85.6**	**79.6**	**99.3**	**89.1**	**81.0**	**86.2**	**85.1**

SAA improves adversarial transferability against defenses.

Attack	HGD	R&P	NIPS-r3	JPEG	FD	RS	Bit-Red	NRP	Diffpure	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$	Avg. ASR (%)
MI	42.5	21.9	25.3	33.9	42.4	23.6	29.3	6.7	13.8	33.3	31.3	23.3	27.3
MI-SAA	**73.3**	**57.8**	**60.4**	**69.9**	**65.6**	**39.6**	**42.3**	**12.0**	**22.4**	**68.9**	**65.5**	**57.9**	**53.0**
SSA-DI-TI-MI	93.7	89.6	90.2	91.9	89.7	82.3	81.7	14.8	71.1	92.5	91.1	89.8	81.5
SSA-DI-TI-MI-SAA	**96.0**	**93.2**	**94.8**	**95.1**	**94.0**	**89.8**	**85.5**	**20.2**	**78.8**	**95.7**	**94.5**	**93.3**	**85.9**

Grad-CAM on target models.