PubDef

Defending Against Transfer Attacks From Public Models

Chawin Sitawarin[1] Jaewon Chang*[1] David Huang*[1]
Wesson Altoyan[2] David Wagner[1]

[1] UC Berkeley
[2] King Abdulaziz City for Science and Technology

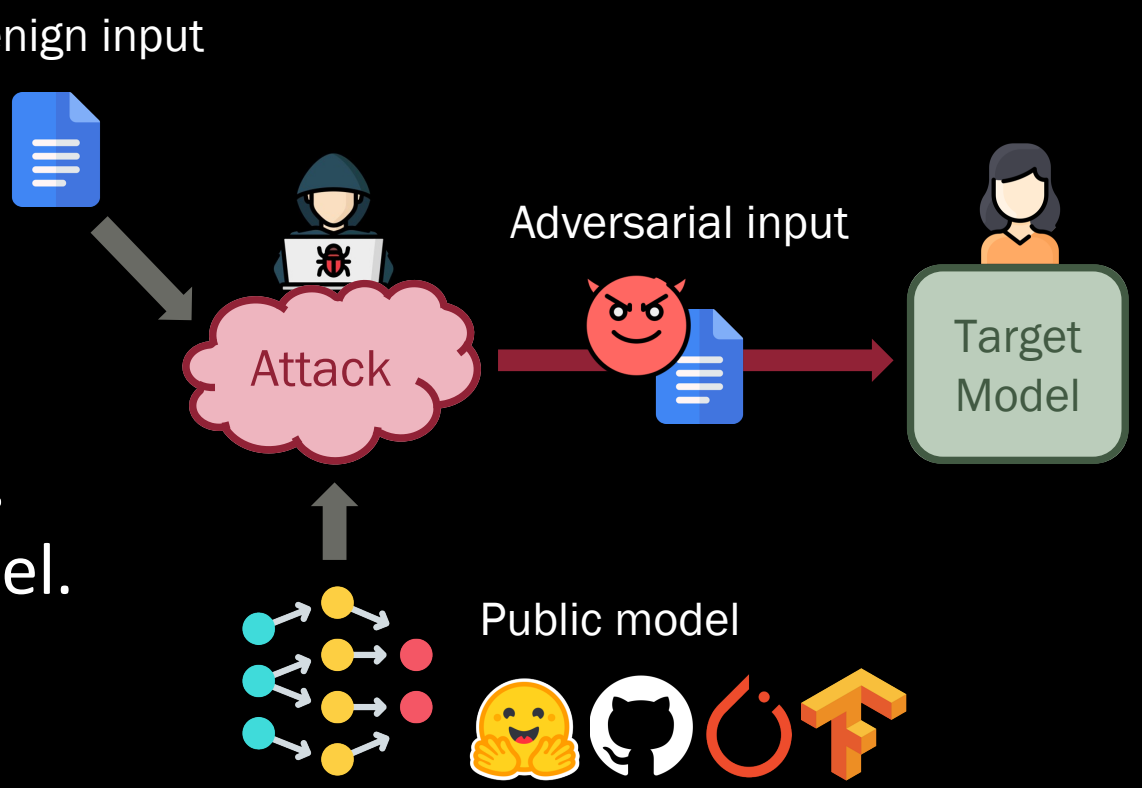
ICLR

KACST
مدينة الملك عبدالعزيز للعلوم والتقنية

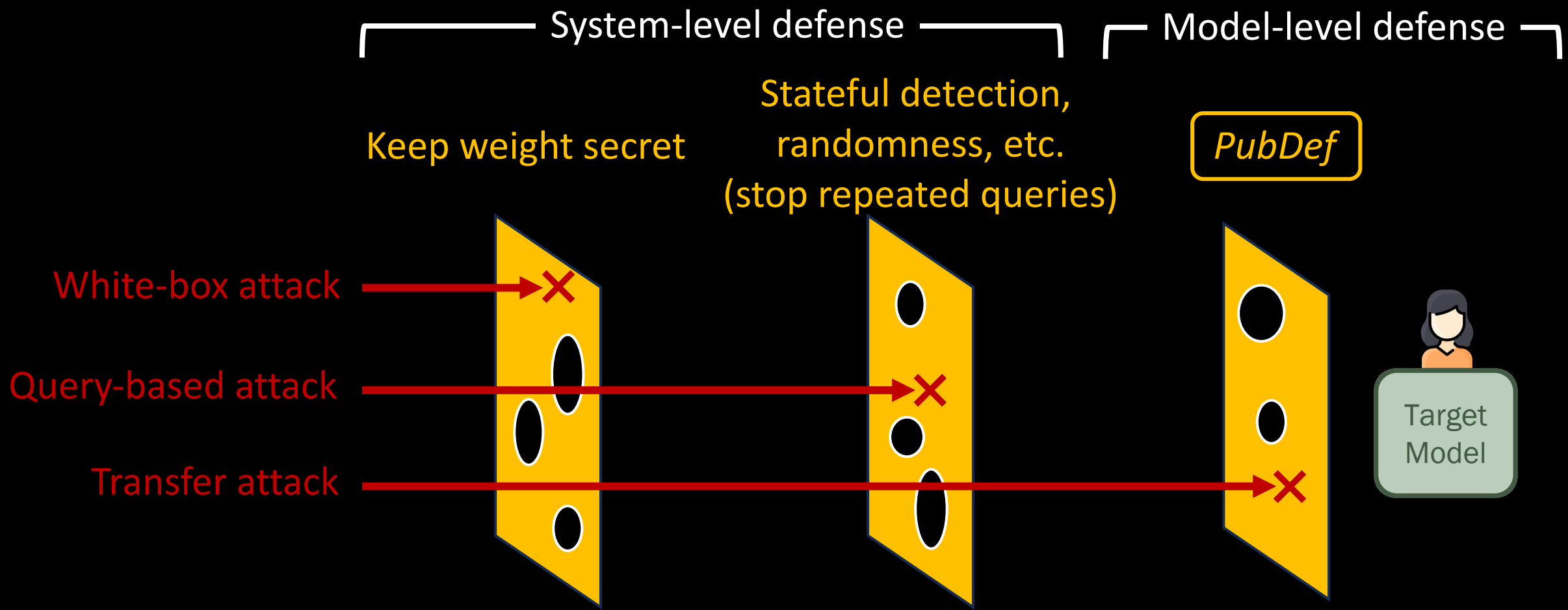
“Reasonable” Threat to Defend Against

Transfer attack

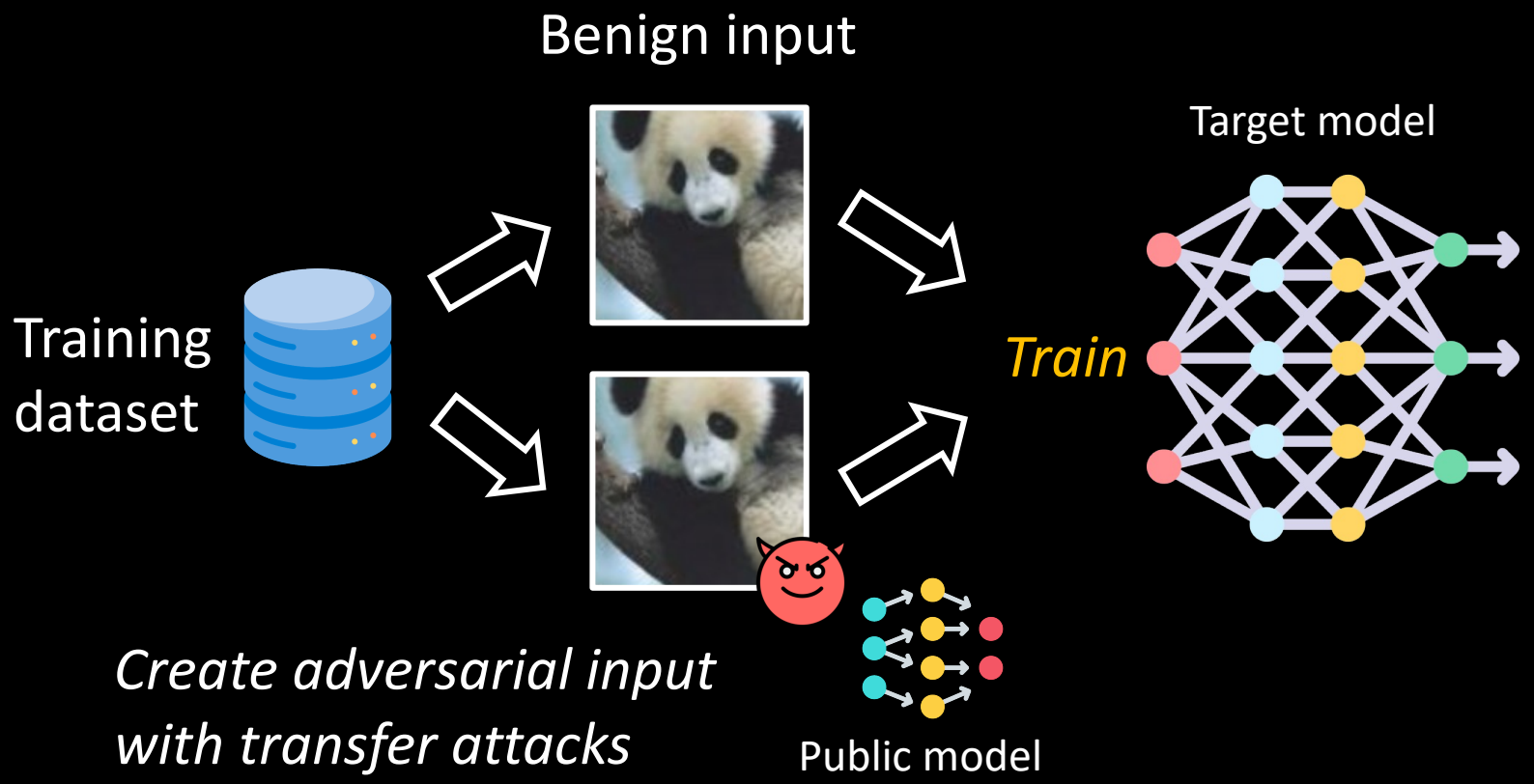
- Create adversarial examples on models with white-box access (*e.g.*, public open-source models).
- Transfer the attacks to target model.
- Only query target model once.



“Swiss Cheese” Model of Our Defense



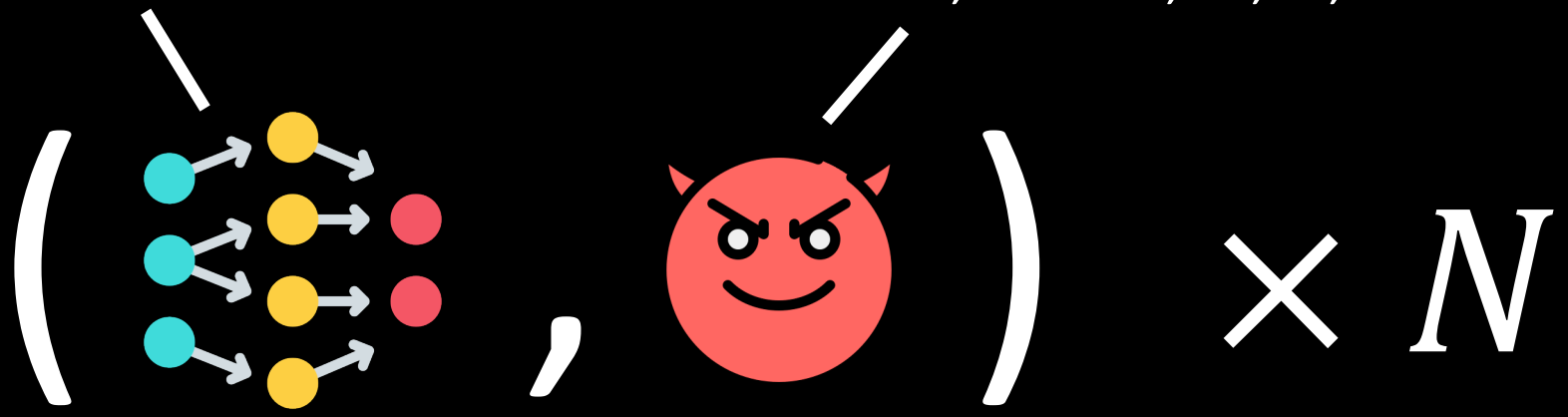
PubDef: Defense Against Transfer Attack



PubDef: Defense Against Transfer Attack

ResNet, ViT, ConvNeXt, …

PGD, M-PGD, DI, TI, …



Public model

Attack algorithm

PubDef: Defense Against Transfer Attack

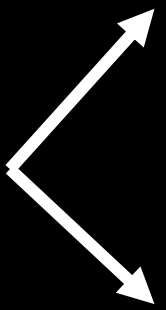
Q: How well does this defense generalize to *unseen* public models and *unseen* attack algorithms?

PubDef: Defense Against Transfer Attack

24 public models.

11 attack algorithms.

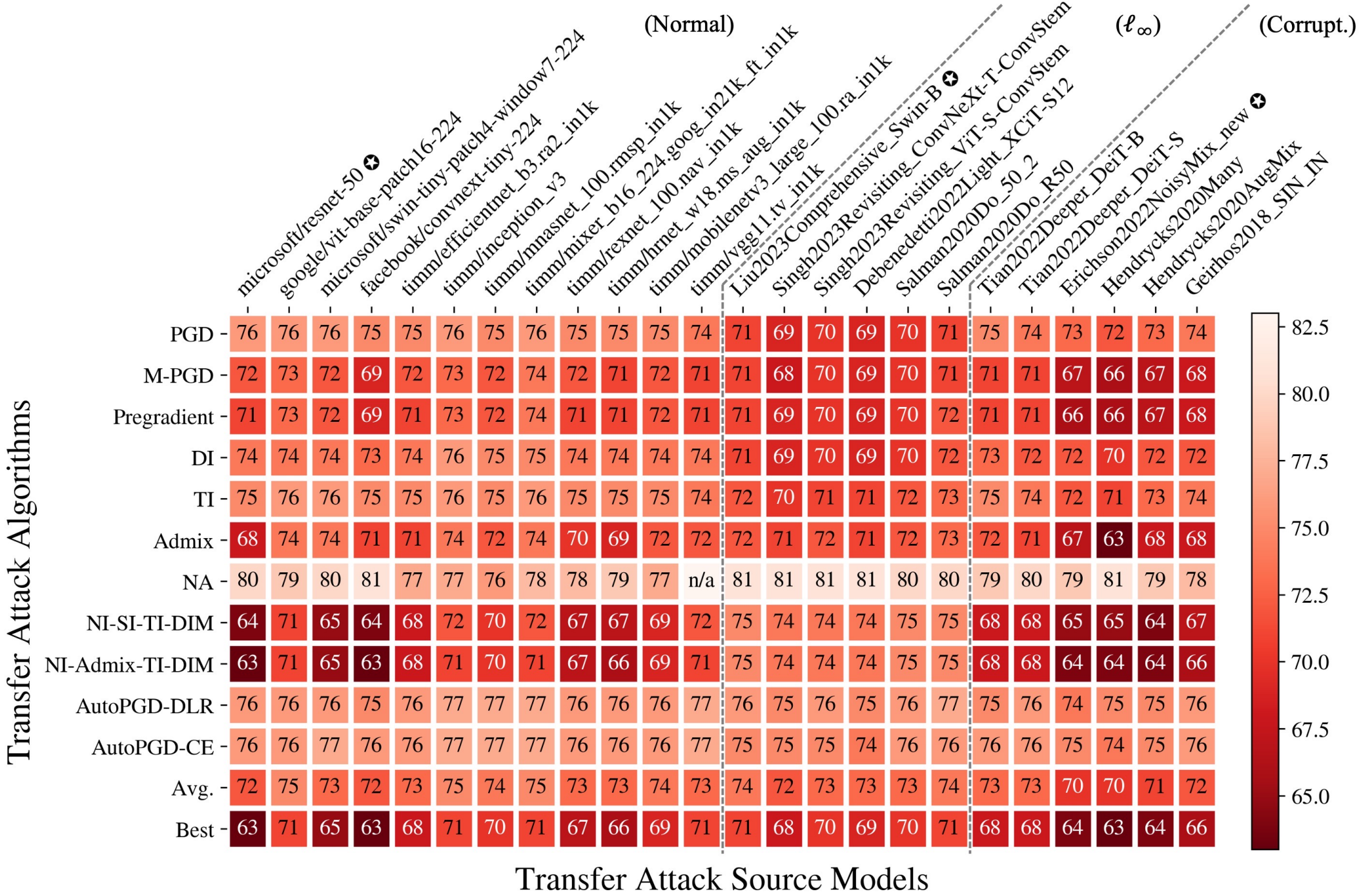
24 x 11 = 264 attacks in total.



Use 4 during training (seen)

260 are unseen

(make sure that they are diverse)


(Normal)
(ℓ_∞)
(Corrupt.)
microsoft/resnet-50 ✪
google/vit-base-patch16-224
microsoft/swin-tiny-patch4-window7-224
facebook/convnext-tiny-224
timm/efficientnet_b3.ra2_in1k
timm/inception_v3
timm/mnasnet_100.rmsp_in1k
timm/mixer_b16_224.goog_in21k_ft_in1k
timm/rexnet_100.nav_in1k
timm/hrnet_w18.ms_aug_in1k
timm/mobilenetv3_large_100.ra_in1k
timm/vgg11.tv_in1k
Liu2023Comprehensive_Swin-B ✪
Singh2023Revisiting_ConvNeXt-T-ConvStem
Singh2023Revisiting_ViT-S-ConvStem
Debenedetti2022Light_XCiT-S12
Salman2020Do_50_2
Salman2020Do_R50
Tian2022Deeper_DeiT-B
Tian2022Deeper_DeiT-S
Erichson2022NoisyMix_new ✪
Hendrycks2020Many
Hendrycks2020AugMix
Geirhos2018_SIN_IN
PGD 76 76 76 75 75 76 75 76 75 75 75 74 71 69 70 69 70 71 75 74 73 72 73 74
M-PGD 72 73 72 69 72 73 72 74 72 71 72 71 71 68 70 69 70 71 71 71 67 66 67 68
Pregradient 71 73 72 69 71 73 72 74 71 71 72 71 71 69 70 69 70 72 71 71 66 66 67 68
DI 74 74 74 73 74 76 75 75 74 74 74 74 71 69 70 69 70 72 73 72 72 70 72 72
TI 75 76 76 75 75 76 75 76 75 75 75 74 72 70 71 71 72 73 75 74 72 71 73 74
Admix 68 74 74 71 71 74 72 74 70 69 72 72 72 71 72 71 72 73 72 71 67 63 68 68
NA 80 79 80 81 77 77 76 78 78 79 77 n/a 81 81 81 81 80 80 79 80 79 81 79 78
NI-SI-TI-DIM 64 71 65 64 68 72 70 72 67 67 69 72 75 74 74 74 75 75 68 68 65 65 64 67
NI-Admix-TI-DIM 63 71 65 63 68 71 70 71 67 66 69 71 75 74 74 74 75 75 68 68 64 64 64 66
AutoPGD-DLR 76 76 76 75 76 77 77 77 76 76 76 77 76 75 76 75 76 77 75 76 74 75 75 76
AutoPGD-CE 76 76 77 76 76 77 77 77 76 76 76 77 75 75 75 74 76 76 76 76 75 74 75 76
Avg. 72 75 73 72 73 75 74 75 73 73 74 73 74 72 73 73 73 74 73 73 70 70 71 72
Best 63 71 65 63 68 71 70 71 67 66 69 71 71 68 70 69 70 71 68 68 64 63 64 66
82.5
80.0
77.5
75.0
72.5
70.0
67.5
65.0
Transfer Attack Algorithms
Transfer Attack Source Models

PubDef: Defense Against Transfer Attack

Models	Normal Accuracy	Accuracy Against Best Attack
Adversarial training	85	69
PubDef (ours)	**96**	**89**

(92 on seen attacks)

- Robust against all 264 attacks (4 seen, 260 unseen).
- Does not sacrifice on normal accuracy: Close to SOTA accuracy.
- Much faster than adversarial training (~2x): Pre-compute the attacks.

PubDef: Defense Against Transfer Attack | Takeaways

1. Don’t always need adversarial training to build a secure ML system in practice.
2. Be clear about the threat model.
3. Both system-level and model-level defenses are necessary. Use them to your advantage.

Future work:

Can we design a better model-level or system-level against sophisticated query-based attacks?