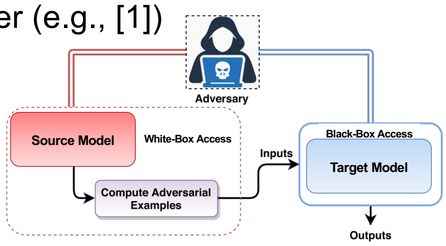
The Ultimate Combo: Boosting Adversarial Example Transferability by Composing Data Augmentations

Zebin Yun (TAU), Achi-Or Weingarten (Weizmann), Eyal Ronen (TAU), Mahmood Sharif (TAU)

מכון ויצמן למדע
WEIZMANN INSTITUTE OF SCIENCE

Motivation

Adversarial examples (AEs) often transfer between models; augmentations boost transfer (e.g., [1])



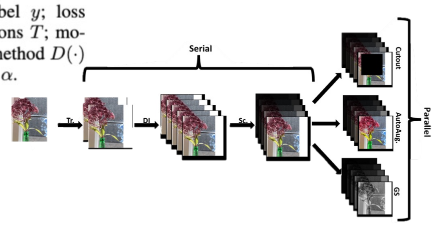
Prior attack only explore limited number of augmentations. ***Can we do better by combining more augmentations?***

New Composition Method

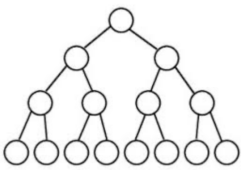
We propose parallel composition to integrate many augmentations into attacks

Algorithm MI-FGSM with data augmentation

Input: Benign sample x; ground-truth label y; loss function $J(\cdot)$; model parameters θ; # iterations T; momentum parameter μ; perturbation norm ϵ; method $D(\cdot)$ producing m augmented samples; step size α.
$\alpha = \epsilon / T$
$\hat{x}_0 = x$
$g_0 = 0$
for $t = 0$ to $T-1$ **do**
  $\bar{g}_{t+1} = \frac{1}{m}\sum_{i=0}^{m-1} \nabla_x \left(J(D(\hat{x}_t)_i, y, \theta)\right)$
  $g_{t+1} = \mu \cdot g_t + \frac{\bar{g}_{t+1}}{\|\bar{g}_{t+1}\|_1}$
  $\hat{x}_{t+1} = \mathrm{Proj}_x^{\epsilon}\left(\hat{x}_t + \alpha \cdot \mathrm{sign}\left(g_{t+1}\right)\right)$
end for
return $\hat{x} = \hat{x}_T$



Finding the *Ultimate Combo*



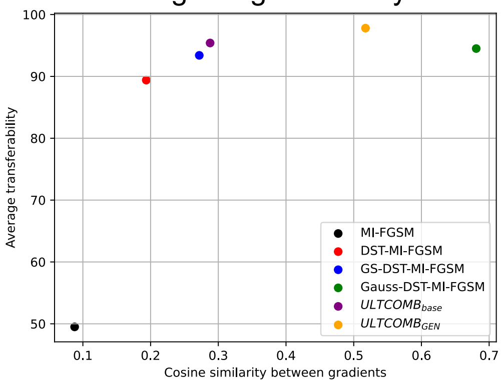
Grid search on a limited search space (2^7 choices) to find the $ULTCOMB_{base}$

Genetic search on the full search space (2^{48} choices) to find the $ULTCOMB_{gen}$

Results

But why some augmentations can help improve transferability whereas others can't?

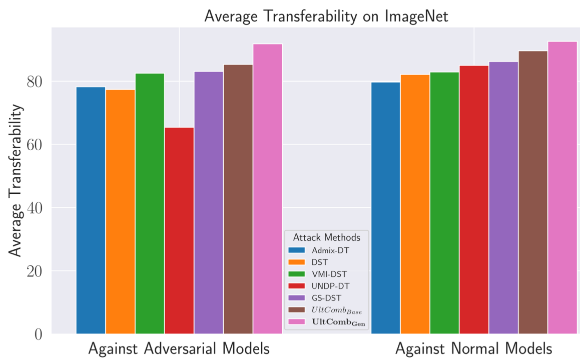
1. Increasing gradient similarity
2. Preserving benign accuracy



For qualified augmentations, we find monotonicity:
more augmentations → high transferability

Ultimate Combo's AEs transfer better than other attacks!

Against normally and adversarially trained targets:

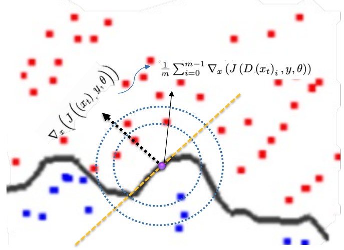


From an ensemble of normally trained surrogates to defended ImageNet models:

Defense	Admix-DT	DST	VMI-DST	UNDP-DT	UltComb$_{Base}$	UltComb$_{Gen}$
Bit-Red	88.6	88.2	94.8	94.9	**96.0**	*95.5*
NRP	51.0	54.9	**80.0**	27.9	*65.3*	55.8
RS	87.3	84.8	90.6	85.5	*88.5*	**95.6**
ARS	65.4	62.9	66.5	61.9	*67.0*	**71.9**

Theoretical Analysis

Some augmentations smoothen the model gradients
(proven with techniques from randomized smoothing)



We expect this reduces the effect of surrogate models' peculiarities on adversarial examples
→ better generalization to unseen models

[1] Xie, Cihang, et al. "Improving transferability of adversarial examples with input diversity." *CVPR*. 2019.