Testing robustness against unforeseen adversaries

Daniel Kang*, **Yi Sun***, Tom Brown,
Dan Hendrycks, Jacob Steinhardt

Columbia University

I. Setting: Unforeseen adversarial attacks

II. Novel and diverse adversarial attacks

III. ImageNet-UA: Measuring robustness to unforeseen attacks

IV. New insights from ImageNet-UA

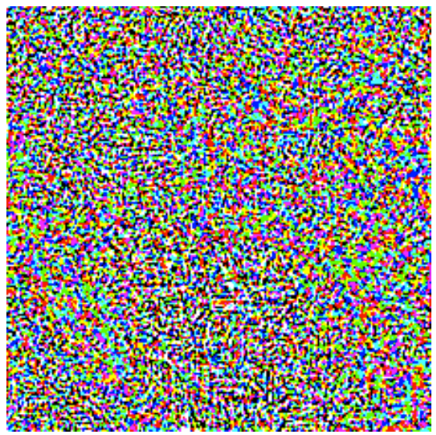
Adversarial attacks for image classification

Neural networks classify clean images well...

$\boldsymbol{x}$

"panda"
57.7% confidence

$+ .007 \times$



$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"nematode"
8.2% confidence

$=$

$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"gibbon"
99.3 % confidence

[Goodfellow-Shlens-Szegedy '14]

... but are vulnerable to adversarially crafted attacks.

Adversarial attacks as distributional shift

Test and train distributions overlap for ordinary evaluation:

Adversarial attacks move test images outside the train distribution.

L_∞ attack

Most adversarial attacks allow a fixed type of distortion, e.g. L_∞:

Original

Attacked ($\varepsilon = 32/255$)

- View input as a pixel vector $x \in \mathbb{R}^{3\times 224\times 224}$
- Attacked image x' maximizes loss under the L_∞-constraint

$$\|x' - x\|_\infty \leq \varepsilon,$$

  where each pixel can change by at most ε.
- Optimize using projected gradient descent (PGD)

Adversarial defense: L_∞ adversarial training

Adversarial training solves the min-max problem:

$$\min_{\theta} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \left[\max_{\|x'-x\|_\infty \leq \varepsilon} \ell(F_\theta(x'), y) \right].$$

[Madry et. al. '17, Xie et. al. '18]

Training time: train on attacked images

- Solve outer minimization with SGD
- Approximate inner maximization with PGD

Test time: evaluate on attacked images

- Distributional shift no longer appears

Requires defender to **know attack** is L_∞.

L_2 attack

Replace L_∞ by L_2 in PGD:

Original

Attacked ($\varepsilon = 4800$)

- Attacked image x' maximizes loss under the L_2-constraint

$$\|x' - x\|_2 \leq \varepsilon,$$

  where the Euclidean distance is at most ε.
- Optimize using projected gradient descent (PGD)

Adversarial defense: L_2 randomized smoothing

Randomized smoothing finds a smoothed classifier via:

$$\min_{\theta} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \left[\mathop{\mathbb{E}}_{x'\sim\mathcal{N}(x,\sigma^2\cdot I)} \left[\ell(F_\theta(x'), y) \right] \right]$$

[Cohen et. al. '19, Salman et. al. '19, Raff et. al. '19]

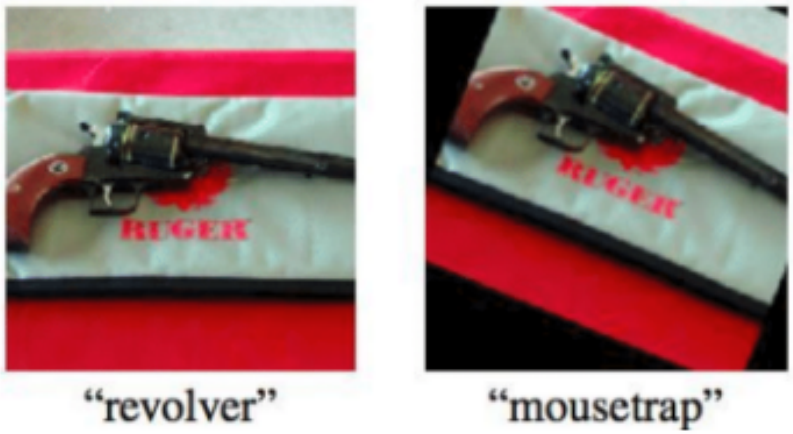
Gives provably robust defense against L_2-attacks:

- Training: use images with average-case Gaussian noise
- Test: evaluate on attacked images

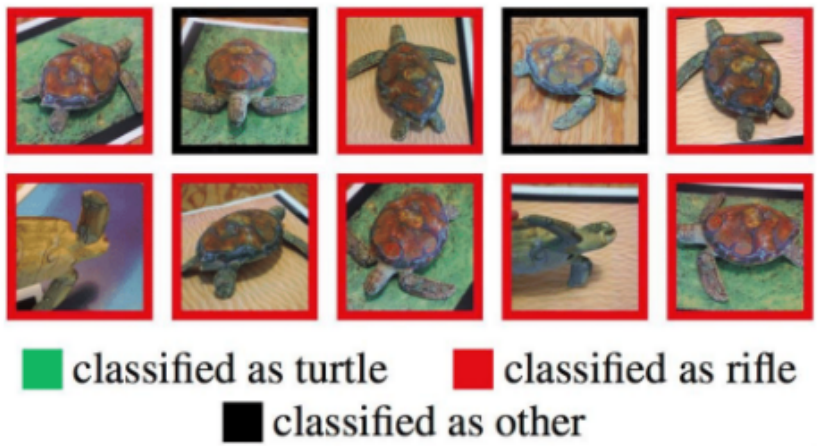
Requires defender to **know attack** is L_2.

Adversarial attacks in practice

Real-world attacks fall outside the L_p paradigm:



Adversarial rotations



3-D printed turtles

[Athalye et. al. '17, Engstrom '18]

In practice: Distortions can even be **hard to specify**!

Adversarial attacks in practice

Adversaries can deploy unforeseen attacks...

```
<a><span>
<span class="c1">Sp</span>
<span class="c2">S</span>
<span class="c1">on</span>
<span class="c2">S</span>
<span class="c1">so</span>
<span class="c2">S</span>
<span class="c1">red</span>
<span class="c2">S</span>
</span></a>
```

```
.c2 { font-size: 0; }
```

[Tramèr et. al. '18]

... leading to an arms race between attack and defense.

Adversarial evaluation: L_∞ evaluation

Best practices in adversarial evaluation include diverse attacks...

- §4.3 Apply a **diverse set of attacks** (especially when training on one attack approach).
  - Do not blindly apply multiple (nearly-identical) attack approaches.

"On evaluating adversarial robustness" [Carlini et. al. '19]

...but papers measure adversarial robustness with L_∞ robustness.

In this paper, we consider the specification that the image predictions should remain the same within an ℓ_∞-ball of an image x, where an allowable maximum perturbation is $\epsilon = 16/255$, relative to the pixel intensity scaled between 0 and 1.

"Towards robust image classification using sequential attention models" [Zoran et. al. '19]

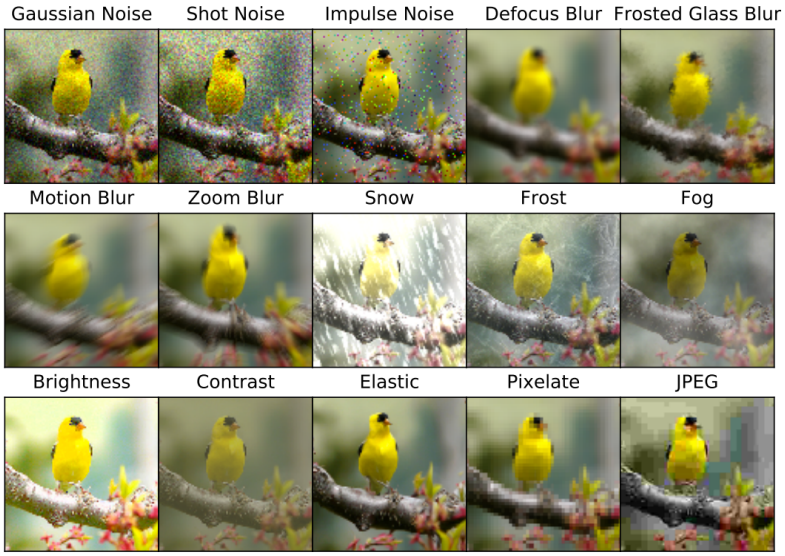
Black-box attacks. In the black-box setting, we ran attacks against the `resnext50_32x4d` model in `torchvision.models`. Note that this model is different from the five models considered in this paper. We set the number of PGD steps to 10 and the step size to 2/225. We varied the total perturbation size of the attack ϵ, defined as the l_∞-norm of the perturbation divided by the l_∞-norm of the clean image: $\epsilon \equiv ||\mathbf{x}_{\text{adv}} - \mathbf{x}||_\infty / ||\mathbf{x}||_\infty$, from 0.01 to 0.1. These attacks against the `resnext50_32x4d` model were highly successful, yielding below 10% top-1 accuracy even for the lowest perturbation size $\epsilon = 0.01$. We then tested the generated adversarial images with the five ResNeXt-101 models considered in this paper.

White-box attacks. In the white-box setting, attacks were run directly against the ResNeXt-101 models. Attack parameters were identical to those described in the previous paragraph. However, since using only a small number of PGD steps can lead to a significant overestimation of the robustness of a model against white-box adversarial attacks (Engstrom et al., 2018), we also ran stronger white-box attacks with up to 50 PGD steps (fixing the total perturbation size to $\epsilon = 0.06$ for these attacks).

"Robustness properties of Facebook's ResNeXt WSL models" [Orhan '19]

Comparison with other robustness evaluations

Non-worst-case robustness is often evaluated using diverse unforeseen datasets



ImageNet-C [Hendrycks et. al. '19]

ImageNet-A [Hendrycks et. al. '19]

... but this is difficult to do for adversarial / worst-case robustness.

ImageNet-UA: a new evaluation framework

Evaluation against unforeseen attacks requires:

- Diverse set of fast adversarial attacks
- Generic way of assessing robustness to each of these attacks

This talk: Framework ImageNet-UA for evaluation against unforeseen attacks.

- Introduce a diverse set of novel adversarial attacks
- Construct ImageNet-UA using these attacks
- Use ImageNet-UA to reveal weaknesses in existing evaluations and defenses

I. Setting: Unforeseen adversarial attacks

II. Novel and diverse adversarial attacks

III. ImageNet-UA: Measuring robustness to unforeseen attacks

IV. New insights from ImageNet-UA

Adversarial attacks for a test suite

For evaluation, need attacks which:

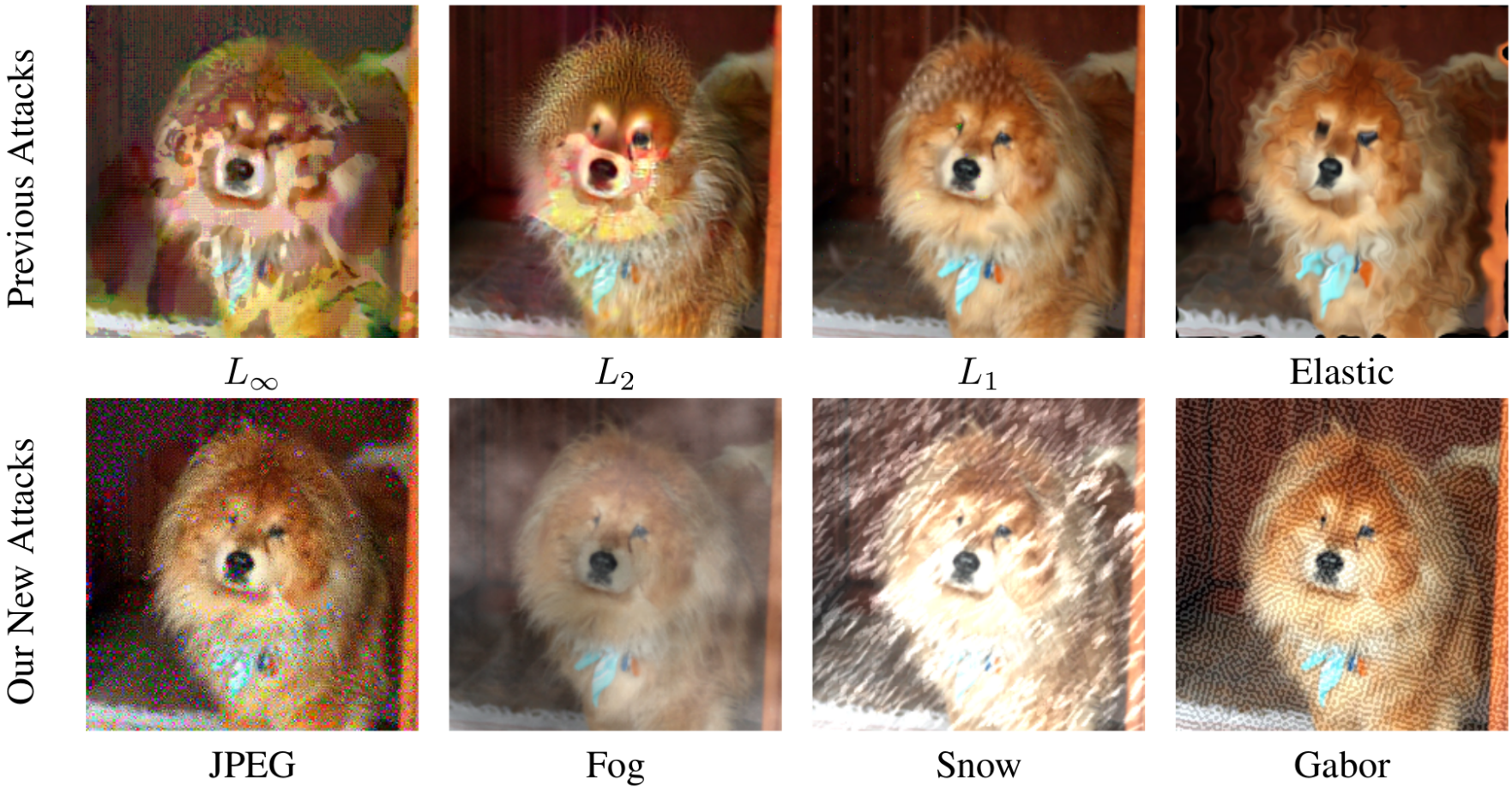
- use a diverse set of distortions, and
- are not used in the construction of the defense

Less obviously, also want attacks which are:

- strong enough to challenge weaker models
- fast and differentiable for ease of evaluation

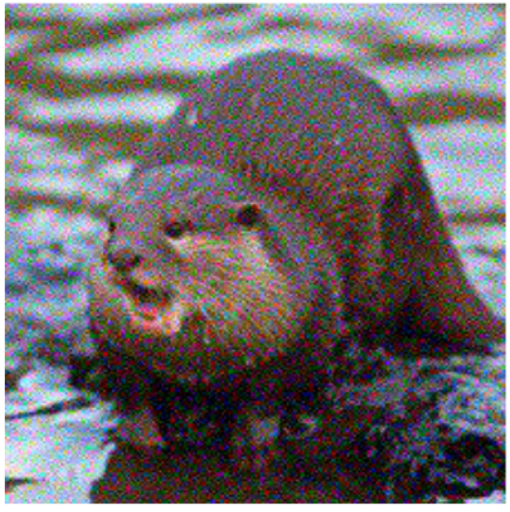
We construct 4 novel attacks satisfying all of these properties.

Diverse and novel adversarial attacks



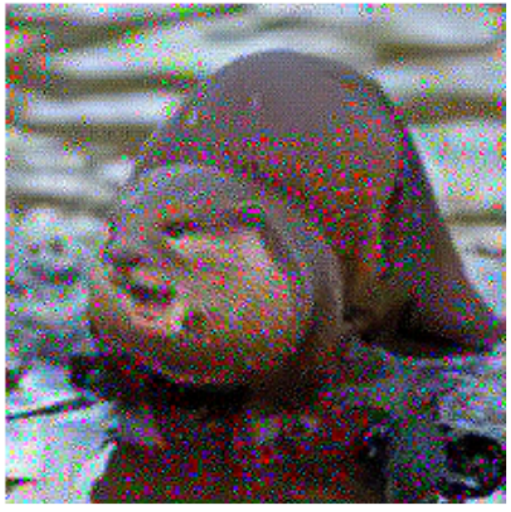
Novel JPEG attack

Randomly Initialized JPEG



Otter (100.0%)

Adversarially Optimized JPEG



Basketball (86.4%)

Uses PGD to optimize x' so that

$$\|\text{JPEG}(x) - \text{JPEG}(x')\|_\infty \leq \varepsilon,$$

where JPEG denotes JPEG compression without quantization.

Differentiable JPEG due to [Shin et. al. '17]

Novel Fog attack

Randomly Initialized Fog

Otter (100.0%)

Adversarially Optimized Fog

Titi Monkey (100.0%)

Creates fog-like distortions by adversarially optimizing parameters in the diamond-square algorithm.

Novel Snow attack

Randomly Initialized Snow

Otter (100.0%)

Adversarially Optimized Snow

Loafer (100.0%)

Generates occlusions of randomly located image regions and adversarially optimizes intensity and direction of these artificial snowflakes.

Novel Gabor attack

Randomly Initialized Gabor

Otter (100.0%)

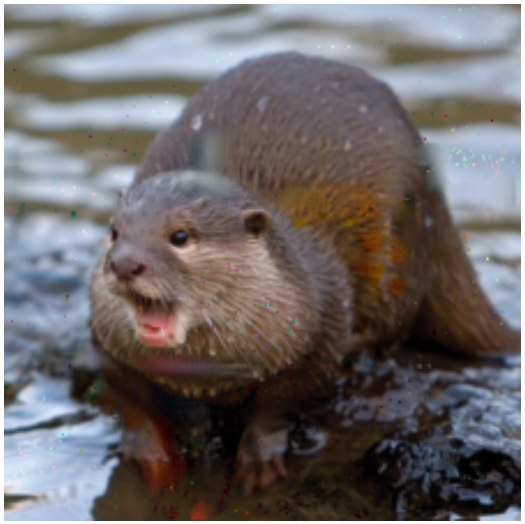
Adversarially Optimized Gabor

Zebra (76.3%)

Spatially occludes the image with procedural Gabor noise with adversarially optimized parameters (orientation, bandwidth).

Existing L_1 and Elastic attacks

Original



L_1

Elastic

L_1 attack: Optimize x' so that $\|x' - x\|_1 \leq \varepsilon$

- Use Frank-Wolfe to improve on [Chen et. al. '18, Tramèr & Boneh '19]

Elastic attack: Allow distortions $x' = \text{Flow}(x, V)$, where

- V is a vector field on pixel space
- V is a Gaussian smoothing of a field W with

$$\|W(i,j)\|_\infty \leq \varepsilon.$$

Modification of attack of [Xiao et. al. '18].

Stochastic vs. adversarial attacks

- Stochastic analogues appear in [Hendrycks et. al. '19, Co et. al. '19]
- Adversarial versions require careful choice of optimization parameters

I. Setting: Unforeseen adversarial attacks

II. Novel and diverse adversarial attacks

III. ImageNet-UA: Measuring robustness to unforeseen attacks

IV. New insights from ImageNet-UA

Evaluation with ImageNet-UA

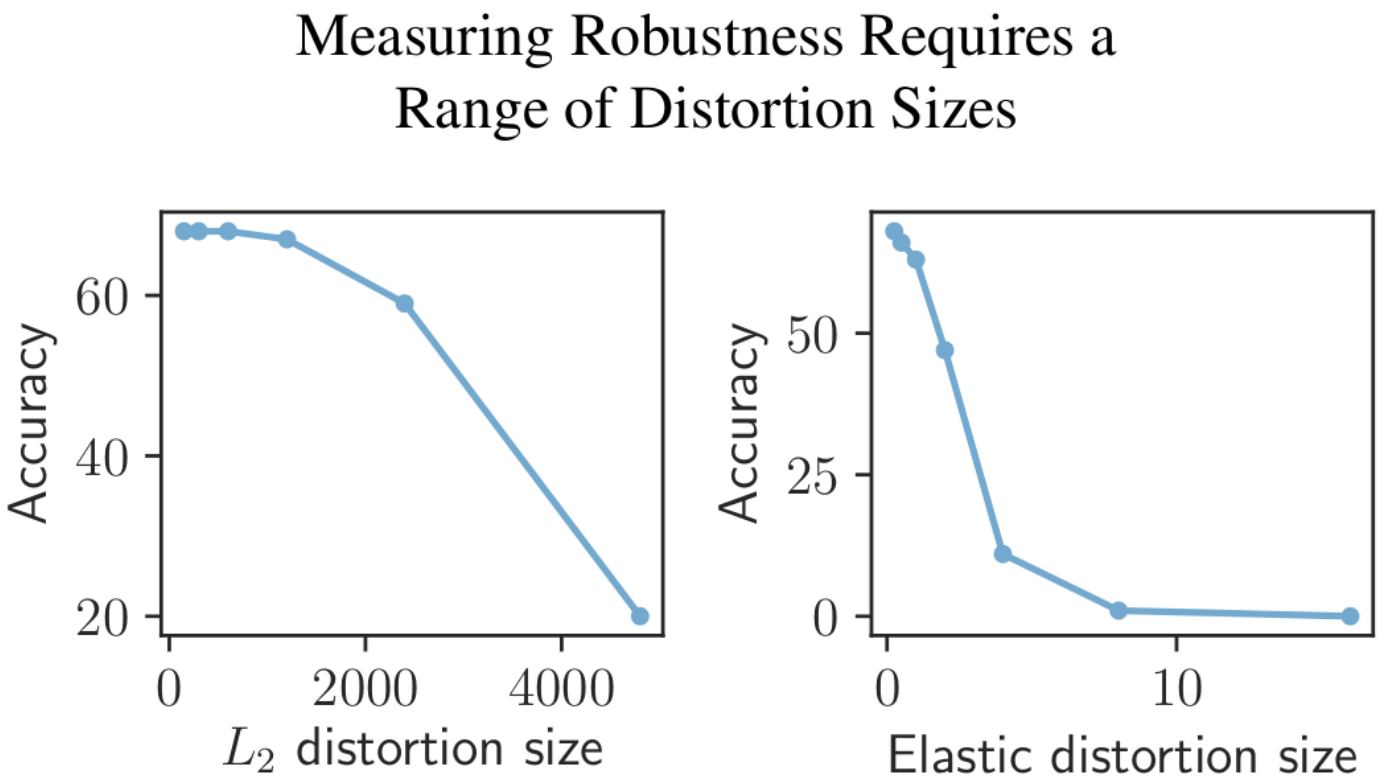
ImageNet-UA evaluation procedure:

- Test against L_1, Elastic, JPEG, Fog, Snow, and Gabor attacks at 6 strengths each
- For each attack, compute a metric UAR (Unforeseen Attack Robustness) assessing the defense
- Average robustness information into the metric

$$\text{mUAR} := \frac{1}{6}\Big[\text{UAR}(L_1) + \text{UAR}(\text{Elastic}) + \text{UAR}(\text{JPEG}) + \text{UAR}(\text{Fog}) + \text{UAR}(\text{Snow}) + \text{UAR}(\text{Gabor})\Big]$$

To define UAR, need to select distortion size...

Distortion size and evaluation



- The model is robust to both L_2 and Elastic for small distortions
- For large distortions, robustness persists for L_2 but vanishes for Elastic

Assessing attack strength

Goal: Choose range of distortion sizes corresponding to range of attack strengths

- Expect attack strength to increase with distortion size ε
- Measure strength by performance against certain models

Undefended models are too weak to measure attack strength

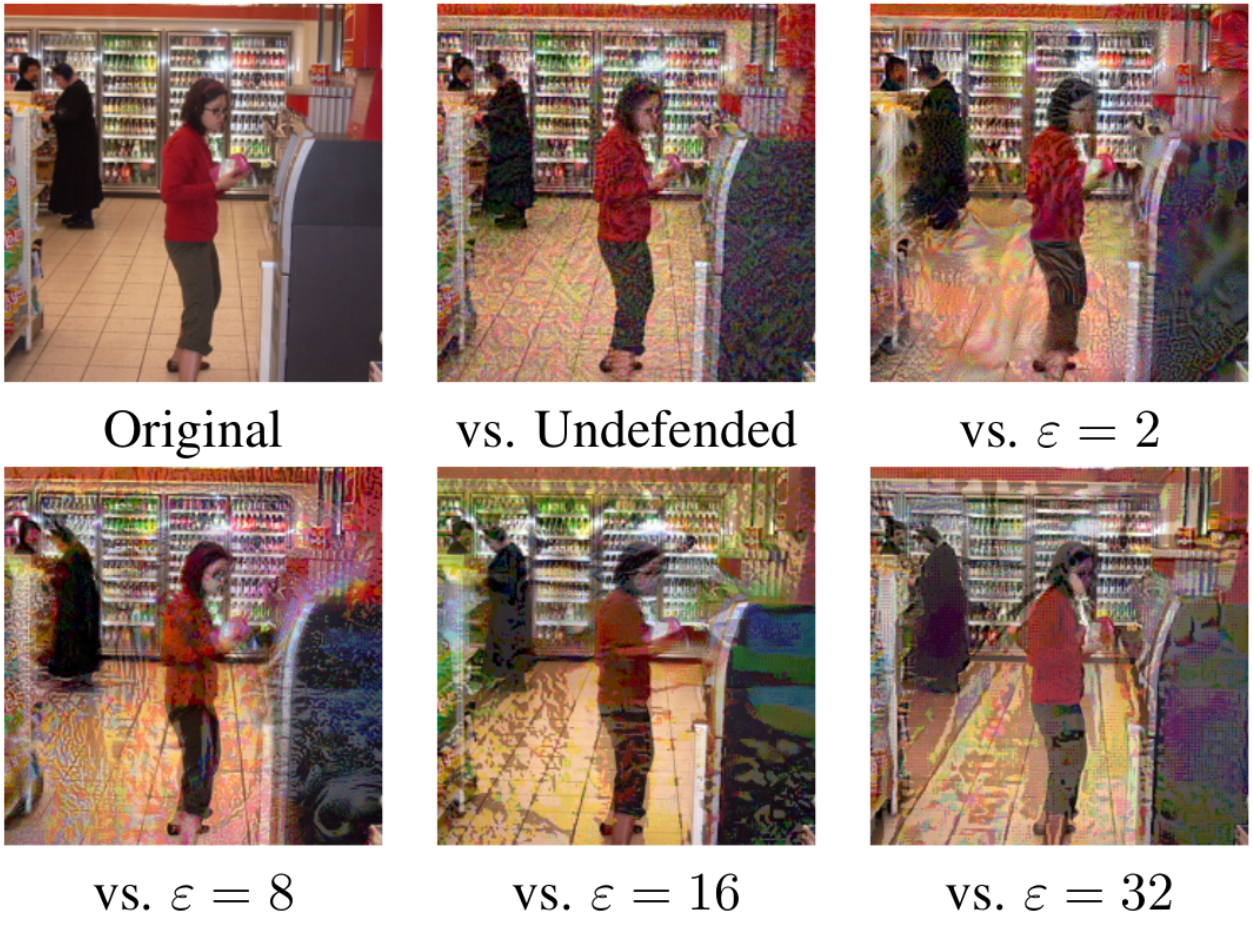
- adversarial accuracy of undefended models is 0 for small ε
- attacking undefended models barely changes images

Consider adversarially trained models instead

- strong generic defense which works for all attacks
- produces close to SOTA results on CIFAR-10, ImageNet

Adversarial training and attack strength

Adversarial Training Makes Adversarial Distortions More Conspicuous



The L_∞ attack at $\varepsilon = 32$ vs. L_∞ adversarially trained models

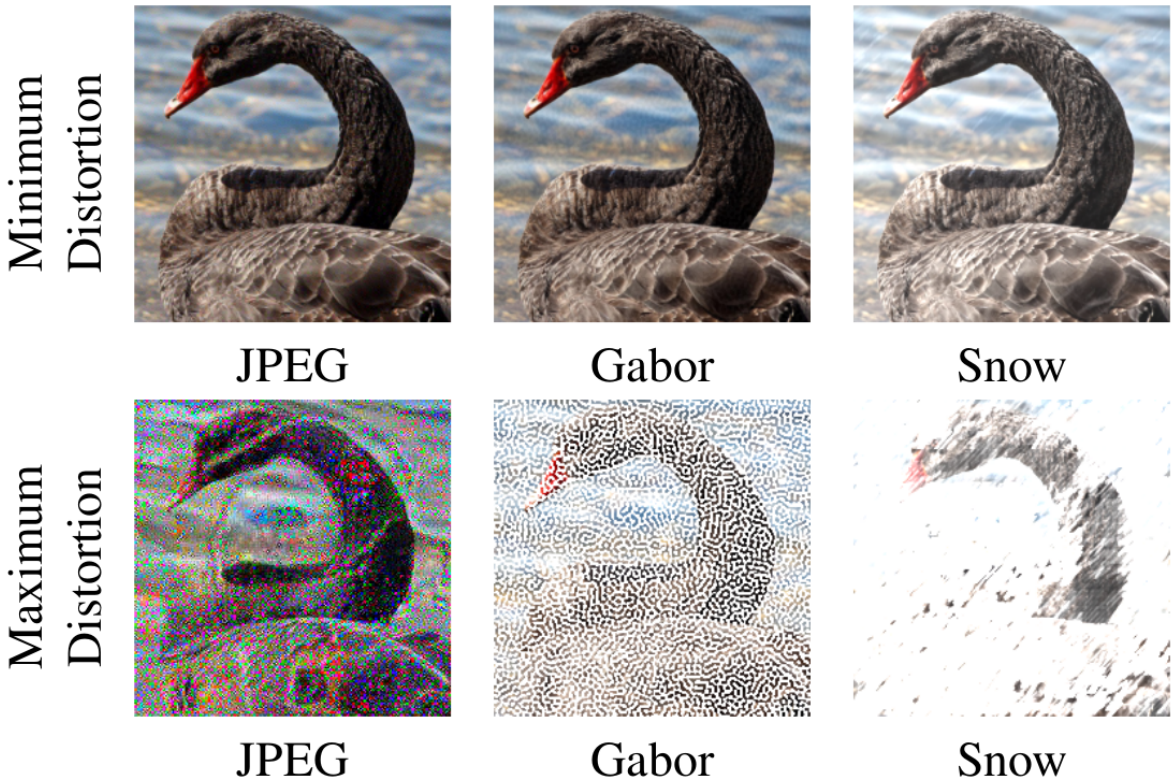
Attack strength: Adversarial Training Accuracy (ATA)

Measure strength of an attack A at distortion size ε:

$$\mathrm{ATA}(A, \varepsilon) := \left\{ \begin{array}{l} \text{best adversarial accuracy against} \\ A \text{ at size } \varepsilon \text{ achieved by adversarial} \\ \text{training against } A \text{ at size } \varepsilon' \end{array} \right\}$$

- Fixed architecture for each dataset (ResNet-50 for ImageNet)
- Estimate value of $\mathrm{ATA}(A, \varepsilon)$ by optimizing over a range of ε'

Adversarial training and distortion sizes



Choose ε_{min} and ε_{max} so that

- ε_{min} is maximal so that ATA$(A, \varepsilon_{\text{min}})$ is comparable to accuracy of an undefended model on clean data
- ε_{max} is minimal so that it reduces an adversarially trained model below 25 or confuses humans

Choosing distortion sizes

Choose $\varepsilon_1, \ldots, \varepsilon_6$ to:

- be between ε_{min} and ε_{max}
- approximately match $\text{ATA}(A, \varepsilon_k)$ values with L_∞ at $\varepsilon = 1, \ldots, 32$

Example for Gabor:

Attack	ε_1	ε_2	ε_3	ε_4	ε_5	ε_6
L_∞	1	2	4	8	16	32
Gabor	6.25	12.5	25	400	800	1600
Attack	ATA_1	ATA_2	ATA_3	ATA_4	ATA_5	ATA_6
L_∞	84.6	82.1	76.2	66.9	40.1	12.9
Gabor	84.0	79.8	79.8	66.2	44.7	14.6

UAR: Robustness against a single attack

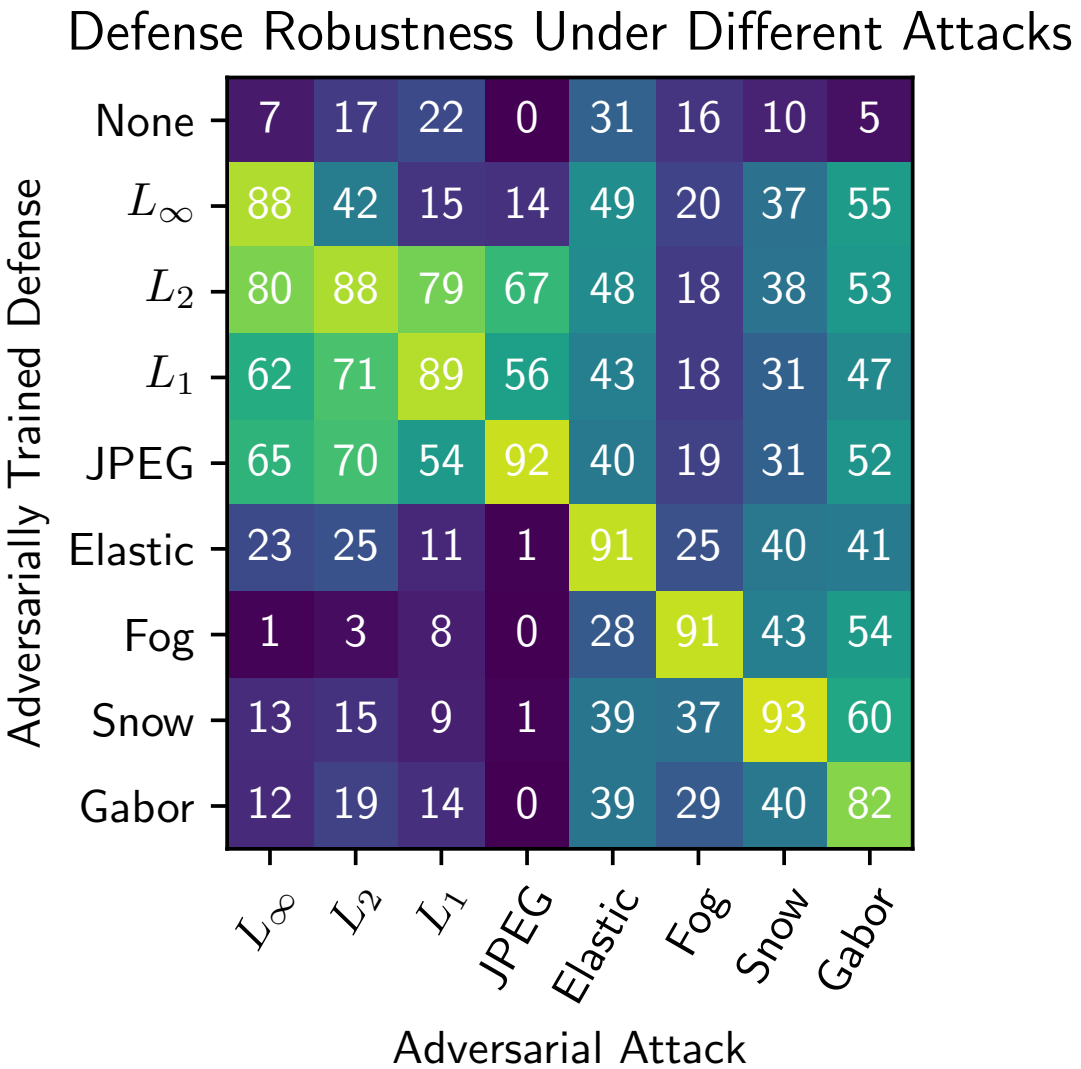
For a model M and six distortion sizes $\varepsilon_1, \ldots, \varepsilon_6$ for attack A:

$$\text{UAR}(A) := 100 \times \frac{\sum_{k=1}^{6} \text{Acc}(A, \varepsilon_k, M)}{\sum_{k=1}^{6} \text{ATA}(A, \varepsilon_k)},$$

where $\text{Acc}(A, \varepsilon_k, M)$ is the accuracy of M under A at size ε_k.

- Compares M to adversarial training against A
  - Challenging to achieve UAR near 100 without knowledge of A.
- Provides a value more commensurable between attacks
- Can be computed using 6 adversarial evaluations (with precomputed ATA values)
  - We provide values for ImageNet-100 (100-class subset of ImageNet) and CIFAR-10

UAR: Robustness against a single attack



UAR values on ImageNet-100 (100 class subset of ImageNet)

mUAR: Mean Unforeseen Attack Robustness

Summarize performance on ImageNet-UA using mUAR:

$$\text{mUAR} := \frac{1}{6}\Big[\text{UAR}(L_1) + \text{UAR}(\text{Elastic}) + \text{UAR}(\text{JPEG}) + \text{UAR}(\text{Fog}) + \text{UAR}(\text{Snow}) + \text{UAR}(\text{Gabor})\Big]$$

We advise against adversarially training against these 6 attacks when evaluating ImageNet-UA.

- Estimates robustness to a broad threat model of 6 attacks
- High mUAR indicates generalization to several held-out attacks

I. Setting: Unforeseen adversarial attacks

II. Novel and diverse adversarial attacks

III. ImageNet-UA: Measuring robustness to unforeseen attacks

IV. New insights from ImageNet-UA

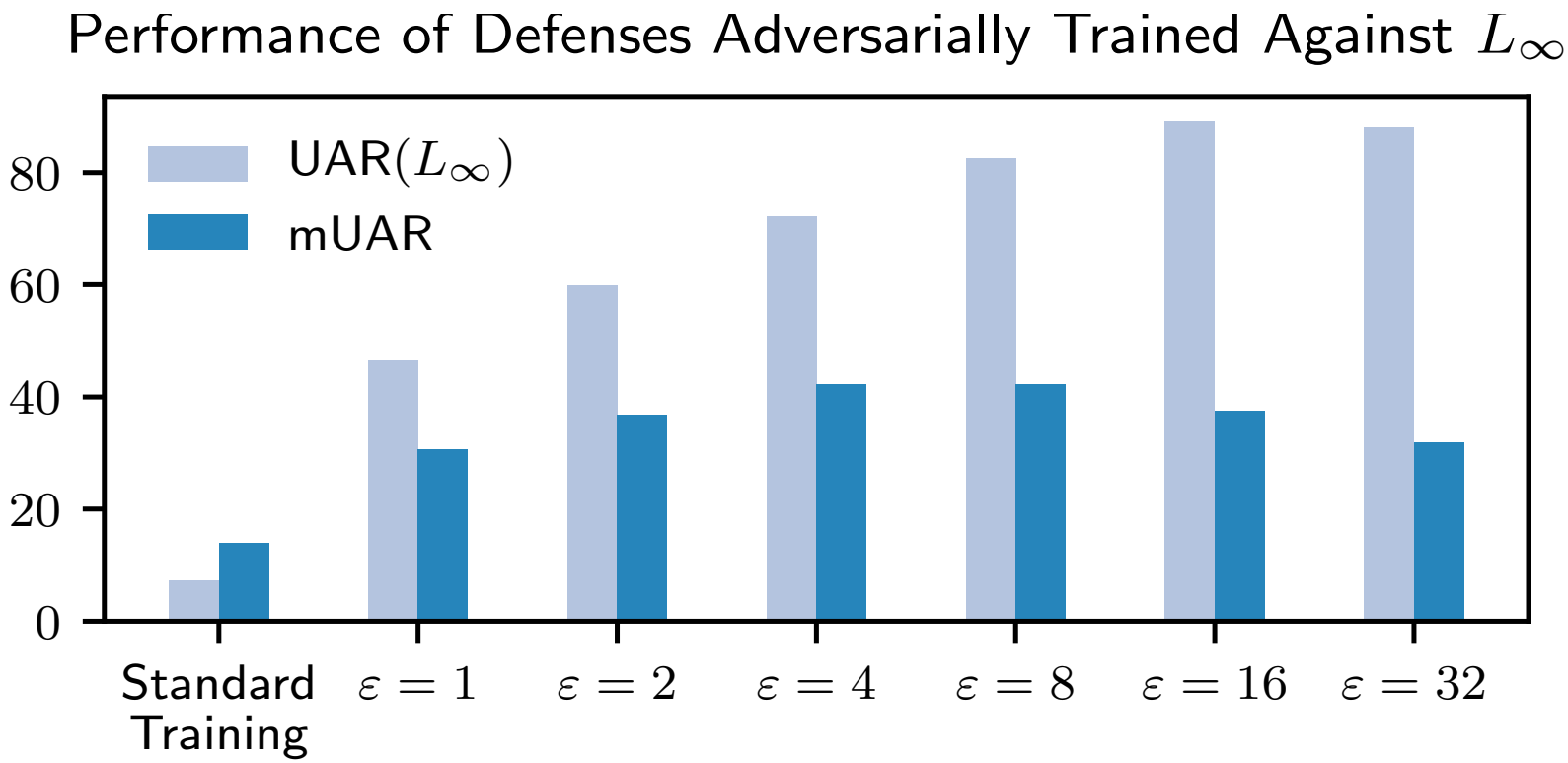
Experimental setup

Adversarially train 48 models against 8 attacks at 6 distortion sizes

- Dataset: ImageNet-100 (100 class subset of ImageNet)
- Apply randomly targeted attacks at both training and test
- Train on 10-step attacks[1], test on 200-step attacks
- Randomly scale distortion size during training only
- Backprop on attacked images only
- Distributed training on 8-GPU machines (at OpenAI)

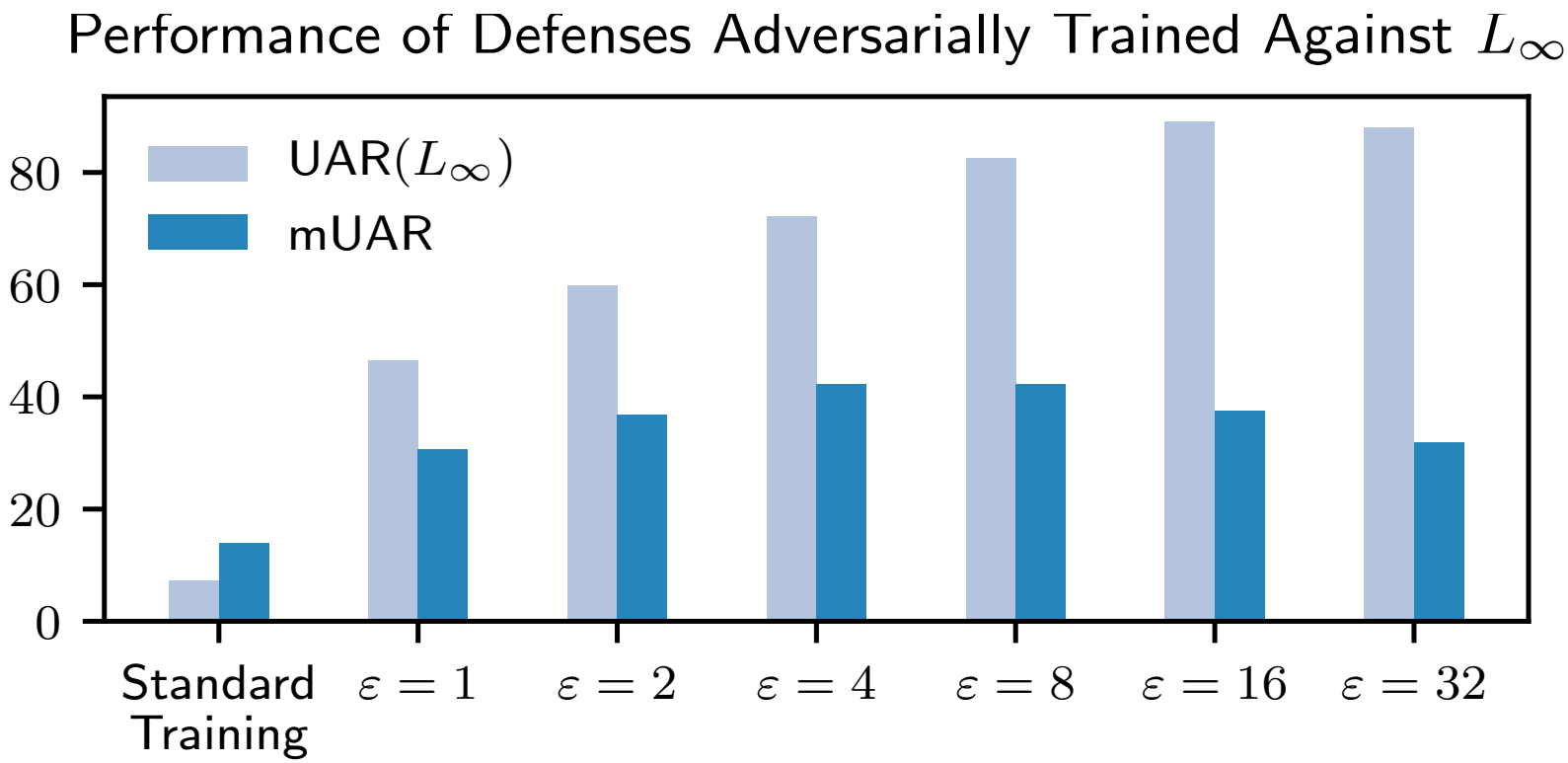
[1] 30 steps for Elastic.

ImageNet-UA reveals weakness in L_∞ training



- For $\varepsilon \leq 4$, L_∞ training improves UAR(L_∞) and mUAR
- For $\varepsilon \geq 8$, L_∞ training overfits to L_∞ at the expense of generalization

ImageNet-UA reveals weakness in L_∞ evaluation



- Evaluating against L_∞ alone does not reveal the mUAR drop at large ε
- L_∞ is a ubiquitous measure of robustness in deep learning

Limits of adversarial training for ImageNet-UA

L_2 training gives a baseline for ImageNet-UA

	Clean Acc.	L_∞	L_2	mUAR
Normal Training	86.7	7.3	17.2	14.0
L_∞ $\varepsilon = 4$	83.9	72.1	73.6	42.3
L_2 $\varepsilon = 2400$	76.8	79.6	88.5	**50.7**

- L_2 training improves mUAR over L_∞ training
- Substantial improvements against all attacks other than Fog, including very different attacks (Elastic, Gabor, Snow)

A significant gap remains between the best L_2 mUAR of 50.7 and the best UAR values against individual attacks (high 80s+)

Limits of adversarial training for ImageNet-UA

Natural next idea: joint adversarial training

- Simultaneously train against multiple attacks
- Backpropagate with respect to gradients induced by image with greater loss
- Hope this covers more of the "space of distortions"

(L_∞, L_2)-training gives marginal improvement at $2\times$ compute cost:

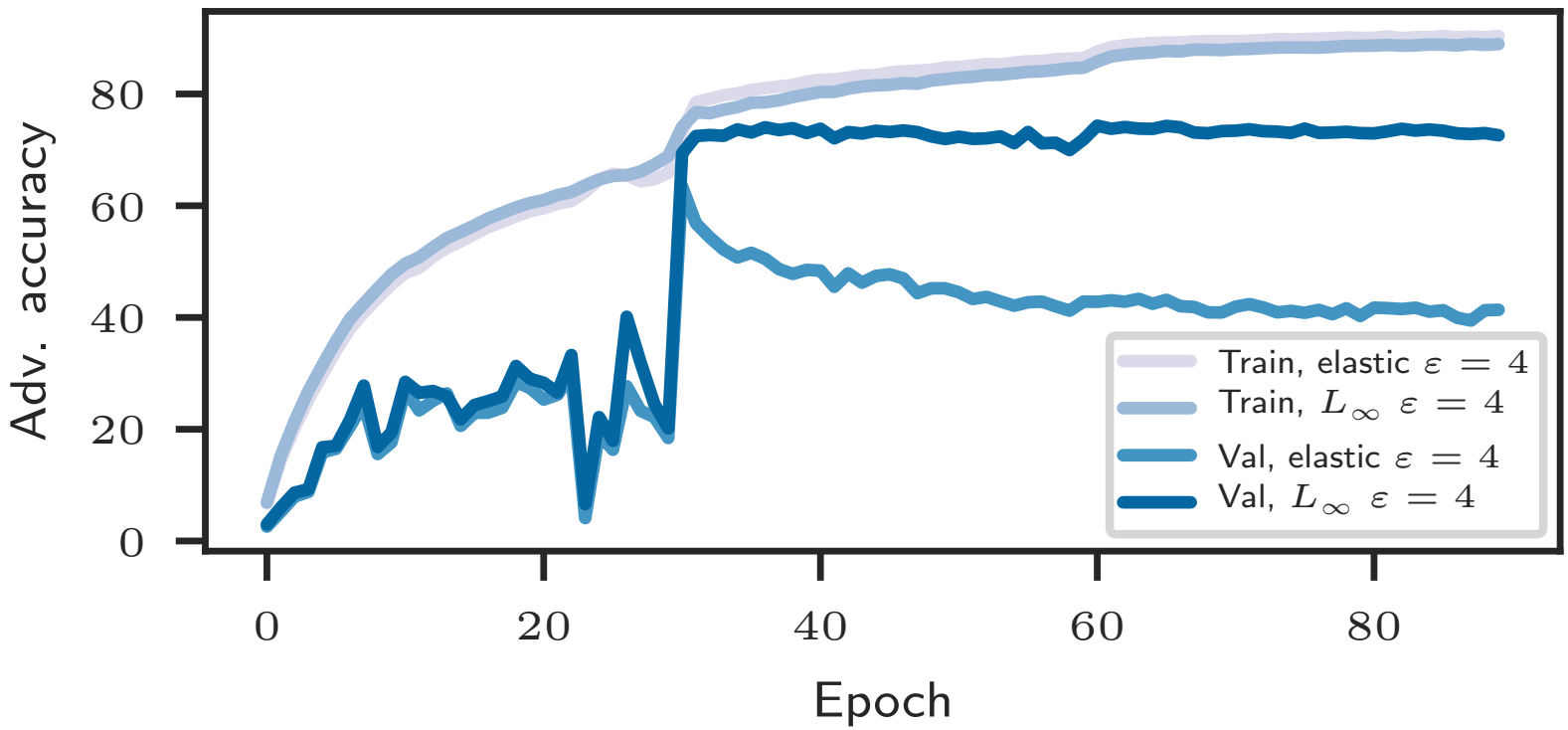
	Clean Acc.	L_∞	L_2	mUAR
Normal Training	86.7	7.3	17.2	14.0
L_2 $\varepsilon = 2400$	76.8	79.6	88.5	50.7
L_∞ $\varepsilon = 16$, L_2 $\varepsilon = 4800$	68.4	81.5	87.9	**50.9**

Joint adversarial training: other attacks

Training parameters	L_∞ train	other train	L_∞ val	other val
$L_\infty\ \varepsilon = 8$, Elastic $\varepsilon = 4$ (1)	90	89	35	74
$L_\infty\ \varepsilon = 8$, Elastic $\varepsilon = 4$ (2)	89	90	47	44
$L_\infty\ \varepsilon = 16, L_1\ \varepsilon = 612000$ (1)	86	87	22	16
$L_\infty\ \varepsilon = 16, L_1\ \varepsilon = 612000$ (2)	88	87	16	24

- Joint adversarial training is **unstable** at higher distortion sizes
- Training runs produce different results with different seeds

Joint adversarial training: overfitting



- **Overfitting** in joint training for conflicting distortions
- Validation accuracy for Elastic **decreases**, while train accuracy increases

ImageNet-UA via non-adversarial defenses

	Clean Acc.	L_∞	L_2	mUAR
AlexNet	82.8	**14.0**	**25.5**	18.5
SqueezeNet	84.1	5.2	11.2	12.8
ResNeXt-101 (32×8d)	95.9	2.5	5.5	13.4
ResNeXt-101 (32×8d) + WSL	**97.1**	3.0	5.7	19.0
ResNet-18	91.6	2.7	8.2	12.0
ResNet-50	94.2	2.7	6.6	13.2
ResNet-50 + Stylized ImageNet	94.6	2.9	7.4	14.6
ResNet-50 + Patch Gaussian	93.6	4.5	10.9	16.2
ResNet-50 + AugMix	95.1	6.1	13.4	**23.2**

Models evaluated by subsetting logits from ImageNet-1000.

- Alternative training methods (weak supervision, data augmentation) improve on mUAR more than L_∞ or L_2
- Surprisingly, vanilla AlexNet is competitive with methods using more data and sophisticated data augmentation

ImageNet-UA via non-adversarial defenses

	Clean Acc.	L_∞	L_2	mUAR
AlexNet	82.8	**14.0**	**25.5**	18.5
SqueezeNet	84.1	5.2	11.2	12.8
ResNeXt-101 (32×8d)	95.9	2.5	5.5	13.4
ResNeXt-101 (32×8d) + WSL	**97.1**	3.0	5.7	19.0
ResNet-18	91.6	2.7	8.2	12.0
ResNet-50	94.2	2.7	6.6	13.2
ResNet-50 + Stylized ImageNet	94.6	2.9	7.4	14.6
ResNet-50 + Patch Gaussian	93.6	4.5	10.9	16.2
ResNet-50 + AugMix	95.1	6.1	13.4	**23.2**

Models evaluated by subsetting logits from ImageNet-1000.

- Robust and clean performance may not relate
- mUAR reveals robustness of data augmentation methods
- mUAR shows robustness of smaller models (SqueezeNet, ResNet-18) is specific to L_∞ and L_2

ImageNet-UA via non-adversarial defenses

	Clean Acc.	L_∞	L_2	mUAR
AlexNet	82.8	**14.0**	**25.5**	18.5
SqueezeNet	84.1	5.2	11.2	12.8
ResNeXt-101 (32×8d)	95.9	2.5	5.5	13.4
ResNeXt-101 (32×8d) + WSL	**97.1**	3.0	5.7	19.0
ResNet-18	91.6	2.7	8.2	12.0
ResNet-50	94.2	2.7	6.6	13.2
ResNet-50 + Stylized ImageNet	94.6	2.9	7.4	14.6
ResNet-50 + Patch Gaussian	93.6	4.5	10.9	16.2
ResNet-50 + AugMix	95.1	6.1	13.4	**23.2**

Models evaluated by subsetting logits from ImageNet-1000.

- Still far away from the mUAR of 50.7 achieved by L_2 training
  - Only AlexNet (2.3) achieves UAR over 1.8 on JPEG
  - Only AugMix (11.1) achieves UAR over 9.0 on Gabor
- Contrast to [Orhan '19], which shows WSL helps tremendously on other robustness benchmarks (ImageNet-C, ImageNet-A)

Summary

New evaluation framework ImageNet-UA enables:

- evaluation beyond L_p with four novel attacks
- a more realistic and broader threat model

Using ImageNet-UA, we find:

- existing evaluation practices using L_∞ are inadequate
- adversarial training establishes a limited but strong baseline
- non-adversarial defenses offer hope for future improvement

Testing robustness against unforeseen adversaries, Daniel Kang*, Yi Sun*, Dan Hendrycks, Tom Brown, and Jacob Steinhardt, arXiv:1908.08016.

Funding: NSF, Simons Foundation, OpenPhil, OpenAI