TESTING ROBUSTNESS AGAINST UNFORESEEN ADVERSARIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Most existing adversarial defenses only measure robustness to L_p adversarial attacks. Not only are adversaries unlikely to exclusively create small L_p perturbations, adversaries are unlikely to remain fixed. Adversaries adapt and evolve their attacks; hence adversarial defenses must be robust to a broad range of *unforeseen attacks*. We address this discrepancy between research and reality by proposing a new evaluation framework called ImageNet-UA. Our framework enables the research community to test ImageNet model robustness against attacks not encountered during training. To create ImageNet-UA's diverse attack suite, we introduce a total of four novel adversarial attacks. We also demonstrate that, in comparison to ImageNet-UA, prevailing L_∞ robustness assessments give a narrow account of adversarial robustness. By evaluating current defenses with ImageNet-UA, we find they provide little robustness to unforeseen attacks. We hope the greater variety and realism of ImageNet-UA enables development of more robust defenses which can generalize beyond attacks seen during training.

1 INTRODUCTION

Neural networks perform well on many datasets (He et al., 2016) yet can be consistently fooled by minor adversarial distortions (Goodfellow et al., 2014). The research community has responded by quantifying and developing adversarial defenses against such attacks (Madry et al., 2017), but these defenses and metrics have two key limitations.

First, the vast majority of existing defenses exclusively defend against and quantify robustness to L_p-constrained attacks (Madry et al., 2017; Cohen et al., 2019; Raff et al., 2019; Xie et al., 2018). Though real-world adversaries are not L_p constrained (Gilmer et al., 2018) and can attack with diverse distortions (Brown et al., 2017; Sharif et al., 2019), the literature largely ignores this and evaluates against the L_p adversaries already seen during training (Madry et al., 2017; Xie et al., 2018), resulting in optimistic robustness assessments. The attacks outside the L_p threat model that have been proposed (Song et al., 2018; Qiu et al., 2019; Engstrom et al., 2017; Evtimov et al., 2017; Sharif et al., 2016) are not intended for general defense evaluation and suffer from narrow dataset applicability, difficulty of optimization, or fragility of auxiliary generative models.

Second, existing defenses assume that attacks are known in advance (Goodfellow, 2019) and use knowledge of their explicit form during training (Madry et al., 2017). In practice, adversaries can deploy *unforeseen attacks* not known to the defense creator. For example, online advertisers use attacks such as perturbed pixels in ads to defeat ad blockers trained only on the previous generation of ads in an ever-escalating arms race (Tramèr et al., 2018). However, current evaluation setups implicitly assume that attacks encountered at test-time are the same as those seen at train-time, which is unrealistic. The reality that future attacks are unlike those encountered during training is akin to a train-test distribution mismatch—a problem studied outside of adversarial robustness (Recht et al., 2019; Hendrycks & Dietterich, 2019)—but now brought to the adversarial setting.

The present work addresses these limitations by proposing an evaluation framework ImageNet-UA to measure robustness against unforeseen attacks. ImageNet-UA assesses a defense which may have been created with knowledge of the commonly used L_∞ or L_2 attacks with six diverse attacks (four of which are novel) distinct from L_∞ or L_2. We intend these attacks to be used at *test-time* only and not during training. Performing well on ImageNet-UA thus demonstrates generalization to a diverse set of distortions not seen during defense creation. While ImageNet-UA

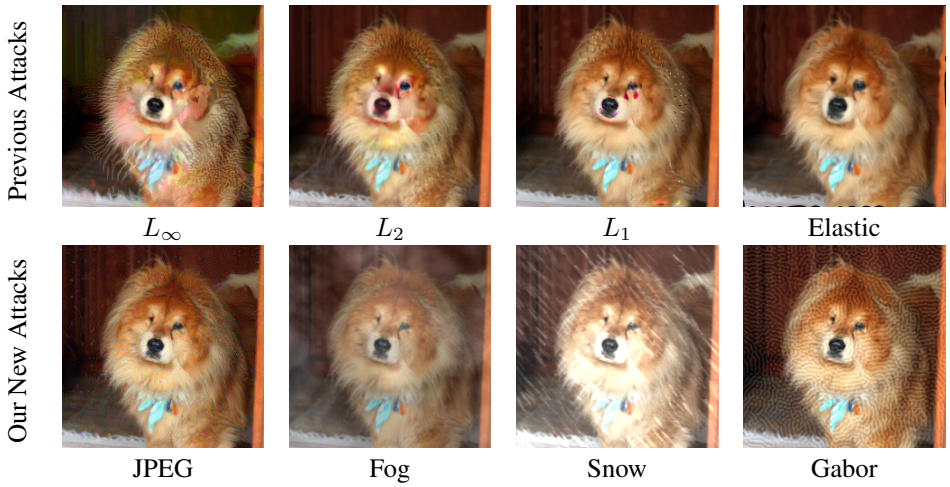


Figure 1: Adversarially distorted chow chow dog images created with old attacks and our new attacks. The JPEG, Fog, Snow, and Gabor adversarial attacks are visually distinct from previous attacks, result in distortions which do not obey a small L_p norm constraint, and serve as unforeseen attacks for the ImageNet-UA attack suite.

does not provide an exhaustive guarantee over all conceivable attacks, it evaluates over a diverse unforeseen test distribution similar to those used successfully in other studies of distributional shift (Rajpurkar et al., 2018; Hendrycks & Dietterich, 2019; Recht et al., 2019). ImageNet-UA works for ImageNet models and can be easily used with our code available at `https://github.com/anon-submission-2020/anon-submission-2020`.

Designing ImageNet-UA requires new attacks that are strong and varied, since real-world attacks are diverse in structure. To meet this challenge, we contribute four novel and diverse adversarial attacks which are easily optimized. Our new attacks produce distortions with occlusions, spatial similarity, and simulated weather, all of which are absent in previous attacks. Performing well on ImageNet-UA thus demonstrates that a defense generalizes to a diverse set of distortions distinct from the commonly used L_∞ or L_2.

With ImageNet-UA, we show weaknesses in existing evaluation practices and defenses through a study of 8 attacks against 48 models adversarially trained on ImageNet-100, a 100-class subset of ImageNet. While most adversarial robustness evaluations use only L_∞ attacks, ImageNet-UA reveals that models with high L_∞ attack robustness can remain susceptible to other attacks. Thus, L_∞ evaluations are a narrow measure of robustness, even though much of the literature treats this evaluation as comprehensive (Madry et al., 2017; Qian & Wegman, 2019; Schott et al., 2019; Zhang et al., 2019). We address this deficiency by using the novel attacks in ImageNet-UA to evaluate robustness to a more diverse set of unforeseen attacks. Our results demonstrate that L_∞ adversarial training, the current state-of-the-art defense, has limited generalization to unforeseen adversaries, and is not easily improved by training against more attacks. This adds to the evidence that achieving robustness against a few train-time attacks is insufficient to impart robustness to unforeseen test-time attacks (Jacobsen et al., 2019; Jordan et al., 2019; Tramèr & Boneh, 2019).

In summary, we propose the framework ImageNet-UA to measure robustness to a diverse set of attacks, made possible by our four new adversarial attacks. Since existing defenses scale poorly to multiple attacks (Jordan et al., 2019; Tramèr & Boneh, 2019), finding defense techniques which generalize to unforeseen attacks is crucial to create robust models. We suggest ImageNet-UA as a way to measure progress towards this goal.

2 Related Work

Adversarial robustness is notoriously difficult to correctly evaluate (Papernot et al., 2017; Athalye et al., 2018a). To that end, Carlini et al. (2019a) provide extensive guidance for sound adversarial robustness evaluation. By measuring attack success rates across several distortion sizes and using a broader threat model with diverse differentiable attacks, ImageNet-UA has several of their recommendations built-in, while greatly expanding the set of attacks over previous work on evaluation.

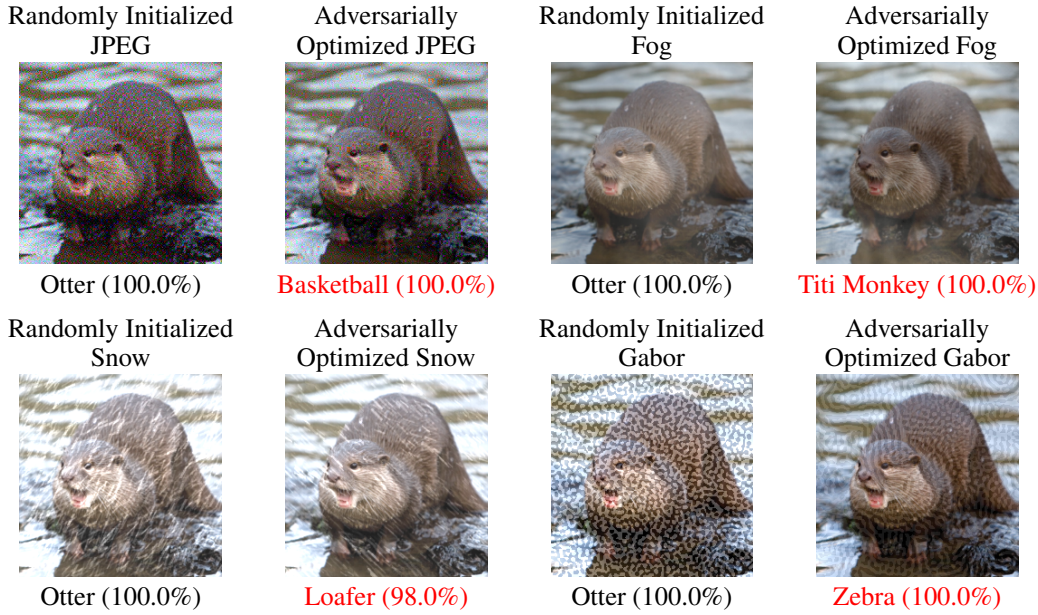


Figure 2: Randomly sampled distortions and adversarially optimized distortions from our new attacks, targeted to the target class in red. Stochastic average-case versions of our attacks affect classifiers minimally, while adversarial versions are optimized to reveal high-confidence errors. The snowflakes in Snow decrease in intensity after optimization, demonstrating that lighter adversarial snowflakes are more effective than heavy random snowfall at uncovering model weaknesses.

A few prior works on evaluate on unforeseen attacks in specific limited circumstances. Wu et al. (2020) evaluate against physically-realizable attacks from Evtimov et al. (2017) and Sharif et al. (2016), though this limits the threat model to occlusion attacks on narrow datasets.

Prior attacks outside the L_p threat model exist, but they lack the general applicability and fast optimization of ours. Song et al. (2018) and Qiu et al. (2019) attack using variational autoencoders and StarGANs, respectively, resulting in weaker attacks which require simple image distributions suitable for VAEs and GANs. Engstrom et al. (2017) apply Euclidean transformations determined by brute-force search. Zhao et al. (2019) use perceptual color distances to align human perception and L_2 perturbations. Evtimov et al. (2017) and Sharif et al. (2016) attack stop signs and face-recognition systems with carefully placed patches or modified eyeglass frames, requiring physical object creation and applying only to specific image types.

3 New Attacks for a Broader Threat Model

There are few diverse, easily optimizable, plug-and-play adversarial attacks in the current literature; outside of Elastic (Xiao et al., 2018), most are L_p attacks such as L_∞ (Goodfellow et al., 2014), L_2 (Szegedy et al., 2013; Carlini & Wagner, 2017), L_1 (Chen et al., 2018). We rectify this deficiency with four novel adversarial attacks: JPEG, Fog, Snow, and Gabor. Our attacks are differentiable and fast, while optimizing over enough parameters to be strong. We show example adversarial images in Figure 1 and compare stochastic and adversarial distortions in Figure 2.

Our novel attacks provide a range of *test-time* adversaries visually and semantically distinct from L_∞ and L_2 attacks. Namely, they cause distortions with large L_∞ and L_2 norm, but result in images that are perceptually close to the original. These attacks are intended as unforeseen attacks not used during training, allowing them to evaluate whether a defense can generalize from L_∞ or L_2 to a more varied set of distortions than current evaluations. Though our attacks are not exhaustive, performing well against them already demonstrates robustness to occlusion, spatial similarity, and simulated weather, which are absent from previous evaluations.

Our attacks create an adversarial image x' from a clean image x with true label y. Let model f map images to a softmax distribution, and let $\ell(f(x), y)$ be the cross-entropy loss. Given a target class $y' \neq y$, our attacks attempt to find a valid image x' such that (1) the attacked image x' is obtained

by applying a distortion (of size controlled by a parameter ε) to x, and (2) the loss $\ell(f(x'), y')$ is minimized. An unforeseen adversarial attack is a white- or black-box adversarial attack unknown to the defense designer which does not change the true label of x according to an oracle or human.

3.1 Four New Unforeseen Attacks

JPEG. JPEG applies perturbations in a JPEG-encoded space of compressed images rather than raw pixel space. After color-space conversion, JPEG encodes small image patches using the discrete cosine transform. It then uses projected gradient descent to find an L_∞-constrained adversarial perturbation in the resulting frequency space. The perturbed frequency coefficients are quantized and reverse-transformed to obtain the image in pixel space. We use ideas from Shin & Song (2017) to make this differentiable. The resulting attack is conspicuously distinct from L_p attacks.

Fog. Fog simulates worst-case weather conditions. Robustness to adverse weather is a safety critical priority for autonomous vehicles, and Figure 2 shows Fog provides a more rigorous stress-test than stochastic fog (Hendrycks & Dietterich, 2019). Fog creates adversarial fog-like occlusions by adversarially optimizing parameters in the diamond-square algorithm (Fournier et al., 1982) typically used to render stochastic fog effects.

Snow. Snow simulates snowfall with occlusions of randomly located small image regions representing snowflakes. It adversarially optimizes their intensity and direction. Making Snow fast and differentiable is non-trivial and hinges on the use of an exponential distribution for snowflake intensities. Compared to synthetic stochastic snow (Hendrycks & Dietterich, 2019), our adversarial snow is faster and includes snowflakes at differing angles. Figure 2 shows adversarial snow exposes model weaknesses more effectively than the easier stochastic, average-case snow.

Gabor. Gabor spatially occludes the image with visually diverse Gabor noise Lagae et al. (2009). Gabor adversarially optimizes semantically meaningful parameters (orientation, bandwidth, etc.) to create different Gabor kernels. While rendering Gabor noise, we use spectral variance normalization (Co et al., 2019) and initialize our optimization parameters with a sparse random matrix.

3.2 Improving Existing Attacks

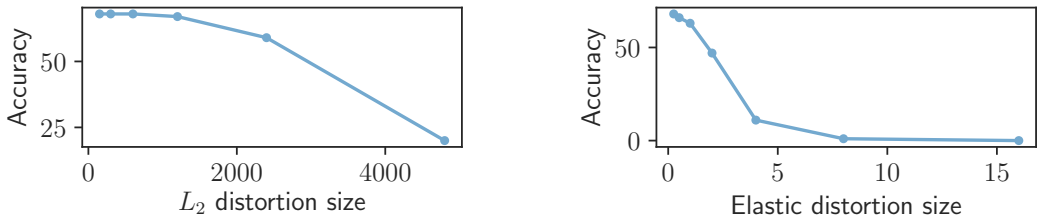
Elastic modifies the attack of Xiao et al. (2018); it warps the image by distortions $x' = \mathsf{Flow}(x, V)$, where $V : \{1, \ldots, 224\}^2 \to \mathbb{R}^2$ is a vector field on pixel space, and Flow sets the value of pixel (i, j) to the bilinearly interpolated original value at $(i, j) + V(i, j)$. We construct V by smoothing a vector field W by a Gaussian kernel (size 25×25, $\sigma \approx 3$ for a 224×224 image) and optimize W under $\|W(i, j)\|_\infty \leq \varepsilon$ for all i, j. The resulting attack is suitable for large-scale images.

The other three attacks are L_1, L_2, L_∞ attacks, but we improve the L_1 attack. For L_∞ and L_2 constraints, we use randomly-initialized projected gradient descent (PGD), which applies gradient descent and projection to the L_∞ and L_2 balls (Madry et al., 2017). Projection is difficult for L_1, and previous L_1 attacks resort to heuristics (Chen et al., 2018; Tramèr & Boneh, 2019). We replace PGD with the Frank-Wolfe algorithm (Frank & Wolfe, 1956), which optimizes a linear function instead of projecting at each step (pseudocode in Appendix D). This makes our L_1 attack more principled than previous implementations.

4 ImageNet-UA: Measuring Robustness to Unforeseen Attacks

We propose the framework ImageNet-UA and its CIFAR-10 analogue CIFAR-10-UA to measure and summarize model robustness while fulfilling the following desiderata: (1) defenses should be evaluated against a broad threat model through a diverse set of attacks, (2) defenses should exhibit generalization to attacks not exactly identical to train-time attacks, and (3) the range of distortion sizes used for an attack must be wide enough to avoid misleading conclusions caused by overly weak or strong versions of that attack (Figure 3).

The ImageNet-UA evaluation framework aggregates robustness information into a single measure, the mean Unforeseen Adversarial Robustness (mUAR). The mUAR is an average over six different attacks of the Unforeseen Adversarial Robustness (UAR), a metric which assesses the robustness of a defense against a specific attack by using a wide range of distortion sizes. UAR is normalized using a measure of attack strength, the ATA, which we now define.



(a) L_2 vs. L_2-training

(b) Elastic vs. L_2-training

Figure 3: Accuracies of L_2 and Elastic attacks at different distortion sizes against a ResNet-50 model adversarially trained against L_2 at $\varepsilon = 9600$ on ImageNet-100. At small distortion sizes, the model appears to defend well against Elastic, but large distortion sizes reveal that robustness does not transfer from L_2 to Elastic.

Adversarial Training Accuracy (ATA). The Adversarial Training Accuracy $\mathsf{ATA}(A, \varepsilon)$ estimates the strength of an attack A against adversarial training (Madry et al., 2017), one of the strongest known defense methods. For a distortion size ε, it is the best adversarial test accuracy against A achieved by adversarial training against A. We allow a possibly different distortion size ε' during training, since this can improves accuracy, and we choose a fixed architecture for each dataset.

For ImageNet-100, we choose ResNet-50 for the architecture, and for CIFAR-10 we choose ResNet-56. When evaluating a defense with architecture other than ResNet-50 or ResNet-56, we recommend using ATA values computed with these architectures to enable consistent comparison. To estimate $\mathsf{ATA}(A, \varepsilon)$ in practice, we evaluate models adversarially trained against distortion size ε' for ε' in a large range (we describe this range at this section's end).

UAR: Robustness Against a Single Attack. The UAR, a building block for the mUAR, averages a model's robustness to a single attack over six distortion sizes $\varepsilon_1, \ldots, \varepsilon_6$ chosen for each attack (we describe the selection procedure at the end of this section). It is defined as

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

where $\mathsf{Acc}(A, \varepsilon_k, M)$ is the accuracy $\mathsf{Acc}(A, \varepsilon_k, M)$ of a model M after attack A at distortion size ε_k. The normalization in (1) makes attacks of different strengths more commensurable in a stable way. We give values of $\mathsf{ATA}(A, \varepsilon_k)$ and ε_k for our attacks on ImageNet-100 and CIFAR-10 in Tables 4 and 5 (Appendix B), allowing computation of UAR of a defense against a single attack with six adversarial evaluations and no adversarial training.

mUAR: Mean Unforeseen Attack Robustness. We summarize a defense's performance on ImageNet-UA with the mean Unforeseen Attack Robustness (mUAR), an average of UAR scores for the L_1, Elastic, JPEG, Fog, Snow, and Gabor attacks:

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

Our measure mUAR estimates robustness to a broad threat model containing six unforeseen attacks at six distortion sizes each, meaning high mUAR requires generalization to several held-out attacks. In particular, it cannot be achieved by the common practice of engineering defenses to a single attack, which Figure 4 shows does not necessarily provide robustness to different attacks.

Our four novel attacks play a crucial role in mUAR by allowing us to estimate robustness to a sufficiently large set of adversarial attacks. As is customary when studying train-test mismatches and distributional shift, we advise against adversarially training with these six attacks when evaluating ImageNet-UA to preserve the validity of mUAR, though we encourage training with *other* attacks.

Distortion Sizes. We explain the ε' values used to estimate ATA and the choice of $\varepsilon_1, \ldots, \varepsilon_6$ used to define UAR. This calibration of distortion sizes adjusts for the fact (Figure 3) that adversarial robustness against an attack may vary drastically with distortion size. Further, the relation between distortion size and attack strength varies between attacks, so too many or too few ε_k values in a certain range may cause an attack to appear artificially strong or weak according to UAR.

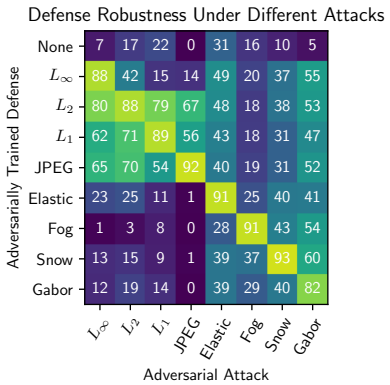


Figure 4: UAR for adv trained defenses (row) against attacks (col) on ImageNet-100. Defenses from L_∞ to Gabor were trained with $\varepsilon = 32$, $4.8k$, $612k$, 2, 16, 8192, 8, and $1.6k$.

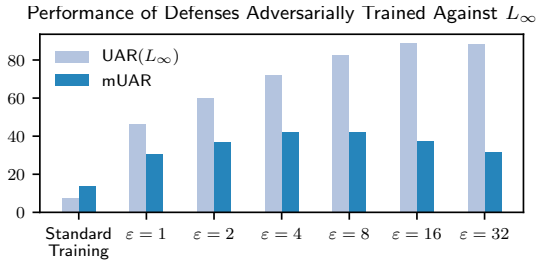


Figure 5: UAR(L_∞) and mUAR for L_∞-trained models at different distortion sizes. Increasing distortion size in L_∞-training improves UAR(L_∞) but hurts the mUAR, suggesting models heavily fit L_∞ at the cost of generalization.

We choose distortion sizes between ε_{min} and ε_{max} as follows. The minimum distortion size ε_{min} is the largest ε for which the adversarial accuracy of an adversarially trained model at distortion size ε is comparable to that of a model trained and evaluated on unattacked data (for ImageNet-100, within 3 of 87). The maximum distortion size ε_{max} is the smallest ε which either reduces adversarial accuracy of an adversarially trained model at distortion size ε below 25 or yields images confusing humans (adversarial accuracy can remain non-zero in this case).

As is typical in recent work on adversarial examples (Athalye et al., 2018b; Evtimov et al., 2017; Dong et al., 2019; Qin et al., 2019), our attacks can be perceptible at large distortion sizes. We make this choice to reflect perceptibility of attacks in real world threat models per Gilmer et al. (2018).

For ATA, we evaluate against models adversarially trained with ε' increasing geometrically from ε_{min} to ε_{max} by factors of 2. We then choose ε_k as follows: We compute ATA at ε increasing geometrically from ε_{min} to ε_{max} by factors of 2 and take the size-6 subset whose ATA values have minimum ℓ_1-distance to the ATA values of the L_∞ attack in Table 4 (Appendix B.1). For example, for Gabor, $(\varepsilon_{\text{min}}, \varepsilon_{\text{max}}) = (6.25, 3200)$, so we compute ATAs at the 10 values $\varepsilon = 6.25, \ldots, 3200$. Viewing size-6 subsets of the ATAs as vectors with decreasing coordinates, we select ε_k for Gabor corresponding to the vector with minimum ℓ_1-distance to the ATA vector for L_∞.

5 New Insights From ImageNet-UA

We use ImageNet-UA to assess existing methods for adversarial defense and evaluation. First, ImageNet-UA reveals that L_∞ trained defenses fail to generalize to different attacks, indicating substantial weakness in current L_∞ adversarial robustness evaluation. We establish a baseline for ImageNet-UA using L_2 adversarial training which is difficult to improve upon by adversarial training alone. Finally, we show non-adversarially trained models can still improve robustness on ImageNet-UA over standard models and suggest this as a direction for further inquiry.

5.1 Experimental Setup

We adversarially train 48 models against the 8 attacks from Section 3 and evaluate against targeted attacks. We use the CIFAR-10 and ImageNet-100 datasets for ImageNet-UA and CIFAR-10-UA. ImageNet-100 is a 100-class subset of ImageNet-1K (Deng et al., 2009) containing every tenth class by WordNet ID order; we use a subset of ImageNet-1K due to the high compute cost of adversarial training. We use ResNet-56 for CIFAR-10 and ResNet-50 from `torchvision` for ImageNet-100 (He et al., 2016). We provide training hyperparameters in Appendix A.

To adversarially train against attack A, at each mini-batch we select a uniform random (incorrect) target class for each training image. For maximum distortion size ε, we apply targeted attack A to the current model with distortion size $\varepsilon' \sim \text{Uniform}(0, \varepsilon)$ and take a SGD step using only the attacked images. Randomly scaling ε' improves performance against smaller distortions.

Table 1: Clean Accuracy, UAR, and mUAR scores for models adv trained against L_∞ and L_2 attacks. L_∞ training, the most popular defense, provides less robustness than L_2 training. Comparing the highest mUAR achieved to individual UAR values in Figure 4 indicates a large robustness gap.

	Clean Accuracy	L_∞	L_2	**mUAR**		Clean Accuracy	L_∞	L_2	**mUAR**
Normal Training	86.7	7.3	17.2	14.0	Normal Training	86.7	7.3	17.2	14.0
L_∞ $\varepsilon = 1$	86.2	46.4	54.2	30.7	L_2 $\varepsilon = 150$	86.6	38.0	49.4	27.1
L_∞ $\varepsilon = 2$	85.5	59.8	64.4	36.9	L_2 $\varepsilon = 300$	85.9	49.7	60.1	33.3
L_∞ $\varepsilon = 4$	83.9	72.1	**73.6**	**42.3**	L_2 $\varepsilon = 600$	84.7	61.9	71.6	40.0
L_∞ $\varepsilon = 8$	79.8	82.6	72.0	42.2	L_2 $\varepsilon = 1200$	82.3	72.9	82.0	46.8
L_∞ $\varepsilon = 16$	74.5	**89.1**	60.0	37.5	L_2 $\varepsilon = 2400$	76.8	79.6	**88.5**	**50.7**
L_∞ $\varepsilon = 32$	70.8	88.1	41.9	31.8	L_2 $\varepsilon = 4800$	68.3	**80.4**	87.7	50.5

Table 2: Clean Accuracy, UAR, and mUAR scores for models jointly trained against (L_∞, L_2). Joint training does not provide much additional robustness.

	Clean Accuracy	L_∞	L_2	**mUAR**
L_∞ $\varepsilon = 1$, L_2 $\varepsilon = 300$	86.1	50.3	60.2	33.6
L_∞ $\varepsilon = 2$, L_2 $\varepsilon = 600$	85.1	62.8	72.5	41.0
L_∞ $\varepsilon = 4$, L_2 $\varepsilon = 1200$	81.3	72.9	81.2	46.9
L_∞ $\varepsilon = 8$, L_2 $\varepsilon = 2400$	76.5	80.0	87.3	50.8
L_∞ $\varepsilon = 16$, L_2 $\varepsilon = 4800$	68.4	**81.5**	**87.9**	**50.9**

We train on 10-step attacks for attacks other than Elastic, where we use 30 steps due to a harder optimization. For L_p, JPEG, and Elastic, we use step size $\varepsilon/\sqrt{\text{steps}}$; for Fog, Gabor, and Snow, we use step size $\sqrt{0.001/\text{steps}}$ because the latent space is independent of ε. These choices have optimal rates for non-smooth convex functions (Nemirovski & Yudin, 1978; 1983). We evaluate on 200-step targeted attacks with uniform random (incorrect) target, using more steps for evaluation than training per best practices (Carlini et al., 2019b).

Figure 4 summarizes ImageNet-100 results. Full results for ImageNet-100 and CIFAR-10 are in Appendix E and robustness checks to random seed and attack iterations are in Appendix F.

5.2 ImageNet-UA Reveals Weaknessess in L_∞ Training and Testing

We use ImageNet-UA to reveal weaknesses in the common practices of L_∞ robustness evaluation and L_∞ adversarial training. We compute the mUAR and UAR(L_∞) for models trained against the L_∞ attack with distortion size ε and show results in Figure 5. For small $\varepsilon \leq 4$, mUAR and UAR(L_∞) increase together with ε. For larger $\varepsilon \geq 8$, UAR(L_∞) continues to increase with ε, but the mUAR decreases, a fact which is not apparent from L_∞ evaluation.

The decrease in mUAR while UAR(L_∞) increases suggests that L_∞ adversarial training begins to heavily fit L_∞ distortions at the expense of generalization at larger distortion sizes. Thus, while it is the most commonly used defense procedure, L_∞ training may not lead to improvements on other attacks or to real-world robustness.

Worse, L_∞ evaluation against L_∞ adversarial training at higher distortions indicates higher robustness. In contrast, mUAR reveals that L_∞ adversarial training at higher distortions in fact hurts robustness against a more diverse set of attacks. Thus, L_∞ evaluation gives a misleading picture of robustness. This is particularly important because L_∞ evaluation is the most ubiquitous measure of robustness in deep learning (Goodfellow et al., 2014; Madry et al., 2017; Xie et al., 2018).

5.3 Limits of Adversarial Training for ImageNet-UA

We establish a baseline on ImageNet-UA using L_2 adversarial training but show a significant performance gap even for more sophisticated existing adversarial training methods. To do so, we evaluate several adversarial training methods on ImageNet-UA and show results in Table 1.

Our results show that L_2 trained models outperform L_∞ trained models and have significantly improved absolute performance, increasing mUAR from 14.0 to 50.7 compared to an undefended model. The individual UAR values in Figure 7 (Appendix E.1) improve substantially against all attacks other than Fog, including several (Elastic, Gabor, Snow) of extremely different nature to L_2.

Table 3: Non-adversarial defenses can noticeably improve ImageNet-UA performance. ResNeXt-101 (32×8d) + WSL is trained on approximately 1 billion images Mahajan et al. (2018). Stylized ImageNet is trained on a modification of ImageNet using style transfer Geirhos et al. (2019). Patch Gaussian augments using Gaussian distortions on small portions of the image Lopes et al. (2019). AugMix mixes simple random augmentations of the image Hendrycks et al. (2020). These results suggest that ImageNet-UA performance may be achieved through non-adversarial defenses.

	Clean Acc.	L_∞	L_2	L_1	Elastic	JPEG	Fog	Snow	Gabor	**mUAR**
SqueezeNet	84.1	5.2	11.2	14.9	25.9	**1.9**	20.1	9.8	4.4	12.8
ResNeXt-101 (32×8d)	95.9	2.5	5.5	20.7	26.5	1.8	14.1	12.4	5.3	13.4
ResNeXt-101 (32×8d) + WSL	**97.1**	3.0	5.7	28.3	29.4	**1.9**	26.2	20.3	8.0	19.0
ResNet-18	91.6	2.7	8.2	13.5	22.6	1.8	20.3	9.5	4.2	12.0
ResNet-50	94.2	2.7	6.6	20.1	24.9	1.8	15.8	11.9	4.9	13.2
ResNet-50 + Stylized ImageNet	94.6	2.9	7.4	22.8	26.0	1.8	16.2	12.5	8.1	14.6
ResNet-50 + Patch Gaussian	93.6	4.5	10.9	27.4	28.2	1.8	23.9	10.5	5.2	16.2
ResNet-50 + AugMix	95.1	**6.1**	**13.4**	**34.3**	**38.8**	1.8	**28.6**	**24.7**	**11.1**	**23.2**

This result suggests pushing adversarial training further by training against multiple attacks simultaneously via *joint adversarial training* (Jordan et al., 2019; Tramèr & Boneh, 2019) detailed in Appendix C. Table 2 shows that, despite using twice the compute of L_2 training, (L_∞, L_2) joint training only improves the mUAR from 50.7 to 50.9. We thus recommend L_2 training as a baseline for ImageNet-UA, though there is substantial room for improvement compared to the highest UARs against individual attacks in Figure 4, which are all above 80 and often above 90.

5.4 ImageNet-UA Robustness through Non-Adversarial Defenses

We find that methods can improve robustness to unforeseen attacks without adversarial training. Table 3 shows mUAR for SqueezeNet (Iandola et al., 2017), ResNeXts (Xie et al., 2016), and ResNets. For ImageNet-1K models, we mask 900 logits to predict ImageNet-100 classes.

A popular defense against average case distortions (Hendrycks & Dietterich, 2019) is Stylized ImageNet (Geirhos et al., 2019), which modifies training images using image style transfer in hopes of making networks rely less on textural features. Table 3 shows it provides some improvement on ImageNet-UA. More recently, Lopes et al. (2019) propose to train against Gaussian noise applied to small image patches, improving the mUAR by 3% over the ResNet-50 baseline. The second largest mUAR improvement comes from training a ResNeXt on approximately 1 billion images (Mahajan et al., 2018). This three orders of magnitude increase in training data yields a 5.4% mUAR increase over a vanilla ResNeXt baseline. Finally, Hendrycks et al. (2020) create AugMix, which randomly mixes stochastically generated augmentations. Although AugMix did not use random nor adversarial noise, it improves robustness to unforeseen attacks by 10%.

These results imply that defenses not relying on adversarial examples can improve ImageNet-UA performance. They indicate that training on more data only somewhat increases robustness on ImageNet-UA, unlike many other robustness benchmarks (Hendrycks & Dietterich, 2019; Hendrycks et al., 2019) where more data helps tremendously (Orhan, 2019). While models with lower clean accuracy (e.g., SqueezeNet and ResNet-18) have higher UAR(L_∞) and UAR(L_2) than many other models, there is no clear difference in mUAR. Last, these non-adversarial defenses have minimal cost to accuracy on clean examples, unlike adversarial defenses. Much remains to explore, and we hope non-adversarial defenses will be a promising avenue toward adversarial robustness.

6 Conclusion

This work proposes a framework ImageNet-UA to evaluate robustness of a defense against *unforeseen attacks*. Because existing adversarial defense techniques do not scale to multiple attacks, developing models which can defend against attacks not seen at train-time is essential for robustness. Our results using ImageNet-UA show that the common practice of L_∞ training and evaluation fails to achieve or measure this broader form of robustness. As a result, it can provide a misleading sense of robustness. By incorporating our 4 novel and strong adversarial attacks, ImageNet-UA enables evaluation on the diverse held-out attacks necessary to measure progress towards robustness more broadly.

REFERENCES

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *arXiv preprint arXiv:1802.00420*, 2018a.

Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 284–293, Stockholmsmssan, Stockholm Sweden, 10–15 Jul 2018b. PMLR. URL http://proceedings.mlr.press/v80/athalye18b.html.

Tom B. Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch. *CoRR*, abs/1712.09665, 2017. URL http://arxiv.org/abs/1712.09665.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 39–57. IEEE, 2017.

Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian G Goodfellow, and Aleksander Madry. On evaluating adversarial robustness: Principles of rigorous evaluations. 2019a.

Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian J. Goodfellow, Aleksander Madry, and Alexey Kurakin. On evaluating adversarial robustness. *CoRR*, abs/1902.06705, 2019b. URL http://arxiv.org/abs/1902.06705.

Pin-Yu Chen, Yash Sharma, Huan Zhang, Jinfeng Yi, and Cho-Jui Hsieh. EAD: Elastic-net attacks to deep neural networks via adversarial examples. In *Thirty-second AAAI conference on artificial intelligence*, 2018.

Kenneth T. Co, Luis Muñoz-González, and Emil C. Lupu. Sensitivity of deep convolutional networks to Gabor noise. *CoRR*, abs/1906.03455, 2019. URL http://arxiv.org/abs/1906.03455.

Jeremy M. Cohen, Elan Rosenfeld, and J. Zico Kolter. Certified adversarial robustness via randomized smoothing. *CoRR*, abs/1902.02918, 2019. URL http://arxiv.org/abs/1902.02918.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. IEEE, 2009.

Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019.

Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. A rotation and a translation suffice: Fooling CNNs with simple transformations. *arXiv preprint arXiv:1712.02779*, 2017.

Ivan Evtimov, Kevin Eykholt, Earlence Fernandes, Tadayoshi Kohno, Bo Li, Atul Prakash, Amir Rahmati, and Dawn Xiaodong Song. Robust physical-world attacks on deep learning models. 2017.

Alain Fournier, Don Fussell, and Loren Carpenter. Computer rendering of stochastic models. *Commun. ACM*, 25(6):371–384, June 1982. ISSN 0001-0782. doi: 10.1145/358523.358553. URL http://doi.acm.org/10.1145/358523.358553.

Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2):95–110, 1956.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bygh9j09KX.

Justin Gilmer, Ryan P. Adams, Ian J. Goodfellow, David Andersen, and George E. Dahl. Motivating the rules of the game for adversarial example research. *ArXiv*, abs/1807.06732, 2018.

Ian J. Goodfellow. A research agenda: Dynamic models to defend against correlated attacks. *ArXiv*, abs/1903.06293, 2019.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch SGD: Training ImageNet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pp. 630–645. Springer, 2016.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. *arXiv preprint arXiv:1907.07174*, 2019.

Dan Hendrycks, Norman Mu, Ekin D. Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. AugMix: A simple data processing method to improve robustness and uncertainty. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.

Forrest N. Iandola, Matthew W. Moskewicz, Khalid Ashraf, Song Han, William J. Dally, and Kurt Keutzer. Squeezenet: AlexNet-level accuracy with 50x fewer parameters and $<$1mb model size. *ArXiv*, abs/1602.07360, 2017.

Jrn-Henrik Jacobsen, Jens Behrmannn, Nicholas Carlini, Florian Tramr, and Nicolas Papernot. Exploiting excessive invariance caused by norm-bounded adversarial robustness, 2019.

Matt Jordan, Naren Manoj, Surbhi Goel, and Alexandros G. Dimakis. Quantifying perceptual distortion of adversarial examples. *arXiv e-prints*, art. arXiv:1902.08265, Feb 2019.

Ares Lagae, Sylvain Lefebvre, George Drettakis, and Philip Dutré. Procedural noise using sparse Gabor convolution. *ACM Trans. Graph.*, 28(3):54:1–54:10, July 2009. ISSN 0730-0301. doi: 10.1145/1531326.1531360. URL http://doi.acm.org/10.1145/1531326.1531360.

Raphael Gontijo Lopes, Dong Yin, Ben Poole, Justin Gilmer, and Ekin Dogus Cubuk. Improving robustness without sacrificing accuracy with patch gaussian augmentation. *ArXiv*, abs/1906.02611, 2019.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. Exploring the limits of weakly supervised pretraining. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (eds.), *Computer Vision – ECCV 2018*, pp. 185–201, Cham, 2018. Springer International Publishing. ISBN 978-3-030-01216-8.

Arkadi Nemirovski and D Yudin. On Cezari's convergence of the steepest descent method for approximating saddle point of convex-concave functions. In *Soviet Math. Dokl*, volume 19, pp. 258–269, 1978.

Arkadi Nemirovski and D Yudin. *Problem Complexity and Method Efficiency in Optimization*. Intersci. Ser. Discrete Math. Wiley, New York, 1983.

A. Emin Orhan. Robustness properties of Facebook's ResNeXt WSL models. *ArXiv*, abs/1907.07640, 2019.

Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pp. 506–519. ACM, 2017.

Haifeng Qian and Mark N. Wegman. L_2-nonexpansive neural networks. In *International Conference on Learning Representations (ICLR)*, 2019. URL https://openreview.net/forum?id=ByxGSsR9FQ.

Chongli Qin, James Martens, Sven Gowal, Dilip Krishnan, Krishnamurthy Dvijotham, Alhussein Fawzi, Soham De, Robert Stanforth, and Pushmeet Kohli. Adversarial robustness through local linearization, 2019.

Haonan Qiu, Chaowei Xiao, Lei Yang, Xinchen Yan, Honglak Lee, and Bo Li. Semanticadv: Generating adversarial examples via attribute-conditional image editing. *ArXiv*, abs/1906.07927, 2019.

Edward Raff, Jared Sylvester, Steven Forsyth, and Mark McLean. Barrage of random transforms for adversarially robust defense. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6528–6537, 2019.

P. Rajpurkar, R. Jia, and P. Liang. Know what you don't know: Unanswerable questions for SQuAD. In *Association for Computational Linguistics (ACL)*, 2018.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *ICML*, 2019.

L. Schott, J. Rauber, W. Brendel, and M. Bethge. Towards the first adversarially robust neural network model on MNIST. May 2019. URL https://arxiv.org/pdf/1805.09190.pdf.

Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K. Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 23rd ACM SIGSAC Conference on Computer and Communications Security*, 2016.

Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K Reiter. A general framework for adversarial examples with objectives. *ACM Transactions on Privacy and Security (TOPS)*, 22(3): 1–30, 2019.

Richard Shin and Dawn Song. JPEG-resistant adversarial images. In *NIPS 2017 Workshop on Machine Learning and Computer Security*, 2017.

Yang Song, Rui Shu, Nate Kushman, and Stefano Ermon. Constructing unrestricted adversarial examples with generative models. In *NeurIPS*, 2018.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Florian Tramèr and Dan Boneh. Adversarial training and robustness for multiple perturbations. *arXiv e-prints*, art. arXiv:1904.13000, Apr 2019.

Florian Tramèr, Pascal Dupré, Gili Rusak, Giancarlo Pellegrino, and Dan Boneh. Ad-versarial: Defeating perceptual ad-blocking. *CoRR*, abs/1811.03194, 2018. URL http://arxiv.org/abs/1811.03194.

Tong Wu, Liang Tong, and Yevgeniy Vorobeychik. Defending against physically realizable attacks on image classification. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=H1xscnEKDr.

Chaowei Xiao, Jun-Yan Zhu, Bo Li, Warren He, Mingyan Liu, and Dawn Song. Spatially transformed adversarial examples. *arXiv preprint arXiv:1801.02612*, 2018.

Cihang Xie, Yuxin Wu, Laurens van der Maaten, Alan Yuille, and Kaiming He. Feature denoising for improving adversarial robustness. *arXiv preprint arXiv:1812.03411*, 2018.

Saining Xie, Ross B. Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5987–5995, 2016.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7472–7482, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL `http://proceedings.mlr.press/v97/zhang19p.html`.

Zhengyu Zhao, Zhuoran Liu, and Marisa Larson. Towards large yet imperceptible adversarial image perturbations with perceptual color distance. *ArXiv*, abs/1911.02466, 2019.

A TRAINING HYPERPARAMETERS

For ImageNet-100, we trained on machines with 8 NVIDIA V100 GPUs using standard data augmentation (He et al., 2016). Following best practices for multi-GPU training (Goyal et al., 2017), we ran synchronized SGD for 90 epochs with batch size 32×8 and a learning rate schedule with 5 "warm-up" epochs and a decay at epochs 30, 60, and 80 by a factor of 10. Initial learning rate after warm-up was 0.1, momentum was 0.9, and weight decay was 10^{-4}. For CIFAR-10, we trained on a single NVIDIA V100 GPU for 200 epochs with batch size 32, initial learning rate 0.1, momentum 0.9, and weight decay 10^{-4}. We decayed the learning rate at epochs 100 and 150.

B CALIBRATION OF ImageNet-UA AND CIFAR-10-UA

B.1 CALIBRATION FOR ImageNet-UA

Calibrated distortion sizes and ATA values are in Table 4.

B.2 CALIBRATION FOR CIFAR-10-UA

The ε calibration procedure for CIFAR-10 was similar to that used for ImageNet-100. We started with small ε_{min} values and increased ε geometrically with ratio 2 until adversarial accuracy of an adversarially trained model dropped below 40. Note that this threshold is higher for CIFAR-10 than ImageNet-100 because there are fewer classes. The resulting ATA values for CIFAR-10 are shown in Table 5.

C JOINT ADVERSARIAL TRAINING

Our joint adversarial training procedure for two attacks A and A' is as follows. At each training step, we compute the attacked image under both A and A' and backpropagate with respect to gradients induced by the image with greater loss. This corresponds to the "max" loss of Tramèr & Boneh (2019). We train ResNet-50 models for (L_∞, L_2), (L_∞, L_1), and $(L_\infty, \text{Elastic})$ on ImageNet-100.

Table 6 shows training against (L_∞, L_1) is worse than training against L_1 at the same distortion size and performs particularly poorly at large distortion sizes. Table 7 shows joint training against

Table 4: Calibrated distortion sizes and ATA values for different distortion types on ImageNet-100.

Attack	ε_1	ε_2	ε_3	ε_4	ε_5	ε_6	ATA_1	ATA_2	ATA_3	ATA_4	ATA_5	ATA_6
L_∞	1	2	4	8	16	32	84.6	82.1	76.2	66.9	40.1	12.9
L_2	150	300	600	1200	2400	4800	85.0	83.5	79.6	72.6	59.1	19.9
L_1	9562.5	19125	76500	153000	306000	612000	84.4	82.7	76.3	68.9	56.4	36.1
Elastic	0.25	0.5	2	4	8	16	85.9	83.2	78.1	75.6	57.0	22.5
JPEG	0.062	0.125	0.250	0.500	1	2	85.0	83.2	79.3	72.8	34.8	1.1
Fog	128	256	512	2048	4096	8192	85.8	83.8	79.0	68.4	67.9	64.7
Snow	0.0625	0.125	0.25	2	4	8	84.0	81.1	77.7	65.6	59.5	41.2
Gabor	6.25	12.5	25	400	800	1600	84.0	79.8	79.8	66.2	44.7	14.6

Table 5: Calibrated distortion sizes and ATA values for ResNet-56 on CIFAR-10

Attack	ε_1	ε_2	ε_3	ε_4	ε_5	ε_6	ATA_1	ATA_2	ATA_3	ATA_4	ATA_5	ATA_6
L_∞	1	2	4	8	16	32	91.0	87.8	81.6	71.3	46.5	23.1
L_2	40	80	160	320	640	2560	90.1	86.4	79.6	67.3	49.9	17.3
L_1	195	390	780	1560	6240	24960	92.2	90.0	83.2	73.8	47.4	35.3
JPEG	0.03125	0.0625	0.125	0.25	0.5	1	89.7	87.0	83.1	78.6	69.7	35.4
Elastic	0.125	0.25	0.5	1	2	8	87.4	81.3	72.1	58.2	45.4	27.8

Table 6: UAR scores for L_1-trained models and (L_∞, L_1)-jointly trained models. At each distortion size, L_1-training performs better than joint training.

	UAR_{L_∞}	UAR_{L_1}
L_∞ $\varepsilon = 2$, L_1 $\varepsilon = 76500$	48	66
L_∞ $\varepsilon = 4$, L_1 $\varepsilon = 153000$	**51**	**72**
L_∞ $\varepsilon = 8$, L_1 $\varepsilon = 306000$	44	62
L_1 $\varepsilon = 76500$	50	70
L_1 $\varepsilon = 153000$	54	81
L_1 $\varepsilon = 306000$	**59**	**87**

Table 7: UAR scores for L_∞- and Elastic-trained models and (L_∞, Elastic)-jointly trained models. No jointly trained model matches a Elastic-trained model on UAR vs. Elastic.

	UAR_{L_∞}	$\text{UAR}_{\text{Elastic}}$
L_∞ $\varepsilon = 4$, Elastic $\varepsilon = 2$	68	63
L_∞ $\varepsilon = 8$, Elastic $\varepsilon = 4$	35	**65**
L_∞ $\varepsilon = 16$, Elastic $\varepsilon = 8$	**69**	43
Elastic $\varepsilon = 2$	**37**	68
Elastic $\varepsilon = 4$	36	81
Elastic $\varepsilon = 8$	31	**91**

$(L_\infty, \text{Elastic})$ also performs poorly, never matching the UAR score of training against Elastic at moderate distortion size ($\varepsilon = 2$).

D THE FRANK-WOLFE ALGORITHM

We chose to use the Frank-Wolfe algorithm for optimizing the L_1 attack, as Projected Gradient Descent would require projecting onto a truncated L_1 ball, which is a complicated operation. In contrast, Frank-Wolfe only requires optimizing linear functions $g^\top x$ over a truncated L_1 ball; this can be done by sorting coordinates by the magnitude of g and moving the top k coordinates to the boundary of their range (with k chosen by binary search). This is detailed in Algorithm 1.

E FULL EVALUATION RESULTS

E.1 FULL EVALUATION RESULTS AND ANALYSIS FOR IMAGENET-100

We show the full results of all adversarial attacks against all adversarial defenses for ImageNet-100 in Figure 6. These results also include L_1-JPEG and L_2-JPEG attacks, which are modifications of the JPEG attack applying L_p-constraints in the compressed JPEG space instead of L_∞ constraints. Full UAR scores are provided for ImageNet-100 in Figure 7.

E.2 FULL EVALUATION RESULTS AND ANALYSIS FOR CIFAR-10

We show the results of adversarial attacks and defenses for CIFAR-10 in Figure 8. We experienced difficulty training the L_2 and L_1 attacks at distortion sizes greater than those shown and have omitted those runs, which we believe may be related to the small size of CIFAR-10 images. Full UAR values for CIFAR-10 are shown in Figure 9.

F ROBUSTNESS OF OUR RESULTS

F.1 REPLICATION

We replicated our results for the first three rows of Figure 6 with different random seeds to see the variation in our results. As shown in Figure 10, deviations in results are minor.

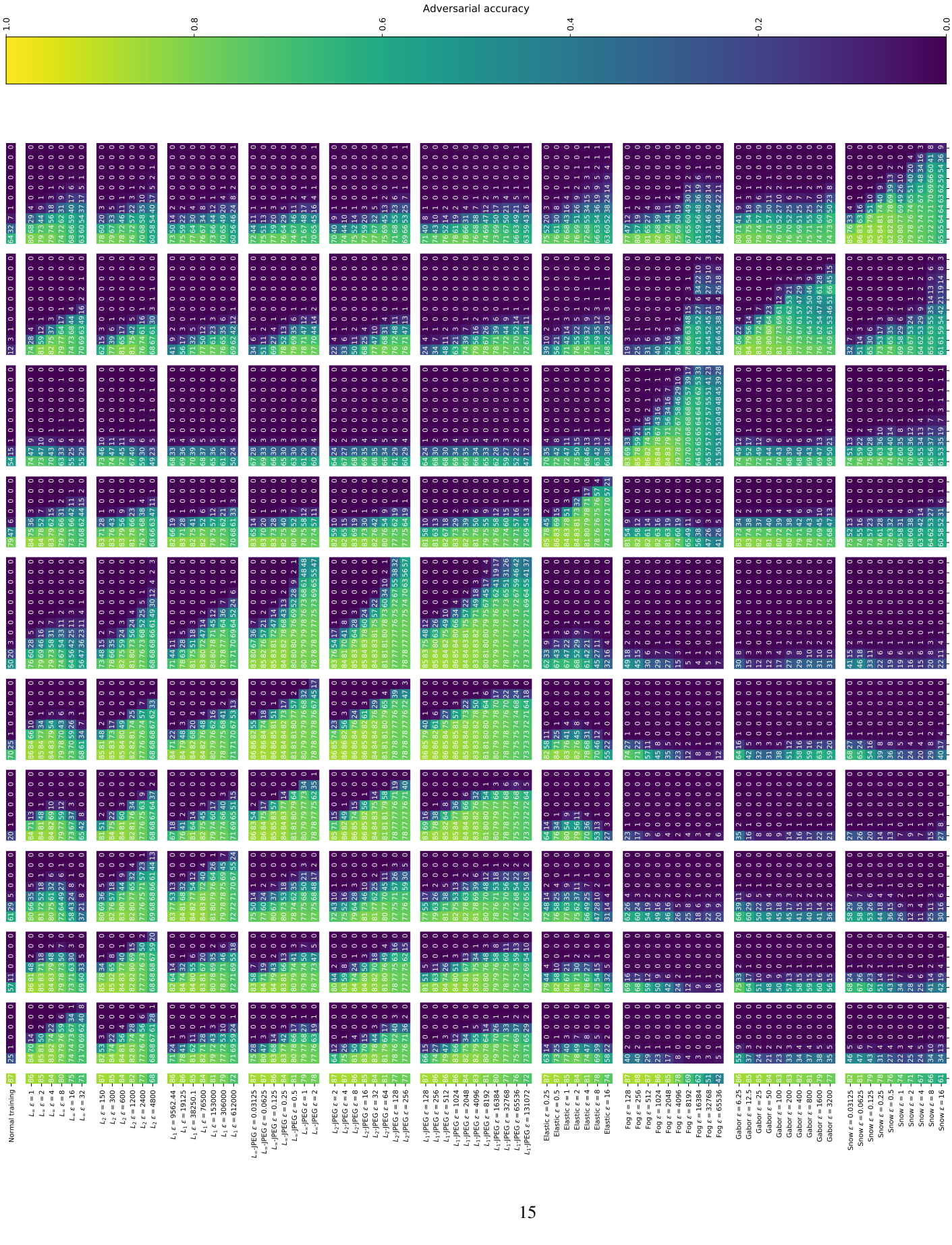


Figure 6: Accuracy of adversarial attack (column) against adversarially trained model (row) on ImageNet-100.

Algorithm 1 Pseudocode for the Frank-Wolfe algorithm for the L_1 attack.

Input: function f, initial input $x \in [0,1]^d$, L_1 radius ρ, number of steps T.
Output: approximate maximizer $\bar{x}$ of f over the truncated L_1 ball $B_1(\rho; x) \cap [0,1]^d$ centered at x.

$x^{(0)} \leftarrow \mathrm{RandomInit}(x)$ {Random initialization}
for $t = 1, \ldots, T$ **do**
  $g \leftarrow \nabla f(x^{(t-1)})$ {Obtain gradient}
  for $k = 1, \ldots, d$ **do**
    $s_k \leftarrow$ index of the coordinate of g by with k^{th} largest norm
  end for
  $S_k \leftarrow \{s_1, \ldots, s_k\}$.

  {Compute move to boundary of $[0,1]$ for each coordinate.}
  for $i = 1, \ldots, d$ **do**
    if $g_i > 0$ **then**
      $b_i \leftarrow 1 - x_i$
    else
      $b_i \leftarrow -x_i$
    end if
  end for
  $M_k \leftarrow \sum_{i \in S_k} |b_i|$ {Compute L_1-perturbation of moving k largest coordinates.}
  $k^* \leftarrow \max\{k \mid M_k \leq \rho\}$ {Choose largest k satisfying L_1 constraint.}

  {Compute $\hat{x}$ maximizing $g^\top x$ over the L_1 ball.}
  for $i = 1, \ldots, d$ **do**
    if $i \in S_{k^*}$ **then**
      $\hat{x}_i \leftarrow x_i + b_i$
    else if $i = s_{k^*+1}$ **then**
      $\hat{x}_i \leftarrow x_i + (\rho - M_{k^*}) \operatorname{sign}(g_i)$
    else
      $\hat{x}_i \leftarrow x_i$
    end if
  end for
  $x^{(t)} \leftarrow (1 - \frac{1}{t}) x^{(t-1)} + \frac{1}{t} \hat{x}$ {Average $\hat{x}$ with previous iterates}
end for
$\bar{x} \leftarrow x^{(T)}$

F.2 Convergence

We replicated the results in Figure 6 with 50 instead of 200 steps to see how the results changed based on the number of steps in the attack. As shown in Figure 11, the deviations are minor.

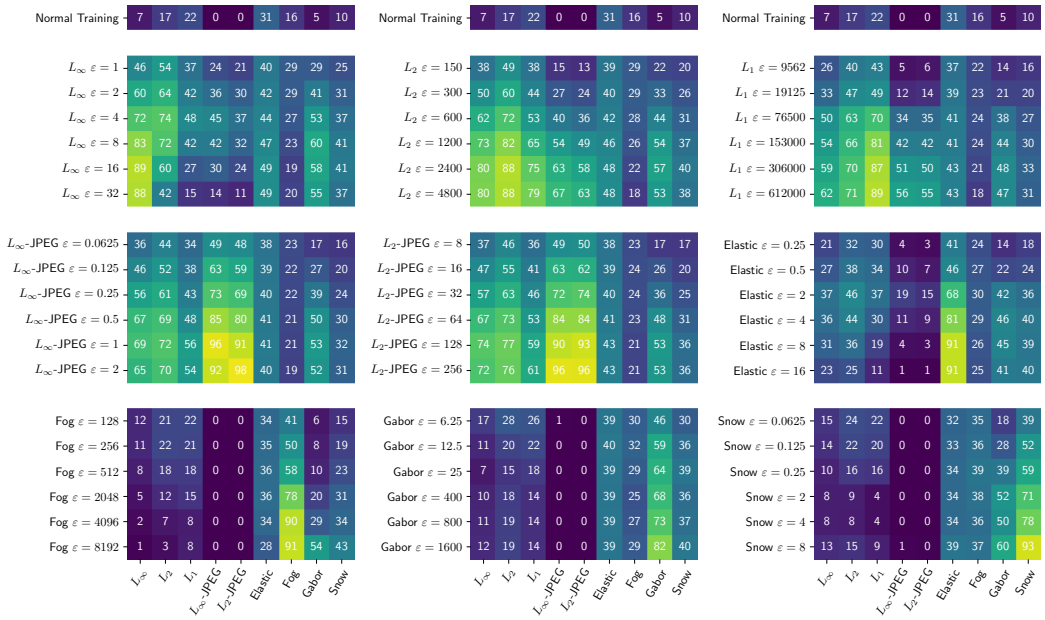


Figure 7: UAR scores for adv. trained defenses (rows) against distortion types (columns) for ImageNet-100.

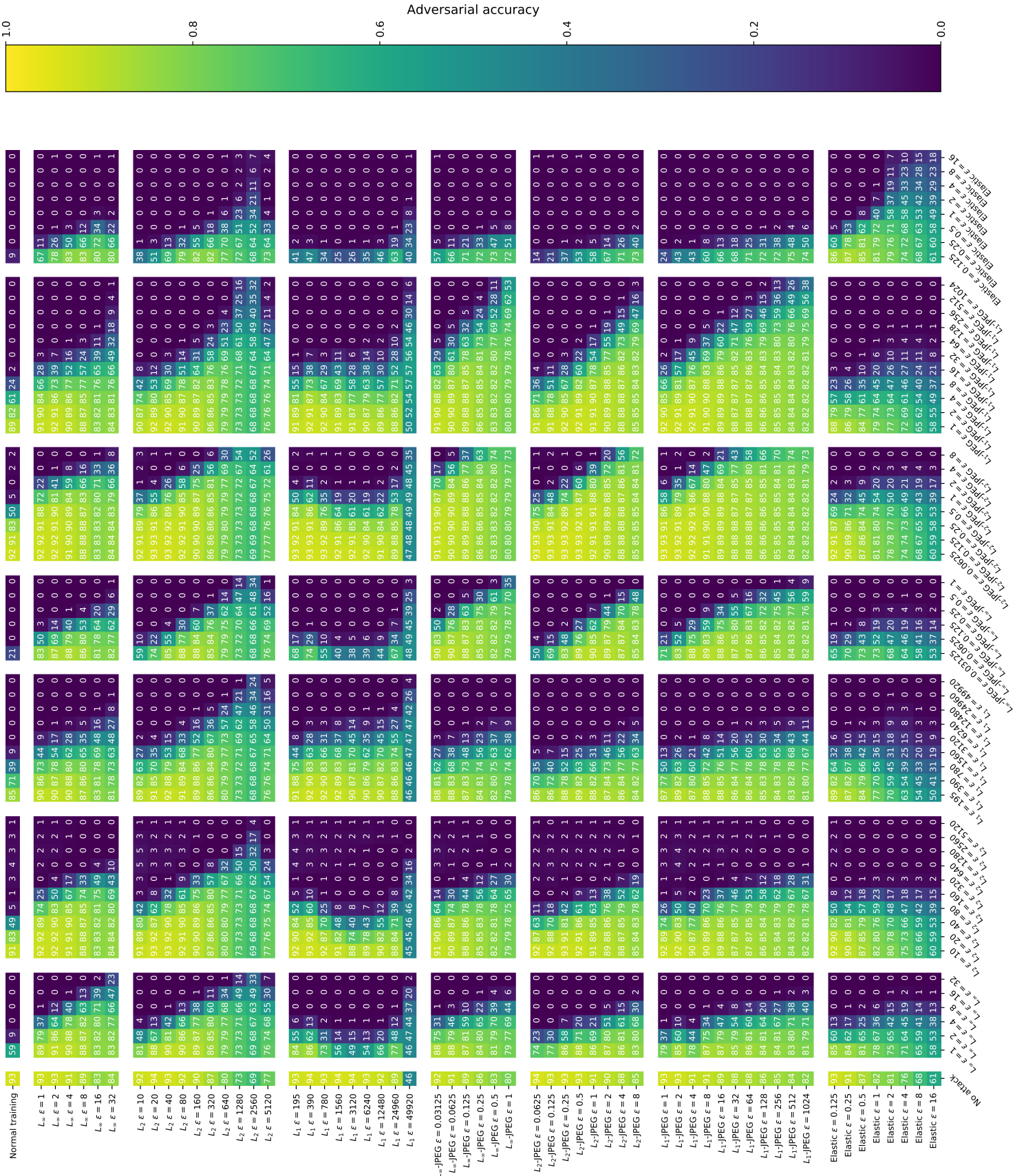


Figure 8: Accuracy of adversarial attack (column) against adversarially trained model (row) on CIFAR-10.

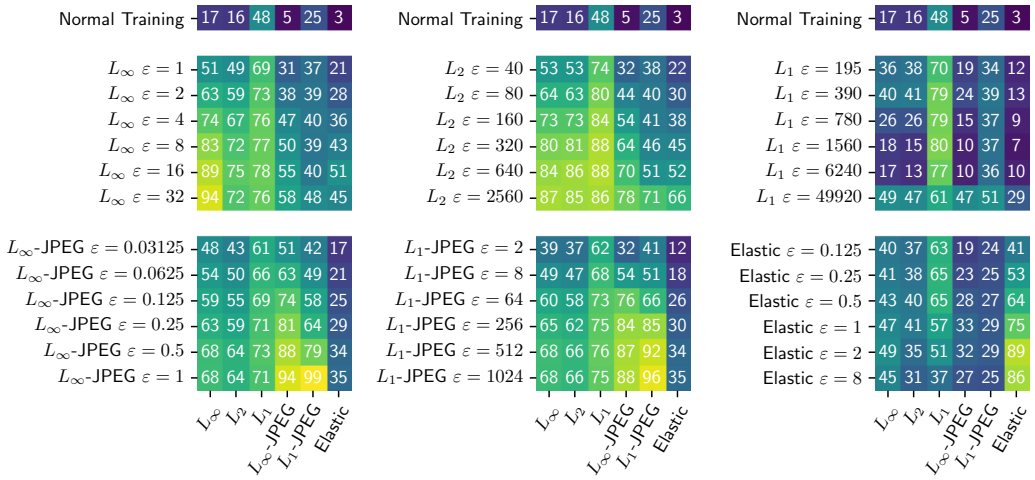


Figure 9: UAR scores on CIFAR-10. Displayed UAR scores are multiplied by 100 for clarity.

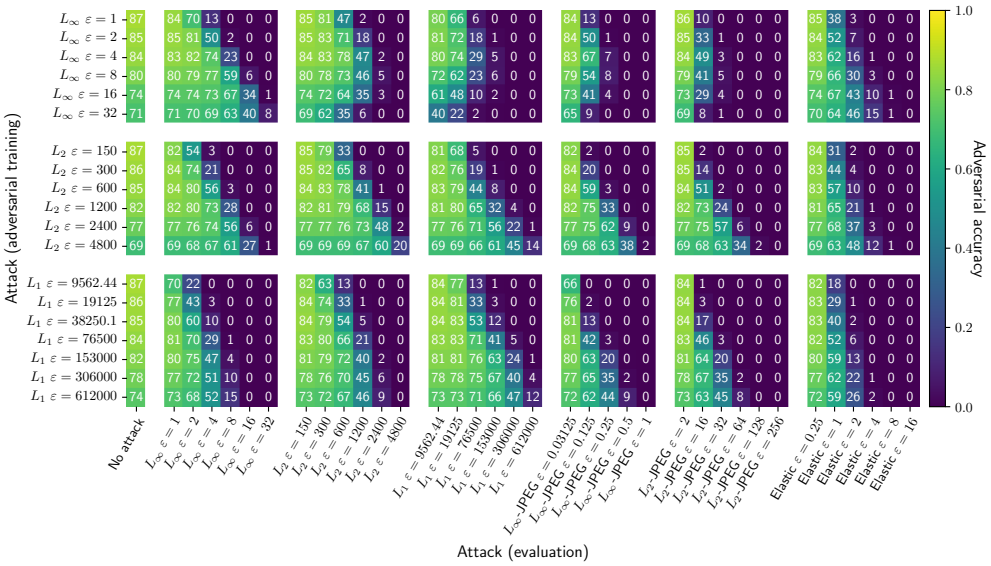


Figure 10: Replica of the first three block rows of Figure 6 with different random seeds. Deviations in results are minor.

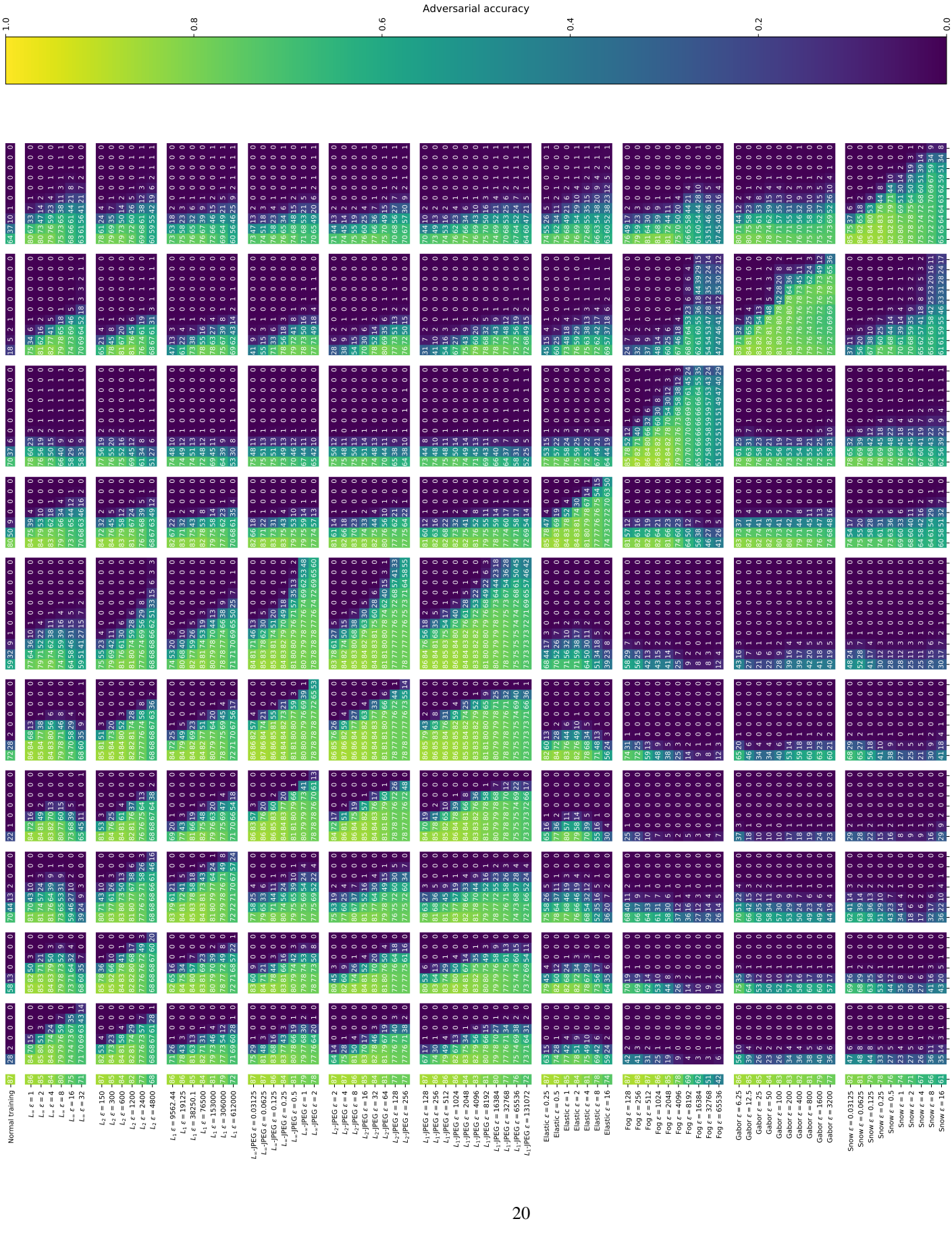


Figure 11: Replica of Figure 6 with 50 steps instead of 200 at evaluation time. Deviations in results are minor.