Fundamental Tradeoffs between Invariance and Sensitivity to Adversarial Perturbations

Florian Tramèr 1  Jens Behrmann 2  Nicholas Carlini 3  Nicolas Papernot 3  Jörn-Henrik Jacobsen 4

Abstract

Adversarial examples are malicious inputs crafted to induce misclassification. Commonly studied sensitivity-based adversarial examples introduce semantically-small changes to an input that result in a different model prediction. This paper studies a complementary failure mode, invariance-based adversarial examples, that introduce minimal semantic changes that modify an input’s true label yet preserve the model’s prediction. We demonstrate fundamental tradeoffs between these two types of adversarial examples. We show that defenses against sensitivity-based attacks actively harm a model’s accuracy on invariance-based attacks, and that new approaches are needed to resist both attack types. In particular, we break state-of-the-art adversarially-trained and certifiably-robust models by generating small perturbations that the models are (provably) robust to, yet that change an input’s class according to human labelers. Finally, we formally show that the existence of excessively invariant classifiers arises from the presence of overly-robust predictive features in standard datasets.

1. Introduction

Research on adversarial examples (Szegedy et al., 2013; Biggio et al., 2013) is motivated by a spectrum of questions, ranging from the security of models in adversarial settings (Tramèr et al., 2019), to limitations of learned representations under natural distribution shift (Gilmer et al., 2018). The broadest accepted definition of an adversarial example is “an input to a ML model that is intentionally designed by an attacker to fool the model into producing an incorrect output” (Goodfellow & Papernot, 2017).

Adversarial examples are commonly formalized as inputs obtained by adding some perturbation to test examples to change the model output. We refer to this class of malicious inputs as sensitivity-based adversarial examples.

To enable concrete progress, the adversary’s capabilities are typically constrained by bounding the size of the perturbation added to the original input. The goal of this constraint is to ensure that semantics of the input are left unaffected by the perturbation. In the computer vision domain, \( \ell_p \) norms have grown to be a default metric to measure semantic similarity. This led to a series of proposals for increasing the robustness of models to sensitivity-based adversaries that operate within the constraints of an \( \ell_p \) ball (Madry et al., 2017; Wong & Kolter, 2018; Raghunathan et al., 2018).

In this paper, we show that optimizing a model’s robustness to \( \ell_p \)-bounded perturbations is not only insufficient to resist general adversarial examples, but also potentially harmful. As \( \ell_p \) distances are a crude approximation to the visual similarity in a given task, over-optimizing a model’s robustness to \( \ell_p \)-bounded perturbations renders the model excessively invariant to real semantics of the underlying task.

Excessive invariance of a model causes vulnerability against invariance adversarial examples (Jacobsen et al., 2019). These are perturbations that change the human-assigned label of a given input but keep the model prediction unchanged. For example in Figure 1 an MNIST image of a digit ‘3’ is perturbed to be an image of a ‘5’ by changing only 20 pixels; models that are excessively invariant do not change their decision and incorrectly label both images as a ‘3’, despite the fact that the oracle label has changed.

This paper exposes a fundamental tradeoff between sensitivity-based and invariance-based adversarial examples. We show that due to a misalignment between formal robustness notions (e.g., \( \ell_p \)-balls) and a task’s perceptual metric, current defenses against adversarial examples cannot prevent both sensitivity-based and invariance-based attacks, and must trade-off robustness to each (see Figure 2). Worse, we find that increasing robustness to sensitivity-based attacks decreases a model’s robustness to invariance-based attacks. We introduce new algorithms to craft \( \ell_p \)-bounded invariance-based adversarial examples, and illustrate the
We further break a

The fundamental tradeoff described in this paper will affect other

vision tasks once we can train strongly robust models on them.

models have been made robust in non-negligible

eXamples, we chose it because it is the

formally and empirically that data augmentation may offer

that such

datasets aim to capture. Our experiments on MNIST show

2019) and to

features (Jo & Bengio, 2017; Ilyas et al., 2019; Yin et al.,

between sensitivity and invariance can be studied rigorously.

Finally, we introduce a classification task where the tradeoff

Oracle Decision-boundary

“robust” decision boundary

“unrobust” decision boundary

Figure 1: Decision boundaries near a real image of a digit ‘3’ and an invariance-based adversarial example labeled as ‘5’.

[Left]: Training a classifier without constraints may learn a decision boundary unrobust to sensitivity-based adversarial examples. [Right]: Enforcing robustness to norm-bounded perturbations introduces erroneous invariance (dashed regions in ε-spheres). We display real data here, the misclassified ‘5’ is an image found by our attack which resides within a typically reported ε-region around the displayed ‘3’ (in the ℓ∞ norm). This excessive invariance of the robust model in task-relevant directions illustrates how robustness to sensitivity-based adversarial examples can result in new model vulnerabilities.

above tradeoff on MNIST.1 We show that state-of-the-art

robust models disagree with human labelers on many of our
crafted invariance-based examples, and that the disagree-

ment rate is higher the more robust a model is. We find

that even models robust to very small perturbations (e.g.,
of ℓ∞-norm below ε = 0.1) have higher vulnerability to
invariance attacks compared to undefended models.

We further break a provably-robust defense (Zhang et al.,
2019) with our attack. This model is certified to have 87% test-accuracy (with respect to the MNIST test-labels) under ℓ∞-perturbations of radius ε = 0.4. That is, for 87% of test inputs (x, y), the model is guaranteed to predict class y for any perturbed input x′ that satisfies ∥x − x′∥∞ ≤ 0.4. Yet, on our invariance-based adversarial examples that satisfy this norm-bound, the model only agrees with human labelers in 60% of the cases for an automated attack, and 12% of the cases for manually-created examples—i.e., no better than chance. The reason is that we can find perturbed inputs x′ that humans no longer classify the same way as x.

Code to reproduce our attacks is available at https://github.com/ftramer/Excessive-Invariance.

Finally, we introduce a classification task where the tradeoff

between sensitivity and invariance can be studied rigorously.

We show that state-of-the-art

robust models disagree with human labelers on many of our
crafted invariance-based examples, and that the disagree-

ment rate is higher the more robust a model is. We find

that even models robust to very small perturbations (e.g.,
of ℓ∞-norm below ε = 0.1) have higher vulnerability to
invariance attacks compared to undefended models.

We further break a provably-robust defense (Zhang et al.,
2019) with our attack. This model is certified to have 87% test-accuracy (with respect to the MNIST test-labels) under ℓ∞-perturbations of radius ε = 0.4. That is, for 87% of test inputs (x, y), the model is guaranteed to predict class y for any perturbed input x′ that satisfies ∥x − x′∥∞ ≤ 0.4. Yet, on our invariance-based adversarial examples that satisfy this norm-bound, the model only agrees with human labelers in 60% of the cases for an automated attack, and 12% of the cases for manually-created examples—i.e., no better than chance. The reason is that we can find perturbed inputs x′ that humans no longer classify the same way as x.

Code to reproduce our attacks is available at https://github.com/ftramer/Excessive-Invariance.

2. Norm-bounded Sensitivity and Invariance

Adversarial examples

We begin by defining a framework to formally describe two

complementary failure modes of machine learning mod-
els, namely (norm-bounded) adversarial examples that arise

from excessive sensitivity or invariance of a classifier.

We consider a classification task with data (x, y) ∈ Rd ×

{1, ..., C} from a distribution D. We assume the existence of a labeling oracle O : Rd → {1, ..., C} ∪ {⊥} that maps any input in Rd to its true label, or to the “garbage class” ⊥ for inputs x considered “un-labelable” (e.g., for a digit classification task, the oracle O corresponds to human-labeling of any image as a digit or as the garbage class). Note that for (x, y) ∼ D, we always have y = O(x).2

The goal of robust classification is to learn a classifier f : Rd → {1, ..., C} that agrees with the oracle’s labels not only in expectation over the distribution D, but also on any rare or out-of-distribution inputs to which the oracle assigns a class label—including adversarial examples obtained by imperceptibly perturbing inputs sampled from D.

At its broadest, the definition of an adversarial example

encompasses any adversarially induced failure in a classi-

er (Goodfellow & Papernot, 2017). That is, an adversarial

example is any input x∗ created such that f(x∗) ̸= O(x∗).

This definition has proven difficult to work with, due to its

inherent reliance on the oracle O. As a result, it has become

customary to study a relaxation of this definition, which re-

stricts the adversary to applying a “small” perturbation to an

input x sampled from the distribution D. A common choice

1While MNIST can be a poor choice for studying adversarial

eXamples, we chose it because it is the only vision task for which

models have been made robust in non-negligible ℓp norm balls.

The fundamental tradeoff described in this paper will affect other

vision tasks once we can train strongly robust models on them.

2We view the support of D as a strict subset of all inputs in Rd
to which the oracle assigns a label. That is, there are inputs for

which humans agree on a label, yet that have measure zero in the
data distribution from which the classifier’s train and test inputs
are chosen. For example, the train-test data is often a sanitized and
normalized subset of natural inputs. Moreover, “unnatural” inputs
such as adversarial examples might never arise in natural data.
We study a complementary failure mode to sensitivity adversarial examples: the well-separated failure modes of a classifier (i.e., invariance adversarial examples) that arise between these two types of adversarial examples, that arise from this assumption being violated. We demonstrate that state-of-the-art robust classifiers do violate this assumption, and (sometimes catastrophically) have low robust error $\mathcal{L}_\varepsilon(f)$ for a norm-bound $\varepsilon$ that does not guarantee that the oracle’s label is preserved. We show that these classifiers actually have high “true” robust error as measured by human labelers.

**Remarks.** Definition 2 is a conscious restriction on a definition of Jacobsen et al. (2019), who define an invariance adversarial example as an *unbounded* perturbation that changes the oracle’s label while preserving a classifier’s output at an intermediate feature layer. As we solely consider the model’s final classification, considering unbounded perturbations would allow for a “trivial” attack: given an input $x$ of class $y$, find any input of a different class that the model misclassifies as $y$. (e.g., given an image of a digit 8, an unbounded invariance example could be any unperturbed digit that the classifier happens to misclassify as an 8).

Definition 2 presents the same difficulty as the original broad definition of adversarial examples: a dependence on the oracle $O$. Automating the process of finding invariance adversarial examples is thus challenging. In Section 4.2, we present some successful automated attacks, but show that a human-in-the-loop process is more effective.

### 3. The Sensitivity and Invariance Tradeoff

In this section, we show that if the norm that is used to define “small” adversarial perturbations is misaligned with the labeling oracle $O$, then the robust classification objective in equation 1 is insufficient for preventing both sensitivity-based and invariance-based adversarial examples under that norm. That is, we show that optimizing a model to attain low robust error on perturbations of norm $\varepsilon$ cannot prevent both sensitivity and invariance adversarial examples.

We begin by formalizing our notion of norm-oracle misalignment. The definition applies to any similarity metric over inputs, of which $\ell_p$-norms are a special case.

**Definition 3 (Distance-Oracle Misalignment).** Let $\text{dist} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a distance measure (e.g., $\|x_1 - x_2\|$). We say that $\text{dist}$ is aligned with the oracle $O$ if for any input $x$ with $O(x) = y$, and any inputs $x_1, x_2$ such that $O(x_1) = y$, $O(x_2) \neq y$, we have $\text{dist}(x, x_1) < \text{dist}(x, x_2)$. $\text{dist}$ and $O$ are misaligned if they are not aligned.

For natural images, $\ell_p$-norms (or other simple metrics) are clearly misaligned with our own perceptual metric. A concrete example is in Figure 3. This simple fact has deep
We now empirically demonstrate and evaluate the trade-off between sensitivity-based and invariance-based adversarial examples. We propose an algorithm for generating invariance adversarial examples, and show that robustified models are disparately more vulnerable to these attacks compared to standard models. In particular, we break both adversarially-trained and certifiably-robust models on MNIST by generating invariance adversarial examples—within the models’ (possibly certified) norm bound—to which the models assign different labels than an ensemble of humans.

4. Generating Invariance-based Adversarial Examples on MNIST

We now empirically demonstrate and evaluate the trade-off between sensitivity-based and invariance-based adversarial examples. We propose an algorithm for generating invariance adversarial examples, and show that robustified models are disparately more vulnerable to these attacks compared to standard models. In particular, we break both adversarially-trained and certifiably-robust models on MNIST by generating invariance adversarial examples—within the models’ (possibly certified) norm bound—to which the models assign different labels than an ensemble of humans.

**Algorithm 1** Meta-algorithm for finding invariance-based adversarial examples. For an input $x$, we find an input $x^*$ of a different class in the dataset $\mathcal{X}$, that is closest to $x$ under some set of semantics-preserving transformations. $\mathcal{T}$.

\[
\textbf{GenInv} (x, y, \mathcal{X}, \mathcal{T}) \\
\begin{align*}
\mathcal{S} &= \{ \tilde{x} : (\tilde{x}, \tilde{y}) \in \mathcal{X}, \tilde{y} \neq y \} \\
\mathcal{X}^* &= \{ t(\tilde{x}) : t \in \mathcal{T}, \tilde{x} \in \mathcal{S} \} \\
\text{return } x^* &= \arg \min_{\tilde{x} \in \mathcal{X}^*} \| \tilde{x} - x \| 
\end{align*}
\]

**Why MNIST?** We elect to study MNIST, the only dataset for which strong robustness to various $\ell_p$-bounded perturbations is attainable with current techniques (Madry et al., 2017; Schott et al., 2019). The dataset’s simplicity is what initially prompted the study of simple $\ell_p$-bounded perturbations (Goodfellow et al., 2015). Increasing MNIST models’ robustness to such perturbations has since become a standard benchmark (Schott et al., 2019; Madry et al., 2017; Wong & Kolter, 2018; Raghunathan et al., 2018). Due to the existence of models with high robustness to various $\ell_p$-bounded attacks, robust classification on MNIST is considered close to solved (Schott et al., 2019).

This paper argues that, contrary to popular belief, MNIST is far from being solved. We show why optimizing for robustness to $\ell_p$-bounded adversaries is not only insufficient, but actively harms the performance of the classifier against alternative invariance-based attacks.

In Section 4.3, we show that complex vision tasks (e.g., ImageNet) are also affected by the fundamental tradeoffs we describe. These tradeoffs are simply not apparent yet, because of our inability to train models with non-negligible robustness to any attacks on these tasks.

4.1. Generating Model-agnostic Invariance-based Adversarial Examples

We propose a model-agnostic algorithm for crafting invariance adversarial examples. Our attack generates minimally perturbed invariance adversarial examples that cause humans to change their classification. We then evaluate these examples against multiple models. The rationale for this approach is mainly that obtaining human labels is expensive, which encourages the use of a single attack for all models. The high-level algorithm we use is in Algorithm 1 and
Figure 4: Process for generating $\ell_0$ invariant adversarial examples. From left to right: (a) the original image of an 8; (b) the nearest training image (labeled as 3), before alignment; (c) the nearest training image (still labeled as 3), after alignment; (d) the $\Delta$ perturbation between the original and aligned training example; (e) spectral clustering of the perturbation $\Delta$; and (f-h) candidate invariance adversarial examples, selected by applying subsets of clusters of $\Delta$ to the original image. (f) is a failed attempt at an invariance adversarial example. (g) is successful, but introduces a larger perturbation than necessary (adding pixels to the bottom of the 3). (h) is successful and minimally perturbed.

described below. It is simple, albeit tailored to datasets where comparing images in pixel space is meaningful, like MNIST.\(^3\)

Given an input $x$, the attack’s goal is to find the smallest class-changing perturbation $x^* = x + \Delta$ (c.f. equation 2) such that $O(x^*) \neq O(x)$. Typically, $x^*$ is not a part of the dataset. We thus approximate $x^*$ via semantics-preserving transformations of other inputs. That is, for the set $S$ of inputs of a different class than $x$, we apply transformations $T$ (e.g., small image rotations, translations) that are known a-priori to preserve input labels. We then pick the transformed input that is closest to our target point under the considered $\ell_p$ metric. In Appendix A, we describe instantiations of this algorithm for the $\ell_0$ and $\ell_\infty$ norms. Figure 4 visualizes the sub-steps for the $\ell_0$ attack, including an extra post-processing that further reduces the perturbation size.

**Measuring Attack Success.** We refer to an invariance adversarial example as *successful* if it causes a change in the oracle’s label, i.e., $O(x^*) \neq O(x)$. This is a model-agnostic version of Definition 2. In practice, we simulate the oracle by asking an ensemble of humans to label the point $x^*$; if more than some fraction of them agree on the label (throughout this section, 70%) and that label is different from the original, the attack is successful. Note that success or failure is independent of any machine learning model.

### 4.2. Evaluation

**Attack analysis.** We generate 100 invariance adversarial examples on inputs randomly drawn from the MNIST test set, for both the $\ell_0$ and $\ell_\infty$ norms. Our attack is slow, with the alignment process taking (amortized) minutes per

\(^3\)Kaushik et al. (2019) consider a similar problem for NLP tasks. They ask human labelers to produce “counterfactually-augmented data” by introducing a minimal number of changes to a text document so as to change the document’s semantics.

<table>
<thead>
<tr>
<th>Attack Type</th>
<th>Success Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clean Images</td>
<td>0%</td>
</tr>
<tr>
<td>$\ell_0$ Attack</td>
<td>55%</td>
</tr>
<tr>
<td>$\ell_\infty$, $\epsilon = 0.3$ Attack</td>
<td>21%</td>
</tr>
<tr>
<td>$\ell_\infty$, $\epsilon = 0.3$ Attack (manual)</td>
<td>26%</td>
</tr>
<tr>
<td>$\ell_\infty$, $\epsilon = 0.4$ Attack</td>
<td>37%</td>
</tr>
<tr>
<td>$\ell_\infty$, $\epsilon = 0.4$ Attack (manual)</td>
<td>88%</td>
</tr>
</tbody>
</table>

Table 1: Success rate of our invariance adversarial examples in causing humans to switch their classification.

Figure 5: Our invariance-based adversarial examples. Top to bottom: original images and our $\ell_0$, $\ell_\infty$ at $\epsilon = 0.3$ and $\ell_\infty$ at $\epsilon = 0.4$ invariance adversarial examples. (left) successful attacks; (right) failed attack attempts.

**Human Study.** We conducted a human study to evaluate whether our invariance adversarial examples are indeed successful, i.e., whether humans agree that the label has been changed. We also hand-crafted 50 invariance adversarial examples for the $\ell_0$ and $\ell_\infty$ norm. The process was quite simple: we built an image editor that lets us change images at a pixel level under an $\ell_p$ constraint. One author then modified 50 random test examples in the way that they perceived as changing the underlying class. We presented all these invariance examples to 40 human evaluators. Each evaluator classified 100 digits, half of which were unmodified MNIST digits, and the other half were sampled randomly from our $\ell_0$ and $\ell_\infty$ invariance adversarial examples.

**Results.** Of 100 clean (unmodified) test images, 98 are labeled identically by all human evaluators. The other 2 images were labeled identically by over 90% of evaluators.

Our $\ell_0$ attack is highly effective: For 55 of the 100 examples at least 70% of human evaluators labeled it the same way, with a different label than the original test label. Humans only agreed with the original test label (with the same 70% threshold) on 34 of the images, while they did not form a
Fundamental Tradeoffs between Invariance and Sensitivity to Adversarial Perturbations

Table 2: Model accuracy with respect to the oracle human labelers on the subset of examples where the human-obtained oracle label is different from the test label. Models which are more robust to perturbation adversarial examples (such as those trained with adversarial training) tend to agree with humans less often on invariance-based adversarial examples. Values denoted with an asterisks * violate the perturbation threat model of the defense and should not be taken to be attacks. When the model is wrong, it failed to classify the input as the new oracle label.

<table>
<thead>
<tr>
<th>Model</th>
<th>Undefended</th>
<th>l0 Sparse</th>
<th>Binary-ABS</th>
<th>ABS</th>
<th>l∞ PGD (ε = 0.3)</th>
<th>l2 PGD (ε = 2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clean</td>
<td>99%</td>
<td>99%</td>
<td>99%</td>
<td>99%</td>
<td>99%</td>
<td>99%</td>
</tr>
<tr>
<td>l0</td>
<td>80%</td>
<td>38%</td>
<td>47%</td>
<td>58%</td>
<td>56%*</td>
<td>27%*</td>
</tr>
<tr>
<td>l∞, ε = 0.3</td>
<td>33%</td>
<td>19%*</td>
<td>0%</td>
<td>14%</td>
<td>0%</td>
<td>5%*</td>
</tr>
<tr>
<td>l∞, ε = 0.4</td>
<td>51%</td>
<td>27%*</td>
<td>8%</td>
<td>18%</td>
<td>16%*</td>
<td>19%*</td>
</tr>
</tbody>
</table>

1 l0 Sparse: (Bafna et al., 2018); ABS and Binary-ABS: (Schott et al., 2019); l∞ PGD and l2 PGD: (Madry et al., 2017)

Consensus on 18 examples. The simpler $l_\infty$ attack is less effective: with a distortion of 0.3 the oracle label changed 21% of the time and with 0.4 the oracle label changed 37% of the time. The manually created $l_\infty$ examples with distortion of 0.4 were highly effective however: for 88% of the examples, at least 70% assigned the same label (different than the test set label). We summarize results in Table 1. In Figure 5 we show sample invariance adversarial examples.

Our attack code, as well as our invariance examples and their human-assigned labels are available at https://github.com/framer/Excessive-Invariance.

To simplify the analysis below, we split our generated invariance adversarial examples into two sets: the successes and the failures, as determined by whether the plurality decision by humans was different than or equal to the original label. We only evaluate models on those invariance adversarial examples that caused the humans to switch their classification.

**Model Evaluation.** Given oracle ground-truth labels for each of the images (as decided by humans), we report how often models agree with the human-assigned label. Table 2 summarizes this analysis. For the invariance adversarial examples, we report model accuracy only on successful attacks (i.e., those where the human oracle label changed between the original image and the modified image). For these same models, Table 3 in Appendix C reports the “standard” robust accuracy for sensitivity-based adversarial examples, i.e., in the sense of equation 1.

The models which empirically achieve the highest robustness against $l_0$ perturbations (in the sense of equation 1) are the $l_0$-Sparse classifier of Bafna et al. (2018), the Binary-ABS model of Schott et al. (2019), and the $l_2$-PGD adversarially trained model (see Table 3 in the Appendix for a comparison of the robustness of these models). Thus, these are the models that are most invariant to perturbations of large $l_0$-norm. We find that these are the models that achieve the lowest accuracy—as measured by the human labelers—on our invariance examples. Moreover, all robust models perform much worse than an undefended ResNet-18 model on our invariance attacks. This includes models such as the $l_\infty$-PGD adversarially trained model, which do not explicitly aim at worst-case robustness against $l_0$ noise. Thus, we find that models that were designed to reduce excessive sensitivity to certain non-semantic features, become excessively invariant to other features that are semantically meaningful.

Similarly, we find that models designed for $l_\infty$ robustness (Binary-ABS and $l_\infty$-PGD) also fare the worst on our $l_\infty$ invariance adversarial examples. Overall, all robust models do worse than the undefended baseline. The results are consistent for attacks with $\epsilon = 0.3$ and with $\epsilon = 0.4$, the latter being more successful in changing human labels.

Note that the Binary-ABS defense of (Schott et al., 2019) boasts 60% (empirical) robust accuracy on $l_\infty$-attacks with $\epsilon = 0.4$ (see (Schott et al., 2019)). Yet, on our invariance examples that satisfy this perturbation bound, the model actually disagrees with the human labelers 92% of the time, and thus achieves only 8% true accuracy on these examples. Below, we make a similar observation for a certified defense.

**Trading Perturbation-Robustness for Invariance-Robustness.** To better understand how robustness to sensitivity-based adversarial examples influences robustness to invariance attacks, we evaluate a range of adversarially-trained models on our invariance examples.

Specifically, we trained $l_\infty$-PGD models with $\epsilon \in [0, 0.4]$ and $l_1$-PGD models (as a proxy for $l_0$-robustness) with $\epsilon \in [0, 15]$. We verified that training against larger perturbations resulted in a monotonic increase in adversarial
While our experiments are on MNIST, similar phenomena thus, while the proof in the paper is very little robustness to PGD with vulnerability to invariance-based examples.

Conversely, the middle image was semantically changed, imperceptible and thus classifiers should be robust to it. However, when we look at the agreement between this model’s predictions with the new labels assigned to perturbations of size 1/128 of ImageNet images: the rightmost perturbation between CIFAR-10 classes. The approach taken by Tsipras et al. (2019) to create these perturbations, which is based on a model with robustness to very small invariance noise, may point towards an efficient way of automating the generation of invariance attacks for tasks beyond MNIST. The work of Sharif et al. (2018) also shows that “small” $\ell_\infty$ noise (of magnitude 25/255) can reliably fool human labelers on CIFAR-10.

4.3. Natural Images

While our experiments are on MNIST, similar phenomena may arise in other vision tasks. Figure 3 shows two perturbations of ImageNet images: the rightmost perturbation is imperceptible and thus classifiers should be robust to it. Conversely, the middle image was semantically changed, and classifiers should be sensitive to such changes. Yet, the $\ell_2$ norm of both perturbations is the same. Hence, enforcing robustness to $\ell_2$-noise of some fixed size $\epsilon$ will necessarily result in a classifier that is either sensitive to the changes on the right, or invariant to the changes in the middle image. Such a phenomenon will necessarily arise for any image dataset that contains small objects, as perturbations of small $\ell_2$ magnitude will be sufficient to occlude the object, thereby changing the image semantics.

This distance-oracle misalignment extends beyond the $\ell_2$-norm. For instance, Co et al. (2018) show that a perturbation of size 16/255 in $\ell_\infty$ can suffice to give an image of a cat the appearance of a shower curtain print, which are both valid ImageNet classes. Yet, a random perturbation of the same magnitude is semantically meaningless.

On CIFAR-10, some recent defenses are possibly already overly invariant. For example, Shaeiri et al. (2020) and Panda et al. (2019) aim to train models that are robust to $\ell_\infty$ perturbations of size $\epsilon = 32/255$. Yet, Tsipras et al. (2019) show that perturbations of that magnitude can be semantically meaningful and can be used to effectively interpolate between CIFAR-10 classes. The approach taken by Tsipras et al. (2019) to create these perturbations, which is based on a model with robustness to very small invariance noise, may point towards an efficient way of automating the generation of invariance attacks for tasks beyond MNIST. The work of Sharif et al. (2018) also shows that “small” $\ell_\infty$ noise (of magnitude 25/255) can reliably fool human labelers on CIFAR-10.

5. The Overly-Robust Features Model

The experiments in Section 4 show that models can be robust to perturbations large enough to change an input’s semantics. Taking a step back, it is not obvious why training such classifiers is possible, i.e., why does excessive invariance not harm regular accuracy. To understand the learning dynamics of these overly-robust models, we ask two questions:

1. Can an overly-robust model fit the training data?
2. Can such a model generalize (robustly) to test data?

For simplicity, we assume that for every point $(x, y) \sim D$, the closest point $x^*$ (under the chosen norm) for which $O(x^*) \neq y$ is at a constant distance $\epsilon^*$. We train a model $f$ to have low robust error (as in equation 1) for perturbations of size $\epsilon > \epsilon^*$. This model is thus overly-robust.

We first ask under what conditions $f$ may have low robust training error. A necessary condition is that there do not exist training points $(x_i, y_i), (x_j, y_j)$ such that $y_i \neq y_j$ and $\|x_i - x_j\| \leq \epsilon$. As $\epsilon$ is larger than the inter-class distance, the ability to fit an overly robust model thus relies on the training data not being fully representative of the space to
which the oracle assigns labels. This seems to be the case in MNIST: as the dataset consists of centered, straightened and binarized digits, even an imaginary infinite-sized dataset might not contain our invariance adversarial examples.

The fact that excessive robustness generalizes (as provably evidenced by the model of Zhang et al. (2019)) points to a deeper issue: there must exist overly-robust and predictive features in the data—that are not aligned with human perception. This mirrors the observations of (Ilyas et al., 2019), who show that excessive sensitivity is caused by non-robust yet predictive features. On MNIST, our experiments confirm the existence of overly-robust generalizable features.

We formalize these observations using a simple classification task inspired by (Tsipras et al., 2019). We consider a binary task where unlabeled inputs \( x \in \mathbb{R}^{d+2} \) are sampled from a distribution \( D_x^k \) with parameter \( k \):

\[
x_2 = \begin{cases} 
  z + w.p. \frac{1+k/2}{2} x_3, \ldots, x_{d+2} \sim \mathcal{N}(\frac{z}{\sqrt{d}}, k) \\
  -z \quad w.p. \frac{1-k/2}{2}
\end{cases}
\]

Here \( \mathcal{N}(\mu, \sigma^2) \) is a normal distribution and \( k > 1 \) is a constant chosen so that only feature \( x_1 \) is strongly predictive of the latent variable \( z \) (e.g., \( k = 100 \) so that \( x_2, \ldots, x_{d+2} \) are almost uncorrelated with \( z \)). The oracle is defined as \( \mathcal{O}(x) = \text{sign}(x_1) \), i.e., feature \( x_1 \) fully defines the oracle’s class label, and other features are nearly uncorrelated with it. Note that the oracle’s labels are robust under any \( \ell_\infty \)-noise with norm strictly below \( \epsilon = 1/2 \).

We model the collection of “sanitized” and labeled datasets from a data distribution as follows: the semantic features (i.e., \( x_1 \)) are preserved, while “noise” features have their variance reduced (e.g., because non-standard inputs are removed). Sanitization thus enhances “spurious correlations” (Jo & Bengio, 2017; Ilyas et al., 2019) between non-predictive features and class labels.\(^5\) We further assume that the data labeling process introduces some small label noise.\(^6\) Specifically, the labeled data distribution \( D \) on which we train and evaluate classifiers is obtained by sampling \( x \) from a sanitized distribution \( D_x^{1+\alpha} \) (for a small constant \( \alpha > 0 \)) where features \( x_2, \ldots, x_{d+2} \) are strongly correlated with the oracle label. The label \( y \) is set to the correct oracle label with high probability \( 1 - \delta \). The consequences of this data sanitization are two-fold (see Appendix E for proofs):

1. A standard classifier (that maximizes accuracy on \( D \))\(^4\)

\[\text{agrees with the oracle with probability at least } 1 - \delta, \text{ but is vulnerable to } \ell_\infty - \text{perturbations of size } \epsilon = O(d^{-1/2}).\]

2. There is an overly-robust model that only uses feature \( x_2 \) and has robust accuracy \( 1 - \alpha/2 \) on \( D \) for \( \ell_\infty \)-noise of size \( \epsilon = 0.99 \). This classifier is vulnerable to invariance attacks as the oracle is not robust to such perturbations.

The Role of Data Augmentation. This simple task suggests a natural way to prevent the training of overly robust models. If prior knowledge about the task suggests that classification should be invariant to features \( x_2, \ldots, x_{d+2} \), then enforcing these invariances would prevent a model from being robust to excessively large perturbations.

A standard way to enforce invariances is via data augmentation. In the above binary task, augmenting the training data by randomizing over features \( x_2, \ldots, x_{d+2} \) would force the model to rely on the only truly predictive feature, \( x_1 \).

We experimented with aggressive data-augmentation on MNIST. For values of \( \epsilon \in [0, 0.4] \), we train models with an adversary that rotates and translates inputs by a small amount and then adds \( \epsilon \)-bounded \( \ell_\infty \)-noise. This attacker

![Figure 7: Robust accuracy of models trained and evaluated on an adversary combining a small spatial data augmentation (rotation + translation) with an \( \ell_\infty \) perturbation bounded by \( \epsilon \).](image1)

![Figure 8: Model accuracy with respect to human labelers on successful invariance adversarial examples of \( \ell_\infty \)-norm bounded by 0.4. Models trained with data augmentation agree more often with humans, and thus are less invariant to semantically-meaningful changes.](image2)
mirrors the process we use to generate invariance adversarial examples in Section 4. Thus, we expect it to be hard to achieve robustness to attacks with large $\varepsilon$ on this dataset, as this requires the model to correctly classify inputs that humans consider mislabeled. Figure 7 confirms this intuition. As $\varepsilon$ grows, it becomes harder to learn a model that is invariant to both spatial data augmentation and $\ell_\infty$ noise. We further find that the models trained with data augmentation agree more often with human labelers on our invariance attacks (see Figure 8). Yet, even with data augmentation, models trained against large $\ell_\infty$-perturbations still perform worse than an undefended model. This simple experiment thus demonstrates that while data-augmentation (over truly invariant features) can help in detecting or preventing excessive invariance to semantic features, even though it is not currently sufficient for training models that resist both sensitivity-based and invariance-based attacks.

6. Discussion

Our results show that solely focusing on robustness to sensitivity-based attacks is insufficient, as mis-specified bounds can cause vulnerability to invariance-based attacks.

On $\ell_p$-norm evaluations. Our invariance attacks are able to find points within the $\ell_p$-ball in which state-of-the-art classifiers are (provably) robust. This highlights the need for a more careful selection of perturbation bounds when measuring robustness to adversarial examples. At the same time, Figure 6 shows that even promoting robustness within conservative bounds causes excessive invariance. The tradeoff explored in Section 3 suggests that aiming for robustness against $\ell_p$-bounded attacks may be inherently futile for making models robust to arbitrary adversarial examples.

Trading Sensitivity and Invariance. We show that models that are robust to small perturbations make excessively invariant decisions and are thus vulnerable to other attacks. Interestingly, Engstrom et al. (2019a) show an opposite effect for models’ internal representations. Denoting the logit layer of a model as $z(x)$, they show that for robust models it is hard to find inputs $x, x^*$ such that $O(x) \neq O(x^*)$ and $z(x) \approx z(x^*)$. Conversely, Sabour et al. (2016) and Jacobsen et al. (2019) show that excessive invariance of feature layers is common in non-robust models. These observations are orthogonal to ours as we study invariances in a model’s classification layer, and for bounded perturbations. As we show in Section 3, robustness to large perturbations under a norm that is misaligned with human perception necessarily causes excessive invariance of the model’s classifications (but implies nothing about the model’s feature layers).

Increasing model robustness to $\ell_p$-noise also leads to other tradeoffs, such as reduced accuracy (Tsipras et al., 2019) or reduced robustness to other small perturbations (Yin et al., 2019; Tramèr & Boneh, 2019; Kang et al., 2019).

7. Conclusion

We have introduced and studied a fundamental tradeoff between two types of adversarial examples, that stem either from excessive sensitivity or invariance of a classifier. This tradeoff is due to an inherent misalignment between simple robustness notions and a task’s true perceptual metric.

We have demonstrated that defenses against $\ell_p$-bounded perturbations on MNIST promote invariance to semantic changes. Our attack exploits this excessive invariance by changing image semantics while preserving model decisions. For adversarially-trained and certified defenses, our attack can reduce a model’s true accuracy to random guessing.

Finally, we have studied the tradeoff between sensitivity and invariance in a theoretical setting where excessive invariance can be explained by the existence of overly-robust features.

Our results highlight the need for a more principled approach in selecting meaningful robustness bounds and in measuring progress towards more robust models.

Acknowledgments

Florian Tramèr’s research was supported in part by the Swiss National Science Foundation (SNSF project P1SKP2_178149). Resources used in preparing this research were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute www.vectorinstitute.ai/partners.

References


Fundamental Tradeoffs between Invariance and Sensitivity to Adversarial Perturbations


