# Fundamental Tradeoffs between Invariance and Sensitivity to Adversarial Perturbations



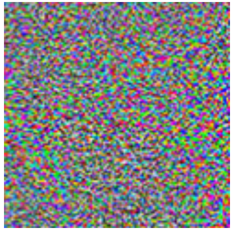


Florian Tramèr Jens Behrmann Nicholas Carlini Nicolas Papernot Jörn-Henrik Jacobsen

Stanford University
Universität Bremen
Google AI
University of Toronto
Vector Institute

# What are Adversarial Examples?

"any input to a ML model that is intentionally designed by an attacker to fool the model into producing an incorrect output"

| "Small" perturbations | "Large" perturbations | Nonsensical inputs | |
|---|---|---|---|
|  |  |  | |
| 99% Guacamole | 99% Guacamole | 99% Guacamole | etc. |

# $L_p$-bounded Adversarial Examples

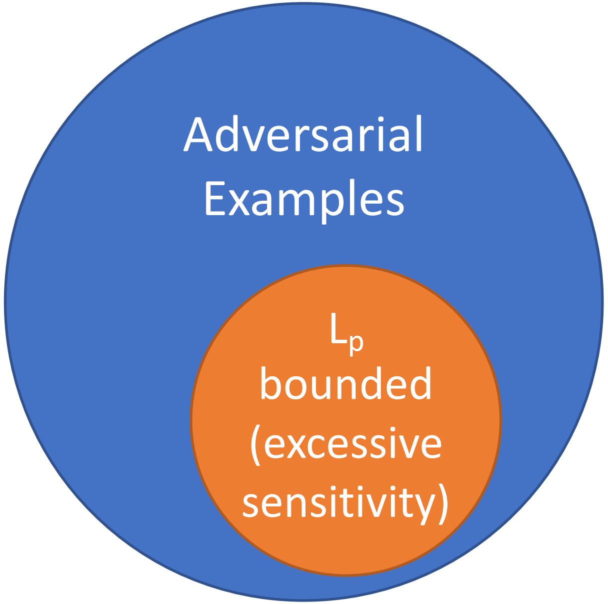
Given input x, find x′ that is misclassified such that $||x' - x|| \leq \varepsilon$

(+) Easy to formalize

(–) Incomplete

**Concrete measure of progress:**

*"my classifier has 97% accuracy for perturbations of $L_2$ norm bounded by $\varepsilon = 2$ "*

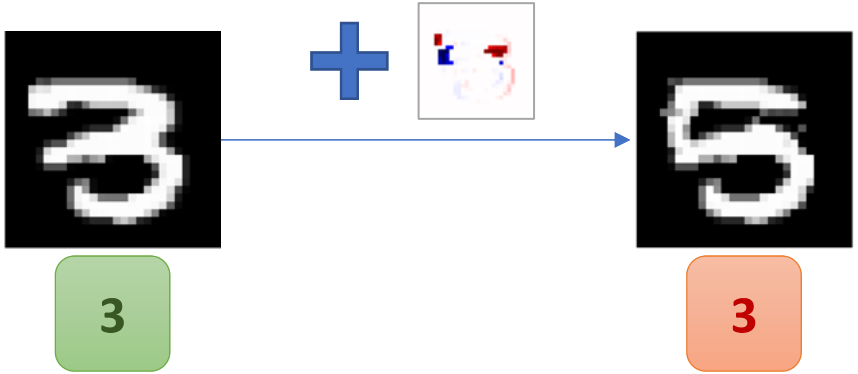


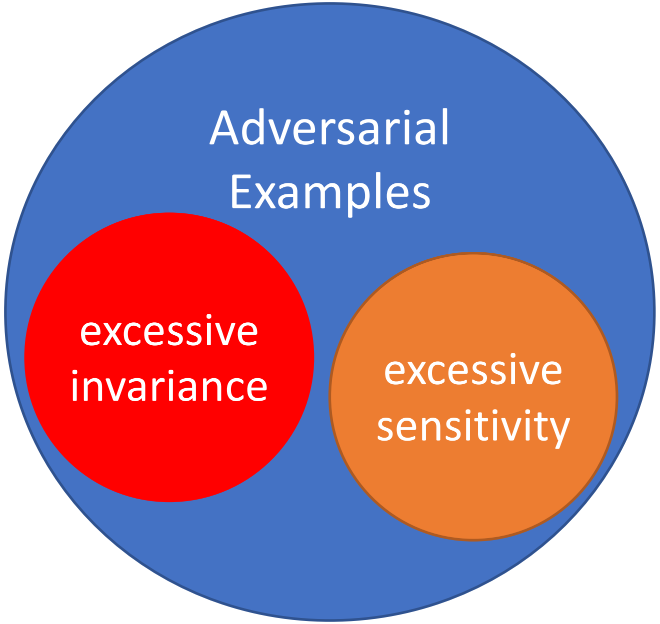
# Goodhart’s Law

“When a measure becomes a target, it ceases to be a good measure”

# New Vulnerability: Invariance Adversarial Examples

*Small semantics-altering* perturbations that *don’t* change classification



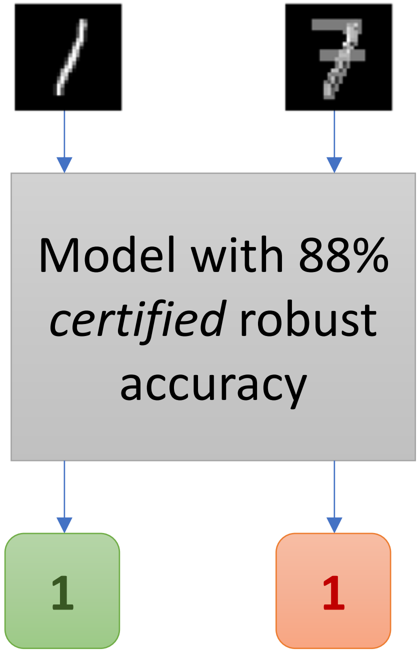
# Our Results

<u>State-of-the-art robust models are *too* robust</u>

Invariance to semantically meaningful features can be exploited

<u>Inherent tradeoffs</u>

Solving excessive sensitivity & invariance implies perfect classifier




**12% agreement with human labels**

# A Fundamental Tradeoff

Hermit-crab

$\|x' - x\|_2 \leq 22$

Guacamole

**OK!** I'll make my classifier robust to $L_2$ perturbations of size 22

(we don't yet know how to do this on ImageNet)

# A Fundamental Tradeoff

$$\|x' - x\|_2 \leq 22$$

**Hermit-crab**

**Hermit-crab**

**OK!** I'll choose a better norm than $L_2$

# A Fundamental Tradeoff

Theorem (informal)

Choosing a "good" norm is as hard as building a perfect classifier

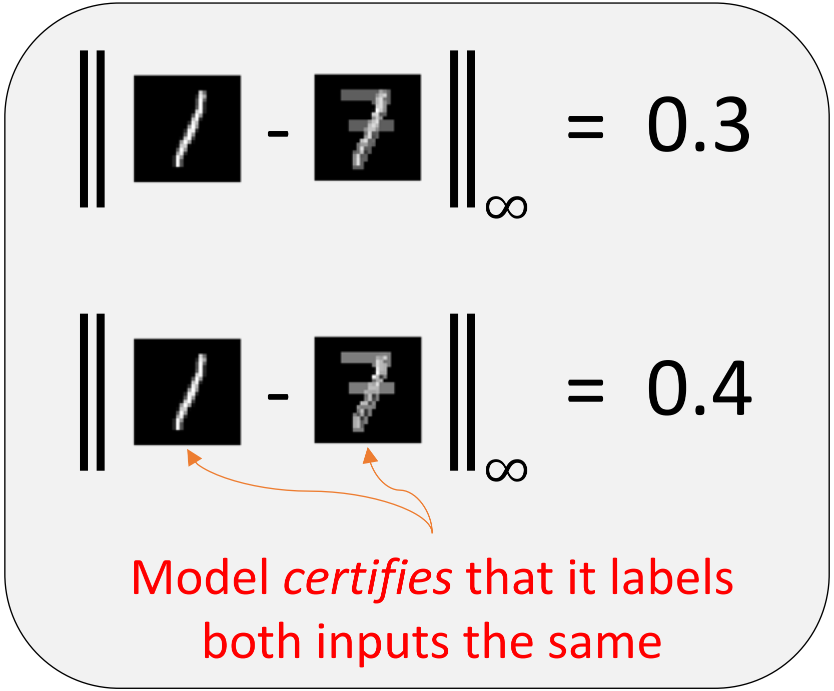
# Are Current Classifiers Already too Robust?

# A Case-Study on MNIST

State-of-the-art certified robustness:

$L_\infty \leq 0.3$: **93% accuracy**

$L_\infty \leq 0.4$: **88% accuracy**



# Automatically Generating Invariance Attacks

**Challenge**: ensure label is changed from human perspective

**Meta-procedure**: alignment via data augmentation


input

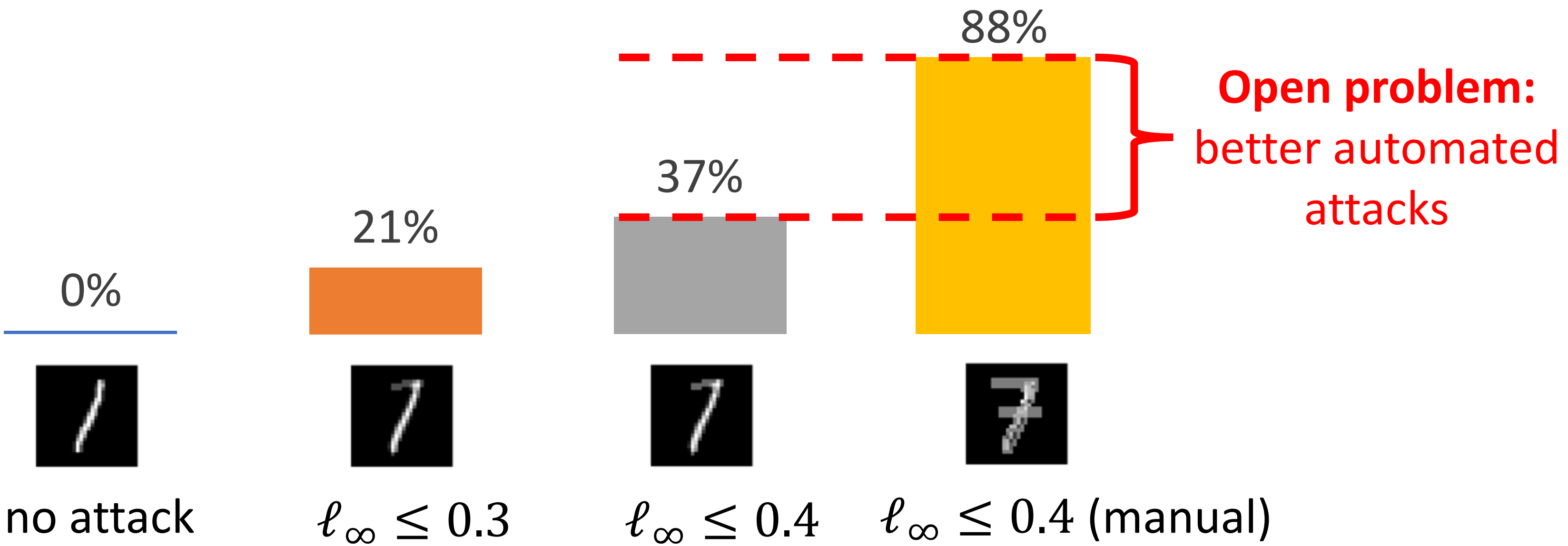
input from other class

semantics-preserving transformation

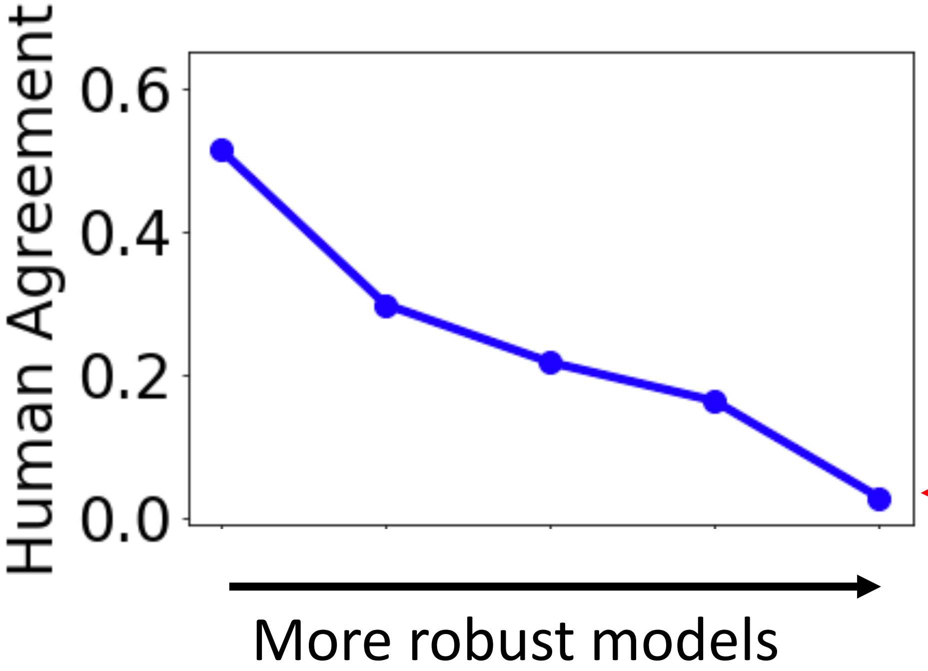
diff

result

# Do our invariance examples change human labels?



# Which models agree most with humans?



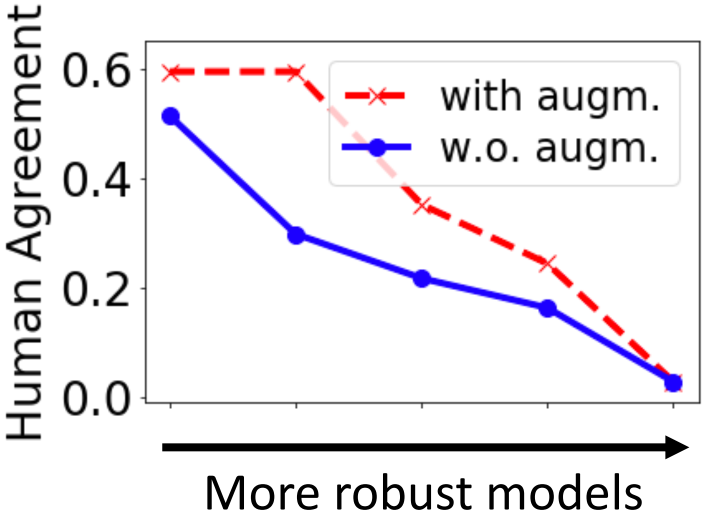

Most robust model *provably* gets all invariance examples wrong!

# Why can models be accurate yet overly invariant?

Or, why can an MNIST model achieve 88% test-accuracy for $\ell_\infty \leq 0.4$ ?

<u>Problem:</u> dataset is not *diverse enough*

<u>Partial solution:</u> data augmentation



# Conclusion

Robustness isn't yet another metric to monotonically optimize!

Max "real" robust accuracy on MNIST: ≈80% at $\ell_\infty = 0.3$
≈10% at $\ell_\infty = 0.4$

$\Rightarrow$ We've already over-optimized!

**Are we really making classifiers more robust,
or just overly smooth?**