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]

[1]Stanford University [2]University of Bremen [3]Google Brain [4]Vector Institute and University of Toronto

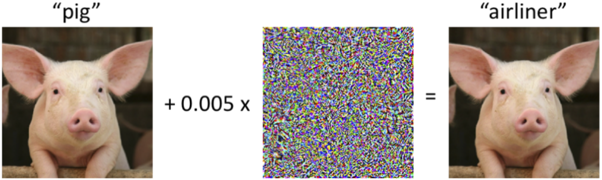
Tea talk - 13th February 2020

Adversarial Examples

Setting: classification for computer vision.

Definition

Malicious inputs (eg, designed by an adversary) that induces misclassification



Adversarial Examples

- classic adversarial examples, **"sensitivity based"**:
  small perturbation (non semantic) of an input that results in *different model prediction*
- this paper studies another kind of adversarial example, **"invariance based"**:
  small perturbation (semantic change) of the input that *does not change the model prediction*.

Is it possible to be robust to both types? There seems to be a fundamental trade-off.

Definition

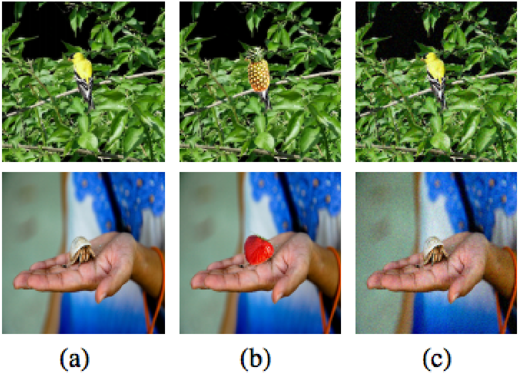
Adversarial training: The adversary abilities are constrained by bounding the size of the perturbation added to the original input (to leave the semantic of the input unchanged)

Formally, the perturbation lives in a l_p-ball where l_p is a norm:

- $l_p(x) = (\sum_{i=1}^{n} x_i^p)^{1/p}$
- $l_\infty(x) = \max_{i=1,\ldots,n} |x_i|$
- $l_0(x)$: number of non zeros coordinates/pixels that differ (not a norm)

Problem : this remains a crude approximation for visual similarity

Example



(a): original image; (b): invariance-based example; (c): sensitivity-based example

(b) and (c) are perturbations of same l_2 norm

Also Co et al. (2018) show that a perturbation of size 16/255 in l_∞ can suffice to give an image of a cat the appearance of a shower curtain print, which are both valid ImageNet classes.

Problems with current adversarial training

Their results: There seems to be a trade off between being robust to sensitivity-based examples and invariance-based examples.

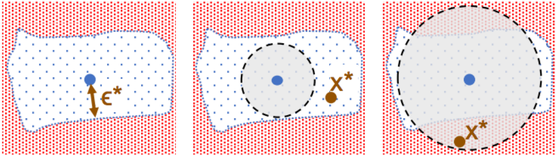
They managed to break *adversarially-trained* (1) and *certifiably robust* (2) models with these invariance-based examples.

- (1): augmenting training data using adversarial examples
- (2) Zhang et al 2019 provide a model certified to have 87% test accuracy under l_∞ perturbations of norm $\epsilon <= 0.4$

Intuition: distance-oracle misalignment

Definition

dist is aligned with the oracle $\mathcal{O}$ if for any x st $\mathcal{O}(x) = y$, and any (x_1, x_2) st $\mathcal{O}(x_1) = y$ and $\mathcal{O}(x_2) \neq y$, we have $dist(x, x_1) < dist(x, x_2)$.

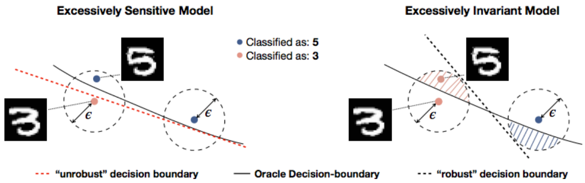


(a) (b) (c)

- (a): a point at distance ϵ^* in a chosen norm
- (b): a model robust to perturbations of norm $\epsilon < \epsilon^*$ is still vulnerable to sensitivity-based attacks (x^*)
- (c) : a model robust to perturbations of norm $\epsilon > \epsilon^*$ has invariant-based adversarial examples (x^*)

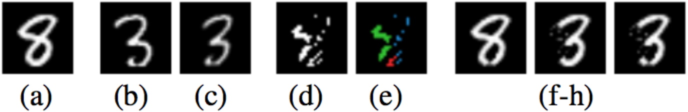
Study of MNIST

- Robust classification on MNIST is considered close to solved, with the existence of models highly robust to various l_p-bounded attacks
- This paper argues that it's far from being the case; and that this training harms the performance of the model against invariance-based attacks



Algorithm to generate Invariance-based examples

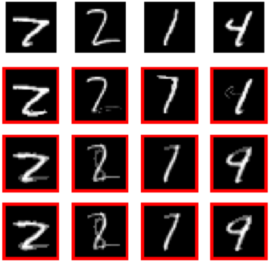
They introduce an algorithm to generate l_0 and l_∞ bounded invariance-based examples:



Process for generating l_0 invariant adversarial examples.
(a) original image;
(b)/(c) the nearest training image (labeled as 3), before/after alignment;
(d) the Δ perturbation between the original and aligned training example; (e) spectral clustering of Δ;
(f-h) candidate invariance adversarial examples, selected by applying subsets of clusters of Δ 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.

Invariance-based examples

Slow process, but they managed to create successful examples of low-distortion ($l_0 = 25$ or $l_\infty = 0.3, 0.4$).



Attack Type	Success Rate
Clean Images	0%
ℓ_0 Attack	55%
ℓ_∞, $\varepsilon = 0.3$ Attack	21%
ℓ_∞, $\varepsilon = 0.3$ Attack (**manual**)	26%
ℓ_∞, $\varepsilon = 0.4$ Attack	37%
ℓ_∞, $\varepsilon = 0.4$ Attack (**manual**)	88%

For evaluation, they use 100 generated IB examples and 50 hand-crafted ones.
They conduct a human-study (40 humans) to check if these examples are successful, ie if humans agree the label has been changed.

Results

Even models robust to small perturbations (l_∞ below $\epsilon < 0.01$) have higher vulnerability to invariance-based attacks compared to original models (without adversarial training).

Agreement between model and humans, for *successful* invariance adversarial examples						
Model:[1]	**Undefended**	ℓ_0 **Sparse**	**Binary-ABS**	**ABS**	ℓ_∞ **PGD**	ℓ_2 **PGD**
Clean	99%	99%	99%	99%	99%	99%
ℓ_0	80%	38%	47%	58%	56%*	27%*
$\ell_\infty, \varepsilon = 0.3$	33%	19%*	0%	14%	0%	5%*
$\ell_\infty, \varepsilon = 0.4$	51%	27%*	8%	18%	16%*	19%*

[1] ℓ_0 Sparse: (Bafna et al., 2018); ABS and Binary-ABS: (Schott et al., 2019); ℓ_∞ PGD and ℓ_2 PGD: (Madry et al., 2017)

$+$ they break certifiably robust models, such as Zhang et al 2019 (the one guaranteed 87% accuracy for l_∞ pert. of norm $\epsilon \leq 0.4$)

Conclusion

- The tradeoff between robustness to sensitivity based (SB) and invariant based (IB) examples is due to the distance misalignment (between the norm chosen and the perception)
- increasing robustnesss to SB decreases robustness to IB
- Discussion: they propose data augmentation (incorporate prior knowledge about invariance to features, or randomize over non-informative features)
- Code available to reproduce attacks