Adversarial Training and Robustness for Multiple Perturbations

Florian Tramèr & Dan Boneh
NeurIPS 2019
Adversarial examples

88% Tabby Cat + 

99% Guacamole

Szegedy et al., 2014
Goodfellow et al., 2015
Athalye, 2017
Adversarial examples

- ML models learn very different features than humans
- This is a safety concern for deployed ML models
- Classification in adversarial settings is hard

88% Tabby Cat + perturbation → 99% Guacamole

Szegedy et al., 2014
Goodfellow et al., 2015
Athalye, 2017
Adversarial training
Adversarial training

1. Choose a set of perturbations: e.g., noise of small $\ell_\infty$ norm:

Szegedy et al., 2014
Madry et al., 2017
Adversarial training

1. Choose a set of perturbations: e.g., noise of small $\ell_\infty$ norm:

2. For each example, find an adversarial example:

3. Train the model on:

4. Repeat until convergence

Szegedy et al., 2014
Madry et al., 2017
How well does it work?
How well does it work?

Adversarial training on CIFAR10, with $\ell_\infty$ noise of norm $\varepsilon = 4/255$

- Accuracy:
  - No attack: 92%
  - $\ell_\infty$ attack: 71%

Engstrom et al., 2017
Sharma & Chen, 2018
How well does it work?

Adversarial training on CIFAR10, with $\ell_\infty$ noise of norm $\varepsilon = 4/255$

<table>
<thead>
<tr>
<th>Attack Type</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>No attack</td>
<td>92%</td>
</tr>
<tr>
<td>$\ell_\infty$ attack</td>
<td>71%</td>
</tr>
<tr>
<td>$\ell_1$ attack</td>
<td>16%</td>
</tr>
<tr>
<td>Rotation attack</td>
<td>9%</td>
</tr>
</tbody>
</table>

Engstrom et al., 2017
Sharma & Chen, 2018
How to prevent other adversarial examples?
How to prevent other adversarial examples?

\[ S_1 = \{ \delta : \| \delta \|_\infty \leq \varepsilon_\infty \} \]

\[ S_2 = \{ \delta : \| \delta \|_1 \leq \varepsilon_1 \} \]

\[ S_3 = \{ \delta : «small rotation» \} \]

Adversary can choose a perturbation type for each input.
How to prevent other adversarial examples?

$S_1 = \{ \delta : \|\delta\|_\infty \leq \varepsilon_\infty \}$

$S_2 = \{ \delta : \|\delta\|_1 \leq \varepsilon_1 \}$

$S_3 = \{ \delta : \text{«small rotation»} \}$

$S = S_1 \cup S_2 \cup S_3$

- Pick worst-case adversarial example from $S$
- Train the model on that example
Does this work?
Does this work?

CIFAR10:

Train/eval on $\ell_\infty$: 71%

Train/eval on $\ell_1$: 66%

Train/eval on both: 61%
Does this work?

CIFAR10:

Train/eval on $\ell_\infty$: 71%
Train/eval on $\ell_1$: 66%
Train/eval on both: 61% -5%

similar results for $\ell_\infty$ and rotations

Adversarial Training and Robustness for Multiple Perturbations

Stanford University
Does this work?

CIFAR10:

We prove: this **robustness tradeoff** is inherent in some classification tasks.

similar results for $\ell_\infty$ and rotations

- Train/eval on $\ell_\infty$: 71%
- Train/eval on $\ell_1$: 66%
- Train/eval on both: 61% -5%

NOT GREAT, NOT TERRIBLE
Does this work?

CIFAR10:

- Train/eval on $\ell_\infty$: 71%
- Train/eval on $\ell_1$: 66%
- Train/eval on both: 61% -5%

MNIST:

- Train/eval on one of $\{\ell_\infty, \ell_1, \ell_2\}$: $\geq$72% -20%
- Train/eval on all: 52%

We prove: this robustness tradeoff is inherent in some classification tasks.

Adversarial Training and Robustness for Multiple Perturbations

Stanford University
What if we combine perturbations?
What if we combine perturbations?

natural image  rotation  $\ell_\infty$ noise  $\frac{1}{2}$ rotation + $\frac{1}{2}$ $\ell_\infty$ noise
What if we combine perturbations?

natural image | rotation | $\ell_\infty$ noise | $\frac{1}{2}$ rotation + $\frac{1}{2}$ $\ell_\infty$ noise

96% | 83% | 71% | 66% | 56% -10%

natural accuracy | rotation | $\ell_\infty$ noise | both attacks | mixed (affine) attack

Adversarial Training and Robustness for Multiple Perturbations
Conclusion

Adversarial training for multiple perturbation sets works, but...

- Significant loss in robustness
- Weak robustness to affine combinations of perturbations

Conclusion

Adversarial training for multiple perturbation sets works, but...

• Significant loss in robustness
• Weak robustness to affine combinations of perturbations

Open questions:

• Preventing gradient masking on MNIST to obtain high $\ell_1, \ell_2$ and $\ell_\infty$ robustness
• Better scaling of adversarial training to multiple perturbations
• How do we enumerate all the perturbation types that we care about?