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

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?
Adversarial training on CIFAR10, with $\ell_\infty$ noise

<table>
<thead>
<tr>
<th>No noise</th>
<th>$\ell_\infty$ noise</th>
</tr>
</thead>
<tbody>
<tr>
<td>96%</td>
<td>70%</td>
</tr>
</tbody>
</table>

Engstrom et al., 2017

Sharma & Chen, 2018
How well does it work?

Adversarial training on CIFAR10, with $\ell_\infty$ noise

<table>
<thead>
<tr>
<th>No noise</th>
<th>$\ell_\infty$ noise</th>
<th>$\ell_1$ noise</th>
<th>Rotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>96%</td>
<td>70%</td>
<td>16%</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?

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

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

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

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

Adversary can choose a perturbation type for each input
Does this work?
Does this work?
Does this work?

A robustness tradeoff is provably inherent in some classification tasks

Increased robustness to one type of noise

⇒ decreased robustness to another

Empirically validated on CIFAR10 & MNIST

NOT GREAT, NOT TERRIBLE
Does this work?

A robustness tradeoff is provably inherent in some classification tasks

Increased robustness to one type of noise
\[ \Rightarrow \text{decreased robustness to another} \]

Empirically validated on CIFAR10 & MNIST

MNIST:
Does this work?

A robustness tradeoff is provably inherent in some classification tasks.

Increased robustness to one type of noise \(\implies\) decreased robustness to another.

Empirically validated on CIFAR10 & MNIST.

MNIST:

For \(\ell_\infty\), \(\ell_1\) and \(\ell_2\) noise:

50% accuracy.
Does this work?

A robustness tradeoff is provably inherent in some classification tasks

Increased robustness to one type of noise ⇒ decreased robustness to another

Empirically validated on CIFAR10 & MNIST

MNIST:

For $\ell_\infty$, $\ell_1$ and $\ell_2$ noise:

50% accuracy

For $\ell_1$ noise:

MNIST: gradient masking

NOT GREAT, NOT TERRIBLE

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?

- No noise: 96%
- One noise type: 70%
- One of two noise types: 65%
- Mixture of two noise types: 55%

Accuracy
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:

- Train a *single* MNIST model with high robustness to any $\ell_p$ noise
- Better scaling of multi-perturbation adversarial training
- Which perturbations do we care about?