Adversarial Robustness for Aligned AI

Ian Goodfellow, Staff Research
NIPS 2017 Workshop on Aligned Artificial Intelligence

Many thanks to Catherine Olsson for feedback on drafts
The Alignment Problem

What can I help you with?

“I’m bleeding really bad can you call me an ambulance ”

From now on, I’ll call you ‘An Ambulance’. OK?

Cancel Yes

(This is now fixed. Don’t try it!)

(Goodfellow 2017)
Main Takeaway

• My claim: if you want to use alignment as a means of guaranteeing safety, you probably need to solve the adversarial robustness problem first
Why the “if”?

• I don’t want to imply that alignment is the only or best path to providing safety mechanisms

• Some problematic aspects of alignment

  • Different people have different values

  • People can have bad values

  • Difficulty / lower probability of success. Need to model a black box, rather than a first principle (like low-impact, reversibility, etc.)

• Alignment may not be necessary

  • People can coexist and cooperate without being fully aligned
Some context: many people have already been working on alignment for decades

• Consider alignment to be “learning and respecting human preferences”

• Object recognition is “human preferences about how to categorize images”

• Sentiment analysis is “human preferences about how to categorize sentences”
What do we want from alignment?

• Alignment is often suggested as something that is primarily a concern for RL, where an agent maximizes a reward
  • but we should want alignment for supervised learning too
• Alignment can make better products that are more useful
• Many want to rely on alignment to make systems safe
  • Our methods of providing alignment are not (yet?) reliable enough to be used for this purpose
Improving RL with human input

• Much work focuses on making RL more like supervised learning
  • Reward based on a model of human preferences
  • Human demonstrations
  • Human feedback

• This can be good for RL capabilities
  • The original AlphaGo bootstrapped from observing human games
  • OpenAI’s “Learning from Human Feedback” shows successful learning to backflip

• This makes RL more like supervised learning and makes it work, but does it make it robust?
Adversarial Examples

Timeline:
“Adversarial Classification” Dalvi et al 2004: fool spam filter
“Evasion Attacks Against Machine Learning at Test Time” Biggio 2013: fool neural nets
Szegedy et al 2013: fool ImageNet classifiers imperceptibly
Goodfellow et al 2014: cheap, closed form attack

(Goodfellow 2017)
Maximizing model’s estimate of human preference for input to be categorized as “airplane”
Sampling: an easier task?

- Absolutely maximizing human satisfaction might be too hard. What about sampling from the set of things humans have liked before?

- Even though this problem is easier, it’s still notoriously difficult (GANs and other generative models)

- GANs have a trick to get more data
  - Start with a small set of data that the human likes
  - Generate millions of examples and assume that the human dislikes them all
Spectrally Normalized GANs

This is better than the adversarial panda,
but still not a satisfying safety mechanism.

(Miyato et al., 2017)
Progressive GAN has learned that humans think cats are furry animals accompanied by floating symbols

(Karras et al, 2017)
Confidence

• Many proposals for achieving aligned behavior rely on accurate estimates of an agents’ confidence, or rely on the agent having low confidence in some scenarios (e.g. Hadfield-Menell et al 2017)

• Unfortunately, adversarial examples often have much higher confidence than naturally occurring, correctly processed examples
Adversarial Examples for RL

(Huang et al., 2017)
Summary so Far

• High level strategies will fail if low-level building blocks are not robust

• Reward maximizing places low-level building blocks under exactly the same situation as adversarial attack

• Current ML systems fail frequently and gracelessly under adversarial attack; have higher confidence when wrong
What are we doing about it?

- Two recent techniques for achieving adversarial robustness:
  - Thermometer codes
  - Ensemble adversarial training
- A long road ahead
Thermometer Encoding: One Hot Way to Resist Adversarial Examples

Jacob Buckman*
Aurko Roy*
Colin Raffel
Ian Goodfellow

*joint first author
Linear Extrapolation

Vulnerabilities
Neural nets are “too linear”

Plot from “Explaining and Harnessing Adversarial Examples”, Goodfellow et al, 2014
<table>
<thead>
<tr>
<th></th>
<th>Real-valued</th>
<th>Quantized</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.13</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td>0.66</td>
<td>0.65</td>
</tr>
<tr>
<td></td>
<td>0.92</td>
<td>0.95</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Discretized (one-hot)</th>
<th>Discretized (thermometer)</th>
</tr>
</thead>
<tbody>
<tr>
<td>[01000000000]</td>
<td>[01111111111]</td>
</tr>
<tr>
<td>[00000010000]</td>
<td>[00000011111]</td>
</tr>
<tr>
<td>[00000000001]</td>
<td>[00000000001]</td>
</tr>
</tbody>
</table>
Large improvements on SVHN direct ("white box") attacks

5 years ago, this would have been SOTA on clean data

(Goodfellow 2017)
Large Improvements against CIFAR-10 direct ("white box") attacks

6 years ago, this would have been SOTA on clean data
Ensemble Adversarial Training

Florian Tramèr
Alexey Kurakin
Nicolas Papernot
Ian Goodfellow
Dan Boneh
Patrick McDaniel
Cross-model, cross-dataset generalization
Ensemble Adversarial Training
Transfer Attacks Against Inception ResNet v2 on ImageNet

![Bar chart showing worst black-box test accuracy for different models: Standard, Adv Trained, Ensemble Adv Trained.](Goodfellow2017)
Competition

AI Fight Club Could Help Save Us from a Future of Super-Smart Cyberattacks

Best defense so far on ImageNet: Ensemble adversarial training.
Used as at least part of all top 10 entries in dev round 3
Future Work

• Adversarial examples in the max-norm ball are not the real problem

• For alignment: formulate the problem in terms of inputs that reward-maximizers will visit

• Verification methods

• Develop a theory of what kinds of robustness are possible

• See “Adversarial Spheres” (Gilmer et al 2017) for some arguments that it may not be feasible to build sufficiently accurate models

(Goodfellow 2017)
Get involved!

https://github.com/tensorflow/cleverhans