Instructor: Ryan Lowe (ryan.lowe@cs.mcgill.ca)

Slides by: Ryan Lowe

Class web page: www.cs.mcgill.ca/~hvanho2/comp551
Announcements

• First round of project presentations is **today!!**
  – 6-7:30pm in this room
  – *If you are not on the list, please let me know*

• **Please fill out course evaluations!**

• TAs (Prasanna + Sanjay) will host a joint office hour for assignment 3 clarifications, from 1-2pm on Thursday in TR 3104
Hot areas of ML research

• Reinforcement learning
• Generative models, GANs
• Meta-learning
• Adversarial examples
• Imitation learning from few examples
Hot areas of ML research

• Reinforcement learning
• Generative models, GANs
• Meta-learning
• Adversarial examples
• Imitation learning from few examples
Recap: generative models

- Given inputs $x$, labels $y$

  **Traditional definition:** A generative model learns a joint probability distribution $p(x,y)$ over the inputs and labels.

- Can later transform to $p(y|x)$ to make predictions.

- Can ‘generate’ data by sampling from $p(x|y)$ or $p(x)$.

- What if our goal is not to classify, but just to generate realistic samples?
  - Useful in video prediction, model-based RL.
Generative adversarial networks
Generative adversarial networks

- **Idea**: have two networks, a *discriminator* and a *generator*

- **Generator**: with random noise as input, generate a realistic image

- **Discriminator**: try to distinguish images from generator and dataset
Generative adversarial networks

\[ \min_G \max_D V(D, G) \]

\[ V(D, G) = \mathbb{E}_{x \sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z)))] \]

\( x \): data, \( G(z) \): generator sample, \( D() \): discriminator evaluation (0-1)
Generative adversarial networks

Karras et al., (2018)
Generative adversarial networks


Karras et al., (2018)
Generative adversarial networks

Nguyen et al., (2016)
Adversarial examples

• Convolutional neural nets (CNNs) perform very well at classifying images
• But do they understand images in the same way that humans do?
• In other words, by making a tiny change, can we fool neural networks into thinking an image is something else?
Adversarial examples

• By making a tiny change, can we fool neural networks into thinking an image is something else?

• **Yes!!!**

• **Idea:** take the gradient of the image with respect to a particular output class
Adversarial examples

\[ x + \epsilon \text{sign}(\nabla_x J(\theta, x, y)) = x + \epsilon \text{sign}(\nabla_x J(\theta, x, y)) \]

- **“panda”**: 57.7% confidence
- **“nematode”**: 8.2% confidence
- **“gibbon”**: 99.3% confidence
Adversarial examples
Adversarial examples in the physical world

- [https://www.youtube.com/watch?v=zQ_uMenoBCk](https://www.youtube.com/watch?v=zQ_uMenoBCk)
Adversarial examples

- **July 12, 2017:** don’t worry, adversarial examples are sensitive to scaling!
Adversarial examples

• July 24, 2017: just kidding

Synthesizing Robust Adversarial Examples

Anish Athalye
OpenAI, MIT

Ilya Sutskever
OpenAI

Abstract

Neural networks are susceptible to adversarial examples: small, carefully-crafted perturbations can cause networks to misclassify inputs in arbitrarily chosen ways. However, some studies have showed that adversarial examples crafted following the usual methods are not tolerant to small transformations: for example, zooming in on an adversarial image can cause it to be classified correctly again. This raises the question of whether adversarial examples are a concern in practice, because many real-world systems capture inputs from multiple angles and perspectives.

This paper shows that adversarial examples can be made robust to distributions of transformations. Our approach produces single images that are simultaneously adversarial under all transformations in a chosen distribution, showing that we cannot rely on transformations such as rescaling, translation, and rotation to protect against adversarial examples.

Examples are not preserved by simple transformations such as rescaling or rotation, implying that adversarial examples are not an issue in practice because many real-world systems capture inputs from multiple angles and perspectives [10, 11]. We attempted to reproduce these results, and we did find that naively-generated adversarial examples are brittle, failing to remain adversarial when subject to small transformations.

However, our results show that adversarial examples are a concern even when inputs are subject to transformations such as zoom, translation, rotation, and noise. We show that it is possible to synthesize adversarial examples that are robust to an entire distribution of transformations.

2 Approach

When finding adversarial examples for the regular case, we are given a classifier $P(y \mid x)$, an input $x$, a target class $\hat{y}$, and a maximum perturbation $\varepsilon$, and we want...
Robust adversarial examples
Robust adversarial examples
Adversarial patches

https://www.youtube.com/watch?v=i1sp4X57TL4
Adversarial examples

- First shown in 2013 (Szegedy et al.)
- Some methods are more robust than others
  - E.g. can use adversarial examples as extra training data
- After 5 years, there is still no method that defeats all forms of adversarial examples
- Could cause problems with increasing use of automatic detection/classification technology
One-shot imitation

- Want robots to be able to do complex tasks
- In some cases, much easier to do by imitating a human demonstration (*imitation learning*) than reinforcement learning
  - Cook a recipe after seeing a human complete it
  - Assemble a new device after seeing it assembled once
- But human demonstrations are expensive
- Can we get a robot to solve a task *after seeing only one human demonstration*?
One-shot imitation

- https://blog.openai.com/robots-that-learn/
Ethical issues in ML/ AI

• Machine learning / AI will have a huge impact on the world
• Many, many ethical issues to consider:
  – Privacy
  – Unemployment & economic inequality
  – Military applications
  – Bias (based on e.g. race, gender, etc.)
  – …
Exercise

• Form a group of 3 with your neighbours
• Take 5-10 minutes to brainstorm how machine learning could impact one of the ethical issues on the previous slide
• Some questions that may spur your discussion:
  – What are some worst case scenarios?
  – How could these problems come about?
  – How could this be averted?
  – Who should be responsible for overseeing efforts to ensure that these issues don’t happen?
AI safety

- **Artificial General Intelligence (AGI):** an AI that can perform any task at least as well as an average human
- **Artificial Superintelligence (ASI):** an AI that greatly outperforms humans on any task
- Several surveys of AI experts indicate a 50% confidence that AGI will be achieved by around 2050 (e.g. Grace et al., (2017))
- *Should we be afraid of AGI and ASI? Or will it be beneficial to humanity?*
AI safety: cons

• AGI is a long way away --- lots of problems current systems
  – Why don’t we focus on the ethical problems we know we’ll encounter instead?
• We have no idea what AGI will look like
  – How can we start preparing if we don’t know what to work on?
• There is no such thing as ‘general intelligence’ --- intelligence is always specialized
• AI safety advocates are fear mongering, Terminator-like scenarios are extremely unlikely
• If an AI is truly intelligent, it will learn our values regardless
AI safety: pros

- We have no idea when AGI could occur
- AGI may follow an existing framework (e.g. RL)
- If we wait until AGI is almost here, will be too hard to catch up
- Can focus on both AI safety *and* ethical issues
- Some risks seem probable no matter what form AGI takes
  - AGI will likely want to increase its capabilities
  - AGIs could accelerate the pace of AI research, leading to an ‘intelligence explosion’
  - *How can we ensure AGI has values that are aligned with humans?*
    - Small difference in values could lead to a big difference in behaviour
Discussion

• How soon after AGI could we have ASI?
• How likely is a ‘race dynamic’ to occur when we get close to AGI? How will that affect safety concerns?
• Orthogonality thesis: an AI’s capabilities (intelligence) and goals can be completely independent. Do you agree with this?
• How could we program an AGI to have human values?
• What would have to go right for AGI to have a positive impact on humanity?
Myth: Superintelligence by 2100 is inevitable
Fact: It may happen in decades, centuries or never. AI experts disagree & we simply don't know.

Myth: Superintelligence by 2100 is impossible

Myth: Only Luddites worry about AI
Fact: Many top AI researchers are concerned.

Myth: AI turning evil
Fact: AI turning competent, with goals misaligned with ours.

Myth: AI turning conscious

Myth: Robots are the main concern
Fact: Misaligned intelligence is the main concern: it needs no body, only an internet connection.

Myth: AI can't control humans
Fact: Intelligence enables control: we control tigers by being smarter.

Myth: Machines can't have goals
Fact: A heat-seeking missile has a goal.

Mythical worry: Superintelligence is just years away
Actual worry: It's at least decades away, but it may take that long to make it safe.

PLAN AHEAD!
Thank you