Adversarial Attacks Against Medical Deep Learning Systems

Deep Learning in Medicine

Collage of some medical imaging applications in which deep learning has achieved state-of-the-art results. From top-left to bottom-right:

- mammographic mass classification
- segmentation of lesions in the brain
- leak detection in airway tree segmentation
- diabetic retinopathy classification (Kaggle)
- prostate segmentation
- nodule classification
- breast cancer metastases detection in lymph nodes
- human expert performance in skin lesion classification
- and state-of-the-art bone suppression in x-rays
-

The U.S. Food and Drug Administration announced the approval of the first computer vision algorithm that can be utilised for medical diagnosis without the input of a human clinician (April 2018)

Adversarial Attacks

Stop signs - With a perturbation in the form of only black and white stickers,we attack a real stop sign, causing targeted miss classification in 100% of the images obtained in lab settings, and in 84.8%of the captured video frames obtained on a moving vehicle(field test) for the target classifier.

Google Inception - We apply our algorithm to complex three-dimensional objects, using 3D-printing to manufacture the first physical adversarial objects.

Video shows a 3D printed turtle. Small histogram is the predictions, green is correct, blue is not bad, red is completely wrong (pause it a few times to discuss)

Why is Medicine important?

- Either sent for unnecessary treatment, or miss required treatment.
- Total global spend on healthcare is more than $7 trillion each year. Corruption takes many forms, depending on the country's level of development and health financing system. The United States, for example, lost between $82bn and $272bn in 2011 to medical embezzlement, mostly related to its health insurance system
- Increasing use in clinical trials...individual cancer drug worth $1.67billion only four years after approval.

Why is Medicine particularly susceptible?

- Ground truth is ambiguous, experts (radiologists) often disagree on well defined tasks.
- Images are highly standardised. Very different to the real word attacks we saw earlier which have varying angles and lighting
- Popular architecture are often used; quite possible that they will be public for transparency...often pretrained ImageNet or Inception
- Many adversaries, at every step of a patient pathway there is someone who could profit

Create adversarial examples

- 2 techniques, one imperceptible and one is not
- FGS, small perturbation in the direction of gradient (up the hill)
- PGD, like FGS but iterative, each move is projected onto allowed set of perturbations
- if we don't have access to the model we can't compute the gradient (white box)
- if we know the architecture, train our own and then we have a gradient (black box)
- otherwise we have to estimate the gradient, depends on what the model outputs...a probability distribution or just a label.

Patch Attack

- Targeted, we choose a target class that we want the classifier to choose
- Universal, we produce one patch that is used as an attack on any scene
- Robust, they work under a large variety of transformations
-

Patch Attack Example

- Designed to be printed
- Up to about 20 degree rotation either direction
- In the video, classify a banana, add picture of toaster and ignore it, add patch and classify as toaster
-

Create Adversarial Patch

- patch application operator takes patch p, applies transformations t, and adds it to image x at location l.
- find p that maximises the expected log probability
- log because optimiser methods perform better

Current Research

- All publicly available data where deep learning has been shown to perform well (eg kaggle)
- nevus; mark on the skin (ie not a melanoma)
- Classifiers built by fine tuning a pre-trained ImageNet Model (with various augmentation)
- nat patch
- Implemented human-imperceptible (PGD because they trained the model themselves) and patch attacks
- White box and Black Box Attack

Our plan

- Break the best systems, ie show they are vulnerable to attack...can the same attack defeat multiple systems?
- Work out why the attack was successful, what is the mathematical reasoning
- Figure out ways to make things more robust, image augmentation etc