Adversarial Attacks Against Medical Deep Learning Systems

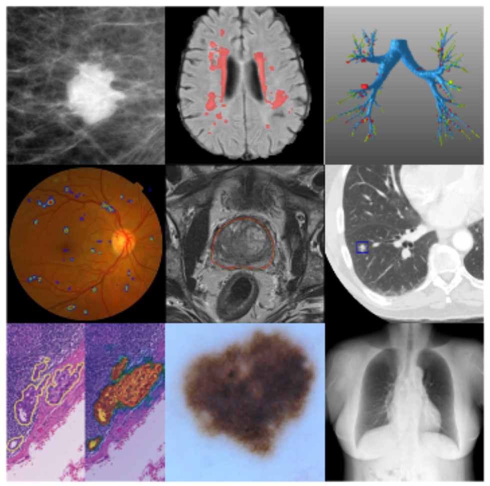
James Campbell

June 27, 2018

Cardiff University, School of Mathematics

Deep Learning in Medicine

Deep Learning in Medicine



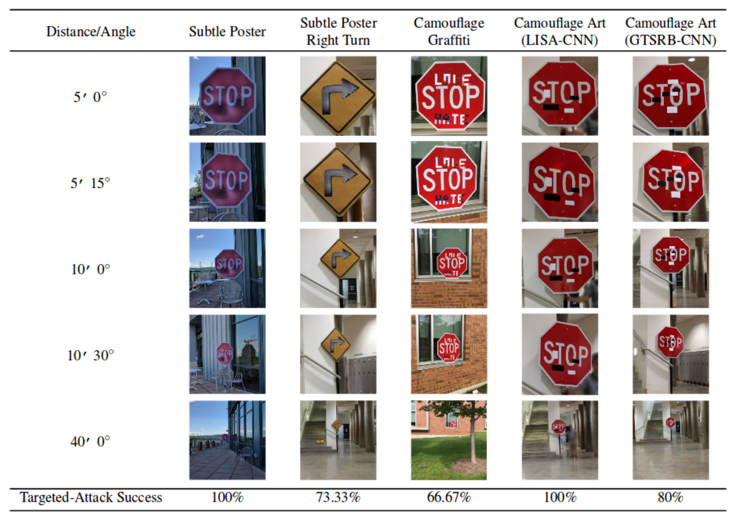
[9]

FDA U.S. FOOD & DRUG ADMINISTRATION

[3]

Adversarial Attacks

Adversarial Attacks



[4]

Adversarial Attacks

Fool Google’s InceptionV3 image classifier video. [1, 7]

Why Medicine?

Why is Medicine important?

- An incorrect diagnosis can be dangerous to patients

Why is Medicine important?

- An incorrect diagnosis can be dangerous to patients
- Healthcare economy is huge and fraud is already a major problem [8]

Why is Medicine important?

- An incorrect diagnosis can be dangerous to patients
- Healthcare economy is huge and fraud is already a major problem [8]
- Increasing use in clinical trials [12]

Why is Medicine particularly susceptible?

- Ambiguous ground truth [11]

Why is Medicine particularly susceptible?

- Ambiguous ground truth [11]
- Images are standardised

Why is Medicine particularly susceptible?

- Ambiguous ground truth [11]
- Images are standardised
- Popular Architectures are often used

Why is Medicine particularly susceptible?

- Ambiguous ground truth [11]
- Images are standardised
- Popular Architectures are often used
- Many potential adversaries

Create Adversarial Examples

Fast Gradient Sign Method

Let θ be the parameters of a model, x an input to the model and y the target associated with x. We also have a well defined loss function $L(\theta, x, y)$.

Fast Gradient Sign Method

Let θ be the parameters of a model, x an input to the model and y the target associated with x. We also have a well defined loss function $L(\theta, x, y)$.

Then FGSM computes an adversarial example as:

$$x + \epsilon \cdot \text{sign}(\nabla_x L(\theta, x, y))$$

[6]

Projected Gradient Descent

PGD make this an iterative process. We specify a set of allowed perturbations $\mathcal{S} \in \mathbb{R}^d$ (commonly the l_∞ ball around x) and compute:

Projected Gradient Descent

PGD make this an iterative process. We specify a set of allowed perturbations $\mathcal{S} \in \mathbb{R}^d$ (commonly the l_∞ ball around x) and compute:

$$x^{t+1} = \Pi_{x+\mathcal{S}}(x^t + \epsilon \cdot \text{sign}(\nabla_x L(\theta, x, y)))$$

[10]

Small Issue

What if we don't have access to the model?

Small Issue

What if we don't have access to the model?

- If we know the architecture, train our own version

Small Issue

What if we don't have access to the model?

- If we know the architecture, train our own version
- If we don't know the architecture, but have access to probabilities, use NES (Natural Evolutionary Strategies) Gradient Estimate [7]

Small Issue

What if we don't have access to the model?

- If we know the architecture, train our own version
- If we don't know the architecture, but have access to probabilities, use NES (Natural Evolutionary Strategies) Gradient Estimate [7]
- If we only have access to predicted class, use a Monte Carlo approximation [7]

Patch Attack

Some major differences:

- Targeted

Patch Attack

Some major differences:

- Targeted
- Universal

Patch Attack

Some major differences:

- Targeted
- Universal
- Robust

[2]

Example Patch Attack

video [2]

Building Adversarial Patches

Given an image x, a patch p, a location l and transformations t (rotation and scaling) we define a *patch application operator* $A(p, x, l, t)$.

Building Adversarial Patches

Given an image x, a patch p, a location l and transformations t (rotation and scaling) we define a *patch application operator* $A(p, x, l, t)$.

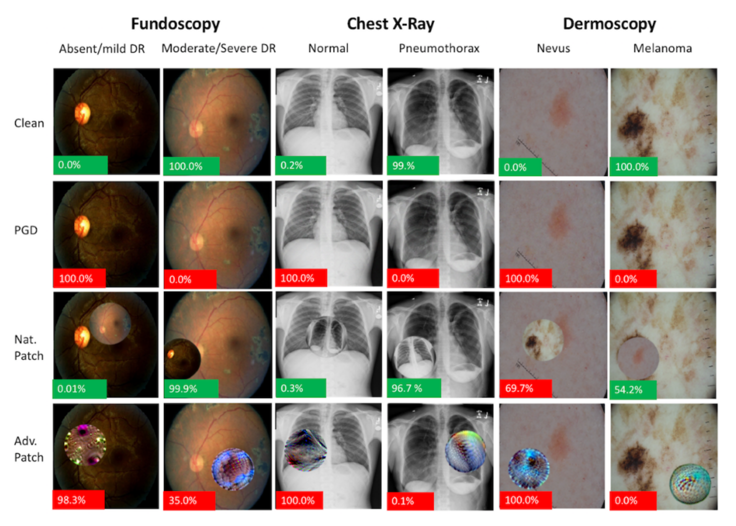
To obtain the final adversarial patch $\hat{p}$ we compute:

$$\hat{p} = \arg\max_{p} \mathbb{E}_{x\sim X, l\sim L, t\sim T}(\log P(\hat{y}|A(p, x, l, t)))$$

where $\hat{y}$ is the target class. [1, 2]

Current Research

Current Research



[5]

Our Plan

Our Plan

- Break the best deep learning systems

Our Plan

- Break the best deep learning systems
- Understand how they were broken

Our Plan

- Break the best deep learning systems
- Understand how they were broken
- Make them more robust

References

Anish Athalye et al. "Synthesizing Robust Adversarial Examples". In: *arXiv:1707.07397 [cs]* (July 2017). arXiv: 1707.07397. URL: `http://arxiv.org/abs/1707.07397`.

Tom B. Brown et al. "Adversarial Patch". In: *arXiv:1712.09665 [cs]* (Dec. 2017). arXiv: 1712.09665. URL: `http://arxiv.org/abs/1712.09665`.

Office of the Commissioner. *Press Announcements - FDA permits marketing of artificial intelligence-based device to detect certain diabetes-related eye problems.* URL: https://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm604357.htm.

Kevin Eykholt et al. "Robust Physical-World Attacks on Deep Learning Models". In: *arXiv:1707.08945 [cs]* (July 2017). arXiv: 1707.08945. URL: http://arxiv.org/abs/1707.08945.

- Samuel G. Finlayson et al. "Adversarial Attacks Against Medical Deep Learning Systems". In: *arXiv:1804.05296 [cs, stat]* (Apr. 2018). arXiv: 1804.05296. URL: `http://arxiv.org/abs/1804.05296`.
- Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. "Explaining and Harnessing Adversarial Examples". In: *arXiv:1412.6572 [cs, stat]* (Dec. 2014). arXiv: 1412.6572. URL: `http://arxiv.org/abs/1412.6572`.
- Andrew Ilyas et al. "Black-box Adversarial Attacks with Limited Queries and Information". In: *arXiv:1804.08598 [cs, stat]* (Apr. 2018). arXiv: 1804.08598. URL: `http://arxiv.org/abs/1804.08598`.

Anita Jain, Samiran Nundy, and Kamran Abbasi. "Corruption: medicine's dirty open secret". In: *BMJ* 348 (June 2014). ISSN: 1756-1833. DOI: `10.1136/bmj.g4184`.

Geert Litjens et al. "A Survey on Deep Learning in Medical Image Analysis". In: *Medical Image Analysis* 42 (Dec. 2017). arXiv: 1702.05747, pp. 60–88. ISSN: 13618415. DOI: `10.1016/j.media.2017.07.005`.

Aleksander Madry et al. "Towards Deep Learning Models Resistant to Adversarial Attacks". In: *arXiv:1706.06083 [cs, stat]* (June 2017). arXiv: 1706.06083. URL: `http://arxiv.org/abs/1706.06083`.

C. F. Njeh. "Tumor delineation: The weakest link in the search for accuracy in radiotherapy". In: *Journal of Medical Physics* 33.4 (Oct. 2008), p. 136. ISSN: 0971-6203. DOI: 10.4103/0971-6203.44472.

Homer H. Pien et al. "Using imaging biomarkers to accelerate drug development and clinical trials". In: *Drug Discovery Today* 10.4 (Feb. 2005), pp. 259–266. ISSN: 1359-6446. DOI: 10.1016/S1359-6446(04)03334-3.