Adversarial Attacks Against Medical Deep Learning Systems

S.G. Finlayson[1], H.W. Chung[1], I.S. Kohane[1], A.L. Beam[1]

[1]Harvard Medical School [2]Massachusetts Institute of Technology

arXiv: 1804.05296

Reviewed by : Bill Zhang
University of Virginia
https://qdata.github.io/deep2Read/

Outline

Introduction

Adversarial Examples

Healthcare System

Demonstration of Adversarial Attacks

Discussion

Conclusion

References

Introduction

Basic Premise and Motivation

- There have been many recent high-profile examples of deep learning achieving parity with human physicians on tasks in radiology, pathology, and opthalmology
- Early 2018, FDA announced approval of first computer vision algorithm that can be used for medical diagnosis without human input
- At the same time, adversarial examples have exposed vulnerabilities within even state-of-the-art learning systems
- As deep learning becomes more common in the healthcare system, adversarial attacks present opportunities for fraud and harm
- Argue that the healthcare system is particularly vulnerable to these adversarial attacks

Adversarial Examples

- Inputs that are specifically crafted to cause model to make classification error
- First discovered by Szegedy et al. and Goodfellow et al.
- Thought to arise from piecewise linear components of complex nonlinear models; not random, not due to overfitting or incomplete training, occupy a small part of feature subspace, are robust to random noise, and have shown to transfer between models
- Most recent works discussing impact of adversarial examples on real-world applications have focused on self-driving cars; not much focus on medical systems

Healthcare System

Background and Incentives for Fraud

- The healthcare economy is huge and fraud is already pervasive
  - Almost 1/5 of U.S. economy
  - Medical fraud costs hundreds of billions of dollars per year
  - Common actions include systemically inflating costs and physicians frequently billing for highest possible amount
- Algorithms will likely make medical reimbursement decisions in the future
  - Ability to influence any ML models used for these applications (as either the provider or payer) could affect movement of billions of dollars in economy
- Algorithms will increasingly determine pharmaceutical and device approvals
  - Clinical trials are expensive, and if decisions to approve of these trials are done by algorithms, trialists could influence models to approve of their new drugs

Healthcare System

Sources of Vulnerability

- Ground truth is often ambiguous
  - Many medical imaging tasks have no clear answer, with disagreement among even trained professional radiologists
  - Perturbing borderline cases would be very difficult to detect
- Medical imaging is highly standardized
  - Adversarial examples do not need to consider variations in positioning and lighting
- Commodity network architectures are often used
  - Most medical computer vision models have the same or similar architecture
  - Easier to make transferable attacks
  - Likely that most medical model architectures will be made public for transparency
- Medical data interchange is limited and balkanized
  - Data sharing is spotty between hospitals

Healthcare System

Sources of Vulnerability

- Hospital infrastructure is hard to update
  - Updating medical software is expensive and time-consuming
  - Vulnerabilities in software likely to persist for years
- Medicine contains a mix of technical and non-technical workers
  - Physicians tend to lack computational expertise involved with creating these ML systems
- Biomedical images carry personal signatures that could be used to defend against many simpler attacks, but not against adversarial examples
  - Many biomedical images (i.e. retinal scans, fingerprints, etc.) contain personal identifiers, making it hard to substitue another person's image
  - However, adversarial examples do not have to modify these personal identifiers
- There are many potential adversaries

Demonstration of Adversarial Attacks

Model Construction

- Developed baseline models to classify referable diabetic retinopathy from retinal fundoscopy (Gulshan et al.), pneumothorax from chest-xray (Wang et al. and Rajpurkar et al.), and melanoma from dermoscopic photographs (Esteva et al.)
- Used publicly available data including Kaggle Diabetic Retinopathy dataset, ChestX-Ray14 (Wang et al.), and International Skin Imaging Collaboration website pictures; all involved slight modifications to labels
- Built classifiers by fine-tuning pre-trained ResNet-50 model using SGD
- Augmented data with rotation, flipping, and Mixup

Demonstration of Adversarial Attacks

- Projected Gradient Descent attacks (Madry et al.)

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

- Adversarial patch attacks (Brown et al.)

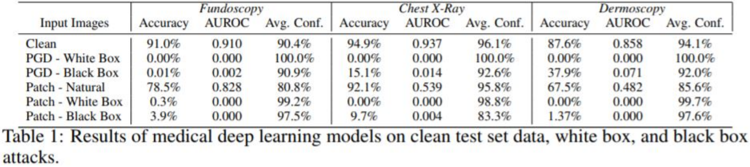
$$\hat{p} = argmax_p E[\log p_{Y|X}(\hat{y}|A(p, X, L, T))]$$

  where $p_{Y|X}$ represents the probability output given input X, L is the location of the patch, T is the transformation, $\hat{y}$ is the target label, and A is the deterministic mapping into adversarially patched image

- Naive patch attacks as control
- Black-box attacks performed by training on independently trained model with same architecture

Demonstration of Adversarial Attacks

Adversarial Attacks



Input Images	*Fundoscopy*			*Chest X-Ray*			*Dermoscopy*		
	Accuracy	AUROC	Avg. Conf.	Accuracy	AUROC	Avg. Conf.	Accuracy	AUROC	Avg. Conf.
Clean	91.0%	0.910	90.4%	94.9%	0.937	96.1%	87.6%	0.858	94.1%
PGD - White Box	0.00%	0.000	100.0%	0.00%	0.000	100.0%	0.00%	0.000	100.0%
PGD - Black Box	0.01%	0.002	90.9%	15.1%	0.014	92.6%	37.9%	0.071	92.0%
Patch - Natural	78.5%	0.828	80.8%	92.1%	0.539	95.8%	67.5%	0.482	85.6%
Patch - White Box	0.3%	0.000	99.2%	0.00%	0.000	98.8%	0.00%	0.000	99.7%
Patch - Black Box	3.9%	0.000	97.5%	9.7%	0.004	83.3%	1.37%	0.000	97.6%

Table 1: Results of medical deep learning models on clean test set data, white box, and black box attacks.

Demonstration of Adversarial Attacks

Adversarial Attacks

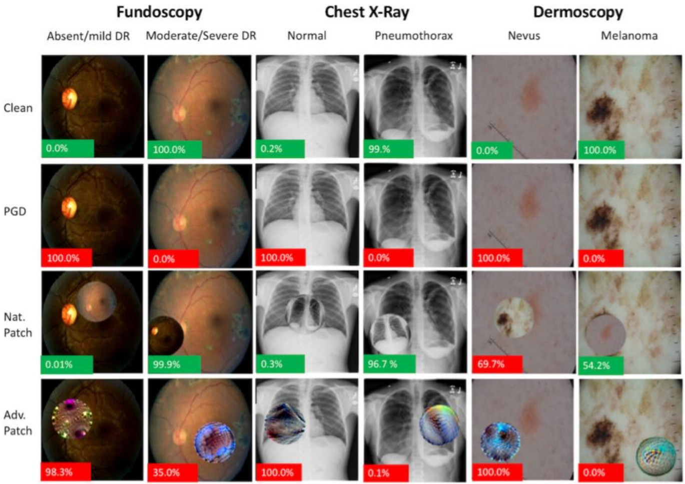


Figure 1: Characteristic results of adversarial manipulation. Each clean image represents the natural image to which the model assigns the highest probability for the given diagnosis. The percentage displayed on the bottom left of each image represents the probability that the model assigns that image of being diseased. Green = Model is correct on that image. Red = Model is incorrect.

Discussion

Hypothetical Examples

- Dermatology
  - Many dermatologists incentivized to perform as many procedures as possible to maximize revenue
  - Insurance company could require that a deep learning system be used on all dermascopy images to determine if surgery is necessary
  - Bad actors could add adversarial noise to only borderline cases to make attacks impossible to detect by human review
- Radiology
  - Thoracic radiology images (i.e. CT scans) are used to measure tumor burden (a secondary endpoint of cancer therapy response)
  - X-ray results can be used to justify heavily reimbursed procedures like biopsies or CT scans

Discussion

Hypothetical Examples

- Ophthalmology
  - Insurance companies are required to cover certain procedures if deemed necessary
  - If a deep learning model was used to determine if procedures were necessary, insurance companies could add slight adversarial noise to barely positive results to save money

Discussion

Possible Research Areas

- Algorithmic defenses: most strategies seem to be limited in scope or data set size, some work has been done on theoretical guarantees of robustness
- In particular, domain-specific defenses have been highly effective - given the standardization of medical procedures, perhaps medical-domain-specific defenses could prove viable
- Infrastructural defenses like storing hashes of new images, having scans done on a 3rd party system to prevent data manipulation; difficult since this would require system-wide standardization
- Ethical tradeoffs: increased robustness to adversaries can lead to lower accuracy

Conclusion

- There is reasonable cause for optimism that deep learning can revolutionize healthcare systems
- It seems inevitable that these systems will become entrenched within the industry
- However, this brings significant opportunity and incentive for fradulent behavior and patient harm
- Outlined the systemic and technological reasons that make the medical system especially vulnerable to adversaries

References

- https://arxiv.org/pdf/1804.05296.pdf