Adversarial Attacks Against Medical Deep Learning Systems

By: Samuel G. Finlayson, Hyung Won Chung, Isaac S. Kohane, Andrew L. Beam

Presented by: Jennifer Fang [Week 05]

Department of Computer Science: University of Virginia

@ https://qdata.github.io/deep2Read/

Adversarial Attacks Against Medical Deep Learning Systems

Overview: Describe how the medical field is uniquely susceptible to adversarial attacks by laying out how and why such attacks could be carried out.

Demonstrate such an attack on a medical system.

Call medical researchers to action by advising ML researchers to take a closer look at these threats.

Key Terms

- **Adversarial examples:** inputs to ML models that have been crafted to force the model to make a classification error
- **Current context**: healthcare costs in the US are huge
- Those costs are largely driven by physician pay.

Incentives for fraud in medical systems

1. Healthcare economy is huge and fraud is already prevalent.
2. Algorithms will most likely make medical reimbursement decisions in the future due to complexity.
3. Algorithms will increasingly determine pharmaceutical and device approvals.

Vulnerabilities of medical systems

1. Actual true diagnosis is often ambiguous.
2. Medical imaging is highly standardized; attackers do not need to account for large amounts of invariance.
3. Commodity network architectures are frequently used; there’s a lack of architectural diversity.
4. Medical data interchange is limited.
5. Hospital infrastructure is hard to update.
6. Medicine is interdisciplinary; includes technical and non-technical workers.

Vulnerabilities of medical systems (cont.)

7. Biomedical images with signatures cannot defend against adversarial attacks.

8. The medical imaging pipeline has many potential attackers that make it vulnerable at many stages.

Creating an attack

1. **Dataset**: Used the Kaggle Diabetic Retinopathy dataset, but only wanted to predict if referable (grade 2 or worse), so they merged/relabeled the dataset.
2. **Adversarial attacks**: implemented both human-imperceptible and patch attacks
3. **Control**: naive patch attack

Attacks

1. **Human-imperceptible**
   - Followed white and black-box PGD strategies
   - PGD attack is an extension of FGSM (Goodfellow)
   - Implemented PGD using CleverHans library

2. **Patch attacks**

   - The learning process of the adversarial patch p’ uses a variant of the expectation over transformation algorithm

Results

All the attacks decreased accuracy dramatically.

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.

Hypothetical examples

1. **Dermatology**: doctors could use ML to increase the number of procedures performed to increase their own profit
2. **Radiology**: ML could be used to guarantee positive trial results to justify heavily reimbursed procedures
3. **Opthamology**: insurers could reduce the # of diagnoses for procedures that are mandatory to be covered to reduce their own costs

Future Work

1. **Algorithmic defenses**: put more effort into researching medical-domain-specific algorithmic defenses

2. **Infrastructural defenses**: imaging devices could immediately store a hashed version of any image they generate as a reference; process raw clinical images on a third-party system to prevent manipulation; implement healthcare system-wide standardization.

3. **Ethical trade-offs**: how does one weigh protecting against adversarial examples against missing true diagnosis?