**Evaluating Evasion Attacks on Super-Resolution Convolutional Neural Networks (SRCNN) in Medical Imaging**

Group Members: Zhaoyu Xu, Letitia Su

**Introduction:**

In the world of medical imaging, accurate interpretation of images is critical. Super-Resolution Convolutional Neural Networks (SRCNN), such as the pre-trained model torchSR, have shown promise in enhancing the resolution of low-quality medical images. However, with the rise of adversarial attacks, there is a need to investigate the robustness of these models.

**Problem Statement:**

Evasion attacks manipulate input data to induce misclassifications or degrade the output of neural networks without changing the model itself. In medical imaging, a successful evasion attack on an SRCNN can cause a high-resolution output image to lose critical details, potentially leading to incorrect diagnoses. The study aims to understand the susceptibility of specific torchSR models to evasion attacks, thereby assessing its reliability in real-world medical applications.

**Significance of the Problem:**

The potential risks of not addressing this problem are dire. A misinterpreted medical image due to adversarial perturbations can lead to:

* Misdiagnosis: Critical conditions might go unnoticed by humans.
* Reduced trust in AI-enhanced medical imaging.

The emerging domain of ML security focuses on ensuring that new technological advancements are not just efficient, but also reliable and robust against malicious activities. As AI models, especially those like SRCNN, become an integral part of critical sectors like healthcare, it's imperative to study their vulnerabilities.

**Key Challenges:**

Variety of Attacks: Different evasion attacks might affect the model in distinct ways. Understanding the nuances of each attack is crucial.

Subtlety of Changes: Perturbations might be indistinguishable to the human eye, making detection challenging.

Domain-specific Nature of Medical Images: Medical images are complex, and the loss of minute details can be significant.

**Plan (Week-by-week for the remaining 8 weeks):**

Week 3 (Current):

* Set up the environment and tools required for the project, ensuring compatibility and functionality.
* Explore similar works of SRCNN in the medical imaging domain to have a better understanding of the training process and evaluation metrics.

Week 4-5:

* Familiarize and re-validate the functioning of the torchSR model with a sample set of medical images to establish a baseline.
* Finetune the pre-trained model on our selected datasets (see dataset section).

Week 6-7:

* Implement and evaluate the FGSM attack on torchSR.
* Implement and evaluate the JSMA and DeepFool attacks.
* Analyze the outcomes and document findings, comparing the efficiency and impact of each method.

Week 8:

* If time allows, explore the implementation of the Carlini & Wagner attack or Limited-memory BFGS method.
* Evaluate the SRCNN's resilience against the previously implemented attacks.
* Have presentable visualizations of the model outputs after perturbations.

Week 9:

* Begin drafting the final report, incorporating findings from the evaluations.
* Ensure detailed documentation of all findings, evaluations, and comparisons.

Week 10:

* Finalize the report.
* Review and refine the report for clarity and comprehensiveness.

Note: While we aim to implement and evaluate at least three attack methods during this project, if time permits, we will extend our investigation to encompass all the aforementioned attack methods, ensuring a comprehensive study.

**Goals and Outcomes:**

By the end of this study, we aim to:

Understand the vulnerabilities of torchSR in the face of evasion attacks.

Propose potential defense mechanisms.

Raise awareness about the importance of security in AI-enhanced medical imaging, emphasizing that advancements should be both innovative and secure.

In essence, this project not only aims to uncover the pitfalls in the current SRCNN models but also paves the way for more robust AI models in medical imaging.

**Datasets:**

BraTS (Brain Tumor Segmentation Challenge)

* Overview: MRI scans of the brain with brain tumor annotations. Though primarily used for segmentation tasks, the data can be used for super-resolution with appropriate pre-processing.
* Link: [BraTS Dataset](https://www.med.upenn.edu/sbia/brats2017/data.html)

IXI Dataset

* Overview: Contains MRI brain scans.
* Link: [IXI Dataset](https://brain-development.org/ixi-dataset/)

LIDC-IDRI (Lung Image Database Consortium and Image Database Resource Initiative)

* Overview: Contains thoracic CT scans annotated for nodules.
* Link: [LIDC-IDRI Dataset](https://www.oasis-brains.org/)

**References:**

Dong, C., Loy, C. C., He, K., & Tang, X. (31 Jul 2015). Image Super-Resolution Using Deep Convolutional Networks. *ArXiv*. /abs/1501.00092

Renkhoff, J., Tan, W., Velasquez, A., Wang, I. Y., Liu, Y., Wang, J., Niu, S., Fazlic, L. B., Dartmann, G., & Song, H. (8 Mar 2023). Exploring Adversarial Attacks on Neural Networks: An Explainable Approach. *ArXiv*. /abs/2303.06032

Mustafa, A., Khan, S. H., Hayat, M., Shen, J., & Shao, L. (2 Sep 2019). Image Super-Resolution as a Defense Against Adversarial Attacks. *ArXiv*. <https://doi.org/10.1109/TIP.2019.2940533>

Carlini, N., & Wagner, D. (22 Mar 2017). Towards Evaluating the Robustness of Neural Networks. *ArXiv*. /abs/1608.04644

Bhambri, S., Muku, S., Tulasi, A., & Buduru, A. B. (7 Feb 2020). A Survey of Black-Box Adversarial Attacks on Computer Vision Models. *ArXiv*. /abs/1912.01667