Attacking CNNs in Histopathology with SNAP: Sporadic and Naturalistic Adversarial Patches (Student Abstract)

Daya Kumar[1], **Abhijith Sharma**[2], **Apurva Narayan**[1,2]

[1]Western University, London, ON, Canada
[2]University of British Columbia, Kelowna, BC, Canada
dkumar55@uwo.ca, sharma86@mail.ubc.ca, apurva.narayan@uwo.ca

Abstract

Convolutional neural networks (CNNs) are being increasingly adopted in medical imaging. However, in the race for developing accurate models, their robustness is often overlooked. This elicits a significant concern given the safety-critical nature of the healthcare system. Here, we highlight the vulnerability of CNNs against a sporadic and naturalistic adversarial patch attack (SNAP). We train SNAP to mislead the ResNet50 model predicting metastasis in histopathological scans of lymph node sections, lowering the accuracy by 27%. This work emphasizes the need for defense strategies before deploying CNNs in critical healthcare settings.

Introduction

Artificial Intelligence (AI) in healthcare is being increasingly integrated into diagnostic and prognostic workflows. Specifically, Convolutional neural networks (CNNs) have become the backbone of various image-processing tasks in computational pathology. However, before deploying them in clinical settings, ensuring their robustness against adversarial attacks is crucial (Foote et al. 2021). An adversarial attack involves adding noise to an image's pixel values to mislead a model's prediction. Although the umbrella of adversarial attacks is quite large, adversarial patch attacks have gained significant traction due to their real-world application (Sharma et al. 2022). Adversarial patches are optimally formulated and localized perturbations in the form of a patch. Given the safety-critical nature of the healthcare system, it is imperative to explore the vulnerabilities of CNNs and develop mitigation strategies (Ghaffari Laleh et al. 2022).

Patch attacks are often visually perceptible to human eyes, making them suspicious and easy to detect by a human-in-the-loop. Although digital attacks (like PGD, FGSM) (Foote et al. 2021) are covert and more potent, the adversarial training of CNNs makes it easier to defend against them. However, the influence of patch attacks is harder to mitigate with adversarial training (Sharma et al. 2022). The existing work on patch attacks mainly considers a single patch, which is relatively large ($\sim$ 5 - 10 % of image). Our study aims to showcase the threat to CNNs from a modified patch attack, designed explicitly by exploiting the fundamental characteristics of histopathological images. We call it Sporadic and Naturalistic Adversarial Patches (SNAP). We propose a generalized framework for designing SNAP to mislead a CNN trained for a histopathological application.

SNAP has three important properties that make it a serious threat to the histopathological domain. First, the sporadic nature of SNAP helps attackers distribute the patch into multiple pieces, textually blending it in the scene. Second, the naturalistic appearance increases the visual fidelity of SNAP in the scene. Third, the current state-of-the-art defenses mostly tackle single patch attacks (Sharma et al. 2022) and the possibility of evading them with sporadic patches is high. Overall, SNAP is designed to evade detection and yet be effective to mislead a CNN.

Methodology

Model Formulation We assume an input image $\boldsymbol{x} \in \mathcal{X} \in \mathbb{R}^{w \times h \times c}$, where w, h, c is the width, height and the number of channels, respectively. The neural network model $\mathcal{F} : \mathcal{X} \rightarrow \mathcal{Y}$ produces the output vector $\vec{y} \in \mathcal{Y} \in \mathbb{R}^n$ for a given image input $\boldsymbol{x}$, where n is the number of classes in the dataset. Each element of $\vec{y}$ is the class classification probability. The class with maximum probability is the final classification label k given as:

$$k = \text{argmax}[\mathcal{F}(\vec{y}|\boldsymbol{x})]\,, \tag{1}$$

We assume k_o is the original label of the image and k_t is the target label to which the attacker wants the model to incorrectly predict on an adversarial image $\boldsymbol{x}' = \boldsymbol{x} + \delta$ for the carefully calculated perturbation δ.

Attack Formulation An adversarial attack is crafted by optimizing a loss function. The loss of classifying the image to the original label $\mathcal{L}(\text{argmax}[\mathcal{F}(\vec{y}|\boldsymbol{x}')], k_o)$ is maximised in an untargeted attack. In a targeted attack, the loss of classifying to the target label $\mathcal{L}(\text{argmax}[\mathcal{F}(\vec{y}|\boldsymbol{x}')], k_t)$ is minimised. The gradient of the loss is used to iteratively update the image's pixel values until satisfactory performance is achieved. We define a mask to limit the perturbation to a small, local region to create an adversarial image as

$$\boldsymbol{x}' = (1 - m) \odot \boldsymbol{x} + m \odot \delta\,, \tag{2}$$

where $\boldsymbol{x}' \in \mathcal{X}, \delta \in [0,1]^{w \times h \times c}$ is the adversarial patch and $m \in M \subset \{0,1\}^{w \times h \times c}$ is the binary mask. The $\odot$ is the Hadamard operator for element-wise matrix multiplication.

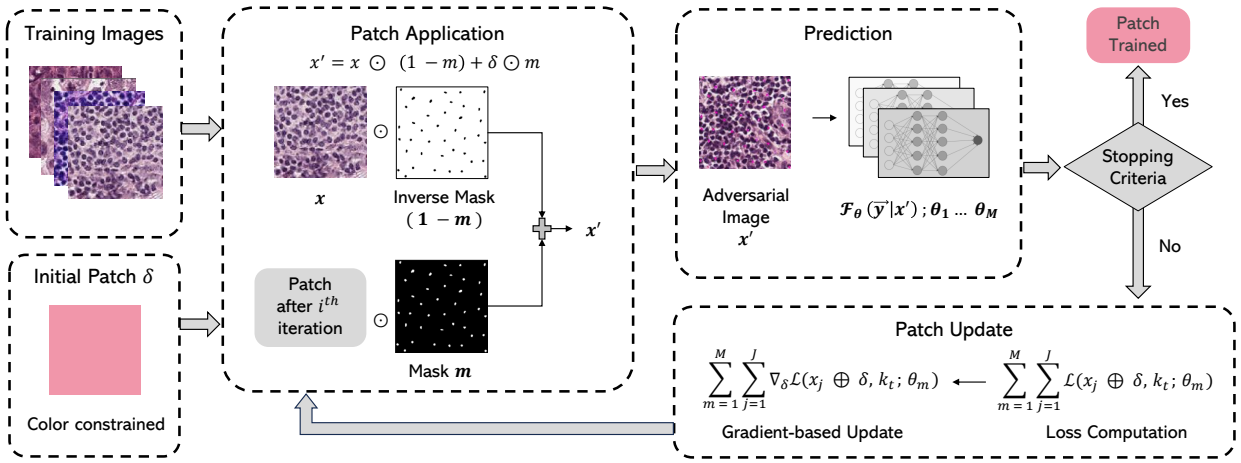

Figure 1: A generalized framework for SNAP training to mislead CNNs used for Histopathological cancer detection.

Mask Design The most important aspect of SNAP is to design an appropriate mask to ensure the sporadic spread of patches all over the image. A sporadic patch approach is used to mitigate the perceptual distortion that arises from training a single patch. Hence, the patch's structure (shape and texture of each patch as a part of the whole sporadic patch) is designed to resemble the shape of a cell to blend into the overall image seamlessly. The mask design depends on the application of interest, and its characteristics.

Patch Training SNAP is a special class of multi-patch attacks designed for histopathological applications. To design patches of severe potency, it is necessary to have simultaneous training of all patches. This helps the patches that are part of SNAP learn collaborative and complementary patterns to perform an attack. As discussed in attack formulation, the patch is trained iteratively to optimize the loss function. The training is carried out until it reaches the maximum iteration or the confidence score of the attack reaches 95%.

An interesting challenge in patch training arises from the unique colors in histopathological images. It is necessary to restrict the color of the trained patch to match the color space of the image for high visual fidelity. During patch optimization, the pixel values are constrained within a specific range determined by the pixel colors in the image. We also found starting with a patch of specific color, based on the median pixel value of the image (pink in our case), rather than random initial values helped in the training process.

Experimental Setup and Results

For this study we use a Histopathologic Cancer Detection (HCD) dataset (Cukierski 2018), containing images of metastatic tissue in histopathologic scans of lymph node sections. Each image is 64 $\times$ 64 and its ground truth label indicates if the central 32 $\times$ 32 region contains at least a single pixel of metastatic tissue. We resize all images to 224 $\times$ 224 for improved resolution, along with transformation for increasing robustness in training. We use 8000 images for training and 2000 for validation. We use ResNet50, a CNN model pre-trained on ImageNet, which we then fine-tune on HCD. The model achieves a training accuracy of 96.28 and a validation accuracy of 89.35. For the attack, the patch is trained to target the class 0, which indicates the absence of metastasis. When the model is attacked using the trained patch it results in the reduction of accuracy from 96.28 to 69.0, which demonstrates the effectiveness of SNAP.

Conclusion

This work demonstrates the threat of adversaries to safety-critical clinical settings. The framework of designing SNAP can be adopted by a malicious attacker to mislead a CNN in a histopathological setting. Some possible directions to extend our work are as follows:

- Extensive empirical evaluation of SNAP against various CNNs and vision transformers (ViT).
- Use of generative techniques (GANs) to train a stronger SNAP with more naturalistic behaviour.
- Develop a defense strategy to mitigate the influence of SNAP in the scene.

References

Cukierski, W. 2018. Histopathologic Cancer Detection. https://kaggle.com/competitions/histopathologic-cancer-detection. Accessed: 2023-02-10.

Foote, A.; Asif, A.; Azam, A.; Marshall-Cox, T.; Rajpoot, N.; and Minhas, F. 2021. Now you see it, now you dont: adversarial vulnerabilities in computational pathology. *arXiv preprint arXiv:2106.08153*.

Ghaffari Laleh, N.; Truhn, D.; Veldhuizen, G. P.; Han, T.; van Treeck, M.; Buelow, R. D.; Langer, R.; Dislich, B.; Boor, P.; Schulz, V.; et al. 2022. Adversarial attacks and adversarial robustness in computational pathology. *Nature communications*, 13(1): 5711.

Sharma, A.; Bian, Y.; Munz, P.; and Narayan, A. 2022. Adversarial patch attacks and defences in vision-based tasks: A survey. *arXiv preprint arXiv:2206.08304*.