Previous research, such as “X-Detect: Explainable Adversarial Patch Detection for Object Detectors in Retail” by Omar Hofman et al. and others, has shown that attackers meticulously select computerized, human-edited inputs—such as 3D image-like patches, for example, toaster-placed in a tabletop scene alongside a banana and notebook—to deceive or manipulate AI models. This type of attack is known as an adversarial patch attack. Attackers deploy adversarial patches with varying characteristics across different scene sections where object detection occurs. We aim to emulate their methodology to understand better how adversarial patches are created and then plan to implement a defense strategy using a Generative Neural Architecture (GNA) with a three-pipeline anomaly detection system to mitigate the use of adversarial patches. This strategy is based on the research “Anomaly Unveiled: Securing Image Classification against Adversarial Patch Attacks” by Nandish Chattopadhyay et al. As the three-step pipeline progresses through the network, the data undergoes a comprehensive process of comparison, analysis, and refinement, all centered around anomaly detection. This process continues until the data is returned in a refined state, effectively neutralizing its potential to function as an adversarial patch.

This approach is accomplished with two neural networks in a generative adversarial setup. Initially, in the segmentation phase, the image is divided into smaller segments, with the GNA model analyzing each segment to detect deviations from the norm, which are indicative of potential adversarial patches. In the second, isolation phase, each segment exhibiting deviations is flagged. Lastly, in the blocking phase, the segments with deviations are modified to contain standardized or neutral values, with the goal being that they would be unable to act as adversarial patches, effectively mitigating their intended effect. This process tests against initial algorithms to determine whether a Generative Neural Architecture can effectively mitigate adversarial patch attacks.

References

Chattopadhyay, N., Guesmi, A., & Shafique, M. (2024). Anomaly Unveiled: Securing Image Classification against Adversarial  Patch Attacks. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2402.06249

Hofman, O., Giloni, A., Hayun, Y., Morikawa, I., Shimizu, T., Elovici, Y., & Shabtai, A. (2023). X-Detect: Explainable adversarial patch detection for object detectors in retail. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2306.08422