Consequences of Adversarial Attacks on AI systems

Michael Christopher – S224830467

One popular use of AI that is highly vulnerable to adversarial attacks is face recognition in identity verification and security systems. To extract and compare face traits with a stored database for authentication or surveillance reasons, facial recognition AI uses deep learning models, most often convolutional neural networks (CNNs). This technology has been widely used in law enforcement to identify suspects, at airports to verify passengers, and in cell phones to unlock them using biometrics. Its main benefits are its speed, automation, and removal of manual verification procedures. However, because it depends on complex pattern recognition, it can be manipulated adversarially, leading to the system misidentifying a person based on subtle, undetectable changes to a picture.

Convolutional neural networks (CNNs) powered facial recognition systems are being used more for security, surveillance, and authentication. However, because of their strong reliance on high-dimensional feature extraction, they are more vulnerable to adversarial attacks, which intentionally create changes to input images that result in incorrect categorization without being noticeable to humans. Attackers can launch such attacks because of the model's decision boundaries' fundamental weaknesses and their capacity to take advantage of cross-model transferability (Kong et al., 2021). This is made worse in the field of face recognition by the ease with which attackers may get facial photos from public sources, allowing them to create input alterations that trick or avoid systems (Vakhshiteh et al., 2021).

Attacks like this have serious consequences. Adversarial instances in security contexts might enable unauthorized people to pose as authorized users, getting around access controls in high-security institutions, financial apps, and cell phones (Alparslan et al., 2020). By creating fake matches, they might help law enforcement catch criminals or sway investigations. When applied to vital infrastructure, the cumulative impact of such vulnerabilities compromises society and reduces confidence in AI-driven authentication systems (Baniecki & Biecek, 2024). Furthermore, even little disruptions can significantly reduce identification accuracy, as Goswami et al. (2018) point out, which means that widespread deployment without strong preventative measures increases the attack surface.

Technical and procedural methods are being used to reduce these hazards. Adversarial training, which involves adding adversarial cases to training datasets, has demonstrated potential for improving model resilience despite its high computing cost (Dong et al., 2019). According to Kong et al. (2021), defensive distillation, feature squeezing, and input preprocessing can help reduce vulnerability by filtering harmful noise and smoothing decision boundaries. To mitigate assaults, real-world systems can use multimodal authentication, which combines face recognition with contextual checks or other biometrics, in addition to model-level protections (Vakhshiteh et al., 2021). In addition to assisting human supervision, transparent explainable AI techniques can identify anomalies in system decision-making that can point to adversary manipulation (Baniecki & Biecek, 2024).

Overall, because of the high stakes involved in their applications as well as technical model flaws, adversarial attacks on AI facial recognition systems pose a serious threat. To combat these attacks, a multi-layered defensive approach is required, combining algorithmic resilience with operational security measures to guarantee AI's continued efficacy and reliability in security-sensitive settings.

References:

Alparslan, Y., Alparslan, K., Keim-Shenk, J., Khade, S., & Greenstadt, R. (2020). Adversarial attacks on convolutional neural networks in facial recognition domain. *arXiv preprint arXiv:2001.11137*.

Baniecki, H., & Biecek, P. (2024). Adversarial attacks and defenses in explainable artificial intelligence: A survey. *Information Fusion*, *107*, 102303.

Dong, Y., Su, H., Wu, B., Li, Z., Liu, W., Zhang, T., & Zhu, J. (2019). Efficient decision-based black-box adversarial attacks on face recognition. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7714-7722).

Goswami, G., Ratha, N., Agarwal, A., Singh, R., & Vatsa, M. (2018, April). Unravelling robustness of deep learning based face recognition against adversarial attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 32, No. 1).

Kong, Z., Xue, J., Wang, Y., Huang, L., Niu, Z., & Li, F. (2021). A survey on adversarial attack in the age of artificial intelligence. *Wireless Communications and Mobile Computing*, *2021*(1), 4907754.

Vakhshiteh, F., Nickabadi, A., & Ramachandra, R. (2021). Adversarial attacks against face recognition: A comprehensive study. *IEEE Access*, *9*, 92735-92756.