SIT 799

Task 4.1: Consequences of Adversarial Attacks on AI systems

Introduction

Surveillance systems where AI based Face Recognition Systems and Face Detection Systems are now used for human surveillance, either without human intervention or with minimal human intervention. Increased awareness of AI based surveillance makes perpetrators utilise various methods to fool such systems including physical and software-based techniques.

Causes of adversarial attacks

Generally, the motivation to fool AI based face recognition system arises from following scenarios:

- **Privacy Concerns**: Due to increased deployment of AI based Face Recognition System in public and private spaces, there is a threat to maintain an individual's privacy which leads to an ongoing debate around privacy vs security. Irrespective of the fact that such solutions pose a threat to privacy, they have proven to be an asset for monitoring agencies to prevent crime. Owing to these fears there is a large sub section of society that does not approve of privacy infringing measures such as Face Recognition Systems. This leads to employing measures such as using adversarial attacks to maintain an individual's privacy.
- **Crime Evasion**: This scenario is the most threating motivation to fool such systems, criminals can use adversarial methods to fool an AI system by misreporting the identity of the perpetrator.
- **Fooling proctoring systems:** With the ongoing COVID-19 pandemic, there is an increase in online assessments, some of these are very critical. To maintain the sanctity of these tests, organizations employ software based proctoring solutions to prevent cheating. Some candidates with dishonest motivations might use adversarial attacks to prevent cheating from being reported and can gain an unfair advantage over other candidates.

Methods and Consequences of adversarial attacks

In [1], [2] and [3] , we can see how adversarial attacks are used to accomplish the goal of fooling face recognition systems.

- **Face Spoofing**: Generally, all face recognition methods are dependent on face detection module, approaches such as using **adversarial patches**, or an **adversarial image** might lead the face detector to believe that there is a face in the frame when there is none. This can be used in online proctoring scenarios where the system cannot monitor the candidate. An adversarial patch or adversarial image consists of an image which leads the model to output a desired result when the object is not present.

In [1], authors show two methods which can be used to fool FRS solutions:

- **Impersonation**: This method uses an adversarial image generated to map a face to a known face, this leads the FRS solution to misidentify the individual.
- **Dodging**: This method is like impersonation, but instead of mapping to a known face the adversarial example is generated to map the face to a random face.

Both these approaches hide the original identity of the individual from the FRS solution. Due to the transferable nature of the features that are learned using Deep learning [2] show how these methods can be used to fool SOTA methods like Arcface.

Potential prevention against such attacks

- **Liveliness check**: For preventing face detection systems from adversarial attacks, liveliness checks, either model based (classify whether the image is live or printed) or action based (request user to perform facial movements by following direction on screen) can be used to ensure there is a valid face present instead of an adversarial image.
- **Adversarial Training**: Face recognition models can be trained on large amounts of adversarial faces to map similar faces to each other by overlooking the perturbations in the image.
- **Model Encryption**: Because adversarial attacks depend on extracting information from the model, encrypting the model so that they can not be used without required authentication can act as a deterrent to such attacks.

Conclusion

This report shows how adversarial attacks can be used to fool Face Recognition and Face detection systems. Adversarial attacks utilise the fact that vision systems are highly sensitive to small changes in the image, this also shows that an increased focus into developing models that have a different strategy to extract features to represent faces is required for a robust AI based Face Recognition solution.

References

[1] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 2016.

[2] Y. Xu, K. Raja, R. Ramachandra, and C. Busch, "Adversarial attacks on Face Recognition Systems," in *Handbook of Digital Face Manipulation and Detection*, Cham: Springer International Publishing, 2022, pp. 139–161.

[3] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *2017 IEEE Symposium on Security and Privacy (SP)*, 2017, pp. 39–57.