Artificial Immune System of Secure Face Recognition Against Adversarial Attacks (Abstract Reprint)

Min Ren[1], **Yunlong Wang**[2], **Yuhao Zhu**[3], **Yongzhen Huang**[1], **Zhenan Sun**[2], **Qi Li**[2], **Tieniu Tan**[2]

[1]School of Artificial Intelligence, Beijing Normal University, Beijing, China
[2]State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS), Institute of Automation, Chinese Academy of Sciences, Beijing, China
[3]Postgraduate Department, China Academy of Railway Sciences, Beijing, China

Abstract Reprint. This is an abstract reprint of the journal article by Ren, Wang, Zhu, Huang, Sun, Li, and Tan (2024).

Abstract

Deep learning-based face recognition models are vulnerable to adversarial attacks. In contrast to general noises, the presence of imperceptible adversarial noises can lead to catastrophic errors in deep face recognition models. The primary difference between adversarial noise and general noise lies in its specificity. Adversarial attack methods give rise to noises tailored to the characteristics of the individual image and recognition model at hand. Diverse samples and recognition models can engender specific adversarial noise patterns, which pose significant challenges for adversarial defense. Addressing this challenge in the realm of face recognition presents a more formidable endeavor due to the inherent nature of face recognition as an open set task. In order to tackle this challenge, it is imperative to employ customized processing for each individual input sample. Drawing inspiration from the biological immune system, which can identify and respond to various threats, this paper aims to create an artificial immune system to provide adversarial defense for face recognition. The proposed defense model incorporates the principles of antibody cloning, mutation, selection, and memory mechanisms to generate a distinct antibody for each input sample, wherein the term antibody refers to a specialized noise removal manner. Furthermore, we introduce a self-supervised adversarial training mechanism that serves as a simulated rehearsal of immune system invasions. Extensive experimental results demonstrate the efficacy of the proposed method, surpassing state-of-the-art adversarial defense methods.

References

Ren, M.; Wang, Y.; Zhu, Y.; Huang, Y.; Sun, Z.; Li, Q.; and Tan, T. 2024. Artificial Immune System of Secure Face Recognition Against Adversarial Attacks. *International Journal of Computer Vision*, 132: 5718–5740.