ABSTRACT

ASSESSING THE EFFECTIVENESS OF K-WINNER-TAKES-ALL ACTIVATION FUNCTION ON DEFENDING DEEPFAKE DETECTOR MODEL FROM WHITEBOX ADVERSARIAL ATTACK

Allen Nathael Ardy

20/457764/PA/19802

This study examines the K-Winner-Takes-All (K-WTA) activation function's effectiveness in enhancing deepfake detection models' robustness against whitebox adversarial attacks. Adversarial attacks, which subtly alter input data to mislead neural network models, pose significant challenges to AI system security. Traditional activation functions like ReLU often fail to defend against these sophisticated attacks. This research utilizes the MesoNet model and the FaceForensics++ dataset to evaluate the K-WTA activation function and its derivatives, including Average-Takes-All and K-Losers-Takes-All.

The results demonstrate that K-WTA and its derivatives significantly enhance model robustness without compromising performance. By introducing gradient discontinuities and enforcing sparsity, K-WTA and its variants limit the effectiveness of gradient-based adversarial attacks, improving neural network security. These findings suggest that integrating K-WTA and its derivatives into deepfake detection models offers a promising defence mechanism against adversarial threats, contributing to the development of more secure AI systems.

Keywords: Deepfake, Adversarial Attack, K-WTA, AI Security, Model Robustness

ABSTRAK

MENILAI EFEKTIVITAS FUNGSI AKTIVASI K-WINNER-TAKES-ALL DALAM MELINDUNGI MODEL DETEKTOR DEEPFAKE DARI SERANGAN ADVERSARIAL WHITEBOX

Allen Nathael Ardy
20/457764/PA/19802

Penelitian ini mengkaji efektivitas fungsi aktivasi *K-Winner-Takes-All* (K-WTA) dalam meningkatkan ketahanan model deteksi deepfake terhadap serangan whitebox adversarial. Serangan adversarial yang mengubah data input secara halus untuk menyesatkan model jaringan neural menimbulkan tantangan signifikan bagi keamanan sistem AI. Fungsi aktivasi tradisional seperti ReLU sering gagal dalam mempertahankan diri dari serangan canggih ini. Penelitian ini menggunakan model MesoNet dan dataset FaceForensics++ untuk mengevaluasi fungsi aktivasi K-WTA dan turunannya, termasuk *Average-Takes-All* dan *K-Losers-Takes-All.*

Hasil penelitian menunjukkan bahwa K-WTA dan turunannya secara signifikan meningkatkan ketahanan model tanpa mengorbankan kinerja. Dengan memperkenalkan diskontinuitas gradien dan penyebaran nilai *tensor*, K-WTA dan variasinya membatasi efektivitas serangan adversarial berbasis gradien, meningkatkan keamanan jaringan neural. Temuan ini menunjukkan bahwa mengintegrasikan K-WTA dan turunannya ke dalam model deteksi deepfake menawarkan mekanisme pertahanan yang menjanjikan terhadap ancaman adversarial, serta berkontribusi pada pengembangan sistem AI yang lebih aman.

Keywords: Deepfake, Serangan Adversarial, K-WTA, Keamanan AI, Ketahanan Model