INNOVATION PARTNERSHIPS
UNIVERSITY OF MICHIGAN

D4: Detection of Adversarial Diffusion Deepfakes Using Disjoint Ensembles

TECHNOLOGY NUMBER: 2024-376

Technology ID
2024-376

Category
Software
MOSS - Michigan Open Source Support

Inventor
Atul Prakash
Ryan Feng
Neal Mangaokar
Ashish Hooda
Somesh Jha
Kassem Fawaz

Further information
Ashwathi Iyer
ashwathi@umich.edu

View online page

OVERVIEW

D4 enhances deepfake detection robustness using disjoint frequency spectrum model ensembles

- Outperforms traditional defenses by reducing adversarial attack effectiveness via frequency partitioning
- Digital forensics, social media content validation, security, authenticity verification, online trust

BACKGROUND

The rapid advancement of generative diffusion models has made it increasingly easy to create convincing deepfake images, posing major challenges to digital content authenticity. Historically, methods to detect deepfakes have largely relied on deep neural networks trained to differentiate between authentic and generated images. Adversarial training has become a standard defense mechanism; it exposes models to perturbed, adversarial examples during training to bolster resilience. However, attackers have adapted, using imperceptible perturbations specifically crafted to evade these detectors, undermining their reliability. Furthermore, these existing techniques tend to overfit to specific attack patterns and often fail to generalize, especially under black-box adversarial settings and against new, unseen generative methods. Given the escalating sophistication of both generative models and adversarial attacks, there is a critical, growing need for detection methods that are robust to such manipulations and can generalize well across diverse attack strategies and data

distributions.

INNOVATION

The Disjoint Diffusion Deepfake Detection (D4) method introduces a fundamentally new approach to adversarially robust deepfake detection. D4 employs an ensemble of detection models, each specialized on a unique, disjoint subset of the image frequency spectrum. Saliency partitioning techniques ensure each model focuses on distinct frequency features, minimizing overlap and redundancy among the ensemble. This setup limits the attacker's ability to simultaneously generate adversarial perturbations that deceive all models, shrinking the adversarial subspace and thereby making successful attacks far more difficult even in challenging black-box scenarios. D4's effectiveness is both theoretically validated (by provably reducing the attackers' effective space) and empirically demonstrated—showing significant improvements over prior state-of-the-art solutions for diffusion-generated deepfake images, including resilience to attacks from unseen generative techniques and data distributions. Real world applications span digital forensics, social media integrity, law enforcement, journalism, and any setting where image authenticity is critical.

ADDITIONAL INFORMATION

PROJECT LINKS:

- D4 Github

DEPARTMENT/LAB:

- Atul Prakash, Computer Science and Engineering (CSE)

LICENSE: