Engane a IA: Desenvolvendo Perturbações Adversariais Contra Classificadores de DeepFake

Eduardo Verissimo Faccio, Guilherme Ferreira Lourenço

[1]Instituto de Ciência e Tecnologia – Universidade Federal de São Paulo (UNIFESP)
São José dos Campos – SP – Brasil

{verissimo.eduardo,gflourenco}@unifesp.br

Resumo. *Este trabalho explora a vulnerabilidade e a robustez de redes neurais diante de ataques adversariais imperceptíveis aos humanos. Utilizando a base de dados "140k Real and Fake Faces", treinou-se uma rede neural convolucional, uma mini UNET e um XGBoost para a classificação de imagens faciais, atingindo alta acurácia em condições normais. Em seguida, desenvolveu-se um ataque adversarial baseado em uma arquitetura UNET modificada, capaz de gerar perturbações sutis que comprometem significativamente a performance do classificador. Os resultados demonstram que, mesmo com perturbações invisíveis a olho nu, a confiabilidade dos sistemas de reconhecimento facial pode ser severamente comprometida, ressaltando a importância de desenvolver estratégias para aumentar a resiliência dos classificadores.*

Palavras-chave*: Ataques adversariais, robustez, redes neurais convolucionais, perturbações imperceptíveis, reconhecimento facial, XGBoost.*

1. Introdução

O avanço das técnicas de aprendizado profundo tem permitido que sistemas de reconhecimento facial alcancem níveis de acurácia sem precedentes, contribuindo para uma ampla gama de aplicações, desde segurança pública até interfaces de usuário baseadas em biometria [Parkhi et al. 2015]. Entretanto, apesar do desempenho elevado, diversos estudos demonstraram que redes neurais são inerentemente vulneráveis a perturbações adversariais - pequenas modificações imperceptíveis a olho nu que podem levar a classificações equivocadas [Szegedy et al. 2014, Goodfellow et al. 2015].

Neste contexto, o presente trabalho investiga a robustez de modelos de classificação de imagens faciais quando expostos a ataques adversariais sutis. Utilizando a base de dados "140k Real and Fake Faces", foram treinados três classificadores distintos - uma rede neural convolucional (CNN), uma versão reduzida de UNET e um classificador baseado em XGBoost [Tianqi and Carlos 2016] - que, sob condições normais, alcançaram alta performance. Em seguida, propõe-se um ataque adversarial que se baseia em uma arquitetura UNET modificada, cujo objetivo é gerar perturbações imperceptíveis que comprometam a confiabilidade dos classificadores.

Ao integrar modelos de naturezas distintas na análise, este estudo busca identificar possíveis diferenças na robustez dos sistemas frente a ataques adversariais, evidenciando as limitações dos métodos de classificação atuais e a necessidade de desenvolver estratégias de defesa que aumentem a resiliência dos sistemas de reconhecimento facial. A abordagem proposta não só contribui para uma compreensão mais aprofundada das vulnerabilidades dos modelos de aprendizado profundo, mas também sugere caminhos

para a implementação de mecanismos de defesa que possam mitigar os impactos desses ataques em aplicações críticas.

2. Conclusão

Referências

Goodfellow, I. J., Shlens, J., and Szegedy, C. (2015). Explaining and harnessing adversarial examples.

Parkhi, O. M., Vedaldi, A., and Zisserman, A. (2015). Deep face recognition. *Proceedings of the British Machine Vision Conference*.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. (2014). Intriguing properties of neural networks.

Tianqi, C. and Carlos, G. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794. ACM.