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 e Carlos 2016] a - 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. Referencial Teórico

2.1. Redes Neurais Convolucionais e Reconhecimento Facial

Redes neurais convolucionais (CNNs) têm se destacado no processamento de imagens devido à sua capacidade de extrair automaticamente características hierárquicas, desde padrões simples, como bordas e texturas, até estruturas complexas, essenciais para o reconhecimento facial. Estudos como o de Parkhi et al. (2015) demonstram que a utilização de arquiteturas profundas permite alcançar altos índices de acurácia, tornando as CNNs a escolha preferencial para aplicações em biometria.

2.2. Ataques Adversariais

Apesar do elevado desempenho, as CNNs são vulneráveis a perturbações adversariais - pequenas modificações intencionais e imperceptíveis na entrada que podem levar a classificações errôneas. Szegedy et al. (2013) foram pioneiros ao evidenciar essa fragilidade, enquanto Goodfellow et al. (2014) apresentaram o Fast Gradient Sign Method (FGSM) para gerar tais perturbações. Esses ataques exploram a alta dimensionalidade dos dados e a complexidade dos espaços de decisão dos modelos, comprometendo a confiabilidade dos sistemas de reconhecimento facial.

2.3. Estratégias de Geração de Perturbações

Diversas técnicas de ataque adversarial têm sido propostas, muitas baseadas na otimização da função de perda do modelo. Entre elas, o FGSM e suas variações mostram alta eficácia na geração de exemplos adversariais. No presente trabalho, adota-se uma abordagem inovadora que utiliza uma arquitetura UNET modificada para criar perturbações sutis. Essa rede adversarial é treinada para gerar modificações que se mantenham dentro de um intervalo pré-definido (por exemplo, [epsilon]), garantindo que as alterações sejam imperceptíveis ao olho humano, mas capazes de degradar significativamente a performance dos classificadores.

2.4. Outros Classificadores e Robustez dos Modelos

Além das CNNs, outros classificadores, como UNET e o XGBoost, têm sido empregados em problemas de classificação de imagens. O XGBoost, conforme descrito por Chen Guestrin (2016), utiliza técnicas de boosting que o tornam robusto e eficiente, especialmente em cenários com dados de alta dimensionalidade. A comparação entre esses modelos permite avaliar a variabilidade na robustez frente a ataques adversariais, demonstrando que, mesmo que a mesma perturbação seja aplicada, a sensibilidade de cada modelo pode variar, evidenciando a necessidade de estratégias de defesa que aumentem a resiliência global dos sistemas de reconhecimento facial.

2.5. Transferibilidade dos Ataques e Estratégias de Defesa

Um aspecto crítico dos ataques adversariais é a sua capacidade de transferência, isto é, a perturbação gerada para um modelo pode afetar outros modelos, mesmo que estes tenham arquiteturas distintas. Essa propriedade aumenta a ameaça dos ataques em cenários do

mundo real, onde o atacante pode não ter acesso direto ao modelo alvo. Compreender e mitigar essa transferência é fundamental para o desenvolvimento de mecanismos de defesa robustos que assegurem a integridade dos sistemas de reconhecimento facial em ambientes críticos.

3. Conclusão

Referências

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

Parkhi, O. M., Vedaldi, A., e 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., e Fergus, R. (2014). Intriguing properties of neural networks.

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