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 o desempenho 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 [Ronneberger et al. 2015] e um classificador baseado em XGBoost [Tianqi e Carlos 2016] a - que, sob condições normais, alcançaram alto desempenho. Em seguida, propõe-se um ataque adversarial que se baseia em uma arquitetura UNET parecida com o classificador criado, 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 defensivas 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, baseadas na otimização da função de perda do modelo. Entre elas, o FGSM e suas variações mostram alta eficácia na geração de exemplos adversariais. Neste trabalho, adota-se uma abordagem inovadora que utiliza uma arquitetura UNET modificada para criar perturbações sutis. A escolha da UNET como base se justifica por sua capacidade comprovada de capturar detalhes contextuais e preservar informações espaciais, características que foram originalmente exploradas por Ronneberger et al. (2015) em seu trabalho seminal. Adaptamos essa estrutura para garantir que as modificações permaneçam num intervalo pré-definido (por exemplo, $[-\epsilon,\epsilon]$), assegurando que as perturbações sejam imperceptíveis ao olho humano, mas capazes de degradar significativamente o desempenho 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.

A versão reduzida da UNET utilizada neste trabalho deriva do modelo originalmente proposto por Ronneberger para segmentação de imagens biomédicas, demonstrando alta eficiência na extração de características relevantes [Ronneberger et al. 2015]. Esse modelo revolucionou o campo da segmentação de imagens ao demonstrar que uma rede com caminhos de contração e expansão, aliada a conexões de *skip*, pode extrair e preservar informações espaciais cruciais.

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 defensivas que aumentem a resiliência global dos sistemas de reconhecimento facial.

2.5. Transferibilidade dos Ataques e Estratégias defensivas

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 [Tramèr et al. 2021]. 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. Metodologia

A abordagem adotada neste trabalho visa investigar a robustez de classificadores de imagens faciais frente a ataques adversariais imperceptíveis. Para isso, a metodologia foi dividida em três etapas principais: (i) preparação e treinamento dos modelos, (ii) desenvolvimento do ataque adversarial e (iii) avaliação dos resultados.

3.1. Preparação dos Dados e Treinamento dos Modelos

A base de dados utilizada foi a *"140k Real and Fake Faces"*, composta por imagens reais e geradas, que foram previamente divididas em conjuntos de treinamento, validação e teste. A seguir, três modelos foram treinados para a classificação de imagens faciais:

- **Rede Neural Convolucional (CNN):** Implementada em PyTorch, esta arquitetura é composta por camadas convolucionais, seguidas por camadas de pooling, normalização e camadas totalmente conectadas.
- **Mini UNET:** Uma versão reduzida da tradicional arquitetura UNET foi desenvolvida para explorar sua capacidade em extrair características e classificar as imagens, mantendo uma complexidade computacional inferior.
- **XGBoost:** Utilizando o framework XGBoost [Tianqi e Carlos 2016], foi treinado um classificador baseado em boosting, que se mostrou eficiente para problemas de alta dimensionalidade.

Todos os modelos foram treinados sob condições normais, atingindo altos índices de acurácia no conjunto de teste.

3.2. Desenvolvimento do Ataque Adversarial

Após o treinamento dos classificadores, desenvolveu-se um ataque adversarial visando gerar perturbações sutis e imperceptíveis que comprometem a confiabilidade dos modelos. Para isso, foi utilizada uma arquitetura UNET modificada, denominada rede adversarial, que recebe como entrada uma imagem normalizada (em escala $[-1, 1]$) e gera uma

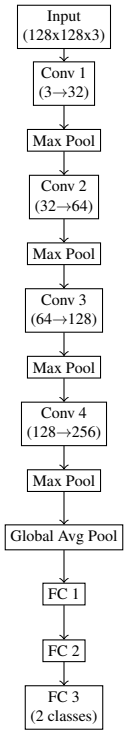

Figura 1. Diagrama das arquiteturas dos modelos utilizados no trabalho.

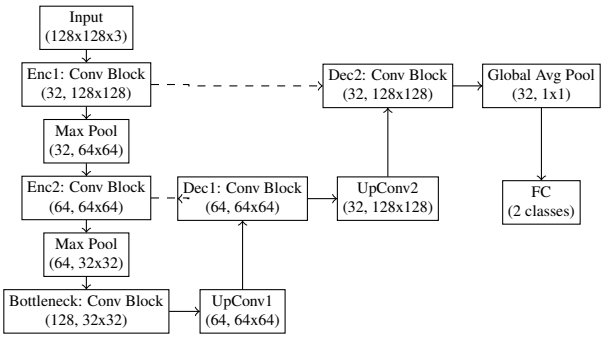

Figura 2. Diagrama da arquitetura da Mini UNET.

perturbação. Essa perturbação é então limitada por um mecanismo de *clamp*, para garantir que as alterações não excedam um intervalo pré-definido $[-\epsilon, \epsilon]$. Formalmente, a imagem adversarial adv_x é obtida através da equação:

$$\text{adv\_x} = \text{clamp}\Big(x + \text{clamp}\big(\text{Perturbação}(x), -[\epsilon], [\epsilon]\big), -1, 1\Big)$$

onde x representa a imagem original. Essa abordagem assegura que as modificações sejam imperceptíveis ao olho humano, mas capazes de induzir erros nos classificadores.

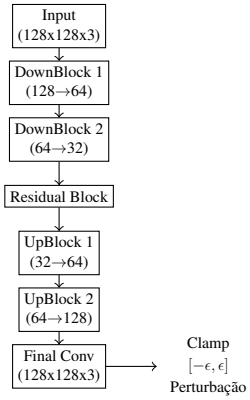

Figura 3. Diagrama das arquiteturas dos modelos utilizados no trabalho.

3.3. Avaliação Cruzada do Ataque Adversarial

Além da avaliação convencional dos ataques adversariais, foi realizado um estudo adicional para investigar a generalização das perturbações geradas pelo ataque e a eficácia dos ataques adversariais em diferentes contextos. Esse estudo consiste em treinar a rede adversarial contra um modelo específico e, posteriormente, testá-la contra outro classificador não visto durante o treinamento do ataque.

O objetivo dessa abordagem é determinar se as perturbações adversariais são específicas ao modelo para o qual foram treinadas ou se possuem propriedades transferíveis, capazes de enganar diferentes arquiteturas de classificação. Para isso, foi realizado o seguinte procedimento experimental:

Treinamento da rede adversarial para enganar um classificador específico, escolhendo entre CNN ou UNET. Geração das imagens adversariais utilizando a rede treinada. Aplicação dessas imagens adversariais no outro classificador para avaliar o impacto e o desempenho na detecção das imagens modificadas. Medição da eficácia do ataque por meio de métricas como acurácia, precisão, recall e F1-score, comparando os resultados pré e pós-ataque em cada cenário.

Com essa avaliação, é possível responder a duas questões fundamentais:

As perturbações adversariais são específicas ao modelo utilizado no treinamento? As perturbações podem ser generalizadas para outros classificadores, sugerindo uma vulnerabilidade compartilhada entre diferentes arquiteturas? Se um ataque treinado contra um modelo não afeta significativamente outro classificador, isso indica que a adversarial

possui um comportamento fortemente dependente da arquitetura alvo. No entanto, se a adversarial treinada em um modelo reduz drasticamente a acurácia do outro, isso sugere a existência de padrões estruturais comuns entre os modelos, tornando-os vulneráveis a ataques transferíveis.

3.4. Avaliação dos Resultados

A robustez dos modelos foi avaliada comparando a acurácia obtida sob condições normais com a acurácia após a aplicação dos ataques adversariais. Adicionalmente, a técnica GradCAM foi utilizada para visualizar as regiões de maior atenção dos modelos, permitindo identificar quais áreas (por exemplo, olhos e lábios) são mais suscetíveis às perturbações [Selvaraju et al. 2017].

A avaliação abrangeu os três classificadores de forma integrada, possibilitando comparar a sensibilidade de cada modelo frente às perturbações geradas pela rede adversarial. Os resultados quantitativos (como a redução percentual da acurácia) e as análises qualitativas (obtidas via GradCAM) forneceram uma visão abrangente sobre as vulnerabilidades dos sistemas de reconhecimento facial.

4. Experimentos e Resultados

4.1. Configuração Experimental

Foram realizados experimentos para avaliar a robustez dos três classificadores (CNN, Mini UNET e XGBoost) sob três condições distintas:

- **Condições Normais:** Treinamento e avaliação dos modelos sem a aplicação de ataques adversariais.
- **Condições Adversariais:** Avaliação dos modelos após a aplicação do ataque adversarial, que utiliza uma arquitetura UNET modificada para gerar perturbações sutis e imperceptíveis.
- **Condições Cruzadas do Ataque Adversarial:** Treino do ataque adversarial contra um modelo específico e depois aplicado a outro classificador não visto durante seu treinamento. Aplicado aos modelos CNN e Mini UNET.

Todos os modelos foram treinados utilizando a base de dados ”140k Real and Fake Faces”, que possui 100 mil dados para treino, 20 mil para validação e 20 mil para teste. Todos os resultados apresentados nas próximas seções foram calculados utilizando os dados do teste.

Para avaliação do desempenho dos modelos, foram consideradas métricas como acurácia, precisão, recall e F1-Score. Além disso, técnicas de visualização (GradCAM) foram empregadas para analisar qualitativamente as regiões de maior atenção dos modelos.

4.2. Resultados Quantitativos

A Tabela 1 apresenta o desempenho dos três modelos avaliados – CNN, Mini UNET e XGBoost – comparando as acurácias obtidas em condições normais e após a aplicação dos ataques adversariais. Observa-se que tanto a CNN quanto a Mini UNET alcançaram acurácias muito elevadas (99.0% e 97.56%, respectivamente) em condições normais, enquanto o XGBoost obteve desempenho consideravelmente inferior (56.20%). Entretanto,

sob ataque adversarial, os modelos baseados em redes neurais (CNN e Mini UNET) sofreram reduções expressivas na acurácia, atingindo 62.78% e 58.48%, correspondendo a reduções de 36.22% e 39.08%, respectivamente. Por outro lado, o XGBoost apresentou uma diminuição mais modesta (redução de 8.20%), considerando que sua acurácia base já estava inferior à dos demais.

Tabela 1. Desempenho dos modelos sob condições normais e adversariais

Modelo	Acurácia Normal (%)	Acurácia Adversarial (%)	Redução (%)
CNN	99.0%	62.78%	36.22%
Mini UNET	97.56%	58.48%	39.08%
XGBoost	56.20%	48.00%	8.20%

A Tabela 2 detalha outras métricas de desempenho dos modelos sob ataque adversarial. Nota-se que, para os três modelos, a precisão permanece muito elevada (próxima de 100%), indicando que, quando os modelos classificam uma imagem como falsa, essa é com certeza uma imagem falsa. Contudo, os valores de recall são significativamente baixos – 25.57% para a CNN, 16.97% para a Mini UNET e 13.92% para o XGBoost – o que reflete uma alta taxa de falsos negativos. Portanto, resultados indicam que a capacidade de identificar corretamente todas as amostras adversariais é comprometida.

Tabela 2. Métricas de desempenho sob ataque adversarial

Modelo	Precisão (%)	Recall (%)	F1-Score (%)
CNN	100%	25.57%	0.4073
Mini UNET	99.94%	16.97%	0.2901
XGBoost	99.70%	13.92%	0.2443

A Tabela 3 apresenta os resultados da avaliação cruzada, na qual o ataque adversarial foi treinado em um modelo e posteriormente aplicado a outro, com o intuito de investigar a transferência das perturbações entre arquiteturas distintas. Observa-se que, embora ambos os modelos de treinamento do ataque tenham impactado os classificadores base, a degradação do desempenho foi menos pronunciada em comparação com os valores apresentados na Tabela 2. Ademais, os resultados sugerem que o modelo CNN demonstra maior robustez, evidenciado pela redução substancialmente menor no desempenho quando submetido ao ataque adversarial.

Tabela 3. Métricas de desempenho utilizando a avaliação cruzada.

Treinado	Usado	Acurácia (%)	Precisão (%)	Recall (%)	F1-Score (%)
CNN	UNET	84.14%	99.53%	68.61%	0.8122
UNET	CNN	93.16%	99.91%	86.39%	0.9265

4.3. Resultados Qualitativos

Para complementar a análise quantitativa, empregou-se a técnica GradCAM para identificar as regiões de atenção dos modelos. A Figura 4 ilustra exemplos de imagens originais e suas versões adversariais, destacando alterações nas áreas de interesse (por exemplo, olhos e lábios).

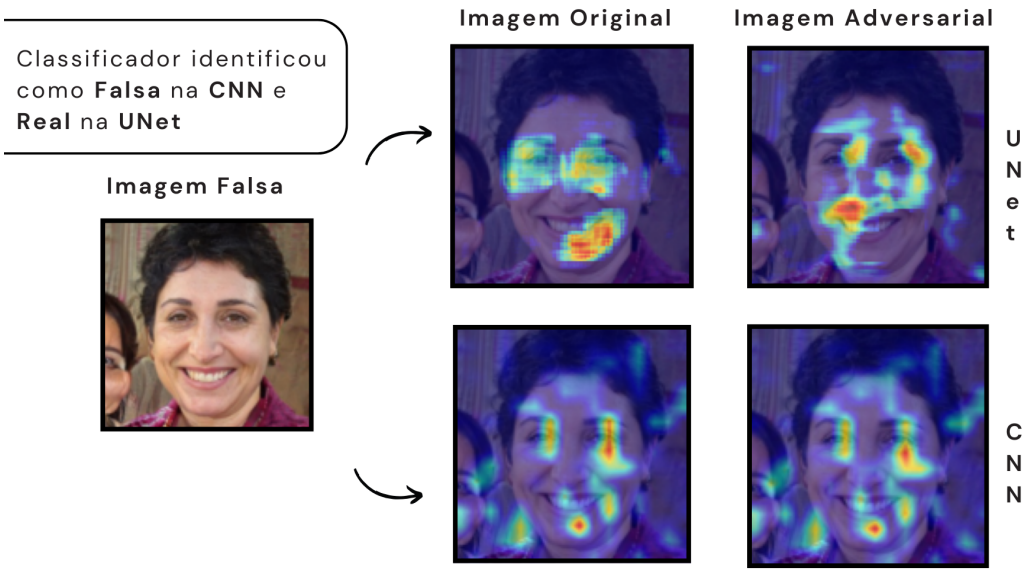

Figura 4. Exemplos de visualização GradCAM para UNET e CNN com comparativo de imagem original e imagem adversarial.

Nota-se que, para a rede neural da UNET, após a imagem ter sido alterada pela rede adversarial, houve uma grande alteração nas áreas de interesse da rede. Isso indica que a rede adversarial possui uma maior facilidade em alterar os resultados esperados pelo classificador. Já no caso da CNN, as regiões de interesse apenas foram realçadas, o que pode indicar uma maior robustez contra ataques, obrigando a rede adversarial a gerar perturbações mais refinadas e focalizadas, de modo a comprometer mesmo as regiões reforçadas pela CNN, o que requer ajustes mais precisos nos parâmetros de otimização para induzir mudanças significativas na decisão final do classificador.

Por fim, a Figura 5 ilustra as diferenças entre a imagem adversarial e a imagem original. Nota-se que ambas as imagens apresentam ruídos imperceptíveis a olho nu, atingindo o objetivo de gerar uma imagem cuja alteração não seja detectada pelo observador. Ademais, observa-se que cada gerador aprende um padrão distinto, o que pode explicar as discrepâncias entre os resultados da avaliação cruzada (Tabela 3) e os obtidos na avaliação direta sob ataque adversarial (Tabela 2).

5. Conclusão

Este trabalho teve como objetivo investigar a vulnerabilidade e a robustez de diferentes classificadores de imagens faciais diante de ataques adversariais imperceptíveis. A aplicação do ataque adversarial demonstrou que perturbações sutis podem degradar significativamente o desempenho dos classificadores, sobretudo das redes neurais, como indicado pelas expressivas reduções de acurácia e pelos baixos índices de recall. Além disso, os mapas de calor gerados via GradCAM evidenciaram que as áreas de interesse dos modelos, tais como olhos e lábios, foram alteradas de maneira distinta, dependendo da arquitetura. Em particular, a CNN mostrou uma robustez relativa maior, pois suas regiões de atenção foram apenas realçadas, enquanto a versão reduzida da UNET evidenciou uma maior facilidade para alterar os resultados esperados, refletindo vulnerabilidades

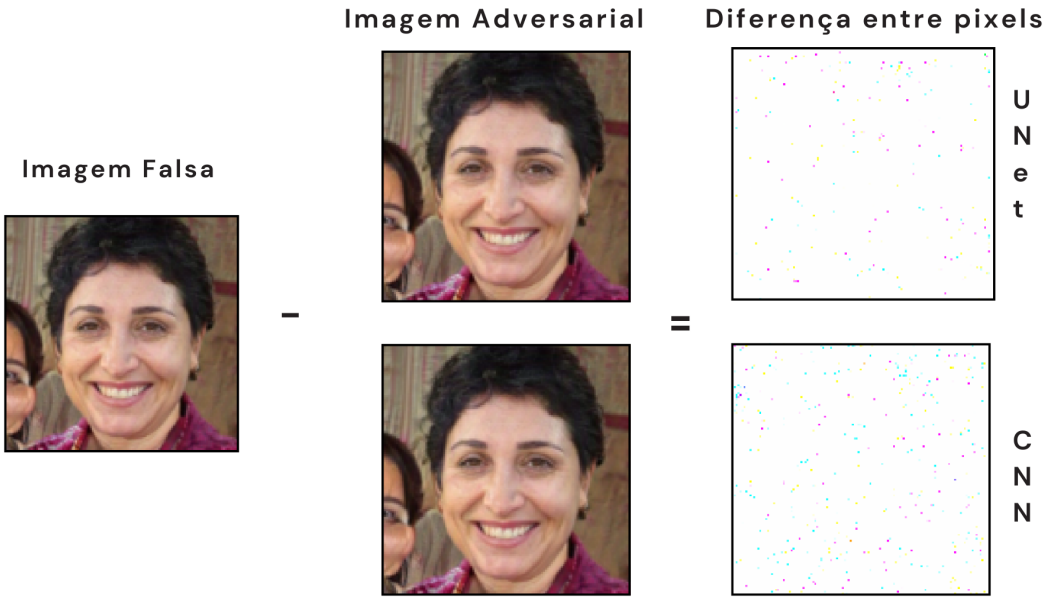

Figura 5. Exemplos de imagens adversariais geradas pela CNN e UNET e a diferença no que foi alterado com base na imagem original

mais acentuadas.

Além disso, a avaliação cruzada dos ataques adversariais revelou que as perturbações possuem propriedades transferíveis entre os modelos, embora a eficácia do ataque varie conforme a arquitetura-alvo. Esses resultados enfatizam não apenas as limitações dos métodos de classificação atuais, mas também a necessidade de desenvolver estratégias defensivas mais robustas que possam mitigar os impactos desses ataques em aplicações críticas, como o reconhecimento facial.

Por fim, este estudo contribuiu para uma compreensão mais aprofundada das vulnerabilidades presentes em sistemas baseados em aprendizado profundo, evidenciando que mesmo perturbações imperceptíveis podem comprometer a integridade dos resultados. Destaca-se a importância de se adotar mecanismos defensivos e de se explorar novas técnicas de treinamento que considerem a adversarialidade como parte integrante do processo de aprendizado, visando aumentar a resiliência dos modelos.

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*.

Ronneberger, O., Fischer, P., e Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., e Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In

Proceedings of the IEEE International Conference on Computer Vision (ICCV), pages 618–626.

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.

Tramèr, F., Carlini, N., Brendel, W., Kurakin, A., Papernot, N., Tsipras, D., e Hendrycks, D. (2021). Transferable adversarial attacks and defenses in the real world.