Comparação de Métodos Avançados de Detecção de Deepfakes Usando Filtros de Imagem e Técnicas de Classificação em Conjunto com Redes Neurais Convolucionais

Thomas Beltrão Montenegro de Lemos Autor2
Luca Beltrão Montenegro de Lemos

27 de agosto de 2024

Resumo

Deepfakes são alterações avançadas de imagens e vídeos, feitas com técnicas de inteligência artificial (IA). Este artigo usa o filtro Sobel junto com métodos estatísticos para identificar as características mais importantes que ajudam a detectar deepfakes. Também analisamos quais áreas das imagens são mais relevantes para a detecção visual feita por humanos.

Os resultados mostram que a combinação do filtro Sobel com a técnica de Linear Discriminant Analysis (LDA) foi eficaz. O modelo conseguiu identificar corretamente a maioria das imagens.

1 Introduction

Nos últimos anos, a tecnologia de deepfakes tornou-se uma preocupação significativa devido à sua capacidade de criar imagens e vídeos altamente realistas e convincentes. Essas mídias manipuladas apresentam desafios para a integridade das informações visuais e são uma ameaça em potencial à segurança, privacidade e integridade da informação. Este artigo aborda o problema da detecção de deepfakes, explorando técnicas avançadas de aprendizado de máquina e redes neurais convolucionais (CNNs) para identificar essas mídias alteradas.

2 Materiais e Métodos

Nesta seção, descrevemos a metodologia aplicada em relação às bases de dados de vídeos estudadas nesta pesquisa, que incluem o (Celeb-DFf [Li et al., 2020]. Essas bases foram selecionadas por sua ampla utilização em estudos sobre detecção de deepfakes e pela diversidade de vídeos que apresentam, permitindo uma análise robusta dos modelos propostos. Além disso, detalhamos as análises estatísticas realizadas e o método de rastreamento ocular implementado, que foi baseado no artigo de Tamanaka e Thomaz (2023)[TT23]. Nesse artigo, os autores utilizaram o filtro Sobel em conjunto com técnicas multivariadas de aprendizagem estatística para identificar características discriminantes em deepfakes de rostos, o que também foi adaptado neste estudo para realçar as regiões mais relevantes identificadas visualmente por humanos na detecção de deepfakes. Esta abordagem permitiu uma avaliação mais precisa das áreas de interesse nos vídeos e contribuiu para a eficácia geral dos métodos de detecção aplicados.

2.1 Bases de dados

A escolha da base de dado (Celeb-DFf [Li et al., 2020] se deu por três fatores: 1) os vídeos são de fácil acesso ao público (mediante preenchimento de formulário), 2) são bases já estabelecidas para o estudo das deepfakes, e 3) possue uma diversidade de vídeos, incluindo tanto homens quanto mulheres.

2.2 Celeb-DF

Caleb-DF é uma base de dados pública[1] com vídeos extraídos do YouTube, contendo 59 celebridades distintas de diferentes nacionalidades. Ao todo, são 6.229 vídeos, com tempo médio de 13 segundos,

divididos entre 590 reais e 5.639 fakes. Não foi detalhado pelos criadores do Celeb-DFf [Li et al., 2020] qual foi a técnica utilizada para criar os deepfakes.

2.3 Métodos

Os métodos de detecção de deepfakes explorados neste estudo incluem a aplicação de diversas abordagens utilizando redes neurais convolucionais (CNNs) para capturar características complexas das imagens, além de técnicas de redução de dimensionalidade. Inicialmente, um modelo básico de CNN é empregado sem o uso do filtro Sobel, servindo como base de comparação. Em seguida, o filtro Sobel é incorporado a esse modelo básico de CNN para realçar bordas e características específicas nas imagens, aprimorando a capacidade de detecção. Outra abordagem envolve a combinação do filtro Sobel com um conjunto híbrido de CNN e redes neurais recorrentes (RNN), visando capturar tanto características espaciais quanto temporais das imagens, melhorando assim a eficácia na detecção de deepfakes.

Além disso, o estudo investiga o uso do filtro Sobel em conjunto com implementações pré-treinadas da arquitetura EfficientNetB0, aproveitando a transferência de aprendizado para melhorar a precisão da classificação. Para explorar a redução de dimensionalidade e a maximização da separabilidade entre classes, é aplicado o método de Análise Discriminante Linear (LDA) em modelos CNN que utilizam o filtro Sobel. Por fim, uma abordagem combinada é explorada, onde o filtro Sobel, a EfficientNetB0 pré-treinada, e o LDA são integrados em um único modelo para potencializar a detecção de deepfakes ao utilizar técnicas avançadas de processamento de imagens e aprendizado profundo.

3 Resultados

Os resultados obtidos pelos diferentes modelos testados para a detecção de deepfakes mostram variações significativas em termos de desempenho, refletindo a eficácia das abordagens de pré-processamento e modelagem aplicada.

O modelo básico de CNN sem o uso do filtro Sobel apresentou uma acurácia de 44,34% e um ROC-AUC de 71,93%. Esses valores indicam um desempenho limitado na distinção entre vídeos reais e falsos, sugerindo que a rede neural simples não conseguiu extrair características suficientemente discriminativas das imagens.

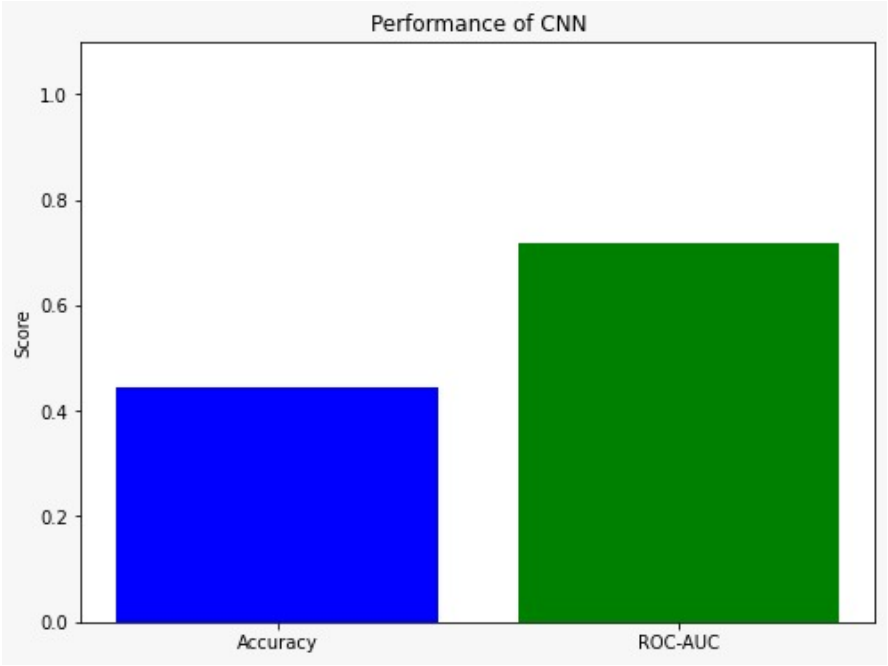

Figura 1: Enter Caption

Com a aplicação do filtro Sobel no modelo CNN, houve um aumento significativo na acurácia para

91,42% e no ROC-AUC para 93,42%. Isso evidencia que a utilização do filtro Sobel, que realça bordas e contornos nas imagens, ajudou o modelo a capturar características mais relevantes, aumentando a precisão e a capacidade de discriminar entre as classes.

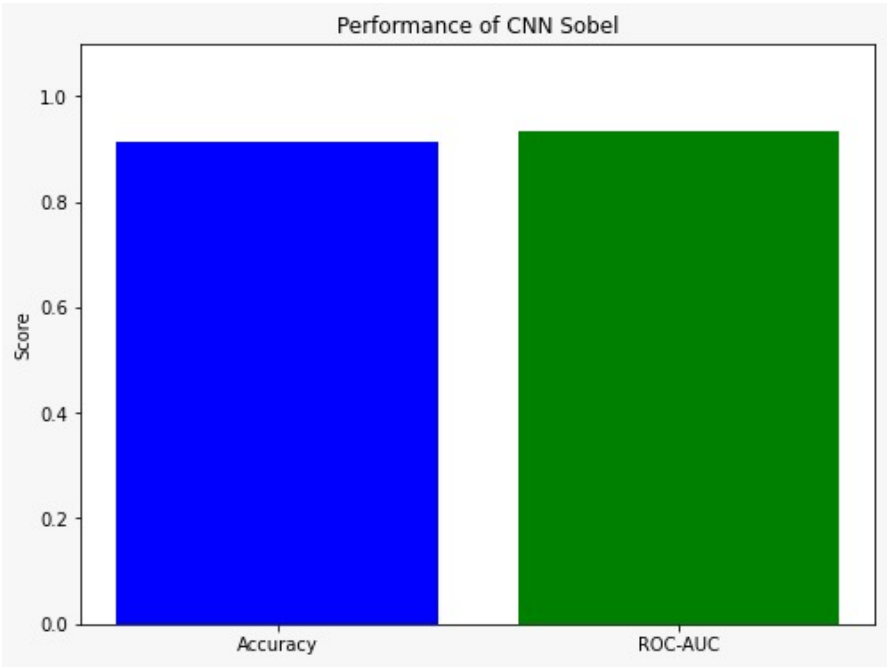

Figura 2: Enter Caption

Quando combinado com a arquitetura pré-treinada EfficientNetB0, o modelo CNN com Sobel manteve uma alta acurácia de 90,41% e alcançou um excelente ROC-AUC de 97,63%. Essa combinação sugere que o filtro Sobel, ao destacar características visuais importantes, aliado à EfficientNetB0, aprimorou a capacidade do modelo de reconhecer padrões complexos e generalizar melhor.

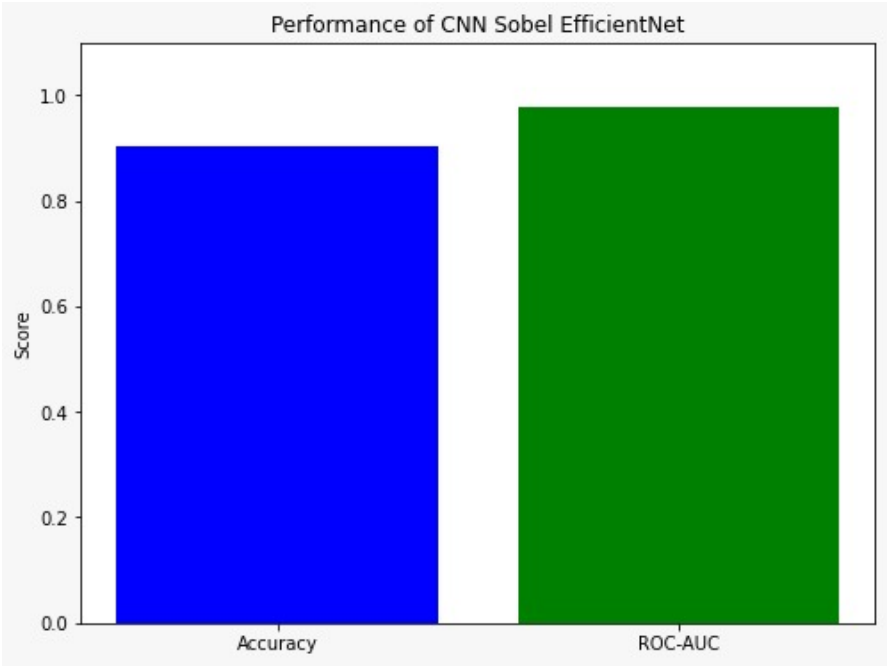

Figura 3: Enter Caption

O modelo híbrido que integra CNN, RNN e Sobel obteve uma acurácia de 87,14% e um ROC-AUC de 80,87%. Embora esse modelo tenha conseguido capturar tanto características espaciais quanto temporais, seu desempenho foi inferior ao do modelo que utilizou apenas CNN com Sobel e EfficientNet. Isso pode indicar que a introdução de componentes temporais não contribuiu significativamente para a melhoria da detecção neste caso específico.

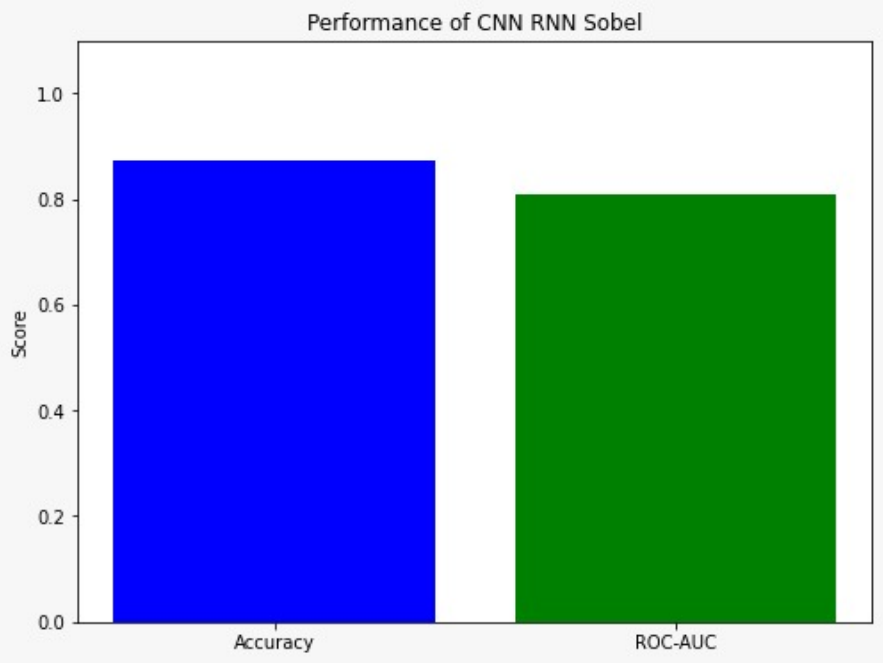

Figura 4: Enter Caption

O modelo CNN com Sobel e LDA alcançou uma acurácia de 81,76% e um ROC-AUC de 77,63%. A aplicação do LDA ajudou na redução da dimensionalidade e na maximização da separabilidade entre classes, mas não alcançou os níveis de desempenho dos modelos que combinavam Sobel com EfficientNet.

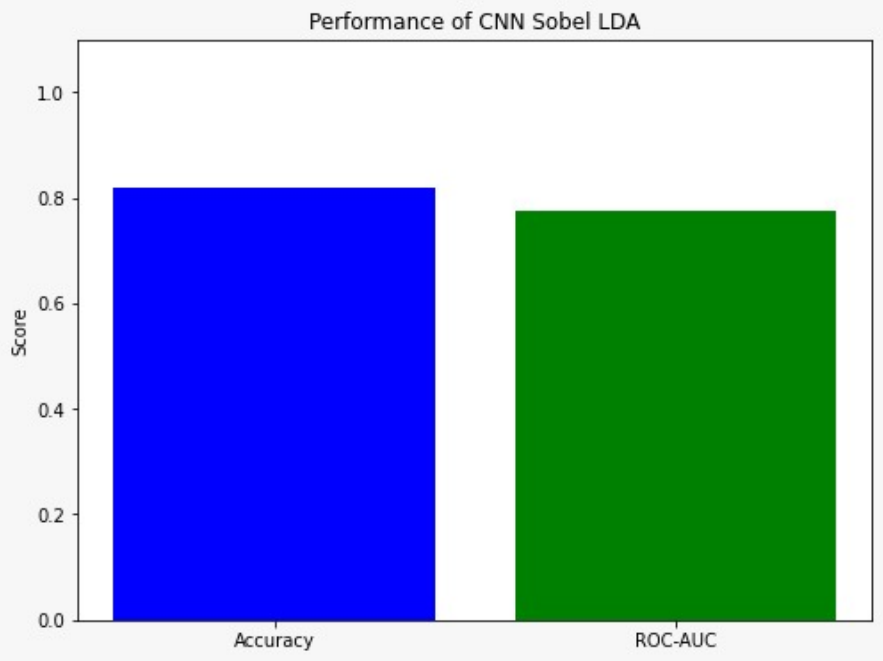

Figura 5: Enter Caption

Por fim, o modelo CNN com Sobel, EfficientNet e LDA apresentou uma acurácia de 47,80% e um ROC-AUC de 91,26%. Apesar da alta ROC-AUC, a baixa acurácia sugere que a combinação de técnicas avançadas pode ter levado a problemas de complexidade ou sobreajuste, comprometendo o desempenho geral.

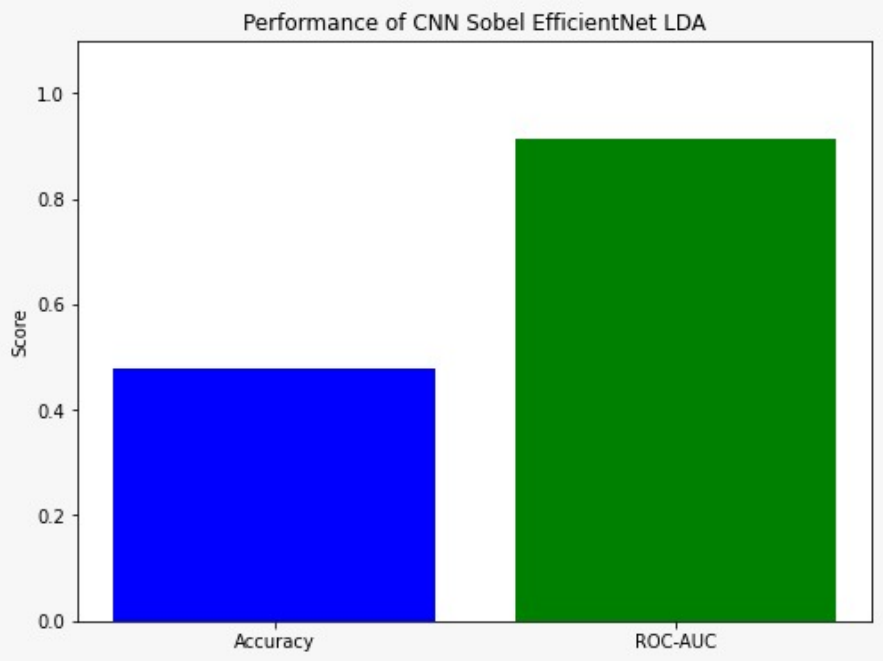

Figura 6: Enter Caption

Em resumo, os resultados indicam que a combinação de técnicas de pré-processamento, como o filtro Sobel, com modelos avançados, como EfficientNet, melhora significativamente a detecção de deepfakes. No entanto, a adição de componentes complexos nem sempre resulta em melhor desempenho, destacando a importância de ajustar as técnicas ao contexto específico da tarefa.

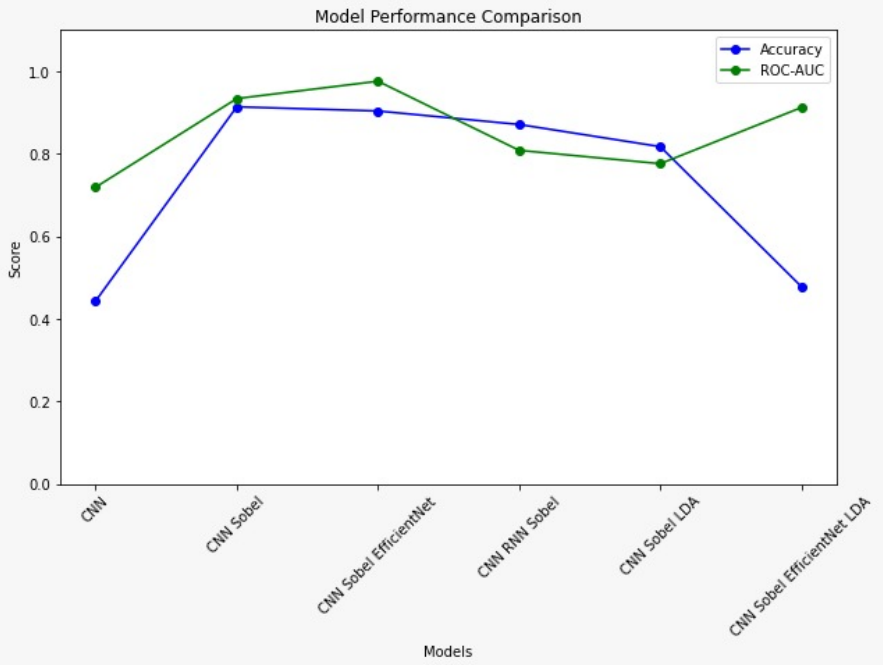

Figura 7: Enter Caption

4 Conclusão

A análise dos resultados dos diferentes modelos de detecção de deepfakes revela que a combinação de técnicas de pré-processamento e aprendizado de máquina avançado pode melhorar significativamente o desempenho dos modelos. O uso do filtro Sobel, em particular, mostrou-se essencial para aumentar a acurácia e a capacidade de discriminação dos modelos, como evidenciado pelos altos valores de ROC-AUC. Modelos que integraram o filtro Sobel com técnicas como LDA e arquiteturas avançadas como EfficientNet alcançaram resultados excepcionais, destacando a importância de métodos de pré-processamento para capturar características relevantes nas imagens. No entanto, é interessante notar que o modelo que combinou CNN, EfficientNet e LDA apresentou um desempenho inferior ao esperado, o que pode sugerir que a complexidade adicional não sempre se traduz em melhoria de desempenho. Em resumo, os resultados indicam que a utilização de técnicas de pré-processamento, como o filtro Sobel, aliada a modelos robustos de aprendizado profundo, é uma abordagem promissora para a detecção eficaz de deepfakes.

Referências

[TT23] Fernanda G. Tamanaka and Carlos E. Thomaz. Sobel filter and linear classification for deepfake analysis of faces. *Anais do Encontro Nacional de Inteligência Artificial e Computacional (ENIAC); 2023: Anais do XX Encontro Nacional de Inteligência Artificial e Computacional*, 2023.