Além do Desempenho: Um Estudo da Confiabilidade de Detectores de Deepfakes

Lucas Lopes, Rayson Laroca, André Gregio

Universidade Federal do Paraná
Pontifícia Universidade Católica do Paraná

Motivação

CNN World Africa Americas Asia Australia China Europe India Middle East More

World / Asia

Finance worker pays out $25 million after video call with deepfake 'chief financial officer'

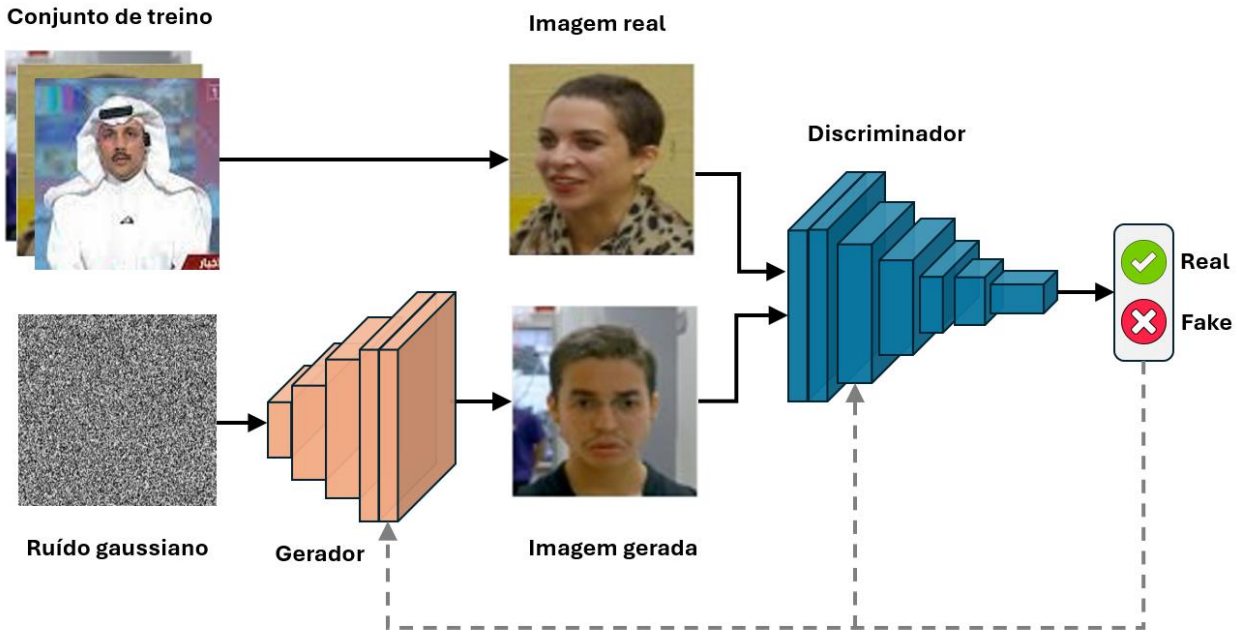
By Heather Chen and Kathleen Magramo, CNN

2 min read · Published 2:31 AM EST, Sun February 4, 2024

SBSeg25 FOZ DO IGUAÇU

Fundamentação

Geração de Deepfakes



Fundamentação

Métodos de Detecção

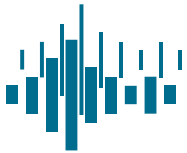
1

Inconsistências Físicas

2

Aprendizado por Dados

3

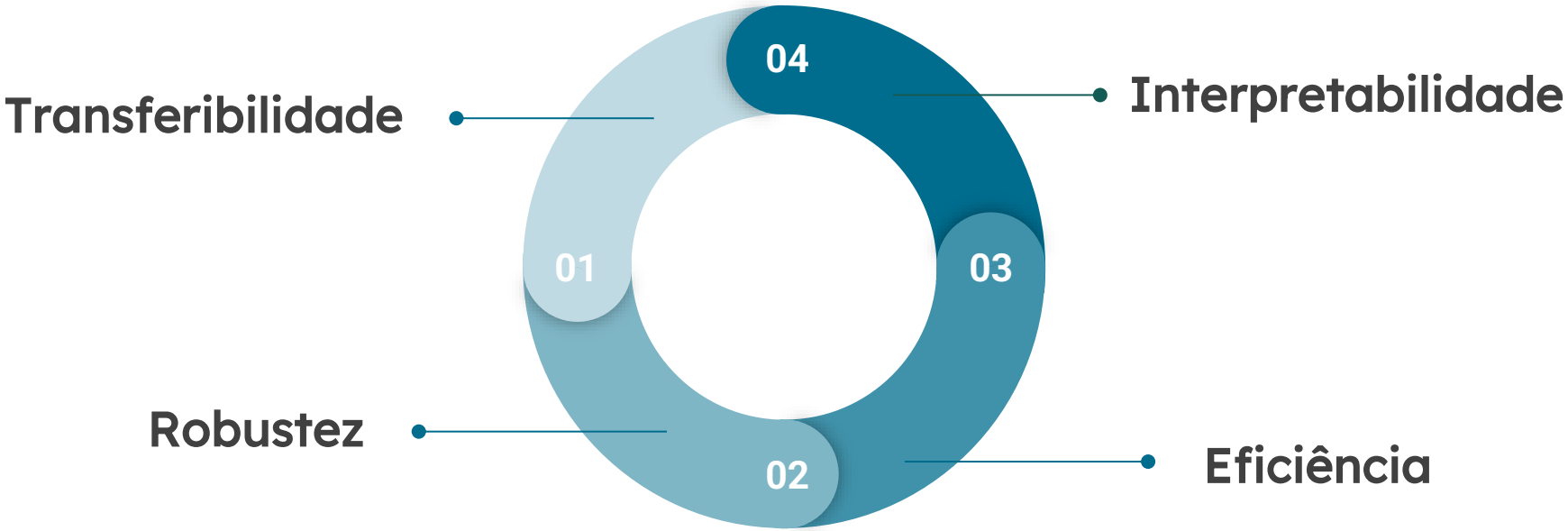


Detecção de Artefatos

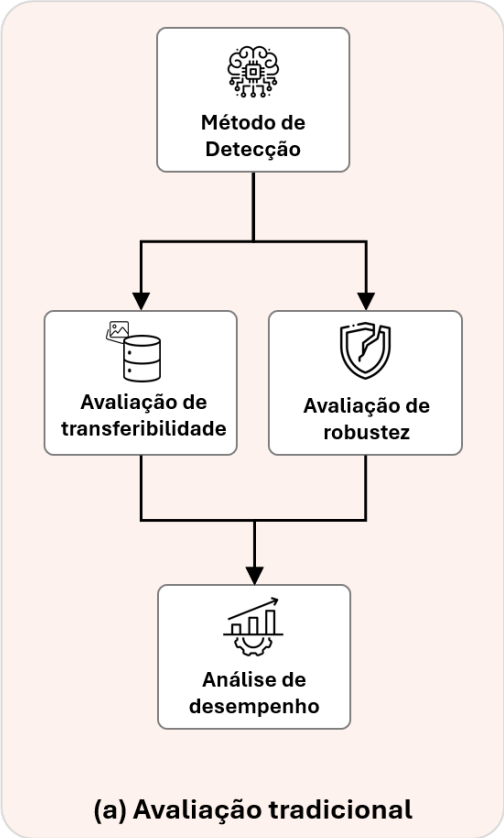
4

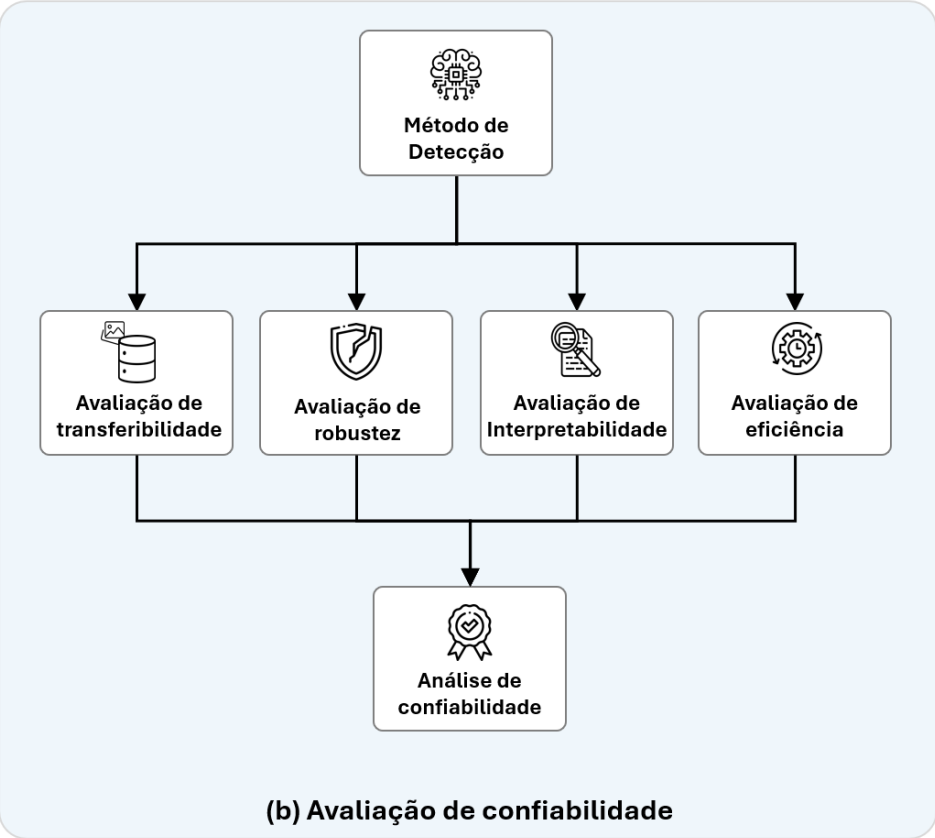
Métodos Híbridos

Desafios


Transferibilidade
Interpretabilidade
Robustez
Eficiência
01
02
03
04

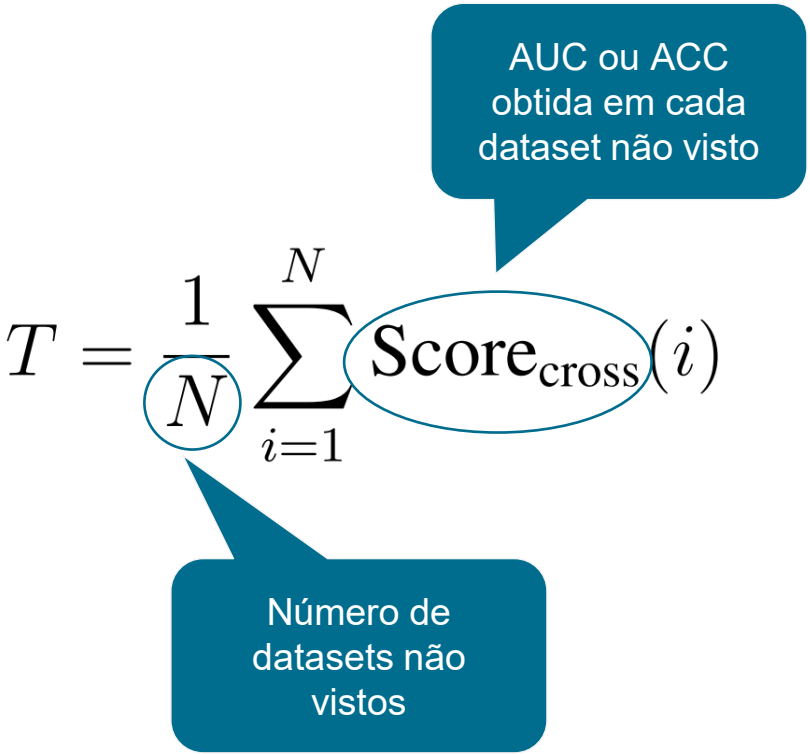
Solução Proposta





Solução Proposta

Transferibilidade



Solução Proposta

Robustez

$$R = \frac{1}{3}\left(\frac{1}{C}\sum_{i=1}^{C}\text{Score}_{\text{comp}}(i) + \frac{1}{P}\sum_{j=1}^{P}\text{Score}_{\text{perturb}}(j) + \frac{1}{A}\sum_{k=1}^{A}\text{Score}_{\text{adv}}(k)\right)$$

Número de testes de perturbação

Número de testes de compressão

Número de testes com ataques adversários

Solução Proposta

Interpretabilidade

Critério	Valor	Descrição
Nenhuma explicação	0,0	Modelo “caixa-preta”, sem qualquer visualização ou justificativa
Visualizações básicas	0,3 – 0,5	Uso de técnicas como Grad-CAM (mapas de saliência) ou t-SNE, sem análise crítica
Análises interpretativas	0,6 – 0,8	Aplicação de métodos como LIME ou SHAP, com explicações baseadas em atributos.
Explicabilidade integrada	0,9 – 1,0	Mecanismos explicativos integrados ao modelo

Solução Proposta

Eficiência Computacional

Número de parâmetros utilizados na etapa de inferência

$$E = \begin{cases} 1.0, & \text{se } P < 10^7 \\ 0.8, & \text{se } 10^7 \leq P < 5 \times 10^7 \\ 0.6, & \text{se } 5 \times 10^7 \leq P < 10^8 \\ 0.4, & \text{se } 10^8 \leq P < 3 \times 10^8 \\ 0.2, & \text{se } 3 \times 10^8 \leq P < 10^9 \\ 0.0, & \text{se } P \geq 10^9 \end{cases}$$

Resultados

Aplicação do framework no estado da arte:

SCLoRA

OSDFD e CFM

FrePGAN

SAMSUNG SDS

TruthLens

Google DeepMind

Resultados

Robustez dos métodos

Método	Métrica	Score (comp)	Score (perturb)	Score (adv)	Robustez (R)
SCLoRA	AUC	0,70	0,00	0,00	0,23
OSDFD	AUC	0,79	0,87	0,00	0,55
CFM	AUC	0,93	0,80	0,00	0,58
FrePGAN	Acurácia	**0,99**	**0,97**	0,00	**0,65**
TruthLens	Acurácia	0,94	0,00	0,00	0,31

Resultados: Avaliação

Avaliação de Confiabilidade

T = Transferibilidade
R = Robustez
I = Interpretabilidade
E = Eficiência Comp.

Método	Métrica	T	R	I	E	SCG
OSDFD	AUC	0,82	0,55	0,62	0,62	0,62
SCLoRA	AUC	0,72	0,23	0,20	0,60	0,44
CFM	AUC	0,84	0,58	0,50	**0,80**	**0,68**
FrePGAN	Acurácia	0,76	**0,65**	0,30	0,58	0,57
TruthLens	Acurácia	**0,94**	0,31	**1,00**	0,00	0,56

$$SCG = \frac{1}{4}(T + R + \mathrm{I} + E)$$

Considerações finais

- Transferibilidade ainda limitada entre domínios distintos
- Robustez pouco explorada (para ataques adversários)
- Interpretabilidade pouco acessível
- Trade-off entre desempenho e custo computacional

Trabalhos futuros

- Expandir análise para mais detectores
- Focar em interpretabilidade e eficiência
- Desenvolvimento de um detector orientado ás métricas de confiabilidade

- Lucas Lopes, Rayson Laroca, André Grégio

- lopes.lucas@ufpr.br