Um Sistema Híbrido para a Detecção de Vídeos Gerados por Inteligência Artificial

Davi Nascimento Mattos
linkedin.com/in/davi-mattos-5a39b3288
github.com/DaviMattosDev

28 de maio de 2025

Resumo

Este trabalho apresenta um sistema híbrido que combina análise visual heurística, aprendizado de máquina e leitura de metadados para detectar vídeos com alta probabilidade de terem sido gerados por inteligência artificial. O modelo utiliza um conjunto de técnicas que incluem detecção de marcos faciais, análise de nitidez (variância Laplaciana), detecção de instabilidade temporal (*jitter*), análise de frequência espacial (FFT) e a aplicação de um modelo XceptionNet pré-treinado. O sistema foi aplicado a vídeos reais e sintéticos, incluindo exemplos do modelo Veo da Google, obtendo resultados promissores mesmo sem acesso a grandes bases de dados para treinamento.

Palavras-chave: IA Generativa, Detecção de Deepfake, Visão Computacional, Sistema Híbrido, Mídia Sintética.

1 Introdução

O avanço exponencial de modelos generativos de vídeo, como o Sora da OpenAI e o Veo da Google, inaugurou uma nova era na criação de conteúdo digital. A capacidade de gerar cenas fotorrealistas a partir de descrições textuais representa um marco tecnológico. Contudo, essa mesma tecnologia introduz desafios críticos relacionados à desinformação, autenticidade e segurança. A distinção entre conteúdo gravado no mundo real e conteúdo gerado sinteticamente tornou-se uma tarefa complexa e de suma importância. Este trabalho propõe uma abordagem técnica e científica para essa detecção, desenvolvendo um sistema híbrido que não depende exclusivamente de um único método, mas sim da confluência de múltiplos indicadores.

2 Metodologia

O sistema proposto adota uma arquitetura modular para avaliar diferentes aspectos de um vídeo. A hipótese central é que, apesar do alto realismo, os vídeos gerados por IA ainda contêm artefatos sutis e inconsistências que podem ser detectados por meio de uma análise multifacetada.

2.1 Fluxo Geral do Sistema

O processo de análise é executado na seguinte ordem:

Extração de Frames e Metadados: O vídeo é decomposto em frames individuais e seus metadados são extraídos para análise inicial. **Análise Facial e Microexpressões:** Detecção de rostos e análise da frequência de piscadas, um indicador biológico frequentemente inconsistente em vídeos sintéticos. **Estimativa de Nitidez (Laplaciano):** Mede a qualidade das bordas. Vídeos de IA tendem a apresentar uma suavidade artificial, resultando em menor variância Laplaciana. **Detecção de Instabilidade Temporal (*Jitter*):** Análise da diferença absoluta entre frames consecutivos para identificar tremores ou transições não naturais. **Análise de Frequência Espacial (FFT):** Uso da Transformada Rápida de Fourier para examinar o espectro de frequência da imagem, onde padrões artificiais podem se manifestar. **Classificação com XceptionNet:** Utilização de um modelo de aprendizado profundo, pré-treinado no dataset FaceForensics++, para classificar a autenticidade de cada frame. **Pontuação Final e Decisão:** Consolidação dos resultados de cada módulo em um sistema de pontuação que classifica o vídeo como "Real"ou "Provavelmente Gerado por IA".

3 Fórmulas e Modelos Matemáticos

Cada módulo do sistema baseia-se em fundamentos matemáticos específicos para extrair características relevantes.

3.1 Eye Aspect Ratio (EAR) para Detecção de Piscadas

A frequência de piscadas é um forte indicador de autenticidade. Utilizamos o EAR, uma métrica simples e eficaz baseada em marcos faciais (facial landmarks).

$$\text{EAR} = \frac{\|p_2 - p_6\| + \|p_3 - p_5\|}{2\|p_1 - p_4\|}$$

Onde $p_1, \ldots, p_6$ são os pontos de referência 2D do olho. Uma queda súbita no valor do EAR (tipicamente abaixo de 0.2) indica uma piscada. A ausência ou frequência anormal de piscadas ao longo do vídeo é um forte indício de origem sintética.

3.2 Análise de Nitidez com Operador Laplaciano

A nitidez é quantificada pela variância do operador Laplaciano aplicado a uma imagem em escala de cinza. O operador de Laplace, ∇^2, realça regiões de rápida mudança de intensidade, como as bordas.

$$\text{fm} = \text{Var}(\nabla^2 I)$$

Onde I é a imagem em escala de cinza e fm (*focus measure*) é a métrica de nitidez. Imagens sintéticas ou desfocadas apresentam um valor de fm significativamente menor.

3.3 Análise de Frequência Espacial (FFT)

A Transformada Rápida de Fourier (FFT) decompõe a imagem do domínio espacial para o domínio da frequência. Vídeos gerados por IA podem exibir padrões de alta frequência anômalos ou uma supressão de certas frequências. Analisamos a magnitude média do espectro de Fourier:

$$\text{Magnitude Média} = \frac{1}{N}\sum_{i=1}^{N} 20\log(|\mathcal{F}(I_i)|)$$

Onde $\mathcal{F}$ é a FFT e I_i é o i-ésimo frame. Valores atípicos podem indicar a presença de artefatos de compressão ou padrões de *upscaling* comuns em modelos generativos.

3.4 Sistema de Pontuação Final

Para chegar a uma decisão final, foi criado um sistema de pontuação ponderado, onde cada evidência contribui para o score final.

Tabela 1: Tabela de Pontuação para Classificação Final

Critério de Análise	**Condição para Pontuar**	**Pontos**
Anomalias Faciais	> 5 frames com falhas de detecção	+2
Frequência de Piscadas	< 2 piscadas detectadas no vídeo	+1
Nitidez Média (fm)	< 100	+1
Instabilidade (*Jitter*)	> 10 (diferença média)	+1
Probabilidade (XceptionNet)	Média $> 60\%$ (fake)	+2
Metadados Suspeitos	Palavras-chave como 'Google', 'Lavf'	+2

Limiar de Decisão: Se o score total for ≥ 4, o vídeo é classificado como **"Provavelmente Gerado por IA"**.

4 Resultados e Análise Prática

O sistema foi testado em um vídeo de demonstração do modelo Veo da Google. A tabela a seguir resume os resultados obtidos.

Tabela 2: Resultados da Análise do Vídeo do Google Veo

Métrica	**Valor Obtido**	**Pontuação Atribuída**
Anomalias Faciais	3 frames	0
Piscadas Detectadas	0	+1
Nitidez Média (fm)	26.32	+1
Jitter Médio	4.92	0
Média FFT (Magnitude)	147.07	-
Probabilidade IA (XceptionNet)	57%	0
Metadados	'Google Inc.', 'Lavf'	+2
Score Total	-	**4**

Classificação Final: Com um score total de 4, o vídeo foi corretamente classificado como **Provavelmente Gerado por IA**.

5 Discussão

O sistema híbrido demonstrou eficácia na detecção de um vídeo de última geração. A força da abordagem reside na sua resiliência: a falha de um único detector (como o XceptionNet, que ficou abaixo do limiar) pode ser compensada por outras evidências fortes, como a ausência total de piscadas e a presença de metadados incriminatórios.

Limitações: O modelo depende de limiares heurísticos (ex: nitidez < 100), que podem necessitar de ajuste para diferentes tipos de vídeo. Além disso, a análise facial é ineficaz em vídeos que não contêm rostos humanos.

Trabalhos Futuros: Para aprimorar o sistema, sugerimos a incorporação de:

4. **Análise de Sinais Fisiológicos:** Detecção de padrões de respiração através de movimentos sutis do tórax.

- **Reconhecimento Óptico de Caracteres (OCR):** Análise de textos e letreiros no vídeo, que são frequentemente distorcidos por modelos de IA.
- **Arquiteturas baseadas em Transformers:** Utilizar modelos como o Vision Transformer (ViT) para capturar relações temporais complexas entre os frames.

6 Conclusão

Este trabalho demonstrou a viabilidade de um sistema híbrido para a detecção de vídeos sintéticos. Ao integrar análises de baixo nível (pixels e frequência) com aprendizado de máquina e análise de metadados, criamos um framework robusto capaz de identificar artefatos que passariam despercebidos por uma análise superficial. Mesmo diante de vídeos altamente realistas, como os gerados pelo modelo Veo, o sistema conseguiu agregar evidências suficientes para uma classificação correta, reforçando que uma abordagem multifacetada é essencial na corrida contra a desinformação digital.

Referências

Chollet, F. (2017). Xception: Deep Learning with Depthwise Separable Convolutions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. arXiv:1610.02357. Rossler, A., et al. (2019). FaceForensics++: Learning to Detect Manipulated Facial Images. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Google Research. (2020). MediaPipe Face Mesh. `https:`

//google.github.io/mediapipe/solutions/face_mesh.html FFmpeg Developers. (2024). FFmpeg Documentation. https://ffmpeg.org/ffmpeg-all.html