Detecção de Deepfakes

Autor: Matheus Peixoto Ribeiro Vieira
Orientador: Pedro Henrique Lopes Silva

22 de setembro de 2024

1 Contextualização e Definição do Problema de Pesquisa

O termo Deepfake é uma combinação dos termos 'deep learning' e 'fake', e refere-se a uma tecnologia que utiliza redes neurais profundas para gerar arquivos de mídia altamente convincentes e realistas, sendo principalmente utilizado para a manipulação de rostos e vozes, gerando fotos, vídeos e arquivos de áudio (MIRSKY; LEE, 2021). Um exemplo de DeepFake pode ser visto em 1, onde o rosto original foi modificado a fim de compor a base de dados Celeb-DF (LI et al., 2020).

Figura 1: Imagem original à esquerda e DeepFake à direita

Fonte: (LI et al., 2020).

Muitas das diferentes formas e técnicas para a criação de DeepFake são *open-source* e estão ficando cada vez mais fáceis de serem acessadas e utilizadas, sem que o usuário tenha um profundo conhecimento em aprendizado de máquina ou processamento de imagens, gerando, portanto, um crescimento muito grande de mídias modificadas com diferentes qualidades ((GONG; LI, 2024; HEIDARI et al., 2024)).

Entre os anos de 2022 e 2023, a presença de vídeos com a presença de deepfakes na internet triplicou, gerando consequências significativas para diferentes indivíduos (ŁABUZ; NEHRING, 2024). Um exemplo notável ocorreu com Muharrem Ince, candidato a presidência da Turquia, que desistiu da corrida presidencial após falsos vídeos sexuais manipulados com a sua face serem compartilhados na internet (KIRBY, 2023; ŁABUZ; NEHRING, 2024).

Em Hong Kong, uma multinacional do setor financeiro transferiu cerca de 25 milhões de dólares para golpistas depois que um dos funcionários da empresa recebeu um email do chefe financeiro da sede e participou de uma videoconferência com ele e outros colaboradores. No entanto, todos os participantes da reunião pareciam e soavam como as pessoas reais, mas eram deepfakes (CHEN; MAGRAMO, 2024).

Dessa forma, visando o combate à disseminação de vídeos manipulados com a presença de DeepFake, modelos de redes neurais artificiais se mostram eficazes para detectarem a presença dos mesmos. Sendo a acurácia uma das principais métricas utilizadas, (HU et al., 2022) obteve 90,47% de acerto nas identificações ao utilizar uma parcela de 20 quadros do vídeo original, o

equivalente a cerca de um segundo de vídeo. Já em (SINGH et al., 2020), com o uso de 30 quadros, equivalente a um segundo, foi obtida uma acurácia de 97,63%.

Melhores resultados foram alcançados em (HERNANDEZ-ORTEGA et al., 2022), onde foi obtida uma acurácia de 99,24% em vídeos com cinco segundos, obtendo uma média das acurácias de cada quadro, e indo até 99,47% de acurácia com sete segundos, utilizando a mesma métrica.

Dessa forma, é possível identificar uma correlação entre o aumento do número de quadros de um vídeo utilizada no treinamento e validação de um modelo com sua acurácia final obtida. Todavia, essa inclusão de mais imagens aumenta o processamento necessário, gerando um maior custo computacional. A partir disso, surge uma questão relevante: as técnicas atuais de detecção de DeepFake podem ser aprimoradas para identificar com maior precisão manipulações em vídeos mais curtos, onde as alterações faciais tendem a ser menos evidentes?

2 Objetivos

Assim, para este trabalho, tem-se, como objetivo principal, demonstrar a possibilidade da criação de um modelo de redes neurais capaz de detectar DeepFakes em vídeos que possuem uma duração de 24 quadros obtendo uma acurácia superior a 98% com a base de dados Celeb-DF, a mesma utilizada em (HERNANDEZ-ORTEGA et al., 2022), sendo este valor uma acurácia competitiva quando comparada a modelos com maior robustez e complexidade.

Dessa forma, para atingir este objetivo, surgem os seguintes objetivos específicos:

- Identificar fatores que influenciam a acurácia dos modelos;
- Levantar diferenças entre as formas de pré-processamento de vídeos;
- Avaliar o uso de diferentes modelos convolucionais e convolucionais recorrentes para detectar DeepFake;
- Analisar, entre diferentes modelos, características em comum que induzem a erros.

3 Hipótese de Trabalho

Redes neurais convolucionais podem atingir uma acurácia superior a 98% na detecção da base de dados da Celeb-DF, utilizando somente 24 quadros dos vídeos originais.

A abordagem da escolha da quantidade de quadros é justificada por ser um valor intermediário entre trabalhos que utilizam menos quadros, como em (HU et al., 2022), que utiliza 20, (SINGH et al., 2020), que utiliza 30 e para trabalhos que se utilizam de um maior número de dados com uma maior duração de vídeo, como em (HERNANDEZ-ORTEGA et al., 2022).

Embora 24 quadros não seja o meio termo exato entre 20 e 30, é uma taxa de quadros universalmente adotada em diferentes mídias digitais (WILCOX et al., 2015), o que a torna um valor prático e eficaz.

Ademais, como discutido em (HEIDARI et al., 2024), há uma demanda por estudos que combinem uma alta acurácia com uma maior eficiência temporal, podendo ser obtida com o uso de 24 quadros para treinar modelos de redes neurais convolucionais, sendo estes os mais comuns entre os trabalhos supracitados.

Por fim, a acurácia definida se mostra um valor competitivo ao ser comparada com os resultados obtidos por (HERNANDEZ-ORTEGA et al., 2022), que possui um modelo mais

robusto e com mais processamento, enquanto supera resultados de modelos mais rápidos, como o (SINGH et al., 2020).

4 Procedimento Metodológico

Para o desenvolvimento da pesquisa, as seguintes etapas são propostas como procedimento metodológico a fim de alcançar o objetivo geral e os objetivos específicos:

1. Mapeamento sistemático da literatura para identificar as principais arquiteturas de redes neurais para detecção de DeepFake e as principais formas de pré-processamento de dados;
2. Analisar as características comuns e destoantes dos pré-processamentos de dados da literatura e pré-processar os dados;
3. Avaliar o uso de diferentes modelos convolucionais e recorrentes para detectar DeepFake.
4. Propor um modelo de Deep Learning para detecção de DeepFake;
5. Avaliar os resultados do modelo proposto com os modelos propostos na literatura.

4.1 Limitações

Com o rápido crescimento de modelos geradores de DeepFake a partir do uso de diferentes técnicas e tecnologias (GONG; LI, 2024), torna-se impraticável a criação de um modelo capaz de detectar, para todas as bases, a presença deste tipo de manipulação em todos os tipos de vídeos.

Assim, considerando tal dificuldade, o estudo proposto se limitará ao uso da base de dados Celeb-DF (LI et al., 2020) para detecção de DeepFake, sendo esta uma base dados altamente utilizada em diferentes estudos, proporcionando um amplo *benchmark* para com diferentes propostas de modelos. O *dataset* é composto por uma ampla variedade de vídeos, sendo 890 sem manipulações e 5639 vídeos com DeepFake, possuindo imagens de celebridades diversas e em alta qualidade.

Referências

CHEN, Heather; MAGRAMO, Kathleen. **Finance worker pays out $25 million after video call with deepfake chief financial officer**. [S.l.]: CNN, fev. 2024. Disponível em: <https://edition.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/index.html>.

GONG, Liang Yu; LI, Xue Jun. A Contemporary Survey on Deepfake Detection: Datasets, Algorithms, and Challenges. **Electronics**, v. 13, n. 3, 2024. ISSN 2079-9292. DOI: 10.3390/electronics13030585. Disponível em: <https://www.mdpi.com/2079-9292/13/3/585>.

HEIDARI, Arash et al. Deepfake detection using deep learning methods: A systematic and comprehensive review. **WIREs Data Mining and Knowledge Discovery**, v. 14, n. 2, e1520, 2024. DOI: https://doi.org/10.1002/widm.1520. eprint: https://wires.onlinelibrary.wiley.com/doi/pdf/10.1002/widm.1520. Disponível em: <https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.1520>.

HERNANDEZ-ORTEGA, Javier et al. DeepFakes Detection Based on Heart Rate Estimation: Single- and Multi-frame. In: **Handbook of Digital Face Manipulation and Detection: From DeepFakes to Morphing Attacks**. Edição: Christian Rathgeb. Cham: Springer International Publishing, 2022. p. 255–273. DOI: 10.1007/978-3-030-87664-7_12. Disponível em: <https://doi.org/10.1007/978-3-030-87664-7_12>.

HU, Juan et al. FInfer: Frame Inference-Based Deepfake Detection for High-Visual-Quality Videos. **Proceedings of the AAAI Conference on Artificial Intelligence**, v. 36, n. 1, p. 951–959, jun. 2022. DOI: 10.1609/aaai.v36i1.19978. Disponível em: <https://ojs.aaai.org/index.php/AAAI/article/view/19978>.

KIRBY, Paul. **Muharrem Ince: Turkish candidate dramatically pulls out before election**. [S.l.]: BBC News, mai. 2023. Disponível em: <https://www.bbc.com/news/world-europe-65560052>.

ŁABUZ, Mateusz; NEHRING, Christopher. On the way to deep fake democracy? Deep fakes in election campaigns in 2023. **European Political Science**, abr. 2024. ISSN 1682-0983. DOI: 10.1057/s41304-024-00482-9. Disponível em: <https://doi.org/10.1057/s41304-024-00482-9>.

LI, Yuezun et al. **Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics**. [S.l.: s.n.], 2020. arXiv: 1909.12962 [cs.CR]. Disponível em: <https://arxiv.org/abs/1909.12962>.

MIRSKY, Yisroel; LEE, Wenke. The Creation and Detection of Deepfakes: A Survey. **ACM Computing Surveys**, Association for Computing Machinery (ACM), v. 54, n. 1, p. 1–41, jan. 2021. ISSN 1557-7341. DOI: 10.1145/3425780. Disponível em: <http://dx.doi.org/10.1145/3425780>.

SINGH, Amritpal et al. DeepFake Video Detection: A Time-Distributed Approach. **SN Computer Science**, v. 1, n. 4, p. 212, jun. 2020. ISSN 2661-8907. DOI: 10.1007/s42979-020-00225-9. Disponível em: <https://doi.org/10.1007/s42979-020-00225-9>.

WILCOX, Laurie M. et al. Evidence that Viewers Prefer Higher Frame-Rate Film. Association for Computing Machinery, New York, NY, USA, v. 12, n. 4, set. 2015. ISSN 1544-3558. DOI: 10.1145/2810039. Disponível em: <https://doi.org/10.1145/2810039>.