Detecção de Vídeos Gerados por GAN (Generative Adversarial Networks)

Mikhael M. F. Maia[1], João Gabriel C. França[2], Joseppe Pedro C. Fellini[3], Wendel Márcio O. Filho[4]

[1] Instituto de Informática – Universidade Federal de Goiás (UFG)

Goiânia– GO – Brasil

jacson@inf.ufg.br - *professor orientador*

Abstract. *This meta-paper presents a consolidated of studies on the detection of videos generated by "Generative Adversarial Networks" (GAN). The main idea of the article is to show that, over time, the number of fake videos, which are a synthesis of human images or sounds, based on artificial intelligence techniques, called "deep fakes", has grown a lot. Thus, it is shown in this article how to detect these videos generated by GAN.*

Resumo. *Este artigo apresenta um consolidado de estudos sobre a detecção de vídeos gerados por "Generative Adversarial Networks" (GAN). A ideia central do artigo é mostrar que, com o passar do tempo, o número de vídeos falsos, que são uma síntese de imagens ou sons humanos, baseadas em técnicas de inteligência artificial, chamados de "deep fakes", cresceu muito. Assim, mostra-se neste artigo como detectar estes vídeos gerados por GAN.*

1. Informações Gerais

1. Informações Gerais

Nos últimos anos, vídeos falsos criados através do uso de tecnologias, como GAN, foram crescendo. Assim, mesmo que as técnicas de criação e edição de vídeos tenham evoluído, ajudando muitos profissionais e empresas sérias, elas também ajudaram no crescimento da desinformação, alterações em resultados políticos, falsidade ideológica e até crimes sexuais pela internet. Tudo isso gerou grandes questionamentos globais sobre como a internet deveria ser usada..

Permitir mudar a identidade de um orador em um conteúdo nas redes sociais não é mais uma tarefa difícil. Antigamente, para editar um conteúdo com um computador, necessitava-se de softwares robustos e de um grande domínio deles previamente. Contudo, estes sistemas e ferramentas de manipulação facial são avançados ao ponto de usuários, sem grandes habilidades ou experiência na área, consigam, facilmente, alterar a imagem o áudio e imagem de um vídeo de uma autoridade, como um político, comprometendo-o com falas inadequadas e interferindo em resultados de eleições.

O foco principal desses vídeos falsos gerados por GAN é a manipulação facial, sendo o tipo de conteúdo mais visto nos famosos "deep fakes", que vem da junção das palavras "deep learning" (aprendizado profundo) e "fake" (falso). Esta técnica usa inteligência artificial para manipular informações ao criar fotos, vídeos e áudios. A

manipulação facial, muitas vezes, passa despercebida. Por isso, o Tribunal Superior Eleitoral (TSE) aprovou doze propostas de resolução em fevereiro de 2024, não permitindo que campanhas utilizassem “deep fakes”.

Por um lado, este avanço tecnológico coopera para o avanço da nossa sociedade, através do uso dessas tecnologias, ajudando em produções de filmes, artes visuais, produção de músicas, etc. Contudo, por outro lado, também facilita a geração de vídeos falsos por usuários maliciosos. A ideia principal deste estudo é mostrar como identificar vídeos falsos gerados por GAN.

Existem padrões que ajudam a entender que um vídeo foi feito por GAN, como falhas em expressões faciais não-naturais, piscar dos olhos anormais, descompasso no movimento dos lábios, iluminação inconsistente, vídeos com bordas irregulares, mudanças abruptas em pixels, frequências de áudio estranhas e tom robótico na voz. Entretanto, vídeos com compressões excessivas, múltiplas operações de edição ao mesmo tempo, redução de amostragem, entre outros pontos, tornam a identificação de traços de um vídeo falso mais difícil. Por esta razão, as técnicas modernas de manipulação facial são muito difíceis de se detectar. Na verdade, muitas manipulações diferentes de faces existem técnicas, ou seja, não há um modelo único explicando essas falsificações). Além disso, estas edições são feitas, normalmente, em algumas partes do corpo e não no corpo todo, ou seja, a face ou parte dela, e não o quadro completo. Consequentemente, esses tipos de vídeos manipulados são normalmente compartilhados através de plataformas sociais que aplicam redimensionamento, dificultando ainda mais o desempenho dos detectores clássicos.

2. Metodologia

Vídeos gerados por GAN são detectados através de quatro etapas, sendo elas “coleta de dados”, “pré-processamento”, “treinamento do modelo” e “avaliação e validação”.

A coleta de dados foca na utilização de datasets públicos, como o primeiro colocado no "*Kaggle Deepfake Detection Challenge*" e o "*Video Face Manipulation Detection Through Ensemble of CNNs*". Já o balanceamento de classes está vinculado à implementação de “*oversampling*” e “*undersampling*”, garantindo uma representação equilibrada de vídeos reais e sintéticos. Por fim, tem-se uma categorização dos tipos de manipulação.

O pré-processamento é dividido em três partes, “extração de frames”, “detecção e alinhamento facial”, “análise de qualidade” e “extração de características específicas”. A extração de frames foca nos segmentos-chave do vídeo, realizando-a em 30 frames por segundo. A detecção e alinhamento facial utiliza MTCNN, *Multi-task Cascaded Convolutional Networks*, para a localização precisa das faces.

O treinamento do modelo ocorre a partir de redes neurais convolucionais (CNNs), explorando arquiteturas pré-treinadas, como *EfficientNet*. Além disso, necessita-se do uso de “*transfer learning*”, “*data augmentation*”, “treinamento com otimização adaptativa”, diminuição do risco de “*overfitting*” e uma função de perda para medir a diferença entre as previsões do modelo e rótulos reais dos vídeos analisados

3. Trabalhos relacionados

Dois trabalhos foram utilizados como referência principal para a modelagem deste sistema, "*Kaggle Deepfake Detection Challenge*" e o "*Video Face Manipulation Detection Through Ensemble of CNNs*".

Kaggle Deepfake Detection Challenge

Este artigo descreve uma solução baseada na detecção de "*deepfakes*", através de uma análise quadro a quadro. utilizando CNNs e técnicas avançadas de pré-processamento e e treinamento.

Os pontos principais deste artigo são detecção de rosto, pré-processamento de imagens, estratégia de predição, aumentação de dados, preparação e processamento de dados, treinamento e inferência.

Video Face Manipulation Detection Through Ensemble of CNNs

4. Arquitetura C4

Diagrama de contexto

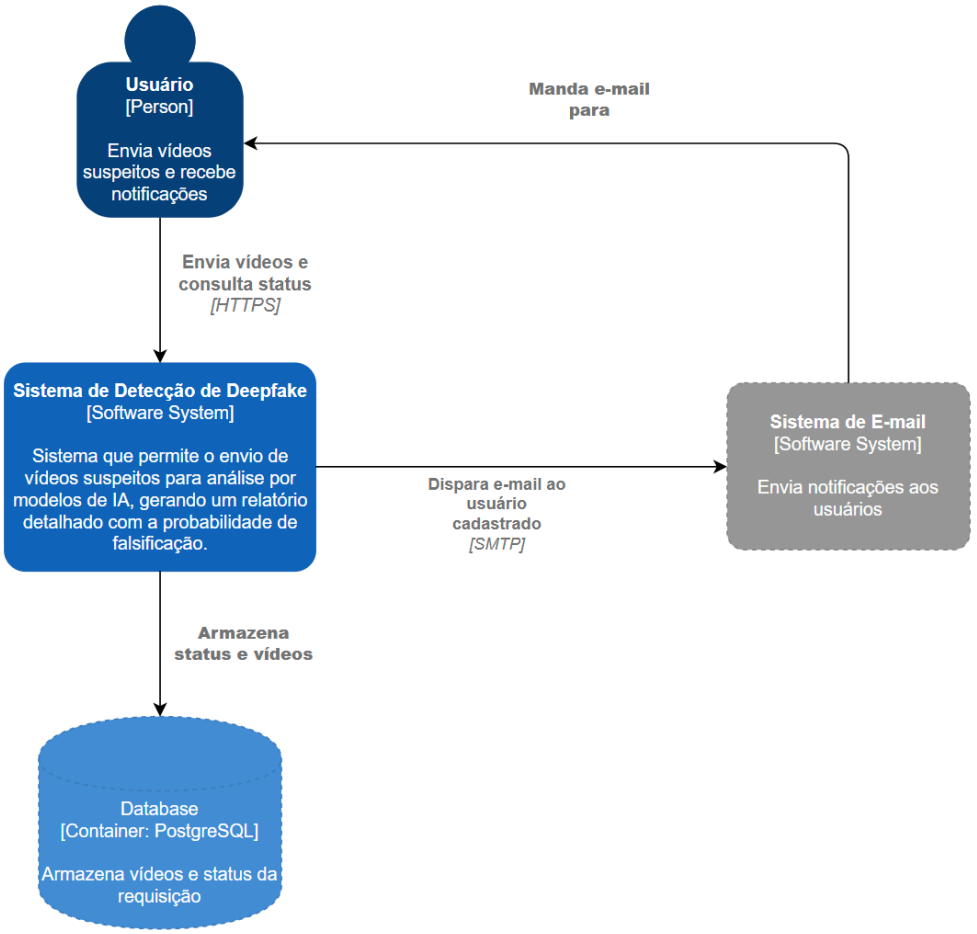


Diagrama de containers

All

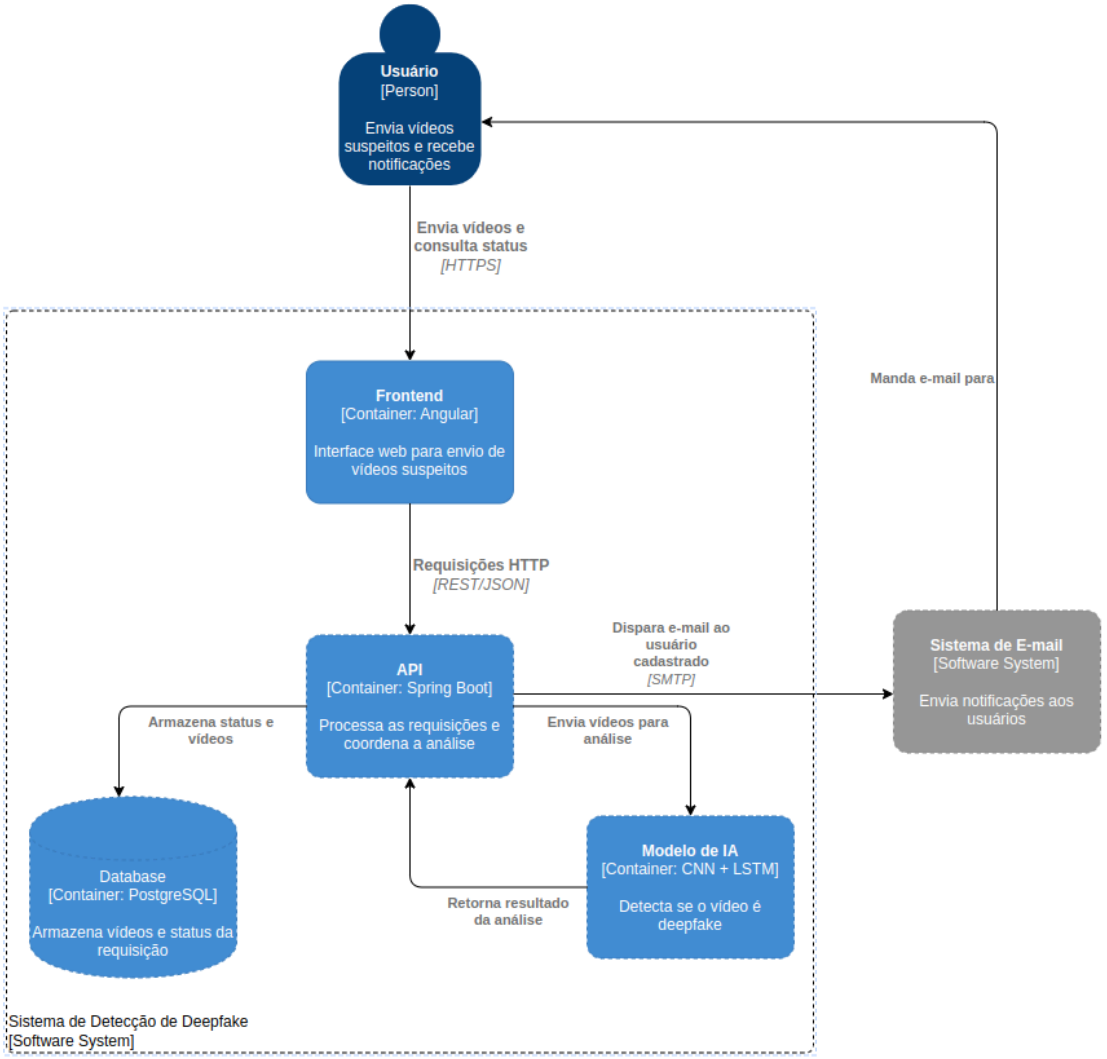
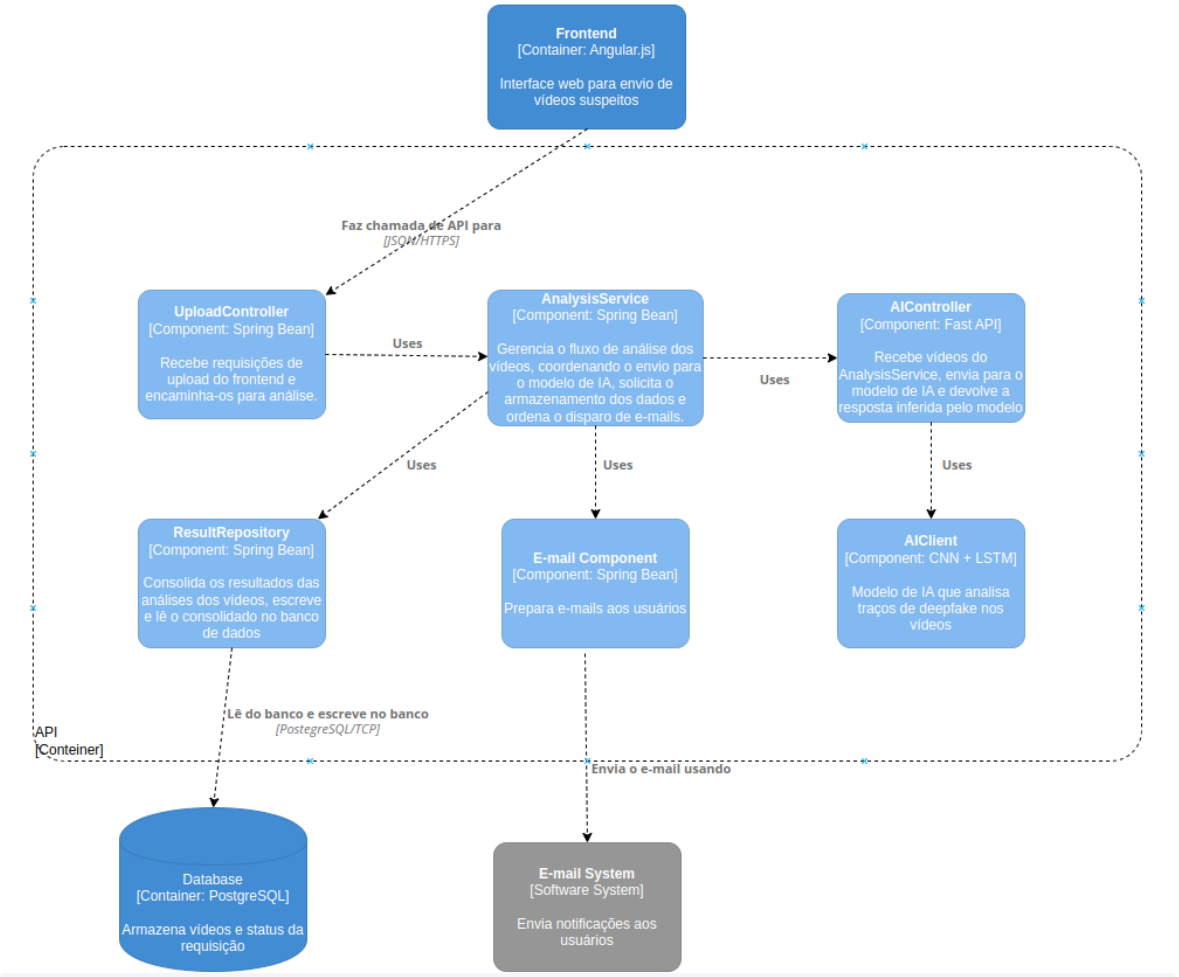


Diagrama de componentes

All


Frontend
[Container: Angular.js]
Interface web para envio de vídeos suspeitos
Faz chamada de API para
[JSON/HTTPS]
UploadController
[Component: Spring Bean]
Recebe requisições de upload do frontend e encaminha-os para análise.
Uses
AnalysisService
[Component: Spring Bean]
Gerencia o fluxo de análise dos vídeos, coordenando o envio para o modelo de IA, solicita o armazenamento dos dados e ordena o disparo de e-mails.
Uses
AIController
[Component: Fast API]
Recebe vídeos do AnalysisService, envia para o modelo de IA e devolve a resposta inferida pelo modelo
Uses
Uses
Uses
ResultRepository
[Component: Spring Bean]
Consolida os resultados das análises dos vídeos, escreve e lê o consolidado no banco de dados
E-mail Component
[Component: Spring Bean]
Prepara e-mails aos usuários
AIClient
[Component: CNN + LSTM]
Modelo de IA que analisa traços de deepfake nos vídeos
Lê do banco e escreve no banco
[PostegreSQL/TCP]
API
[Conteiner]
Envia o e-mail usando
Database
[Container: PostgreSQL]
Armazena vídeos e status da requisição
E-mail System
[Software System]
Envia notificações aos usuários

References

Ball´ e, Johannes, Laparra, Valero, and Simoncelli, Eero P. Density modeling of images using a generalized normalization transformation. CoRR, abs/1511.06281, 2015. URL http://arxiv.org/abs/1511.06281.

Bengio, Yoshua, Mesnil, Grégoire, Dauphin, Yann, and Rifai, Salah. Better mixing via deep representations. CoRR, abs/1207.4404, 2012. URL http://arxiv.org/abs/1207.4404.

Blundell, C., Uria, B., Pritzel, A., Li, Y., Ruderman, A., Leibo, J. Z, Rae, J., Wierstra, D., and Hassabis, D. Model-Free Episodic Control. ArXiv e-prints, June

2016.

Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., and Abbeel, P. InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets. ArXiv e-prints, June 2016.

Denton, Emily L., Chintala, Soumith, Szlam, Arthur, and Fergus, Robert. Deep generative image models using a laplacian pyramid of adversarial networks. CoRR, abs/1506.05751, 2015. URL http://arxiv.org/abs/1506.05751.

Dinh, Laurent, Sohl-Dickstein, Jascha, and Bengio, Samy. Density estimation using real NVP. CoRR, abs/1605.08803, 2016. URL http://arxiv.org/abs/1605.08803.

Donahue, J., Kr¨ ahenb¨ uhl, P., and Darrell, T. Adversarial Feature Learning. ArXiv e-prints, May 2016.

Dumoulin, V., Belghazi, I., Poole, B., Lamb, A., Arjovsky, M., Mastropietro, O., and Courville, A. Adversarially Learned Inference. ArXiv e-prints, June 2016.

Sung, W., Shin, S., Hwang, K.: Resiliency of deep neural networks under quantization. arXiv preprint arXiv:1511.06488 (2015)

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. (2016)

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. (2016)

Yang, J., Shen, X., Xing, J., Tian, X., Li, H., Deng, B., Huang, J., Hua, X.s.:Quantization networks. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. (2019)

Yoo, J., Cho, M., Kim, T., Kang, U.: Knowledge extraction with no observable data. In: Proc. Adv. Neural Inf. Process. Syst. pp. 2701–2710 (2019)

Zeng, R., Huang, W., Tan, M., Rong, Y., Zhao, P., Huang, J., Gan, C.: Graph convolutional networks for temporal action localization. In: Proc. IEEE Int. Conf. Comp. Vis. (2019)

Zhou, S., Wu, Y., Ni, Z., Zhou, X., Wen, H., Zou, Y.: Dorefa-net: Training low bitwidth convolutional neural networks with low bitwidth gradients. arXiv preprintarXiv:1606.06160 (2016)

Zhuang, B., Liu, L., Tan, M., Shen, C., Reid, I.: Training quantized neural networks with a full-precision auxiliary module. In: Proc. IEEE Conf. Comp. Vis. Patt.Recogn. (2020)

Zhuang, B., Shen, C., Tan, M., Liu, L., Reid, I.: Towards effective low-bitwidth convolutional neural networks. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. (2018)