Reconhecimento facial em risco: Vulnerabilidades e defesas contra ataques adversariais.

Gabriela Cristina Moreira dos Santos
Instituto de Ciência e Tecnologia
Universidade Federal de São Paulo
São José dos Campos - Brasil
Email: gabriela.moreira@unifesp.br

Lucas Guilherme Silva de Carvalho
Instituto de Ciência e Tecnologia
Universidade Federal de São Paulo
São José dos Campos - Brasil
Email: lgscarvalho@unifesp.br

***Resumo*—Este estudo se concentra na compreensão dos ataques adversariais em sistemas de reconhecimento facial baseados em Inteligência Artificial (IA) e na implementação de estratégias de defesa. Os sistemas de reconhecimento facial, muitas vezes baseados em aprendizado profundo e redes neurais artificiais, são susceptíveis a ataques adversariais que envolvem perturbações quase imperceptíveis inseridas nas imagens de entrada. A metodologia empregada abrange uma revisão extensa da literatura, análise de conjuntos de dados de imagens de rosto para treinar e testar modelos, e a avaliação de diferentes abordagens de detecção e defesa. Os modelos são treinados utilizando técnicas de aprendizado profundo e, em seguida, são testados sob diversas condições, incluindo ataques adversariais. Variadas estratégias de defesa contra esses ataques são implementadas e testadas, com seus resultados subsequentemente analisados. Além disso, o estudo também procura incentivar a colaboração na área de segurança em IA, de forma a enfrentar os desafios associados aos ataques adversariais.**

Palavras-chave—Reconhecimento Facial, Ataques Adversariais, Inteligência Artificial, Estratégias de Defesa, Aprendizado Profundo

I. Introdução

À medida que a tecnologia de reconhecimento facial baseada em Inteligência Artificial ganha relevância em áreas como segurança, controle de acesso e autenticação biométrica, surgem desafios significativos relacionados à segurança desses sistemas [1]. Entre esses desafios, os ataques adversariais - técnicas que inserem perturbações quase imperceptíveis nas imagens de entrada para enganar modelos de IA - representam uma ameaça crescente, aumentando a necessidade de compreender e mitigar tais vulnerabilidades [5]. Este trabalho, motivado pelo imperativo de garantir a confiabilidade e a segurança dos sistemas de reconhecimento facial, busca aprofundar o entendimento sobre esses ataques adversariais e explorar possíveis estratégias de defesa.

II. Conceitos Fundamentais

A. Reconhecimento Facial

Este é um método de identificação ou verificação da identidade de uma pessoa usando seu rosto. Os sistemas de reconhecimento facial capturam, analisam e comparam padrões faciais para a identificação ou autenticação [6].

B. Aprendizado Profundo (Deep Learning)

Uma subcategoria do aprendizado de máquina, o aprendizado profundo é baseado em redes neurais artificiais com várias camadas (daí o termo "profundo"). Muitos sistemas modernos de reconhecimento facial são baseados em técnicas de aprendizado profundo [10].

C. Ataques Adversariais

Estas são entradas cuidadosamente projetadas para enganar modelos de aprendizado de máquina, fazendo com que eles façam classificações incorretas. No contexto do reconhecimento facial, ataques adversariais geralmente envolvem a manipulação sutil de imagens para enganar os sistemas de reconhecimento [7].

D. Perturbações Adversariais

São pequenas alterações feitas em imagens ou outros dados de entrada que levam os modelos de aprendizado de máquina a fazerem classificações incorretas. Essas perturbações são projetadas para serem indetectáveis ou quase imperceptíveis para os humanos [5].

E. Defesa contra Ataques Adversariais

Refere-se a técnicas usadas para tornar os modelos de aprendizado de máquina mais robustos contra ataques adversariais. Isso pode incluir a criação de redes neurais mais resistentes a perturbações adversariais, a detecção de entradas adversariais e a correção de previsões em caso de ataques [9].

F. Avaliação e Verificação de Modelos

São métodos usados para testar a eficácia e a robustez dos modelos de aprendizado de máquina. No contexto de reconhecimento facial, isso pode incluir a verificação da capacidade de um sistema de lidar com variações nas imagens dos rostos, como diferentes expressões faciais, iluminação e ângulos, e também sua resistência a ataques adversariais [3].

G. Segurança de IA

Um campo de pesquisa dedicado à proteção de sistemas de IA contra diferentes tipos de ameaças, incluindo ataques adversariais [11].

H. Redes Neurais Artificiais

As redes neurais artificiais são modelos computacionais baseados no funcionamento do cérebro humano que desempenham um papel fundamental na área da inteligência artificial. Compostas por neurônios artificiais interconectados em camadas, essas redes recebem inputs, realizam operações matemáticas e geram outputs [2].

Durante o treinamento, os pesos das conexões entre os neurônios são ajustados para minimizar os erros e permitir que a rede se generalize e aprenda com os dados [8]. Essa capacidade de aprendizado e o poder de lidar com problemas complexos tornam as redes neurais artificiais ferramentas poderosas em diversas áreas, como

reconhecimento de padrões, classificação de dados e processamento de linguagem natural.

I. Redes Generativas Adversariais

GANs são Redes Gerativas Adversariais (do inglês, Generative Adversarial Networks). Elas são uma classe de modelos de aprendizado profundo usados em aprendizado de máquina. A ideia fundamental é criar um cenário competitivo onde dois modelos (o gerador e o discriminador) competem entre si, permitindo a geração de saídas muito realistas [12].

Gerador (Generator): Este modelo aprende a criar dados realistas (por exemplo, imagens) a partir de um vetor de ruído aleatório. A ideia é que as saídas deste modelo sejam indistinguíveis dos dados reais.

Discriminador (Discriminator): Este modelo aprende a distinguir entre as saídas do gerador e os dados reais. A ideia é que este modelo seja capaz de reconhecer se uma entrada é uma amostra real ou gerada.

Durante o treinamento, o gerador busca melhorar sua capacidade de criar dados realistas para "enganar" o discriminador, enquanto o discriminador busca melhorar sua capacidade de distinguir entre dados reais e dados gerados. Este processo iterativo leva a melhorias simultâneas em ambos os modelos [12].

J. Validação Cruzada

A validação cruzada (do inglês, cross-validation) é uma técnica estatística utilizada para avaliar a capacidade de generalização de um modelo em um conjunto de dados independente, ou seja, avaliar quão bem um modelo de aprendizado de máquina é capaz de prever novos dados que não foram usados no treinamento. A ideia é garantir que o modelo não esteja apenas memorizando os dados de treinamento (overfitting), mas aprendendo padrões que permitam prever corretamente novos dados.

O procedimento de validação cruzada envolve a divisão do conjunto de dados em subconjuntos, e então o modelo é treinado em alguns desses subconjuntos (chamados conjuntos de treinamento) e validado/avaliado nos subconjuntos restantes (chamados conjuntos de validação ou teste). O processo é repetido várias vezes, com diferentes subconjuntos usados para treinamento e validação a cada vez. As métricas de desempenho são então agregadas para fornecer uma estimativa mais robusta do desempenho do modelo [13].

A validação cruzada é uma técnica amplamente utilizada na aprendizagem de máquina e é fundamental para prevenir o problema do overfitting, ou seja, quando um modelo se ajusta muito bem aos dados de treinamento, mas não generaliza bem para novos dados [14].

K. Dlib

Dlib é uma biblioteca de software moderna, de código aberto e multiplataforma, escrita em C++, mas com bindings extensivos para Python, que oferece uma variedade de algoritmos de aprendizado de máquina e ferramentas para desenvolvimento de software.

A biblioteca Dlib contém algoritmos para segmentação de imagens, detecção de características, transformações de perspectiva, reconhecimento de padrões e muitos outros.

Dlib se destaca pelo seu design flexível e eficiente, permitindo que os desenvolvedores utilizem as estruturas de dados e algoritmos do Dlib em seus programas como se fossem partes nativas da linguagem. [15]

L. VGG-Face

VGG-Face é um modelo de aprendizado profundo, especificamente uma rede neural convolucional (CNN), desenvolvido pelo Visual Geometry Group (VGG) da Universidade de Oxford para reconhecimento facial. O VGG-Face é amplamente reconhecido por sua eficácia no reconhecimento de faces em imagens e vídeos.

O modelo VGG-Face é derivado da arquitetura VGG-16, um dos modelos de CNN mais influentes no campo do reconhecimento de imagens. O VGG-Face foi treinado usando um conjunto de dados de 2,6 milhões de imagens de rostos, abrangendo 2.622 identidades.

O modelo se destaca pela sua precisão em tarefas de reconhecimento facial, ultrapassando muitos outros modelos pré-existentes no momento de sua publicação. Ele tem sido usado em uma ampla gama de aplicações, incluindo autenticação biométrica, análise de sentimentos e outras tarefas de processamento de imagem relacionadas a rostos.

M. The All Convolutional Net

"The All Convolutional Net" [16] é um modelo de aprendizado de máquina, especificamente uma rede neural convolucional (CNN), desenvolvido por Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox e Martin Riedmiller da Universidade de Freiburg para reconhecimento de objetos. A proposta desta rede convolucional é questionar e simplificar os componentes essenciais das CNNs, desafiando a necessidade de funções de ativação complexas, normalização de resposta ou max-pooling.

A principal inovação deste modelo é que ele substitui o max-pooling, geralmente encontrado em CNNs, por uma camada convolucional com passo aumentado, sem perda na precisão em vários benchmarks de reconhecimento de imagens. A estrutura da rede consiste exclusivamente em camadas convolucionais, o que a torna uma arquitetura simplificada em comparação aos modelos tradicionais.

III. Trabalhos Relacionados

Serão utilizados dois artigos como referência, ambos abordam diretamente o tema dos ataques adversariais. O trabalho intitulado "Why are Generative Adversarial Networks so Fascinating and Annoying?" explora as Redes Generativas Adversariais (GANs) como modelos generativos desafiadores, mas extremamente eficazes na geração de dados complexos. Ele apresenta as variações do GAN e discute suas aplicações na indústria do entretenimento, como transferência de estilo e tradução de imagens, além de abordar medidas de avaliação de qualidade de imagens geradas. Os autores Fabio Augusto Faria e Gustavo Carneiro são os responsáveis por esse trabalho [3].

O segundo trabalho intitulado "Generating Master Faces for Dictionary Attacks with a Network-Assisted Latent Space Evolution" aborda o problema dos ataques de dicionário em sistemas de autenticação baseados em

reconhecimento facial. Ele propõe uma abordagem baseada em algoritmos evolutivos e utiliza o espaço latente da rede geradora de faces StyleGAN para otimizar "faces mestras" que possam passar na autenticação de identidade facial para uma grande parte da população. O trabalho compara diferentes estratégias evolutivas e propõe o uso de uma rede neural para direcionar a busca em direção a amostras promissoras. Os autores deste trabalho são Ron Shmelkin, Tomer Friedlander e Lior Wolf [4].

Os dois estudos em questão, cuidadosamente selecionados, oferecem uma investigação profunda e multifacetada dos ataques adversariais, cada um esmiuçando um domínio particular de aplicação. O primeiro discurso se debruça sobre a esfera de ataques adversariais em modelos generativos, conduzindo o leitor através das complexidades e nuances desta área. O segundo, por sua vez, focaliza sua atenção em ataques de dicionário direcionados aos sistemas de reconhecimento facial, fornecendo um exame detalhado das estratégias e contramedidas pertinentes.

Deve ser mencionado que, embora exista uma vastidão de literatura disponível que poderia ser referenciada, os mencionados anteriormente foram especificamente escolhidos por seu conteúdo abrangente e relevante. A escolha por limitar a discussão a esses dois trabalhos não se baseia em uma falta de material disponível, mas em uma intenção consciente de focar na profundidade e qualidade da análise, em vez da quantidade de fontes.

A contribuição desses estudos à compreensão dos ataques adversariais é indiscutível. Eles elucidam os desafios inerentes a essa esfera de estudo, delineiam possíveis abordagens e iluminam o caminho para futuras pesquisas. Em última análise, esses trabalhos são de vital importância para qualquer um que busca explorar a problemática dos ataques adversariais de maneira aprofundada e consciente.

IV. Objetivo

O objetivo deste estudo é explorar e compreender os ataques adversariais direcionados a sistemas de reconhecimento facial baseados em Inteligência Artificial. Isso envolve analisar as técnicas utilizadas nos ataques, os impactos que podem ter nos sistemas de reconhecimento facial e as vulnerabilidades exploradas pelos adversários. Além disso, busca-se propor e avaliar estratégias de defesa contra esses ataques, com o objetivo de fortalecer a segurança e a confiabilidade dos sistemas de reconhecimento facial. O estudo envolverá a análise de casos conhecidos de ataques adversariais, a investigação de métodos de detecção e a proposição de soluções de defesa. Espera-se que essas contribuições ajudem a aumentar a conscientização sobre a segurança em IA e promovam a colaboração entre diferentes partes interessadas para enfrentar os desafios associados aos ataques adversariais em sistemas de reconhecimento facial.

V. Metodología Experimental

A metodologia deste estudo envolve duas etapas: geração de imagens via Rede Generativa Adversarial (GAN) e reconhecimento facial com a Convolutional Neural Network (CNN).

Utilizando a base de dados LFW (Labeled Faces in the Wild), processada para recorte e redimensionamento, a GAN gera imagens sintéticas. Esta rede é formada por um gerador, que cria imagens a partir de um código aleatório, e um discriminador, treinado para distinguir imagens reais das sintéticas, melhorando a qualidade das imagens geradas.

Na fase de reconhecimento facial, a CNN é usada, com uma arquitetura de camadas convolucionais, de agrupamento e classificação. Os dados LFW são divididos em treino e teste, e uma técnica de aumento de dados é aplicada para aprimorar a generalização do modelo. A performance do modelo é avaliada pela acurácia, matriz de confusão, taxa de verdadeiros positivos (recall) e taxa de identificação correta (precisão). Além disso, é adotado o protocolo de validação cruzada para avaliar a capacidade de generalização do modelo em diferentes conjuntos de dados.

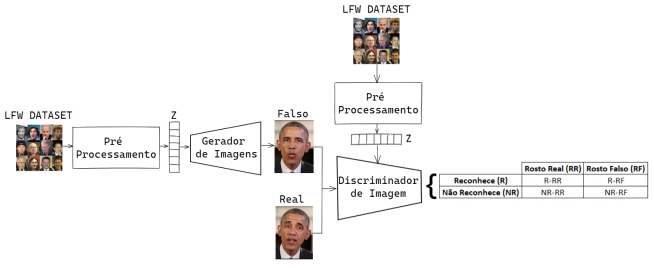
Os resultados são analisados qualitativa e quantitativamente, observando-se a acurácia, a matriz de confusão, a taxa de verdadeiros positivos (recall) e a taxa de identificação correta (precisão). Essas medidas de avaliação fornecem informações sobre o desempenho do sistema de reconhecimento facial, sua capacidade de corretamente identificar indivíduos e a precisão das suas previsões.

A etapa de geração de imagens sintéticas serve para simular ataques adversariais ao sistema de reconhecimento facial. A eficácia do discriminador em distinguir imagens reais das sintéticas é avaliada, visando melhorias caso seja facilmente enganado, aumentando assim a robustez do sistema.

Logo, a metodologia adotada avalia a eficácia geral do sistema de reconhecimento facial e sua robustez contra ataques adversariais, contribuindo para o desenvolvimento de sistemas mais eficientes e seguros. A utilização do protocolo de validação cruzada permite uma avaliação mais robusta e confiável do desempenho do modelo em diferentes conjuntos de dados, garantindo uma análise abrangente de sua capacidade de generalização.

Dessa forma, na Figura 1 é possível ver um diagrama de blocos que ilustra o funcionamento da parte experimental do projeto que foi resumido nos parágrafos acima.

Figura 1 - Diagrama de blocos do funcionamento do projeto.



Fonte: Autores.

A. *Revisão da literatura*

Será realizada uma revisão extensa da literatura para entender o estado atual dos ataques adversariais e das defesas em sistemas de reconhecimento facial. Este estudo abordará as técnicas de ataques adversariais mais recentes e as estratégias de defesa existentes.

B. Base de Dados

Será utilizado um conjunto de dados pré-existente, que contém imagens de rostos que serão usadas para treinar e testar os modelos de reconhecimento facial. Este conjunto de dados será analisado para garantir sua adequação ao estudo, e incluirá exemplos de ataques adversariais para permitir uma avaliação efetiva dos modelos.

Dessa forma, a base de dados utilizada neste trabalho será a LFW (Labelled Faces in the Wild) que é um banco de imagens amplamente utilizado na área de reconhecimento facial e aprendizado de máquina. O LFW é composto por uma coleção de mais de 13.000 imagens de rostos de pessoas coletadas da internet. Essas imagens não são controladas, ou seja, foram obtidas em condições reais, com variações significativas de poses, expressões faciais, iluminação e oclusões.

Cada imagem no LFW é acompanhada de um rótulo correspondente, que indica a identidade da pessoa retratada. O banco de dados contém um total de 5.749 identidades diferentes, com um número variável de imagens para cada pessoa.

Uma das características distintivas do LFW é a sua natureza desafiadora. As imagens contêm uma ampla variedade de variações, como diferentes ângulos de visão, expressões faciais complexas e oclusões parciais. Isso torna o reconhecimento facial uma tarefa difícil e estimula o desenvolvimento de algoritmos mais robustos.

C. Treinamento e Avaliação de Modelos

Neste estudo, empregamos como métricas de avaliação a acurácia e a precisão. Essas medidas, tradicionalmente utilizadas no campo do reconhecimento facial, desempenham uma função vital na avaliação da performance do modelo estudado.

A acurácia, representando a porcentagem de classificações corretamente realizadas pelo modelo em relação ao universo total de amostras avaliadas, é indicativa da capacidade do modelo em identificar apropriadamente as faces dos indivíduos. Um percentual elevado de acurácia sinaliza uma performance superior do modelo. No código disponibilizado para análise, essa métrica é calculada e empregada para avaliar a performance do modelo de reconhecimento facial.

A precisão é uma outra métrica de importância adotada nesta pesquisa. Ela é definida como a proporção entre as amostras positivas corretamente classificadas e o conjunto total de amostras que o modelo categorizou como positivas. No campo do reconhecimento facial, a precisão demonstra a taxa de identificações corretas das faces dos indivíduos. Uma precisão elevada implica em um menor número de falsos positivos, isto é, um menor número de faces erroneamente identificadas como pertencentes a uma pessoa específica. No código disponibilizado, a precisão é calculada e utilizada como uma das métricas de avaliação para examinar a performance do modelo de reconhecimento facial.

A união da acurácia e precisão viabiliza uma avaliação abrangente e robusta da performance do modelo de reconhecimento facial. Enquanto a acurácia fornece uma métrica global de acertos do modelo, a precisão complementa tal informação ao focar na taxa de identificação correta, reduzindo a ocorrência de falsos positivos. Ambas as métricas são cruciais para verificar a eficácia do modelo e contribuir para o desenvolvimento de sistemas de reconhecimento facial mais acurados e confiáveis.

D. Discriminador de Imagem (Reconhecimento Facial)

No estudo em questão, a parte de reconhecimento facial é realizada por meio de uma Convolutional Neural Network (CNN). A CNN é uma arquitetura de rede neural amplamente utilizada em tarefas de processamento de imagens, incluindo o reconhecimento facial.

A arquitetura da CNN adotada consiste em camadas convolucionais, camadas de agrupamento e camadas densas. Essas camadas são projetadas para extrair características significativas dos rostos humanos e realizar a classificação com base nessas características.

No início da arquitetura, as camadas convolucionais aplicam filtros espaciais nas imagens de rostos, visando detectar padrões e características locais, como bordas, texturas e detalhes faciais relevantes. Essas camadas são seguidas pelas camadas de agrupamento, que reduzem a dimensionalidade das características extraídas, preservando as informações mais importantes.

Após as camadas convolucionais e de agrupamento, temos camadas densas que são responsáveis por realizar a classificação final dos rostos. Essas camadas transformam as características extraídas em uma representação adequada para a classificação e utilizam uma função de ativação softmax para produzir as probabilidades de pertencer a cada classe de pessoa.

Durante o treinamento da CNN no estudo, é utilizada a função de perda conhecida como categorical cross-entropy, que compara as probabilidades produzidas pelo modelo com os rótulos verdadeiros dos dados e penaliza as previsões incorretas. O otimizador Nadam é empregado para otimizar os pesos da rede neural com base nessa função de perda.

No contexto do estudo, o desempenho da CNN é avaliado por meio de medidas de avaliação como acurácia e precisão. Essas medidas de avaliação fornecem informações valiosas sobre o desempenho do modelo de reconhecimento facial e sua capacidade de realizar classificações precisas. No contexto do estudo, elas são utilizadas para analisar a eficácia do discriminador em distinguir imagens reais de imagens sintéticas geradas pela GAN, contribuindo para a avaliação e aprimoramento do sistema de reconhecimento facial proposto.

E. Análise e Discussão dos Resultados

O estudo consistiu na geração de imagens de rostos humanos utilizando uma Rede Generativa Adversarial (GAN) e no reconhecimento facial por meio de uma Convolutional Neural Network (CNN).

Na etapa de geração de imagens, a GAN foi treinada para criar imagens sintéticas a partir de códigos aleatórios. A qualidade das imagens geradas foi avaliada visualmente, observando-se a capacidade da GAN em produzir imagens realistas e convincentes. No entanto, é importante ressaltar que a avaliação qualitativa é subjetiva e pode variar de acordo com o observador.

Em seguida, o foco foi direcionado para a etapa de reconhecimento facial utilizando a CNN. A CNN foi treinada com imagens da base de dados LFW (Labeled Faces in the Wild). O modelo foi avaliado utilizando medidas de desempenho, como a acurácia e a precisão.

A acurácia média obtida durante os testes foi de 85%, o que indica uma boa capacidade do modelo em realizar classificações corretas. A matriz de confusão foi utilizada para analisar a distribuição das classificações do modelo em relação aos rótulos verdadeiros. Essa análise permitiu identificar possíveis padrões de erros e avaliar o desempenho do modelo em cada classe.

A precisão média obtida foi de 63%, o que demonstra uma alta taxa de identificação correta dos rostos positivos. Isso significa que o modelo apresenta um baixo número de falsos positivos, ou seja, de casos em que identifica erroneamente um rosto como pertencente a uma determinada pessoa.

Além disso, foi observado que a CNN apresentou um bom desempenho geral durante o experimento, sendo capaz de realizar classificações precisas e confiáveis. No entanto, é importante destacar que o desempenho pode variar dependendo das condições de captura das imagens, como variações na iluminação, ângulos de visão e expressões faciais.

Em suma, os resultados obtidos demonstraram que a geração de imagens utilizando a GAN e o reconhecimento facial por meio da CNN são técnicas promissoras na área de reconhecimento facial. A acurácia e a precisão foram medidas eficazes para avaliar o desempenho do modelo. Esses resultados contribuem para o avanço do desenvolvimento de sistemas de reconhecimento facial mais eficientes e precisos, com aplicações em segurança, identificação pessoal e análise de dados visuais.

VI. Entregas

No final deste experimento, os seguintes entregáveis serão produzidos:

A. Relatório de Análise de Dados

Um relatório detalhado apresentando a análise do conjunto de dados de imagens de rosto utilizadas, incluindo uma descrição das características das imagens e a identificação de qualquer padrão relevante.

B. Avaliação dos Modelos de Reconhecimento Facial

Uma avaliação completa dos modelos de reconhecimento facial treinados, incluindo seu desempenho sob condições normais e ataques adversariais. Este relatório incluirá métricas detalhadas de desempenho.

C. Análise das Estratégias de Defesa

Um relatório detalhado sobre as estratégias de defesa implementadas, incluindo sua eficácia na detecção e mitigação de ataques adversariais. Dessa forma, pretende-se melhorar o discriminador do reconhecimento facial, utilizando GANs, para que ele aceite menos rostos ao se utilizar uma face mestra, ou seja, diminuir a porcentagem de aceitação dos rostos genéricos como rostos válidos.

D. Discussão e Recomendações

Um documento que resume as descobertas do estudo, discute suas implicações e fornece recomendações para futuras pesquisas e desenvolvimento na área de reconhecimento facial e segurança de IA.

E. Código-fonte e Dados

Todo o código-fonte usado no estudo, juntamente com os dados gerados durante o experimento, serão disponibilizados para futuras pesquisas e replicação de resultados.

Links dos relatórios de acompanhamento do experimento no Google Colab:

- Reconhecimento Facial (CNN): https://colab.research.google.com/drive/19zRXRn-CxL-EsCC5Mpt6TO1tbMVI7LGS?usp=sharing
- Geração de Imagens (GAN): https://colab.research.google.com/drive/1y1hMOVLjoaLPU7lqssf8B71hmFssmkGN?usp=sharing

Estes entregáveis fornecerão um registro completo das descobertas do estudo, além de servir como base para futuras pesquisas e desenvolvimento na área de defesa contra ataques adversariais em sistemas de reconhecimento facial.

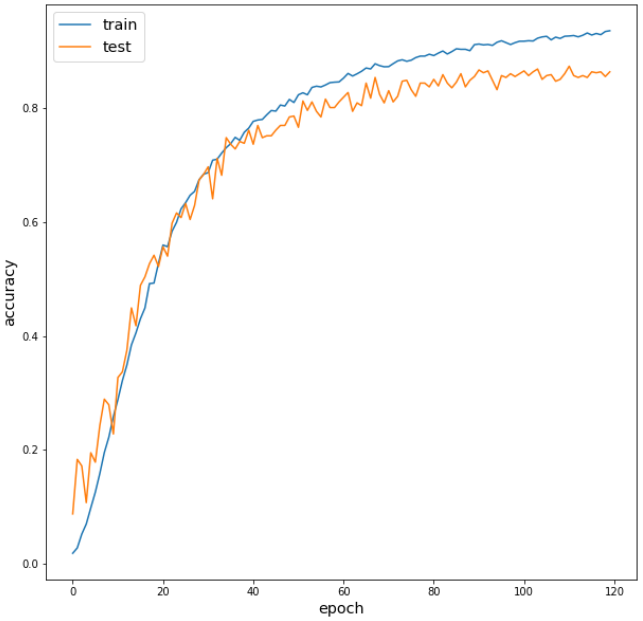
VII. Experimentos

A. Experimentos com Reconhecimento Facial

No início da investigação experimental, foram empregados o dlib e o VGG face na tentativa de replicar o estudo principal, Master Face [4]. No entanto, a implementação usando o Google Colaboratory se mostrou impraticável, devido à alta demanda computacional e à complexidade da rede neural convolucional (CNN) envolvida.

Desta forma, a metodologia foi alterada para The All Convolutional Net [16]. Com essa abordagem, foi possível alcançar uma acurácia de 85% utilizando a base de dados LFW. Este resultado encorajou a mudança de metodologia para a continuação do experimento e para a subsequente implementação da rede generativa adversarial (GAN).

Figura 2 - Gráfico de Treino e Teste da CNN utilizando o método The All Convolutional Net com 85% de acurácia.



B. *Experimentos com Rede Generativa Adversarial (GAN)*

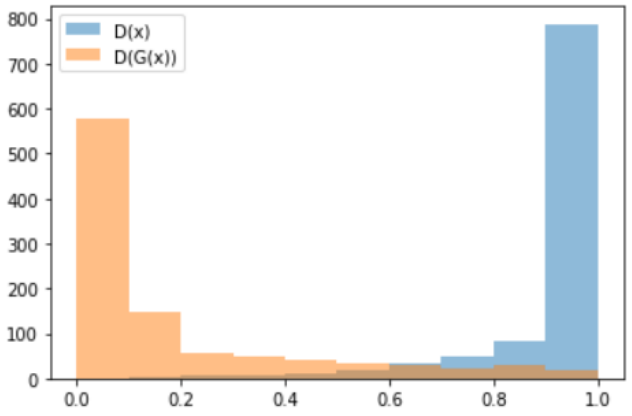
A Rede Generativa Adversarial (GAN), neste estudo, desempenha um papel crucial na melhoria da qualidade das imagens geradas artificialmente. O objetivo principal é otimizar a acurácia dessas imagens, buscando torná-las o mais semelhante possível às imagens reais.

A etapa inicial do experimento com a GAN evidenciou resultados notáveis, porém com espaço considerável para melhorias. A acurácia do reconhecimento facial das imagens originais atingiu a marca de 90% em 80% das instâncias testadas. Em contraste, apenas 2% das imagens geradas pela GAN alcançaram essa mesma marca. Esses resultados iniciais destacaram a necessidade de aprimorar o desempenho da GAN.

A questão primordial desta pesquisa reside em melhorar o discriminador de reconhecimento facial. Busca-se aprimorar a sua capacidade de distinção para que possa resistir com sucesso a ataques adversários. Em outras palavras, o foco é aumentar a eficácia do discriminador em identificar corretamente as imagens, sejam elas reais ou artificialmente geradas, mesmo quando as últimas são otimizadas pela GAN para se assemelharem fortemente às verdadeiras.

Para avaliar o sucesso desse empreendimento, o experimento será mensurado com base na acurácia obtida pelo discriminador. Essa métrica atua como um reflexo direto da capacidade do discriminador em resistir aos ataques da GAN e em classificar corretamente as imagens, sejam elas autênticas ou artificialmente geradas. Portanto, os esforços futuros serão direcionados para otimizar esse aspecto do experimento, para aprimorar o desempenho tanto da GAN quanto do discriminador.

Figura 3 - Histograma de confiança do discriminador em classificar imagens reais D(x) comparado a confiança da classificação de imagens geradas artificialmente D(G(x)).



VIII. Discussão de Resultados

Neste estudo, enfrentamos desafios significativos, sendo os principais a otimização da qualidade das imagens produzidas pela Rede Generativa Adversarial (GAN) e o aprimoramento do discriminador de reconhecimento facial.

No estágio inicial do experimento com a GAN, houve um contraste marcante entre a acurácia de reconhecimento das imagens autênticas e das artificialmente geradas. Enquanto as imagens originais obtiveram um índice de reconhecimento superior a 90% em 80% dos casos, as imagens geradas pela GAN alcançaram essa mesma taxa de acurácia em somente 2% dos casos. Essa discrepância ressalta claramente a necessidade de otimizar a GAN.

A finalidade principal desta pesquisa é fortalecer a capacidade do discriminador para que ele possa identificar corretamente imagens reais e falsas. Isso se aplica mesmo quando as imagens falsas são aprimoradas pela GAN para se parecerem muito com as verdadeiras. Portanto, o desenvolvimento de um discriminador robusto o suficiente para resistir a ataques adversários bem-sucedidos foi um dos maiores desafios que enfrentamos.

A pesquisa também enfrentou obstáculos relacionados à capacidade computacional disponível para o treinamento do modelo. A utilização do Google Colab gratuito para processamento não conseguiu atender às demandas computacionais do modelo.

No entanto, apesar desses desafios, os resultados obtidos até o momento são promissores. Eles indicam a viabilidade de continuar a investigação em prol da melhoria do discriminador, o que potencialmente pode aumentar a segurança de mecanismos de reconhecimento facial. Esta perspectiva também ressalta a necessidade de uma estratégia de alocação de recursos mais eficaz e de um investimento maior em termos de capacidade computacional para futuros experimentos.

Referências

[1] TEXEIRA, R. F. da S. ; Rafael B. JANUZI, R. B. ; A. FARIA, F. A. Os Dados dos Brasileiros sob Risco na Era da Inteligência Artificial? Encontro Nacional de Inteligência Artificial e Computacional (ENIAC), 2022

[2] J. Zaki and W. Meira Jr., "Data Mining and Machine Learning: Fundamental Concepts and Algorithms," 2nd ed., Cambridge University Press, Mar. 2020. [Online]. Disponível: https://dataminingbook.info/book_html/ ISBN: 978-1108473989.

[3] FARIA, F A.; CARNEIRO, G. Why are Generative Adversarial Networks so Fascinating and Annoying?. 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), 2020, pp. 1-8.

[4] R. Shmelkin, T. Friedlander, and L. Wolf, "Generating Master Faces for Dictionary Attacks with a Network-Assisted Latent Space Evolution," in Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2021,pp.1-9.

[5] L. da Silva Biff, “Um Método de Ataque Adversarial a Redes Neurais Convolutivas para Reconhecimento Facial,” Monografia, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2020. [Online]. Disponível: https://www.lume.ufrgs.br/bitstream/handle/10183/212988/001117221.pdf .

[6] J. C. Santos e J. B. Silva, “Reconhecimento facial utilizando redes neurais artificiais,” Caderno de Graduação - Ciências Exatas e Tecnológicas - UNIT - SERGIPE, vol. 3, no. 3, pp. 135-144, 2016. [Online]. Disponível: https://periodicos.set.edu.br/cadernoexatas/article/view/1761/1021

[7] I. J. Goodfellow et al., “Generative Adversarial Networks,” arXiv preprint arXiv:1406.2661, 2014. [Online]. Disponível: https://arxiv.org/pdf/1406.2661.pdf

[8] L. Fleck, M. H. F. Tavares, E. Eyng, A. C. Helmann e M. A. de M. Andrade, “Redes neurais artificiais: princípios básicos,” Revista Eletrônica Científica Inovação e Tecnologia, vol. 7, no. 15, 2016. [Online]. Disponível: https://periodicos.utfpr.edu.br/recit/article/view/4330/Leandro

[9] Z. Yuan, K. He, Z. Zhang, “Adversarial Attacks and Defenses in Deep Learning,” Engineering, vol. 5, no. 4, pp. 693-702, 2019. [Online]. Disponível: https://www.sciencedirect.com/science/article/pii/S209580991930503X

[10] T. Dettmers, “Deep Learning in a Nutshell: History and Training,” NVIDIA Developer Blog, Dec 16, 2015. [Online]. Disponível: https://developer.nvidia.com/blog/deep-learning-nutshell-history-training/

[11] R. Raimundo and A. Rosário, “The Impact of Artificial Intelligence on Data System Security: A Literature Review,” Sensors (Basel), vol. 21, no. 21, p. 7029, Nov. 2021. [Online]. Disponível: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8586986/

[12] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y. (2014). Generative Adversarial Networks. ArXiv, abs/1406.2661. [Online]. Disponível: https://arxiv.org/abs/1406.2661

[13] Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In Proceedings of the 14th

international joint conference on Artificial intelligence - Volume 2 (pp. 1137-1143). Montreal Quebec, Canada: Morgan Kaufmann Publishers Inc.

[14] Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. Statistics Surveys, 4, 40-79. doi:10.1214/09-SS054. [Online]. Disponível: https://arxiv.org/abs/0907.4728

[15] King, D.E. (2009). Dlib-ml: A Machine Learning Toolkit. Journal of Machine Learning Research, 10, 1755-1758.

[16] Springenberg, J. T., Dosovitskiy, A., Brox, T., Riedmiller, M. (2015). Striving for Simplicity: The All Convolutional Net. ArXiv, abs/1412.6806. [Online]. Disponível: https://arxiv.org/pdf/1412.6806