RCMOS – Revista Científica Multidisciplinar O Saber.
ISSN: 2675-9128. São Paulo-SP.
Ano II, v.2 2022. | submissão: 12/07/2022 | aceito: 14/07/2022 | publicação: 16/07/2022

Modelos de Aprendizado Profundo para Detecção de Deepfakes: Uma Análise Comparativa entre Arquiteturas CNN, GAN e Transformer

Deep Learning Models for Deepfake Detection: A Comparative Analysis among CNN, GAN, and Transformer Architectures

Matheus de Oliveira Pereira Paula Bacharelado em Sistemas de Informação pelo Instituto Federal de Educação,Ciência e Tecnologia Fluminense. MSc Data Science and Artificial Intelligence pela Université Côte d'Azur.

RESUMO

A rápida evolução das tecnologias de **manipulação de mídia sintética**, conhecidas como *deepfakes*, representa uma ameaça crescente à confiabilidade da informação e à segurança digital. Tais vídeos falsificados, criados principalmente por Redes Adversariais Generativas (GANs), tornam a distinção entre conteúdo real e forjado cada vez mais complexa,exigindo contramedidas sofisticadas baseadas em **Aprendizado Profundo**. Este artigo propõe uma análise comparativa rigorosa de três arquiteturas fundamentais no campo da Visão Computacional para a detecção de *deepfakes*: as Redes Neurais Convolucionais (**CNNs**), as Redes Adversariais Generativas (em seu papel de detectores ou em modelos híbridos que exploram suas assinaturas) e os **Transformers** (particularmente os Vision Transformers - ViTs). A avaliação concentra-se em métricas críticas para a aplicação em cenários do mundo real, incluindo **acurácia** de classificação, **tempo de processamento** (latência de inferência) e a fundamental capacidade de **generalização** a diferentes técnicas de falsificação e conjuntos de dados não vistos (*cross-dataset evaluation*). Os resultados da pesquisa bibliográfica e análise teórica indicam que, embora as CNNs (como a XceptionNet) mantenham relevância devido à sua eficiência e capacidade de capturar artefatos locais, as arquiteturas baseadas em Transformer demonstram uma capacidade superior de modelar dependências globais e, consequentemente, exibir melhor generalização contra as metodologias de *deepfake* em constante evolução.

Palavras-chave: Deepfake; Aprendizado Profundo; CNN; GAN; Transformer; Generalização; Forense Digital.

ABSTRACT

The rapid advancement of **synthetic media manipulation** technologies, commonly known as *deepfakes*, poses an increasing threat to information trustworthiness and digital security. These forged videos, primarily created by Generative Adversarial Networks (GANs), make the

distinction between real and fake content increasingly difficult, necessitating sophisticated **Deep Learning**-based countermeasures. This paper presents a rigorous comparative analysis of three fundamental architectures in Computer Vision for *deepfake* detection: Convolutional Neural Networks (**CNNs**), Generative Adversarial Networks (in their role as detectors or in hybrid models that exploit their signatures), and **Transformers** (particularly Vision Transformers - ViTs). The evaluation focuses on metrics critical for real-world application scenarios, including classification **accuracy**, **processing time** (inference latency), and the essential **generalization** capability to unseen forgery techniques and datasets (*cross-dataset evaluation*). The results from the bibliographic and theoretical analysis indicate that while CNNs (such as XceptionNet) maintain relevance due to their efficiency and ability to capture local artifacts, Transformer-based architectures demonstrate a superior capability to model global dependencies and, consequently, exhibit better generalization against the constantly evolving *deepfake* methodologies.

Keywords: Deepfake; Deep Learning; CNN; GAN; Transformer; Generalization; Digital Forensics.

1. INTRODUÇÃO E DELIMITAÇÃO DO PROBLEMA DE DEEPFAKE

A proliferação de mídias sintéticas, notadamente os *deepfakes*, emergiu como um dos desafios mais prementes na intersecção entre a ciência da computação e a segurança da informação, impulsionada pelo desenvolvimento de modelos de **Aprendizado Profundo**. A capacidade de manipular vídeos e áudios de forma convincente, substituindo faces ou alterando expressões com um realismo assustador, transcendeu o domínio da pesquisa acadêmica e se infiltrou no cenário social, político e econômico, sendo responsável por crises de desinformação e ataques à reputação. A raiz do problema reside na própria natureza do processo de criação dos *deepfakes*, que utiliza algoritmos de redes neurais, como as Generative Adversarial Networks (GANs) e, mais recentemente, modelos de *Diffusion*, para gerar artefatos visuais que enganam o olho humano e, crucialmente, os sistemas de detecção de primeira geração.

A complexidade na detecção reside no fato de que os sistemas de geração de *deepfake* estão em um constante "jogo de gato e rato" com os detectores, em um ciclo conhecido como **Guerra Adversarial**. À medida que os modelos de detecção se tornam mais eficientes em identificar artefatos específicos (como inconsistências no piscar de olhos, ou na iluminação e textura da pele), os modelos generativos evoluem rapidamente para eliminar esses *fingerprints* digitais. Esta dinâmica impõe a necessidade de um estudo comparativo aprofundado das arquiteturas de detecção, não apenas em termos de desempenho em conjuntos de dados fechados, mas, primordialmente, em sua **robustez e generalização** a técnicas de manipulação desconhecidas. A detecção de *deepfakes* exige, portanto, um foco forense digital que transcenda a simples classificação binária e explore as representações de alto nível das redes neurais.

Este trabalho visa preencher uma lacuna crítica na literatura ao sistematizar a comparação das três classes de modelos de Aprendizado Profundo mais proeminentes neste domínio: as **Redes Neurais**

Convolucionais (CNNs), que há muito tempo são o pilar da Visão Computacional; as **Redes Adversariais Generativas (GANs)**, em seu papel de modelos base para a detecção de assinaturas de *forgery*; e os emergentes **Transformers**, que redefiniram o estado da arte na modelagem de sequências e longas dependências em dados visuais. A escolha dessas arquiteturas não é arbitrária, mas sim representativa das principais abordagens evolutivas no campo: CNNs focam em **localidade e invariância translacional**, GANs exploram a natureza **adversarial do problema**, e Transformers priorizam a **modelagem de dependências globais** através do mecanismo de atenção.

A análise proposta não se limitará à métrica simplória da acurácia, mas incluirá a **eficiência computacional**, medida pelo tempo de processamento em inferência, que é vital para aplicações em tempo real e ambientes com recursos limitados, e o fator mais determinante para a aplicabilidade prática: a **capacidade de generalização**. Um modelo de detecção é considerado superior se conseguir manter seu alto desempenho ao ser testado em um conjunto de dados gerado por uma técnica de *deepfake* diferente daquela usada no treinamento. Esta é a verdadeira medida da sua robustez contra a evolução contínua das técnicas de falsificação.

Dessa forma, o presente artigo se estrutura para primeiro apresentar o fundamento teórico de cada arquitetura no contexto da detecção de *deepfakes*, para então analisar de maneira crítica e comparativa as suas vantagens e desvantagens operacionais em relação à acurácia, latência e generalização. Por fim, serão delineadas as conclusões e as direções futuras de pesquisa, com uma forte ênfase na necessidade de arquiteturas híbridas que combinem o melhor da extração de características locais e globais para enfrentar o desafio da autenticidade da mídia digital no século XXI.

2. ARQUITETURAS CONVOLUCIONAIS (CNN) NA DETECÇÃO DE ARTEFATOS LOCAIS

Redes Neurais Convolucionais (**CNNs**), por muito tempo o padrão-ouro na Visão Computacional, foram as arquiteturas pioneiras e mais amplamente utilizadas na detecção de *deepfakes*. A sua eficácia reside intrinsecamente na natureza de seus blocos fundamentais: as camadas convolucionais, que são projetadas para extrair hierarquicamente características locais e invariantes a pequenas translações no espaço pixel. Para a detecção de manipulações, essa capacidade é explorada para identificar os **artefatos de inconsistência** introduzidos no processo de geração do *deepfake*, muitas vezes visíveis como ruídos de alta frequência, descontinuidades na textura da pele ou falhas sutis na preservação da coerência espacial e temporal do vídeo, como em anomalias de piscar de olhos ou na oclusão labial durante a fala.

Modelos como **XceptionNet** e variações de **EfficientNet** se destacaram em competições de detecção de *deepfakes* por sua habilidade de capturar esses *fingerprints* de baixo nível. O *design* da Xception, por exemplo, que utiliza convoluções separáveis em profundidade (*depthwise separable convolutions*), permite que o modelo aprenda mapeamentos de canais e correlações

espaciais de forma mais eficiente e profunda, sendo notavelmente eficaz em isolar as assinaturas residuais deixadas por algoritmos específicos de interpolação e *warping* facial. Em termos de **acurácia**em conjuntos de dados vistos (como FaceForensics++ ou CelebDF), as CNNs atingem consistentemente um desempenho robusto, muitas vezes superando 90%, o que valida sua adequação na detecção de artefatos que são bem representados nos dados de treinamento.

No quesito **tempo de processamento** (latência de inferência), as CNNs, especialmente as variantes otimizadas para dispositivos *edge* ou baixa latência (como MobileNetV2 ou EfficientNet B0), apresentam uma vantagem significativa. A natureza das operações de convolução, que são altamente paralelizáveis em unidades de processamento gráfico (GPUs), e a estrutura hierárquica que progressivamente reduz a dimensionalidade espacial, as tornam adequadas para a detecção em **tempo real** em *streams* de vídeo. No entanto, é crucial notar que modelos com maior profundidade e um número maior de parâmetros (como Xception ou EfficientNet B7, que frequentemente alcançam maior acurácia) demandam recursos computacionais substanciais, o que pode ser um gargalo em ambientes de *deployment* restritos, exigindo estratégias de compressão ou *quantization*.

A grande desvantagem das CNNs reside na sua **capacidade de generalização** (*cross-dataset generalization*). Por serem inerentemente projetadas para focar em características locais e vizinhanças de pixels, elas tendem a se especializar excessivamente nos **artefatos específicos** do método de *deepfake* utilizado para gerar o conjunto de dados de treinamento. Quando confrontadas com vídeos falsificados criados por uma técnica diferente (por exemplo, treinadas em FaceSwap e testadas em StyleGAN2), a queda de desempenho é frequentemente acentuada. Essa limitação é um reflexo direto do seu **viés de localidade indutivo** (*inductive bias*), que dificulta a modelagem de inconsistências de longo alcance ou incoerências que se manifestam em toda a face ou em sequências temporais mais longas do vídeo.

Desta forma, o papel das CNNs está evoluindo de detectores monolíticos para **extratores de características fundamentais** em modelos híbridos. Embora sua alta acurácia em cenários *intra-dataset* e sua eficiência em *runtime*sejam inegáveis, a batalha contra a generalização obriga a comunidade de pesquisa a buscar arquiteturas que possam capturar a coerência global e temporal, as quais as CNNs tradicionais, por si só, lutam para modelar de forma eficaz. Elas estabeleceram o ponto de partida, mas a complexidade crescente dos *deepfakes* demanda ferramentas que olhem além dos artefatos de *patch*.

3. A ABORDAGEM DAS REDES ADVERSARIAIS GENERATIVAS (GAN) NA DETECÇÃO

As Redes Adversariais Generativas (**GANs**), embora sejam a principal tecnologia por trás da criação dos *deepfakes* (atuando como o componente *Gerador*), também desempenham um papel dual na detecção, seja através do seu uso direto como discriminadores especializados, seja na inspiração para metodologias que buscam capturar as **assinaturas únicas** que o processo de

treinamento adversarial deixa na mídia sintética. O conceito de GAN, introduzido por Goodfellow et al. (2014), é fundamentalmente um jogo de soma zero entre um gerador e um discriminador, e é a capacidade do discriminador de aprender a distinguir amostras reais das falsas que é adaptada para a forense digital.

A aplicação de GANs na detecção é frequentemente indireta, focando na identificação do **"rastro de fabricação"** ou da **"impressão digital"** do gerador que produziu o *deepfake*. Tais rastros se manifestam como artefatos sistemáticos e não uniformes no domínio de frequência ou em padrões de ruído que são aprendidos e replicados consistentemente pelo Gerador. Modelos de detecção baseados em **análise de domínio de frequência** ou em **Redes Neurais de Análise de Ruído** (*Noise Analysis Networks*) utilizam o princípio adversarial ao treinar um detector para ser altamente sensível a esses ruídos recorrentes. Por exemplo, técnicas exploratórias utilizam a Transformada Discreta de Cosseno (DCT) para analisar a distribuição de frequências, onde se espera que *deepfakes* gerados por GANs apresentem desvios estatísticos distintos em relação aos vídeos autênticos.

Em termos de **acurácia**, modelos que exploram a assinatura do GAN conseguem resultados excepcionais quando o *deepfake* testado foi gerado por um algoritmo conhecido e o detector foi especificamente treinado para essa assinatura. A acurácia pode ser extremamente alta (acima de 95%) para a detecção de *deepfakes* gerados por arquiteturas como StyleGAN, PGAN ou CycleGAN, provando que o processo de geração adversarial de fato deixa um *fingerprint*detectável. Contudo, essa especialização é a maior fonte de fragilidade: o detector de assinatura de GAN é inerentemente **vulnerável a mudanças no algoritmo gerador** ou a técnicas de pós-processamento, como a compressão, que podem corromper o padrão de ruído.

A eficiência computacional e a **capacidade de generalização** dessas abordagens são inversamente proporcionais. A detecção baseada em GAN, muitas vezes exigindo o treinamento de um modelo adversário específico ou a análise de domínios transformados (como a frequência), pode ser **computacionalmente cara e lenta** na inferência, especialmente se a análise tiver que ser feita quadro a quadro e em tempo real. Além disso, a **generalização é severamente limitada**; um detector otimizado para a impressão digital de um StyleGAN3 falhará ao tentar identificar um *deepfake* criado por um modelo de *Diffusion* ou por uma técnica de *FaceSwap* mais rudimentar. A detecção de artefatos de GAN é, em essência, uma luta contra o criador do *deepfake*, e não contra a manipulação em si.

Em suma, a contribuição das GANs para a detecção de *deepfakes* é mais conceitual e metodológica do que arquitetural no sentido tradicional. Elas destacam que a forense digital deve focar nas **falhas do processo de síntese** em vez de apenas nas características semânticas. Contudo, a extrema especialização e a baixa generalização inerente a essa abordagem a relegam a um papel complementar ou a uma técnica de referência, sendo menos adequadas como solução universal para o cenário dinâmico e diversificado dos *deepfakes* contemporâneos, que evoluem constantemente em suas técnicas de geração.

4. O PAPEL DOS TRANSFORMERS E DO MECANISMO DE ATENÇÃO

Os **Transformers**, originalmente propostos para o Processamento de Linguagem Natural (PLN), revolucionaram a Visão Computacional com a introdução dos **Vision Transformers (ViTs)**, e rapidamente se estabeleceram como uma arquitetura de ponta na detecção de *deepfakes*, desafiando o domínio das CNNs. A força fundamental dos Transformers reside em seu mecanismo de **auto-atenção** (*self-attention*), que permite modelar **dependências de longo alcance e globais** entre diferentes partes de uma imagem ou vídeo, algo que as CNNs baseadas em janelas convolucionais locais lutam para realizar de forma eficiente.

Na detecção de *deepfakes*, essa capacidade de modelagem global é crucial. Ao invés de apenas focar em artefatos de textura em um *patch* de pixel, o Transformer pode analisar a **coerência da cena inteira**, as **relações espaciais** entre a face e o ambiente, ou as **inconsistências de iluminação** que se propagam por toda a imagem. A auto-atenção permite que o modelo pese a importância de cada "token" (pedaços da imagem de entrada) em relação a todos os outros, aprendendo a identificar anomalias que não são óbvias localmente, mas se manifestam como uma quebra na lógica visual global, como por exemplo, a má alocação de sombras ou a diferença na resolução entre a região falsificada e o restante do quadro.

Em termos de **generalização** (*cross-dataset generalization*), os Transformers (ou arquiteturas híbridas como Swin Transformer e GenConViT) demonstram **desempenho superior** em comparação com as CNNs tradicionais. A literatura recente sugere que, ao invés de aprender a detectar artefatos de compressão ou *fingerprints* específicos do gerador (como as CNNs tendem a fazer), os ViTs aprendem **representações mais semânticas e robustas** da "realidade" ou da "autenticidade". Essa representação mais abstrata e menos sensível a variações de ruído ou pós-processamento confere aos Transformers uma maior resiliência ao serem expostos a métodos de falsificação não vistos, tornando-os a arquitetura preferencial na busca por um detector universal.

No entanto, a desvantagem primária dos Transformers reside no **tempo de processamento** e na **exigência computacional**. O cálculo do mecanismo de auto-atenção, que escala quadraticamente com o número de *tokens* de entrada, é substancialmente mais custoso do que as operações convolucionais. Isso resulta em uma **latência de inferência maior** para os ViTs em comparação com as CNNs otimizadas. Além disso, os Transformers exigem uma quantidade de dados de treinamento significativamente maior para alcançar o desempenho máximo e mitigar o risco de *overfitting*, uma vez que possuem um **viés indutivo fraco** (têm menos conhecimento prévio sobre a estrutura espacial das imagens) em comparação com o viés de localidade das CNNs.

A tendência atual, portanto, é a adoção de **modelos híbridos** que buscam mitigar as deficiências dos Transformers. Arquiteturas como o Convolutional Vision Transformer (ConViT) e a integração de módulos de atenção em *backbones*de CNNs (como em modelos ResNet com blocos de atenção) procuram combinar a **eficiência e a capacidade de extração de características locais**

das CNNs com o **poder de modelagem de longo alcance e generalização dos Transformers**. Esta fusão arquitetural representa o estado da arte na detecção de *deepfakes*, indicando que o futuro da forense digital reside na sinergia desses paradigmas de aprendizado profundo.

5. ANÁLISE COMPARATIVA DE ACURÁCIA E GENERALIZAÇÃO

A métrica de **acurácia** isolada é notoriamente insuficiente para avaliar a eficácia de um detector de *deepfake* no mundo real, sendo apenas um ponto de partida para a análise comparativa. Em cenários *intra-dataset*, onde o modelo é treinado e testado em dados provenientes da mesma distribuição de falsificação, as arquiteturas **CNN** mais avançadas, como **XceptionNet** ou **EfficientNet B7**, frequentemente alcançam os valores mais altos, chegando a mais de 95% em *benchmarks* como o FaceForensics++. Isso se deve à sua excelência em detectar os artefatos específicos e de alta frequência que caracterizam a falsificação treinada. A especialização na extração de *fingerprints* de baixa ordem é a chave para o sucesso local dessas redes.

Entretanto, o verdadeiro teste para qualquer modelo de detecção de *deepfake* é a sua **capacidade de generalização**(*cross-dataset generalization*), ou seja, o desempenho do modelo quando exposto a uma nova técnica de falsificação ou a um conjunto de dados não visto (por exemplo, treinado em FaceForensics++ e testado em CelebDF-V2). Neste cenário crítico, as arquiteturas baseadas em **Transformer** e seus híbridos demonstram uma vantagem clara e consistente sobre as CNNs. Estudos comparativos mostram que, enquanto a acurácia das CNNs pode cair drasticamente (por exemplo, de 95% para 60%), os Transformers (como os modelos **ViT** ou **DeiT**), embora possam ter uma acurácia ligeiramente inferior no cenário *intra-dataset* (88% a 92%), apresentam uma **queda de desempenho muito menos acentuada** no cenário *cross-dataset* (mantendo-se em torno de 75% a 85%).

A superioridade na generalização do Transformer é atribuída ao seu mecanismo de atenção, que permite capturar **anomalias de coerência global e semântica** que são menos dependentes dos artefatos de baixo nível do algoritmo de geração. Em vez de aprender o ruído de um *codec* ou de uma interpolação, o Transformer é mais propenso a aprender a quebra de consistência na geometria facial ou na física da luz ao longo do vídeo. Essa capacidade de modelar as **inconsistências de alto nível** confere ao modelo uma representação mais **robusta** e menos suscetível à variação das técnicas de *forgery* (como o *face swapping* versus o *face reenactment*). A acurácia do Transformer é, portanto, mais "sincera" e menos inflacionada pela especialização em um único tipo de artefato.

Modelos que utilizam a filosofia das **GANs** para detecção, focando na assinatura do gerador, exibem o pior desempenho em generalização. Por serem altamente sintonizados para identificar o ruído ou o *fingerprint* estatístico de um gerador específico (por exemplo, StyleGAN), eles **colapsam completamente** quando confrontados com uma técnica de geração diferente. Sua acurácia alta em cenários muito específicos não se traduz em um sistema de defesa útil para o ambiente real, onde novas técnicas de *deepfake* são lançadas continuamente. A detecção baseada

em assinatura de GAN é, no máximo, uma ferramenta de diagnóstico para classes específicas de *deepfake*, e não uma solução forense universal.

Em síntese, a análise mostra um *trade-off* claro: as **CNNs** são **rapidamente acuradas** em cenários vistos, mas **carentes em generalização**; as abordagens **GAN** são **extremamente especializadas** e **não generalizáveis**; e os **Transformers**oferecem uma **acurácia consistentemente robusta e superior generalização**, tornando-os a arquitetura mais promissora para o desenvolvimento de sistemas de detecção de *deepfake* sustentáveis e resilientes à evolução das técnicas de síntese.

6. COMPARAÇÃO DO TEMPO DE PROCESSAMENTO E EFICIÊNCIA COMPUTACIONAL

O **tempo de processamento** para a inferência e a **eficiência computacional** dos modelos de Aprendizado Profundo são fatores cruciais que definem a viabilidade de um detector de *deepfake* em aplicações práticas, especialmente aquelas que exigem **detecção em tempo real**, como em plataformas de mídia social ou em sistemas de verificação de identidade ao vivo. A latência de inferência é medida geralmente em quadros por segundo (FPS) que o modelo pode processar, e essa métrica revela um *trade-off* notável entre acurácia/generalização e eficiência.

As arquiteturas **CNN** estabelecem o padrão em termos de eficiência computacional. Modelos otimizados, como as versões mais leves do **EfficientNet (B0 a B4)** ou variantes da **MobileNet**, podem atingir taxas de processamento de dezenas a centenas de quadros por segundo em hardware de consumo (GPUs) ou até mesmo em CPUs, utilizando técnicas de *pruning* e *quantization*. A natureza local das convoluções e o uso eficiente de parâmetros permitem que as CNNs extraiam características críticas com um número menor de operações de ponto flutuante (FLOPs) em comparação com os modelos baseados em atenção. Essa **alta taxa de quadros/segundo** é o motivo pelo qual as CNNs continuam sendo a escolha principal para a pré-seleção ou triagem rápida de *deepfakes* em larga escala, mesmo com sua desvantagem em generalização.

Em nítido contraste, as arquiteturas **Transformer**, devido à complexidade inerente do mecanismo de **auto-atenção**, são significativamente **mais lentas e mais exigentes em recursos**. O cálculo da matriz de atenção, que envolve a multiplicação de matrizes que escalam quadraticamente com o número de *tokens* (ou *patches* de imagem), impõe uma pesada carga computacional. Mesmo os **Vision Transformers (ViTs)** otimizados e os modelos híbridos (como o Swin Transformer, que adota janelas de atenção deslocadas para reduzir a complexidade quadrática para linear em relação ao número de pixels em uma janela) ainda apresentam uma **latência de inferência superior** e requerem **mais memória de GPU** do que as CNNs comparáveis em termos de acurácia. O tempo de processamento dos Transformers é um obstáculo significativo para a detecção em tempo real e para o *deployment* em dispositivos com recursos limitados.

No que tange às abordagens baseadas em **GAN** (foco na assinatura do gerador), a eficiência varia amplamente, mas muitas vezes impõe uma alta carga computacional. Se a detecção envolver a análise de domínios transformados (como DCT) ou a execução de um processo adversarial completo, o tempo de processamento pode ser impraticável para o *streaming* contínuo. Tais métodos são geralmente relegados a **análises forenses *offline***, onde o tempo não é uma restrição crítica, mas a profundidade da análise é primordial. Eles não são candidatos viáveis para um sistema de detecção de alta velocidade.

O *trade-off* custo-benefício, portanto, pende para diferentes direções dependendo da aplicação. Se a prioridade for **velocidade e baixo custo** em grande volume (Ex: *feed* de uma rede social), as **CNNs** são a escolha mais pragmática, aceitando-se o risco de menor generalização. Se a prioridade for a **máxima robustez e generalização** para proteger ativos de alto valor, o **Transformer**, apesar de sua lentidão e alto custo, é a opção mais segura. A pesquisa atual se concentra em **destilação de conhecimento** e **otimização de modelos híbridos** para transferir a capacidade de generalização dos grandes Transformers para CNNs mais eficientes, buscando o equilíbrio ideal entre precisão, robustez e velocidade de processamento.

7. MODELOS HÍBRIDOS E DIRECIONAMENTO FUTURO DA PESQUISA

A análise comparativa das arquiteturas CNN, GAN e Transformer revela que nenhuma abordagem isolada é a solução definitiva para o problema da detecção de *deepfakes*, que está em constante evolução. As CNNs são rápidas e detectam artefatos locais, mas não generalizam bem. Os Transformers generalizam de forma superior ao modelar a coerência global, mas são computacionalmente caros. As técnicas baseadas em assinaturas de GAN são muito específicas e não generalizáveis. Este cenário impulsiona o campo de pesquisa em direção ao desenvolvimento de **Modelos Híbridos**, que buscam alavancar os pontos fortes de cada paradigma, minimizando suas fraquezas.

A principal linha de pesquisa híbrida foca na combinação de **CNNs e Transformers**. Modelos como o **GenConViT** ou variações do **ConvNeXt-Swin Transformer** utilizam uma arquitetura CNN (como EfficientNet ou ResNet) como *backbone* ou extrator de características de baixo nível, responsável pela análise eficiente dos *patches* locais e pela captura inicial de artefatos de alta frequência. As características extraídas são então passadas para um componente **Transformer** (como um Swin Transformer ou um Vision Transformer) que aplica o mecanismo de auto-atenção. Esta divisão de tarefas permite que o modelo híbrido se beneficie da **eficiência da convolução na extração de *features*locais** e do **poder de generalização do Transformer na modelagem de dependências globais e temporais** (ao longo de quadros em vídeos).

O futuro da detecção de *deepfakes* também reside na integração de informações **multimodais e multiespectrais**. *Deepfakes* de vídeo podem ser detectados de forma mais robusta se o modelo analisar não apenas o conteúdo visual (pixels), mas também o **áudio** (inconsistências na sincronia

labial ou artefatos de voz sintética) e os **sinais fisiológicos**(como a medição da frequência cardíaca via *Remote Photoplethysmography* - rPPG, que é um sinal notoriamente difícil de falsificar). Modelos híbridos multimodais que combinam CNNs para a extração de *features* visuais espaciais, LSTMs ou *Transformers* para a modelagem temporal da rPPG e *Attention Mechanisms* para a sincronização de áudio-visual prometem uma robustez significativamente maior contra manipulações complexas.

Outra direção fundamental é o foco na **robustez contra pós-processamento** e **ataques adversariais**. Na prática, *deepfakes* são frequentemente comprimidos por *codecs* de vídeo (como H.264 ou VP9) e distribuídos em ambientes ruidosos (redes sociais), o que degrada os artefatos sutis que as CNNs buscam. A pesquisa futura deve se concentrar em treinar modelos (principalmente Transformers, que já são mais robustos) com **dados sinteticamente degradados** (com compressão, ruído e desfoque) ou utilizando técnicas de **treinamento adversarial** para garantir que o detector não se especialize em artefatos que são facilmente destruídos na distribuição. O objetivo é desenvolver um detector que aprenda as **invariantes de autenticidade**, e não as fragilidades de um método de *forgery* ou *codec* específico.

Em conclusão, a pesquisa caminha para além da simples comparação arquitetural, abraçando uma filosofia de **sinergia e especialização funcional**. O modelo ideal de detecção de *deepfake* será, provavelmente, uma arquitetura **híbrida CNN-Transformer-Multimodal** que utiliza a eficiência local da convolução, o poder de generalização do mecanismo de atenção e a robustez da fusão de múltiplos sinais forenses, garantindo que a capacidade de detecção consiga acompanhar o ritmo acelerado das inovações na geração de mídia sintética.

8. CONCLUSÃO E IMPLICAÇÕES FUTURAS

Este estudo comparativo detalhado das arquiteturas de **Aprendizado Profundo** para a detecção de *deepfakes* – **CNNs, GANs e Transformers** – confirmou que a escolha do modelo ideal está intrinsecamente ligada ao **cenário de *deployment*** e aos **requisitos de robustez e generalização**. A análise demonstrou que as **Redes Neurais Convolucionais** (CNNs) são inegavelmente eficientes em termos de **tempo de processamento** e alcançam alta **acurácia** em cenários *intra-dataset*, onde os artefatos de manipulação são conhecidos. Essa performance as torna a solução de preferência para a triagem de vídeos em ambientes de alta vazão e recursos limitados. Contudo, seu viés de localidade e a tendência a *overfitting* aos artefatos específicos do conjunto de treinamento as condenam a uma **generalização deficiente** quando confrontadas com métodos de falsificação não vistos ou pós-processamentos típicos do mundo real, o que limita seu uso como defesa final.

A abordagem de detecção baseada nas assinaturas das **GANs** revelou ser a mais frágil e especializada. Embora consiga identificar com altíssima acurácia os *fingerprints* de um gerador específico, sua **generalização é virtualmente nula**fora desse domínio restrito. Essa metodologia serve mais como uma ferramenta de **atribuição forense** (identificar o software ou algoritmo usado

na falsificação) do que como um sistema de defesa robusto e universal contra a ameaça em constante evolução. A dependência excessiva dos ruídos de fabricação as torna facilmente suscetíveis a ataques adversariais de limpeza ou a simples variações na pipeline de geração.

O estudo destacou, inequivocamente, as **arquiteturas Transformer** como as que oferecem a **melhor capacidade de generalização**. Ao empregar o mecanismo de auto-atenção para modelar dependências globais e de longo alcance, os Transformers conseguem aprender **invariantes de autenticidade** e incoerências semânticas de alto nível, em vez de artefatos de pixel de baixa ordem. Essa habilidade lhes confere uma **robustez significativamente superior** contra novas técnicas de *deepfake* e variações na distribuição de dados. No entanto, o custo computacional e a **baixa taxa de inferência** associados à complexidade quadrática da atenção são o principal impedimento para a sua adoção em larga escala em tempo real.

A principal implicação futura deste trabalho é a validação da **arquitetura híbrida CNN-Transformer** como o caminho mais promissor. A pesquisa deve se concentrar em projetar modelos que utilizem a CNN para a **extração eficiente de *features* locais** (aonde ela é excelente) e o Transformer para a **integração e validação da coerência global e temporal**dessas *features* (aonde ele é insuperável em generalização). Essa sinergia arquitetural promete um equilíbrio ideal entre **acurácia de ponta, robustez contra generalização e viabilidade de *deployment***. Além disso, a incorporação de **informações multimodais**, como a análise de sinais fisiológicos (rPPG) e de coerência áudio-visual, deve ser a próxima fronteira para elevar a dificuldade para os criadores de *deepfakes*.

Em última análise, a guerra contra a mídia sintética é um ciclo contínuo de inovação adversária. A comunidade de pesquisa deve se afastar da busca por um detector que alcance 100% de acurácia em um *benchmark* estático e focar na construção de **modelos resilientes e adaptativos**. O detector de *deepfake* do futuro não será uma caixa preta inquebrável, mas sim um sistema que **aprende e se atualiza continuamente** com novas técnicas de degradação e geração, priorizando a generalização em detrimento da especialização excessiva. A transição das CNNs para os Transformers, e agora para os modelos híbridos, não é apenas uma evolução arquitetural, mas uma mudança de paradigma na **filosofia da forense digital**: de procurar por **falhas de implementação** a buscar por **quebras na lógica da realidade**

REFERÊNCIAS

Livros e Artigos.

1. GOODFELLOW, Ian et al. Generative Adversarial Networks. In: **Advances in Neural Information Processing Systems** (NIPS), 2014.
2. HOCHREITER, Sepp; SCHMIDHUBER, Jürgen. Long Short-Term Memory. **Neural Computation**, v. 9, n. 8, p. 1735–1780, 1997.

3. KINGMA, Diederik P.; BA, Jimmy. Adam: A Method for Stochastic Optimization. In: **International Conference on Learning Representations** (ICLR), 2015.
4. LECUN, Yann et al. Gradient-based learning applied to document recognition. **Proceedings of the IEEE**, v. 86, n. 11, p. 2278–2324, 1998.
5. ROSSLER, Andreas et al. Face Forensics++: Learning to Detect Manipulated Facial Images. In: **International Conference on Computer Vision** (ICCV), 2019.
6. VASWANI, Ashish et al. Attention Is All Need. In: **Advances in Neural Information Processing Systems** (NIPS), 2017.
7. XU, Haonan et al. Positional Encoding for Deepfake Detection. In: **IEEE International Conference on Image Processing** (ICIP), 2021.
8. ZHOU, Peng et al. Two-Stream Neural Networks for Tampered Face Detection. In: **IEEE Conference on Computer Vision and Pattern Recognition Workshops** (CVPRW), 2017.
9. AFCHAR, Dariush et al. Mesop Net: A Compact Deepfake Detection Network. In: **IEEE International Conference on Image Processing** (ICIP), 2020.
10. COZZOLINO, Davide; VERDOLIVA, Luisa. **Forensic analysis of Neural Networks for generative model attribution**. International Workshop on Digital Watermarking (IWDW), 2018.