Análise Comparativa de Representações de Imagens em Classificação de Figuras Sintéticas

Ada Maris Pereira Mário

[1]Relatório desenvolvido para a
Disciplina de Mineração de Dados não Estruturados - ICMC-USP

Resumo. *Este projeto investiga a eficácia de diferentes representações de imagem na classificação de figuras reais e sintéticas, usando o CIFAKE Dataset. Compararam-se três abordagens: histogramas de cores, descritores de textura de Haralick e embeddings de Residual Networks. O algoritmo k-Nearest Neighbors foi utilizado para avaliar o desempenho em identificar corretamente a origem das imagens com base nas suas características visuais. A hipótese avaliada via teste de Wilcoxon sugere que técnicas mais sofisticadas, como a ResNet-18, superam abordagens baseadas em características clássicas. Este estudo destaca a importância de escolher a representação adequada para lidar com os desafios modernos da mídia gerada por IA, especialmente em contextos que exigem autenticidade visual.*

1. Identificação do Problema

Os avanços em aprendizado de máquina permitiram recentemente a síntese hiperrealista de prosa, imagens, dados de áudio e vídeo, no que é chamado de mídia gerada por inteligência artificial (IA). Essas técnicas oferecem novas oportunidades para criar interações com retratos digitais de indivíduos que podem nos inspirar e intrigar [Pataranutaporn et al. 2021]. Contudo, tais inovações proporcionaram uma maior facilidade na geração dos chamados *deepfakes* (da junção de *"deep learning"* e *"fake"* — mídias que são geradas ou editadas usando ferramentas de IA que, em geral, representam elementos em situações inverídicas [Seow et al. 2022].

No que diz respeito às imagens geradas por tais métodos sintéticos em contextos digitais, muitas são visualmente convincentes ao ponto de mimetizarem características de imagens reais. Tais figuras, embora produzidas artificialmente, contêm atributos sofisticados que podem incluir, por exemplo, texturas complexas e reflexos fotorrealistas, gerando desafios na distinção entre imagens verdadeiras e sintéticas [Bird and Lotfi 2024]. Essa problemática torna-se ainda mais relevante à medida que tais imagens podem ser utilizadas em contextos que requerem autenticidade, como redes sociais, marketing, política e até investigações forenses, uma vez que a precisão na identificação da origem das imagens é crucial para assegurar a confiabilidade da informação visual apresentada.

O surgimento de técnicas avançadas, como redes generativas adversariais e modelos de difusão, contribuiu significativamente para a geração de imagens que simulam realismo, resultando na necessidade de desenvolvimento de métodos eficazes para detectar a diferença entre imagens reais e sintéticas. Isso tem despertado crescente atenção no campo da ciência de dados, onde abordagens automatizadas de mineração de imagens buscam identificar características que possam distinguir, de forma robusta, imagens geradas por IA das imagens capturadas em cenários reais [Corvi et al. 2023].

Nesse sentido, este projeto objetiva avaliar se, para diferentes técnicas de extração de características de imagens — neste caso, técnicas baseadas em *low-level features* como histogramas de cores e descritores de texturas, ou em redes neurais convolucionais como *Residual Networks* —, há diferenças significativas no desempenho das tarefas de classificação das imagens, estruturando o modelo a partir da própria representação dos dados ao utilizar o algoritmo *k*-Nearest Neighbors. Para avaliar as hipóteses (1), será utilizado o teste de Wilcoxon para dados pareados ao nível $\alpha = 5\%$ de significância, de modo a testar par a par e unilateralmente à direita as métricas obtidas com cada representação. Como resultado é esperado um desempenho dos modelos de pelo menos 80% de acurácia no conjunto de teste, com técnicas de extração de características mais sofisticadas apresentando desempenho superior em relação às representações clássicas (ou seja, espera-se rejeitar a hipótese nula).

$$
\begin{aligned}
&E(D) : \text{Esperança da diferença entre as métricas do 2° menos do 1° modelo} \\
&H_0 : E(D) \leq 0 \quad \text{(O desempenho do 2° modelo é igual ou menor que o 1°)} \\
&H_1 : E(D) > 0 \quad \text{(Há diferenças nos desempenhos dos modelos — o 2° modelo é melhor)}
\end{aligned} \tag{1}
$$

O conjunto de dados utilizado no desenvolvimento deste projeto se trata do CIFAKE Dataset [Bird and Lotfi 2024], contendo 60 mil imagens geradas sinteticamente e 60 mil imagens reais coletadas do conjunto CIFAR-10 [Krizhevsky et al. 2009]. Esse último consiste em um dataset com figuras de dimensão 32×32 e divididas em dez classes, que variam basicamente entre animais e meios de transporte, com 6 mil arquivos em cada uma. Os autores do CIFAKE geraram as imagens falsas utilizando difusão latente, de modo a espelhar as dez classes do dataset original, fornecendo um conjunto contrastante de imagens para comparação com fotografias reais, tendo sido capazes de gerar atributos visuais complexos, como reflexos fotorrealistas na água. Os dados foram disponibilizados em um conjunto de treino, com 100 mil imagens, e um conjunto de teste, com 20 mil imagens, ambos com as classes a serem aqui analisadas (real ou *fake*) igualmente distribuídas.

2. Pré-processamento

Inicialmente, os arquivos foram processados de modo a organizar suas informações e passar seus *labels* para binários (0 para real e 1 para *fake*), para em seguida iniciar a representação das imagens de forma vetorial, como será descrito nas seções a seguir.

2.1. Histograma de Cores

O histograma de cores é uma representação que quantifica a distribuição das intensidades de cor em uma imagem, caracterizando as frequências relativas das cores presentes. Em termos formais, um histograma de cores de uma imagem em RGB com três canais (vermelho, verde e azul) pode ser representado pela função $H(r, g, b)$, onde cada valor $H(i, j, k)$ indica o número de pixels na imagem que apresentam uma intensidade r para o canal vermelho, g para o canal verde e b para o canal azul [Swain and Ballard 1991]. A técnica utilizada neste projeto segue a fórmula geral para o cálculo de tais histogramas,

dada por (2).

$$H(c) = \sum_{i=1}^{N} \delta(I_i, c) \tag{2}$$

Em (2), c é a cor especificada, N é o total de pixels na imagem, I_i é o valor de cor do pixel i, e $\delta(I_i, c)$ é uma função indicadora que retorna 1 quando o valor do pixel coincide com c, e 0 caso contrário. Para capturar mais nuances nas imagens, os histogramas foram computados separadamente para os três canais de cor e agrupados em uma estrutura tridimensional, com 8 intervalos (*bins*) para cada canal. Esse número de intervalos permite representar as cores com detalhes, mas evita uma dimensionalidade excessiva.

As imagens foram convertidas para o espaço de cor BGR e os histogramas de cor gerados foram padronizados e transformados em vetores unidimensionais, facilitando a integração com o modelo de classificação. Assim, a representação final dos dados consistiu em vetores com 512 atributos (8 intervalos para cada um dos três canais, resultando em $8 \times 8 \times 8 = 512$ características por imagem).

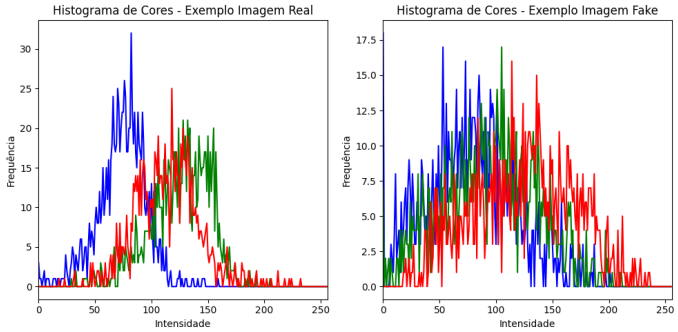

Figura 1. Histogramas de cores de imagens real e *fake*

Na Figura 1, tem-se dois exemplos para as distribuições de cores resultantes para duas imagens do conjunto de teste, uma verdadeira e outra sintética. Observa-se, para estes exemplares, que o histograma possui caudas leves e intensidades médias diferentes para todas as cores na figura real, também possuindo uma faixa maior de valores de frequências, com cada cor assumindo frequências relativamente distintas em suas distribuições. Por outro lado, a figura falsa tem caudas um pouco mais pesadas e os histogramas têm formatos muitos similares entre si, com medianas de intensidade quase iguais.

2.2. Descritores de Texturas de Haralick

Os descritores de texturas baseados em estatísticas de Haralick são amplamente utilizados para capturar padrões e distribuições de intensidade em imagens, oferecendo uma maneira de caracterizar as texturas presentes. Desenvolvidos por Haralick et al. [Haralick et al. 1973], esses descritores são derivados de uma matriz de coocorrência de níveis de cinza (GLCM, do inglês *Gray Level Co-occurrence Matrix*), que representa a frequência relativa de pares de pixels com intensidades específicas a uma certa distância e direção. A partir da GLCM, extraem-se diversas características estatísticas que resumem

as relações de textura, incluindo contraste, correlação, variância e homogeneidade, que refletem a uniformidade, repetitividade e complexidade da textura na imagem.

Para a sua implementação neste projeto, as imagens foram primeiro convertidas para escala de cinza, de forma a gerar a GLCM. Em seguida, calcularam-se as características de Haralick com cada imagem sendo representada por 13 descritores, como média dos valores calculados ao longo das direções de análise (0°, 45°, 90°e 135°). Posteriormente, a representação vetorial resultante foi padronizada para ter média $\mu = 0$ e variância $\sigma^2 = 1$.

2.3. *Residual Networks*

As *Residual Networks* (ou *ResNets*) são uma arquitetura de rede neural profunda projetada para resolver problemas comuns ao treinamento de redes muito profundas, como o desaparecimento e a explosão de gradientes, muitas vezes resultando em perda de informações [He et al. 2016].

O conceito central das *ResNets* é o uso de blocos residuais, que introduzem "atalhos" ou conexões de salto (*skip connections*) entre as camadas da rede. Em um bloco residual típico, a entrada original é combinada diretamente com a saída da camada, permitindo que a rede aprenda uma função residual, que é mais fácil de ajustar do que a função original. Formalmente, a ideia básica pode ser representada como:

$$\mathbf{y} = F(\mathbf{x}, W_i) + \mathbf{x} \tag{3}$$

onde $\mathbf{x}$ é a entrada do bloco residual, W_i representa os pesos da camada, e $F(\mathbf{x}, W_i)$ é a função residual aprendida pela camada intermediária. Essa operação permite que a rede aprenda uma transformação $F(\mathbf{x})$, preservando a informação original da entrada $\mathbf{x}$.

No projeto, foi utilizada uma *ResNet-18*, versão compacta composta por 18 camadas de convolução. As principais camadas nessa versão incluem camadas de convolução para captura de características locais; camadas de normalização para estabilizar o treinamento; camadas de pooling para reduzição de dimensionalidade e blocos residuais que permitem o aprendizado das transformações complexas das imagens.

A *ResNet-18* foi pré-treinada no ImageNet [Deng et al. 2009], o que significa que ela aprendeu um grande conjunto de características visuais. Para este projeto, foi utilizada a *embedding* extraída da penúltima camada, a camada de *pooling* global média, que gera uma representação de 512 dimensões para cada imagem, contendo informações visuais aprofundadas que representam nuances como texturas, bordas e variações de cor. Assim como os vetores das outras representações, tais dados foram padronizados para média $\mu = 0$ e variância $\sigma^2 = 1$.

3. Extração de Padrões

Nesta seção, descrevem-se as técnicas aplicadas para extrair padrões das imagens com base em suas representações vetoriais, obtidas a partir dos descritores de cor, textura e *embeddings* de redes neurais convolucionais. A técnica de classificação empregada foi o *k*-Nearest Neighbors (KNN), escolhida por sua simplicidade e eficácia em problemas

onde a proximidade no espaço vetorial representa a similaridade entre amostras. No contexto deste trabalho, o KNN é particularmente vantajoso, pois permite avaliar a similaridade entre as representações vetoriais geradas para cada imagem, dispensando suposições de distribuição específica para os dados.

O *k*-Nearest Neighbors é um classificador baseado em instâncias, onde a previsão para uma nova amostra é realizada a partir dos rótulos das amostras de treinamento mais próximas. No caso deste trabalho, as distâncias entre os vetores de características foram interpretadas como medidas de similaridade entre imagens, possibilitando o agrupamento de imagens visualmente semelhantes.

Para otimizar o desempenho do KNN, procedeu-se ao ajuste de hiperparâmetros. Realizou-se uma busca aleatória sobre a quantidade de vizinhos (k) no intervalo $[1, 20]$, assim como sobre diferentes métricas de distância. Foram testadas as distâncias euclidiana, Manhattan e a similaridade do cosseno, escolhidas para refletir diferentes aspectos das representações vetoriais: enquanto as distâncias euclidiana e Manhattan medem a dissimilaridade geométrica entre vetores, a similaridade do cosseno mede o ângulo entre eles, o que se mostrou útil para *embeddings* onde a orientação no espaço vetorial importa mais do que a magnitude. Esses parâmetros são apresentados de forma resumida na Tabela 3.

Tabela 1. Melhores parâmetros para KNN em diferentes representações

Representação	**Métrica**	***k***
Histograma de Cores	Manhattan	8
Haralick	Manhattan	20
ResNet-18	Euclidiana	13

4. Pós-processamento

Para avaliar os resultados obtidos na classificação das imagens, foram calculadas as métricas de desempenho dos modelos aplicados a cada representação vetorial. As métricas analisadas incluíram acurácia, precisão, recall, F1-score e AUC-ROC. Cada uma dessas métricas forneceu uma perspectiva complementar sobre o desempenho do modelo: acurácia mede a proporção geral de classificações corretas, precisão avalia o nível de exatidão entre as previsões positivas, recall mensura a capacidade de identificar corretamente as instâncias positivas, enquanto o F1-score pondera precisão e recall em uma única métrica harmonizada. O AUC-ROC, por outro lado, quantifica a capacidade do modelo em discriminar entre as classes de forma robusta, especialmente útil para avaliar os métodos mais sofisticados aplicados ao contexto de imagens.

Para garantir que os resultados fossem consistentes e representativos, foi realizada uma validação cruzada em cinco subdivisões para cada técnica de representação, calculando a média de desempenho para cada abordagem. A Tabela 2 apresenta os valores das métricas mencionadas encontradas para os modelos a partir do conjunto de teste.

Os resultados iniciais mostram que o modelo baseado em *ResNet* apresentou o melhor desempenho global, especialmente no que se refere à AUC-ROC, que alcançou $0,9738$, destacando-se na discriminação entre classes. A técnica de descritores de textura Haralick, embora com desempenho inferior à *ResNet*, mostrou-se superior à representação

Tabela 2. Métricas de Desempenho dos Modelos

Representação	Acurácia	Precisão	Recall	F1-Score	AUC-ROC	CV
Cores	$0,7683$	$0,7874$	$0,7351$	$0,7603$	$0,8514$	$0,7307$
Haralick	$0,8262$	$0,8197$	$0,8363$	$0,8279$	$0,9043$	$0,8029$
ResNet-18	$0,9151$	$0,9205$	$0,9087$	$0,9146$	$0,9738$	$0,8872$

baseada em cores, sugerindo uma captura mais eficaz dos padrões visuais relevantes para a tarefa. O modelo de cores, embora adequado para diferenciações de base, apresentou uma taxa de acurácia de 0.7683, com pontuações inferiores nas demais métricas, refletindo uma capacidade reduzida de discriminação em comparação com as abordagens de textura e rede neural.

Para avaliar se essas diferenças são estatisticamente significativas, realizou-se o teste de Wilcoxon para as comparações de desempenho entre as representações, conforme descrito nas hipóteses estabelecidas em 1 com um nível de significância $\alpha = 0,05$. Esse teste não-paramétrico é apropriado para verificar se há uma diferença significativa entre duas amostras pareadas de desempenho sob diferentes condições de modelagem. A Tabela 3 apresenta os níveis descritivos obtidos com os testes entre as combinações de modelos.

Tabela 3. Resultados do Teste de Wilcoxon entre Representações

Diferença entre métricas - D	**Valor-*p*** - $Pr(D \geq d)$
Haralick $-$ Cores	$0,03125$
ResNet-18 $-$ Cores	$0,03125$
ResNet-18 $-$ Haralick	$0,03125$

As comparações indicam que as diferenças entre todas as representações são significativas ao nível $\alpha = 0,05$, especialmente em relação ao desempenho da *ResNet-18*, que se destaca em relação às outras duas técnicas. Esse resultado sugere que a abordagem baseada em redes neurais é significativamente mais eficaz para a tarefa de classificação de imagens neste conjunto de dados, possivelmente pela sua capacidade de capturar nuances e estruturas visuais mais complexas, que não são detectadas com a mesma precisão pelos métodos baseados em cores ou textura.

Com exceção do desempenho do modelo baseado no histograma de cores, os resultados estão conforme o que se esperava inicialmente do estudo, com todas as hipóteses nulas sendo rejeitadas e reiterando a superioridade das representações mais complexas. Para refinar tais resultados, pode-se ajustar os parâmetros de acordo com as particularidades dos dados, com uma maior faixa de possibilidade e exploração de outras métricas de análise, de modo a superar os objetivos estabelecidos.

5. Uso do Conhecimento

A capacidade de distinguir entre imagens reais e sintéticas é crucial para garantir a autenticidade das informações visuais em uma série de contextos sensíveis. A análise aprofundada de representações vetoriais desses dados, como histogramas de cores, descritores de textura e *embeddings* de redes neurais profundas, não apenas permite que mo-

delos automatizados de detecção de imagens sintéticas sejam mais precisos, mas também traz à tona padrões e peculiaridades das imagens geradas artificialmente.

O conhecimento obtido por meio dessa análise põe à disposição o desenvolvimento de técnicas de detecção mais robustas e a implementação de políticas de segurança digital. Ao compreender melhor as diferenças estruturais entre imagens reais e sintéticas evidenciadas em cada representação é possível criar sistemas de detecção que discriminem com eficácia entre esses tipos de imagens, bem como se adaptem às inovações e complexidades das técnicas de geração de imagens artificiais. Além disso, a aplicação de métodos de comparação de representações visuais contribui para o aprimoramento da confiabilidade em áreas que utilizam intensivamente imagens como base para tomada de decisões, fortalecendo a defesa contra a disseminação de informações visuais enganosas.

Para melhorias futuras, seria essencial investigar os padrões nos erros recorrentes de cada técnica de classificação, especialmente as imagens mal categorizadas por todos os modelos. Uma análise detalhada desses casos poderia revelar limitações específicas na detecção de certos elementos visuais, inspirando o desenvolvimento de técnicas de pré-processamento e de representação mais robustas para este contexto.

Referências

Bird, J. J. and Lotfi, A. (2024). Cifake: Image classification and explainable identification of ai-generated synthetic images. *IEEE Access*.

Corvi, R., Cozzolino, D., Zingarini, G., Poggi, G., Nagano, K., and Verdoliva, L. (2023). On the detection of synthetic images generated by diffusion models. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Haralick, R. M., Shanmugam, K., and Dinstein, I. H. (1973). Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics*, (6):610–621.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Krizhevsky, A., Hinton, G., et al. (2009). Learning multiple layers of features from tiny images.

Pataranutaporn, P., Danry, V., Leong, J., Punpongsanon, P., Novy, D., Maes, P., and Sra, M. (2021). Ai-generated characters for supporting personalized learning and well-being. *Nature Machine Intelligence*, 3(12):1013–1022.

Seow, J. W., Lim, M. K., Phan, R. C., and Liu, J. K. (2022). A comprehensive overview of deepfake: Generation, detection, datasets, and opportunities. *Neurocomputing*, 513:351–371.

Swain, M. J. and Ballard, D. H. (1991). Color indexing. *International journal of computer vision*, 7(1):11–32.