Detecção de Anomalias Visuais com Autoencoder Convolucional e SSIM no Dataset MVTec AD

Vinicius Pereira Tavares de Sousa
Universidade Tecnológica Federal do Paraná
viniciustavares.sousa@gmail.com

***Abstract*—This paper presents an approach for visual anomaly detection based on a Convolutional Autoencoder with a hybrid loss function combining mean squared error and Structural Similarity Index (SSIM). Applied to the *bottle* category of the MVTec AD dataset, the model demonstrates high sensitivity in identifying and localizing defects in industrial images. Qualitative and quantitative results confirm the effectiveness of the methodology for automated inspection.**

***Index Terms*—Detecção de anomalias, autoencoder convolucional, SSIM, visão computacional, inspeção industrial, MVTec AD.**

I. Introdução

A inspeção visual automática tem papel fundamental no controle de qualidade industrial, sendo responsável por detectar e classificar defeitos em produtos manufaturados. Este processo tradicionalmente é realizado manualmente por operadores treinados, o que pode gerar custos elevados e inconsistências devido à fadiga e subjetividade. Com o avanço da inteligência artificial e do aprendizado profundo, métodos automáticos têm ganhado destaque por sua capacidade de análise precisa e em alta velocidade.

O desafio central na detecção de anomalias está na raridade e diversidade dos defeitos, que tornam inviável a criação de conjuntos balanceados para treinamento supervisionado. Assim, métodos baseados em aprendizado não supervisionado, como autoencoders, têm sido amplamente estudados. Esses modelos aprendem uma representação compacta dos dados normais e detectam anomalias quando uma entrada apresenta alto erro de reconstrução.

Este trabalho propõe o uso de um autoencoder convolucional com função de perda híbrida combinando erro quadrático médio (MSE) e índice estrutural de similaridade (SSIM). Essa combinação visa preservar detalhes estruturais importantes durante a reconstrução, melhorando a sensibilidade na identificação e localização dos defeitos. A aplicação foi realizada na categoria *bottle* do dataset MVTec AD, um benchmark amplamente utilizado para avaliação de métodos de detecção de anomalias visuais.

II. Trabalhos Relacionados

A literatura sobre detecção de anomalias visuais é vasta e crescente. Modelos baseados em autoencoders simples têm sido usados com sucesso, explorando a capacidade de compressão e reconstrução dos dados normais [1], [2]. Porém, a função de perda padrão, geralmente MSE, pode falhar em capturar diferenças perceptuais importantes, especialmente em contextos industriais onde pequenas alterações estruturais são cruciais [3].

Para contornar essas limitações, trabalhos recentes têm incorporado métricas perceptuais na função de perda, como o SSIM, que avalia similaridade em termos de luminância, contraste e estrutura [4], alinhando-se melhor com a percepção humana. Além disso, abordagens adversariais combinam autoencoders com GANs para produzir reconstruções mais realistas, embora com maior complexidade computacional e dificuldade de treinamento [5].

Neste cenário, a proposta deste trabalho é uma solução intermediária, combinando MSE e SSIM de forma simples e eficaz, oferecendo uma boa qualidade de reconstrução sem elevar a complexidade do modelo.

III. Metodologia

A. Arquitetura do Autoencoder

O modelo proposto é dividido em duas partes principais: encoder e decoder. O encoder é responsável por extrair uma

representação latente compacta da imagem de entrada, enquanto o decoder reconstrói a imagem original a partir dessa representação.

A entrada do modelo são imagens monocromáticas redimensionadas para 256×256 pixels, com um canal único. O encoder possui sete camadas convolucionais sequenciais com filtros de tamanho 5×5. As primeiras seis camadas utilizam stride 2 para redução progressiva da dimensão espacial, enquanto a última mantém stride 1 para preservar informações. Cada camada é seguida de ativação LeakyReLU, que ajuda a evitar o problema do "morto" de neurônios comuns em ReLU, e normalização por lotes para acelerar o treinamento e estabilizar a rede.

O vetor latente obtido possui dimensão 128, permitindo um balanceamento entre capacidade expressiva e compactação dos dados.

O decoder realiza a reconstrução da imagem por meio de camadas transpostas convolucionais (Conv2DTranspose), que realizam upsampling espacial. A arquitetura espelha o encoder para facilitar a reconstrução. A última camada utiliza ativação sigmoidal para normalizar a saída entre 0 e 1, compatível com a escala das imagens de entrada.

TABLE I: Arquitetura detalhada do encoder

Camada	Filtros	Saída (H, W, C)	Observações
Entrada	-	(256, 256, 1)	Imagem original
Conv2D + BN	8	(128, 128, 8)	Stride 2
Conv2D + BN	16	(64, 64, 16)	Stride 2
Conv2D + BN	32	(32, 32, 32)	Stride 2
Conv2D + BN	64	(16, 16, 64)	Stride 2
Conv2D + BN	128	(8, 8, 128)	Stride 2
Conv2D + BN	256	(4, 4, 256)	Stride 2
Conv2D + BN	512	(4, 4, 512)	Stride 1
Flatten	-	(8192,)	-
Dense	128	(128,)	Vetor latente

TABLE II: Arquitetura detalhada do decoder

Camada	Filtros	Saída (H, W, C)	Observações
Entrada	-	(128,)	Vetor latente
Dense + BN	-	(4, 4, 512)	Reshape
Conv2DTranspose + BN	256	(8, 8, 256)	Stride 2
Conv2DTranspose + BN	128	(16, 16, 128)	Stride 2
Conv2DTranspose + BN	64	(32, 32, 64)	Stride 2
Conv2DTranspose + BN	32	(64, 64, 32)	Stride 2
Conv2DTranspose + BN	16	(128, 128, 16)	Stride 2
Conv2DTranspose + BN	8	(128, 128, 8)	Stride 1
Conv2DTranspose (sigmoid)	1	(256, 256, 1)	Stride 2

B. Função de Perda Híbrida

A função de perda usada combina dois termos importantes:

- **Erro Quadrático Médio (MSE):** Mede a diferença pontual média entre a imagem original e a reconstruída, incentivando fidelidade pixel a pixel.
- **Índice Estrutural de Similaridade (SSIM):** Avalia a similaridade considerando luminância, contraste e estrutura, promovendo preservação dos detalhes estruturais e texturas, essenciais para detectar pequenas anomalias.

A perda é dada por:

$$\mathcal{L}(x, \hat{x}) = 0.5 \times \text{MSE}(x, \hat{x}) + 0.5 \times (1 - \text{SSIM}(x, \hat{x}))$$

Este balanço foi escolhido empiricamente para maximizar a qualidade perceptual das reconstruções, fundamental para o sucesso na detecção de defeitos.

C. Pré-processamento dos Dados

Todas as imagens foram convertidas para escala de cinza e redimensionadas para 256×256 pixels, permitindo padronização para o modelo. A normalização foi feita para intervalo $[0, 1]$, melhorando estabilidade numérica no treinamento.

Durante o treinamento, apenas imagens sem defeitos foram usadas para que o modelo aprendesse o padrão normal. Para avaliação, imagens defeituosas foram processadas para medir a capacidade do modelo em identificar anomalias.

D. Configurações de Treinamento

O treinamento foi realizado utilizando o otimizador Adam com taxa de aprendizado 0.002, batch size de 8, durante 500 épocas. O uso de batch normalization em todas as camadas convolucionais auxilia na aceleração do treinamento e evita overfitting.

A função de perda híbrida foi monitorada para garantir convergência, e o modelo final foi salvo para uso em testes e análise.

IV. Experimentos

A. Dataset

O dataset MVTec AD é um benchmark público para detecção de anomalias visuais, contendo imagens reais de defeitos industriais. A categoria *bottle* possui 209 imagens normais usadas para treinamento. Para teste, foram utilizadas apenas as 19 imagens da subcategoria *broken_large*, que apresentam defeitos evidentes como quebras grandes, além das imagens normais do conjunto de treinamento para comparação.

B. Avaliação Quantitativa

Para avaliar o desempenho, utilizamos o índice SSIM médio entre imagens originais e reconstruídas, comparando imagens normais e defeituosas.

Na Tabela III, observa-se que as imagens com defeito apresentam SSIM significativamente menor, indicando que o modelo captura adequadamente as anomalias.

TABLE III: Média e desvio padrão do SSIM nas imagens de teste

Categoria	Média SSIM	Desvio Padrão
Imagens normais	0.9583	0.0029
Imagens com defeito	0.8315	0.0302

C. Visualização dos Resultados

A Figura 1 exemplifica imagens com e sem defeito, suas reconstruções e os mapas de erro, evidenciando a capacidade do modelo em localizar anomalias apenas nas imagens defeituosas.

V. Discussão

Os resultados demonstram que a combinação da perda MSE e SSIM promove reconstruções que preservam detalhes estruturais e texturais relevantes para a identificação das anomalias. O mapa de erro gerado facilita não só a detecção mas também a localização dos defeitos, potencializando aplicações práticas em inspeção automatizada.

Limitações do modelo incluem menor sensibilidade a defeitos muito sutis e a dependência da qualidade das imagens de entrada. Ruídos, variações de iluminação e diferentes condições de aquisição podem afetar o desempenho, o que exige cuidados no pré-processamento.

VI. Análise de Robustez e Limitações

Embora o modelo apresente desempenho robusto na detecção de defeitos visuais evidentes, sua capacidade de identificar anomalias muito sutis, como pequenas manchas ou fissuras pouco contrastadas, pode ser limitada. Isso ocorre porque tais defeitos podem gerar erro de reconstrução baixo, dificultando a distinção entre normalidade e anomalia.

Além disso, a abordagem atual depende fortemente da qualidade do pré-processamento e padronização das imagens. Condições adversas de iluminação, ruído e variações de pose podem comprometer a qualidade da reconstrução, levando a falsos positivos ou negativos.

Outro ponto crítico é o custo computacional. A arquitetura convolucional profunda e as operações de SSIM implicam em tempo considerável de treinamento e inferência, o que pode ser um desafio para sistemas de inspeção em tempo real. Investigações futuras podem explorar otimizações e arquiteturas mais eficientes.

VII. Perspectivas Futuras

Para ampliar a aplicabilidade do método, propõe-se a integração com técnicas de atenção espacial, que destacam regiões relevantes da imagem, potencializando a detecção de defeitos localizados. Além disso, o uso de redes adversariais (GANs) para melhorar a qualidade da reconstrução pode aumentar a sensibilidade a anomalias sutis.

Explorar aprendizado multimodal, combinando dados visuais com informações de sensores industriais, também pode enriquecer o diagnóstico.

Repositório e Código

Todo o código e os experimentos descritos neste artigo estão disponíveis no GitHub:
https://github.com/ViniciusTavaresSousa/Deteccao-de-Anomalias-Visuais-com-Autoencoder-Convolucional-e-SSIM-no-Dataset-MVTec-AD

Agradecimentos

Agradeço à equipe do dataset MVTec AD e à Universidade Tecnológica Federal do Paraná pelo suporte.

References

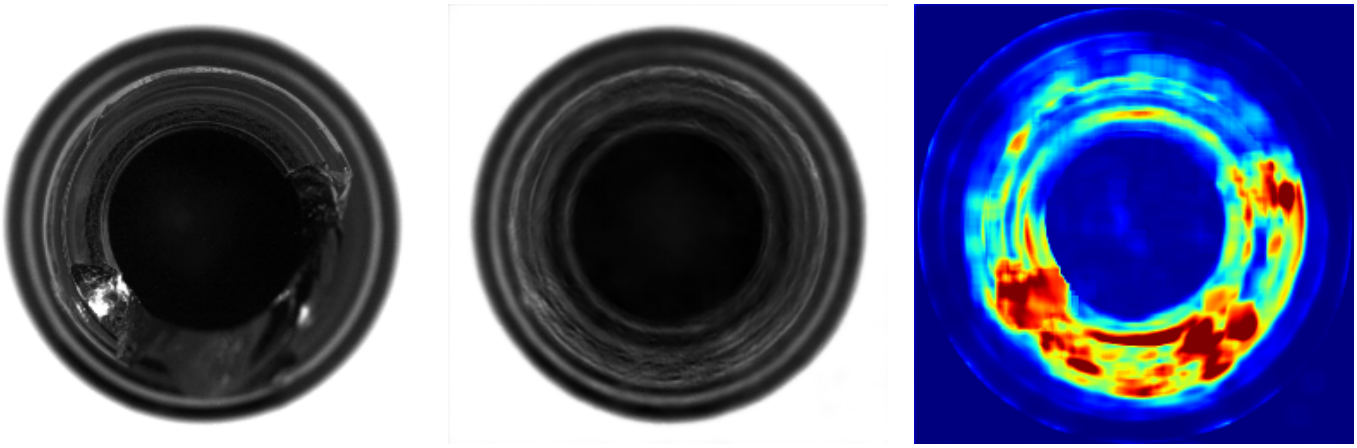
[1] M. Sakurada and T. Yairi, "Anomaly detection using autoencoders with nonlinear dimensionality reduction," *Proceedings of the MLSDA*, 2014.

[2] Y. Xia, W. Liu, and W. Wang, "Learning discriminative reconstructions for unsupervised outlier removal," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 3519–3531, 2015.

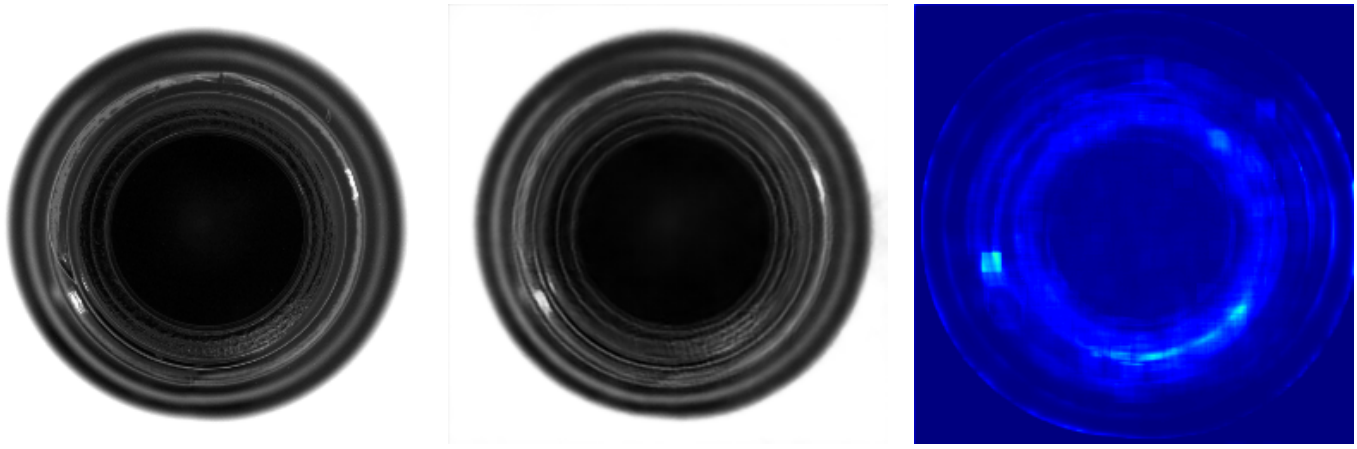
[3] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Loss functions for image restoration with neural networks," *IEEE Transactions on Computational Imaging*, vol. 3, no. 1, pp. 47–57, 2017.

[4] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[5] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs, "Unsupervised anomaly detection with generative adversarial networks to guide marker discovery," in *Information Processing in Medical Imaging (IPMI)*, 2017.



(a) Imagem com defeito (b) Reconstrução da imagem com defeito (c) Mapa de erro (1-SSIM) com defeito



(d) Imagem sem defeito (e) Reconstrução da imagem sem defeito (f) Mapa de erro (1-SSIM) sem defeito

Fig. 1: Visualização comparativa das imagens com defeito (primeira linha) e sem defeito (segunda linha), suas reconstruções e mapas de erro obtidos pelo modelo. Nota-se que o mapa de erro destaca claramente as anomalias apenas na imagem com defeito.