Benchmark of XAI Methods for Image Classification and Extension of Optimization-Based Explanations to CLIP Models

Rheidner A. C. S. Fernandes [1], **Leonardo N. Matos**[1], **André B. Carvalho**[1], **Marcelo H. L. Barreto**[2], **Flávio Arthur O. Santos**[2]

[1]Departamento de Computação – Universidade Federal de Sergipe (UFS)
São Cristóvão, SE – Brasil

[2]Centro de Informática – Universidade Federal de Pernambuco (UFPE)
Recife, PE – Brasil

{rheidner.achiles,andre,leonardo}@dcomp.ufs.br,{mhlb,faos}@cin.ufpe.br

Abstract. *Explainable Artificial Intelligence (XAI) techniques are used to understand the decisions of black-box models. In image classification, gradient-based and optimization-based methods aim to identify the regions most relevant to model predictions. This work presents a quantitative benchmark of Saliency Maps, Integrated Gradients, Guided Backpropagation, Grad-CAM, Guided Grad-CAM, and IPHA-based methods. Additionally, an extension of optimization-based methods to multimodal models based on the CLIP architecture is proposed and evaluated qualitatively in an exploratory manner. The results indicate that optimization methods using continuous regions, particularly those based on inpainting, achieve superior performance while also demonstrating their flexibility through application to the CLIP model.*

Resumo. *Técnicas de Inteligência Artificial Explicável (XAI) são utilizadas para compreender as decisões de modelos black-box. Na classificação de imagens, métodos baseados em gradientes e otimização buscam identificar as regiões mais relevantes para as predições dos modelos. Neste trabalho, é realizado um benchmark quantitativo envolvendo Saliency Maps, Integrated Gradients, Guided Backpropagation, Grad-CAM, Guided Grad-CAM e variantes da família IPHA. Também é proposta uma extensão dos métodos de otimização para modelos multimodais baseados na arquitetura CLIP, avaliada qualitativamente de forma exploratória. Os resultados indicam vantagem para métodos por otimização que utilizam regiões contínuas, especialmente aqueles baseados em inpainting, além de demonstrar a flexibilidade desses métodos por meio de sua aplicação ao modelo CLIP.*

1. Introdução

Modelos de inteligência artificial vêm sendo cada vez mais utilizados em diversas áreas, como saúde, finanças e segurança, com cada vez mais decisões importantes sendo tomadas com base em suas previsões. A área de classificação de imagens é uma das aplicações fundamentais para visão computacional, na qual os modelos são treinados para identificar e classificar objetos em imagens [Chen et al. 2021]. No entanto, a complexidade desses

modelos pode dificultar a compreensão de como eles tomam suas decisões pois, em sua maioria, são considerados caixas-pretas e possuem milhões de parâmetros, tornando-os ininteligíveis para humanos [Hassija et al. 2024].

Nesse contexto, técnicas de Inteligência Artificial Explicável (XAI, do inglês *explainable aritificial intelligence*) buscam tornar os modelos mais transparentes e compreensíveis, permitindo identificar quais informações influenciam suas decisões [Lin et al. 2021, Hassija et al. 2024]. Em classificação de imagens, técnicas de XAI são utilizadas para identificar quais áreas da imagem são mais relevantes para a decisão do modelo, algo útil para entender o seu comportamento. Os métodos baseados em gradientes estão entre os mais difundidos, dado que grande parte dos modelos de classificação de imagens são baseados em redes neurais convolucionais, que são modelos diferenciáveis, como Grad-CAM [Selvaraju et al. 2017] e Integrated Gradients [Sundararajan et al. 2017]. No entanto, outras técnicas, que tem como base a otimização de uma função de perda, surgiram como alternativas à análise dos gradientes, apresentando resultados promissores na geração de mapas de saliência [Santos et al. 2021, Barreto et al. 2025, Fernandes et al. 2026].

Apesar dos avanços, a comparação sistemática entre as diversas técnicas de XAI ainda representa um desafio devido à diversidade de métodos, métricas, conjuntos de dados e arquiteturas utilizados. Com isso, este trabalho tem como objetivo a realização de uma avaliação quantitativa entre diferentes métodos de explicabilidade baseados em gradientes e em otimização para classificação de imagens, avaliando sua robsutez e generalização em diferentes cenários. Para isso, é utilizado uma métrica padrão Feature Impact Index (FII) [Santos et al. 2021], além dos diferentes tipos de conjuntos de dados ImageNette [Deng et al. 2009], ImageWoof [Deng et al. 2009] e Stanford Cars [Krause et al. 2013] e arquiteturas Resnet-18 [He et al. 2016], SqueezeNet [Iandola et al. 2016] e VGG-16-BN [Simonyan and Zisserman 2014]. Além disso, este trabalho também propõe uma extensão dos métodos baseados em otimização para modelos multimodais VLM basaeados na arquitetura CLIP, investigando de forma exploratória a viabilidade da aplicação dessas técnicas em tarefas de alinhamento de imagens com descrições textuais dinâmicas.

2. Fundamentação Teórica

Este trabalho avalia métodos de explicabilidade para classificação, que podem ser divididos aqui em duas categorias: baseados em gradiente ou otimização. Embora ambos tenham o objetivo de identificar quais as partes da imagem mais influenciam a decisão de um modelo, suas diferenças são significativas quanto aos seus fundamentos teóricos e forma de geração das explicações. Exemplos de explicações produzidas pelos métodos avaliados são apresentados na Figura 1.

2.1. Métodos baseados em gradientes

Métodos baseados em gradientes utilizam informações obtidas diretamente do processo de diferenciação da rede neural, que mede a taxa de variação da classificação obtida em relação a pequenas alterações nos atributos (pixels) da imagem. Dessa forma, os gradientes são interpretados como uma estimativa da sensibilidade local do modelo. Percebe-se também que estes são métodos *white-box*, por necessitarem de acesso a mecanismos e

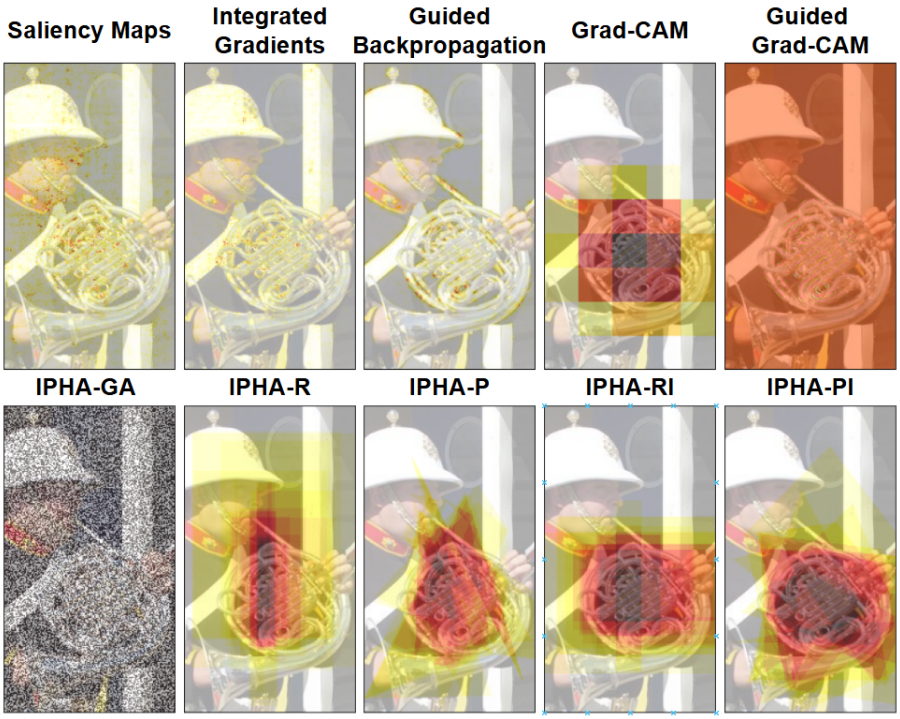


Figura 1. Mapas de calor para explicações geradas para classe *trompa*.

valores internos à arquitetura do modelo. Neste trabalho serão avaliados os métodos Saliency Map, Integrated Gradients, Guided Backpropagation, Grad-CAM e Guided Grad-CAM.

A abordagem Saliency Map [Simonyan et al. 2013], considerada uma das mais simples, calcula a importância de cada pixel a partir do módulo do gradiente da saída em relação à imagem de entrada. Pixels que são associados a gradientes de maior valor são considerados relevantes para a decisão do modelo. Apesar de sua simplicidade, o método produz mapas sensíveis a pequenas pertubações da imagem. Integrated Gradients [Sundararajan et al. 2017], por sua vez, utiliza além do gradiente local, uma entrada de referência, que servirá para integrar os gradientes em um caminho que irá desta referência até a imagem analisada. No contexto de classificação de imagens, a entrada de referência geralmente é definida como uma imagem completamente preta, que busca representar a ausência de informações.

O método de explicabilidade Guided Backpropagation [Springenberg et al. 2014], modifica o processo tradicional de retropropagação das redes neurais. Durante o cálculo dos gradientes, apenas ativações e gradientes positivos são propagados através das funções RELU, o que suprime contribuições negativas. Essa modificação reduz o ruído e realça evidências positivas à decisão do modelo. O Gradient-weighted Class Activation Mapping, ou Grad-CAM [Selvaraju et al. 2017], por outro lado, é uma técnica de explicabilidade desenvolvida para redes neurais convolucionais que utiliza os gradientes da camada de saída da classe de interesse em relação a camada convolucional escolhida, geralmente a última disponível. Os mapas gerados por essa explicação possuem resolução inferior à da imagem original e são interpolados para o tamanho da imagem original para que se tenha a máscara de saliência final. Por fim, o Guided Grad-CAM [Selvaraju et al. 2017]

combina as atribuições produzidas pelo Guided Backpropagation com a máscara final de saliência gerada pelo Grad-CAM. Essa combinação é feita multiplicando-se elemento a elemento entre os dois resultados, com o objetivo de unir os benefícios dos dois métodos.

2.2. Métodos baseados em otimização

Os métodos baseados em otimização formulam a geração da explicação como um problema de busca. Como dependem apenas das saídas produzidas pelo modelo, podem ser considerados *black-box*. Neste trabalho são avaliados cinco variações desta abordagem: IPHA-GA, IPHA-R, IPHA-RI, IPHA-P e IPHA-PI. Todos estas derivam da técnica IPHA [Santos et al. 2021], que busca determinar uma máscara que otimize sua avaliação da função objetivo $f(x)$, definida a partir da probabilidade atribuída pelo modelo à classe originalmente predita. Durante a otimização, pixels marcados pelas máscaras são ocultados, e a redução resultante na confiança do classificador é utilizada como medida de relevância. Assim, o trabalho argumenta que os pixels cuja remoção provoca maior degradação da predição são mais importantes para a decisão do modelo. O método IPHA-GA [Barreto et al. 2025] funciona do mesmo modo que o IPHA, buscando máscaras pixel a pixel, formadas por $m \times n$ variáveis binárias, diferenciando-se na utilização de um algoritmo genético próprio em substituição do algoritmo de *hill climbing*.

Os métodos IPHA-R e IPHA-P [Fernandes et al. 2026] modificam a representação pixel por pixel do IPHA por regiões contínuas, descritas por retângulos (2 pontos, 4 variáveis) ou polígonos ($n \geq 3$ pontos, $2n$ variáveis), respectivamente. Além disso, é introduzida uma segunda função objetivo que busca minimizar a área da região de máscara, o que torna o problema multiobjetivo. Como resultado, o processo produz um conjunto de soluções ótimas, e as máscaras deste conjunto são sobrepostas para geração de um mapa de atribuição para cada pixel, gerando assim a máscara de explicação final. Por último, os métodos IPHA-RI e IPHA-PI [Fernandes et al. 2026] seguem a mesma formulação dos métodos anteriores, mas aplicam a técnica de *inpaint* ao invés de utilizar uma máscara preta na ocultação dos pixels durante a otimização, visando preencher as regiões removidas das imagens utilizando informações de seu contexto.

3. Formulação dos Experimentos

Para a realização dos experimentos de comparação, os algoritmos de interpretabilidade apresentados foram implementados para explicar as decisões de três arquiteturas de modelos de classificação de imagens, ResNet, SqueezeNet e VGG-16-BN, em três conjuntos de dados, ImageNette, ImageWoof e Stanford Cars. Os experimentos foram implementados em Python com uso das bibliotecas FastAI [Howard and Gugger 2020], PyTorch [Paszke et al. 2019], PyMOO [Blank and Deb 2020] e Captum [Naritomy et al. 2020]. O código fonte está disponibilizado publicamente em `https://github.com/racs4/framework-xai-optimization`. Todos experimentos foram executados em um computador com processador Intel i9-14900K, 64 GB de RAM e GPU NVIDIA GeForce RTX 5090 32GB.

3.1. Parametrização

Os métodos por otimização foram implementados utilizando-se o algoritmo genético NSGA-II [Deb et al. 2002] com uma população de 200 indivíduos e 10 gerações. Foram utilizados o crossover de simulação binária (SBX) e mutação polinomial, ambos com

probabilidade 1.0 para alta exploração do espaço de busca. O algoritmo foi inicializado com uma população de indivíduos gerados aleatoriamente. Deve-se notar que esta é uma diferença em relação a implementação do IPHA-GA, que em seu trabalho original foi desenvolvido com um algoritmo genético próprio e genérico.

Para os métodos de otimização baseados em polígonos, o número de vértices do polígono foi definido como 5; todos os algoritmos de otimização que não utilizaram *inpaint* foram utilizados com máscara de cor preta; os que utilizaram a técnica de *inpaint* foram implementados através do método de Telea [Telea 2004], com raio de contexto igual a 3; o *threshold* para os algoritmos de otimização que utilizaram uma função multiobjetivo foi definido como 0.1, o que significa que, para um pixel ser considerado na máscara final, ele deve aparecer em pelo menos 10% das soluções encontradas. Devido à alta resolução dos conjuntos avaliados, o IPHA-GA foi adaptado para utilizar uma representação baseada em *packed bitstrings*. Nesse esquema, grupos de 16 pixels são representados por uma única variável inteira de 16 bits, reduzindo a dimensionalidade do problema sem alterar o formato final das máscaras.

Com relação às técnicas de gradientes, as mesmas foram implementadas utilizando os parâmetros padrões de cada técnica. As que exigiam a escolha de uma camada do modelo, como é o caso do Grad-CAM, foram implementadas utilizando a última camada convolucional de cada modelo, dado que esta é a camada mais próxima da saída do modelo e, portanto, a que mais influencia a decisão final e a que mais retém informações relevantes sobre a imagem e a decisão do modelo. O resultado da aplicação destas técnicas é uma máscara de relevância, com pesos que variam entre 0 e 1. Para escolha do limiar e formulação das imagens explicativas a partir destas máscaras, escolheu-se uma porcentagem p de pixels mais relevantes, que foi definido como a média do número de pixels relevantes encontrados pelas técnicas de otimização, visando manter uma comparação justa entre os dois tipos de técnicas.

3.2. Conjuntos de dados e métrica

Foram utilizados os conjuntos de dados ImageNette e ImageWoof, ambos subconjuntos do ImageNet [Deng et al. 2009], e, também, o conjunto Stanford Cars. ImageNette possui 10 classes distintas sendo mais simples que os outros dois, ImageWoof com 10 classes somente de cães e Stanford Cars com 196 classes somente de carros. Os subconjuntos do ImageNet foram utilizados com imagens limitadas a 160 pixels em seu maior lado, enquanto o Stanford Cars foi limitado a 256 pixels também em seu maior lado, mantendo ainda assim uma alta resolução. Todos os conjuntos foram obtidos por meio da biblioteca FastAI e divididos em 80-20% para treinamento e teste. Mil imagens da partição de testes de cada conjunto de dados foram utilizadas para a obtenção das métricas avaliadas.

A principal métrica utilizada foi a Feature Impact Index (FII), que avalia a capacidade da explicação de preservar a decisão original do modelo quando apenas os pixels considerados relevantes são mantidos. Ela é calculada como $(S_{original} - S_{mascara})$, que representa a pontuação do classificador para a imagem original e para a imagem contendo somente a região destacada pela máscara de explicação. Quanto menor o valor de FII, melhor a qualidade da explicação, já que o valor de $S_{mascara}$ se aproxima do original. Também foram avaliados o tempo de execução e o tamanho da área encontrada pelos métodos, definido como a porcentagem de pixels considerados relevantes em relação ao

total. Valores de áreas menores, associados à analise da pontuação de FII, podem indicar quais algoritmos são mais precisos em suas explicações.

4. Resultados

Todos os resultados apresentados nesta seção podem ser vistos na Figura 2, que exibe uma matriz de mapas de calor contendo os valores obtidos no experimento. Os dados estão agrupados de acordo com os três conjuntos de dados e as três arquiteturas utilizadas. Em todos os mapas, tonalidades mais claras indicam melhor desempenho da técnica e estas encontram-se ordenadas de forma crescente com base na média geral de FII. Letras ao lado dos valores indicam se os resultados são estatisticamente diferentes entre si, com base no teste de Nemenyi, feito após a realização do teste de Kruskal-Wallis. Letras iguais indicam que os resultados não são significativamente diferentes ($p\text{-}value > 0.05$) e vice-versa. As letras estão ordenadas de forma crescente, com a letra A indicando o melhor desempenho e a letra I o pior.

4.1. Feature Impact Index

A linha superior da Figura 2 apresenta os resultados da métrica FII, principal indicador utilizado neste trabalho para quantificar a qualidade das explicações geradas. Os algoritmos IPHA-RI e IPHA-PI, que utilizam a técnica de *inpainting*, aparecem como os melhores posicionados em todas as combinações de conjuntos de dados e arquiteturas avaliadas. Isto sugere que a substituição dos pixels utilizando informações do contexto da imagem pode contribuir de forma significativa para melhores soluções por parte dos algoritmos que utilizam otimização.

De forma geral, os métodos baseados em regiões contínuas (retângulos e polígonos) obtiveram os melhores resultados, ocupando quatro das cinco primeiras posições no *ranking* médio. Esse desempenho está diretamente relacionado a forma como a métrica FII é calculada, já que a imagem é reavaliada pelo classificador. Como essas abordagens produzem regiões contínuas, a imagem resultante da aplicação da máscara preserva melhor sua estrutura, evitando a adição de corpos estranhos ou a alta quantidade de pixels pretos espalhados na reavaliação. O algoritmo que mais sofre com isso é justamente o método por otimização IPHA-GA. Como suas explicações são baseadas na ocultação independente de pixels variados, suas máscaras tendem a ser altamente descontínuas, prejudicando sua reavaliação. Outro fator para o seu desempenho ruim é a alta dimensionalidade das soluções de sua otimização, ocasionada pelo uso de imagens de maior resolução.

Entre os métodos que utilizam gradientes, o Grad-CAM foi o melhor posicionado, mas esse resultado também pode ser explicado pela produção de uma região contínua, que suas atribuições geram ao serem interpoladas para o tamanho original da imagem. Os demais métodos por gradientes apresentaram desempenho semelhantes entre si, sem diferenças expressivas. Nos conjuntos de dados ImageNette e ImageWoof observa-se um aumento acentuado dos valores de FII a partir da sexta colocação, onde se encontram quatro de cinco dos métodos por gradiente, indicando que esses tiveram maior dificuldade do que os métodos por otimização. Este fato não acontece com relação ao Stanford Cars, que apresentou um maior equilíbrio entre todas as abordagens. Isto ocorre devido à sua maior complexidade, relacionada ao tamanho das imagens e quantidade de classes, que

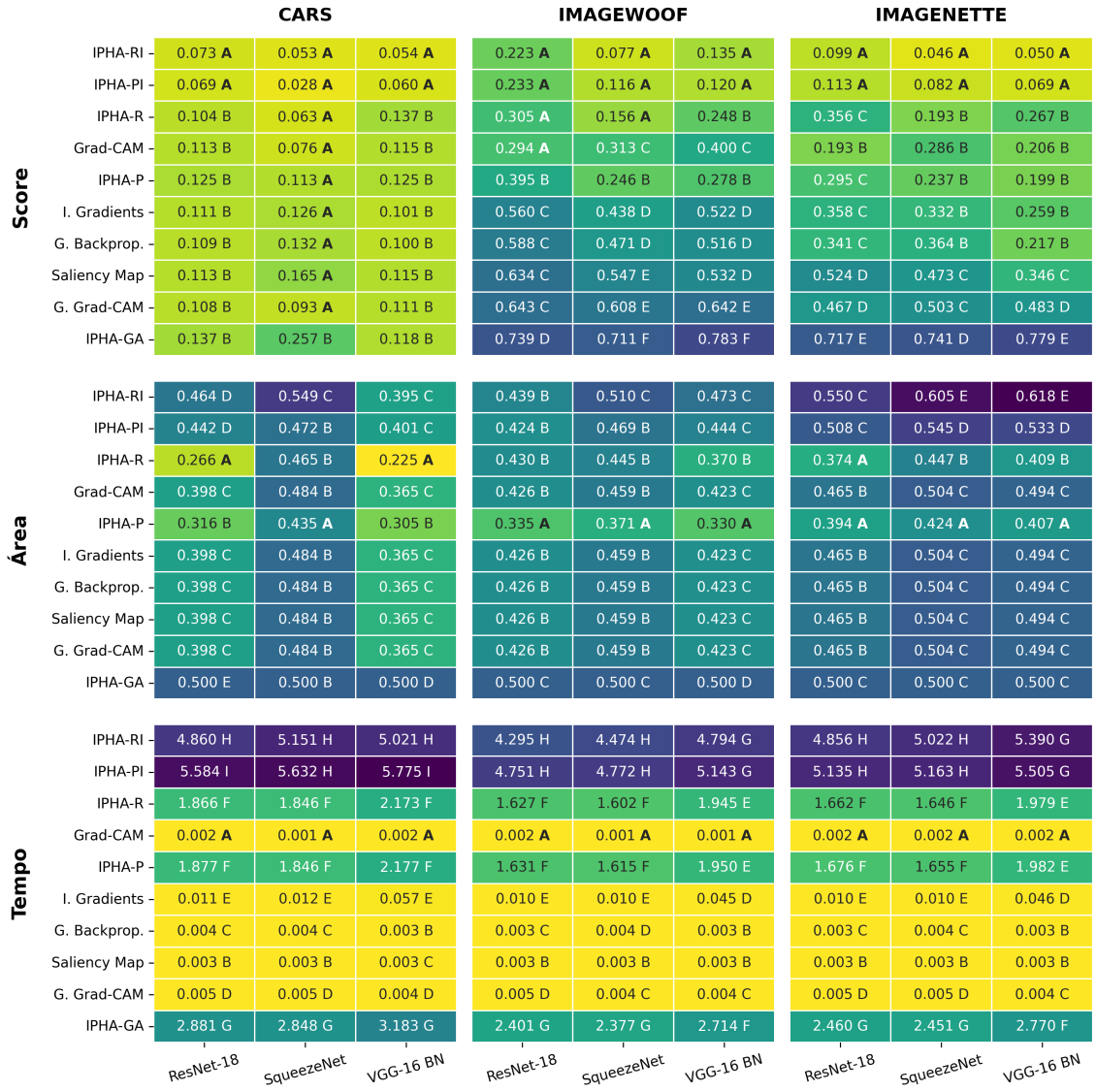


Figura 2. Resultado quantitativos para todas métricas obtidas. Tonalidades mais claras representam melhor desempenho.

faz com que sua pontuação original seja baixa, tornando possível melhores pontuações de FII para todos os algoritmos.

4.2. Área e Tempo de Execução

Para a métrica de área, observa-se um grande equilíbrio entre todos os métodos, resultado do limite imposto aos métodos por gradiente, que utilizam a média dos métodos por otimização e da restrição de uso de 0.5 de área para o IPHA-GA. Entre os quatro algoritmos restantes, é possível identificar dois grupos definidos pelo uso ou não de *inpainting*. Os métodos que o utilizaram tendem a usar áreas consideravelmente maiores, o que contribui para seu baixo FII, como visto anteriormente. Neste cenário, pode-se destacar o método IPHA-R que, em média, apresentou os menores valores de área, mantendo valores expressivos de FII. Por fim, nota-se uma tendência de uso maior de área para o conjunto de dados Imagenette, que possui classes de imagens mais diferentes entre si e

também para a arquitetura SqueezeNet, a mais simples entre as três utilizadas, indicando maior necessidade de evidências visuais nesses cenários.

Os resultados de tempo de execução mostram uma vantagem crítica para os métodos por gradiente em relação aos por otimização. Como dependem apenas de uma etapa de propagação e retropropagação na rede, os métodos por gradiente geram explicações de forma instântanea. Os métodos por otimização, por outro lado, precisaram passar por um processo com múltiplas gerações de um algoritmo evolutivo, exigindo um grande número de avaliações e aumentando consideravelmente seu tempo. Métodos que utilizaram *inpainting* foram ainda mais lentos devido à demanda de tempo do processo de substituição com contexto. Entre as abordagens por otimização, o destaque novamente vai para o IPHA-R, que dentre os algoritmos por otimização apresentou bons resultados de tempo, ainda com baixo FII. Por último, uma sutil diferença é vista para o conjunto de dados Stanford Cars, que apresenta um tempo um pouco maior em razão de suas imagens maiores e uma leve desvantagem da arquitetura VGG-16-BN quando comparada às demais arquiteturas avaliadas.

4.3. Síntese dos Resultados

Considerando todas as métricas avaliadas, observa-se um compromisso entre qualidade da explicação, avaliada pelo FII, área utilizada e custo de tempo de execução. Os métodos por otimização apresentaram melhores valores de FII, motivado pelo uso de regiões contínuas e *inpainting*, indicando maior capacidade de identificar trechos da imagem que são importantes para a decisão do modelo numa reavaliação. Métodos por gradientes destacaram-se pelo tempo de execução instantâneo e apresentaram medidas qualitativas de explicação médias. Entre todos, o IPHA-R apresenta um equilíbrio entre as três medidas avaliadas, combinando explicações bem avaliadas pelo FII com áreas reduzidas, além de menor tempo dentre os métodos por otimização.

5. Extensão para modelo CLIP

Além dos experimentos realizados com classificadores tradicionais, foi desenvolvida uma extensão da aplicação de métodos por otimização para um modelo do tipo Contrastive Language-Image Pre-training (CLIP). Diferente dos classificadores tradicionais, o CLIP realiza suas predições a partir da similaridade entre as imagens com descrições textuais dinâmicas passadas. Este modelo multimodal é considerado um Visual Language Model (VLM), capaz de realizar tarefas de classificação de imagens sem a necessidade de treinamento específico para cada conjunto de dados, sendo assim mais geral e flexível. Como a saída do modelo CLIP é também um score de probabilidade para cada rótulo dinâmico passado, seu uso em técnicas por otimização pode ser realizada de maneira igual aos classificadores tradicionais. O modelo CLIP escolhido para esta exploração foi o CLIP ViT-B/32 [Radford et al. 2021], modelo desenvolvido em 2021 pela Open AI, pré-treinado em um grande conjunto de imagens e textos e adquirido através da biblioteca Transformers [Wolf et al. 2020].

A Figura 3 mostra um exemplo da aplicação da técnica IPHA-RI em uma mesma imagem, retirada do ImageNette, mas com o conjunto de rótulo $\{\textit{fish}, \textit{flag}\}$ passado para o modelo CLIP na sequência de imagens superior e com o conjunto de rótulos $\{\textit{face}, \textit{golf_ball}\}$ para a sequência inferior. É possível observar que as imagens da primeira

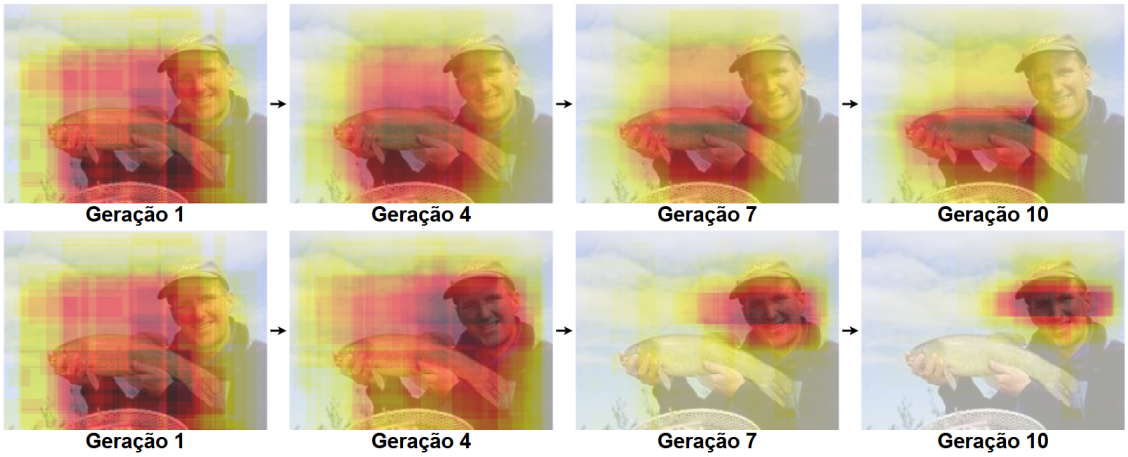


Figura 3. Explicação gerada por IPHA-RI da decisão do modelo CLIP com opção de rótulo *fish* (acima) e com opção de rótulo *face* abaixo.

geração compartilham as mesmas soluções, mas estas vão se diversificando através dos scores passados pelo modelo na função objetivo e chegam no objeto que melhor representa a imagem: o peixe, para a sequência superior e o rosto humano para a inferior. Também é interessante ressaltar que enquanto *Tench*, um tipo de peixe, é uma das classificações possíveis dos rótulos presentes no ImageNette, o rótulo *face* não possui uma contraparte neste conjunto de dados. Mesmo sem este rótulo para treinamento, o modelo conseguiu distinguir o rosto humano e a explicação gerada pela técnica de otimização conseguiu, qualitativamente, detectar isso.

A Figura 4 mostra outro exemplo de imagem com máscara gerada pelo IPHA-RI, para os mesmos conjuntos. Percebe-se um comportamento similar ao que ocorreu na imagem da Figura 3, mas desta vez relacionado aos rótulos de *flag* e *golf_ball* nas sequências superior e inferior, respectivamente. É possível observar como as soluções retangulares se direcionaram ao canto da imagem, onde está a bandeira que indica o buraco do golf, o que sugere uma boa explicação para o rótulo *flag*. Mas para a sequência inferior, apesar da máscara ter conseguido chegar até a bola de golf, a área da explicação gerada está muito abrangente, e acaba por incluir grandes áreas da imagem que não condiz somente com a bola. Isto sugere que ajustes ainda precisam ser feitos ao algoritmo, em princípio em sua segunda função objetivo, relacionada a área da máscara, para que se busque uma máscara final com maior enfoque no objeto mencionado pelo rótulo.

Os outros métodos de otimização apresentaram resultados qualitativos semelhantes, com exceção do IPHA-GA que possui maior diferença em seu funcionamento. A Figura 5 mostra um exemplo destas explicações. Os algoritmos que utilizam métodos multiobjetivo e máscaras contínuas obtiveram resultados similares e sobrepostos ao IPHA-RI. O IPHA-GA, por sua vez, apresenta uma máscara consideravelmente diferente e menos interpretável em relação à percepção humana devido à sua máscara não contínua e ao alto número de variáveis que o algoritmo precisa otimizar. Por fim, é importante observar que métodos por gradiente também podem ser aplicados ao modelo CLIP, mas sua aplicação depende do acesso às camadas internas do codificador visual e pode exigir adaptações adicionais em arquiteturas baseadas em Vision Transformers. Assim, apesar de ser possível a aplicação de técnicas de gradientes para explicação de modelos CLIP, ela não ocorre de

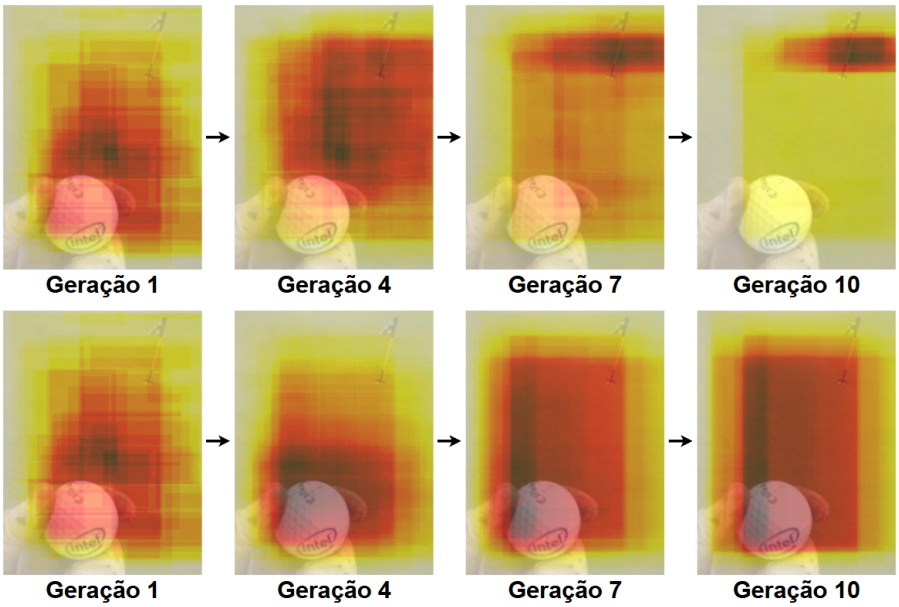


Figura 4. Explicação gerada por IPHA-RI da decisão do modelo CLIP com opção de rótulo *flag* (acima) e com opção de rótulo *golf_ball* abaixo.

forma tão direta e simples quanto a aplicação de técnicas de otimização.

6. Conclusão

Os resultados obtidos demonstraram uma vantagem consistente dos métodos baseados em otimização que utilizam regiões contínuas para a métrica Feature Impact Index, com destaque para os métodos IPHA-RI e IPHA-PI, que utilizam *inpainting*. Embora os métodos por gradiente tenham apresentado tempos de execução significativamente menores, os resultados indicam que as abordagens por otimização utilizadas foram mais eficazes na identificação de regiões que preservam a decisão original dos modelos. Como contribuição adicional, foi apresentada uma extensão dos métodos por otimização para modelos multimodais VLM baseados na arquitetura CLIP. Foi demonstrada a viabilidade dessa aplicação, em cenários que envolvem alinhamento entre imagens e descrições textuais dinâmicas. Isto ilustra a característica agnóstica dos métodos por otimização e sua vantagem na fácil utilização com modelos complexos como o CLIP, que também lida com reconhecimento textual, sem que seja necessário fazer nenhuma adaptação adicional.

Os experimentos de extensão para o modelo CLIP realizados tiveram caráter exploratório, mas os resultados iniciais indicam que a abordagem pode ser adaptada para além da tarefa de classificação tradicional de imagens. Como trabalhos futuros, pretende-se ampliar a avaliação em arquiteturas para diferentes tipos de tarefas como Zero-shot Classification e Visual Question Answering. Também pode-se investigar novas formulações que aproveitam o uso de *inpainting* na geração de explicações de máscaras, para que se possa diminuir seu custo de tempo, além da avaliação do mesmo em conjuntos de dados que possuam anotações de regiões.

Agradecimentos

O trabalho do autor Leonardo Matos recebeu apoio da Fundação CAPES através da bolsa Processo nº 88887.318495/2026-00 do Programa PRAPG.

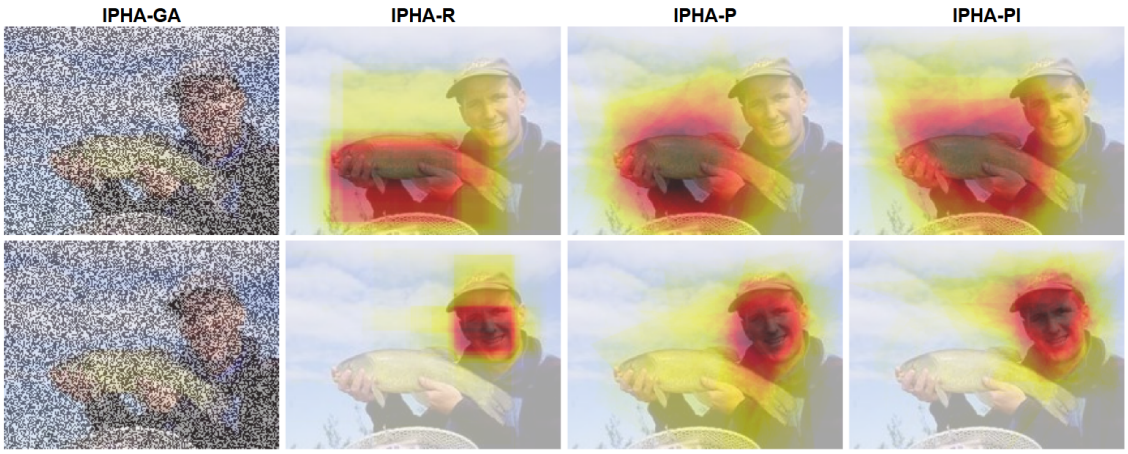


Figura 5. Explicação gerada pelos algoritmos de otmização restantes para decisão do modelo CLIP com opção de rótulo *fish* (acima) e com opção de rótulo *face* abaixo.

Referências

Barreto, M. H., Oliveira, C. L., Santos, F. A., Novais, P., Matos, L. N., and Britto, A. (2025). Genetic algorithm to understand image classification. In *World Conference on Information Systems and Technologies*, pages 541–553. Springer.

Blank, J. and Deb, K. (2020). pymoo: Multi-objective optimization in python. *IEEE Access*, 8:89497–89509.

Chen, L., Li, S., Bai, Q., Yang, J., Jiang, S., and Miao, Y. (2021). Review of image classification algorithms based on convolutional neural networks. *Remote Sensing*, 13(22):4712.

Deb, K., Pratap, A., Agarwal, S., and Meyarivan, T. (2002). A fast and elitist multiobjective genetic algorithm: Nsga-ii. *IEEE transactions on evolutionary computation*, 6(2):182–197.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255.

Fernandes, R., Matos, L., Andre Carvalho, M. B., and Santos, F. (2026). Optimization methods for generating saliency masks in image classifiers. In *IEEE World Congress on Computational Intelligence (WCCI)*. In press.

Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., Scardapane, S., Spinelli, I., Mahmud, M., and Hussain, A. (2024). Interpreting black-box models: a review on explainable artificial intelligence. *Cognitive Computation*, 16(1):45–74.

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.

Howard, J. and Gugger, S. (2020). fastai: A layered api for deep learning. *Information*, 11(2):108.

Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., and Keutzer, K. (2016). Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size. *arXiv preprint arXiv:1602.07360*.

Krause, J., Stark, M., Deng, J., and Fei-Fei, L. (2013). 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561.

Lin, K.-Y., Liu, Y., Li, L., and Dou, R. (2021). A review of explainable artificial intelligence. In *IFIP International Conference on Advances in Production Management Systems*, pages 574–584. Springer.

Naritomy, O., Bhardwaj, S., et al. (2020). Captum: A unified and generic model interpretability library for pytorch.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Santos, F. A. O., Zanchettin, C., Silva, J. V. S., Matos, L. N., and Novais, P. (2021). A hybrid post hoc interpretability approach for deep neural networks. In *International Conference on Hybrid Artificial Intelligence Systems*, pages 600–610. Springer.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626.

Simonyan, K., Vedaldi, A., and Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Springenberg, J. T., Dosovitskiy, A., Brox, T., and Riedmiller, M. (2014). Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*.

Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

Telea, A. (2004). An image inpainting technique based on the fast marching method. *Journal of graphics tools*, 9(1):23–34.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.