Explicações Visuais usando Grad-CAM: Uma Análise Abrangente e Validação Quantitativa

João Henrique Lessa
Centro de Informática
Universidade Federal de Pernambuco
Recife, PE – Brasil
jhmrl@cin.ufpe.br

Pierre Oriá
Centro de Informática
Universidade Federal de Pernambuco
Recife, PE – Brasil
pco2@cin.ufpe.br

***Abstract*—Este artigo apresenta uma análise abrangente e uma validação quantitativa do Gradient-weighted Class Activation Mapping (Grad-CAM), uma técnica independente de arquitetura usada para gerar explicações visuais, discriminativas entre classes, em redes neurais convolucionais. Revisamos a motivação para interpretabilidade em deep learning, as limitações de métodos anteriores de visualização e a formulação matemática que fundamenta o Grad-CAM. Realizamos, ainda, uma análise qualitativa visual, por meio da aplicação dessa técnica em diversos cenários. Por fim, avaliamos sua robustez do Grad-CAM, através de métricas de inserção e deleção (*Drop in Confidence*), propostas por *Petsiuk et al.*. Para mitigar o viés de seleção de hiperparâmetros (*Adaptive Overfitting*), os experimentos foram conduzidos no conjunto de dados ImageNet-V2. Nossos resultados destacam a utilidade do Grad-CAM para compreender o comportamento dos modelos e, consequentemente, aumentar a confiança em sistemas de deep learning.**

***Index Terms*—Explainability, Grad-CAM, Visualization, Deep Learning, Interpretability, Saliency Maps, Convolutional Neural Networks.**

I. Introdução

Modelos de aprendizado profundo atingiram desempenho de estado da arte em diversas tarefas de visão computacional e imagem médica. Entretanto, sua falta de interpretabilidade limita a adoção em domínios sensíveis à segurança. Historicamente, observa-se um trade-off entre complexidade e transparência: modelos simples oferecem explicações claras, porém com capacidade limitada, enquanto redes neurais profundas apresentam alto poder representacional, mas são difíceis de interpretar.

Técnicas iniciais de visualização, como *saliency maps*, deconvolução e *guided backpropagation*, destacam regiões relevantes da entrada, mas não produzem explicações discriminativas por classe. O *Class Activation Mapping* (CAM) [1] solucionou esse problema, porém exigia modificações na arquitetura da rede e ré-treino. Nesse contexto, o Grad-CAM [2] supera essas restrições ao utilizar informações de gradiente para gerar mapas de calor, específicos por classe, em qualquer arquitetura convolucional.

Embora essas técnicas gerem mapas de calor visualmente intuitivos, sua avaliação carece de rigor científico. A validação puramente qualitativa é propensa ao viés de confirmação humano e não garante que a explicação reflita fielmente o raciocínio matemático do modelo. Surge, portanto a necessidade de métricas objetivas que quantifiquem a fidelidade das explicações.

Este trabalho apresenta uma análise estruturada da formulação e das aplicações do Grad-CAM, além de uma validação quantitativa da sua robustez. Diferentemente de trabalhos anteriores, usamos o conjunto de dados **ImageNet-V2** [3] para obter as métricas de inserção e deleção, propostas por *Petsiuk et al.* [4]. A escolha do conjunto de dados tem por objetivo mitigar o fenômeno *Adaptive Overfitting*, que consiste em um viés na seleção dos hiperparâmetros e arquiteturas, motivado pela otimização intensa, por parte comunidade científica, do desempenho de modelos em *benchmarks* estáticos específicos.

II. Contexto e Trabalho relacionado

A. Explicabilidade em Aprendizagem Profunda

Historicamente, modelos de deep learning sofreram com um *trade-off* entre capacidade e interpretabilidade. Modelos simples, como regressões lineares ou árvores de decisão, oferecem transparência, mas não possuem a mesma capacidade de generalização de redes neurais profundas. Por outro lado, CNNs modernas apresentam excelente desempenho, mas são notoriamente difíceis de interpretar [5].

Nesse contexto, surge o Grad-CAM (Gradient-weighted Class Activation Mapping), proposto por Selvaraju et al. (2017), como uma técnica robusta de visualização que proporciona explicações discriminativas entre classes sem exigir mudanças na arquitetura do modelo.

B. Interpretabilidade e Métodos de Visualização

A interpretabilidade desempenha um papel fundamental na adoção de modelos de aprendizado profundo em aplicações sensíveis. Em primeiro lugar, a capacidade de visualizar por que um modelo tomou determinada decisão aumenta a confiança de usuários e especialistas, permitindo que a predição seja contextualizada no domínio do problema. Além disso, explicações visuais — como mapas de ativação ou *saliency maps* — são valiosas no diagnóstico de erros e na identificação de vieses, evidenciando se a rede está baseando sua decisão em regiões sem relevância semântica ou em correlações espúrias presentes no conjunto de dados. Como

destacado por *Petsiuk et al.*, um sistema de IA pode se comportar de maneira muito diferente de um humano, aprendendo a usar pistas do plano de fundo (por exemplo, usar a presença de grama para detectar vacas) que não são intuitivas para nós. Avaliações puramente visuais podem falhar em detectar esses comportamentos se a explicação apenas imitar a anotação humana em vez de revelar a verdadeira causa da decisão

C. Class Activation Mapping (CAM) e Grad-CAM

O Class Activation Mapping (CAM), proposto por Zhou et al., foi um dos primeiros métodos a produzir explicações verdadeiramente discriminativa entre classes, revelando quais regiões da imagem contribuem para a predição de uma classe específica. Entretanto, sua utilização é limitada a arquiteturas que encerram a porção convolucional com uma operação de *Global Average Pooling* (GAP), seguida de uma única camada totalmente conectada. Essa restrição estrutural exige modificar o modelo original, o que inviabiliza o uso de CAM em arquiteturas comuns que possuem múltiplas camadas densas após as convoluções, como VGG, CNNs tradicionais ou variantes com blocos adicionais.

O Grad-CAM, introduzido por Selvaraju et al., estende o CAM ao substituir a dependência estrutural por uma dependência baseada em gradientes. Em vez de exigir GAP, o método utiliza os gradientes do score da classe em relação aos mapas de ativação convolucionais para estimar a importância de cada canal. Dessa forma, Grad-CAM pode ser aplicado a qualquer rede que contenha camadas convolucionais, independentemente da arquitetura adotada ou da presença de múltiplas camadas densas.

Além disso, Grad-CAM supera limitações de métodos anteriores como guided backpropagation e deconvolução. Embora essas técnicas produzam mapas de sensibilidade de alta resolução, elas não são classe-discriminativas e podem refletir apenas padrões de alta ativação. O Grad-CAM, por outro lado, combina informações de ativação com informação de gradiente, resultando em mapas que destacam especificamente as regiões que contribuem para a classe de interesse.

D. Formulação do Grad-CAM

Nesta subseção, apresentamos a formulação matemática do Grad-CAM, seguindo a descrição original do método. Seja y_c o escore pré-softmax correspondente à classe c. Denotamos por $A_k \in \mathbb{R}^{u \times v}$ o k-ésimo mapa de ativação de uma dada camada convolucional.

Cálculo dos Gradientes: O Grad-CAM parte da derivada espacial do escore da classe em relação aos mapas de ativação:

$$\frac{\partial y_c}{\partial A_k}. \tag{1}$$

Pesos de Importância via Global Average Pooling: Para cada canal k, os gradientes são agregados por meio de uma média global, produzindo um peso que quantifica a relevância daquele mapa para a classe considerada:

$$\alpha_k^c = \frac{1}{uv} \sum_{i=1}^{u} \sum_{j=1}^{v} \frac{\partial y_c}{\partial A_{k,ij}}. \tag{2}$$

Combinação Linear Ponderada e Aplicação da ReLU: Os pesos são utilizados para formar uma combinação linear dos mapas de ativação, seguida de uma operação de ReLU, de modo a manter apenas contribuições positivas:

$$L^c_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_k \alpha_k^c A_k\right). \tag{3}$$

O resultado $L^c_{\text{Grad-CAM}}$ é então redimensionado para o tamanho da entrada e sobreposto à imagem original, produzindo um mapa visual que destaca as regiões mais discriminativas para a classe c. Este mecanismo permite interpretar decisões de redes convolucionais sem exigir modificações estruturais no modelo.

E. Limitações da Avaliação Humana e Métricas Causais

Embora métodos como o Grad-CAM gerem mapas de saliência visualmente coerentes, a avaliação dessas explicações permanece um desafio em aberto. Tradicionalmente, a qualidade das explicações é medida através da avaliação humana ou da localização de objetos (*Pointing Game*).

Contudo, a localização é apenas um passo intermediário para a explicação e pode não capturar corretamente o que explica a decisão do modelo.*Petsiuk et al.*. argumentam que métricas dependentes de humanos são inadequadas para avaliar se a explicação é a verdadeira causa da decisão, pois focam em quão bem a explicação imita a intuição humana, e não na fidelidade ao processo interno da rede.

Para mitigar essa subjetividade, foram propostas métricas causais de inserção e deleção. A métrica de deleção mede a queda na probabilidade da classe-alvo conforme os pixels importantes são removidos, assumindo que uma boa explicação deve identificar as regiões necessárias para a predição. Em contraposição, a métrica de inserção mede o aumento da probabilidade ao introduzir os pixes mais importantes, avaliando a suficiência da explicação. Essas métricas, por serem agnósticas ao humano, são mais aptas a avaliar explicações causais.

III. Metodologia

A metodologia deste trabalho foi estruturada para validar as técnicas de XAI sob múltiplas dimensões: fidelidade visual, consistência matemática, robustez a ataques e capacidade de auditoria de viés. O pipeline experimental divide-se em análise qualitativa (inspeção visual e contrafactual) e análise quantitativa (métricas causais), aplicadas a três arquiteturas distintas: VGG16, ResNet50 e GoogLeNet.

A. Arquiteturas e Conjuntos de Dados

Os experimentos principais utilizam modelos pré-treinados no ImageNet. A arquitetura GoogLeNet exigiu a implementação de *hooks* para a extração correta de gradientes em seus módulos Inception.

Para a avaliação quantitativa, adotou-se o conjunto de dados **ImageNet-V2** para evitar o viés de *Adaptive Overfitting* presente no conjunto de validação original.

Para o experimento de auditoria de viés, desenvolvemos a **SimpleCNN**, uma arquitetura convolucional com 4 camadas

treinada sobre um conjunto de dados sintético enviesado, construído especificamente para induzir correlações espúrias entre cor e classe (detalhado na Seção III-D).

B. Técnicas de Explicação Visual

A técnica central avaliada é o **Grad-CAM**, aplicada na última camada convolucional de cada uma das arquiteturas, conforme proposto por Selvaraju et al. Além da visualização padrão, exploramos duas variações:

1) Guided Backpropagation: Modifica a retropropagação, ao suprimir gradientes negativos nas ativações ReLU. O método isola detalhes de alta frequência, produzindo visualizações nítidas em nível de pixel, mas carece da capacidade de discriminação de classe inerente ao Grad-CAM.

2) Guided Grad-CAM: Combina a localização do Grad-CAM com os detalhes de alta frequência do *Guided Backpropagation* via produto de Hadamard ($\odot$), gerando visualizações de alta resolução e discriminativas entre classes.

C. Protocolo de Avaliação Quantitativa

Para medir a fidelidade das explicações, aplicamos o framework RISE [4] sobre o ImageNet-V2:

- **Deletion Score (Drop in Confidence):** Mede a necessidade. Removemos progressivamente os pixels mais relevantes (segundo o Grad-CAM) e calculamos a Área Sob a Curva (AUC) da queda de confiança. Uma queda rápida indica alta fidelidade.
- **Insertion Score:** Mede a suficiência. Inserimos progressivamente os pixels relevantes em um fundo neutro. Uma subida rápida na confiança indica que a explicação isolou o contexto mínimo necessário.

D. Protocolo de Avaliação de Robustez e Viés

Além da fidelidade, avaliamos a utilidade prática das explicações em cenários críticos:

1) Análise de Impacto Adversarial: Investigamos a estabilidade das explicações frente a ataques adversariais. Imagens foram perturbadas intencionalmente (utilizando método baseado em gradiente - PGD) para induzir erros de classificação. O Grad-CAM foi então aplicado às imagens perturbadas para verificar se a explicação visual muda drasticamente ou se revela a manipulação (focando no ruído adversarial em vez do objeto).

2) Auditoria de Viés (Bias): Treinamos a SimpleCNN em um conjunto de dados onde a classe "Gato" estava fortemente correlacionada à cor branca e "Cachorro" à cor preta. O modelo foi então testado em amostras *Out-of-Distribution* (ex: Cachorro Branco). O Grad-CAM foi utilizado para auditar se a rede aprendeu as características morfológicas (orelhas, focinho) ou se baseou a decisão puramente na correlação espúria de cor/textura.

IV. Métricas de Avaliação

A. Insertion Score

A métrica *Insertion* avalia a qualidade de um método de explicabilidade o aumento da confiança do modelo na classe predita conforme regiões consideradas importantes, para essa predição, pelo método são progressivamente adicionadas a uma imagem inicialmente degradada (por exemplo, borrada ou preenchida com um valor neutro).

O procedimento consiste em ordenar os pixels de acordo com sua relevância e inseri-los de forma cumulativa na imagem. Em cada etapa, calcula-se a probabilidade predita para a classe-alvo. Explicações de alta qualidade tendem a produzir um aumento rápido na confiança, indicando que as regiões inseridas de fato influenciam fortemente a decisão do modelo.

A pontuação final para uma única amostra é dada pela área sob a curva gerada ao longo do processo:

$$\text{AUC-Insertion} = \int_0^1 p(t)\,dt,$$

onde $p(t)$ é a confiança do modelo após inserir a fração t dos pixels mais relevantes. Quanto maior a AUC, melhor a explicação.

B. Drop in Confidence

A métrica *Drop in Confidence* avalia a qualidade de um método de explicabilidade medindo a queda da confiança do modelo na classe predita conforme regiões consideradas importantes, para essa predição, pelo método são progressivamente removidas da imagem original.

O cálculo realizado neste trabalho é um pouco diferente do que é feito no *framework* RISE: Utilizamos o *Drop* normalizado pela confiança original para garantir que a métrica avalie a fidelidade da explicação independentemente da calibração ou confiança inicial do modelo na classe predita. Em cada etapa, calcula-se a confiança do modelo na classe predita C antes da remoção de uma porcentagem dos pixels p^C_{orig} e a confiança da mesma classe após a deleção dos pixels p^C_{new}. O *Drop in confidence* para determinada amostra é dado pela seguinte expressão:

$$D^C = \frac{\max(0, p^C_{orig} - p^C_{new})}{p^C_{orig}} \tag{4}$$

Um valor alto de D^C indica que a máscara removeu regiões cruciais para a decisão do modelo, validando a qualidade da explicação. Em contrapartida, um valor baixo sugere que a explicação os pixels removidos não foi tão determinante na predição da classe C.

A pontuação final para uma única amostra é dada pela área sob a curva gerada ao longo do processo:

$$\text{AUC-Deletion} = \int_0^1 D(t)dt,$$

onde $D(t)$ é o tamanho da perda de confiança do modelo na classe predita, dado por D^C, após deletada a fração t dos pixels mais relevantes. Quanto maior a AUC, melhor a explicação.

V. Experimentos

Nesta seção descrevemos o protocolo experimental adotado para avaliar a qualidade das explicações geradas pelo Grad-CAM. Para ambas as análises, qualitativa e quantitativa, utlizamos três arquiteturas clássicas disponíveis na biblioteca `torchvision.models`: VGG16, GoogLeNet e ResNet50. Todas as redes foram carregadas com pesos pré-treinados do conjunto ImageNet1K_V1. O conjunto de dados usado para a obtenção das métricas *Insertion Score* e *Deletion Score* foi o ImageNetV2, composto por 10 000 imagens.

A. Análise Qualitativa das Explicações Visuais

Na nossa análise qualitativa, aplicamos o Grad-CAM, Guided Backpropagation e Guided Grad-CAM às mesmas arquiteturas pré-treinadas e analisamos os resultados visualmente.

Como é possível observar na Fig. 1, o Grad-CAM produz mapas explicitamente discriminativos entre classes, ressaltando regiões-chave associadas à decisão do modelo. Por outro lado, a Guided Backpropagation resulta em mapas de alta resolução, com fortes detalhes espaciais, mas sem discriminação entre classes, conforme descrito originalmente por [6]. Já a combinação Guided Grad-CAM oferece o melhor dos dois mundos: mapas de alta resolução com informações discriminativas, servindo como ferramenta útil para inspeção fina do comportamento da rede.

B. Estudo de Caso: Identificação de Viés em Classificação de Cães e Gatos

Para demonstrar o uso prático da interpretabilidade na depuração de modelos, conduzimos um experimento com uma rede convolucional simplestreinada para classificação binária entre cães e gatos.

O treinamento foi realizado combinando um subconjunto, selecionado manualmente, de uma base de dados do Kaggle[1] com imagens adicionais coletadas automaticamente e verificadas manualmente. O conjunto final contém 263 imagens (135 gatos e 128 cachorros), porém com um desbalanceamento proposital na distribuição das características: aproximadamente 90% dos gatos possuíam pelagem clara, enquanto cerca de 82% dos cachorros apresentavam pelagem predominantemente preta.

As visualizações com Grad-CAM evidenciaram que o modelo havia aprendido um atalho espúrio. Em vez de focar em regiões anatômicas relevantes — como olhos, focinho ou orelhas — a rede ativava quase exclusivamente as áreas de alto contraste associadas à cor predominante da pelagem. Isso indica que, durante o treinamento, o modelo internalizou a correlação espúria cor = classe, resultando em decisões equivocadas.

[1]Conjunto de dados com cães e gatos

Fig. 2: Exemplo de Grad-CAM aplicado a um gato com pelagem escura. As ativações se concentram em regiões claras irrelevantes, evidenciando que o modelo ainda associa claridade à classe “gato”.

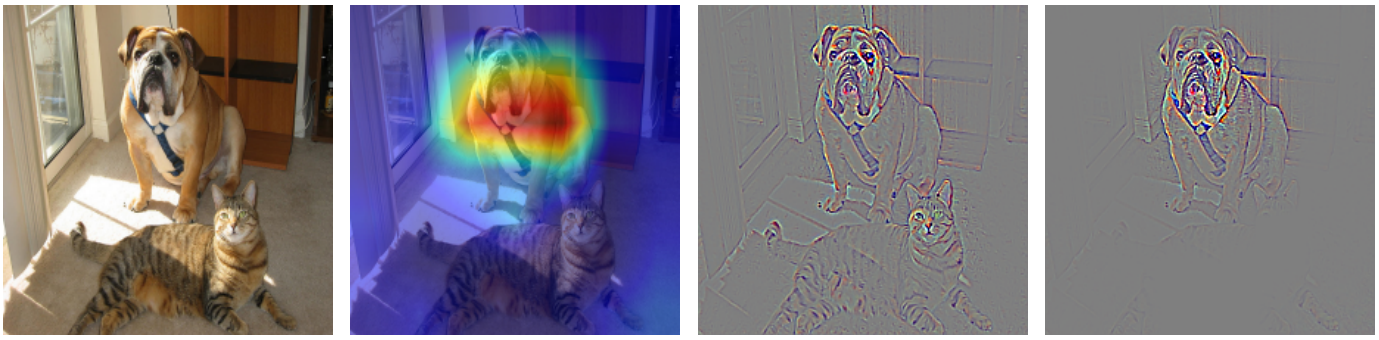
Fig. 3: Exemplo de Grad-CAM aplicado a um cachorro com pelagem branca. As ativações destacam áreas escuras do fundo em vez de características do animal, mostrando dependência indevida da cor.

C. Ataques Adversariais e o Comportamento do Grad-CAM

As Figuras 2 e 3 ilustram esse fenômeno. Embora o gato da primeira imagem tenha pelagem escura — contrariando o padrão majoritário do conjunto de dados — o Grad-CAM destaca regiões claras irrelevantes no fundo, sinalizando que o modelo ainda tenta apoiar-se no “padrão de gato claro”. De modo análogo, na imagem do cachorro branco, observamos ativações focadas em áreas escuras periféricas, também sem relação com o animal em si. Esses exemplos reforçam que as regiões apontadas pelo modelo como relevantes para a predição não correspondem às características do próprio animal, mas sim às cores claras ou escuras presentes na cena.

Ataques adversariais consistem em perturbações deliberadas, de pequena magnitude e geralmente imperceptíveis ao olho humano, aplicadas a uma imagem com o objetivo de induzir o modelo a produzir uma predição incorreta e altamente confiante. Entre os métodos mais influentes na literatura encontra-se o *Projected Gradient Descent* (PGD), um procedimento iterativo que generaliza o ataque FGSM (*Fast Gradient Sign Method*) ao realizar múltiplos passos de gradiente dentro de um conjunto permitido de perturbação.

O ataque PGD pode ser formulado como um processo de maximização do erro do modelo sob uma restrição normada.



(a) Imagem original (b) Grad-CAM para “Bull mastiff” (c) Guided Backprop para “Bull mastiff” (d) Guided Grad-CAM para “Bull mastiff”

Fig. 1: Comparação qualitativa entre resultados obtidos da aplicação do Grad-CAM, Guided Backprop e Guided Grad-CAM na VGG16. O Grad-CAM (b) revela de forma clara *onde* a rede busca evidências para a classe “Bull mastiff”, produzindo um mapa de calor de baixa resolução, porém semanticamente interpretável. O Guided Backprop (c) destaca *quais* pixels geram ativações fortes, resultando em um mapa altamente detalhado, porém não necessariamente relacionado à classe final. O Guided Grad-CAM (d) combina ambos: mantém a granularidade fina do Guided Backprop, mas filtrada pela atenção de classe do Grad-CAM, evidenciando apenas os padrões realmente relevantes para a predição “Bull mastiff”.

Dado um classificador f e uma imagem limpa $\mathbf{x}$, o ataque busca uma perturbação adversarial δ tal que:

$$\mathbf{x}_{\text{adv}} = \mathbf{x} + \delta,$$

com $\|\delta\|_\infty \leq \epsilon$. A atualização iterativa típica do PGD é dada por:

$$\mathbf{x}^{t+1} = \Pi_{B_\epsilon(\mathbf{x})} \left(\mathbf{x}^t + \alpha \cdot \text{sign} \left(\nabla_{\mathbf{x}} \mathcal{L}(f(\mathbf{x}^t), y) \right) \right),$$

onde $\Pi_{B_\epsilon(\mathbf{x})}$ projeta o ponto resultante no conjunto de admissibilidade definido por uma bola L_∞ de raio ϵ, e α é o passo de atualização. Esse processo garante que a imagem adversarial permanece próxima da imagem original, apesar de poder alterar substancialmente a predição do modelo.

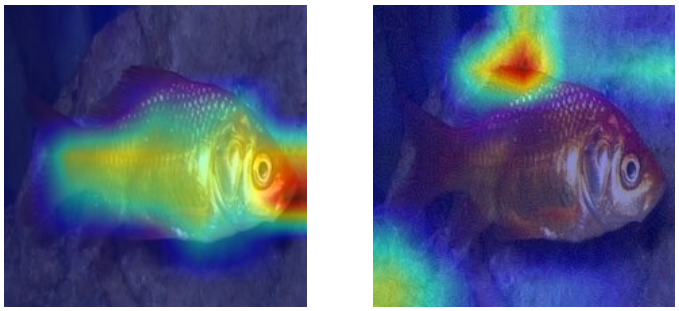
Geração de exemplos adversariais e análise via Grad-CAM. Neste estudo, aplicamos PGD para gerar entradas adversariais a partir de imagens saudáveis. As perturbações produzidas apresentam baixa perceptibilidade visual, mas ainda assim são suficientes para induzir mudanças drásticas no comportamento do modelo. Em um dos casos analisados, a imagem original foi corretamente classificada como *goldfish* (*Carassius auratus*), com confiança de 0.5978. Após a aplicação do ataque adversarial, a mesma imagem passou a ser classificada como *Lakeland terrier*, com confiança de 0.7001.

Sem ferramentas de interpretabilidade, o usuário final não possui meios de explicar tal comportamento. Contudo, a inspeção com Grad-CAM revela que, na versão adversarial, o modelo desloca fortemente sua atenção para regiões da imagem que não têm qualquer relação semântica com a classe verdadeira. Essa discrepância evidencia que a perturbação adversarial não apenas altera a saída numérica do modelo, mas também reorganiza suas regiões de atenção internas, levando a explicações completamente incongruentes com o conteúdo visual.

Visualização dos resultados. As Figuras 4 e 5 apresentam, respectivamente, as imagens original e adversarial, seguidas dos mapas Grad-CAM correspondentes. Observa-se que a imagem saudável produz um mapa consistente, focado no corpo do peixe, enquanto a imagem adversarial apresenta um mapa totalmente deslocado.

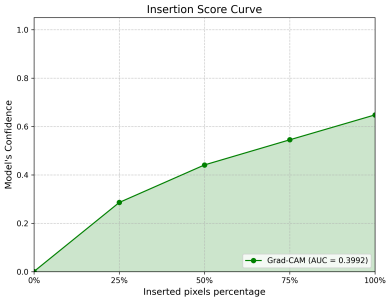
(a) Imagem saudável (classificada como *goldfish*). (b) Imagem adversarial (classificada como *Lakeland terrier*).

Fig. 4: Comparação entre imagem original e adversarial.

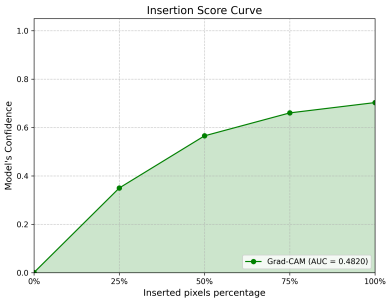


(a) Grad-CAM da imagem saudável (b) Grad-CAM da imagem adversarial

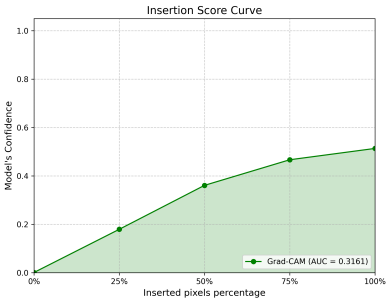
Fig. 5: Mapas de ativação evidenciam o impacto do ataque adversarial.



(a) Curva de Inserção - VGG16



(b) Curva de Inserção - ResNet50



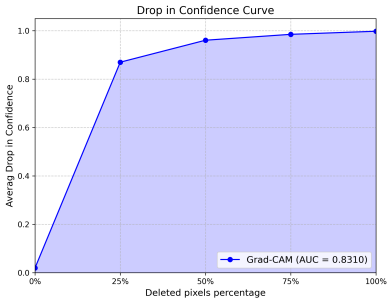
(c) Curva de Inserção - GoogLeNet

Fig. 6: Resultados da métrica de inserção para os três modelos pré-treinados. À medida que os pixels mais relevantes são progressivamente reintroduzidos, a confiança do modelo aumenta consistentemente. Isso evidencia que o Grad-CAM atribuiu importâncias aos pixels de maneira congruente à suficiência real de cada região para a predição.

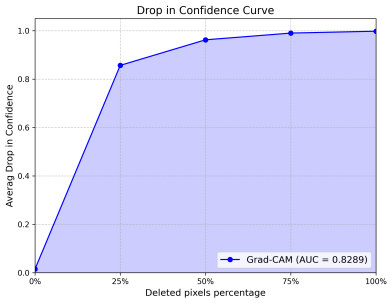
D. Avaliação via Insertion Score

O *Insertion Score* mede a suficiência das regiões consideradas relevantes pela explicação, por meio da inserção progressiva de pixels indicados pelo mapa de calor. Em nossos experimentos, utilizamos taxas de inserção iguais a $[0, 0.25, 0.5, 0.75, 1.0]$. A cada nível, calcula-se a confiança do modelo na classe-alvo e, ao final, a área sob a curva é utilizada como indicador quantitativo da qualidade da explicação.

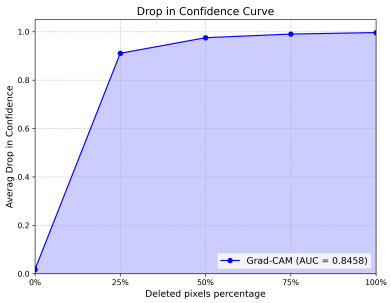
Os resultados na Fig.6 mostram que os mapas de calor do Grad-CAM fornecem representações suficientes para recuperar a probabilidade da classe-alvo de maneira consistente. Observamos diferenças em comparação ao trabalho do RISE, devido ao uso de taxas mais espaçadas, adotadas por limitações computacionais, o que desfavorece os resultados obtidos. Além disso, a mudança do conjunto de dados é um ponto fundamental: *Petsiuk et al.* usam **ImageNetVal**, historicamente sujeito a *Adaptive Overfitting*, nossos experimentos no **ImageNet-V2** mitigam esse viés de seleção, oferecendo uma estimativa mais fiel da capacidade de generalização do modelo.



(a) Curva de Deleção - VGG16



(b) Curva de Deleção - ResNet50



(c) Curva de Deleção - GoogLeNet

Fig. 7: Resultados da métrica de deleção para os três modelos pré-treinados. A queda média de mais de 80% na confiança do modelo para todas as arquiteturas evidencia que as regiões destacadas pelo Grad-CAM são causalmente importantes para a predição do modelo

E. Avaliação via Deletion Score

A métrica *Deletion Score* avalia o grau de causalidade entre os pixels destacados pelo Grad-CAM e a predição final do modelo, por meio da deleção progressiva dos mesmos. Dessa forma, avalia-se o quanto a confiança do modelo diminui conforme regiões consideradas relevantes são efetivamente retiradas.

No experimento conduzido, utilizamos frações de remoção definidas em $[0, 0.25, 0.5, 0.75, 1.0]$. Para cada fração, calculamos a queda da confiança, como descrito anteriormente, para cada imagem do conjunto de dados e extraímos a média.

Como é possível observar na Fig.7, mesmo sob o cenário desafiador do ImageNetV2, os três modelos exibiram um decaimento consistente na confiança. Isso evidencia que o Grad-CAM atribuiu importâncias aos pixels de maneira congruente à relevância causal real que cada região exerce sobre a predição. Os resultados corroboram as tendências reportadas no trabalho original do RISE — ainda que utilizemos um *Deletion Score* normalizado —, fortalecendo a validade da métrica e a fidelidade das explicações produzidas.

F. Conclusões Gerais dos Experimentos

Os resultados quantitativos e qualitativos convergem para demonstrar que:

- O Grad-CAM produz mapas discriminativos entre classes, úteis tanto para auditoria quanto para compreensão de decisões individuais;
- A Guided Backpropagation fornece detalhes espaciais finos, embora não seja discriminativa;
- A combinação Guided Grad-CAM oferece alta resolução e discriminação, sendo particularmente eficaz para inspeção detalhada;
- A análise de viés demonstra a importância prática dessas técnicas, permitindo identificar atenuar comportamentos espúrios aprendidos pelo modelo.
- Os resultados obtidos, para as métricas de *Deletion Score* e *Insertion Score*, para diferentes arquiteturas atestam a robustez do Grad-CAM

G. Conclusões Principais

A principal contribuição experimental deste estudo é a análise sistemática das métricas de Insertion e Deletion aplicadas no contexto do ImageNetV2, um conjunto mais robusto que o ImageNet original de 2012. Observamos que:

- O *Insertion Score* confirma a capacidade do Grad-CAM de identificar regiões suficientes para reconstruir a confiança do modelo;
- O *Deletion Score* evidencia a forte relação causal entre os pixels destacados e a predição final, mesmo em um conjunto de dados mais difícil;
- Os três modelos avaliados — VGG16, GoogLeNet e ResNet50 — apresentaram desempenho consistente, reforçando a generalidade das métricas e das explicações produzidas.

Esses resultados indicam que o Grad-CAM produz mapas de calor efetivos para interpretação, permitindo compreender o comportamento dos modelos e oferecendo uma base sólida para futuras extensões em aplicações sensíveis.

VI. Discussão

Os experimentos quantitativos e qualitativos apresentados revelam propriedades fundamentais do Grad-CAM. A análise do *Insertion Score* e do *Deletion Score* confirma que os mapas de calor não apenas identificam regiões visualmente relevantes, mas capturam relações causais ligadas à predição. O ganho rápido de confiança na inserção indica que o método preserva a suficiência da informação, enquanto as quedas acentuadas na deleção demonstram a necessidade causal das regiões destacadas, validando a fidelidade da explicação através de diferentes arquiteturas.

A comparação entre técnicas de explicabilidade sugere que a complementaridade pode potencializar os resultados obtidos. Enquanto o Grad-CAM oferece um alto potencial discriminativo entre classes, o *Guided Backpropagation* proporciona maior refinamento e resolução espacial. A fusão no Guided Grad-CAM integra as qualidades de ambos, gerando uma visualização de alta resolução, bem como discriminativa entre classes, o que maximiza a capacidade explicativa dos dois métodos

Além disso, o estudo de caso de viés reforça que a interpretabilidade deve ser parte integrante do ciclo de desenvolvimento. O fato de uma rede simples recorrer a correlações espúrias (cor da pelagem) em vez de características anatômicas, mesmo em uma tarefa trivial, serve de alerta. Isso evidencia que métricas de acurácia são insuficientes para garantir a confiabilidade de um modelo, tornando ferramentas como o Grad-CAM indispensáveis para a auditoria e a garantia de qualidade dos dados.

VII. Conclusões

Neste trabalho apresentamos uma análise abrangente do Grad-CAM, destacando sua formulação, propriedades e utilidade prática na interpretação de modelos convolucionais modernos. Demonstramos que, ao utilizar gradientes para ponderar mapas de ativação, o método produz explicações visualmente consistentes e discriminativas por classe, sem impor restrições estruturais às arquiteturas — uma limitação presente em abordagens anteriores como o CAM original.

Avaliamos o método de forma quantitativa utilizando as métricas de *Insertion Score* e *Deletion Score* aplicadas ao conjunto ImageNetV2, confirmando que os mapas produzidos são ao mesmo tempo suficientes para reconstruir a confiança do modelo e causalmente relevantes para a predição. As curvas obtidas em três arquiteturas distintas (VGG16, GoogLeNet e ResNet50) reforçam a robustez e a estabilidade do Grad-CAM em cenários variados.

A análise qualitativa — envolvendo Grad-CAM, Guided Backpropagation e Guided Grad-CAM — evidenciou o caráter complementar dessas técnicas, mostrando que o Grad-CAM fornece discriminação sem perder foco semântico, enquanto métodos guiados adicionam granularidade espacial. Finalmente, o estudo de caso sobre detecção de viés revelou a importância prática dessas ferramentas na auditoria de modelos, permitindo identificar correlações espúrias e apontando caminhos para a melhoria da qualidade dos dados e da generalização.

Por fim, a análise do comportamento, frente a ataques adversariais, destacou uma propriedade diagnóstica crucial do Grad-CAM. Observamos que perturbações imperceptíveis ao

olho humano são capazes de provocar uma drástica dissociação semântica nos mapas de ativação, desviando o foco da rede do objeto central para regiões irrelevantes. Esse comportamento evidencia um papel investigativo fundamental do Grad-CAM: ele revela, através de seus mapas de calor, que o sucesso do ataque adversarial se deve a uma discrepância no processo de atenção visual do modelo, permitindo ao usuário distinguir entre falhas naturais de generalização e ruídos maliciosos.

Como direções futuras, destacamos a exploração de variantes recentes de métodos baseados em gradientes, abordagens contrastivas de explicação e integrações multimodais que ampliem a capacidade de interpretar modelos cada vez mais complexos. Em conjunto, os resultados obtidos reforçam a relevância do Grad-CAM como técnica central no ecossistema de interpretabilidade em visão computacional, contribuindo para modelos mais confiáveis, transparentes e alinhados a aplicações em domínios sensíveis.

REFERENCES

[1] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2921–2929.

[2] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626.

[3] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar, "Do imagenet classifiers generalize to imagenet?" in *International Conference on Machine Learning (ICML)*, 2019.

[4] V. Petsiuk, A. Das, and K. Saenko, "Rise: Randomized input sampling for explanation of black-box models," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2018.

[5] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[6] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," in *International Conference on Learning Representations (ICLR)*, 2015.