UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
CENTRO DE ENGENHARIA ELÉTRICA E INFORMÁTICA
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

RENNAN ROCHA DE FREITAS

INTERPRETABILIDADE DE REDES NEURAIS CONVOLUCIONAIS APLICADAS A IMAGENS DE RESSONÂNCIA MAGNÉTICA

CAMPINA GRANDE - PB

2024

RENNAN ROCHA DE FREITAS

INTERPRETABILIDADE DE REDES NEURAIS CONVOLUCIONAIS APLICADAS A IMAGENS DE RESSONÂNCIA MAGNÉTICA

Trabalho de Conclusão Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

Orientador : Eanes Torres Pereira

CAMPINA GRANDE - PB

2024

RENNAN ROCHA DE FREITAS

INTERPRETABILIDADE DE REDES NEURAIS CONVOLUCIONAIS APLICADAS A IMAGENS DE RESSONÂNCIA MAGNÉTICA

Trabalho de Conclusão Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

BANCA EXAMINADORA:

Eanes Torres Pereira

Orientador – UASC/CEEI/UFCG

Heman Martins Gomes

Examinador – UASC/CEEI/UFCG

Francisco Vilar Brasileiro

Professor da Disciplina TCC – UASC/CEEI/UFCG

Trabalho aprovado em: 15 de maio de 2024.

CAMPINA GRANDE - PB

RESUMO

As redes neurais convolucionais atingiram acurácia similar à humana em diversas tarefas de visão computacional, porém a complexidade desses modelos, assim como o número crescente de parâmetros, criam representações de conhecimento e decisões que não são facilmente compreensíveis. Portanto, essas redes estão sendo usadas, na maioria das vezes, como algoritmos de caixa-preta. Dessa forma, é difícil a adoção de tais modelos em ambientes críticos que necessitem de explicações sobre seus resultados, como o contexto médico. Este estudo tem como objetivo treinar um classificador de imagens de tumores cerebrais a partir de um dataset com imagens de ressonância magnética, além de aplicar, avaliar e comparar técnicas de interpretabilidade nesse classificador. Como resultado, obtivemos um classificador com taxa de acurácia de 95% e uma parte das imagens do conjunto de teste foram explicadas através de 11 técnicas de interpretabilidade na vertente de atribuição de características. Em seguida as técnicas foram comparadas de forma subjetiva e objetiva revelando que a técnica RISE obteve a melhor pontuação objetiva dentre as técnicas avaliadas.

INTERPRETABILITY OF CONVOLUTIONAL NEURAL NETWORKS APPLIED TO MAGNETIC RESONANCE IMAGES

ABSTRACT

Convolutional neural networks have achieved human-like accuracy in various computer vision tasks. However, the complexity of these models, along with the increasing number of parameters, creates knowledge representations and decisions that are not easily comprehensible. Therefore, these networks are often used as black-box algorithms. As a result, it is challenging to adopt such models in critical environments that require explanations of their results, such as the medical context. This study aims to train a brain tumor image classifier using a dataset of magnetic resonance imaging (MRI) images, and to apply, evaluate, and compare interpretability techniques on this classifier. As a result, we obtained a classifier with an accuracy rate of 95%, and part of the test set images were explained using 11 feature attribution interpretability techniques. Subsequently, the techniques were compared subjectively and objectively, revealing that the RISE technique achieved the best objective score among the evaluated techniques.

Interpretabilidade de Redes Neurais Convolucionais aplicadas a imagens de ressonância magnética

Rennan Rocha de Freitas (Aluno)
rennan.freitas@ccc.ufcg.edu.br
Universidade Federal de Campina Grande
Departamento de Sistemas e Computação
Campina Grande, Paraíba

Eanes Torres Pereira (Orientador)
eanes@computacao.ufcg.edu.br
Universidade Federal de Campina Grande
Departamento de Sistemas e Computação
Campina Grande, Paraíba

RESUMO

As redes neurais convolucionais atingiram acurácia similar à humana em diversas tarefas de visão computacional, porém a complexidade desses modelos, assim como o número crescente de parâmetros, criam representações de conhecimento e decisões que não são facilmente compreensíveis. Portanto, essas redes estão sendo usadas, na maioria das vezes, como algoritmos de caixa-preta. Dessa forma, é difícil a adoção de tais modelos em ambientes críticos que necessitem de explicações sobre seus resultados, como o contexto médico. Este estudo tem como objetivo treinar um classificador de imagens de tumores cerebrais a partir de um *dataset* com imagens de ressonância magnética, além de aplicar, avaliar e comparar técnicas de interpretabilidade nesse classificador. Como resultado, obtivemos um classificador com taxa de acurácia de 95% e uma parte das imagens do conjunto de teste foram explicadas através de 11 técnicas de interpretabilidade na vertente de atribuição de características. Em seguida as técnicas foram comparadas de forma subjetiva e objetiva revelando que a técnica RISE obteve a melhor pontuação objetiva dentre as técnicas avaliadas.

1 INTRODUÇÃO

A era da aprendizagem profunda tem testemunhado uma ascensão significativa, sendo amplamente adotada para resolver desafios que anteriormente careciam de algoritmos com altas taxas de acerto, mas que dispunham de enormes volumes de dados. Áreas como visão computacional e processamento de linguagem natural têm se beneficiado dessas inovações [23]. A visão computacional, particularmente baseada na extração de características por meio de Redes Neurais Convolucionais (*Convolutional Neural Network* ou CNN), desbloqueou a solução para problemas outrora considerados intransponíveis. Questões como reconhecimento facial, veículos autônomos e assistência médica inteligente tornaram-se viáveis [12]. No entanto, apesar das conquistas notáveis, as CNNs permanecem opacas em relação ao processo de tomada de decisões, funcionando como modelos de caixa-preta. Essa falta de transparência resulta em classificações que carecem de explicabilidade, minando a confiança nos modelos. Além disso, mesmo diante de altas taxas de acurácia para um problema específico, levantam-se questões sobre a capacidade da rede em manter seu funcionamento esperado ao longo do tempo. Pois, ao aprender características incorretas, ela pode classificar corretamente em determinados conjuntos de dados devido a correlações em vez de causalidades. Portanto, para garantir a validade do modelo em diferentes distribuições de dados, torna-se imperativo a utilização de métodos para compreender quais características a rede está empregando para tomar suas decisões. A falta de confiança e a opacidade do modelo não são desafios significativos para todas as aplicações, mas tornam-se críticos em contextos de alto risco, como diagnósticos médicos, operações de carros autônomos e situações que envolvem classificação de indivíduos, como refugiados rotulados como terroristas. Além de aumentar a confiança do modelo, a interpretabilidade se faz necessária para atender conformidades legais que asseguram o direito de explicação, como por exemplo a regulação GDPR (*General Data Protection Regulation*) [6], definida pela União Europeia em maio de 2018, concedendo direito à explicação para decisões tomadas por sistemas computacionais. O aprimoramento da explicabilidade das CNNs torna-se, portanto, um ponto crucial a ser explorado para garantir a confiabilidade e aceitação em áreas sensíveis.

2 FUNDAMENTAÇÃO TEÓRICA

A interpretabilidade em modelos de *Deep Learning* ainda é um campo em aberto no domínio científico e busca esclarecer o conhecimento adquirido por esses modelos, seja no domínio de dados tabulares, texto ou imagens. Esta fundamentação teórica está focada no domínio de imagens, e, por consequência, nos estudos sobre interpretabilidade de CNNs, que são as redes mais utilizadas no contexto escolhido.

É comum pensar nas CNNs como redes que fazem tanto o papel de extração de características, como de classificação, sendo as primeiras camadas da rede, normalmente convolucionais, as consideradas responsáveis pela extração das características. Dessa forma, no campo de pesquisa relacionado à interpretabilidade de CNNs existem dois principais métodos de explicações com diversas variações, estes são: atribuição de características e visualização de características.

2.1 Atribuição de características

Os métodos focados em fornecer explicações sob a ótica de atribuição de características objetivam fornecer explicações para uma predição específica de uma imagem, esses relacionam regiões da imagem de entrada com a sua importância para a predição do modelo, dessa forma, gerando um mapa de calor na imagem de entrada.

Formalmente, o mapa de calor gerado pelas técnicas explica uma imagem I produzindo uma imagem S de mesmo tamanho, onde $S_{h,w}$ indica a contribuição do pixel $I_{h,w}$ [25]. Em vários trabalhos a noção de contribuição foi definida como sensibilidade [19], relevância [11], influência local [16], valores *Shapley* [13], ou ativações de filtro [18].

As duas principais abordagens existentes para gerar mapas de importância são a abordagem por perturbação da entrada e a abordagem baseada em retropropagação dos gradientes. Na abordagem por perturbação da entrada, são introduzidos ruídos na imagem

com o objetivo de encontrar as regiões (pixels ou super-pixels) que são mais importantes para a inferência da classe de interesse. Enquanto nas abordagens baseadas na retropropagação de gradientes são considerados os pesos internos do modelo, calculando a responsabilidade de um pixel na predição da classe escolhida ao propagar um sinal desde a camada final até a camada de entrada.

Como ainda não existe um consenso sobre a definição de "atribuição" nesse contexto, existem muitos métodos concorrentes e com pouca avaliação sistemática [25]. Os métodos utilizados neste estudo são: Gradient [19], SmoothGrad [20], InputGrad, IntegratedGradients [22], GuidedIG [10], IGSmoothGrad, BlurIG [24], GradCam [18], RISE [15], XRAI [9] e LIME [16]. Descrições sobre as diferenças dos métodos serão apresentadas na seção 3.3.

2.2 Visualização de características

Os métodos que seguem a linha de visualização das características procuram entender que tipo de entrada poderia causar determinado comportamento, como a ativação de um neurônio ou um comportamento final. Essa técnica é comparada a um método da neurociência que busca entender o cérebro através do estudo dos principais estímulos que ativam determinados grupos de células [14].

Uma abordagem comum é otimizar imagens para que elas maximizem uma saída desejada na rede neural, conhecida como *Activation Maximization*, conforme introduzido por [3]. No entanto, um desafio significativo nesse campo é produzir imagens sem ruídos de alta frequência, que sejam compreensíveis para os seres humanos, já que padrões não reconhecidos ou úteis podem emergir. A Figura 1 mostra um exemplo de *Activation Maximization* partindo-se de uma imagem aleatória em direção ao gradiente de ativação da classe "Pimentão" do modelo CaffeNet [8], apesar da alta ativação no neurônio, ainda existem ruídos de alta frequência na imagem a ponto da imagem não ser reconhecida por humanos. Abordagens para visualização de características são detalhadas com mais profundidade por [14].

3 MATERIAIS E MÉTODOS

Nesta seção encontram-se a descrição da base de dados utilizada, as ferramentas e a arquitetura do modelo treinado, a descrição das técnicas de interpretabilidade aplicadas e o procedimento de avaliação das técnicas.

3.1 Descrição da base de dados

Imagem de ressonância magnética (IRM) é uma técnica de imagem médica usada na radiologia para produzir imagens bidimensionais ou tridimensionais de alta qualidade do cérebro e tronco cerebral [1]. A base de dados utilizada se chama "Crystal Clean: Brain Tumors MRI Dataset"[7] e possui IRMs do cérebro. A base de dados é uma modificação da "Brain tumor classification MRI"[1]. A base de dados utilizada inclui três classes de tumores (Pituitary, Glioma, Meningioma), e uma classe representando um cérebro normal. Cada classe é composta de imagens em escala de cinza incluindo três diferentes visões do cérebro: axial, coronal e sagital. O número de imagens para cada classe estão dispostos na Tabela 1:

A base de dados Crystal Clean foi escolhida porque a base de dados original carece de limpezas e padronizações. Para realizar a limpeza de dados, Hashemi, et al. [7] aplicaram uma sequência de operações nas imagens, incluindo:

- Remoção de exemplos duplicados: foi utilizado um método de comparação por vetorização nas imagens para remover imagens duplicadas.
- Correção de rótulos: usando conhecimento do domínio, foram inspecionados e corrigidos os rótulos de imagens com rótulos incorretos.
- Redimensionamento das imagens: todas as imagens da base de dados foram redimensionadas para um tamanho padrão aceitável e eficiente em memória de 224×224 pixels.

Para aumentar a diversidade e robustez da base de dados, Hashemi et al. [7] aplicaram as seguintes técnicas de aumento de dados:

- Adição de ruído: pixels aleatórios foram escolhidos e suas intensidades foram definidas para 0 ou 255.
- Equalização de histograma: foi aplicada a equalização de histograma para melhorar o contraste e os detalhes das imagens.
- Rotação: as imagens foram rotacionadas nos sentido horário e anti-horário em ângulos específicos.
- Ajuste de brilho: foi feita a adição e subtração das intensidades dos pixels das imagens.
- Espelhamento horizontal e vertical das imagens.

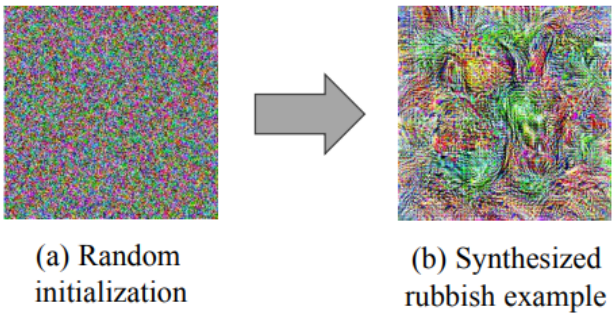

Figura 1: Exemplo de *activation maximization* sem imagem à priori. Começando de uma imagem aleatória (a), iterativamente muda-se a imagem em direção ao gradiente para maximizar a ativação de um determinado neurônio, neste caso a classe "Pimentão" da rede CaffeNet. Apesar da alta ativação do neurônio e ser classificado como "Pimentão", a imagem (b) tem altas frequências e não é reconhecida por humanos. Fonte: [14]

Tabela 1: Número de imagens para cada classe

Classe	N° de exemplos	Porcentagem
Normal	3066	14%
Glioma	6307	29%
Meningioma	6391	29%
Pituitary	5908	27%

Após o aumento de dados, o número total de imagens na base de dados aumentou em 563%, melhorando assim a robustez da base. A Figura 2 mostra algumas imagens da base de dados com seus rótulos na parte superior:

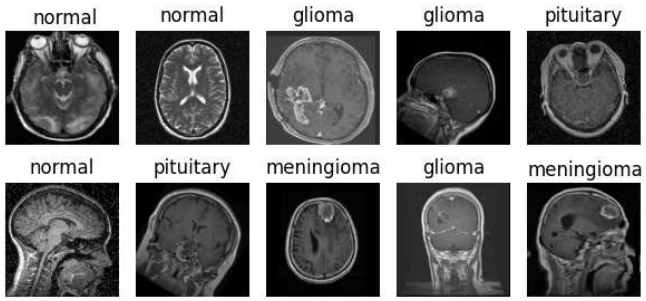

Figura 2: Exemplos de imagens da base de dados utilizada e seus rótulos.

A base de dados possui diversas imagens para o mesmo indivíduo devido ao aumento de dados aplicado, então, para evitar a contaminação do conjunto de teste com imagens do mesmo indivíduo que foram utilizadas no conjunto de treinamento, primeiro foram selecionados aleatoriamente 80% dos indivíduos para o conjunto de treinamento e 20% para o conjunto de teste, em seguida, a divisão das imagens foi aplicada de acordo com a divisão dos indivíduos feita anteriormente. Após a divisão, os conjuntos de teste e treinamento possuem os seguintes números de imagens mostrados na Tabela 2:

Tabela 2: Número de imagens para o conjunto de treinamento e teste.

Classe	Treino	Teste
Normal	2471	595
Glioma	5033	1274
Meningioma	5110	1281
Pituitary	4669	1239

Apesar de todas as imagens já estarem pré-redimensionadas para 224×224 pela própria base de dados, resolvemos redimensioná-las novamente para 150×150 objetivando reduzir o custo computacional e o tempo de treinamento, sem comprometer a qualidade dos dados.

3.2 Treinamento do modelo

Para chegar ao modelo proposto foi utilizado o *framework* para *Deep Learning* em Python 3 chamado Keras [2], que é baseado no Tensorflow [4]. O modelo proposto utiliza as seguintes camadas implementadas no Keras:

- Camada Convolucional: camada Conv2D do Keras define um *kernel* com pesos compartilhados e treináveis que realizará convoluções na matriz de entrada e em seguida aplicará o resultado numa função de ativação, que no caso do modelo proposto é a função Unidade Linear Retificada (ou ReLU), definida por $f(x) = max(0, x)$. Essa função de ativação é amplamente utilizada em redes neurais porque ajuda a resolver o problema de desaparecimento do gradiente, que pode ocorrer em funções saturadas como a função sigmóide, acelerando assim o aprendizado da rede.
- Camada de *Pooling*: as camadas de *pooling* são usadas como uma maneira de reduzir a dimensionalidade e conseguir invariância espacial utilizando uma janela retangular (2×2 no modelo proposto) que percorre a imagem em um determinado passo horizontal e vertical (2×2) e então é calculado o máximo ou a média dos elementos da janela. Ao reduzir a dimensionalidade, por consequência, também se diminui o número de operações na rede [17].
- Camada de *Dropout*: a utilização de camadas de *Dropout* é um dos métodos mais comuns para reduzir o *overfitting*. Nesta camada, alguns neurônios da rede aleatoriamente têm sua saída zerada durante a etapa *forward*, fazendo com que estes neurônios não contribuam para o cálculo das ativações das camadas posteriores. Ao utilizar o *Dropout*, a rede é forçada a generalizar melhor através de mais neurônios, evitando reter o conhecimento em neurônios específicos, isso também ajuda de maneira significativa a acelerar a fase de treinamento [21]. Nos testes realizados, foi observado que a taxa de *Dropout* de 30% nas últimas camadas da rede foi a mais efetiva no modelo proposto.
- Camada *Flatten*: é utilizada para transformar a matriz de entrada em um vetor de apenas uma dimensão, a fim de conectar em camadas densas posteriores.
- Camada Totalmente Conectada: é utilizada para conectar todos os neurônios de uma camada a outra ($N \times M$ conexões) e em seguida aplicar uma função de ativação. Na rede proposta foram utilizadas 2 camadas totalmente conectadas, uma com ativação ReLU, e outra com ativação Softmax, sendo a saída da *Softmax* definida pela Equação 1. A ativação *Softmax* foi utilizada para classificação na última camada da rede, pois a função limita os valores de saída entre 0 e 1, e a soma total da camada é igual a 1. Dessa forma, é obtida a probabilidade normalizada de cada classe.

$$y(z)_j = \frac{e^{z_j}}{\sum_{k=1}^{k} e^{z_k}} \tag{1}$$

Foram criadas 3 versões de arquiteturas de modelos testando diversos hiperparâmetros, cada versão melhorando a acurácia no conjunto de teste. Os valores testados para cada hiperparâmetro estão presentes na Tabela 3.

A Figura 3 mostra a estrutura da CNN proposta. Ela possui 14 camadas, começando da camada de entrada que recebe uma matriz de tamanho 150×150, em seguida 4 blocos que incluem uma camada de convolução com função de ativação ReLU e uma camada de *pooling*. Na última camada de *pooling* foi utilizado o *AveragePooling* e nas outras camadas o *MaxPooling*, pois foi a configuração que demonstrou melhor acurácia no problema. Em seguida uma camada totalmente conectada com ativação ReLU e uma camada de classificação com ativação *softmax* para produzir a probabilidade normalizada de cada classe.

Tabela 3: Diferentes arquiteturas e hiperparâmetros testados antes de chegar ao modelo final.

Hiperparâmetro	Valores
Número de camadas convolucionais + ReLU	3, 4
Número de camadas de *dropout*	0, 1, 2
Número de camadas totalmente conectadas	1, 2
Número de kernels convolucionais	32, 64, 128, 256, 512
Tamanho do *kernel*	2, 3, 5
Camadas de *pooling*	*AveragePooling*, *MaxPooling*
Tamanho da janela de *pooling*	2
Otimizador	Adam
Tamanho do *mini-batch*	32, 64
Taxa de *dropout*	0,1, 0,2, 0,3, 0,5
Taxa de aprendizado inicial	0,005, 0,001, 0,0001
Fator de decaimento da taxa de aprendizagem	0,75

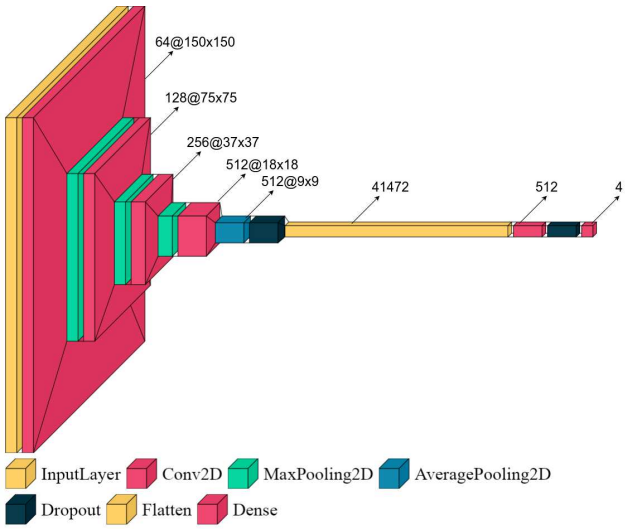

Figura 3: Arquitetura do modelo proposto.

Para prevenir o *overfitting*, duas camadas de *dropout* com taxa de 30% foram adicionadas após a última camada de *pooling* e a primeira totalmente conectada, além da utilização de regularização do tipo L2 nas últimas duas camadas da rede com o objetivo de adicionar uma penalidade à função de custo e introduzir uma diminuição na magnitude dos pesos, como mostrado na Equação 2:

$$J'(W) = J(W) + \lambda \sum_{i=1}^{k} w_i^2 \tag{2}$$

onde J é a função de custo original, λ é o hiperparâmetro de regularização e w são os pesos correspondentes para $i = 1, \ldots, k$. Por último, empregamos a técnica de *early stopping* para monitorar a perda no conjunto de validação a cada época do treinamento. Essa abordagem visa interromper o processo de treinamento antes de completar todas as épocas caso seja detectada uma estabilização, visando evitar o *overfitting*. Para a função de custo, foi utilizada a perda de entropia cruzada, que pode ser estimada pela Equação 3, onde p é o vetor dos rótulos verdadeiros, e $q(x)$ é o vetor de saída da camada *softmax*.

$$H(p, q) = -\sum_{x} (p(x) * \log(q(x))) \tag{3}$$

Os hiperparâmetros que apresentaram melhores resultados e estão presentes no modelo proposto estão dispostos na Tabela 4:

Tabela 4: Hiperparâmetros utilizados no modelo proposto

Hiperparâmetro	Valor
Otimizador	Adam
Tamanho de *minibatch*	32
Taxa de aprendizagem inicial	0.0005
Função de custo	Entropia cruzada
Número de épocas	70
Callbacks	*EarlyStopping* (paciência = 5), *ModelCheckpoint*

3.3 Técnicas de interpretabilidade

As técnicas de interpretabilidade utilizadas são de atribuição de características, ou seja, explicam a inferência para uma instância específica. A seleção de técnicas levou em conta, principalmente, a facilidade de reprodução, a maioria está disponibilizada através da biblioteca python Saliency, já as técnicas RISE e LIME foram utilizadas através de códigos e bibliotecas disponibilizados pelos autores. As técnicas utilizadas foram:

- **Gradiente:** é uma técnica simples proposta por [19] que consiste em gerar um mapa de saliência encontrando a derivada da imagem de entrada com respeito a uma classe através de *back-propagation*.
- **SmoothGrad:** é uma técnica baseada na Gradiente que busca melhorar a visualização do mapa de saliência aplicando suavizações [20].
- **Input-Grad:** a técnica consiste em processar a imagem gerada pela técnica Gradiente reutilizando a imagem original.
- **Integrated Gradients:** a técnica proposta por [22] consiste em definir uma imagem *baseline* e calcular os gradientes cumulativos de todos os pontos ao longo do caminho entre a *baseline* e a imagem de entrada. Além disso, a técnica busca satisfazer dois axiomas julgados importantes por [22]: sensibilidade e invariância à implementação.
- **Guided Integrated Gradients:** modificação da Integrated Gradients que busca modificar o caminho entre a *baseline* e a imagem de entrada a fim de diminuir os ruídos gerados no mapa de saliência [10].
- **Blur Integrated Gradients:** também é uma modificação da Integrated Gradients que visa suavizar o mapa de saliência aplicando sucessivos desfoques gaussianos [24].

- **GradCam:** técnica proposta por [18] que utiliza os pesos da última camada convolucional da rede para gerar o mapa de saliência.
- **RISE:** técnica que se baseia em perturbação, utilizando máscaras aleatórias aplicadas às imagens de entrada, a fim de identificar quais regiões influenciam nas predições quando essas regiões são ocultadas [15].
- **XRAI:** baseando-se na técnica Integrated Gradients, esta abordagem emprega um segmentador de regiões para gerar um mapa de atribuição que transmite as atribuições por meio de regiões segmentadas [9].
- **LIME:** técnica que explica a saída do modelo aprendendo um modelo interpretável localmente em torno da instância de entrada. A técnica modifica a imagem de entrada e observa o impacto resultante na saída do modelo. Seu mapa de saliência gerado é baseado em super-pixels [16].

Para a etapa de interpretabilidade foram separados 60 exemplos de cada classe do conjunto de teste, não utilizados na fase de treinamento, para serem explicados pelas técnicas a serem avaliadas. Após a seleção, as imagens foram submetidas para cada biblioteca, então foram gerados e salvos os mapas de calor em uma pasta com o nome da respectiva técnica. Além disso, também foi mensurado o tempo de execução de todas as técnicas.

3.4 Avaliação das técnicas

O método de avaliação utilizado para comparar as técnicas é baseado em perturbações na imagem de entrada e foi sugerido, com código aberto, por [5]. O método de perturbações consiste em primeiramente selecionar as regiões de pixels mais intensas no mapa de calor da imagem de entrada gerado pela técnica de interpretabilidade e sucessivamente substituir essas regiões na imagem original por ruído aleatório e então medir a queda da confiança do modelo naquela classe. Uma queda maior significa um método de atribuição mais acurado, visto que as atribuições foram capazes de identificar as regiões da entrada que mais explicam a saída do modelo.

Para avaliar as técnicas foram selecionadas 60 imagens aleatórias do conjunto de teste, em seguida foi gerado o mapa de calor de cada imagem para todas as técnicas testadas. O procedimento de avaliação por perturbações é descrito formalmente da seguinte maneira por [5]: Primeiro, é utilizada uma janela deslizante de tamanho $k \times k$ no mapa de saliência gerado para encontrar uma sequência ordenada $S = (r_1, r_2, ..., r_L)$ que contém as L regiões não sobrepostas mais salientes. A ordenação das regiões é baseada na média absoluta dos pixels da região em cada janela deslizante, da maior para menor. Uma região r_i com uma média alta significa a presença de uma região que explica melhor o modelo, segundo a técnica utilizada. Depois, para cada região de tamanho $k \times k$ na sequência S é gerado um ruído de pixels aleatórios e a região na imagem original é substituída pela perturbação. Por fim, é feita predição no modelo com a imagem perturbada. Para cada imagem em uma técnica específica será gerado um vetor com L pontos representando a confiança do modelo na classe original ao longo das L perturbações.

Os resultados das perturbações nas imagens foram representados por meio de vetores, sobre os quais foi aplicada uma operação de média para gerar um vetor médio correspondente a cada técnica de interpretabilidade avaliada. Posteriormente, esses vetores médios foram utilizados para traçar linhas representativas de cada técnica em um gráfico. Além disso, também foi calculada a métrica de área abaixo da curva média. Essa abordagem permitiu a comparação entre as diferentes técnicas de interpretabilidade. Na Figura 4 está representado um diagrama do experimento.

4 RESULTADOS

4.1 Métricas do Modelo

As Figuras 5 e 6 apresentam, respectivamente, os gráficos com as métricas de acurácia e perda dos conjuntos de treinamento e validação ao longo do processo de treinamento. O eixo das abscissas representa as épocas, enquanto o eixo das ordenadas representa os valores das métricas, sendo que as linhas azul e laranja correspondem, respectivamente, aos conjuntos de treinamento e validação. Observa-se que o modelo alcançou a convergência em um número reduzido de épocas, com o treinamento sendo interrompido pelo *callback* de *early stopping* na época 34, onde é possível observar o início do afastamento entre as linhas depois da estabilização. A acurácia final no conjunto de validação (que é o mesmo de teste), chegou a 94,71%, e as linhas de acurácia dos dois conjuntos apresentaram uma discrepância final de 6%. Notavelmente, ao realizar testes sem procedimentos de regularização na versão inicial do modelo, a diferença entre as linhas de acurácia chegou a ser de 11%, indicando a presença de *overfitting*, logo, a versão final demonstrou melhorias através da regularização implantada via *dropout* e L2.

A Figura 7 apresenta a matriz de confusão do modelo proposto ao submeter o conjunto de teste. Observa-se que o modelo apresentou alguns erros ao distinguir as classes Glioma/Meningioma, assim como as classes Normal/Glioma. Também foi observado que as imagens classificadas incorretamente entre os pares Glioma/Meningioma e Normal/Glioma incluiram todas as variações da mesma imagem feitas através do aumento de dados.

Por fim, a Tabela 5 apresenta as estatísticas de precisão, revocação e F1-Score. Percebe-se que a classe Pituitary obteve o melhor resultado de F1-Score e as médias de F1-Score ficaram em 0,95.

Tabela 5: Estatísticas sobre o modelo proposto no conjunto de teste

	Precisão	Revocação	F1-Score
Normal	0,91	0,96	0,93
Glioma	0,94	0,90	0,92
Meningioma	0,94	0,94	0,94
Pituitary	0,98	1,00	0,99
Acurácia			0,95
Média macro	0,94	0,95	0,95
Média ponderada	0,95	0,95	0,95

4.2 Aplicação de técnicas de interpretabilidade

Para os testes das técnicas de interpretabilidade do tipo atribuição de características foram utilizadas as bibliotecas Saliency e LIME, além do código disponibilizado pelo autor da técnica RISE. Foram selecionadas 60 imagens aleatórias do conjunto de teste e cada uma

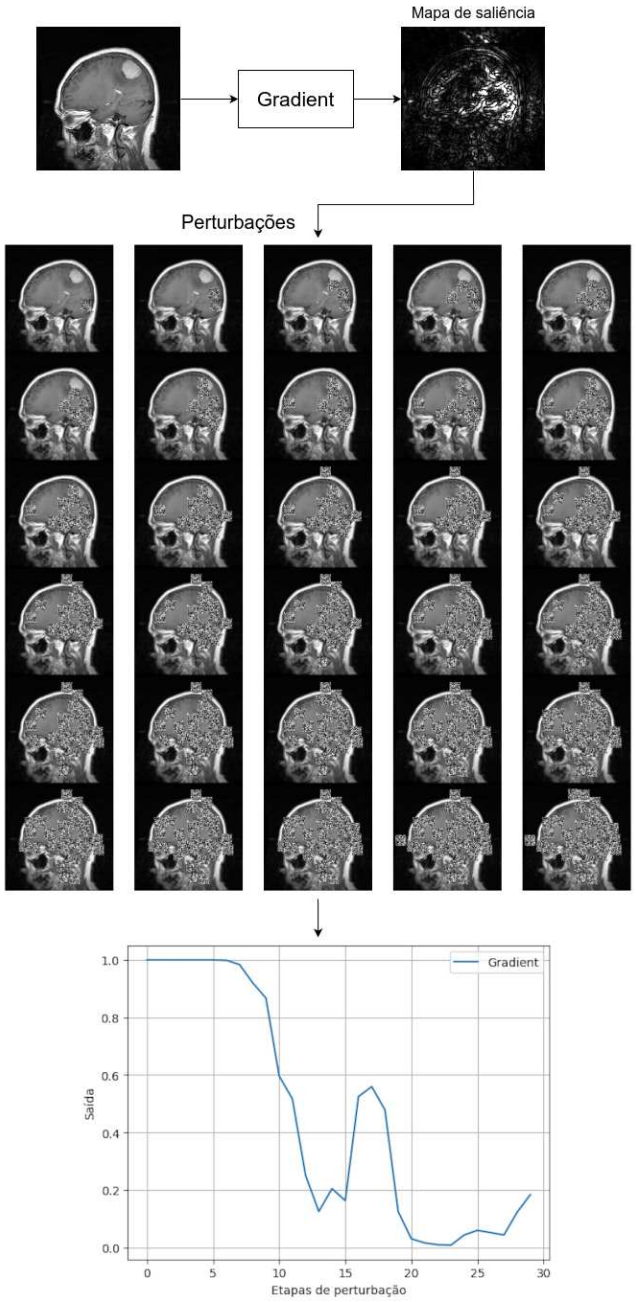

Figura 4: Diagrama do experimento para a avaliação das técnicas de interpretabilidade. A partir do mapa de saliência, 30 etapas de perturbações geram regiões com ruídos aleatórios na imagem, em seguida é observado, através de um gráfico, como as perturbações afetam a predição do modelo.

submetida às 11 técnicas de interpretabilidade estudadas, gerando 11 mapas de saliências por imagem. A Figura 8 apresenta detalhadamente os resultados para uma imagem de cada rótulo juntamente com seus mapas de saliência. Cada linha representa uma imagem

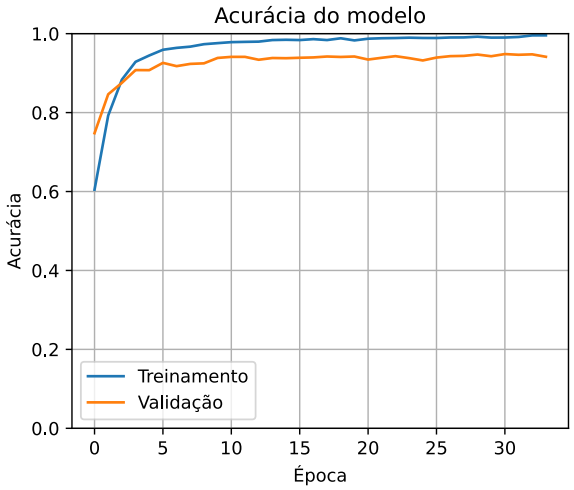

Figura 5: Gráfico de acurácia do modelo proposto na fase de treinamento ao longo das épocas.

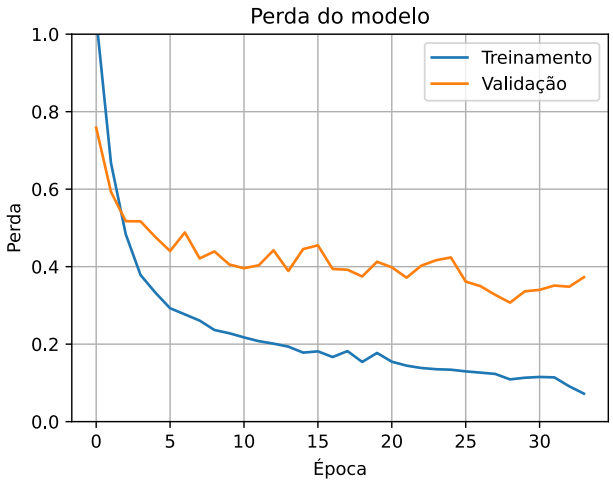

Figura 6: Gráfico de perda do modelo proposto na fase de treinamento ao longo das épocas.

escolhida, estando a imagem original na primeira coluna e os mapas nas demais. Na esquerda estão localizados os rótulos para cada imagem.

4.3 Avaliação das técnicas de interpretabilidade

Na Figura 9, estão apresentados os resultados da comparação das técnicas utilizando a avaliação por perturbações. O eixo das abscissas representa a quantidade de regiões perturbadas nas imagem, enquanto o eixo das ordenadas representa a saída do modelo para a classe de maior confiança. No gráfico, cada curva representa uma técnica diferente.

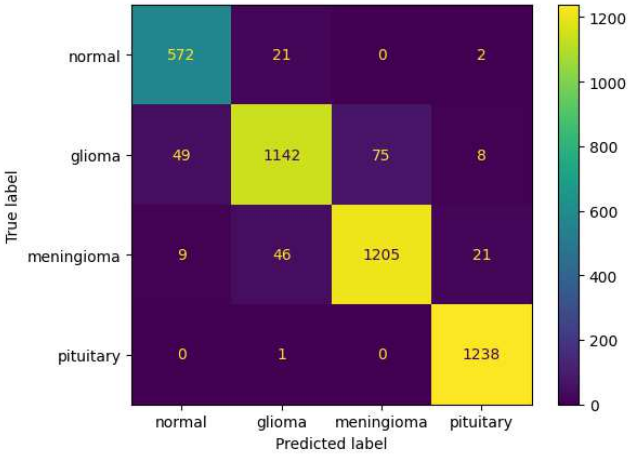

Figura 7: Matriz de confusão do conjunto de teste submetido ao modelo proposto.

As técnicas XRAI, GradCam e IntegratedGradients se destacaram negativamente na avaliação obtida. Enquanto as técnicas, RISE, Gradient e GuidedIG obtiveram os melhores resultados. Um destaque maior se dá à técnica RISE, que obteve uma pontuação de área sob a curva média distante das demais técnicas e ocupou o primeiro lugar na avaliação.

À medida que as perturbações aumentam no eixo horizontal, a confiança média nas previsões tende a diminuir no eixo vertical. Podemos avaliar as técnicas comparativamente usando a métrica da área sob a curva, que é fornecida para cada técnica na Tabela 6. A ideia é que quanto mais acentuada for a queda na confiança das previsões, menor será a área sob a curva. Isso indica que a técnica é melhor, pois as perturbações causadas nas áreas de maior interesse indicadas pela técnica têm um impacto negativo maior nas previsões do modelo.

Alguns dos problemas enfrentados durante esta avaliação foram a complexidade de implementação e custo computacional elevado, visto que, além de para cada imagem ser necessário o ranqueamento das regiões não sobrepostas mais importantes, também foi necessária a inferência do modelo para cada etapa de perturbação em todas as imagens, repetindo-se para cada técnica avaliada.

5 CONCLUSÃO

É possível observar que os métodos de interpretabilidade de atribuição de características utilizados funcionam, na maioria das vezes, como detectores de borda e do objeto de interesse, pois, além de destacar os tumores no mapa de saliência, também é visualizada toda a borda do cérebro. Dessa forma, pode ser valioso o estudo desses métodos para segmentação ou rotulação de *bounding boxes* de forma automática.

Apesar de uma das razões para se utilizar os métodos de interpretabilidade ser o aumento da confiança na classificação do modelo, nem sempre esses métodos são confiáveis, como observado no exemplo da Figura 8, onde na segunda imagem a técnica LIME destaca incorretamente o tumor, contrastando com a maioria das outras técnicas. Enquanto na primeira imagem a técnica RISE destaca incorretamente o tumor, mas as técnicas IGSmoothGrad, BlurIG, Input-Grad e Gradient destacam com precisão.

Tabela 6: Área sob a curva média da avaliação por perturbações de cada técnica (Quanto menor, melhor).

Técnica	Área abaixo da curva média
RISE	0,568
Gradient	0,647
GuidedIG	0,650
Input-Grad	0,662
SmoothGrad	0,674
IGSmoothGrad	0,686
BlurIG	0,693
LIME	0,704
IntegratedGradients	0,721
GradCam	0,722
XRAI	0,729

Além disso, como ameaça à validação do experimento de comparação das técnicas, pode-se considerar que as regiões com ruído aleatório introduzidas na imagem de entrada torna diferente a distribuição estatística das imagens de teste em relação à distribuição estatística das imagens que foram utilizadas no treinamento do modelo, o que pode ser uma das causas da queda da saída do modelo ao longo das perturbações.

6 AGRADECIMENTOS

Quero expressar minha profunda gratidão aos meus pais, cujo apoio inabalável foi a luz que guiou cada passo do meu caminho. À minha noiva, meu eterno agradecimento por sua constante inspiração e encorajamento, que dissiparam minhas dúvidas e fortaleceram minha determinação. Ao estimado professor Rohit Gheyi, sou imensamente grato pelo apoio e inúmeras oportunidades concedidas ao longo da minha graduação. Por fim, também quero agradecer ao professor Eanes Torres por sua valiosa orientação e por me introduzir ao mundo da pesquisa científica.

REFERÊNCIAS

[1] Sartaj Bhuvaji, Ankita Kadam, Prajakta Bhumkar, Sameer Dedge, and Swati Kanchan. 2020. Brain Tumor Classification (MRI). https://doi.org/10.34740/KAGGLE/DSV/1183165
[2] François Chollet et al. 2015. Keras. https://keras.io.
[3] Dumitru Erhan, Y. Bengio, Aaron Courville, and Pascal Vincent. 2009. Visualizing Higher-Layer Features of a Deep Network. *Technical Report, Univeristé de Montréal* (01 2009).
[4] Martín Abadi et al. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. https://www.tensorflow.org/ Software available from tensorflow.org.
[5] Gary S. W. Goh, Sebastian Lapuschkin, Leander Weber, Wojciech Samek, and Alexander Binder. 2021. Understanding Integrated Gradients with SmoothTaylor for Deep Neural Network Attribution. In *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE. https://doi.org/10.1109/icpr48806.2021.9413242
[6] Travis Greene, Galit Shmueli, Soumya Ray, and Jan Fell. 2019. Adjusting to the GDPR: The Impact on Data Scientists and Behavioral Researchers. *Big Data* 7, 3 (2019), 140–162. https://doi.org/10.1089/big.2018.0176 arXiv:https://doi.org/10.1089/big.2018.0176 PMID: 31033336.
[7] Seyed Mohammad Hossein Hashemi. 2023. Crystal Clean: Brain Tumors MRI Dataset. https://doi.org/10.34740/KAGGLE/DS/3505991
[8] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe: Convolutional Architecture for Fast Feature Embedding. arXiv:cs.CV/1408.5093

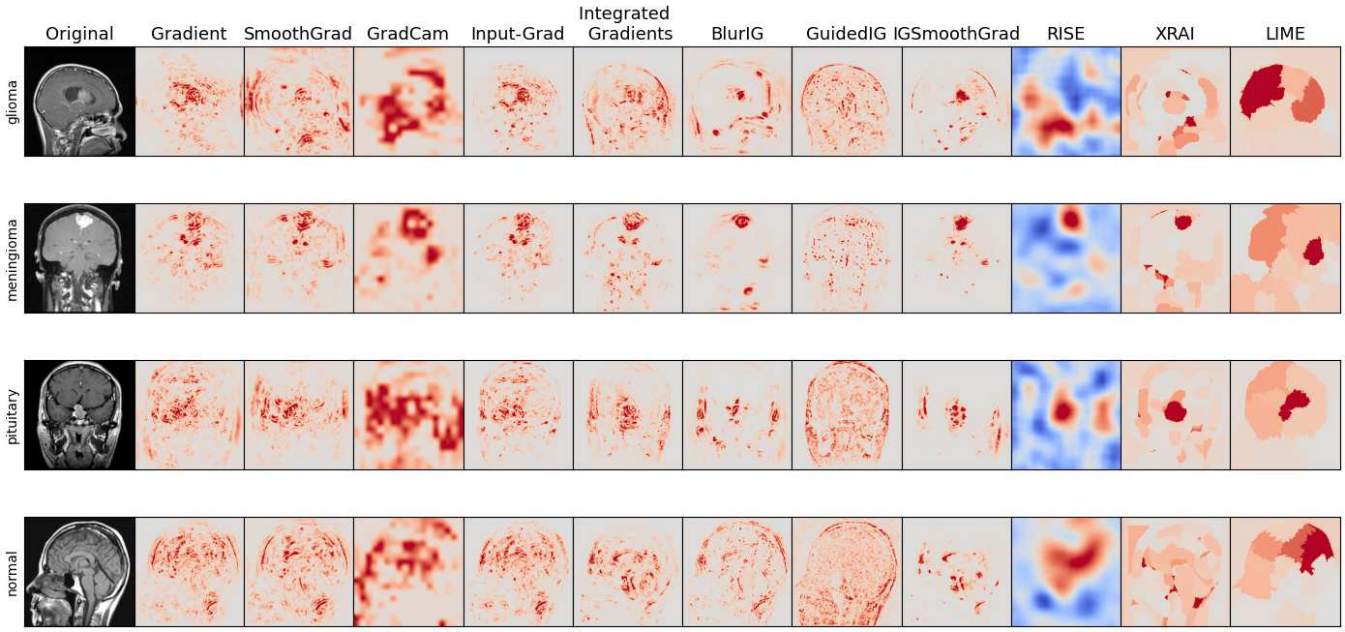

Figura 8: Aplicação de técnicas de interpretabilidade do tipo atribuição de características em imagens aleatórias do conjunto de teste. Cada linha indica uma imagem e os mapas de saliência gerados por todas as técnicas utilizando a imagem original como entrada. A legenda do lado esquerdo indica o rótulo da imagem.

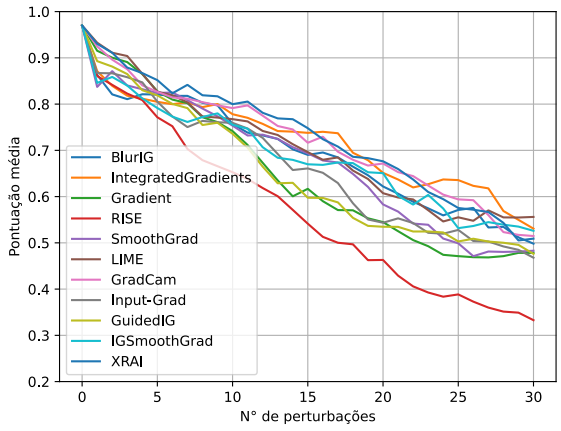

Figura 9: Avaliação das técnicas de interpretabilidades guiada por perturbações na imagem baseadas no mapa de calor de cada técnica (quanto mais cedo a queda da curva, melhor).

[9] Andrei Kapishnikov, Tolga Bolukbasi, Fernanda Viégas, and Michael Terry. 2019. XRAI: Better Attributions Through Regions. arXiv:cs.CV/1906.02825

[10] Andrei Kapishnikov, Subhashini Venugopalan, Besim Avci, Ben Wedin, Michael Terry, and Tolga Bolukbasi. 2021. Guided Integrated Gradients: An Adaptive Path Method for Removing Noise. arXiv:cs.CV/2106.09788

[11] Sebastian Lapuschkin, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLoS ONE* 10 (07 2015), e0130140. https://doi.org/10.1371/journal.pone.0130140

[12] Zewen Li, Wenjie Yang, Shouheng Peng, and Fan Liu. 2020. A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects. arXiv:cs.CV/2004.02806

[13] Scott Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. arXiv:cs.AI/1705.07874

[14] Anh Nguyen, Jason Yosinski, and Jeff Clune. 2019. Understanding Neural Networks via Feature Visualization: A survey. arXiv:cs.LG/1904.08939

[15] Vitali Petsiuk, Abir Das, and Kate Saenko. 2018. RISE: Randomized Input Sampling for Explanation of Black-box Models. arXiv:cs.CV/1806.07421

[16] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. arXiv:cs.LG/1602.04938

[17] Dominik Scherer, Andreas Müller, and Sven Behnke. 2010. Evaluation of Pooling Operations in Convolutional Architectures for Object Recognition. In *Artificial Neural Networks – ICANN 2010*, Konstantinos Diamantaras, Wlodek Duch, and Lazaros S. Iliadis (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 92–101.

[18] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2019. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision* 128, 2 (Oct. 2019), 336–359. https://doi.org/10.1007/s11263-019-01228-7

[19] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. arXiv:cs.CV/1312.6034

[20] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. 2017. SmoothGrad: removing noise by adding noise. arXiv:cs.LG/1706.03825

[21] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research* 15, 56 (2014), 1929–1958. http://jmlr.org/papers/v15/srivastava14a.html

[22] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic Attribution for Deep Networks. arXiv:cs.LG/1703.01365

[23] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2014. Going Deeper with Convolutions. arXiv:cs.CV/1409.4842

[24] Shawn Xu, Subhashini Venugopalan, and Mukund Sundararajan. 2020. Attribution in Scale and Space. arXiv:cs.CV/2004.03383

[25] Yilun Zhou, Serena Booth, Marco Tulio Ribeiro, and Julie Shah. 2021. Do Feature Attribution Methods Correctly Attribute Features? arXiv:cs.LG/2104.14403