A Study on Class Activation Map Methods to Detect Masses in Mammography Images using Weakly Supervised Learning

Vicente Sampaio[1], **Filipe R. Cordeiro** [1]

[1]Visual Computing Lab, Department of Computing
Universidade Federal Rural de Pernambuco, Recife, Brazil

vicentegalencar@gmail.com, filipe.rolim@ufrpe.br

Abstract. *In the last years, weakly supervised models have aided in the mass detection using mammography images, decreasing the need for pixel-level annotations. Most of the models in literature are based on CAM activation maps, not exploring the use of other activation methods for detection in mammography images. This works presents a study of different state-of-the-art activation maps methods, applied to weakly supervised training in mammography images. In this study, we compare the methods CAM, GradCAM, GradCAM++, XGradCAM and LayerCAM, using the GMIC model to detect the presence of mass in mammography images. The evaluation is done using the dataset VinDr-Mammo, using the metrics Accuracy, TPR, FNR and FPPI. Results show that the use of different strategies of activation maps during training and test stages lead to an improvement of the model. With this strategy, we improve the results of the GMIC method, decreasing the FPPI value and increasing TPR.*

Resumo. *Nos últimos anos, modelos de aprendizado fracamente supervisionado têm auxiliado na detecção de lesões em imagens de mamografia, diminuindo a necessidade de anotações de pixels na imagem. A maioria dos modelos na literatura se baseia no uso de mapas de ativação CAM, não sendo explorado o uso de outros modelos de ativação para detecção em imagens de mamografia. Este trabalho apresenta um estudo do uso de diferentes métodos de mapas de ativação do estado da arte, aplicados para treinamento fracamente supervisionado em imagens de mamografia. Neste estudo, comparamos os métodos CAM, GradCAM, GradCAM++, XGradCAM e LayerCAM, utilizando a rede GMIC para detectar a presença de lesões em imagens de mamografia. A avaliação é feita utilizando a base VinDr-Mammo, utilizando as métricas de Acurácia, TPR, FNR e FPPI. Resultados mostram que o uso de diferentes estratégias de mapas de ativação nas etapas de treino e teste melhoram o resultado do modelo. Com isso, melhoramos os resultados do método GMIC, reduzindo o valor de FPPI e aumentando o valor de TPR.*

1. Introdução

O câncer de mama é o tipo mais frequente de câncer entre mulheres no mundo todo, sendo responsável pelo maior número de mortes relacionadas ao câncer no sexo feminino, correspondendo a cerca de 15.5% dos óbitos por câncer [1]. A detecção precoce tem um impacto importante no sucesso do tratamento do tumor, uma vez que o tratamento torna-se mais difícil em estágios avançados [2]. No entanto, a interpretação da mamografia digital

pode ser uma tarefa difícil até mesmo para um especialista, uma vez que a análise é afetada por diversos fatores, tais como a qualidade da imagem, experiência do radiologista, tipo de tecido e de lesão [3]. Portanto, o uso de ferramentas computacionais de auxílio ao diagnóstico (CAD) para detecção de lesões pode auxiliar radiologistas a localizar as lesões e definir suas regiões de fronteira, fornecendo uma ferramenta adicional ao médico e melhorando a precisão do diagnóstico em imagens de mamografia.

Abordagens computacionais baseadas em *Deep Learning*, utilizando Redes Convolutivas (*CNN - Convolutional Neural Networks*, em inglês), têm sido aplicadas com bastante sucesso em tarefas de classificação e segmentação de imagens na área médica [4, 5]. Em aplicações de diagnóstico de câncer, a interpretabilidade das imagens é conseguida através da localização de regiões da imagem que determinam a classe atribuída pela saída do modelo [6], auxiliando o diagnóstico médico. Apesar dos avanços na área de segmentação semântica em imagens médicas, as abordagens existentes ainda são bastante dependentes de bases de treino com grande quantidade de imagens e anotações de alta qualidade para que o treino do modelo de segmentação de imagens seja eficiente [7]. Contudo, a aquisição de bases grandes com anotações é um grande desafio na área médica, onde as anotações de lesões requerem o conhecimento especialista para anotar a localização das lesões nos mamogramas, o que torna o processo extremamente trabalhoso e custoso [8].

Dados esses desafios, a área de aprendizado fracamente supervisionado tem sido bastante estudada nos últimos anos [9, 10], explorando estratégias para extrair informações dos dados com anotações escarças ou anotações fracas [11].Apesar de haver níveis diferentes de aprendizagem fracamente supervisionada, neste estudo consideramos uma base de dados fracamente anotada como sendo aquela em que as imagens possuem anotação apenas referente à classe da imagem (normal ou com lesão), mas que não possui anotação referente à localização ou contorno da lesão. O uso dessa abordagem torna mais acessível o treinamento das redes convolutivas, tornando mais barato e viável a construção e treinamento em imagens de mamografia, pois tornam o treinamento menos dependente da anotação do especialista em relação à localização das lesões. Apesar dos avanços na área de aprendizado fracamente supervisionado, esse ainda é um problema em aberto e novos estudos buscam aproximar os resultados em relação aos métodos fortemente supervisionados.

Métodos de treinamento fracamente supervisionados existentes utilizam o método de mapa de ativação de classe (CAM - *Class Activation Map*, em inglês) [12] para encontrar lesões em imagens de mamografia digital. No entanto, novos métodos baseados em mapa de ativação têm sido propostos nos últimos anos e que não foram explorados no contexto de detecção de lesões em imagens de mamografia. Neste trabalho, é proposto o estudo da utilização de mapas de ativação de classe do estado na arte, para detecção de lesões, utilizando aprendizado fracamente supervisionado. Para o estudo, comparamos os métodos CAM [12], GradCAM [13], GradCAM++ [14], XGradCAM [15] e LayerCAM [16]. Os mapas de ativação são avaliados utilizando o modelo do estado da arte GMIC [17], utilizando a base de dados VinDr-Mamo [18]. As principais contribuições do trabalho são listadas abaixo:

- Estudo do impacto da utilização de diferentes métodos de mapa de ativação aplicados para aprendizado fracamente supervisionado em imagens de mamografia

digital;

- Análise de modelo de detecção de lesões na base VinDr-Mammo;
- Melhoria do modelo do GMIC utilizando diferentes mapas de ativação para treino e teste.

2. Trabalhos relacionados

Nos últimos anos, vários trabalhos têm sido propostos na área de aprendizado fracamente supervisionado para detecção de anomalias em imagens de mamografia digital [17]. Os principais modelos de detecção fracamente supervisionados aplicados em imagens de mamografia se baseiam no método CAM [12] para realizar a identificação da região de interesse.

Shen et al. [6] propõem o modelo GMIC, que utiliza um modelo de rede convolutiva utilizando características locais e globais da imagem. Esse modelo primeiro usa uma rede de baixa capacidade em toda a imagem para identificar as regiões mais informativas. Em seguida, é utilizada outra rede de maior capacidade para coletar detalhes das regiões escolhidas. Por fim, é utilizado um módulo de fusão que agrega informações globais e locais para fazer uma previsão. O modelo é treinado apenas com informação de classe da imagem e as regiões de interesse são obtidas utilizando o método CAM.

Liu et al. propõem o método GLAM, que é uma modificação do GMIC para realizar uma segmentação refinada usando apenas anotação a nível de imagem. A principal ideia do GLAM é selecionar regiões informativas (*patches*) e em seguida realizar a segmentação em regiões selecionadas. Nessa abordagem também é utilizado o método CAM para encontrar as regiões de interesse.

Liang et al. [19] propõem o uso de mapa de ativação CAM para substituir modelos antigos com modelos de atenção, utilizando também uma estratégia de auto-treinamento, que consiste em observar as saídas das camadas intermediárias do modelo. Bakalo et al. [20] utilizam uma abordagem de janela deslizante, utilizando uma rede VGG pré-treinada para identificar regiões de interesse para a classe do problema. Apesar do modelo funcionar bem para imagens menores, essa abordagem tem um custo computacional alto para grandes bases de dados com imagens de resolução alta e treinamento de modelos profundos.

Zhu et al [21] realizam a detecção da região de interesse através da geração de um mapa de características reduzido, obtido a partir das camadas de convolução e *max pooling*. A identificação da classe da imagem é obtida através do uso de aprendizado de múltiplas instâncias (MIL) [22].

Fora do contexto de imagens médicas, vários métodos de geração de mapas de ativação têm sido propostos para utilização em aprendizado fracamente supervisionado [13, 15, 14, 16]. No entanto, apenas o método CAM tem sido utilizado no domínio de mamografias digitais. Nosso trabalho realiza a análise de diferentes métodos baseados em CAM, propostos na literatura nos últimos anos, mostrando que o mapa de ativação é um fator de otimização importante no processo de aprendizado fracamente supervisionado.

3. Materiais e Métodos

3.1. Métodos de Mapas de Ativação

O objetivo da detecção de objeto utilizado aprendizado fracamente supervisionado (WSOD - *weakly supervised object detection*, em inglês) é realizar a identificação da região contendo o objeto da imagem dada apenas a classe da imagem, sem nenhum tipo de supervisão a nível de *pixel*. Abordagens de WSOD utilizam métodos baseados em mapas de ativação para gerar a região delimitante do objeto na imagem para os valores acima de um limiar definido [23]. A região obtida é então redimensionada para o tamanho original da imagem.

Métodos baseados em mapas de saliência têm sido propostos na literatura como forma tentar explicar a relação entre região da imagem observada pelo modelo e a classe presente na imagem [17]. Esses métodos ajudam tanto na explicabilidade dos modelos propostos quanto em problemas de aprendizado fracamente supervisionado. Métodos de saliência baseados em ativação, tal como o CAM [12] se baseiam em observar a ativação da camada final do modelo para identificar as regiões responsáveis pela ativação de cada classe. Métodos baseados em ativação têm sido propostos em tarefas de classificação em imagens médicas para auxiliar na interpretabilidade dos modelos utilizados [24, 25]. No contexto de aprendizado fracamente supervisionado aplicado para detecção de lesões em imagens de mamografia, apenas a utilização do modelo CAM tem sido investigado. No entanto, outras abordagens têm sido propostas na literatura nos últimos anos e são analisados neste trabalho.

Seja f uma rede convolutiva com um classificador e c a classe de interesse. Dada uma imagem x e uma camada convolutiva l_i, onde i é a *i-ésima* camada convolutiva de f, o mapa de ativação de classe (CAM) de x em relação à c é definido pela combinação linear do mapa de ativação l_i, como mostrado a seguir [26]:

$$CAM_c(x) = ReLU\left(\sum_{k=1}^{N_l} \alpha_k A_k\right), \tag{1}$$

onde N_l é o número de canais da camada convolutiva l_i, A_k é o *k-ésimo* canal de ativação, α_k é o peso indicando a importância da camada de ativação em relação à classe c. A função de ativação ReLU é aplicada para considerar apenas as características com influência positiva na classe de interesse. Para abordagens baseadas em CAM, normalmente é feito um redimensionamento no mapa de ativação para ser o mesmo do tamanho da imagem de entrada. Desta forma, a identificação da região de interesse pode ser realizada através da multiplicação do mapa de ativação pela imagem de entrada. Em redes convolutivas com camada de *pooling* de média global, os valores de α_k são os pesos da camada final de classificação [12]. A Figura 1 ilustra o processo de obtenção do mapa de ativação, onde os pesos da camada final da rede são utilizadas para gerar os mapas de ativação.

O método Grad-CAM [27] determina o coeficiente do mapa de ativação através do cálculo de média de gradientes em todos os neurônios de ativação do mapa. O Método Grad-CAM++ [14] é uma versão modificada do Grad-CAM, em que foca nas influências positivas dos neurônios, considerando derivadas de segunda ordem. O método XGrad-CAM [15] também é baseado no Grad-CAM, mas escala os gradientes utilizando as

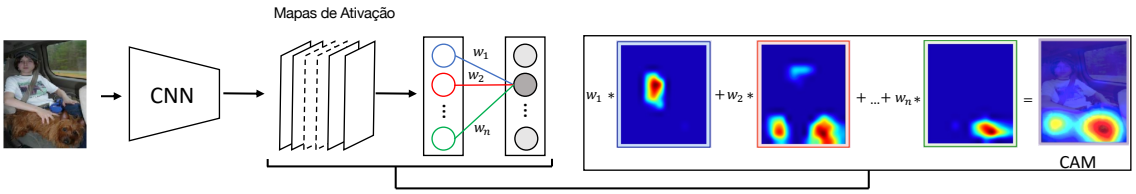


Figura 1. Mapa de Ativação CAM. Imagem adaptada de [12].

ativações normalizadas. O método LayerCAM [16] realiza uma combinação de mapas de ativação de diferentes camadas. Segundo os autores, as camadas iniciais ajudam a capturar melhor os detalhes de informação sobre o objeto (contornos e variação de contraste), enquanto as camadas mais profundas detectam a localização dos objetos de interesse.

3.2. Treinamento

Para realizar o estudo do trabalho, utilizamos a rede GMIC [17], que é o estado da arte em WSOD para detecção de lesões em imagens de mamografia. A Figura 2 mostra o funcionamento do modelo GMIC. O treinamento do modelo GMIC possui um modelo para extração de características globais, no módulo global, os quais utilizam o CAM para identificar as regiões de interesse. A partir das regiões encontradas, é utilizado um modelo de extração de características locais, no módulo local, para extrair o vetor de características para cada região. Por fim, o modelo é treinado a partir da junção das características locais e globais, no módulo de fusão.

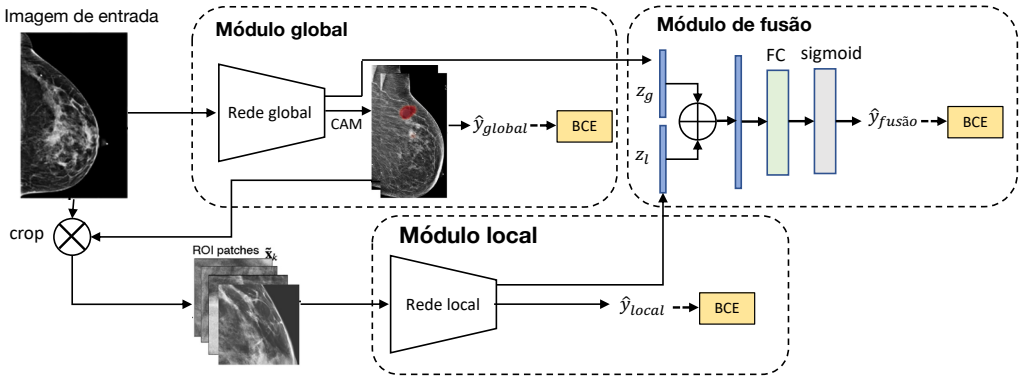


Figura 2. Modelo GMIC. Imagem adaptada de [17].

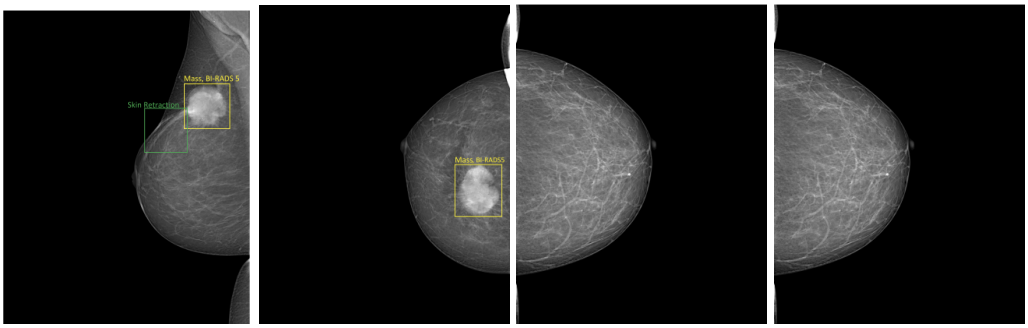
A função de perda do modelo GMIC é defina pela Equação 2 [17]:

$$L(y, \hat{y}) = \sum_c BCE(y^c, \hat{y}^c_{local}) + BCE(y^c, \hat{y}^c_{global}) + BCE(y^c, \hat{y}^c_{fusao}) + \beta L_{reg}(A^c), \quad (2)$$

onde BCE é a entropia cruzada binária, y^c é a saída esperada em relação à classe c, $\hat{y}^c_{local}$ é a saída observada para o modelo local, $\hat{y}^c_{global}$ é a saída observada pelo modelo global, $\hat{y}^c_{fusao}$ é a saída observada pelo modelo global após a fusão de características locais e globais, β é um coeficiente de regularização que utiliza o mapa de ativação A^c, segundo a função L_{reg}. A função de regularização L_{reg} é definida por $L_{reg} = \sum_{i,j} |A^c_{i,j}|$, onde i e j correspondem às linhas e colunas do mapa de ativação.

3.3. Ambiente Experimental

Para a avaliação do modelo, foi utilizada a base de dados pública VinDr-Mammo [18]. A base de dados consiste em 5000 exames de mamograma, onde cada exame possui 4 imagens associadas, com dois planos (médio-lateral e cranio caudal) para cada mama. As imagens da base foram obtidas através do método de mamografia totalmente digital (FFDM - *Full-field digital mammography*). A base VinDr-Mammo fornece informação de classe da anomalia (massa, calcificação, assimétrica, etc) e localização. Para o nosso trabalho, a informação de localização foi utilizada apenas para validar o modelo. Durante o treinamento foi utilizada apenas a classe da imagem. Neste trabalho foram consideradas duas classes: normal e *massa*. A classe normal indica que a imagem não possui nenhuma massa ou lesão, enquanto a classe *massa* indica a presença da lesão, que pode estar associada ao tumor. Para o conjunto de treinamento foram utilizadas 1978 imagens e para o conjunto de teste 474 imagens. Ambos os conjuntos são balanceados. A Figura 3 mostra imagens da base.



(a) MLO-D **(b)** CC-D **(c)** CC-E **(d)** MLO-E

Figura 3. Exemplo de imagens da base VinDr-Mamo. CC-D e CC-E significa visão Cranicaudal da mama direita e esquerda, respectivamente. MLO-D e MLO-E correspondem à visão mediolateral-oblíqua esquerda e direita, respectivamente. As imagens são retiradas de [18].

3.4. Implementação

As imagens utilizadas da base ViDr-Mamo foram redimensionadas para o tamanho 2944×1920 e para o conjunto de treino são aplicadas as operações de aumentado de dados básicas: *flip* horizontal, *random crop* e normalização, conforme utilizado em [17]. Para o conjunto de treino e teste foi uilizada a divisão disponibilizada pela base, fazendo a filtragem das imagens que continham a classe *massa* e *normal*. Para o treinamento do modelo GMIC, foi utilizado o modelo pré-treinado na base *NYU Breast Cancer Screening* [28] e feita a transferência de aprendizado para a base VinDr-Mammo. Para o modelo do GMIC, utilizamos uma rede ResNet-22 [29] para o modelo global e uma rede ResNet-18 [29] para o modelo local, conforme usado em [17]. O treinamento é feito utilizando 50 épocas para treino, valor de $\beta = 3.26$ e tamanho de *batch* igual a 6. Para o restante dos parâmetros do modelo foram utilizados os valores originais utilizados pelos autores. O código foi desenvolvido em Python, utilizando como base o código fonte disponibilizado no *github* dos autores.

Para os modelos de mapa de ativação, os modelos GradCAM, GradCAM++,

XGradCAM e LayerCAM, foi utilizado como base o código disponível em [30]. Para o modelo CAM foi utilizado o código original desenvolvido em [17].

3.5. Métricas de Avaliação

Para a avaliação do modelo foram utilizadas as métricas de acurácia, Área sob a curva ROC (AUC), Taxa de verdadeiros Positivos (TPR - *True Positive Rate*, em inglês), Taxa de verdadeiros Negativos (TNR- *True Negative Rate*, em inglês) e número de falso positivo por imagem (FPPI - *False Positive per Image*, em inglês).

Para a tarefa de classificação, nós utilizamos a área sob a curva ROC (AUC), TPR E TRN, conforme utilizado na literatura [31, 32]. A métrica de TPR diz a proporção da classe positiva que foi classificada corretamente, enquanto a TRN mostra a classificação correta em relação à classe negativa. As métricas TPR e TNR são representadas pelas equações 3 e 4 respectivamente:

$$TPR = \frac{VP}{VP + FN}, \tag{3}$$

$$TNR = \frac{VN}{VN + FP}, \tag{4}$$

onde VP, FN, VN e FP, são a quantidade verdadeiros positivos, falsos negativos, verdadeiros negativos e falsos positivos. Para a análise de classificação um verdadeiro positivo acontece quando a classe da imagem é de *massa* e o modelo realiza a predição corretamente.

Para a análise de detecção utilizamos as métricas TPR e FFPI, conforme utilizado na literatura [33, 34]. Para o modelo de detecção, uma predição de localização é considerada verdadeira positiva quando a interseção sobre união (IoU) das áreas da região predita com a região ouro é maior que 0.3. A métrica de FPPI mede a média de detecções falso-positivas por imagem. O esperado é que o modelo maximize a taxa de TPR e minimize a taxa de FPPI.

4. Resultados

Para a avaliação do mapas de ativação foram analisados dois cenários utilizando o modelo GMIC. No primeiro cenário, foi realizado o treinamento do GMIC original, em que é o utilizado o método CAM para obtenção das regiões de interesse durante o treinamento do modelo. Porém, durante a fase de teste foram analisados diferentes métodos de mapa de ativação para fazer a inferência da localização da região de interesse. No segundo cenário, é alterado o método de mapa de ativação durante o treino e também na etapa de teste. Para ambos os cenários, foram utilizados o mesmo conjunto de treino e teste da base VinDr-Mammo. Após o treinamento do modelo GMIC original, obtemos os valores de métricas mostrados na Tabela 1.

Todo o treinamento do modelo foi realizado utilizando apenas informação da classe da imagem (i.e. normal ou com *massa*). Essa análise foi feita para verificar a qualidade da classificação do modelo. Uma acurácia de 80% indica que o modelo consegue classificar corretamente a maioria das imagens. Além disso, essa é a primeira análise de um modelo fracamente supervisionado para a base VinDr-Mammo.

Tabela 1. Resultados do modelo GMIC treinados na base VinDr-Mammo para classificação das imagens nas classes Normal e Massa.

Modelo	Acurácia	AUC	TPR	FNR
GMIC	80.12	87.22	71.88	88.52

Para realizar a detecção da região de lesão foram utilizados os modelos CAM, GradCAM, GradCAM++, XGradCAM e LayerCAM. Para a avaliação da qualidade de detecção foram analisadas apenas as imagens de teste contendo *massa*. A Figura 4 mostra as segmentações obtidas por cada método para duas imagens da base de teste. A primeira coluna mostra a imagem original, com a marcação ouro da localização em verde. As colunas 2-6 refere-se aos modelos CAM, GradCAM, GracCAM++, XGradCAM e LayerCAM, respectivamente.

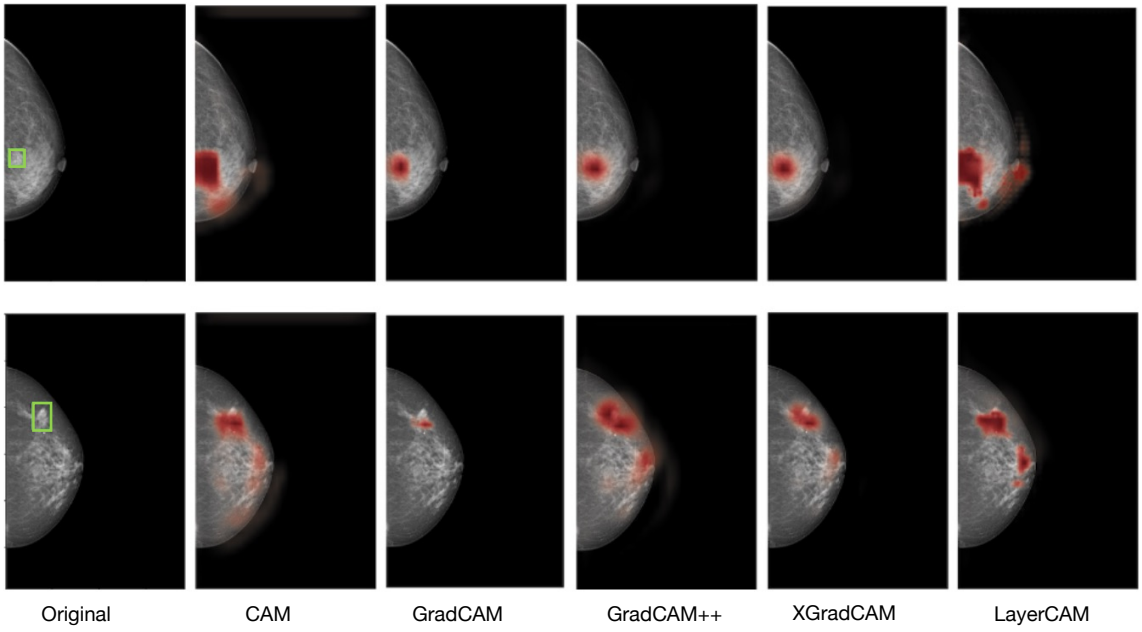


Figura 4. Segmentação das regiões de lesão utilizando diferentes métodos de mapas de ativação.

Pode-se observar pela Figura 4 que apesar de todos os métodos conseguirem encontrar a região de interesse associada à lesão, o método CAM tende a gerar mais falso positivos, contendo uma área maior de regiões segmentadas. No método GradCAM essa região é bem menor, mas em algumas imagens pode obter uma região menor do que esperada, em casos de lesões maiores. Apesar dos métodos GradCAM, GradCAM++ e XGradCAM apresentarem resultados bem próximos na Figura 4, quando analisado todo o conjunto de testes, há uma diferença maior entre os métodos analisados.

A Tabela 2 mostra os resultados de TPR@PFFI, que indicam a taxa de TPR com o valor de FPPI. Para essas métricas foram considerados os maiores valores de TPR obtidos. Nos resultados da tabela 2 analisamos diferentes cenários de treino e teste. Os modelos definidos nas linhas GMIC (CAM), GMIC (GradCAM++) e GMIC (XGradCAM) representam os resultados do modelo GMIC utilizando os métodos CAM, GradCAM++ e XGrad-

CAM durante o treino, respectivamente. Os valores das colunas representam os métodos de mapa de ativação utilizados durante o processo de inferência no teste. O modelo GMIC original corresponde à combinação GMIC(CAM)-CAM. Observando a primeira linha, vemos que o modelo original GMIC(CAM)-CAM possui a maior taxa de TPR, porém com um alto valor de FPPI. Quando mudamos o método de mapa de ativação, conseguimos reduzir a taxa de FPPI sem reduzir muito o TPR, como quando substituímos o CAM por XGradCAM. Quando treinamos o GMIC utilizando outros métodos de mapa de ativação para encontrar as regiões de interesse durante o treino, obtemos diferentes resultados na fase de teste. Para essa análise, a combinação GMIC (XgradCAM)-GradCAM++ foi a que geral os melhores resultados, considerando uma alta taxa de TPR e FPPI, comparado com os outros métodos.

Uma observação interessante deste estudo é que usar métodos diferentes nas fases de treino e teste pode levar a resultados melhores do que usar apenas um modelo para as duas etapas. Nós acreditamos que isso se deve ao fato de que na fase de treinamento é mais importante ter menores valores de detecção com falso positivo, tornando o modelo mais confiável no processo de obtenção de regiões para extrair características. No entanto, na fase de testes, utilizar um método que gera uma região maior de predição leva a um valor mais alto valor de TPR.

Além disso, conseguimos melhorar o desempenho do modelo GMIC através da substituição do método CAM durante o treino e utilização do modelo GradCAM++ durante o teste. Com isso conseguimos reduzir a taxa de FPPI de 1.55 para 0.88, aumentando a taxa de TPR.

Tabela 2. Resultados da detecção de Massas, representado por TPR@FPPI, utilizando diferentes métodos de mapa de ativação.

Método	CAM	GradCAM	GradCAM++	XGradCAM	LayerCAM
GMIC (CAM)	0.69@1.55	0.60@0.63	0.67@1.43	0.68@1.05	0.68@1.62
GMIC (GradCAM++)	0.71@2.89	0.13@0.06	0.65@1.12	0.13@0.06	0.67@1.59
GMIC (XGradCAM)	0.64@0.75	0.60@0.43	**0.70@0.88**	0.60@0.42	0.72@4.19

5. Conclusão

Este trabalho apresentou um estudo da utilização de diferentes métodos de mapas de ativação aplicados para detecção de lesões em mamografia digital, utilizando aprendizado fracamente supervisionado. O estudo mostrou como a utilização de diferentes métodos de ativação pode melhorar os resultados do modelo.

Nos experimentos realizados, concluímos que a estratégia utilizada de mapa de ativação influencia bastante nas taxas de verdadeiro positivo e de falso positivo por imagem do modelo. Também mostramos que a utilização de diferentes mapas de ativação nas fases de treino e teste leva a um melhor desempenho na inferência do modelo, ao invés de usar o mesmo método em ambas as fases. Com a substituição do método CAM no treinamento do modelo GMIC pelo método XGradCAm, e a substituição na fase de teste pelo método GraCAM++, conseguimos reduzir a taxa de FPPI, aumentando a taxa de FPI do modelo.

Em trabalhos futuros, iremos explorar o uso da detecção obtida para treinar o modelo de forma supervisionado. Também pretendemos explorar técnicas de anotações ruidosas para lidar com as detecções incorretas durante o treino.

6. Agradecimentos

Os autores agradecem à FACEPE (APQ-1046-1.03/21, BIC-0067-1.03/22) pelo incentivo financeiro e apoio às atividades de pesquisa.

Referências

[1] World Health Organization, "Breast cancer screening," 2017, disponivel em: http://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/, acessado em 08 dez. 2017.

[2] P. Autier, M. Boniol, R. Middleton, J.-F. Doré, C. Héry, T. Zheng, and A. Gavin, "Advanced breast cancer incidence following population-based mammographic screening," *Annals of Oncology*, pp. 600–633, 2011.

[3] D. S. Deshpande, A. M. Rajurkar, and R. M. Manthalkar, "Medical image analysis an attempt for mammogram classification using texture based association rule mining," in *Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG), 2013 Fourth National Conference on*. IEEE, 2013, pp. 1–5.

[4] M. H. Hesamian, W. Jia, X. He, and P. Kennedy, "Deep learning techniques for medical image segmentation: achievements and challenges," *Journal of digital imaging*, vol. 32, no. 4, pp. 582–596, 2019.

[5] S. S. Yadav and S. M. Jadhav, "Deep convolutional neural network based medical image classification for disease diagnosis," *Journal of Big Data*, vol. 6, no. 1, pp. 1–18, 2019.

[6] Y. Shen, N. Wu, J. Phang, J. Park, K. Liu, S. Tyagi, L. Heacock, S. G. Kim, L. Moy, K. Cho *et al.*, "An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization," *Medical image analysis*, vol. 68, p. 101908, 2021.

[7] S.-T. Tran, C.-H. Cheng, T.-T. Nguyen, M.-H. Le, and D.-G. Liu, "Tmd-unet: Triple-unet with multi-scale input features and dense skip connection for medical image segmentation," in *Healthcare*, vol. 9, no. 1. Multidisciplinary Digital Publishing Institute, 2021, p. 54.

[8] X. Xie, J. Chen, Y. Li, L. Shen, K. Ma, and Y. Zheng, "Instance-aware self-supervised learning for nuclei segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 341–350.

[9] A. Diba, V. Sharma, A. Pazandeh, H. Pirsiavash, and L. Van Gool, "Weakly supervised cascaded convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[10] X. Zhang, Y. Wei, J. Feng, Y. Yang, and T. S. Huang, "Adversarial complementary learning for weakly supervised object localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1325–1334.

[11] X. Ouyang, Z. Xue, Y. Zhan, X. S. Zhou, Q. Wang, Y. Zhou, Q. Wang, and J.-Z. Cheng, "Weakly supervised segmentation framework with uncertainty: A study on pneumothorax segmentation in chest x-ray," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 613–621.

[12] 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*, 2016, pp. 2921–2929.

[13] 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*, 2017, pp. 618–626.

[14] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *2018 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2018, pp. 839–847.

[15] R. Fu, Q. Hu, X. Dong, Y. Guo, Y. Gao, and B. Li, "Axiom-based grad-cam: Towards accurate visualization and explanation of cnns," 2020.

[16] P.-T. Jiang, C.-B. Zhang, Q. Hou, M.-M. Cheng, and Y. Wei, "Layercam: Exploring hierarchical class activation maps for localization," *IEEE Transactions on Image Processing*, vol. 30, pp. 5875–5888, 2021.

[17] Y. Shen, N. Wu, J. Phang, J. Park, K. Liu, S. Tyagi, L. Heacock, S. G. Kim, L. Moy, K. Cho *et al.*, "An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization," *Medical image analysis*, vol. 68, p. 101908, 2021.

[18] H. T. Nguyen, H. Q. Nguyen, H. H. Pham, K. Lam, L. T. Le, M. Dao, and V. Vu, "Vindr-mammo: A large-scale benchmark dataset for computer-aided diagnosis in full-field digital mammography," *medRxiv*, 2022. [Online]. Available: https://www.medrxiv.org/content/early/2022/03/10/2022.03.07.22272009

[19] G. Liang, X. Wang, Y. Zhang, and N. Jacobs, "Weakly-supervised self-training for breast cancer localization," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2020, pp. 1124–1127.

[20] R. Bakalo, R. Ben-Ari, and J. Goldberger, "Classification and detection in mammograms with weak supervision via dual branch deep neural net," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE, 2019, pp. 1905–1909.

[21] W. Zhu, Q. Lou, Y. S. Vang, and X. Xie, "Deep multi-instance networks with sparse label assignment for whole mammogram classification," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2017, pp. 603–611.

[22] M. Dundar, B. Krishnapuram, R. Rao, and G. Fung, "Multiple instance learning for computer aided diagnosis," *Advances in neural information processing systems*, vol. 19, 2006.

[23] Z. Qin, D. Kim, and T. Gedeon, "Neural network classifier as mutual information estimator," https://github.com/ZhenyueQin/Research-Softmax-with-Mutual-Information, 2021.

[24] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2097–2106.

[25] P. Rajpurkar, J. Irvin, R. L. Ball, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. P. Langlotz *et al.*, "Deep learning for chest radiograph diagnosis: A retrospective comparison of the chexnext algorithm to practicing radiologists," *PLoS medicine*, vol. 15, no. 11, p. e1002686, 2018.

[26] S. Poppi, M. Cornia, L. Baraldi, and R. Cucchiara, "Revisiting the evaluation of class activation mapping for explainability: A novel metric and experimental analysis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2299–2304.

[27] 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*, 2017, pp. 618–626.

[28] N. Wu, J. Phang, J. Park, Y. Shen, S. G. Kim, L. Heacock, L. Moy, K. Cho, and K. J. Geras, "The nyu breast cancer screening dataset v1. 0," *New York Univ., New York, NY, USA, Tech. Rep*, 2019.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[30] J. Gildenblat and contributors, "Pytorch library for cam methods," https://github.com/jacobgil/pytorch-grad-cam, 2021.

[31] D. Ribli, A. Horváth, Z. Unger, P. Pollner, and I. Csabai, "Detecting and classifying lesions in mammograms with deep learning," *Scientific reports*, vol. 8, no. 1, pp. 1–7, 2018.

[32] L. Shen, L. R. Margolies, J. H. Rothstein, E. Fluder, R. McBride, and W. Sieh, "Deep learning to improve breast cancer detection on screening mammography," *Scientific reports*, vol. 9, no. 1, pp. 1–12, 2019.

[33] H. Jung, B. Kim, I. Lee, M. Yoo, J. Lee, S. Ham, O. Woo, and J. Kang, "Detection of masses in mammograms using a one-stage object detector based on a deep convolutional neural network," *PloS one*, vol. 13, no. 9, p. e0203355, 2018.

[34] R. Agarwal, O. Díaz, M. H. Yap, X. Lladó, and R. Martí, "Deep learning for mass detection in full field digital mammograms," *Computers in biology and medicine*, vol. 121, p. 103774, 2020.