



Investigating the Impact of Attention on Mammogram Classification

Marc Berghouse^(✉) , George Bebis , and Alireza Tavakkoli 

University of Nevada, Reno, Reno, NV 89557, USA
{mberghouse, bebis, tavakkol}@unr.edu

Abstract. Attention, one of the most important features of modern CNNs, has been shown to improve the performance of mammogram classification, but our understanding of why attention offers improvements is rather limited. In this paper, we present the first comprehensive comparison of different combinations of baseline models and attention methods at multiple resolutions for whole mammogram image classification of masses and calcifications. Our findings indicate that attention generally helps to improve the baseline model scores, but the benefits are variable depending on the resolution and abnormality type. Furthermore, we find that pooling and overall model architecture (i.e., combination of baseline and attention) significantly impact mammogram classification scores. Specifically, scores are generally improved by architectural features that allow the model to retain as much information as possible while still focusing on relevant features. We also find that attention improves the correlation between model performance and LayerCAM activation in the region of interest. Our work provides insightful information to help guide the future construction of attention-based models for mammogram classification.

Keywords: Mammogram Classification · Attention · Deep Learning

1 Introduction

Attention has shown great promise in computer vision by guiding a model to focus on task-relevant local regions and channels [1]. In particular, it has been shown to increase performance in various classification [2], detection [3], and segmentation tasks [4]. Attention has also been shown to generally improve computer vision tasks in the field of medical imagery [2,5], where the region of interest often only comprises a small portion of the image.

Mammogram classification is an active area of research that has potentially life-saving consequences [6]. It has been extensively studied, and a wide variety of deep learning model architectures have been proposed for mammogram classification [7,8]. However, relatively little research has been performed on the impact of attention in mammogram classification. Although it is generally accepted that attention may improve model performance by focusing on relevant

features [2,9], there have not been any rigorous studies that confirm this in the context of mammogram classification. Although many studies have shown that attention improves classification scores over the respective non-attention “baseline” models [10,11], none of them have compared combining different attention methods with various baseline models. Thus, it is unclear what attention models work best, and if attention generally leads to improvements in classification regardless of the baseline model. Furthermore, no research has been performed on the impact of mammogram resolution or abnormality type on attention performance. Therefore, we don’t currently have sufficient understanding of how attention generally impacts whole image mammogram classification.

To better understand the impact of attention on whole image mammogram classification, we have performed extensive experiments with three baseline models and three attention methods (i.e., 12 distinct models in total). The rest of the paper is organized as follows: Section 2 data and methods, Sect. 2.1 data selection and preprocessing, Sect. 2.2 selection of models, Sect. 2.3 selection of attention models, Sect. 2.4 training and testing process, Sect. 3 results and discussion, Sect. 3.1 impact of attention on CNN performance, Sect. 3.2 impact of model architecture on performance differences, Sect. 3.3 impact of resolution on attention, Sect. 3.4 impact of abnormality type on attention, Sect. 3.5 relationship between model activation and AU-ROC, and Sect. 4 conclusions.

2 Data and Methods

2.1 Data Selection and Preprocessing

We used the CBIS-DDSM dataset [12] to analyze the impacts of attention on mammogram classification (benign or malignant). In order to understand how attention impacts model performance for different abnormality types, we trained and tested masses and calcifications separately. This dataset contains 1592 (1231 train, 361 test) images of masses and their respective segmented regions of interest and pathology, and 1513 (1227 train, 286 test) images of calcifications and their respective segmented regions of interest and pathology. Although CBIS-DDSM contains a relatively small amount of data, we chose this dataset because it uses an official train-test split, which we felt was appropriate for our rigorous analysis on the impacts of different attention methods.

Mammogram classification is highly dependent on image resolution where malignant masses or calcifications may be only a couple pixels wide [13]. To understand how the resolution of mammograms impacts attention methods, we tested models at resolutions of 500×300 and 1000×600 (height \times width) pixels. Furthermore, we trained and tested separately on masses and calcifications.

Images were preprocessed according to standard methods in classification of whole-image mammograms with deep learning [14]. The images were normalized, segmented, cropped, flipped and enhanced with CLAHE. The preprocessed images were then resized to the target size for the respective experiment (either 500×300 or 1000×600 pixels). For training, we used brightness, rotation, contrast, and flipping augmentations.

2.2 Selection of Models

To understand how attention influences the classification of mammograms, we first selected three baseline models (Resnet50 [15], Densenet169 [16], Regnetx64 [17]). For each baseline model family, we trained and tested three different model sizes (eg. Resnet38, Resnet50, and Resnet101), and selected the particular model that performed the best. We chose Resnet50 and Densenet169 due to their high popularity for mammogram classification [18,19]. Regnetx64 is an architecture developed by Meta which represents a more state of the art version of Resnet50. Regnetx64 uses a quantized linear function to determine the width and depth of each stage in the network, which the authors claim result in significant improvements [17]. All of our baseline models (as well as the attention methods) come from the Huggingface PyTorch Image Models (TIMM) repository. For each of these baselines, we considered three different attention methods, yielding 12 different models in total.

2.3 Selection of Attention Methods

The main criteria for our choice of attention methods were modularity and popularity. Based on initial experiments, we identified three top performing attention methods that could be easily integrated with our baseline models: Squeeze and Excitation (SE), Efficient Channel Attention (ECA), and the Convolutional Bottleneck Attention Module (CBAM) [20–22]. SE and ECA utilize channel attention to appropriately weight task-relevant channels (Fig. 2b). ECA was developed to address the issue of dimensionality reduction for attention modules such as SE and CBAM, and it has been shown to outperform both of these attention methods [21]. CBAM uses both channel and spatial attention. For each baseline and attention combination, we performed ablation studies at low resolution to determine the best placement of the attention module within the baseline model.

2.4 Training and Testing Process

All models were trained with fine-tuning. Specifically, the last layer of each model was removed and replaced with a dropout layer, a fully connected (FC) layer with 1624, 1664, or 2048 input nodes (based on width of baseline network) and 64 output nodes, leaky ReLU activation with a slope of 0.1, another FC layer with 64 input nodes and 1 output node, and sigmoid activation in the final layer to produce a binary prediction (malignant or benign). We used pre-trained weights from Imagenet to initialize the models. Although some baseline+attention combinations had pre-trained weights in the TIMM repository, we used the respective baseline model weights to initialize all attention models to keep the comparison fair. For Resnet50, we trained and tested two different model weights, since we found that the standard Timm weights significantly under-performed when compared to the imagenet1k_V2 weights for Resnet50 found in the PyTorch library. We experimented with freezing certain parts of the models for fine-tuning, but

found that it either resulted in no significant difference, or a slight decrease, in model scores. Therefore, no layers were frozen in our final experiments.

Hyperparameter tuning was done in a two step process. In the first step, we used Optuna to tune a large search space through Bayesian methods. These results narrowed the search space, then final tuning was done manually for each model. After manually tuning, we trained and tested each model within an approximate range of best hyperparameters over 40 times. We then used the top 30 scores for each model to calculate the results. The low resolution (500×300) models were primarily trained on a machine with an RTX 3070 (8 GB VRAM), and the high resolution (1000×600) models were all trained on a machine with an RTX 3090 (24 GB VRAM). We used a batch size of 10–16 (depending on VRAM constraints), the Adam optimizer, and cosine annealing with warm restarts for our scheduler.

3 Results and Discussion

3.1 Impact of Attention on CNN Performance

We have found that all attention methods generally improve baseline performance, but ECA and SE provide more consistent improvements than CBAM (Fig. 1a). By breaking the results down into model averages, we can observe CBAM yielded the largest performance increase for Densenet169, but it was the only attention method that decreased performance for Regnetx64 (Fig. 1b). Furthermore, all Densenet169 and Regnetx64 models generally outperform all Resnet50 models. We also found that all attention methods result in increased performance for Densenet169 and Resnet50, whereas only ECA and SE result in increased performance for Regnetx64, indicating that attention might be slightly less beneficial for Regnetx64. The only models that showed significant ($p < .05$) increases in AU-ROC over all model variations were Densenet169+CBAM and Densenet169+SE (Fig. 1b). Furthermore, the average differences between model AU-ROCs are relatively small. Thus, although choice of model architecture and attention module does matter, it should not have a large impact on results.

The number of trainable parameters is larger for Resnet50 and Regnetx64 than for Densenet169 (Table 1). Furthermore, the addition of CBAM and SE modules to Resnet50 and Regnetx64 leads to a larger increase in the number of parameters than when they are added to Densenet169. Thus, the poor performance of Resnet50, and Regnetx64+CBAM may be partially explained by the large number of parameters. Models with a large number of parameters have high complexity, and they are more likely to overfit than less complex models [23, 24].

The poor performance of CBAM may be additionally explained by the fact that it has shown inconsistent performance, resulting in decreased scores compared to the baseline for some fine-grained tasks [25]. Since mammogram classification is a highly fine-grained task [26], it is possible that CBAM is helpful for a smaller range of models, abnormality types, and input image resolutions than the other attention methods are. We thus recommend that rigorous testing

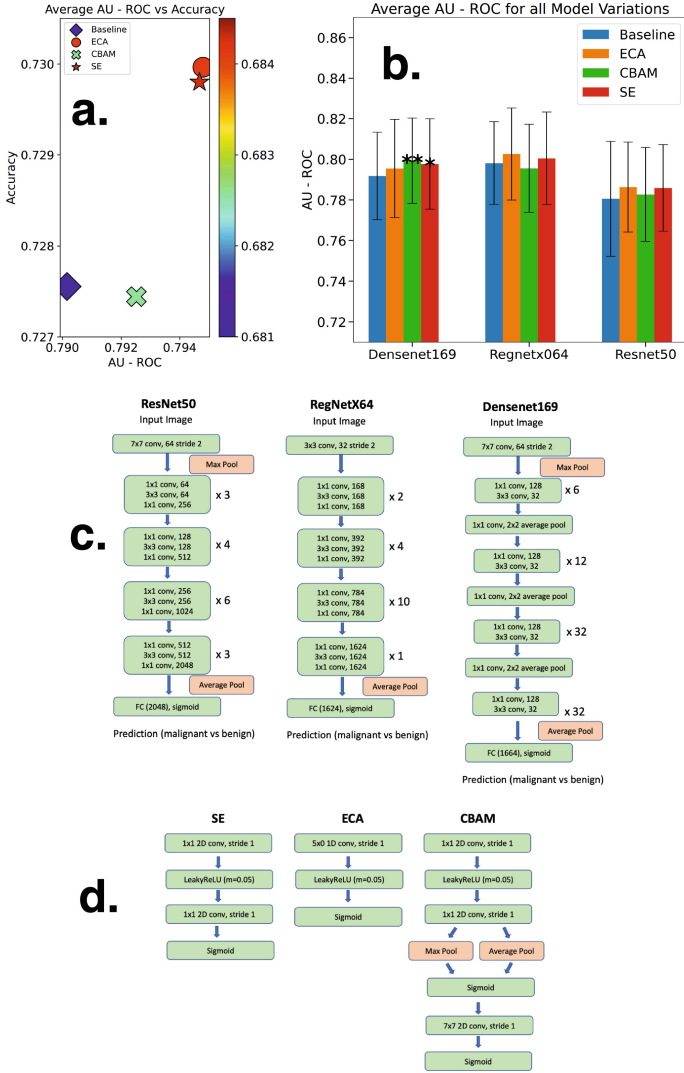


Fig. 1. Average results for all experiments, and architectures for baseline networks and attention methods. (a) Average AU-ROC vs accuracy with F1 colorbar. (b) Average AU-ROC for all model variations (network-specific breakdown of a). (c) Architecture comparison for Resnet50, Densenet169 and Regnetx64. (d) Architecture comparison for SE, ECA and CBAM attention modules. a&b show general increases in all scores due to attention. c&d are designed to highlight the differences between architectures used in this study, and are not a complete representation of the architectures.

be employed before using CBAM for mammogram classification (especially for calcifications).

In general, we found that CBAM was significantly harder to train than the baseline, SE, or ECA. CBAM took a significantly longer time to train than the other models, and the range of hyperparameters that produced good results was generally smaller (Table 1). However, the baseline models had the highest standard deviation between AU-ROC scores in different learning rate bins. Thus, attention results in less variation due to learning rate, which generally implies that models with attention will be less sensitive to hyperparameter choices. Similarly, Resnet50 was more difficult to train than Regnetx64 or Densenet169. Although Resnet50 trained faster than Densenet169, it was more prone to overfitting, and there was a smaller range of hyperparameters that would produce good results.

Table 1. This table contains various miscellaneous information that has been grouped together to save space. The first row gives the number of model parameters times 10^7 . The second row gives the average run time for each model. The third row gives the standard deviation of AU - ROC for each model between groups of learning rates.

DN 169	RG 64	RN 50	CBAM DN 169	CBAM RG 64	CBAM RN 50	ECA DN 169	ECA RG 64	ECA RN 50	SE DN 169	SE RG 64	SE RN 50
1.26	2.47	2.36	1.28	4.36	6.39	1.26	2.47	2.36	1.28	4.36	6.39
20.12	14.53	10.05	19.64	28.26	21.40	18.55	15.19	12.84	20.00	16.04	13.83
.0042	.0066	.0069	.0043	.0060	.0038	.0045	.0045	.0048	.0036	.0041	.0050

Figure 2 shows the AU-ROC for each model variation trained separately on two different resolutions and for two different abnormality types. Based on this figure, the most significant increases in AU - ROC score due to attention were for Regnetx64 and Resnet50 for the 300×500 resolution masses, and Densenet169 for the 600×1000 resolution masses and the 300×500 resolution calcifications. The best specific model performances were for Densenet169 with any attention method for the high resolution masses, and Regnetx64 with ECA for the high resolution masses. Figure 2 also reinforces the finding that ECA is the best attention method (greatest number of significant score increases compared to baseline), followed by SE then CBAM.

3.2 Impact of Model Architecture on Performance Differences

To further investigate these differences in performance, we analyzed the differences in model architecture between each baseline method (Fig. 1c). Densenet169 has a slower/more gradual increase in the number of feature maps (channels) than the other baseline networks do. Furthermore, all Densenet169 feature maps for each layer within a block are connected, meaning that no relevant information is lost within a block [16]. Densenet169 also ends up with less channels than

Resnet50 (1664 vs 2048). Regnetx64 has less changes in the number of channels than Resnet50, and also ends up with less channels (1624 vs 2048). Given the smaller growth rates of channel size and the similar final channel size for Densenet169 and Regnetx64, the relatively poor performance of Resnet50 might be due to its relatively wide architecture and/or bigger changes in number of channels between each layer. Although wide models may have a larger chance to overfit due to overparameterization, most research has shown increased performance for wider networks [27]. However, mammogram classification is an especially fine-grained task, so it’s possible that a slightly larger tendency to overfit makes wider models perform worse in mammogram-related tasks. Furthermore, wide networks take significantly longer to train than thin networks. Thus, the poor performance of Resnet50 may be due to its relatively wide layers that potentially lead to overfitting, or the requirement of a longer training time than what was used in our experiments.

For the attention methods (Fig 1d.), we can see that the architectures of ECA and SE are relatively similar. The only difference is that SE has two 2D convolutional layers, and ECA only has one 1D convolutional layer. The first part of CBAM (channel attention) is nearly identical to the SE module. However, CBAM employs max and average pooling before sigmoid activation, which may restrict the predictive power of small features. After channel attention, the new feature map is passed to a spatial 7×7 2D convolutional filter followed by another sigmoid activation. This spatial convolution with a high kernel size further reinforces prediction of only the most salient features, reducing the input of small or less relevant features that may still be important for mammogram classification. Furthermore, spatial convolution has been shown to result in overfitting [28]. Thus, the relatively poor performance of CBAM may be partially due to its large spatial convolution since spatial convolution. However, we have also noted that CBAM was more difficult to train (due to larger training time and more sensitivity to hyperparameter choice), so we can’t confidently conclude that the spatial convolution component of CBAM will always provide poor performance relative to other attention methods. Furthermore, we did some minor testing of CBAM with only the channel attention module, but this didn’t result in a score increase over vanilla CBAM. More experimentation is needed to understand the exact shortcomings of CBAM (relative to SE and ECA), but we theorize that the explanation primarily lies with the max and average pooling, the large spatial convolution, or the difficulty in finding optimal hyperparameters for training.

Generally, our results seem to suggest that wider networks that use more pooling are likely to result in overfitting for mammogram classification. Although Densenet169 uses more pooling than any other architecture, the results are better than for Resnet50 because of the inter-connectedness of all feature maps, which ensures that small-but-important features are not lost. Thus, we posit that pooling is problematic insofar as it limits the amount of relevant information the model can use. If the underlying network architecture promotes strong connectivity between layers, then pooling does not seem to cause decreases in model performance for mammogram classification.

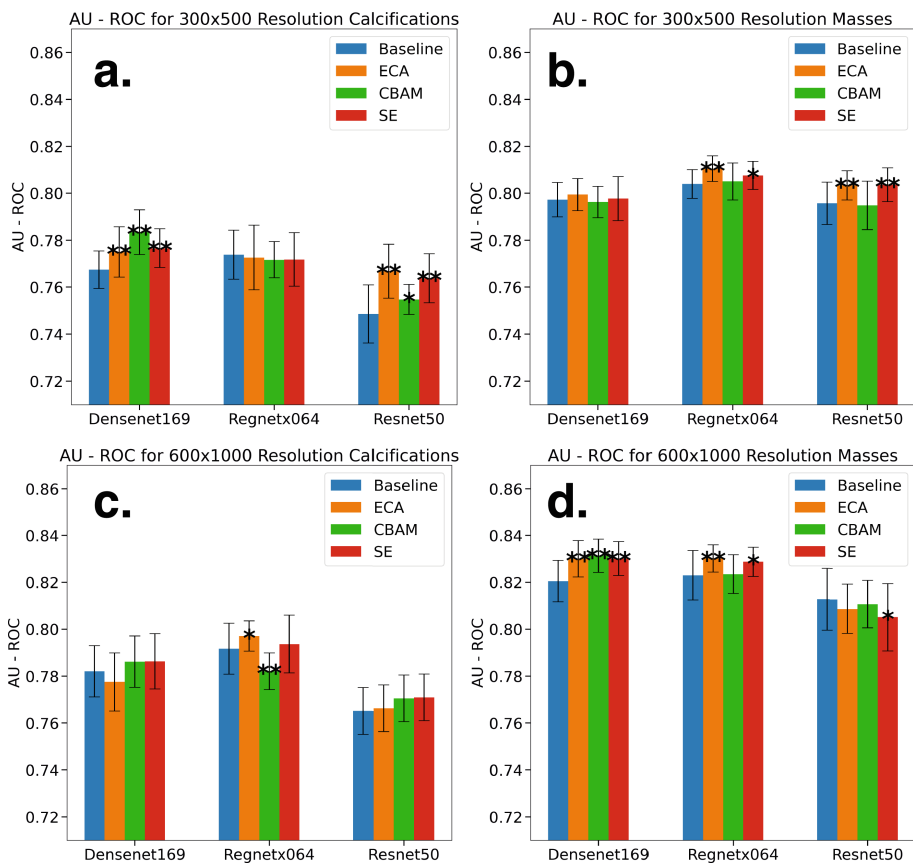


Fig. 2. Average AU-ROC for each model variation under each training/testing scenario for mammogram classification. One star indicates a significant ($p < .05$) difference from the baseline. Two stars indicate a very significant ($p < .01$) difference from the baseline. (a) Low resolution calcifications. (b) Low resolution masses. (c) High resolution calcifications. (d) High resolution masses. This figure generally shows improved performance due to attention, especially for ECA and SE.

3.3 Impact of Attention on Resolution

In addition to our general investigation of the impact of attention on mammogram classification, we also specifically investigated the impacts of attention on classification for different image resolutions. Our results indicate that attention has a greater impact at low resolution (Fig. 3a), although this is largely due to the much more significant increase in AU-ROC due to attention for Resnet50 at low resolution (Fig. 3b). Regnetx64, however, shows a slightly greater increase in AU - ROC when combined with ECA and SE at high resolution than it does at lower resolution (Fig. 3c).

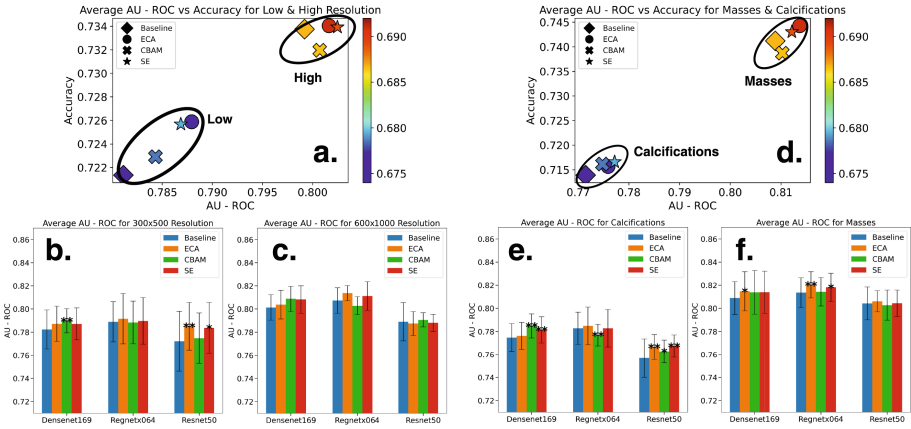


Fig. 3. Average scores of each model for each resolution. (a) Average AU-ROC vs. accuracy with F1 colorbar. (b) Average AU-ROC of each model variation for low resolution. (c) Average AU-ROC of each model for high resolution. a-c generally show higher scores at high resolution, and a higher impact of attention at low resolution. (d) Average AU-ROC vs. accuracy with F1 colorbar. (e) Average AU-ROC of each model variation for calcifications. (f) Average AU-ROC of each model for masses. d-f generally show higher scores for mass classification, and a higher impact of attention for calcifications.

One potential explanation for the relatively poor performance of Resnet50 is that the model is too wide, which resulted in overfitting. Although the results are not shown here due to lack of space, we did find that Resnet50 was more likely to result in overfitting than the other two baseline models. The addition of attention mechanisms helps to prevent this overfitting at low resolution. At high resolution, overfitting was generally less of an issue, so it’s possible that attention is not able to provide much further improvement. However, like CBAM, it is also plausible that the “best” set of hyperparameters were not well determined for Resnet50 at high resolution, especially since the Resnet50 baseline showed the highest standard deviation of score based on choice of learning rate (Table 1). Although significant effort was made to ensure the best hyperparameters for each model were used, given the large number of models tested, it is conceivable that some were not trained in a way to produce the best possible scores.

Besides looking at how attention improves scores at different resolutions, we can also observe some general trends in the impact of resolution on mammogram classification. Clearly, classification scores are generally better at a higher resolution. Furthermore, there is much less variability for all model scores at high resolution. This has important implications for mammogram classification, since images are often heavily downsized to fit in various models [11, 13].

3.4 Impact of Attention on Abnormality Type

We have also investigated the impact of attention on classification for different abnormality types. In general, we have found that SE and CBAM result in greater improvements in scores for calcifications than for masses (Fig. 3d). However CBAM also results in a large decrease in AU-ROC for calcifications when combined with Regnetx64. Thus, the performance of CBAM for calcifications is more reliant on baseline model architecture than for masses. ECA doesn't show any significant difference between scores for calcifications and masses. For Densenet169 and Resnet50, we found that attention results in greater score increases for the classification of calcifications than of masses (Figs. 3e and 3f). However, for Regnetx64, the opposite is true. One possible explanation for this trend is that Regnetx64 uses much less pooling than Densenet169 and Resnet50, meaning its architecture may be more favorable for small features such as calcifications. Thus, due to the constraints of resolution, attention is unable to offer as much of an impact in directing an already-good model to relevant features.

For calcification classification, the model scores may be significantly limited by the resolution of the image. Even at a resolution of 1000×600 , some calcifications may only comprise a couple pixels. Consequently, Resnet50 and Densenet169, which both use a large amount of pooling, do not see much of a performance gain between low and high resolution (Fig 3). Regnetx64, which uses less pooling, sees much more significant increases in scores at high resolution. These results suggest that even at high resolution, too much pooling can result in a loss of relevant information. Densenet169, which uses the largest amount of pooling, shows a relatively small increase in scores for calcifications going from low to high resolution, but the largest increase in scores for masses going from low to high resolution (Fig. 2). The large amount of pooling may allow the model to focus on more relevant features for masses, but with calcifications, the pooling causes too much of a loss of information, even with the interconnectedness of the dense layers. This loss of information likely doesn't significantly impact Densenet169 at low resolution because the small features that would be lost to pooling have already been lost during downsizing.

3.5 Relationship Between Model Activation and AU-ROC

To investigate the impact of attention on mammogram classification at a deeper level, we considered the relation between class activation maps and AU-ROC. Specifically, we used LayerCAM [29] to calculate class activation maps for each test image and for each model variation. We then binarized the resulting heatmap and calculated the Jaccard index, or IOU, between the binarized heatmap and the respective mask for each image. We used a simple ascending search to determine the threshold for heatmap binarization that results in the highest IOU score. We found that ECA generally has the most consistent improvement in IOU score over the baseline (9/12 increases), then SE (7/12 increases), then CBAM (2/12 increases) (results not shown due to lack of space). We also observed that

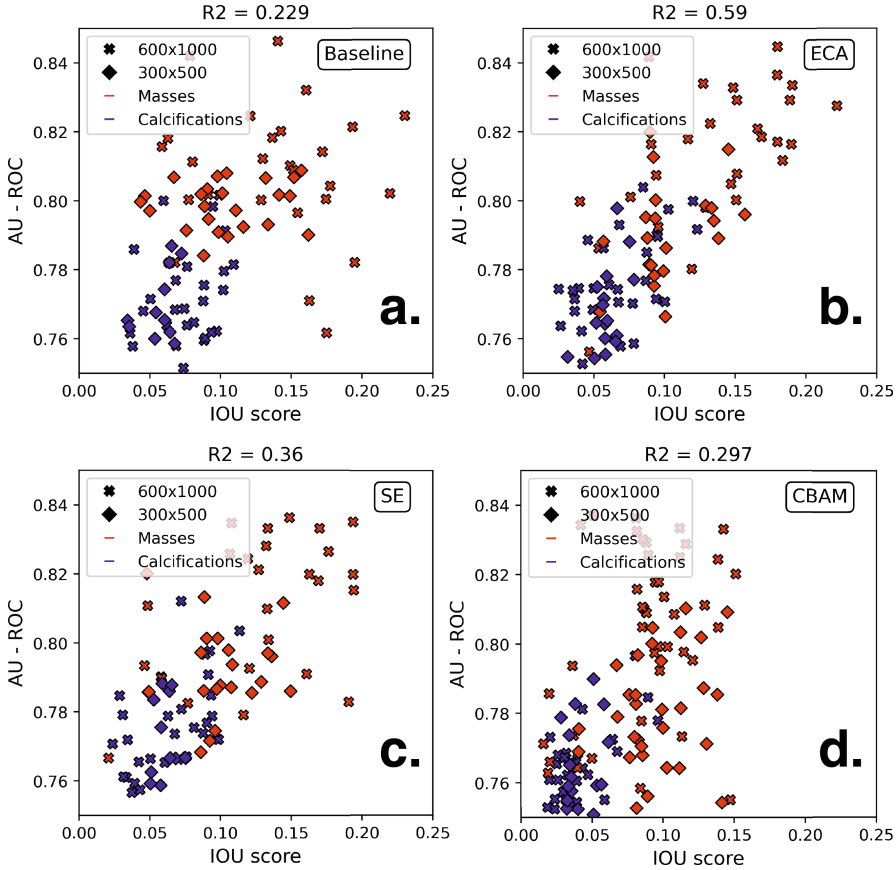


Fig. 4. Relationship between IOU and AU-ROC. Correlation coefficient (R^2) is given at the top of each graph. (a) All baseline models. (b) All models with ECA. (c) All models with SE. (d) All models with CBAM. This figure generally shows that attention methods provide a stronger correlation between IOU and AU-ROC, which indicates that attention helps the models focus on task-relevant features.

IOU scores increase at higher resolution, and are higher for masses than for calcifications.

We further used scatter plots and linear regression to more precisely identify this relationship (Fig. 4). Our results indicate that all attention methods show more correlation between IOU and AU-ROC than the baseline models (Fig. 4a), but that this relationship is much more significant for ECA (Fig. 4b). This implies that attention generally causes a given model to focus on more relevant features in mammograms. Moreover, mass classification in high resolution mammograms provides the greatest IOU scores. Although not shown here due to a lack of space, the correlation between AU-ROC and IOU score increases at high resolution for all attention models. Thus, at higher resolutions, attention should theoretically

provide a greater increase in scores. However, this is not what we have observed from resolution-averaged model performances (Fig. 3). Further investigation is required to determine the source of this contradiction.

4 Conclusions

Our work attempted to better understand how attention impacts mammogram classification. Our findings indicate that attention generally improves model performance. Furthermore, we found that ECA and SE significantly outperform CBAM. Between the baseline models, we found that Regnetx64 had the best performance, but may offer the lowest score increase due to attention. We also found that Densenet169 offers the highest increase due to attention, while Resnet50 gives the lowest scores. We theorize that differences in model architecture provide reasonable explanations for the differences we observed in model performance. Specifically, we found that wide architectures, and architectures with significant amounts of pooling, are more likely to overfit than thin models with less pooling. This impact was more noticeable for calcifications than for masses, which implies that the small features found in mammograms are often lost during pooling.

We have also presented evidence showing a correlation between IOU and AU-ROC scores. This correlation improves due to attention, with ECA showing the most significant correlation among the three attention methods examined. While others have used activation maps to show improvements due to attention, an IOU analysis providing an average score for all test images has not been presented. Our analysis shows that at a statistically relevant sample size, attention improves focus on task-relevant regions for mammogram classification. Future studies could leverage results from this study to develop a loss function that uses segmentation to make predictions for classification.

Our results could also be used to improve the design of future networks for mammogram classification and segmentation. Our findings indicate that model architecture and complexity are important considerations when designing networks for mammogram classification, as wide architectures and excessive pooling may result in overfitting. Furthermore, we have found that attention, especially ECA, results in significant performance increases for mammogram classification by helping the models to focus on task-relevant regions of the mammogram.

References

1. Guo, M.H., et al.: Attention mechanisms in computer vision: a survey. *Comp. Visual Media* **8**(5), 331–368 (2022)
2. Datta, S.K., Shaikh, M.A., Srihari, S.N., Gao, M.: Soft attention improves skin cancer classification performance. In: Reyes, M., et al. (eds.) *IMIMIC/TDA4MedicalData -2021*. LNCS, vol. 12929, pp. 13–23. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87444-5_2

3. Zhang, K., Wang, W., Lv, Z., Fan, Y., Song, Y.: Computer vision detection of foreign objects in coal processing using attention CNN. *Eng. Appl. Artif. Intell.* **102**, 104242 (2021)
4. Huang, Z., Wang, X., Huang, L., Huang, C., Wei, Y., Liu, W.: CCNet: criss-cross attention for semantic segmentation. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 603–612, Seoul, Korea (South) (2019)
5. Anaya-Isaza, A., Mera-Jiménez, L., Zequera-Diaz, M.: An overview of deep learning in medical imaging. *Inform. Med. Unlocked* **26**, 100723 (2021)
6. Hassan, N.M., Hamad, S., Mahar, K.: Mammogram breast cancer CAD systems for mass detection and classification: a review. *Multimed Tools Appl* **81**, 20043–20075 (2022)
7. Altan, G.: Deep learning-based mammogram classification for breast cancer. *Int. J. Intell. Syst. Appl. Eng.* **8**(4), 171–176 (2020)
8. Shen, L., Margolies, L.R., Rothstein, J.H., et al.: Deep learning to improve breast cancer detection on screening mammography. *Sci. Rep.* **9**, 12495 (2019)
9. Azad, R., Asadi-Aghbolaghi, M., Fathy, M., Escalera, S.: Attention Deeplabv3+: multi-level context attention mechanism for skin lesion segmentation. In: Bartoli, A., Fusiello, A. (eds.) *ECCV 2020. LNCS*, vol. 12535, pp. 251–266. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-66415-2_16
10. Lou, Q., Li, Y., Qian, Y., Lu, F., Ma, J.: Mammogram classification based on a novel convolutional neural network with efficient channel attention. *Comput. Biol. Med.* **150**, 106082 (2022)
11. Xu, C., Lou, M., Qi, Y., Wang, Y., Pi, J., Ma, Y.: Multi-Scale Attention-Guided Network for mammograms classification. *Biomed. Signal Process. Control* **68**, 102730 (2021)
12. Lee, R., et al.: A curated mammography data set for use in computer-aided detection and diagnosis research. *Sci Data* **4**, 170177 (2017)
13. Saunders, R.S., Jr., Baker, J.A., Delong, D.M., Johnson, J.P., Samei, E.: Does image quality matter? Impact of resolution and noise on mammographic task performance. *Med. Phys.* **34**(10), 3971–81 (2007)
14. Abdel-Nasser, M., Melendez, J., Moreno, A., Puig, D.: The impact of pixel resolution, integration scale, preprocessing, and feature normalization on texture analysis for mass classification in mammograms. *Int. J. Optics* **2016**, 1370259 (2016)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778 (2016)
16. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4700–4708 (2017)
17. Radosavovic, I., Kosaraju, R.P., Girshick, R., He, K., Dollár, P.: Designing network design spaces. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10425–10433, Seattle, WA, USA (2020)
18. Chen, Y., Zhang, Q., Wu, Y., Liu, B., Wang, M., Lin, Y.: Fine-tuning resnet for breast cancer classification from mammography. In: Wu, C.Q., Chyu, M.-C., Lloret, J., Li, X. (eds.) *ICHSE 2018. LNEE*, vol. 536, pp. 83–96. Springer, Singapore (2019). https://doi.org/10.1007/978-981-13-6837-0_7
19. Al-Antari, M.A., Al-Masni, M.A., Kim, T.S.: Deep learning computer-aided diagnosis for breast lesion in digital mammogram. *Adv. Exp. Med. Biol.* **1213**, 59–72 (2020)

20. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7132–7141, Salt Lake City, UT, USA (2018)
21. Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., Hu, Q.: ECA-Net: efficient channel attention for deep convolutional neural networks. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11531–11539, Seattle, WA, USA (2020)
22. Woo, S., Park, J., Lee, J.-Y., Kweon, I.S.: CBAM: convolutional block attention module. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11211, pp. 3–19. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01234-2_1
23. Kebria, P.M., et al.: Deep imitation learning: the impact of depth on policy performance. In: Cheng, L., Leung, A.C.S., Ozawa, S. (eds.) ICONIP 2018. LNCS, vol. 11301, pp. 172–181. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-04167-0_16
24. Rice, L., Wong, E., Kolter, Z.: Overfitting in adversarially robust deep learning. In International Conference on Machine Learning, pp. 8093–8104, PMLR (2020)
25. Li, Z., Gu, T., Li, B., Xu, W., He, X., Hui, X.: ConvNeXt-based fine-grained image classification and bilinear attention mechanism model. *Appl. Sci.* **12**(18), 9016 (2022)
26. Hang, W., Liu, Z., Hannun, A.: GlimpseNet: attentional methods for full-image mammogram diagnosis. *Proceedings* (2017)
27. Doimo, D., Glielmo, A., Goldt, S., Laio, A.: Redundant representations help generalization in wide neural networks. *Adv. Neural. Inf. Process. Syst.* **35**, 19659–19672 (2020)
28. Yu, S., Jia, S., Xu, C.: Convolutional neural networks for hyperspectral image classification. *Neurocomputing* **219**, 88–98 (2017)
29. Jiang, P.T., Zhang, C.B., Hou, Q., Cheng, M.M., Wei, Y.: LayerCAM: exploring hierarchical class activation maps for localization. *IEEE Trans. Image Process.* **30**, 5875–5888 (2021)