Studying the Effects of Self-Attention for Medical Image Analysis

Adrit Rao[1] Jongchan Park[2] Sanghyun Woo[3] Joon-Young Lee[4] Oliver Aalami[5]
[1]Palo Alto High School [2]Lunit Inc. [3]KAIST [4]Adobe Research [5]Stanford University
adrit.rao@gmail.com jcpark@lunit.io shwoo93@kaist.ac.kr jolee@adobe.com aalami@stanford.edu

Abstract

When the trained physician interprets medical images, they understand the clinical importance of visual features. By applying cognitive attention, they apply greater focus onto clinically relevant regions while disregarding unnecessary features. The use of computer vision to automate the classification of medical images is widely studied. However, the standard convolutional neural network (CNN) does not necessarily employ subconscious feature relevancy evaluation techniques similar to the trained medical specialist and evaluates features more generally. Self-attention mechanisms enable CNNs to focus more on semantically important regions or aggregated relevant context with long-range dependencies. By using attention, medical image analysis systems can potentially become more robust by focusing on more important clinical feature regions. In this paper, we provide a comprehensive comparison of various state-of-the-art self-attention mechanisms across multiple medical image analysis tasks. Through both quantitative and qualitative evaluations along with a clinical user-centric survey study, we aim to provide a deeper understanding of the effects of self-attention in medical computer vision tasks.

1. Introduction

The ability to leverage deep learning and computer vision-based techniques and methods for the automated, accurate, robust, and interpretable classification of medical images has been widely studied [24,25,31]. By doing so robustly, we can potentially increase diagnostic accuracy and increase screening efficiency and productivity [19]. When developing computer vision-based systems to aid physicians, it is important to design the underlying task to be as similar as possible to the medical specialists. Additionally, when deploying these systems, interpretability is critical for clinical decision-making. Therein lies the value in making computer vision systems perform computation similar to human cognition. However, computational systems do not necessarily perform similar to humans in terms of visual cognition and it is important to integrate this capability.

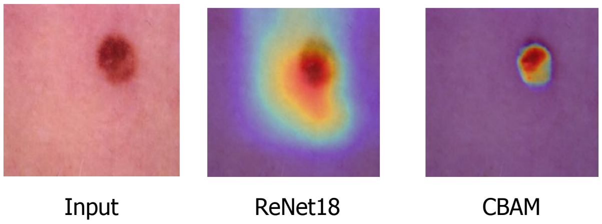


Figure 1: **Heat-map comparison between standard ResNet-18 and ResNet-18 + CBAM** on a benign skin image. The minimal addition of the visual attention mechanism (CBAM [34]) in the residual block (Sec. 4.2) enabled more focus onto the significant pigmented portion of the image making the final classification represent a region of more clinical importance while disregarding other features.

Cognitive Visual Attention One of the main cognitive capabilities which the trained domain specialist leverages is attention or "focus" [8]. Attention is the ability by which the human brain processes visual information while also evaluating the relevance of input features. This cognitive process allows for selective concentration onto a discrete stimulus while disregarding other perceivable stimuli [3]. As a general example, we can take a hypothetical scenario in which a person is viewing a landscape. Within this landscape is a certain object of interest. In order to effectively classify the object, the human brain applies this method of visual attention in order to use the majority of brain capacity to carry out the process on a certain region of interest (ROI). From a medical standpoint, in a chest CXR interpretation scenario, for example, the medical specialist will likely use attention subconsciously to automatically evaluate the clinical relevance of visual features with the advantage of also knowing the presenting symptoms of the patient and indication for the imaging study. With this, they are able to disregard unnecessary features (background, noise, etc.) and make a final diagnostic decision based solely on more relevant factors (abnormal lesions, opacity, etc.) [4]. This leads to a more robust final diagnosis. Replicating this capability within medical computer vision solutions is very important.

Visual Attention Mechanisms However, in comparison to the cognitive capability of attention, the standard and widely used convolutional neural network (CNN) classifier analyzes features more generally and is not guaranteed to extract relevant clinical information similar to the trained domain specialists subconscious [2]. The ability to artificially replicate or mimic this capability within neural networks can enable similar evaluation of clinical feature relevance in images which can potentially lead to more robustness in analysis and increased accuracy. A notable innovation in the computer vision field is the self-attention mechanism [7, 15, 16, 32–34]. Without any explicit supervision, these attention mechanisms learn to focus on important feature values in a data-driven manner. Self-attention mechanisms are trained end-to-end together with the original CNN backbone architectures without any changes in the training process. Thus, using self-attention mechanisms within the standard CNN can have many benefits on medical visual recognition tasks in terms of accuracy, interpretability, and robustness. Fig. 1 shows a visualization of activation heat-maps (Grad-CAM [30]) from a standard CNN (ResNet-18 [13]) and an attention-augmented CNN (ResNet-18 + CBAM [34]) on a skin cancer dataset sample. ResNet-18 is a widely used CNN architecture employed for many image classification tasks. The attention-augmented ResNet-18 has the sole modification of added visual attention mechanisms (CBAM) subsequently following convolution before the skip connection in the residual blocks. Notice how the addition of attention enables the model to focus fully on the important mole region while the standard CNN is less attentive to the mole and has more of a distributed attention.

This study aims to understand the value of self-attention mechanisms within standard computer vision models for approaching the ability to mimic the medical specialist's ability to evaluate the importance of features and their clinical relevancy. By doing so, we aim to understand the value of attention and how it can enable models to focus on more important clinical features. We present an experimental setting in which we use the standard ResNet-18 backbone [13] for performing multiple experiments across medical imaging datasets (multiple data modalities) by augmenting the residual blocks to accommodate 3 state-of-the art attention mechanisms as per original placement (CBAM [34], SE [15], GC [7]). Each of these mechanisms performs differently in terms of computation. We perform validation through a standard quantitative accuracy measure (average AUC-ROC across 3 tests) and qualitative heat-map visualizations. We notice significant increases in accuracy and through a visualization, we also observe significant changes in heat-map activation over more clinically relevant features. To further understand the benefits from a clinical standpoint, we have expert medical specialists (dermatologists and radiologists) interpret the visualizations and provide insight into which models and attention mechanisms focus on the most relevant clinical features and lesions.

2. Related Work

2.1. Medical Image Classification

Computer vision aided medical diagnosis is an important topic in the research community [24, 25, 31]. The ability to automate medical imaging diagnosis at the point-of-care in a robust fashion can lead to more accurate and objective clinical evaluation as well as increased screening efficiency and quality control [11]. Applications of computer vision include chest x-ray [27], CT scan [5, 12, 22], MRI classification [21, 26] and many more. Several systems can accurately classify images at near medical specialist precision. However, most vision-based standard CNN analysis methods do not necessarily employ similar evaluation techniques as the trained specialist. Specialists are trained over many years to understand the clinical relevance of features. Thus, while interpreting images, they are able to disregard features that are not important in making a targeted diagnosis. The ability to develop vision systems that are coherent with this capability is important. The main cognitive capability which the trained specialist employs is attention. Attention enables the ability to evaluate the importance (clinically) of input visual information and perform classification through analysis of a certain region of interest. Similarly, state-of-the-art self-attention mechanisms are able to mimic this capability to a certain extent computationally. Thus, there could be potential value in the integration of visual self-attention within medical computer vision systems to approach specialist evaluation, increase accuracy, and all around the robustness of clinical prediction systems.

2.2. Attention Mechanisms

Following the human visual perception process [9, 17, 28], the most intuitive way of modeling visual attention is to relatively scale up important information (i.e. pay attention) and scale down less important information. The initial works [15, 32, 34] of attention in visual recognition use attention maps to dynamically scale intermediate feature values in CNNs. One of the pioneering approaches is the Residual Attention Network (RAN) [32]. RAN uses an additional attention branch with downsampling convolutions and upsampling layers to generate the attention mask which is the same size of the intermediate feature map. This direct computation is simple and intuitive and improves baseline performance yet the computational cost is quite high. The Squeeze-and-Excitation (SE) [15] network is also a prominent approach that focuses on channel attention. For each given intermediate feature map, an SE mechanism generates a per-channel attention value from the global-average-pooled features. SE has been shown to improve performance with minimal overhead [15]. The Style-based Recalibration Module (SRM) [23] is a simple yet powerful channel attention module that accounts for channel statistics (mean and standard deviation) when scaling the channel values. After pooling the statistics, SRM uses a channel-wise fully connected (CFC) layer where each channel's attention weights are computed by a linear combination of the two statistics. The CFC layer in SRM is extremely efficient making both the computational and parametric overhead minimal. The Convolutional Block Attention Module (CBAM) [34] is a computationally efficient method that decomposes the heavy attention generation into separate dimensions. Specifically, while RAN directly generates full-sized attention maps, CBAM generates 2D spatial and 1D channel attention maps. CBAM has been shown to improve performance in various tasks consistently and reduce the overall computational overhead [34]. In the NLP field, the majority of self-attention mechanisms use attention maps to utilize the long-range dependencies among semantic tokens. Recent self-attention mechanisms in visual recognition models [7, 16, 33] are also equipped with such long-range dependencies. Non-local Neural Networks (NL) [33] is the first piece of work in the visual recognition field to model the long-range dependencies among spatial locations. Most of the previous methods computed attention with limited context, however, NL [33] uses the attention map to softly aggregate the information for all the points in the feature map. That is, all the relevant information in the entire feature map can be added to each individual point in the feature map. However, one of the drawbacks of NL [33] is the high computational cost, and to resolve this, CCNet [16] proposed to approximate the full attention process into separate cross-shaped processes. Aggregating two cross-shaped attention can effectively approximate the effect of NL [33]. Following CCNet [16], GCNet [7] also solved the high-computation issue of NL [33] by simplifying the NL block and inheriting the bottleneck structure of SE [15]. Through experimentation, GCNet achieves superior performance in comparison to NL in object detection, segmentation, classification, and action recognition tasks.

The objective of this study is to empirically verify the effectiveness of state-of-the-art self-attention mechanisms in various medical image analysis tasks. With attention mechanisms, neural networks can potentially start to approach the capability of understanding the clinical importance of features. We choose 3 self-attention mechanisms to compare: SE [15], CBAM [34], and GC [7]. SE is chosen because it is one of the most widely used self-attention mechanisms with the main focus on channel scale re-calibration, CBAM is chosen because it considers both channel and spatial dimensions, and GC is chosen because it is one of the strongest method with the non-local long-range dependency modeling. In the following sections, we will cover each method in detail and the results derived through validation.

3. Attention Mechanisms in Detail

3.1. General Formulation

Most of the self-attention mechanisms, including the three we compare in this study, are self-contained. *Self-contained* means that the inputs and the outputs of the mechanism are the same allowing them to be integrated at any location in a CNN architecture without modifying other parts. Any self-contained self-attention mechanism would fit into the following equation shown below (Eq. 1):

$$\begin{aligned} F^{l+1} &= SA(F^l) \\ F^l &\in \mathbb{R}^{C\times H\times W} \\ F^{l+1} &\in \mathbb{R}^{C\times H\times W} \end{aligned} \tag{1}$$

where F indicates the intermediate feature of a typical 2D CNN, l indicates the current layer index, and $\{C, H, W\}$ indicate the size of channel, height, and width respectively.

The self-attention allows the model to discover the most important task-relevant feature points. In practice, we dynamically compute the attention map of the feature map using its pooled [7, 15, 34] or raw features [33]. At each layer, SE [15] uses the global-average-pooled feature as the statistics of each input and computes the scale factor for each feature channel, CBAM [34] uses a similar manner with both global average and global maximum statistics in both channel and spatial dimensions and GC [7] computes a softmax attention map to aggregate the global statistics over all the spatial locations and then computes a context feature to be added to the input feature map. Technical details for each of the self-attention mechanisms used in this study will be elaborated in the following sub-sections (3.2, 3.3, 3.4).

3.2. Squeeze-and-Excitation [15]

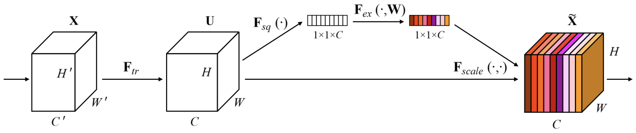


Figure 2: **Squeeze-and-Excitation (SE) network**. Architecture diagram taken from the original paper [15]

Hu *et al.* [15] proposed Squeeze-and-Excitation (SE), a self-attention mechanism focused on the channel dimension. SE is referred to as a 'channel recalibration' method. That is, each channel's magnitude can be explicitly tuned according to the values in other channels. As illustrated in Fig. 2, the first step of SE is to gather the global information in each channel through global average pooling (GAP). The term 'global' indicates that the spatial dimensions (height and width) are reduced and the method only utilizes one pooled value for each channel. The second step of SE is to use the globally-pooled feature vector and two consecutive fully-connected (dense) layers with one ReLU layer in between. The intermediate channel size is $\frac{1}{16}$ of the input channel size and last channel size is equal to the input channel size. The last step of the SE block is to apply sigmoid to the last feature vector and multiply it back to the original full feature map. Eq. 2 contains the described steps:

$$\begin{aligned} f_{ch} &= GAP(F) \\ a_{ch} &= \sigma(FC_{\frac{c}{r}\to c}(ReLU(FC_{c\to\frac{c}{r}}(f_{ch})))) \\ SA(F) &= F * a_{ch} \end{aligned} \tag{2}$$

where σ indicates the sigmoid function, GAP indicates the global average pooling function, r indicates the reduction ratio for the intermediate channel, and FC indicates fully-connected layers with input/output channels specified. Similar to Eq. 1, $F \in \mathbb{R}^{C\times H\times W}$, $f_{ch} \in \mathbb{R}^{C\times 1\times 1}$, and $a_{ch} \in \mathbb{R}^{C\times 1\times 1}$. The final multiplication shown in Eq. 2 is broadcasted along the spatial dimensions.

3.3. Convolutional Block Attention Module [34]

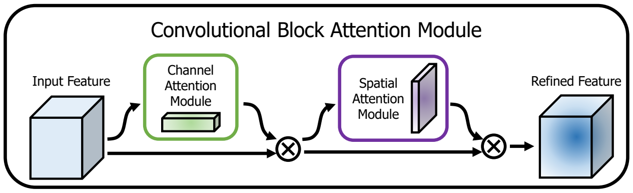


Figure 3: **Overall Convolutional Block Attention Module (CBAM)**. Diagram taken from the original paper [34].

Woo *et al*. [34] proposed Convolutional Block Attention Module (CBAM), a self-attention mechanism designed to make use of both the channel and spatial dimensions. The direct computation of a 3D attention tensor is quite heavy, similar to RAN [32], roughly doubling the overall computation. CBAM decomposes the 3D attention tensor into 1D channel attention and 2D spatial attention and applies them sequentially to the input feature in order to reduce the computational overhead of the attention mechanism. The design is illustrated in Fig. 3. SE utilizes the global average pooled statistics to calculate the attention weights while CBAM utilizes two statistics: global average and global maximum. The two statistics are experimentally shown to be complementary as using both statistics is better than using a single statistic. The channel and spatial attentions are sequentially applied to the input feature map F as shown in Eq. 3:

$$SA(F) = SA_{sp}(SA_{ch}(F)). \tag{3}$$

where SA_{ch} denotes the channel attention sub-module and SA_{sp} denotes the spatial attention sub-module which follows after the channel attention sub-module computation.

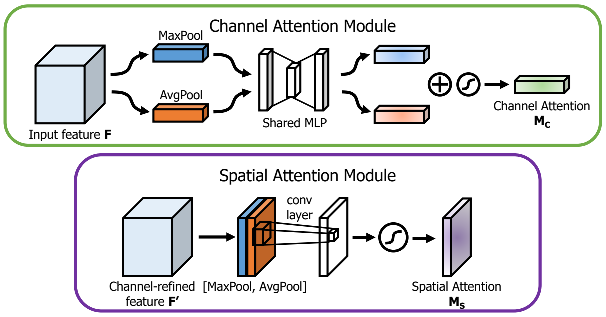


Figure 4: **Channel and spatial attention sub-modules**. Diagram taken from the original paper [34].

Channel attention module The structure of the channel attention module is illustrated in Fig. 4 (top). The attention tensor for the channel dimension is a 1D tensor. To efficiently calculate the 1D tensor, the global average and the global maximum values along each channel are pooled. After this, the 1D features are fed into a 2-layer multi-layered perceptron with a sigmoid layer at the end. The mathematical notation of the channel attention calculation is:

$$\begin{aligned} f_{ch-avg} &= GAP(F) \\ f_{ch-max} &= GMP(F) \\ SA_{ch}(F) &= \sigma(MLP(f_{ch-avg}) + MLP(f_{ch-max})) * F \end{aligned} \tag{4}$$

where σ denotes the sigmoid function, *MLP* is the 2-layer multi-layered perceptron with two fully-connected layers and one ReLU layer in between, f_{ch-avg} and f_{ch-max} are global average pooled / global max pooled features along the channel dimension, where f_{ch-avg} and $f_{ch-max}\in \mathbb{R}^{C\times 1\times 1}$. The final output of the channel attention module is the original 3D CNN feature multiplied by the 1D attention tensor with broadcasting along the spatial dimension.

Spatial attention module The structure of the spatial attention module is illustrated in Fig. 4 (bottom). The architecture follows the same structure as the channel attention module the only difference being the fact that the spatial attention module focuses on the spatial dimension. The mathematical notation for the spatial attention calculation is:

$$\begin{aligned} F_{ch} &= SA_{ch}(F) \\ f_{sp-avg} &= GAP_{sp}(F_{ch}) \\ f_{sp-max} &= GMP_{sp}(F_{ch}) \\ SA_{sp}(F_{ch}) &= \sigma(Conv_{7\times 7}([f_{sp-avg}, f_{sp-max}]) * F_{ch}. \end{aligned} \tag{5}$$

Note that the input to the spatial attention module is the output from the channel attention module, F_{ch}. As written in Eq. 5, the spatially avg/max pooled feature f_{sp-avg}

$f_{sp-max} \in \mathbb{R}^{1\times H\times W}$ are fed into a convolutional layer to compute the spatial attention tensor $M_s \in \mathbb{R}^{1\times H\times W}$.

3.4. Global Context Network [7]

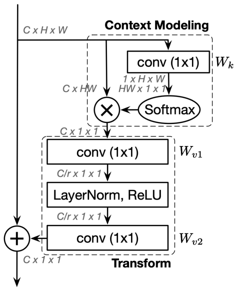


Figure 5: **Global Context Block architecture**. Diagram taken from the original paper [7]

.

Cao *et al.* [7] proposed GCNet (Global Context Network), an efficient yet powerful self-attention mechanism with context aggregation. Prior to GCNet, Wang *et al.* proposed Non-Local Neural Networks (NL) [33] for modeling contextual information for long-range dependencies. NL uses the key-query-value architecture of typical memory architectures and models the dense key-query affinity matrix between all pixel pairs in the feature map. All the values are aggregated using the affinity matrix thus, all the relevant information from all pixel locations are added to each pixel in the feature map. However, the dense affinity matrix modeling requires heavy computations and a large number of parameters to train. GCNet simplifies the architecture of NL and further reduces computation with the bottleneck architecture of SE [15]. The architecture is illustrated in Fig. 5. The forward path of GC is composed of the following steps:

$$SA_{gc}(F) = SA_{tfm}(SA_{agg}(F)) + F \tag{6}$$

where agg indicates context aggregation, tfm indicates feature transform for context information. As shown in Eq. 6, GCNet computes a context feature to be added back to the input feature map F. ctx stands for 'context'. As shown in Eq. 7, the context computation is composed of context aggregation and context transformation:

$$\begin{aligned} SA_{agg}(F) &= \text{Softmax}(\text{Conv}(F)) * F \\ SA_{tfm}(F) &= \underset{\frac{c}{r}\rightarrow c}{\text{Conv}}(\text{LN}(\underset{c\rightarrow\frac{c}{r}}{\text{Conv}}(F))). \end{aligned} \tag{7}$$

The first convolution in the context aggregation stage generates a softmax map to aggregate the feature values at each pixel locations. The resulting context vector is transformed with SA_{tfm} which has the bottleneck structure as SE [15]. The dense relationship modeling in NL [33] is reduced to a simple convolution layer and a softmax layer. After the context aggregation, a single global feature vector is computed and is transformed before adding back to the original input feature. One downside would be the lack of the location-specific context aggregation due to the simplification and all spatial locations will have the same context added back.

Throughout the paper, we strictly follow the original design and the hyper-parameters of each attention mechanism. However, we do not use the same backbone as the original papers. As all the attention mechanisms are of self-contained self-attention, they can be placed at any point within the backbone. Thus, to keep placement consistency, we position all mechanisms at the same location within a standard CNN backbone. Details are elaborated in Sec. 4.2.

4. Methods

In this section, we describe and cover the methodology used to develop our experimental setting environment. The goal of our study is to understand the value of self-attention mechanisms and the potential performance increases and robustness which it provides for medical computer vision systems. We first cover the different medical datasets we use (4.1) and then go over the model architectures and attention placement (4.2) with implementation details (4.3).

4.1. Datasets

We perform our study across 4 medical image datasets. This is done to provide a more robust and comprehensive comparison of self-attention across different data modalities (skin, CXR, MRI, CT). For training, we use a standard 80% training and 20% validation split. We perform a quantitative statistical evaluation across all datasets and a qualitative user study across the skin cancer and CXR datasets.

Skin Dataset The Skin Cancer Dataset consists of 3,297 processed skin images of mole lesions split into malignant (disease) and benign (normal) classes. The dataset was originally collected by the The International Skin Imaging Collaboration (ISIC) organization [1] and made open-source. Differentiating factors between classes are mainly visual feature differences in the pigmented mole lesions [18].

CXR Dataset The CXR (chest radiograph) dataset consists of 5,863 chest x-ray (anterior and posterior) images of normal and pneumonia classes from the open-source Chest X-Ray Images for Classification repository (UCSD) [20]. Differentiating factors between image classes include hazy shadowing in an pneumonia labeled CXR image [10].

MRI Dataset The MRI (magnetic resonance imaging) image dataset consists of 3,264 images of the human brain split into the classes of tumorous and no tumor from an open-source repository on Kaggle [29]. The main visual differentiating factor between the classes are the tumorous lesions which are typically circular and in a different shade compared to the other parts of the brain MRI [6].

CT Dataset The CT (computed tomography) dataset consists of 812 CT scan images spanning the classes of COVID-19 positive and negative. The dataset is from the open-source UCSD COVID-CT repository [35]. The main visual differentiating factor between the classes are the ground-glass opacity, vascular enlargement and white/hazy shadowing within a COVID-19 positive CT scan [14].

4.2. Model Architectures

Across all experiments, we use the ResNet-18 [13] architecture as the backbone. ResNet-18 acts as as good backbone architecture due to it being widely used in a multitude of classification tasks. Following the PyTorch implementation of ResNet-18 for ImageNet[1], we have made a minor modification on the final pooling layer and use a global average pooling instead of the fixed sized average pooling.

For the SE [15] implementation, we used a third party PyTorch implementation[2], for CBAM [34], we used the official PyTorch implementation[3] and for GC [7], we used the official PyTorch implementation[4]. As illustrated in Fig. 6, all the attention mechanisms (SE, CBAM, GC) are placed in each convolutional block in ResNet-18 right before the residual connection. The addition of attention within the ResBlock is the only difference between the standard and attention-augmented ResNet-18 model architectures.

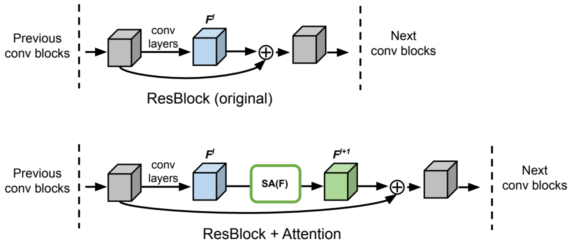


Figure 6: **Attention mechanism placement in ResBlocks.** Original ResBlock (top) and attention-augmented ResBlock (bottom) architectures. Placement consistency is maintained with a single modified ResBlock architecture.

4.3. Implementation Details

The models with and without the attention mechanisms are randomly initialized with *He* initialization. The models have 2 output logits for each input and are trained end-to-end with the cross entropy loss. The stochastic gradient descent (SGD) optimizer is used across all experiments with a momentum of 0.9 and weight decay of $1e^{-4}$. For all experiments, we use a batch size of 64 and a initial learning rate 0.05. The learning rate is decayed by 0.1 at 30%, 60% and 90% of the total epochs. For all datasets, we trained models for 100 epochs except the COVID-CT dataset which we trained across 500 epochs due to the limited sample count.

5. Experimental Results

In this section, we cover the experimental results derived through our validation of each model across the medical image datasets. The following text describes the quantitative and qualitative experiments conducted before presenting the results for each model and the corresponding datasets.

Quantitative In the quantitative experiments, we run each model variant across 3 individual tests and calculate the average area under the curve of receiver operating characteristic (AUC-ROC). AUC-ROC is a robust metric used to understand the performance of binary classification models. Averaging this metric across 3 tests can provide a more concrete reading which is not falsely elevated or deflated.

Qualitative For the qualitative experiments, we visualize Grad-CAM [30] activation heat-maps across all model variants on different dataset image samples. We then conduct a anonymized user study in which different medical specialists are asked to interpret the visualizations and provide clinical insight onto which model is focusing on the most clinically relevant region(s) of the image. Questions inquired about for the visualizations are: (1). *Which model focuses on the most important region?*, (2). *Explain the answer in condensed clinical terminology?*. We perform this study for both the skin and CXR datasets with radiologists and dermatologists in order to gain clinical understanding.

5.1. Skin Dataset

5.1.1 Quantitative

The quantitative results for the skin cancer dataset are reported in Table 1. Row 1 depicts the baseline performance of ResNet-18 [13]. ResNet-18 produced an average AUC-ROC of 93.28% across 3 tests. The SE [15] modified ResNet-18 model (Row 2) received an average AUC-ROC of 95.06% (+1.78 over ResNet-18). The CBAM [34] modified ResNet-18 model (Row 3) received an average AUC-ROC score of 95.09% (+1.81 over ResNet-18, +0.03 over

[1] https://github.com/pytorch/vision/blob/master/torchvision/models/resnet.py

[2] https://github.com/StickCui/PyTorch-SE-ResNet

[3] https://github.com/Jongchan/attention-module

[4] https://github.com/xvjiarui/GCNet

Model	Test 1	Test 2	Test 3	Mean AUC-ROC
ResNet-18 [13]	93.50%	93.77%	92.59%	93.28%
ResNet-18 + SE [15]	95.20%	**94.83**%	95.15%	95.06% (+1.78)
ResNet-18 + CBAM [34]	**95.28**%	94.73%	**95.26**%	**95.09**% (+1.81)
ResNet-18 + GC [7]	93.32%	93.86%	94.28%	93.82% (+0.54)

Table 1: Quantitative Skin Cancer Dataset Results.

Column (#)	Best Model	Description
1	ResNet-18 + CBAM [34]	”Malignant melanomas have asymmetrical shapes, irregular borders and changes in color”
2	ResNet-18 + CBAM [34]	”Malignant melanomas have asymmetrical shapes, irregular borders and changes in color + the heat map for model 2 for row 2 covers the greatest area of the actual lesion”
3	ResNet-18 + SE [15]	”Malignant melanomas have asymmetrical shapes, irregular borders and changes in color + the heat map for model 4 for Column 3 covers the greatest area of the actual lesion.”
4	ResNet-18 + GC [7]	”Heat map focuses on irregularly raised area as well as irregularly pigmented segments of lesion.”
5	ResNet-18 + CBAM [34]	”Malignant melanomas have asymmetrical shapes, irregular borders and changes in color + the heat map for model 2 for column 2 covers the greatest area of the actual lesion.”
6	ResNet-18 + CBAM [34]	”focuses on lesion. Model 4 looks great but focuses on border as well, which is not as favorable. Not clear how much that affects result.”

Table 2: **Skin dataset user-study survey results.** ’Best Model’ is the model which focuses on the most clinically relevant region and ’Description’ is a clinical explanation.

ResNet-18 + SE). The GC [7] modified ResNet-18 model (Row 4) received an average AUC-ROC of 93.82% (+0.54 over ResNet-18). The CBAM modified ResNet-18 model received the highest AUC-ROC score among the others. Additionally, through each individual test, CBAM and SE increase over baseline performance. Through this validation, we notice significant increases in AUC-ROC serving as a preliminary understanding of the benefits of attention.

5.1.2 Qualitative

We use 6 images from the skin dataset and visualize ground-truth activation heat-maps using Grad-CAM [30] for each model (baseline, SE model [15], CBAM model [34], GC model [7]). We also report the softmax probability percentage for each prediction label. In a preliminary nature, it is evident that attention significantly changes the activation heat-map and increases prediction probability (Fig. 7).

After generating the visualizations, we carried out the user study (Table 2). A trained dermatologist interpreted the visualizations and answered the survey questions. The first piece of information is the ”Best Model”. This is what model the clinician feels has the best and greatest amount of activation over a clinically relevant region. The first observation that can be made is that all models selected across all columns were never the baseline (ResNet-18) and the majority of models selected was ResNet-18 + CBAM. The dermatologist described that ”Malignant melanomas have asymmetrical shapes, irregular borders and changes in color” and that the models with attention ”covers the greatest area of the actual lesion”. Also mentioned was that models with attention focus on ”irregularly raised area as well as irregularly pigmented segments of lesion”. In summary, attention outperformed the baseline and ResNet-18 + CBAM had the highest AUC-ROC score and promising clinical user study results. The clinician also provided a summary in which they mentioned that self-attention ”covered larger area of lesion/mole including border to normal skin” and was ”not getting distracted by surrounding skin”.

Model	Test 1	Test 2	Test 3	Mean AUC-ROC
ResNet-18 [13]	98.57%	97.85%	95.40%	97.27%
ResNet-18 + SE [15]	98.93%	**99.15**%	**99.30**%	**99.13**% (+1.86)
ResNet-18 + CBAM [34]	**99.22**%	98.98%	99.16%	99.12% (+1.85)
ResNet-18 + GC [7]	98.16%	96.43%	96.01%	96.87% (-0.4)

Table 3: Quantitative CXR Dataset Results.

Column (#)	Best Model	Description
1	ResNet-18 + CBAM [34]	”Model is focusing on lung, highlighting unilateral patchy areas of consolidation, nodular opacities, bronchial wall thickening and pleural effusions, not highlighting normal appearing lung parenchyma”
2	ResNet-18 + CBAM [34]	”Model is focusing on lung, highlighting unilateral patchy areas of consolidation, nodular opacities, bronchial wall thickening and pleural effusions, not highlighting normal appearing lung parenchyma. Model does an excellent job here”
3	ResNet-18 + CBAM [34]	”Model has the best ratio of highlighting actual lung tissue vs. non-lung tissue”
4	None	”The area of clinical is the left upper lobe and none of the models do a great job of highlighting this are. Most models, except Model 2 highlight a large portion non-lung tissue”
5	ResNet-18 + CBAM [34]	”Does a fantastic job of focusing on right bronchial thickening and patchy areas in right lower and middle lobes”
6	ResNet-18 + CBAM [34]	”Model 2 is doing a great job at picking the windows between the ribs to identify the clear lung parenchyma. Still focusing a bit outside of the lung but not more than the other models”

Table 4: **CXR dataset user-study survey results.** ’Best Model’ is the model which focuses on the most clinically relevant region and ’Description’ is a clinical explanation.

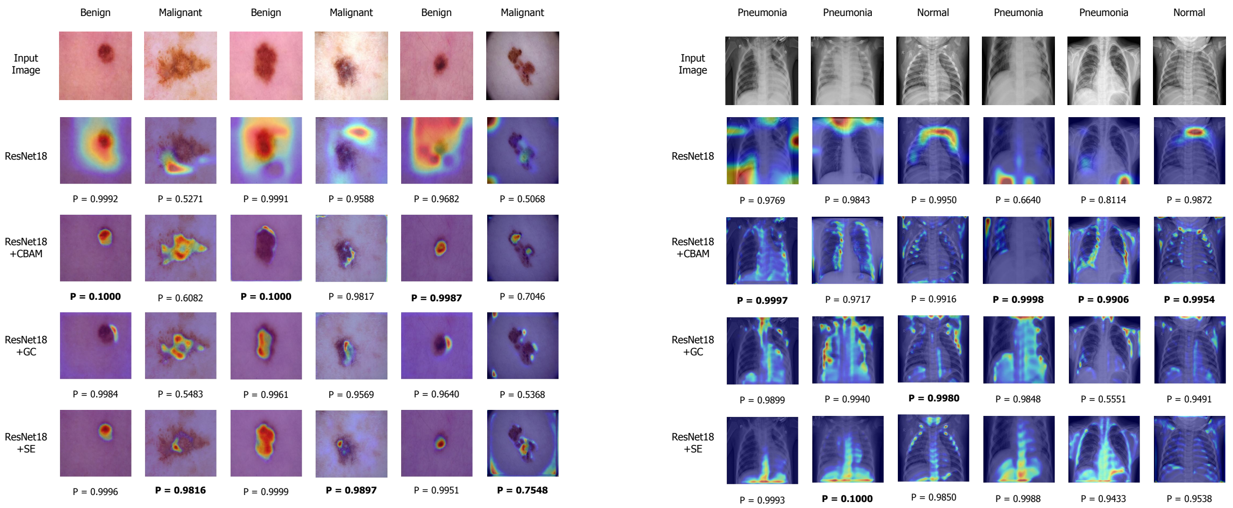
5.2. CXR Dataset

5.2.1 Quantitative

The quantitative results for the CXR dataset are reported in Table 3. The standard ResNet-18 model [13] (Row 1) received an average AUC-ROC of 97.27% across 3 tests. The SE [15] modified ResNet-18 model (Row 2) received an average AUC-ROC of 99.13% (+1.86 over ResNet-18). The CBAM [34] modified model (Row 3) received an average AUC-ROC of 99.12% (+1.85 over ResNet-18, -0.01 lower than ResNet-18 + SE). The GC [7] modified ResNet-18 model (Row 4) received an average AUC-ROC of 96.97% (-0.4 lower than ResNet-18). CBAM and SE performed higher than the baseline. This validation showed that self-attention still increases performance however, GC performed lower (-0.4 lower than ResNet-18) and was to be further investigated in the following dataset experiments.

5.2.2 Qualitative

Similar to the skin dataset qualitative results, the qualitative results for the CXR dataset are shown in Fig. 7. Again, as shown, attention increases classification probability and changes the heat-map in comparison to the ResNet-18 baseline. We can also see that the baseline is focusing on regions



(a) Skin Dataset

(b) CXR Dataset

Figure 7: **Grad-CAM [30] activation heat-map visualization from each model on the skin cancer and CXR datasets.** The visualization is generated from the last convolutional outputs. *P* denotes the softmax classification percentage for the ground-truth prediction. Notice differences in heat-map and prediction percentages between each model.

Model	Test 1	Test 2	Test 3	Mean AUC-ROC
ResNet-18 [13]	93.15%	92.49%	95.37%	93.68%
ResNet-18 + SE [15]	**98.43**%	96.29%	93.34%	96.03% (+2.35)
ResNet-18 + CBAM [34]	95.74%	**97.30**%	**97.25**%	**96.77**% (+3.09)
ResNet-18 + GC [7]	93.01%	87.44%	93.28%	91.25% (-2.43)

Table 5: Quantitative MRI Dataset Results.

Model	Test 1	Test 2	Test 3	Mean AUC-ROC
ResNet-18 [13]	87.79%	87.61%	78.99%	84.80%
ResNet-18 + SE [15]	85.43%	90.86%	88.13%	88.14% (+3.34)
ResNet-18 + CBAM [34]	**90.37**%	**93.25**%	**89.45**%	**91.02**% (+6.22)
ResNet-18 + GC [7]	84.68%	82.36%	85.66%	84.23% (-0.57)

Table 6: Quantitative CT Dataset Results.

outside of the lung while attention starts to move the heat-map closer to the more important lung regions. The user-study results for the dataset are shown in Table 4.

5.3. MRI Dataset

The quantitative results for the MRI dataset are reported in Table 5. The standard ResNet-18 model [13] received an average AUC-ROC of 93.68%. The SE [15] modified ResNet-18 model has a average AUC-ROC of 96.03 (+2.35 over ResNet-18). The CBAM [34] modified ResNet-18 model received an average AUC-ROC 96.77% (+3.09 over ResNet-18, +0.74 over ResNet-18 + SE). The GC [7] modified ResNet-18 model received an average AUC-ROC of 91.25 (-2.43 lower than ResNet-18). Through this validation, we again notice increases in performance with SE and CBAM however, GC performed poorly in comparison to ResNet-18 proving that not all self-attention mechanisms will have a positive impact on medical classification tasks.

5.4. CT Dataset

The quantitative results for the CT dataset are shown in Table 6. The standard ResNet-18 model [13] received an average AUC-ROC of 84.80% (Row 1). The SE [15] modified ResNet-18 model (Row 2) received an average AUC-ROC of 88.14% (+3.34 over ResNet-18). The CBAM [34] modified ResNet-18 (Row 3) received an average AUC-ROC of 91.02% (+6.22 over ResNet-18, +2.88 over ResNet-18 + SE). The GC [7] modified ResNet-18 (Row 4) received average AUC-ROC of 84.23% (-0.57 lower than ResNet-18).

6. Conclusion

In this paper, we evaluate various self-attention mechanisms within medical computer vision systems. Self-attention enables standard CNN models to focus more on semantically important or aggregated relevant content within features. The use of attention improved the AUC-ROC for medical vision task accuracy on dermatologic melanoma images, CXR images, brain MRI images and COVID-19 CT scans. Clinical user-study survey reviews conferred greater clinical agreement with feature focus of self-attention mechanism heat-maps. Further validation with other datasets and attention is required to further validate the improved accuracy trend observed with attention.

References

[1] Isic archive.

[2] S. Albawi, T. A. Mohammed, and S. Al-Zawi. Understanding of a convolutional neural network. In *2017 International Conference on Engineering and Technology (ICET)*, pages 1–6. Ieee, 2017.

[3] A. Allport. Visual attention. 1989.

[4] R. Bertram, J. Kaakinen, F. Bensch, L. Helle, E. Lantto, P. Niemi, and N. Lundbom. Eye movements of radiologists reflect expertise in ct study interpretation: A potential tool to measure resident development. *Radiology*, 281(3):805–815, 2016.

[5] A. Bhandary, G. A. Prabhu, V. Rajinikanth, K. P. Thanaraj, S. C. Satapathy, D. E. Robbins, C. Shasky, Y.-D. Zhang, J. M. R. Tavares, and N. S. M. Raja. Deep-learning framework to detect lung abnormality–a study with chest x-ray and lung ct scan images. *Pattern Recognition Letters*, 129:271–278, 2020.

[6] D. Bhattacharyya and T.-h. Kim. Brain tumor detection using mri image analysis. In *International Conference on Ubiquitous Computing and Multimedia Applications*, pages 307–314. Springer, 2011.

[7] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[8] M. M. Chun and J. M. Wolfe. Visual attention. *Blackwell handbook of perception*, 272310, 2001.

[9] G. Corbetta M., Shulman. Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience*, 3:201–215, 2002.

[10] D. Cozzi, M. Albanesi, E. Cavigli, C. Moroni, A. Bindi, S. Luvarà, S. Lucarini, S. Busoni, L. N. Mazzoni, and V. Miele. Chest x-ray in new coronavirus disease 2019 (covid-19) infection: findings and correlation with clinical outcome. *La radiologia medica*, 125:730–737, 2020.

[11] A. Esteva, K. Chou, S. Yeung, N. Naik, A. Madani, A. Mottaghi, Y. Liu, E. Topol, J. Dean, and R. Socher. Deep learning-enabled medical computer vision. *NPJ digital medicine*, 4(1):1–9, 2021.

[12] M. Grewal, M. M. Srivastava, P. Kumar, and S. Varadarajan. Radnet: Radiologist level accuracy using deep learning for hemorrhage detection in ct scans. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 281–284. IEEE, 2018.

[13] 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*, pages 770–778, 2016.

[14] X. He, X. Yang, S. Zhang, J. Zhao, Y. Zhang, E. Xing, and P. Xie. Sample-efficient deep learning for covid-19 diagnosis based on ct scans. *medrxiv*, 2020.

[15] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[16] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu. Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 603–612, 2019.

[17] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998.

[18] A. F. Jerant, J. T. Johnson, C. D. Sheridan, and T. J. Caffrey. Early detection and treatment of skin cancer. *American family physician*, 62(2):357–368, 2000.

[19] J. Ker, L. Wang, J. Rao, and T. Lim. Deep learning applications in medical image analysis. *Ieee Access*, 6:9375–9389, 2017.

[20] D. Kermany, K. Zhang, M. Goldbaum, et al. Labeled optical coherence tomography (oct) and chest x-ray images for classification. *Mendeley data*, 2(2), 2018.

[21] S. Korolev, A. Safiullin, M. Belyaev, and Y. Dodonova. Residual and plain convolutional neural networks for 3d brain mri classification. In *2017 IEEE 14th international symposium on biomedical imaging (ISBI 2017)*, pages 835–838. IEEE, 2017.

[22] S. Lakshmanaprabu, S. N. Mohanty, K. Shankar, N. Arunkumar, and G. Ramirez. Optimal deep learning model for classification of lung cancer on ct images. *Future Generation Computer Systems*, 92:374–382, 2019.

[23] H. Lee, H.-E. Kim, and H. Nam. Srm: A style-based recalibration module for convolutional neural networks. In *ICCV*, 2019.

[24] J.-G. Lee, S. Jun, Y.-W. Cho, H. Lee, G. B. Kim, J. B. Seo, and N. Kim. Deep learning in medical imaging: general overview. *Korean journal of radiology*, 18(4):570–584, 2017.

[25] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

[26] J. Liu, Y. Pan, M. Li, Z. Chen, L. Tang, C. Lu, and J. Wang. Applications of deep learning to mri images: A survey. *Big Data Mining and Analytics*, 1(1):1–18, 2018.

[27] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.

[28] R. A. Rensink. The dynamic representation of scenes. *Visual cognition*, 7(1-3):17–42, 2000.

[29] Sartaj. Brain tumor classification (mri), May 2020.

[30] 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*, pages 618–626, 2017.

[31] D. Shen, G. Wu, and H.-I. Suk. Deep learning in medical image analysis. *Annual review of biomedical engineering*, 19:221–248, 2017.

[32] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang. Residual attention network for image classification. In *Proceedings of the IEEE conference on*

computer vision and pattern recognition, pages 3156–3164, 2017.

[33] X. Wang, R. Girshick, A. Gupta, and K. He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018.

[34] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[35] X. Yang, X. He, J. Zhao, Y. Zhang, S. Zhang, and P. Xie. Covid-ct-dataset: a ct scan dataset about covid-19. *arXiv preprint arXiv:2003.13865*, 2020.