Computers, Materials & Continua
DOI: 10.32604/cmc.2023.032364
Article

Tech Science Press

Deep Attention Network for Pneumonia Detection Using Chest X-Ray Images

Sukhendra Singh[1], Sur Singh Rawat[2], Manoj Gupta[3], B. K. Tripathi[4], Faisal Alanzi[5], Arnab Majumdar[6], Pattaraporn Khuwuthyakorn[7] and Orawit Thinnukool[7,*]

[1]Department of Information Technology, JSS Academy of Technical Education, Noida, India
[2]JSS Academy of Technical Education, Noida, India
[3]Department of Electronics and Communication Engineering, JECRC University Jaipur, Rajasthan, India
[4]Harcourt Butler Technological University Kanpur, India
[5]Department of Electrical Engineering, Prince Sattam Bin Abdulaziz University, College of Engineering, Al Kharj, 16278, Saudi Arabia
[6]Faculty of Engineering, Imperial College London, London, SW7 2AZ, UK
[7]College of Arts, Media and Technology, Chiang Mai University, Chiang Mai, 50200, Thailand
*Corresponding Author: Orawit Thinnukool. Email: orawit.t@cmu.ac.th
Received: 15 May 2022; Accepted: 22 June 2022

Abstract: In computer vision, object recognition and image categorization have proven to be difficult challenges. They have, nevertheless, generated responses to a wide range of difficult issues from a variety of fields. Convolution Neural Networks (CNNs) have recently been identified as the most widely proposed deep learning (DL) algorithms in the literature. CNNs have unquestionably delivered cutting-edge achievements, particularly in the areas of image classification, speech recognition, and video processing. However, it has been noticed that the CNN-training assignment demands a large amount of data, which is in low supply, especially in the medical industry, and as a result, the training process takes longer. In this paper, we describe an attention-aware CNN architecture for classifying chest X-ray images to diagnose Pneumonia in order to address the aforementioned difficulties. Attention Modules provide attention-aware properties to the Attention Network. The attention-aware features of various modules alter as the layers become deeper. Using a bottom-up top-down feedforward structure, the feedforward and feedback attention processes are integrated into a single feedforward process inside each attention module. In the present work, a deep neural network (DNN) is combined with an attention mechanism to test the prediction of Pneumonia disease using chest X-ray pictures. To produce attention-aware features, the suggested network was built by merging channel and spatial attention modules in DNN architecture. With this network, we worked on a publicly available Kaggle chest X-ray dataset. Extensive testing was carried out to validate the suggested model. In the experimental results, we attained an accuracy of **95.47%** and an F- score of **0.92,** indicating that the suggested model outperformed against the baseline models.

This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Keywords: Attention network; image classification; object detection; residual networks; deep neural network

1 Introduction

Recent developments in artificial intelligent (AI) have improved the quality of services and people's experience in many diverse domains. Clinical Practice is one of them. Clinicians take assistance from these deep learning-based prediction tools which save their time, making their prediction much more reliable. Furthermore, these models can provide an accurate prediction of disease and interpret the severity of infection [1] in case of disease. Pneumonia is a disease that is caused by infections in the lungs and it is detected by examining chest X-ray radiographs. Pneumonia disease is one of the leading causes of death worldwide, especially among the younger and old age population. Pneumonia is an infection of the lungs caused by microbes, with inflammation as a result.

The inflammation causes the liquid to enter the lung tissue, making breathing more difficult. If in the acute case, it may cause death in the absence of proper timely medication. Pneumonia is detected by radiologists by examining a chest X-ray radiograph (CXR). If they are inconclusive about infection in CXR, then they recommend computer tomography (CT) scan of the lungs to examine finer details. Apart from examining the CXR, radiologists use computer-aided detection systems [2–4] which run on deep learning (DL) techniques which can recognize patterns in images as shown in Fig. 1. Such a system saves time and increases the efficiency of radiologists and it minimizes disagreement [5] among multiple radiologists in some cases. Fig. 2 shows samples of chest radiographs of a healthy person and the person suffering from Pneumonia.

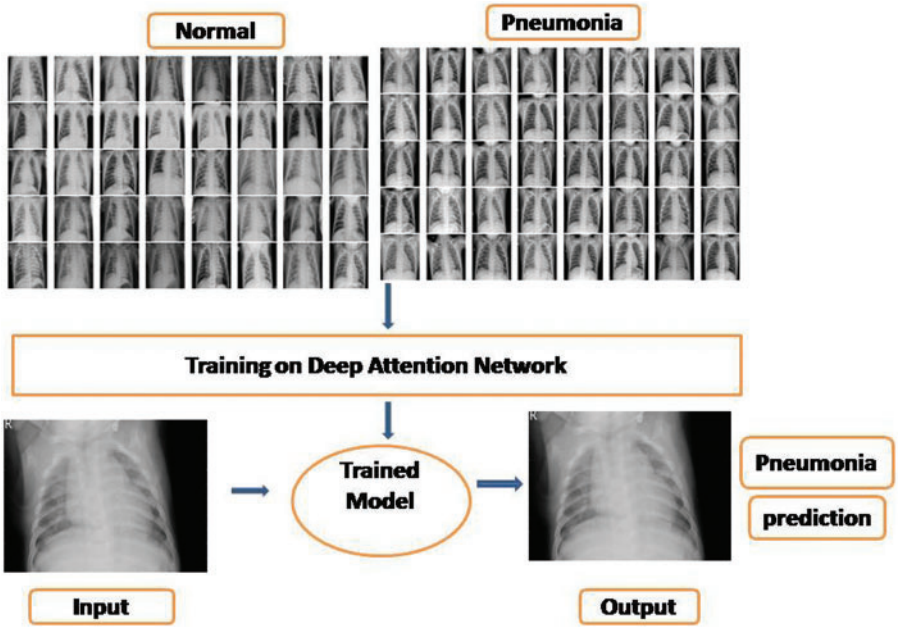

Figure 1: Proposed architecture' high level view

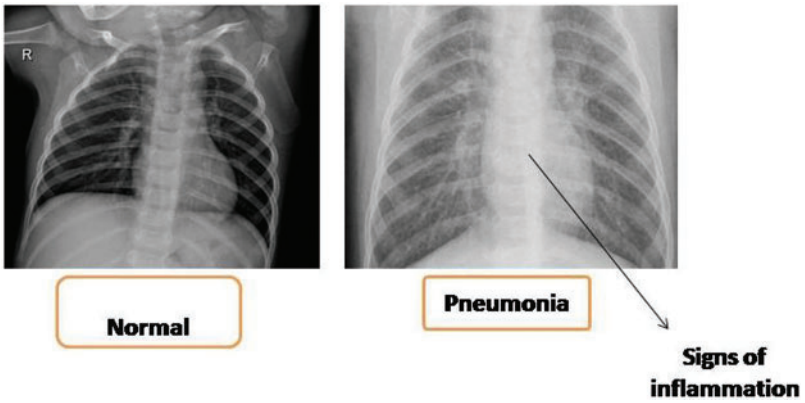

Figure 2: Sample of chest radiograph of normal and pneumonia patient

Previously researchers used hand-crafted feature [6] extraction to convert raw images into relevant features for classifying a medical image for easier classification. This necessitates domain knowledge to do feature engineering, which entails extracting relevant features, changing them, and deleting unnecessary ones. Decision trees, random forest [7], support vector machines [8] have delivered noticeable performance. Deep neural networks [9–11] have substituted feature engineering with feature learning where better features are learned, and it has demonstrated more promising results. Complex features are extracted from input images and further categorized in the architecture's final levels. Transfer learning has been applied with pre-trained classifiers [12,13] with better performance. Although the attention mechanism [14–17] was primarily used for language processing-related tasks like image captioning [18–22], language translation [23].

The motivation behind the proposed work is inspired by the recent advances in deep learning-based segmentation models, a classical machine learning-based end-to-end model which was used to classify chest X-ray images using a hand-crafted set of features defining the local look. Also, the suggested design extracts attention-aware features, hence improving classification performance.

In this paper, we are proposing to apply the attention mechanism in baseline Convolution Neural Network (CNN) and Residual CNN for binary image classification tasks to classify CXR images into normal or Pneumonia. Normally, in CNN, though, each convolution layer processes the entire image regardless of what is foreground and what is background. Attention mechanism enables the model to learn only meaning features by only focusing on the main part of the image and ignoring the rest of the unnecessary part of the image so that it does not learn any feature from the least important portion of the image. The proposed model learns only attention-aware features which makes classification more accurate.

Attention is a mechanism that is inspired by the human brain's complex and powerful cognitive power to detect something. When the human eye observes some image, text, or object, it does not focus uniformly on the entire thing. It focuses mainly on the selected region, and the remaining is ignored while processing. With this same mechanism, only the most relevant features extracted from the selected region of the image have a role to play in classification while ignoring irrelevant parts of the image. In the proposed architecture, many attention modules yield attention-aware features. As the layers become deeper, different modules' attention-aware features modify themselves to provide greater performance.

The novelty of the proposed network is that it captures local, global, and spatial information from chest X-ray images in order to enhance diagnostic performance. By incorporating the channel and spatial attention modules, it has slightly increased the complexity of the architecture but the number of trainable parameters remain the same which does not impose any extra overhead on performance on the contrary it increases the diagnostic accuracy by an appreciable amount.

The following is the summary of contributions in this paper.

i. The paper presents attention-aware CNNs to classify CXR images for the detection of Pneumonia. This attention-aware CNN architecture leads to higher accuracy in comparison to architecture without an attention mechanism. We have applied attention mechanisms in baseline CNN and Residual CNN architecture and compared the performances.
ii. We assessed the performance of pre-trained classifiers on the same dataset compared to our findings.

The rest of the paper is structured as follows. Section 2 summarizes the findings of recent research studies in related areas. Section 3 discusses the material and methods in which we have described our chosen dataset al. ong with its characteristics. Section 4 is about working principle and proposed architecture and the Section 5 provides details of experimentation and results from the analysis. Lastly, the Section 6 discusses the conclusion and its future scope.

2 Related Work

The motivation behind mimicking human attention was first seen in the domain of natural language processing, computer vision [24] in a view of reducing the computation complexity while processing an image. Moreover, to improve the performance, a model was introduced that would mainly focus on the specific region of interest in an image instead of the entire scene. The attention mechanism was further refined to machine translation model to address the issue involved. In the recent, it has been employed in a large number of DL models across various domains and tasks like text classification, image captioning, sentiment analysis and speech recognition [25,26].

Two types of attention mechanisms have arisen in the literature, both inspired by the human visual system. The first is a top-down technique, in which an iterative process selects the appropriate region from a pool of records about the scene. The bottom-up technique, on the other hand, identifies the most relevant and conspicuous locations along the visual path. The top-down strategy is iterative, and it is slower than the single-pass bottom-up approach. Furthermore, the bottom-up strategy selects the most relevant regions from the input data progressively, but the errors produced by these sequential processes increase with the depth of the process.

The attention mechanism has become a hot research area due to several reasons. Firstly, the attention mechanism involved in any model is impressive performance against state-of-the-art approaches. Secondly, the attention model can be jointly trained with a base recurrent neural network [27] or the CNNs by making use of the backpropagation approach. In addition, the induction of the transformer model was very much adopted in the tasks like image processing, video processing and recommendation system which has increased the performance of the attention model and the parallelized issue involved in the recurrent neural networks can be circumvented.

It is well known that the neural network involved in classification models the data as the numeric vector which comprises the low-level features, in which all the features are assigned the same weights irrespective of their capabilities. This issue has been addressed in the attention model [28] wherein the variable was assigned to different features according to their importance or in other words. The

attention model computes the weight distribution based on the input features and assigns higher values to the features with higher rank, which means that the attention model computes the weight distribution based on the input features and assigns higher values to the features with a higher rank.

To elaborate further, the attention layers in the attention mechanism are the alignment layer, attention weight, and the context vector. The working of the attention layer is to compute the alignment score between the encoded vector h = { $h_1, h_2, \ldots . h_n$) and a vector v. The softmax is applied to calculate the probability distribution $_1$ by normalizing over all the n elements of h where i = 1, 2, . . . , n as given shown in the Eqs. (1) and (2).

$$_i = \frac{exp(^{h_i v})}{\sum_{j=1}^{n} exp(^{h_i v})} \tag{1}$$

$$O = \sum_{i=1}^{n} {}_i \, h_i \tag{2}$$

From the above equations, the larger value of $_i$ means that h_i contributes important information to vector v. Also the output **O** of the attention mechanism is a weighted sum of all elements in the encoded vector h_i.

In machine vision challenges, we now have a plethora of attention and DL algorithms, such as (1) attention-based CNN, (2) CNN transformer pipelines, and (3) Hybrid transformer [29,30]. The primary idea of the attention-based CNN was to discover the most important components of the feature maps in CNN so that redundancy could be reduced for machine vision applications. DL [31–33] has used the attention mechanism in a variety of machine vision applications, including object detection, picture captioning, and action recognition. The CNN is utilized to provide the features map to a transformer and act as a teacher to the transformer in the CNN transformer pipelines.

The attention mechanism has evolved as a key mechanism for improving the precision and efficiency of the system by focusing on the most informative input while ignoring clutter and noise.

2.1 Attention Mechanism

Attention is a useful technique for improving the performance of the Encoder-Decoder architecture on neural network-based machine translation applications. It's a connection between the encoder and the decoder that allows the decoder to read data from every encoder's secret state. The model can selectively focus on valuable bits of the input sequence using this framework, and thus learn the relationship between them [34]. This makes it easier for the model to deal with extended sentences. A neural network is a simplified model of the brain. In deep neural networks (DNNs), the attention mechanism is an attempt to emulate the similar behavior of selectively concentrating on a few relevant things while ignoring others. Attention is used by neural networks to improve input data [35]. Other sections of the data are diminished as a result of the effect, with the idea being that the network should focus more on that data. Gradient descent [36] is used to learn which parts of the data are more significant than others, which is dependent on the context. In the 1990 s, attention-like processes such as multiplicative modules, sigma pi units, and hyper networks were introduced.

An attention function is a mapping of a query and a set of key-value pairs to output, where the query, keys, values, and output are all vectors. The output is a weighted sum of the values, with the weights assigned to each value set by the query's key compatibility function.

2.2 Attention Mechanism for Image Classification

As the human eye rapidly scans the entire image, image attention is the act of locating a target region that requires attention. This target region is given additional weight (distribution) in order to acquire the most exact data about the target while suppressing other irrelevant data. Soft attention is the most popular method since it is a fully differentiable procedure that allows CNN models to train from start to finish. Most soft attention models build an attention template, which they then utilize to locate unique components, in order to align the weights of discrete segments in a sequence or an image.

The hard attention mechanism, unlike soft attention, is a random, non-differentiable process that determines the significance of particular regions one at a time rather than identifying the image's key features as a whole. The weight of the arithmetic mean of attention can be calculated from an image's attention spectrum when using attention learning for picture categorization. In the same way that standard natural language processing may be used to gather image-based attention, the technique can be used to gather image-based attention.

Because it effectively pulls characteristics from the data, the DNN is suitable for pixel-wise categorization of images. The attention mechanism, which mimics how people perceive images, is important for quickly and precisely acquiring crucial information. With CNNs [37], all convolution layers process all of the image's features and details. To average the features and details of the entire image, use numerous convolution layers and a global average pooling [38] in the final layer. The last affine fully connected layer in this network estimates the image category. The more the background and other non-required information affect the classification results, the smaller the area of the full image taken. A huge amount of data is used to eliminate errors in findings, and the neural network learns not to output background features, thus global average pooling is unaffected.

One method is to branch the output of one layer in the CNNs, resulting in a single image produced by two or more convolutional layers. Here, we set sigmoid, which serves as the convolution output's activation function, to produce a value between 0.0 and 1.0 for each pixel. Sigmoid is a function that saves input values between 0 and 1. The output of the convolution function is multiplied by the original output. Two additional layers estimate the original output's focal areas. The values approaching 0 aren't worth focusing on. We're interested in areas where the values are multiplied by values close to one because the output is close to the original number.

Configuring processing in this way results in most of the sigmoid values that are near 0 not being used at all in the downstream recognition process, as those values would be completely disregarded. Configuring a neural network to estimate the area of focus using the result of which is the most used approach of using attention for image classification [39].

3 Materials and Methods

3.1 Dataset

The dataset is divided into three folders (train, test, and val), each of which contains a subdirectory for each image type (Pneumonia/Normal). There are 5,856 JPEG X-Ray images in total, divided into two groups (Pneumonia/Normal). From retrospective cohorts, anterior-posterior chest X-ray images from children aged one to five years old at Guangzhou Women and Children's Medical Center in Guangzhou were chosen. Chest X-ray imaging was performed as part of the patients' routine medical care. Quality control was performed on all chest radiographs, with any scans that were poor quality or unreadable being discarded. Two professionals then evaluated the diagnostic for the photographs before they were accepted for use in the AI system. The dataset for each class is shown in Tab. 1.

Table 1: Class wise distribution of dataset

Class	No. of images
Pneumonia (P)	4273
Normal (N)	1583

We have divided the dataset 75% into training set and 80% of the remaining for test set and 20% for validation set as shown in the Tab. 2.

Table 2: Splitting of dataset into train, test and validation

	No. of images	No. of images from P class	No. of images from N class
Training data	4392	3205	1187
Validation data	292	226	66
Test data	1172	842	330

3.2 Data Augmentation

We must purposefully extend our dataset to avoid the risk of overfitting. We have the ability to expand the amount of your current dataset. Small adjustments to the training data are made to reproduce the variations. Methods for changing the array representation while keeping the label the same while changing the training data are known as data augmentation strategies. Grayscales, horizontal and vertical flips, random crops, colour hiccups, translations, rotations, and a variety of additional enhancements are common. By making just a few of these changes to our training data, we can quickly double or quadruple the amount of training examples and construct a very robust model.

We used data augmentation transformation techniques such as 30 degree sequence rotation, zooming within a range of 0.2, width shift by 0.1, height shift by 0.1, and horizontal flipping.

3.3 Implementation Environment and Tools Used

Python 3.6 was used here in this work for the implementation. We have utilized Google Colaboratory (https://colab.research.google.com/), which uses the Jupyter notebook environment, because we needed faster GPUs for our tests. Colaboratory provided a Tesla K80 GPU to speed up the processing. In addition, the Google Colaboratory environment provides 12 GB of GDDR5 VRAM and 13 GB of RAM for free. We utilised the Keras and Tensorflow libraries in versions 2.2.2 and 1.10.0, respectively, to create DL models. The parameter settings utilized during the experiment are listed in Tab. 3.

Table 3: Parameter settings used in entire experimentation

Parameter	Value
Image size	150 × 150
Loss function	Binary cross entropy

(Continued)

Table 3: Continued

Parameter	Value
Dropout	0.2
Stride	1
Batch size	16
Optimizer	Adam
Epochs	50
Learning rate reduction	ReduceLROnPlateau(monitor = 'val_accuracy', patience = 2, verbose = 1, factor = 0.8)

4 Proposed Architecture

4.1 Working Principle

Image attention is the act of locating a target place that requires attention when the human eye rapidly sweeps the global image. This target region is given additional weight (distribution) in order to acquire the most exact data about the target while suppressing other irrelevant data. The most prevalent strategy is soft attention, which is a fully differentiable procedure that allows CNN models to train from beginning to end. In order to align the weights of discrete segments in a sequence or an image, the most soft attention models create an attention template, which they subsequently use to discover unique components. Unlike soft attention [36–39], the hard attention mechanism is a non-differentiable, random process that decides the significance of individual regions one at a time rather than identifying the image's main elements as a whole.

The weight of the arithmetic mean of attention can be determined from an image's attention spectrum when using attention learning for image categorization. The attention model calculates and selects the most relevant region of the feature spectrum for the final classification job by learning the attention of the CNN feature spectrum, and supplies the maximum attention input (weight distribution). When the attention weight is added to the last layer of CNN features, the original attributes are muted to varying degrees. The original feature spectrum is blended with the weighted feature spectrum to solve this suppression. The completely connected layer receives the fusion spectrum after that.

In the second fully connected layer, the attention feature spectrum is coupled with the fully connected feature spectrum in the channel direction before being delivered to the classification layer for classification, which has been adjusted by the global average pooling dimension. The Softmax layer's output error is back propagated during the back propagation process, and the parameters are modified using a random gradient descent approach to lower the final loss function value and bring the network closer to convergence.

The output of the CNN's final convolutional layer is acquired using the soft attention technique. This data is entered into the attention model, which then generates the proper attention spectrum. The following network uses the output attention feature spectrum as an input, and the attention spectrum is utilized to weight the original feature spectrum.

Here I is the input image in Fig. 3. Each of the n parameters in the attention model (a1, a2, . . . ai, ... an) describes a different element of the image. The importance of each ai in relation to the input I determines the return value of the model's attention spectrum (more precisely, the weight

values of the n parameters). You can locate the area that requires the most attention by filtering the input image through this output. Our proposed model's layer-by-layer structure is shown in Fig. 4.

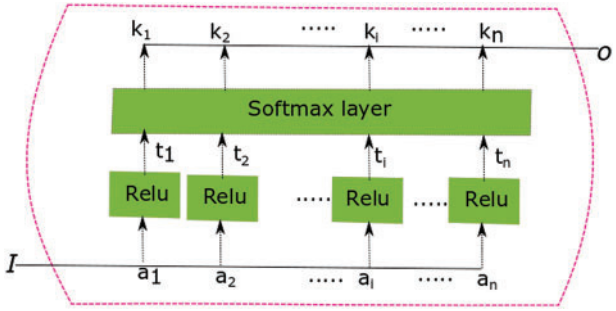

Figure 3: Block diagram of attention mechanism

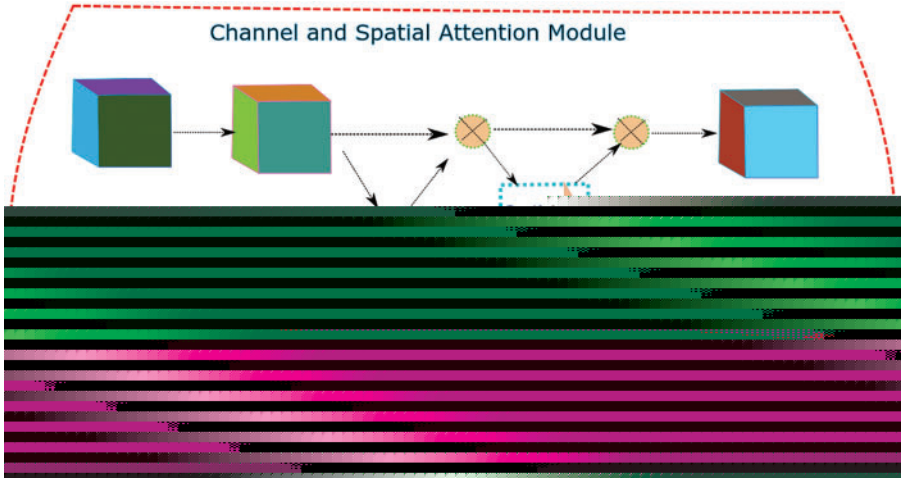

Figure 4: System design of proposed architecture

4.2 Layer Wise Organization of the Proposed Model

Convolutional Block: It extracts feature map in the form of high-level features and then these are passed to channel attention and spatial attention blocks to enhance the feature maps.

Batch Normalization: Batch normalization [40] is a network layer that allows each layer to learn in a more autonomous manner. It's used to normalize the output of the previous layers. When batch normalization is employed, learning becomes more efficient, and it can also be used to prevent model overfitting.

Channel attention: A Channel of Attention A CNN module that focuses on channel-based attention is Module. Using the inter-channel relationship of features, we create a channel attention map. Channel attention focuses on 'what' is significant given an input image since each channel of a feature map is viewed as a feature detector. It compresses the spatial dimension of the input feature map to compute channel attention effectively. We initially perform average-pooling and max-pooling processes to aggregate spatial information from a feature map, resulting in two separate spatial context