applied sciences

MDPI

*Article*

# Attention-Based Transfer Learning for Efficient Pneumonia Detection in Chest X-ray Images

**So-Mi Cha †, Seung-Seok Lee † and Bonggyun Ko ***

Department of Mathematics and Statistics, Chonnam National University, 77, Yongbong-ro, Buk-gu, Gwangju 61186, Korea; 198267@jnu.ac.kr (S.-M.C.); 207924@jnu.ac.kr (S.-S.L.)

* Correspondence: bonggyun.ko@jnu.ac.kr

† These authors contributed equally to this work.

**Abstract:** Pneumonia is a form of acute respiratory infection commonly caused by germs, viruses, and fungi, and can prove fatal at any age. Chest X-rays is the most common technique for diagnosing pneumonia. There have been several attempts to apply transfer learning based on a Convolutional Neural Network to build a stable model in computer-aided diagnosis. Recently, with the appearance of an attention mechanism that automatically focuses on the critical part of the image that is crucial for the diagnosis of disease, it is possible to increase the performance of previous models. The goal of this study is to improve the accuracy of a computer-aided diagnostic approach that medical professionals can easily use as an auxiliary tool. In this paper, we proposed the attention-based transfer learning framework for efficient pneumonia detection in chest X-ray images. We collected features from three-types of pre-trained models, ResNet152, DenseNet121, ResNet18 as a role of feature extractor. We redefined the classifier for a new task and applied the attention mechanism as a feature selector. As a result, the proposed approach achieved accuracy, F-score, Area Under the Curve(AUC), precision and recall of 96.63%, 0.973, 96.03%, 96.23% and 98.46%, respectively.

**Keywords:** transfer learning; attention mechanism; computer aided diagnosis; chest X-ray; pneumonia detection



**Citation:** Cha, S.-M.; Lee, S.-S.; Ko, B. Attention-Based Transfer Learning for Efficient Pneumonia Detection in Chest X-ray Images. *Appl. Sci.* **2021**, *11*, 1242. https://dx.doi.org/10.3390/app11031242

Academic Editor: Yahan Hu
Received: 4 January 2021
Accepted: 25 January 2021
Published: 29 January 2021



## 1. Introduction

Pneumonia is a disease that causes inflammatory reactions and hardening of the lung tissue in areas below the respiratory tract along the breathing path as a result of disease-causing bacteria [1]. Pneumonia can occur in anyone and threatens the lives of people of all ages depending on the progress of pneumonia [2]. It is common to start treatment with a diagnosis of pneumonia from chest photos [3], usually accompanied by coughing, sputum, and fever. However, for identification with a similar non-infective disease, it is a clear method of diagnosis to detect pathogens caused by the lungs or to prove reasonable pathological findings. Howeveer, empirical diagnosis and treatment are still important because not all patients can be tested like this. Medical experts perform additional computed tomography (CT) when a possibility of pneumonia exists based on the chest X-ray results. When the examination does not detect pneumonia, it leads to misdiagnosis and puts the patient in a dangerous condition [4]. In chest X-ray images, blurry parts of the bronchial tubes can be expressed depending on the value of each pixel. At this time, blurry parts indicate suspected pneumonia due to inflammation or the discharge of inflammation. Values corresponding to each pixel are classified as normal or pneumonia through a layer of weights inside the model. The weights used can be numerically identified as weights for pre-trained models. However, it is not easy to diagnose pneumonia through X-ray images without the help of experts. The chest X-ray allows medical experts to diagnose pneumonia or gain confidence with further examinations. However, although there are criteria for diagnosing pneumonia, there are many aspects of pneumonia. It can be difficult for medical professionals to grasp all of the

criteria [5]. Recently, research on detecting and predicting diseases with deep learning has been actively conducted in Computer Aided Diagnosis(CAD) and achieved good performance in various medical fields [6,7]. It obtained successful results in medical areas, including radiology. Although CAD cannot replace the role of medical experts, they can aid medical experts to save time and effort as an auxiliary tool. Computer-aided diagnosis has an advantage in terms of improving reliability and accuracy in a doctor's diagnosis as a decision support system rather than being directly used for pneumonia diagnosis.

Convolutional Neural Network (CNN) has shown advantages in CAD to extract the majority of the informative features of images. CNN has been considered the current state-of-the-art image classification technique. As a result, several pre-trained models based on CNN have won the classification competitions. Since the first large-scale deep neural network AlexNet [8], various CNN models such as VGGNet [9], ResNet [10], and DenseNet [11] have been proposed, and several studies have been conducted to diagnose pneumonia using CNN [12,13]. AlexNet [8] is the first CNN architecture that won the competition. Building a deep neural network by stacking multiple layers is the most widely known method of improving network performance [9,14]. All models that recorded a high score for the challenging ImageNet competition since 2012, have been deep CNN models [9,14]. VGGnets [9] showed that deeper layers can improve the performance. In general, as the layers of CNN get deeper, the extracted features have more information. However, when layers of CNN reach a certain point, the model performance becomes saturated and quickly decreases. ResNets [10] solve the gradient vanishing problem caused by stretched layers using a residual connection with batch normalization. DenseNet [11] densely connects all layers to ensure maximum flow of the information entered. Each layer receives an additional input from all of previous layers and passes the current feature map to all subsequent layers. In DenseNet, the output of the convolutional layer is the concatenated value of the input feature maps. ResNet adds feature maps together, but DenseNet [11] makes concatenation between feature maps. Constructing a CNN model in computer vision, most of them are not trained from scratch(random initialization). They just simply used structures of the pre-trained model for a new task. In the medical field, it is difficult to build sufficient datasets with labels because it is expensive to acquire ground truth labels such as ImageNet [15]. From a machine learning perspective, the ground truth is the representation of the original or actual value of the data we want to train. The ground truth labels of the data we used are normal and pneumonia.

Transfer learning refers to applying the knowledge accumulated by Artificial Intelligence to another task. It is a process of adapting the model trained on one data set to another data set [16]. As a result, it is possible to train faster at an early stage than the model trained from scratch [17]. In transfer learning, a CNN model which pre-trained on a large and diverse dataset such as ImageNet can be applied to perform a new classification task [15]. When the dataset is insufficient for building and training a new model from scratch, transfer learning can be applied as the solution. Additionally, the amount of data required for training is not as big as that from the scratch model. Transfer learning makes the model more stable. So, Instead of building a model from scratch, simply adapt the structure of pre-trained models that were trained on the ImageNet [15]. Pre-trained CNN models are used to diagnose disease in fields with small datasets. Fine-tuning is the method of transfer learning that changes the classifier to fit a new task. The classifier removes the last fully connected layer from the pre-trained model and changes it by adding a new fully connected layer for the new task. This method slightly modifies existing weights during the training.

Afterwards, many studies have focused on further improvement of the performance of CNN and show significant improvements over the years. However, the CNN model only considers the correlation between spatially adjacent pixels in the receptive field defined by the filter size. Therefore, it is difficult to find correlations with distant pixels. Recently, attempts to apply attention mechanisms are followed to solve this problem. Attention is a mean of finding and focusing on the most informational part of the data. The

attention model has shown good performance in various fields of computer vision [18,19]. With an attention mechanism, integrating channel operations into convolution blocks showed great potential in improving performance [20,21]. As a result, attention is used as a method to complement or replace the existing CNN structure. The squeeze-and-excitation Network(SENet) [20] was suggested to incorporate global information into the decision process of a network with channel-wise interdependency for each convolution block. The key idea behind the Squeeze-And-Excitation block is to incorporate global information into the network's decision process. SENet learns the interactions between channels of convoluted features. While CNN only sees local information of a certain size, the Squeeze-And-Excitation block collects information from all receptive fields. SENet uses the Squeeze-And-Excitation building block for feature recalibration. This method can flexibly apply to existing CNN models and requires little additional computation. It is performed by adding computational blocks to all or some stages of the network such as ResNet [10], Inception [22] and ResNeXt [23]. The Efficient Channel Attention Model(ECA-Net) [21] used $1D$ Convolution to demonstrate distinct performance improvements with a small number of parameters through cross-channel interactions without dimension reduction. ECA-Net highlighted the problem of the increasing model complexity of SENet. ECA-Net also uses channel-wise-attention, which is the same as SENet. However, ECA-Net excludes the process of dimensionality reduction in the fully connected layer. Therefore, they can improve performance, while reducing model complexity.

There are hybrid approaches that combined pre-trained models with optimization methods for medical image classification. MobileNet-AEO [24] proposed hybrid approaches that combined MobileNet [25] with Artificial Ecosystem-based Optimization [26] to detect tuberculosis. Pre-trained MobileNet [25] trained on ImageNet takes on the role of feature extractor. Artificial Ecosystem-based Optimization(AEO) takes a role of feature selector. AEO includes the majority of relevant features and exclude the irrelevant features. Inception Fractional-order Marine Predators Algorithm(IFM) [27] also proposed a combined approach using Inception [14] mode as a feature extractor. They used Fractional-order Marine Predators Algorithm(FO-MPA) as a robust feature selection method. FO-MPA was adopted to select efficiently relevant feature vectors and eliminate unnecessary features. IFM achieved high performance with reduced resource consumption and storage capacity, which minimizes processing time.

In this paper, we propose a novel framework which combines several pre-trained CNN models as a tool to extract feature vectors from chest X-ray Images from patients at Guangzhou Women and Children's Medical Center. We then concatenate each feature vector and apply the attention mechanisms. We propose to create a model that can be applied to any task by collecting pre-trained models from various domains. With our framework, we obtained a compliant model that achieved good performance in the field of the insufficient dataset. So, we expect this insight can be applied to a wide range of fields that have small datasets, thus require domain adaptation.

## 2. Materials and Methods

We used models pretrained on datasets from various domains to implement transfer learning using the attention mechanism. Then, we concatenated the feature vectors extracted from each of the pre-trained models—ResNet152, DenseNet121 and ResNet18—to the same dimension vectors. Then, feature vectors extracted from each model are combined and concatenated into the same dimensional vector. We can utilize the advantages of each model as feature extractors. Then, apply the different type of attention mechanism of SE, ECA, Self Attention. The output vectors from the attention operation were passed to the classifier (the last fully connected layer), which changed for the current task. Figure 1 shows the overall framework based on transfer learning with attention mechanism.

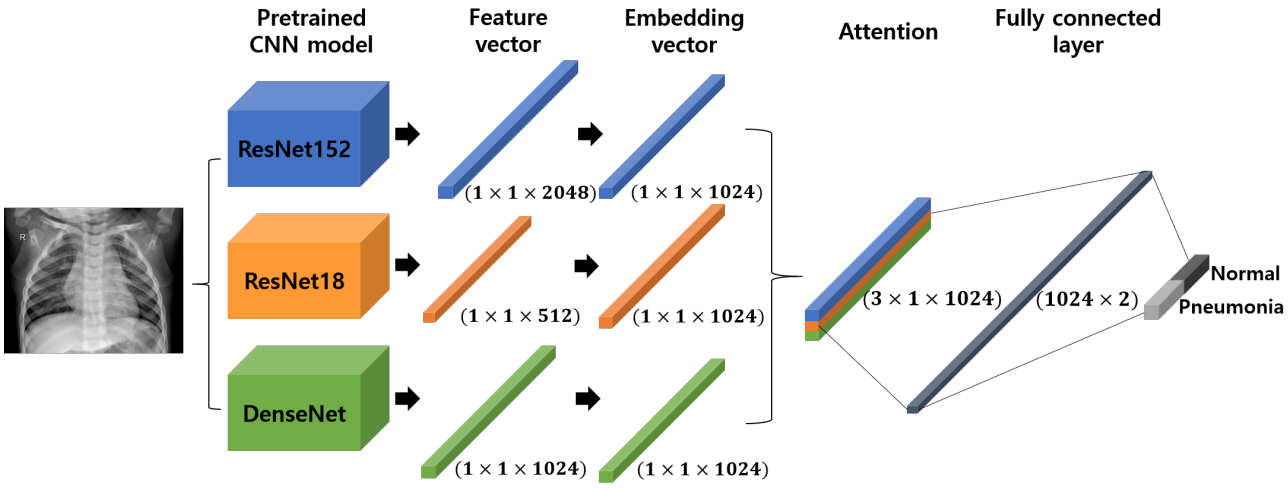


**Figure 1.** The framework based on transfer learning with attention mechanism with pneumonia.

### *2.1. Data Preprocessing*

We utilized pre-trained models from 3 different domains, ImageNet [15], ChestX-ray14 datasets [28] and Custom dataset [29] to gain the generalization ability of our framework. The three pre-trained models we used are quite large and have many parameters compared to the test data. So, we apply some data preprocessing to avoid over-fitting. We used image augmentation methods for chest X-ray images. Data preprocessing is a series of processes that makes input data suitable for specific analysis. For example, Crop means cropping the pixels at the edges when the center of the image is important. Rotating rotates images at a random angle. Flip means flipping images left and right or up and down to make more data. It also includes normalization, which converts the data range to a value between 0 and 1 in order to reduce the impact on the relative size of data. We used crop, rotate, and flip and normalization in data preprocessing. First, we modified the image size to $(224 \times 224)$, rotated randomly selected images, flipped those images horizontally and applied normalization to make each dimension in data have values within the same range. This improves the ability to generalize the model. In case low-quality images are given as input, our approach can reach the compliance performance through data pre-processing. However, the performance can improve when higher quality images are given as input.

### *2.2. Pre-Trained CNN Models*

CNN is a model that improves the problem of Deep Neural Network(DNN). DNN model uses only one-dimensional data. However, image data is two-dimensional. Using DNN for two-dimensional data creates a large loss problem in the process of changing to one-dimensional data. CNN is a proposed model that can be trained on image data by applying filters of a certain size to the image. We utilized pre-trained models that were trained on three different domains, ImageNet, Chest X-ray14 datasets and Custom datasets. The three models are ResNet152, Dense121, ResNet18.

#### 2.2.1. ResNet

ResNet [10] was proposed to overcome the performance degradation caused by vanishing gradient. As the depth of the model increases, the performance deteriorates and then suddenly drops. ResNet learns from the residual using a skip-connection that adds inputs at the end so that the gradient is at least 1. Adding a short-cut connection changes the problem to calculate the residual as how much it changes from the previous value. Since the output value of the current layer and the output value of the previous layer are added to receive as input, only the residual (the difference between the output of the previously learned layers and the output of the added layer) needs to be learned. Through these methods, this network solves the gradient vanishing problem. We use these ResNet

architectures and only change the number of layers and last fully connected layer(classifier) to fit our task. Figure 2 shows the overall structure of ResNet. Conv block is a block that aligns channels in the shortcut path when the channel in the main path through the input differs from the channel in the shortcut path through which the residuals pass. Identity block is a block that performs only simple addition when the shortcut path channel and the main path channel match. Flatten is the process of converting one-dimensional data into a fully connected layer (FC). We used two types of ResNet—ResNet152 and ResNet18. Each model consists of 152 layers and 18 layers, respectively. ResNet solves the issue of performance degradation caused by vanishing gradient. ResNet performs better as the layer depth increases. ResNet18 uses 3 × 3 filters, whereas ResNet152 uses 1×1 filters, showing structural differences.

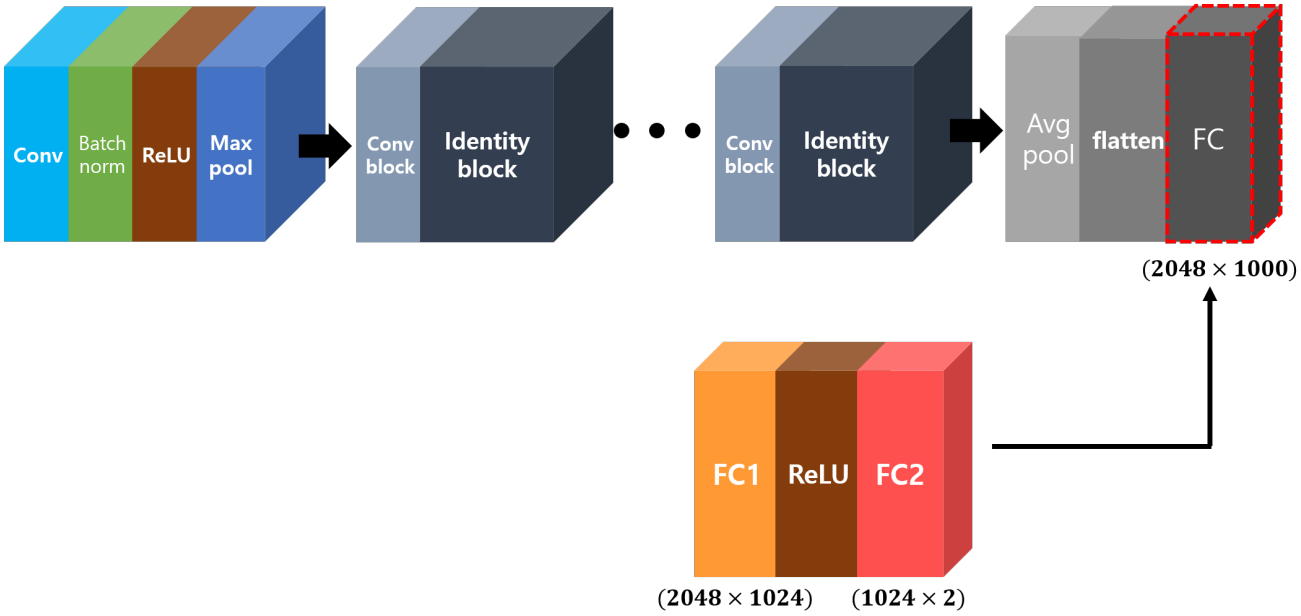


**Figure 2.** Pretrained ResNet with changed classifier on ImageNet.

#### 2.2.2. DenseNet

While ResNet [10] is a way to proceed by adding the previous information, DenseNet [11] is a way to stack all the parts that have been passed through while learning. In order to maximize the information flow between layers of the network, each of the layers with the same feature map size are directly connected. DenseNet(shown in Figure 3) is a structure in which all layers are connected in a feed-forward manner. The convolution layer of each DenseBlock is concatenated with the outputs of all previous convolution layers. This concatenated feature map is passed to the convolutional operation. The feature map of the previous layer is continuously connected with the input of the next layer, and the connection is a method of concatenation rather than addition of the feature maps. Each layer receives an additional input from all previous layers and it delivers the current feature map to the next layer to maintain the feed-forward feature. DenseNet compensates for the loss of initial information as the layer deepens, and each layer directly approaches the gradient obtained from the loss function and input, making learning easier. Additionally, filters are more densely located on each layer than in other models, which efficiently take information and reduce the number of parameters. However, the deeper the layer, the more computational power required. We use the weights of pre-trained DenseNet, but we redefine the dimension of the last fully connected layer of DenseNet121 to fit the number of our classification classes.

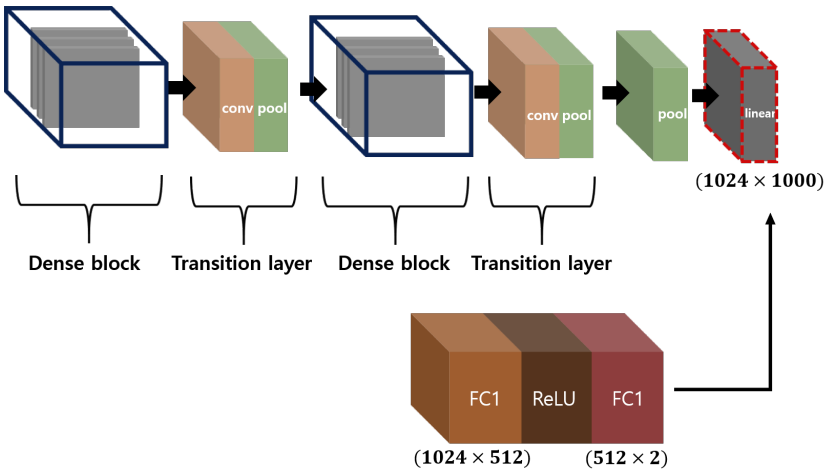


**Figure 3.** Pretrained DenseNet with changed classifier on Chest X-ray14 of NIH.

### *2.3. Transfer Learning*

Transfer learning refers to applying pre-trained models based on large datasets to do tasks in similar fields. Our goal is to utilize chest X-ray images to more accurately diagnose pneumonia. Therefore, we intend to use CNN-based models with a similar objective. All of the pre-trained models we used are classification models, each used 1,281,167 ImageNet data with 1000 classes, 112,120 National Institute Of Health(NIH) chest X-ray data with 14 classes, and 43,956 custom data with 11 classes. The new model can perform well in pneumonia diagnosis despite the lack of data to train the model properly using a pre-trained model. We adopt a method to change the last fully connected layer that fits our task, while we continue to use the weights of the layers of the pre-trained models have. For example, ResNet152 [10] is a CNN model that is trained on ImageNet and classified into 1000 classes. We changed the end layer of the pretrained ResNet152 as the new classifier so that it can be classified into 2 classes—as pneumonia images and normal images, as seen Figure 4. However, when performing transfer learning using pre-trained models, there is a disadvantage of training degradation if the domains of data used by users differ significantly from those of data used for pre-training.

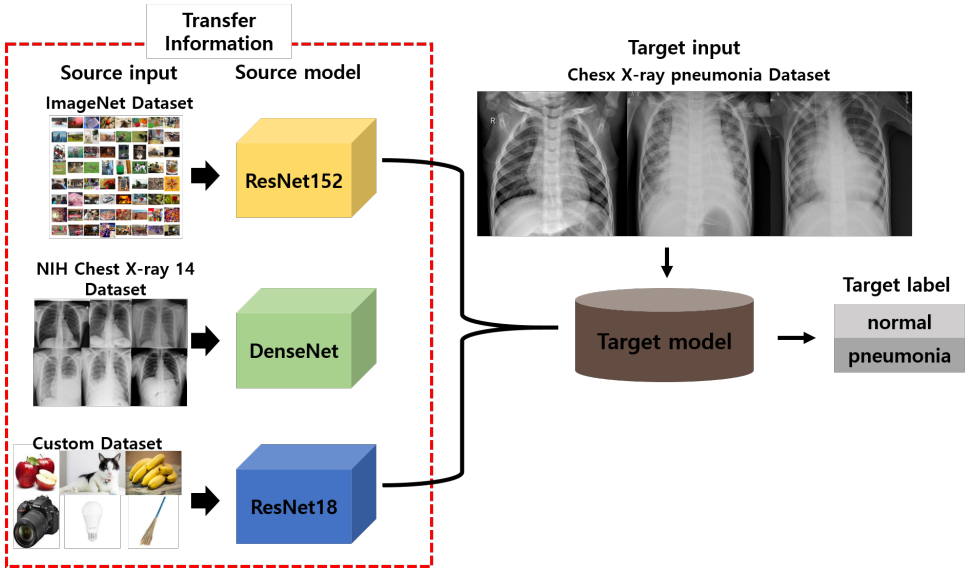


**Figure 4.** Overall process of transfer learning.

### *2.4. Attention Mechanism*

The attention mechanism was first introduced in a sequence-to-sequence translation model based on an encoder–decoder [30]. Attention allowed to summarize context-based information at variable lengths in the input sequence. Self-attention applies attention to a single context rather than multiple contexts, enabling direct long-distance interdependency. In computer vision, the attention mechanism is often used to complement the CNN model. It focuses on a certain feature that is important to classify. The main purpose of the attention mechanism is to make the model focus on important parts [30] of input data.The attention mechanism looks for correlations with distant pixels by applying a weight filter of the same size as the original image. It helps to find areas that show a difference in classifying pneumonia compared to normal X-rays, i.e., areas that contain a lot of information for classifying pneumonia. We will focus on the area surrounding the lungs throughout the image, and pneumonia may be suspected when the surrounding of the lungs looks blurry. When diagnosing a specific disease, it is important to know exactly which part of the image to look at, and attention is effective because it automatically finds and focuses on important parts of the image. We propose the application of the attention mechanism to feature vectors in the computer vision field.

#### 2.4.1. Self-Attention

Previous convolution operations have a problem in identifying relationships with pixels in distant locations. Each of the features is only contained in certain local fields. Self-Attention [31] was proposed to overcome these problems. The computation of Self-Attention is similar to that of convolution operations but it is easy to identify relationships with remote regions. Figure 5 gives a visual representation of Self-Attention. Given an input, it is divided into Query, Key and Value of the identical shape as $(B_{size} \times L_{Query} \times D_{model})$. These mean batch size, query length, and model dimension in that order. Here, we use Multi-Head-Attention that performs Self-Attention from several perspectives. First, we get the mapped score from the convolution operation of query and key, and then pass that score to the soft-max layer. $(S = softmax(Q \otimes K^T))$. The output of the attention operation is the convolution operation of the Score and Value$(O = S \otimes V)$. Classification is performed in a fully connected layer using the final output of Self-Attention.

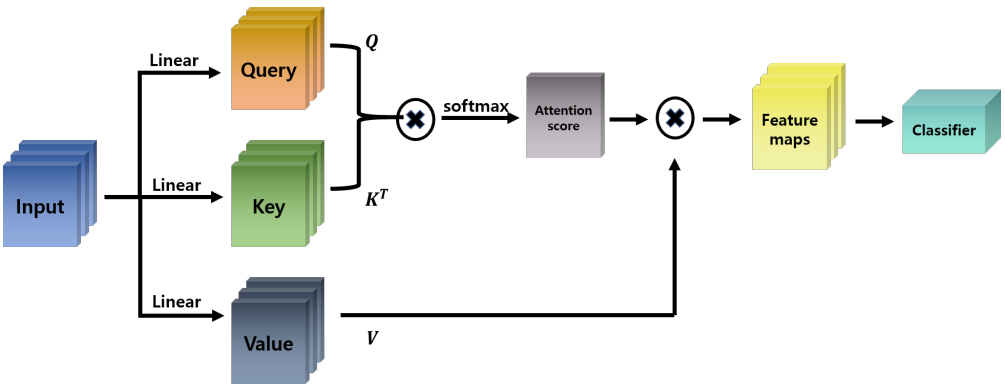


**Figure 5.** Overall process of Self-Attention.

#### 2.4.2. SENet

SENet [20] uses the interdependency between the channels to improve the quality of the generated feature representation by the network. The goal of SENet is to recalibrate the features that were obtained through the convolution process with the importance per channel. SENet uses channel-based attention to selectively adjust the weights of CNN channels. SENet improves the performance of the CNN model by attaching SE blocks after the convolution operation. This model consists of two parts, squeeze and excitation. The squeeze operation compresses the entire information to embed global information. Excitation operation scales the importance of each feature map. In this step, the squeezed

important information is recalibrated. (shown in Figure 6) These two parts are tied together and called SE blocks. The first step is a simple convolution operation that converts the dimension of $X(H' \times W' \times C')$ to $U(H \times W \times C)$. It converts two-dimensions of $(H \times W)$ feature maps of $C$ channels into $(1 \times 1)$-sized $C$ feature maps. Using a Global Average Pooling(GAP), each two-dimensional feature map is averaged to one value. It compress the global information from each channel. This operation was referred in Equation (1), where $u_c(i, j)$ represents the output of the convolutional operation $X(H' \times W' \times C')$ with $c$ filters. In $u_c(i, j)$, $i$ and $j$ are the mean indices of $H$ and $W$, respectively

$$z_c = \frac{1}{H \times W} \sum_{i=0}^{H} \sum_{j=0}^{W} u_c(i, j) \tag{1}$$

The excitation operation (shown in Equation (2)) computes channel-wise dependency by adjusting the fully connected layer and the non-linear function. Dimension reduction is performed in the middle to reduce the amount of computation. In the end, apply a non-linear function to the outputs from the fully connected layer.

$$s = Sigmoid(W_2 ReLU(W_1)) \tag{2}$$

$W_1$ and $W_2$, respectively, refer to the fully connected layers. After all operations are performed, multiply each of the C feature maps before GAP and print them out (shown in Equation (3)).

$$\tilde{x}_c = s_c \bullet u_c \tag{3}$$

As a result, a feature map of where all of these values are scaled by the importance of channel with values between 0 and 1.

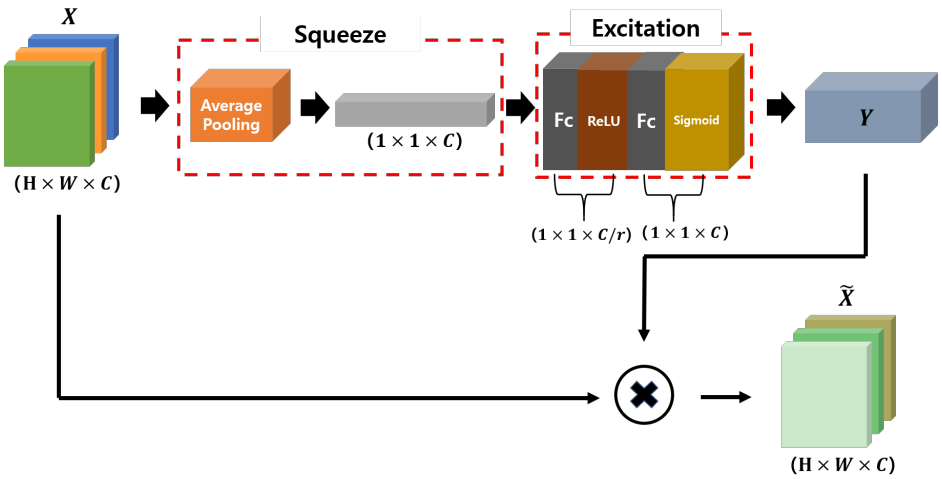


**Figure 6.** Overall process of SENet.

### 2.4.3. Efficient Channel Attention

Previous CNN models have greatly improved their performance by using the self-attention mechanism. However, the complexity of the model is too high. SENet computes weights using two FC layers. Through two fully-connected layers, dimension reduction is performed to reduce the amount of computation. Important features are highlighted via non-linear activation functions. In ECA [21], the channel weight is obtained by performing channel attention without dimensionality reduction using $1D$ convolution of kernel size $k$ instead of two FC layers (shown in Figure 7) Model complexity is reduced by a filter that considers only the surrounding local area with $1D$ convolution. ECA improves the balance between performance and complexity. Kernel size $k$ is determined by cross-channel interaction through the channel dimension $C$ function. Considering that the function between $k$ and $C$ is $C = \phi(k)$ and channel dimension $C$ is generally set as a multiple of

2, it can be set as a non-linear function such as $C = \phi(k) = 2^{\gamma \times k - b}$ This function can be expressed in terms of $k$(shown in Equation (4)). ECA modules are applied to CNN models by replacing existing SE blocks.

$$k = \psi(C) = |\frac{log_2 C}{\gamma} + \frac{b}{\gamma}| \quad (4)$$

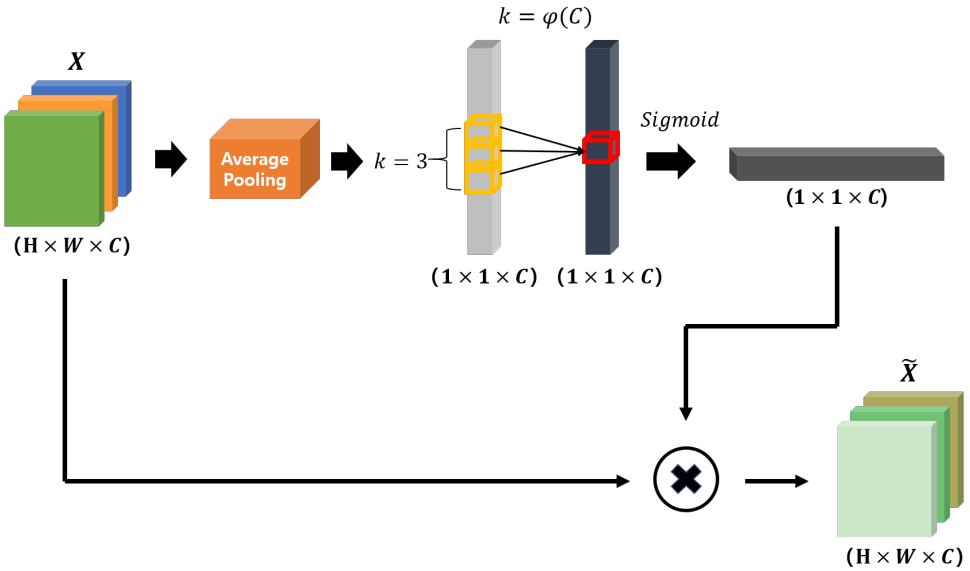


**Figure 7.** Overall process of ECA.

### *2.5. Datasets*

For the training, we employed three different domains of datasets. We employed the structure of pre-trained models that were each trained on ImageNet [15], Chest X-ray14 [28], Custom datasets [29]. The ImageNet dataset is well known for being easily accessible and large. It contains 1,281,167 data classified as 1000 classes. The Chest X-ray 14 dataset is the largest publicly available chest X-ray dataset released by the National Institute Of Health(NIH). It contains 112,120 X-ray images of 30,805 patients and classified into 14 chest pathology labels, including pneumonia. The Chest X-ray 14 dataset has a similar domain as our test dataset. The Custom dataset is classified into 11 classes and contains a total of 43,956 images. This dataset was used to demonstrate the generalization ability of our framework. With our framework, even pretrained models in unrelated domains can also be used to implement the target task. For the testing, we used publicly available chest X-ray images collected from pediatric patients between 1 to 5 years old from Guangzhou Women and Children's medical center [32], which was not used for the pre-trained model. This dataset was collected by researchers that have achieved approvals from the Institutional Review Board (IRB) and Ethics Committee on data collection and experimentation. The dataset consists of 1583 normal images and 4273 pneumonia images. Of the 5856 images, 5216 images are used for framework training, and 16 images are used as data to evaluate models during training. The size of the validation set was too small, so it was used for reference only, and cross-validation was not performed separately. Cross-validation is used to avoid overfitting. However, this process takes a lot of time. So, we chose not to use the cross-validation procedure. Instead of cross-validation, we used the validation set to determine the optimal number of epochs to avoid overfitting. Finally, 624 images were used to conduct tests on the framework (shown in Table 1). Pneumonia is divided into bacterial and viral pneumonia. However, we focus on the binary classification task of classifying normal and pneumonia by defining two pneumonia categories as one pneumonia category due to lack of data. Images of the normal class are much smaller than images of the pneumonia class. We used F-score to evaluate our approach properly given

the class-imbalanced dataset. Although we used a class-imbalanced dataset, there was no major problem with model performance. A sample image of the dataset is shown in Figure 8.

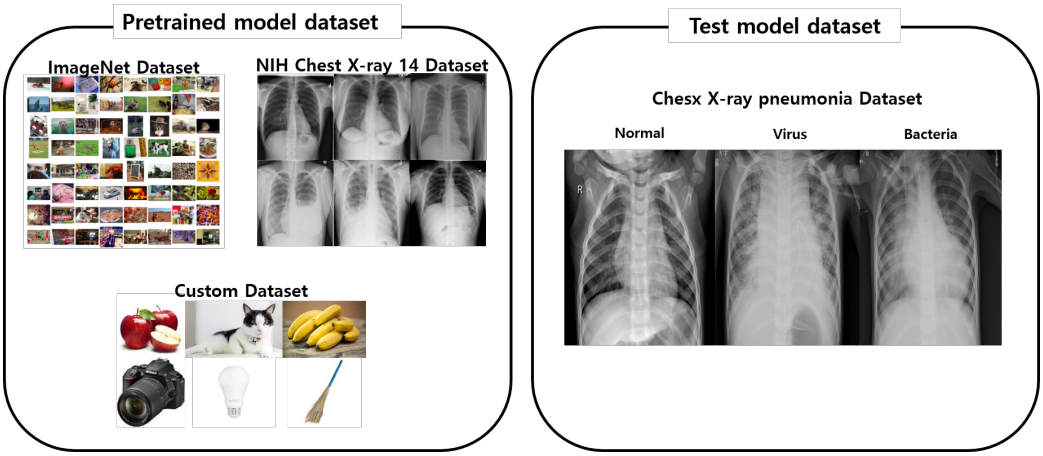


**Figure 8.** A sample of dataset.

**Table 1.** The splits of the pneumonia datasets from Guangzhou Women and Children's Medical Center.

| Category | Train | Validation | Test |
|---|---|---|---|
| Normal | 1341 | 8 | 234 |
| Pneumonia | 3875 | 8 | 390 |
| Total | 5216 | 16 | 624 |

## 3. Results

### *3.1. Results*

The primary goal of our approach is to improve the accuracy of pneumonia diagnosis using chest X-ray images. We prepared three types of pre-trained models based on a different domain to check the generalization of our framework. Feature vectors were extracted using the backbone structures of pre-trained ResNet152, DenseNet121, ResNet18 to confirm the performance of the classification of pneumonia and normal. Each dimension of the feature vector of the model is 2048, 1024, 512. To verify our framework, we concatenate each feature vector with the same 1024 embedding dimensions. We applied three types of attention mechanisms, Self-Attention, SE, ECA. The proposed approach has been trained on 5216 images from Guangzhou Women and Children's medical center, while the rest of the 624 images were used for testing the performance of the model. We redefined the final layer of each of the pre-trained models with a new classifier suitable for classifying pneumonia and normal images. For the experiment, we implemented the Adam optimizer [33] and the NLL (Negative Log-Likelihood) loss function. The learning rate started at 0.0001 and was reduced by $\gamma = 0.1$ after every 10 epochs. The epoch represents the number of training sessions for training the datasets. The computations in the neural network are divided into forward pass, which goes through the calculation process of each layer's weight from input to output, and backward pass, which goes back to the calculation process and modifies the weight. Once the two processes are completed, one epoch was performed. All models were trained within 30 epochs. The results from each model are shown in Table 2. We used Recall, Precision, Area Under the Curve(AUC), and Test Accuracy as our evaluation metrics. The following can be obtained by calculating the Recall and Precision by referring to Figure 9 and Equation (5).There are a total of four possible outcomes from the binary classification we performed. It is divided into True Positive(TP), True Negative(TN), False Positive(FP), False Negative(FN). True and False

means whether the model was correctly or incorrectly predicted. The prediction of normal by the model means Positive, while predicting pneumonia means Negative.

$$\begin{aligned} \text{Recall} &= \frac{\text{TP}}{\text{TP} + \text{FN}} \\ \text{Precision} &= \frac{\text{TP}}{\text{TP} + \text{FP}} \end{aligned} \tag{5}$$

AUC uses Fall-out and Recall indicators. Fall-out is obtained as follows (shown in Equation (6)), and a graph resulting from Fall-out on the x-axis and Recall on the y-axis is called the Receiver Operating Characteristic(ROC) curve, and the area of the graph is called AUC. F-score is the harmonic mean of Precision and Recall (shown in Equation (7)), and is an indicator of the performance of the model if the data labels are unbalanced. Test accuracy indicates classification accuracy using test datasets (shown in Equation (8))

$$\text{Fall-out} = \frac{\text{FP}}{\text{TN} + \text{FP}} \tag{6}$$

$$\text{F score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{7}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}} \tag{8}$$

The proposed models were evaluated by metrics such as recall, precision, AUC, F-score and accuracy. SE achieved an accuracy, F-score, AUC, precision and recall of 96.63%, 0.973%, 96.03%, 96.23%, and 98.46%, respectively. SE achieved the best performance in terms of accuracy, followed by ECA and Self-Attention.

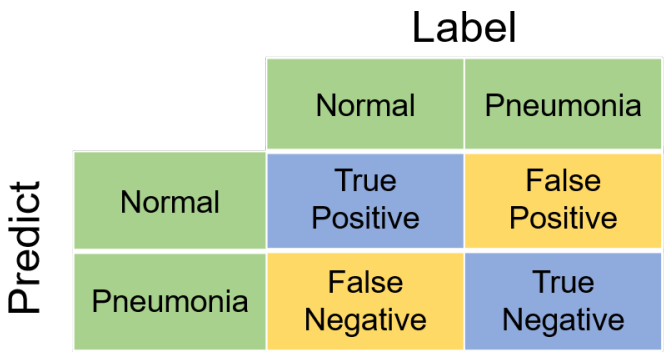


**Figure 9.** Matrix of relationships between actual and model-predicted answers.

**Table 2.** Comparison of simple transfer learning models and models with different attention mechanisms on the test dataset in terms of performance metrics.

| Model | Data | Epoch | Recall (%) | Precision (%) | AUC (%) | F-Score | Accuracy (%) |
|---|---|---|---|---|---|---|---|
| ResNet152 | ImageNet dataset | 30 | 98.97 | 93.46 | 93.72 | 0.961 | 95.03 |
| DenseNet121 | NIH dataset | 30 | 98.72 | 94.13 | 94.23 | 0.964 | 95.35 |
| ResNet18 | Custom dataset | 30 | 98.97 | 93.24 | 93.50 | 0.960 | 94.87 |
| Self-attention | - | 30 | 98.72 | 96.46 | 93.72 | 0.976 | 95.03 |
| ECA | - | 30 | 98.21 | 95.99 | 95.68 | 0.971 | 96.31 |
| SE-Attention | - | 30 | 98.46 | 96.24 | 96.03 | 0.973 | 96.63 |

### *3.2. Comparison*

In this subsection, we compared the performance of our models against recent CNN-based models that support the diagnosis of pneumonia. We compared the models that have the same domain and dataset of chest X-rays from Guangzhou Women and Children's

Medical Center. The results are summarized in Table 3. Kermany et al. [32] used a pre-trained InceptionV3 model, which was trained on the ImageNet. The model was initialized with the weights of the pre-trained InceptionV3 rather than random initialization. They trained the model by unfreezing and updating the pretrained weights on chest X-ray images as fine-tuning. They obtained a classification accuracy and recall of 92.8%, 93.2%, respectively. Cohen et al. [34] used the DenseNet121 [11] architecture, which was shown to perform well on chest X-ray images with a CheXnet DenseNet121 model [28]. They used Adam optimization with default parameters values ($b1$ = 0.9 and $b2$ = 0.999), a learning rate of 0.001, and a learning rate decay of 0.1 when the validation accuracy converged. They achieved an AUC of 98.4%.

**Table 3.** Comparative results for other models on same test dataset. Bold numbers indicate best performance.

| Model | Recall (%) | Precision (%) | AUC (%) | F-Score | Accuracy (%) |
|---|---|---|---|---|---|
| Kermany et al. [32] | 93.2 | - | 96.8 | - | 92.8 |
| Cohen et al. [34] | - | - | 98.4 | - | - |
| Rajaraman et al. [35] | 96.2 | **97.7** | **99.3** | 0.970 | 96.2 |
| Sahlol et al. [24] | 87.22 | - | - | - | 94.18 |
| Saraiva et al. [36] | 94.85 | 95.72 | - | 0.953 | 95.07 |
| Ayan and Über [37] | 82 | - | - | - | 87 |
| Sharma H. et al. [38] | - | - | - | - | 90.68 |
| **Our model(SE-Attention)** | **98.46** | 96.24 | 96.03 | **0.973** | **96.63** |

Rajaraman et al. [35] used customized models based on a pre-trained CNN model as the feature selector. They selected specific regions of interest (ROI) on chest X-ray images to perform the classification of pneumonia and normal images. They evaluated models with two types of dataset, original data as baseline and cropped ROI data. The best model was customized VGG16 and we outperformed their model in terms of accuracy, F-score, and recall. Cohen et al. and Rajaraman et al. attempted to detect pneumonia with the customized CNN model. Sahlol et al. [24] used a MobileNet as the feature extractor and the AEO algorithm as the feature selector. The AEO algorithm finds only the relevant features from a lot of features that are extracted from a MobileNet. Saraiva et al. [36] used two neural networks, the multilayer perceptron and neural network to detect and classify the pneumonia. They achieved best results in recall, precision, F-score and accuracy, of 94.85%, 95.72%, 0.953% and 95.07%, respectively. Ayan and Über [37] also used deep-learning-based methods Xception and VGG16 for pneumonia classification. The VGG16 model outperformed Xception with an accuracy of 87%. Sharma et al. [38] proposed two CNN architectures that were designed from scratch with or without a dropout layer. They used deep CNN architectures as the feature extractor. The accuracy of the results was 90.68%. Despite other papers not providing as detailed results as ours for evaluation, we outperformed other methods in terms of classification accuracy, recall and F-score.

## 4. Discussion

It is widely known that as the model gets deeper, the number of parameters increases. Thus, the flexibility of the model increases and the training error decreases. However, the generalization error increases due to over-fitting to the test data. Sufficient data are needed for the model to classify pneumonia and normal. Distinguishing between normal organs and pathological signs requires medical professionals that have the relevant knowledge. Therefore, large medical datasets with correct answer labels are expensive. To solve this problem, we used three pre-trained models from different domains (ImageNet, Chest X-ray14, Custom dataset). DenseNet121 was trained to predict 14 anomalies including pneumonia on chest X-ray images. Among the above pre-trained models, it was confirmed that using the model based on DenseNet121 had the highest accuracy. The domain of the test dataset is similar to that of the chest X-ray14 data set. When combining the models with

an attention mechanism, it means that the final model will also show good performance if the combined transfer learning model is pre-trained with data from a domain similar to the domain of the target task. The test data for this paper was 5,856 images, which is insufficient to train a model that used over a million images such as ResNet and DenseNet. With data preprocessing, our model can train with low-quality input images and reach good performance. However, good quality input images as input can improve the performance of our model. There are limitations where there may not be a pre-trained model using data from domains similar to our target task. In this case, it is expected that the performance can be improved either by using pre-trained models trained on various datasets, or by using more models. Future work on our study will add a study related to domain adaptation that can achieve domain changes by maintaining high performance in the original domain. We will also have to study how many pre-trained models are desirable to use. Our framework is not the only one used to diagnose pneumonia. Simply, our framework is to look at chest X-rays and classify them as pneumonia or normal with the hope that it could be used as an aid to medical expert diagnosis. At the same time, our framework is expected to add credibility to the diagnosis of pneumonia.

## 5. Conclusions

In this paper, we proposed an attention-based transfer learning framework for efficient pneumonia detection in chest X-ray Images. We used various models which were pre-trained in various domains and these models take on the role of feature extractor. After concatenating feature vectors extracted from pre-trained models, we applied the attention mechanism to classify pneumonia and normal. As for the transfer learning result of the single model, ResNet152, DenseNet121, and ResNet18 showed 95.03%, 95.35%, and 94.87% accuracy, respectively, whereas the result of using SE-Attention by combining all these three models was 96.63% for the Guangzhou Women and Children's Medical Center dataset. In other words, according to the methodology we have proposed, we can create a model that can be applied to new tasks by collecting pre-trained models from various domains and using the attention mechanism. We expect this insight can be applied to a wide range of fields that require domain adaptation. This study may be used in studies or products that highlight areas of suspected pneumonia or produce pneumonia-related annotations.

**Author Contributions:** Conceptualization, B.K.; methodology, B.K.; software, S.-M.C., S.-S.L.; validation, S.-M.C., S.-S.L.; formal analysis, S.-M.C., S.-S.L.; investigation, S.-M.C., S.-S.L.; resources, B.K.; data curation, S.-M.C.; writing—original draft preparation, S.-M.C.; writing—review and editing, S.-M.C.; visualization, S.L.; supervision, B.K.; All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. 2019R1G1A110070412) and also by Chonnam National University under Grant number: 2020-2010.

**Institutional Review Board Statement:** The NIH chest X-ray datasets were approved by the National Institutes of Health Institutional Review Board(IRB), and the Guangzhou chest X-ray datasets were exempt from IRB review. The requirement for informed consent was waived.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Restrictions apply to the availability of these data. Data was obtained from ImageNet, National Institute Of Health, GitHub repository and are available http://image-net.org/download, https://stanfordmlgroup.github.io/projects/chexnet/, https://github.com/anilsathyan7/pytorch-image-classification with the permission of ImageNet, National Institute Of Health, anilsathyan7.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References


1. Gilani, Z.; Kwong, Y.D.; Levine, O.S.; Deloria-Knoll, M.; Scott, J.A.G.; O'Brien, K.L.; Feikin, D.R. A literature review and survey of childhood pneumonia etiology studies: 2000–2010. *Clin. Infect. Dis.* **2012**, *54*, S102–S108. [CrossRef]
2. World Health Organization. WHO Reveals Leading Causes of Death and Disability Worldwide: 2000–2019. Available online: https://www.who.int/news/item/09-12-2020-who-reveals-leading-causes-of-death-and-disability-worldwide-2000-2019 (accessed on 26 December 2020).
3. Esayag, Y.; Nikitin, I.; Bar-Ziv, J.; Cytter, R.; Hadas-Halpern, I.; Zalut, T.; Yinnon, A.M. Diagnostic value of chest radiographs in bedridden patients suspected of having pneumonia. *Am. J. Med.* **2010**, *123*, 88.e1–88.e5. [CrossRef]
4. ReliasMedia. Misdiagnosis of Flu Instead of Pneumonia Results in Death for 10-Year-Old Girl. Available online: https://www.reliasmedia.com/articles/17222-misdiagnosis-of-flu-instead-of-pneumonia-results-in-death-for-10-year-old-girl (accessed on 11 January 2021).
5. Elemraid, M.A.; Muller, M.; Spencer, D.A.; Rushton, S.P.; Gorton, R.; Thomas, M.F.; Eastham, K.M.; Hampton, F.; Gennery, A.R.; Clark, J.E.; et al. Accuracy of the interpretation of chest radiographs for the diagnosis of paediatric pneumonia. *PLoS ONE* **2014**, *9*, e106051. [CrossRef] [PubMed]
6. Kallianos, K.; Mongan, J.; Antani, S.; Henry, T.; Taylor, A.; Abuya, J.; Kohli, M. How far have we come? Artificial intelligence for chest radiograph interpretation. *Clin. Radiol.* **2019**, *74*, 338–345. [CrossRef] [PubMed]
7. Wang, X.; Peng, Y.; Lu, L.; Lu, Z.; Bagheri, M.; Summers, R.M. ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2097–2106.
8. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet Classification with Deep Convolutional Neural Networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
9. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
10. 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, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
11. 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, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
12. Varshni, D.; Thakral, K.; Agarwal, L.; Nijhawan, R.; Mittal, A. Pneumonia detection using CNN based feature extraction. In Proceedings of the 2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), Coimbatore, India, 20–22 February 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1–7.
13. Liang, G.; Zheng, L. A transfer learning method with deep residual network for pediatric pneumonia diagnosis. *Comput. Methods Programs Biomed.* **2020**, *187*, 104964. [CrossRef] [PubMed]
14. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
15. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
16. Bengio, Y. Deep learning of representations for unsupervised and transfer learning. In Proceedings of the ICML Workshop on Unsupervised and Transfer Learning, Washington, DC, USA, 2 July 2012; pp. 17–36.
17. Pratt, L.Y.; Mostow, J.; Kamm, C.A.; Kamm, A.A. *Direct Transfer of Learned Information Among Neural Networks*; AAAI: Menlo Park, CA, USA, 1991; Volume 91, pp. 584–589.
18. Cao, C.; Liu, X.; Yang, Y.; Yu, Y.; Wang, J.; Wang, Z.; Huang, Y.; Wang, L.; Huang, C.; Xu, W.; et al. Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 2956–2964.
19. Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 2048–2057.
20. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
21. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11534–11542.
22. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826.
23. Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; He, K. Aggregated residual transformations for deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1492–1500.
24. Sahlol, A.T.; Abd Elaziz, M.; Tariq Jamal, A.; Damaševičius, R.; Farouk Hassan, O. A novel method for detection of tuberculosis in chest radiographs using artificial ecosystem-based optimisation of deep neural network features. *Symmetry* **2020**, *12*, 1146. [CrossRef]

25. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.
26. Zhao, W.; Wang, L.; Zhang, Z. Artificial ecosystem-based optimization: A novel nature-inspired meta-heuristic algorithm. *Neural Comput. Appl.* **2019**, 1–43. [CrossRef]
27. Sahlol, A.T.; Yousri, D.; Ewees, A.A.; Al-Qaness, M.A.; Damasevicius, R.; Abd Elaziz, M. COVID-19 image classification using deep features and fractional-order marine predators algorithm. *Sci. Rep.* **2020**, *10*, 1–15. [CrossRef] [PubMed]
28. Rajpurkar, P.; Irvin, J.; Zhu, K.; Yang, B.; Mehta, H.; Duan, T.; Ding, D.; Bagul, A.; Langlotz, C.; Shpanskaya, K.; et al. Chexnet: Radiologist-level pneumonia detection on chest X-rays with deep learning. *arXiv* **2017**, arXiv:1711.05225.
29. Anil Sathyan. Pytorch-Image-Classification. 2019. Available online: https://github.com/anilsathyan7/pytorch-image-classification (accessed on 20 December 2020).
30. Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv* **2014**, arXiv:1409.0473.
31. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In *Advances In Neural Information Processing Systems*; Curran Associates Inc.: Red Hook, NY, USA, 2017; pp. 5998–6008.
32. Kermany, D.S.; Goldbaum, M.; Cai, W.; Valentim, C.C.; Liang, H.; Baxter, S.L.; McKeown, A.; Yang, G.; Wu, X.; Yan, F.; et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* **2018**, *172*, 1122–1131. [PubMed]
33. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
34. Cohen, J.P.; Bertin, P.; Frappier, V. Chester: A Web Delivered Locally Computed Chest X-ray Disease Prediction System. *arXiv* **2019**, arXiv:1901.11210.
35. Rajaraman, S.; Candemir, S.; Kim, I.; Thoma, G.; Antani, S. Visualization and interpretation of convolutional neural network predictions in detecting pneumonia in pediatric chest radiographs. *Appl. Sci.* **2018**, *8*, 1715. [CrossRef] [PubMed]
36. Saraiva, A.A.; Santos, D.; Costa, N.J.C.; Sousa, J.V.M.; Ferreira, N.M.F.; Valente, A.; Soares, S. Models of Learning to Classify X-ray Images for the Detection of Pneumonia using Neural Networks. In Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2019), Prague, Czech Republic, 22–24 February 2019; pp. 76–83.
37. Ayan, E.; Ünver, H.M. Diagnosis of Pneumonia from Chest X-ray Images Using Deep Learning. In Proceedings of the 2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT), Istanbul, Turkey, 24–26 April 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1–5.
38. Sharma, H.; Jain, J.S.; Bansal, P.; Gupta, S. Feature Extraction and Classification of Chest X-ray Images Using CNN to Detect Pneumonia. In Proceedings of the 2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 29–31 January 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 227–231.