Attention based X-ray Image Denoising with Non-Uniform Noise Distribution

G011 (S2569758, s2534699)

Abstract

Medical imaging as a field has gained significant traction in recent times due to its applications, but one of the biggest problems that remain in these images is the aberrations that appear due to apparatus limitations or other environmental factors. X-rays are also not immune to such imperfections, but getting rid of this noise can result in efficient diagnostics, aiding the enhancement of image quality and accuracy in medical assessments. Previous attempts at generating noisy datasets for X-rays make an assumption of a uniform scatter of noise on the image. We induce an uneven scatter of noise in images, more towards the centre of the X-ray, where the interaction of X-rays with tissues is greater. We train multiple incremental updates to our model on the NIH X-ray dataset after inducing noise using the core concepts of encoder-decoder architecture. We also propose an end-to-end pipeline stacking the two best-performing models. Despite training on a subset of only 10k images, the resultant denoised images seem to be promising. The objective evaluation of the model is carried out using the Mean Squared Error (MSE), the Structural Similarity Index (SSIM), and the Peak Signal-to-Noise Ratio (PSNR).

1. Introduction

Computer vision models in the field of medical imaging have made great strides in recent times due to the emergence of high-power computation devices and the abundance of data. Image denoising, in particular, is crucial because it helps address the issues related to medical images, like noise reduction and aiding resolution limitation. X-ray images are formed by electromagnetic radiation passing through the body and the intensity being attenuated based on the composition and the density of tissues. X-rays are often noisy because of a multitude of factors like exposure settings, detector noise, and scatter radiations, which contribute to noise and reduced contrast of the image. X-ray denoising can thus help generate better images that can further be trained for classification and detection purposes.

The usual noise patterns that frequently appear in the X-ray images are Gaussian blurring, Poisson noise, Salt and Pepper noise, and speckle noise. We chose the NIH X-ray dataset to retrieve X-ray images and induced the noise specified above in them manually to simulate problems that might occur in real-life X-rays. Conventionally, image denoising was dealt with using filtering techniques such as median filtering and average filtering, but deep learning approaches to denoising have surpassed these conventional models. With deep learning, the models are very modular; thus, the same model can deal with noises from different types and still generalise well to generate a denoised output.

Deep Convolutional Neural Networks (CNN) like AlexNet (Krizhevsky et al., 2012) and ResNet (He et al., 2016) became the base for deep learning-based image denoising. The advantage of using a CNN is that it can deal with high-density noises very well due to their local receptive fields. Variations of vanilla CNN architectures are often used to upscale the images and get rid of the noise in the image. Encoder-decoder architectures offer great avenues to deal with such tasks in which the input is an image, and the expected output from the model is also an image of similar dimensions. The encoder-decoder architecture has the ability to learn underlying patterns in the data and thus help in re-forming the image efficiently.

We initially train a simple model using shallow encoders and decoders to define a performance baseline for our experimentation. We then proceed to explore more complex architectures like SkiDNet (Dutta et al., 2019), which was designed with the specific task of denoising images. To further improve the quality of the presented image, we push the output of the SkidNet model through another architecture which utilised Deep Attention (Li et al., 2022) to upscale the image and create a final reconstructed image. The details of both the model architectures, along with the working of the entire pipeline, are described in Section 4.

The advancements in X-ray denoising can significantly improve diagnostic accuracy while also supporting the development of efficient assisted automation tools. The generation of clean and precise X-ray images helps in the automation of tasks like disease detection, organ segmentation, and anomaly detection. This automation will significantly reduce the load on radiologists and medical professionals, allowing them to tend to more patients and provide them with better care. Moreover, by integrating the DL-based denoising techniques into medical imaging systems, real-time processing of the images becomes possible, reducing the need to create a separate task and making the entire process more streamlined.

2. Related work

Due to the importance of X-rays in the fields of medicine, security, and industrial inspection, over the years, continuous work has been done to improve the quality of X-rays. X-ray images are often corrupted by noise, which can degrade image quality and hinder the accuracy of subsequent tasks. Along with the improvement of the existing technology to take X-ray scans, several denoising techniques aim to mitigate this issue by removing unwanted noise while preserving the important features of the image.

Before the use of Deep Learning, diverse methods for reducing noise while maintaining key image properties were provided by math-based image denoising techniques, which may accommodate a range of noise characteristics and application scenarios. (Rudin et al., 1992) proposed Total Variance (TV) denoising, which smoothes regions with low gradient values while maintaining sharp edges by minimising the total variance of the picture gradient. The K-SVD algorithm (Elad & Aharon, 2006) is an example of a sparse representation-based denoising technique that makes use of the sparsity quality of natural images. Iteratively learning a dictionary of atoms, the algorithm effectively represents image patches while promoting sparsity in representation coefficients. Non-Local Means (NLM) (Buades et al., 2005) denoising uses weighted averages of similar patches to eliminate noise from images while retaining fine features and textures. The techniques of wavelet shrinkage, which adaptively thresholds wavelet coefficients based on their statistical properties, enable effective noise reduction while preserving image details.

Some of the recent techniques to denoise images include utilising Deep Convolutional Neural Networks (CNNs) and residual learning (Zhang et al., 2017). The methodology involves training a deep CNN to directly learn the residual mapping between noisy and clean images, enabling the network to focus on modelling the complex noise distribution. The proposed network architecture consists of multiple convolutional layers with skip connections, facilitating the extraction of both low-level and high-level image features. Another method introduced by uses a Deep Denoising network (Chen et al., 2018), which is integrated with residual learning and perceptual loss to prioritise edge preservation using the high-frequency layer of noisy images as input. A residual mapping predicts the difference between clean and noisy images, guided by a joint loss function that emphasises edge reconstruction alongside pixel-to-pixel Euclidean loss. The perceptual loss leverages a well-trained convolutional neural network to capture semantic information, encouraging the network to reconstruct edges and details effectively rather than merely matching low-level pixel values.

Talking specifically about medical images, a unique approach was established that used spatial filtering and wavelet domain filtering (Dong et al., 2020) to address different types of noise in X-ray images. Using the biorthogonal double wavelet transform, the method soft thresholds wavelet coefficients to create a new coefficient matrix for image reconstruction, resulting in less noise and more picture detail. Low-dose CT has sparked interest in the medical field, prompting a method to denoise the image obtained from the CT scan. This resulted in the residual encoder-decoder convolutional neural network (RED-CNN) (Chen et al., 2017), which combines an autoencoder, a deconvolution network, and shortcut connections to provide low-dose CT imaging and upscale the output image. This network can be retrained to upscale X-ray images as well. Another strategy based on CT image denoising is employing a generative adversarial network (GAN) (Goodfellow et al., 2014) with Wasserstein distance and perceptual similarity (Yang et al., 2018). Wasserstein distance improves GAN performance by exploiting optimal transport theory, whereas perceptual loss compares denoised output features to ground truth to preserve essential information. The method incorporates visual perception knowledge into denoising, resulting in excellent noise reduction while keeping important image details.

3. Dataset and task

For the task of denoising, we take X-ray images from the NIH X-ray dataset (Wang et al., 2017). The dataset consists of approximately 110k frontal-chest x-ray images from around 30k unique patients. The data is diverse since it has multiple scans for a patient, including the left and right lung scans, along with scans which are obstructed by medical equipment like tubes. The typical dimension of an X-ray image is 3000×3000, but this imposes challenges in terms of computing hardware, and thus, the dataset is downscaled to the size 1024×1024. For our use case, though, we try to work on a sample of 10,000 images and downsample the images further, as covered in Section 4.

Some of the X-rays have some medical equipment like tubes in them. They do not appear at a consistent location in all the images and sometimes are over a particular part of a lung as well. A few images also have a label of L or R in either the left or right top corner, which helps the doctor identify which side of the lung is represented on which side. Similar to labels, a select few also include certain symbols like an up arrow or a serial number in them. A selection of the X-rays has not been able to capture one lung properly, which results in the X-ray having one complete lung and one distorted or missing lung in the scan. Numerous X-rays are already unclear and have very little detail about the lungs. Some of the X-rays are not centred in the image, creating a lot of empty space in them. A combination of these discrepancies ensures that the model trained on them will work on a variety of X-ray formats.

4. Methodology

4.1. Data Processing

The NIH dataset consists of mostly clear chest X-rays with no noise, which might only be the case sometimes. X-ray imaging is often plagued with noises, as mentioned earlier, due to the very nature of how it is generated. The absorption patterns of the X-ray can have distortions due to multiple factors like the exposure settings of the environment or even the apparatus being used. Thus, we preprocess the X-ray dataset by simulating noise frequently occurring in medical images. We incorporate the following noises in the dataset after downscaling them to 256 × 256 before feeding it to our model.

4.1.1. Gaussian Noise

Gaussian noise is the most frequent type of noise that appears in signals. It is a statistical noise with the probability distribution function of the normal form, as shown in the equation below -

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \tag{1}$$

4.1.2. Poisson Noise

The photons that strike the receptor surface cannot be forced to be evenly distributed, i.e. a particular region might receive more. The noise density in this disparity often forms a Poisson distribution.

4.1.3. Salt and Pepper (SAP) Noise

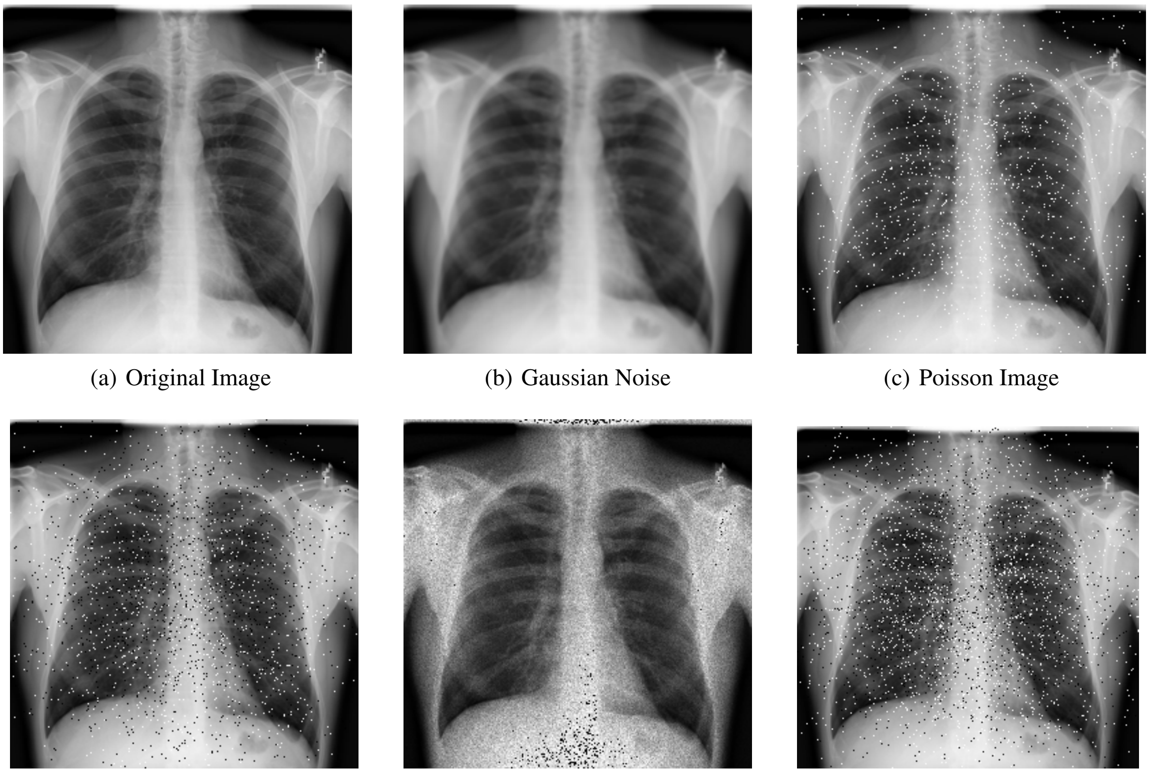
SAP noise is one of the most frequently occurring noises in images. It can arise due to malfunctioning machines, errors in transmissions, or even conversion errors induced from analogue to digital. The image generated is filled with random white and black pixels scattered across it.

4.1.4. Speckle Noise

Speckle noise obscures the underlying features in an image by introducing a change in the brightness or intensity at the pixel level in a random fashion.

All previous approaches to X-ray denoising (Krizhevsky et al., 2012) talk about scattering the noise over the entire image. The impact of the rays usually happens around the centre of the image, and thus, the absorption would be distorted more towards the centre. Therefore, the noise we induce should also follow a similar pattern.

In order to achieve this, we try to multiply all our noise with a Gaussian mask. We generate this mask using a 2-D Gaussian plot of the same dimension as that of the image with centre at the pixel coordinates x=112, y=112 and σ = 56 px. Once this Gaussian mask is generated, we generate a mask with random values from 0 to 1 and the same dimensions as our image. Based on the values of the random mask, we binarise the Gaussian mask if, at a given pixel, the Gaussian value exceeds the random mask value. Since the Gaussian would have higher values towards the centre, our noise, when multiplied with the binary Gaussian mask, would be more focused towards the centre.



(a) Original Image (b) Gaussian Noise (c) Poisson Image

(d) Salt & Pepper Noise (e) Speckle Noise (f) Random Noise

Figure 1. Original Image and different kinds of noise applied on it

4.2. Types of Models

The idea of Convolution Neural Networks (CNN) had existed since the late 80s(LeCun et al., 1998) and was primarily pitched for the problem statement of image classification. The vanilla CNN architectures thus end with a fully connected layer, which gives the output probability distribution over all classes. For our use case, since the task at hand is to regenerate a denoised image, directly using the vanilla approach is not feasible. Variations of the conventional techniques, though, can result in upscaling and producing similar-sized output images. In order to generate images, one of the most common architecture forms used is an encoder-decoder architecture. The primary inspiration for an encoder-decoder architecture came from the Sequence-to-sequence models developed for the purpose of machine translation (Sutskever et al., 2014). The core idea for the problem of denoising is that an encoder model would run on the image input and condense the representation of the image in a fixed-dimension feature vector. The decoder model then takes this feature vector to generate an image as close to the ground truth as possible.

The two major architectures which have been primarily used for the task of image denoising are Fully Convolutional Neural Networks (FCN) (Long et al., 2015) and U-Net (Ronneberger et al., 2015), which were proposed for the task of image segmentation. The fully connected layers from a CNN model are trimmed from the model and replaced with up-convolution layers to reconstruct the dimensions of the original image. This vanilla encoder-decoder architecture takes no input from the encoder architecture and thus results in valuable information getting lost, resulting in bad generations. With the inspirations from the skip connections introduced in the ResNet (He et al., 2016), skip connections were introduced to merge every convolution in the encoder to the corresponding deconvolution layer in the decoder, giving a more feature-rich representation to regenerate images (Mao et al., 2016). These skip connections help in the training of the Deep CNN as they enhance the capability of the model to capture the low-level features like the textures, edges and colour gradients. These features are then combined with high-level features like shapes and structures. These high-level features are captured during the initial layers of the encoder.

Both FCN and U-Net exhibit similar architectural details except for the process of combining these feature representations from encoder to decoder. FCN incorporates a simple element-wise summation of the two feature maps as both are alike in dimensions(Long et al., 2015). U-Net, however, combines it using a concatenation operation (Ronneberger et al., 2015). The problem with summation operation is that it can lead to information loss, while concatenation does a much better job at preserving spatial information. We try to train a shallow encoder-decoder network with summation and concatenation operations and prove that in this context, the U-Net-based architecture outperforms the FCN architecture. We propose incremental updates to the U-Net architecture to refine our denoising model further, as discussed below, and combine two best-performing models to come up with an aggregated model, more of which is discussed in Section 5.

4.2.1. Skip Image Denoising Networks for X-Rays (SkidNet)

SkidNet was proposed by Dutta et al. for the task of x-ray image denoising (Dutta et al., 2019). The model was trained on a uniform distribution of noise across the image, which we tweaked to train the model with uneven distribution focused on the centre of the x-ray frame. The encoder architecture of SkidNet reduces the spatial dimensions of the image by a series of convolution and max pooling blocks. Precisely five convolution layers with kernel size 3×3 are used with filters altering between 32 and 64. A block of two convolutions is followed by a Max Pooling layer of dimension 2×2, which helps in reducing the spatial dimensions of the image by 16 times. We use the Rectified Linear Unit to induce non-linearity in our model.

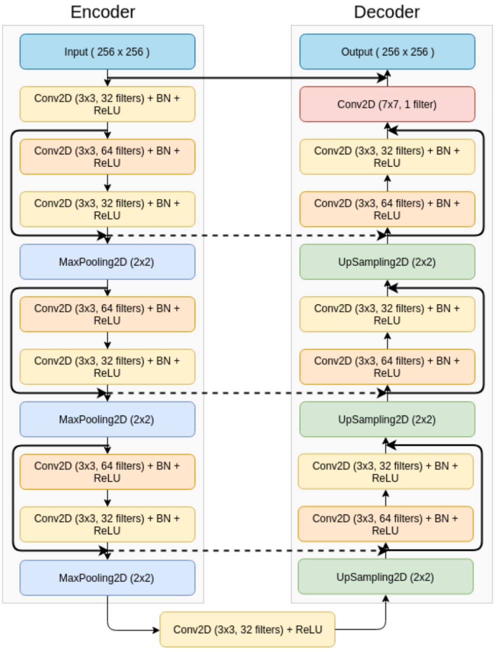

Encoder
Decoder
Input (256 x 256)
Conv2D (3x3, 32 filters) + BN + ReLU
Conv2D (3x3, 64 filters) + BN + ReLU
Conv2D (3x3, 32 filters) + BN + ReLU
MaxPooling2D (2x2)
Conv2D (3x3, 64 filters) + BN + ReLU
Conv2D (3x3, 32 filters) + BN + ReLU
MaxPooling2D (2x2)
Conv2D (3x3, 64 filters) + BN + ReLU
Conv2D (3x3, 32 filters) + BN + ReLU
MaxPooling2D (2x2)
Output (256 x 256)
Conv2D (7x7, 1 filter)
Conv2D (3x3, 32 filters) + BN + ReLU
Conv2D (3x3, 64 filters) + BN + ReLU
UpSampling2D (2x2)
Conv2D (3x3, 32 filters) + BN + ReLU
Conv2D (3x3, 64 filters) + BN + ReLU
UpSampling2D (2x2)
Conv2D (3x3, 32 filters) + BN + ReLU
Conv2D (3x3, 64 filters) + BN + ReLU
UpSampling2D (2x2)
Conv2D (3x3, 32 filters) + ReLU

Figure 2. SkidNet Model Architecture (Dutta et al., 2019)

The decoder architecture takes the final representation generated by the encoder and takes it through a series of convolution and Upsampling layers. Similar to the Max Pooling, Upsampling is applied after every alternate convolution block. The final convolution layer in the decoder is of dimensions 7×7 to give a resultant image of the same dimension as the input image. The dimensions of the convolution kernel and the number of them remain consistent with the encoder to ensure the preservation of spatial information since we use the skip connections to concatenate the corresponding feature maps from encoder to decoder (Dutta et al., 2019). These skip connections also help in better propagation of gradients through the network, aiding the problem of vanishing gradients.

Additionally, SkidNet also induces short-term skip connections, which are connections formed from within the encoder network to the encoder network, and the same is true for the decoder. These connections are exactly similar to the ResNet (He et al., 2016) and help further in enhancing the training of the model and giving it a faster convergence. We add short-range skip connections after every two convolution layers, as shown in Figure 2. These shot range skip connections inspired by ResNet perform a summation operation between the two feature maps. Further details of the experimentation are listed in Section 5 .

4.2.2. Deep Attention Architecture

This model has a similar architecture to SkidNet. The filter size and max pool size remain the same across both the models, but the number of filters in this model varies from 3 to 96, doubling at every convolution block (Li et al., 2022). Each encoder convolution block has two sets of convolution layers: batch normalization layer, an activation using leaky Rectified Linear Unit (Xu et al., 2015), and dropout as the regularization metric.

This model, though, removes the short-range skip connections, and the long-range skip connections pass through attention gates(AG) (Oktay et al., 2018) to have a say in the decoder architecture. Attention gates were primarily introduced for language modelling-related tasks (Vaswani et al., 2017) but have found great relevance in the field of computer vision as well. As we move through the layers, the receptive field, i.e. the area of the image a neuron sees, increases. Attention gates help to focus on the relevant parts of an image while suppressing the background parts of images. Since, in our case, noise is more scattered towards the centre of the picture, attention gates seem like an ideal solution to deal with this type of noise.

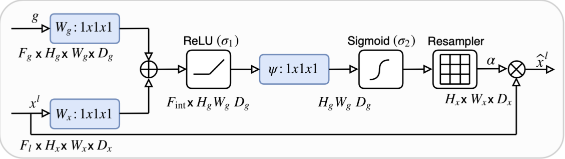


Figure 3. Attention Mechanism

We find attention coefficients that help identify salient image regions. Once we have an idea of the important regions, the irrelevant information is pruned using the attention gate. The gate only retains the activations which are identified as important by the coefficients. The output of this attention gate is a multiplication of each value in the feature map with the corresponding attention coefficient. Before we merge this long-range skip connection with the main path, AGs ensure that only relevant activations are passed. The neuron activations are filtered using AGs both during the forward and backward passes of training. The gradients from the background regions are downweighted, thus ensuring that the shallow layer gets updates based on only the relevant regions. Since a process of pruning is performed, the computational load is reduced, thereby resulting in faster training.

We design multiple experiments to find the optimal hyperparameters and test the viability of the models discussed above, more of which are discussed in Section 5.

4.2.3. Proposed Methodology

Although the SkidNet model's primary task is to denoise the X-ray images, it requires significant training to achieve its desired goal. On the other hand, the Deep Attention model's primary use is to increase the resolution of the images, with the attention mechanism helping it prioritise the sections of the images it needs to focus on. This model has significantly fewer parameters, and the attention gate mostly aids its performance. Both models require significant training to fine-tune the weights to achieve a proper denoised image.

We propose an aggregated model pipeline which is able to combine the best ability of both models while also being resource-efficient while training. We first train the SkidNet model on the artificially noised images to generate the original image for 30 epochs to get a basic model. This model was then used to create a dataset of partially denoised images for the Deep Attention model. This new dataset was then used to train the Deep Attention model, which further denoises the partially denoised images from the SkidNet model to create a proper denoised image output. The advantage of this pipeline lies in utilising both the denoising capability of the SkidNet model and the upscaling power of the Deep Attention model. This pipeline has been tested rigorously against individual models, and the results are displayed in Section 5.

4.3. Evaluation

We need ways to define how well a model performs when contrasted with one another. The metrics that can be assessed for the task of denoising can be defined at three levels: at a pixel level, in terms of the structural arrangement of generated regions in terms of how they are perceived, and in terms of semantic consistency.

4.3.1. Mean Squared Error (MSE)

MSE is a very commonly used metric to compare the differences between the images. It measures the average squared difference between the corresponding pixel values of the original image and the processed image. MSE is widely employed in image processing tasks such as image denoising, compression, and restoration. For our task, the MSE is calculated as follows:

$$\text{MSE}_x = \frac{1}{N}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}[S(i,j) - R(i,j)]^2, N = m * n \qquad (2)$$

Where S and R are the noisy image and cleaner output image, and x represents the channel.

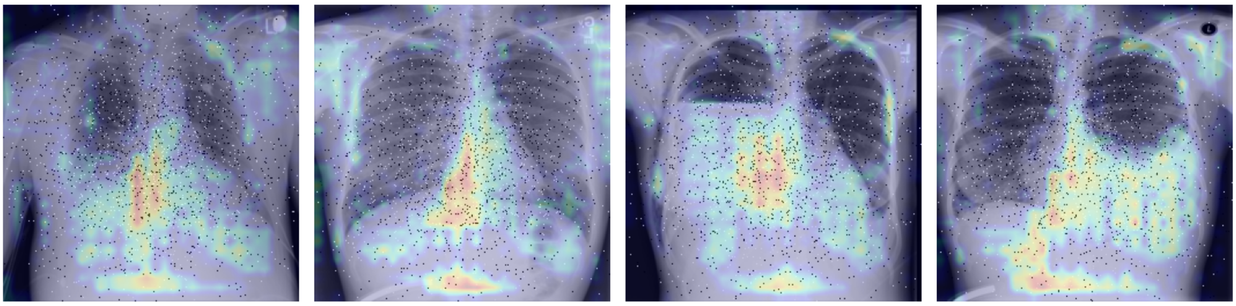


Figure 4. Focus of Attention Mechanism in Deep Attention Model (Li et al., 2022)

4.3.2. Peak Signal to Noise Ration (PSNR)

PSNR is one of the most commonly available metrics to evaluate a denoising model, and it deals with the evaluation at a pixel level. The first step is to calculate the mean squared error (MSE) using Equation 2. Next, we just use a logarithmic transformation on the sum of MSE values, as shown below

$$\text{MSE}_t = \sum_{x \in channels} \text{MSE}_x \tag{3}$$

$$\text{PSNR} = 10 * \log_{10} \frac{(\text{MAX}_I)^2}{\text{MSE}_t} \tag{4}$$

The higher the PSNR score, the better the predicted image. However, since it is at a pixel level, the results are difficult to align with a human's perception of the output.

4.3.3. Structural Similarity Index Measure (SSIM)

SSIM aligns better with human perception of image closeness. The SSIM metric takes similarity using the contrast, luminance, and pixel arrangement. Luminance tries to measure the difference in average brightness. The pixel arrangement is calculated using the correlations and patterns between the two images. The higher the score of SSIM, the better the structural similarity between the two images. The SSIM between 2 images is calculated using the formula:

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{5}$$

where, x and y are the images being compared, μ_x and μ_y are the average luminance of images x and y respectively, σ_x^2 and σ_y^2 are the variances of images x and y respectively, σ_{xy} is the covariance of images x and y and c_1 and c_2 are constants to stabilize the division with weak denominators.

5. Experiments

We have tested a sample of the NIH dataset (10,000 images out of 120,000) on a variety of different models to establish current baseline performance and then improve on it with our proposed architecture. We designed a simple Autoencoder model with one encoder and one decoder layer to establish a baseline performance. We used ReLU activation and Max Pooling at the end of each layer to control the gradients. The performance was not great and it was barely able to denoise the lungs in the X-ray images.

We then train the SKidNet model, which was designed with the denoising of medical images in mind. Another architecture that was also tested was the Deep Attention model, which used the attention mechanism in a UNet architecture model to improve the model's performance. The hypothesis of using attention was that our noise is not uniformly spread across the image, and attention does a better job of focusing on specific areas of the image. The hypothesis is validated with Figure 4, which is a plot of the average gradients after normalisation from the last attention module. Both the model architectures are described above in Section 4.2. Finally, the models were used one after the other to form an aggregated model. This step was done to create an end-to-end pipeline which helps us use each model for a specified task – the SkidNet model for removing the noise in the image and the Deep Attention model to further enhance the images generated by the SkidNet model. Both models were trained separately for efficiency and not to let the parameters of one model affect the parameters of the other. Both the models were trained for 25 epochs to generate the denoised images. This is a fraction of what they were trained on in their respective original papers.

Since we are trying to teach the model to generate accurate images based on the input, we use the Mean Squared Error

Table 1. Comparison of Metrics across various models with random noise

	Mean Squared Error (MSE)		Peak Signal-to-Noise Ratio (PSNR)		Structural Similarity (SSIM)	
	Validation	Test	Validation	Test	Validation	Test
SkidNet	4.743×10^{-3}	5.381×10^{-3}	25.922	25.708	0.8870	0.8832
Deep Attention	1.150×10^{-3}	1.241×10^{-3}	25.589	25.527	0.8666	0.8661
Aggregated Model	$\mathbf{1.959 \times 10^{-4}}$	$\mathbf{1.965 \times 10^{-4}}$	**26.348**	**26.317**	**0.9450**	**0.9433**

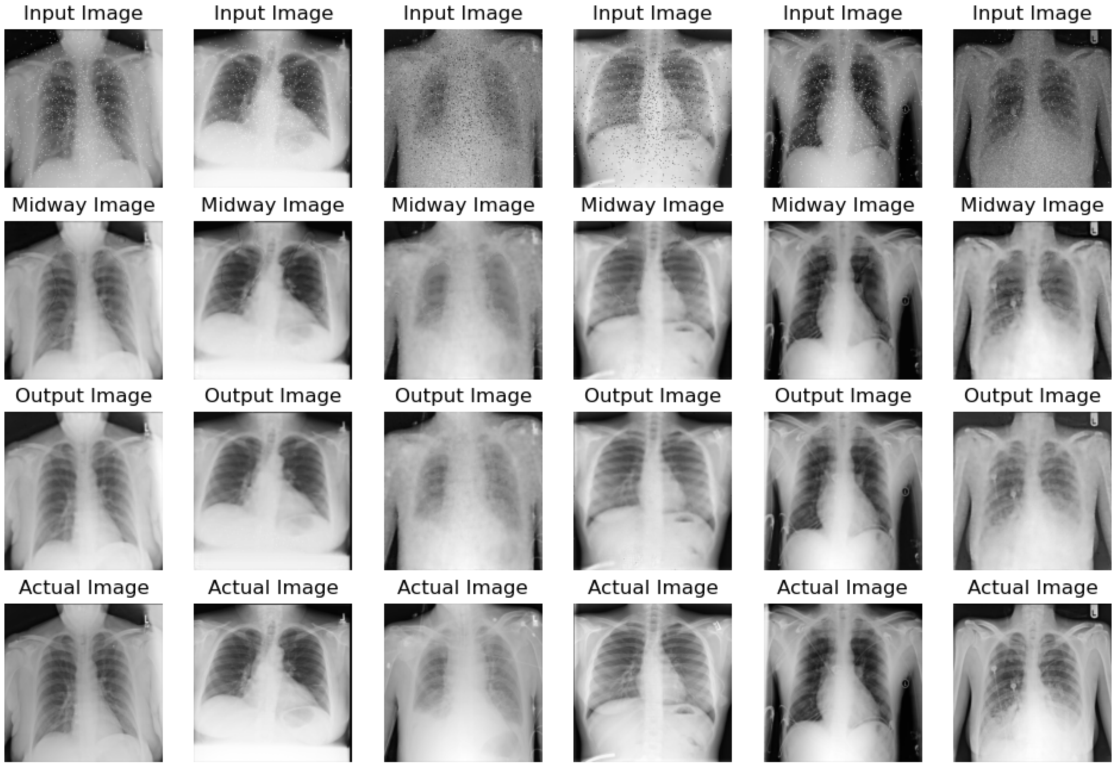


Figure 5. The entire pipeline output for six different scans. The input image represents the artificially noised image. The midway image represents the output of the SkidNet model. The output image is the final output image after the Midway image is pushed through the Deep Attention model. Finally, the last row shows the actual image.

Loss (MSE) in all the models to compare each pixel from the generated image to the source image. Finally, we go with the Adam optimiser for the model as it dynamically modifies the learning rate based on the momentum of the training. The initial learning rate of the models is set to 0.001 to ensure the model takes small jumps in the beginning while trying to reduce the loss. We use the PSNR metric to evaluate the model's performance on the development set to make sure that the model does not overfit the training set and generalise well to all sorts of variations in the data. We also used SSIM to track the quality of the generated images.

Table 1 shows the performance of different models trained on identical randomly noised data evaluated on the evaluation set. These models were then tested on a test set to see the true capabilities of all the models. Set seeds were used to control the randomness of selecting the image for each set and applying the same kind of random noise to them. We can see that the aggregated model outperforms the lightly trained SkidNet and Deep Attention models.

The validation loss is helpful in the training process by letting the model know how to adjust the weights to achieve the best possible output. However, the loss and the other metrics alone can't be the defining metric for how we judge the model because the input and output are grayscale images. This means that the generated images will have very similar pixel values, resulting in low loss throughout. Thus, we manually look at actual images generated by the model based on an image given to it as input. This gives us a better understanding of how the model has learned the features of the image and how effective it is in recreating it.

From the generated images (Figure 5), we can see that the aggregated model is very adept at removing the noise from the images. It is able to capture a lot of detail in recreating the image while still removing the noise from the X-rays. Naturally, the input images with less noise generally have more detail in their reconstructed form, including the labels of *L* or *R*, if they originally had them. The rib cage is mostly discernible in all of the reconstructed images, except when the original X-ray was already distorted, and the addition of noise made the X-ray completely unclear. The model is very able to handle blank spaces in the image if the X-ray is not centred and recreates the X-ray in the same part of the image. Poisson and Speckle noise generally change the tone of the image to a slightly different shade, but the model is able to revert it to its original tone while still reconstructing the blurred details in the case of the Poisson noise. In the case of Salt & Pepper noise, the model is clearly able to learn from pixels around a noisy pixel and fill it with an appropriate colour. In case other bones (like arm bones) are present in the image, the model has no trouble recreating them because most of our noise is in the centre of the image, where the most distortion would occur naturally. The model struggles the most with Speckle noise, failing to recreate the details of the lungs and only managing to recreate a rough outline of them with lots of distortion still available

Table 2. Ablation Study on Aggregated Model

	Mean Squared Error (MSE)		Peak Signal-to-Noise Ratio (PSNR)		Structural Similarity (SSIM)	
	Validation	Test	Validation	Test	Validation	Test
No Noise	5.151×10^{-5}	5.139×10^{-5}	25.575	25.565	0.9902	0.9902
Only Gaussian Noise	1.496×10^{-4}	1.437×10^{-4}	26.045	26.029	**0.9751**	**0.9753**
Only Poisson Noise	$\mathbf{8.672 \times 10^{-5}}$	$\mathbf{8.591 \times 10^{-5}}$	26.196	26.183	0.9725	0.9725
Only Speckle Noise	2.265×10^{-4}	2.249×10^{-4}	26.537	26.537	0.9300	0.9298
Only Salt & Pepper Noise	1.215×10^{-4}	1.197×10^{-4}	26.387	26.380	0.9603	0.9604
All Noises Combined	4.315×10^{-4}	4.225×10^{-4}	**26.660**	**26.647**	0.8839	0.8839

in the recreated image.

We then carried out an ablation study to check how the model performs with each kind of noise in the image to quantitatively justify the qualitative results we saw while manually looking at the image. Table 2 shows how the aggregated model performs when there is no noise in the images, a specific kind of noise, or all the noises are present together. The process of artificially noising the image was modified so that each type of noise was applied to every image instead of randomly applying it to certain images.

We can see that, individually, the model struggles the most with Speckle Noise. It distorts the image comprehensively, and the model is relatively unable to generate the level of detail it can with other noises. Looking at the numbers of Salt & and Pepper noise, we can deduce that the model is able to correct most of these pixels with a value very close to the actual value required in the pixel but not the same. This tiny variation across several salt and pepper pixels might contribute to a small performance dip compared to other noises. The model handles Gaussian and Poisson noise well, recreating the details and correcting the tone of the image. It means that the model has learned to determine the correct pixel range of the image and adjust all pixel values accordingly. This leads to high PSNR and SSIM numbers and lower MSE loss. When all of the noises are applied together to an image, the model is still able to recreate the lungs with a high PSNR but dips to the lowest SSIM. The major reason behind this can be a mixture of the noises impacting the true value of a particular pixel in comparison to other pixels. Despite these challenges, the model has still learned how to denoise an image from a combination of these noises.

6. Conclusions

We have presented a new approach to combine the powers of two separate architectures to denoise chest X-ray images. The SkidNet architecture uses an encoder-decoder UNet architecture with varying lengths of skip connections to denoise the image significantly. After that, the Deep Attention model uses its Attention Mechanism to enhance the output of the SkidNet model. We demonstrate the ability of this aggregated pipeline to effectively adapt to different noises, namely Gaussian, Salt and Pepper, Poisson and Speckle noise, even when they are combined in one image. The most impressive feat here is the efficiency of this solution, with the models boasting comparable MSE, PSNR & SSIM numbers in very little time, utilising a fraction of the resources or the data required by the original models. Thus, we achieved our task of denoising chest X-ray images using deep learning techniques, which is a prevalent problem in the medical industry.

Both models have impressive mechanisms which can be utilised for other applications as well. The next step to improve the model would be to introduce short-term skip connections with attention gates in both the encoder and decoder architectures to make an end-to-end trainable pipeline, making the process more efficient. Other optimisation techniques can further be explored to make this model more efficient. Even though this aggregated model is evaluated on X-ray images, with the process of transfer learning, this pipeline can also be used to denoise X-ray images of different body parts or even other kinds of medical images. We implemented this on a small batch of brain MRI images, and the results are shown in Figure 6.

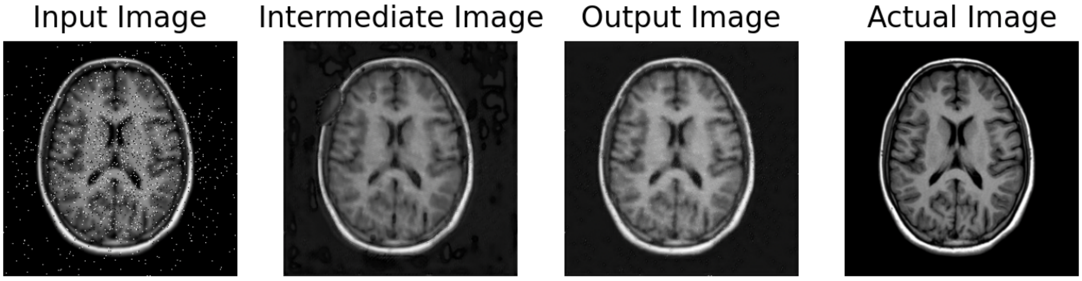


Figure 6. The entire pipeline evaluated on brain MRI images by the process of transfer learning.

References

Buades, A., Coll, B., and Morel, J. M. A review of image denoising algorithms, with a new one. *Multiscale modeling simulation*, 4(2):490–530, 2005. ISSN 1540-3459.

Chen, Hu, Zhang, Yi, Kalra, Mannudeep K., Lin, Feng, Chen, Yang, Liao, Peixi, Zhou, Jiliu, and Wang, Ge. Low-dose ct with a residual encoder-decoder convolutional neural network. *IEEE transactions on medical imaging*, 36(12):2524–2535, 2017. ISSN 0278-0062.

Chen, Xi, Zhan, Shu, Ji, Dong, Xu, Liangfeng, Wu, Congzhong, and Li, Xiaohong. Image denoising via deep network based on edge enhancement. *Journal of ambient intelligence and humanized computing*, 14(11):14795–14805, 2018. ISSN 1868-5137.

Dong, Hanlei, Zhao, Liguo, Shu, Yunxing, and Xiong, Neal N. X-ray image denoising based on wavelet transform and median filter. *Applied mathematics and nonlinear sciences*, 5(2):435–442, 2020. ISSN 2444-8656.

Dutta, Sandipan, Chaturvedi, Shaurya, Kumar, Swaraj, and Bhatia, MPS. Skidnet: Skip image denoising network for x-rays. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2019. ISBN 9781728119854.

Elad, Michael and Aharon, Michal. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image Processing*, 15(12): 3736–3745, 2006. doi: 10.1109/TIP.2006.881969.

Goodfellow, Ian J, Pouget-Abadie, Jean, Mirza, Mehdi, Xu, Bing, Warde-Farley, David, Ozair, Sherjil, Courville, Aaron, and Bengio, Yoshua. Generative adversarial networks. *arXiv.org*, 2014. ISSN 2331-8422.

He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, and Sun, Jian. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778. IEEE, 2016. ISBN 9781467388511.

Krizhevsky, Alex, Sutskever, Ilya, and Hinton, Geoffrey E. Imagenet classification with deep convolutional neural networks. In Pereira, F., Burges, C.J., Bottou, L., and Weinberger, K.Q. (eds.), *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

LeCun, Yann, Bottou, Léon, Bengio, Yoshua, and Haffner, Patrick. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, volume 86, pp. 2278–2324, 1998. URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.42.7665.

Li, Bryan M., Castorina, Leonardo V., Valdés Hernández, Maria del C., Clancy, Una, Wiseman, Stewart J., Sakka, Eleni, Storkey, Amos J., Jaime Garcia, Daniela, Cheng, Yajun, Doubal, Fergus, Thrippleton, Michael T., Stringer, Michael, and Wardlaw, Joanna M. Deep attention super-resolution of brain magnetic resonance images acquired under clinical protocols. *Frontiers in Computational Neuroscience*, 16, 2022. ISSN 1662-5188. doi: 10.3389/fncom.2022.887633. URL https://www.frontiersin.org/articles/10.3389/fncom.2022.887633.

Long, Jonathan, Shelhamer, Evan, and Darrell, Trevor. Fully convolutional networks for semantic segmentation, 2015.

Mao, Xiaojiao, Shen, Chunhua, and Yang, Yu-Bin. Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/0ed9422357395a0d4879191c66f4faa2-Paper.pdf.

Oktay, Ozan, Schlemper, Jo, Folgoc, Loic Le, Lee, Matthew, Heinrich, Mattias, Misawa, Kazunari, Mori, Kensaku, McDonagh, Steven, Hammerla, Nils Y, Kainz, Bernhard, Glocker, Ben, and Rueckert, Daniel. Attention u-net: Learning where to look for the pancreas, 2018.

Ronneberger, Olaf, Fischer, Philipp, and Brox, Thomas. U-net: Convolutional networks for biomedical image segmentation, 2015.

Rudin, Leonid I., Osher, Stanley, and Fatemi, Emad. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1):259–268, 1992. ISSN 0167-2789. doi: https://doi.org/10.1016/0167-2789(92)90242-F. URL https://www.sciencedirect.com/science/article/pii/016727899290242F.

Sutskever, Ilya, Vinyals, Oriol, and Le, Quoc V. Sequence to sequence learning with neural networks, 2014.

Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N, Kaiser, Łukasz, and Polosukhin, Illia. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Wang, Xiaosong, Peng, Yifan, Lu, Le, Lu, Zhiyong, Bagheri, Mohammadhadi, and Summers, Ronald M. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3462–3471. IEEE, 2017. ISBN 1538604574.

Xu, Bing, Wang, Naiyan, Chen, Tianqi, and Li, Mu. Empirical evaluation of rectified activations in convolutional network, 2015.

Yang, Qingsong, Yan, Pingkun, Zhang, Yanbo, Yu, Hengyong, Shi, Yongyi, Mou, Xuanqin, Kalra, Mannudeep K., Zhang, Yi, Sun, Ling, and Wang, Ge. Low-dose ct image denoising using a generative adversarial network with wasserstein distance and perceptual loss. *IEEE transactions on medical imaging*, 37(6):1348–1357, 2018. ISSN 0278-0062.

Zhang, Kai, Zuo, Wangmeng, Chen, Yunjin, Meng, Deyu, and Zhang, Lei. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE transactions on image processing*, 26(7):3142–3155, 2017. ISSN 1057-7149.