Exploring Layerwise Relavance Propagation algorithm on chest radiography

Jón Rúnar **Mahbub Ul Alam** **Yuxia Wang**

Abstract

Explainable deep networks have increasingly been studied and applied in various fields in recent years, in which the Layer-wise Relevance Propagation (LRP) algorithm is a typical representation of visualization. This project aims to explore this algorithm from reproducing the reference paper experiments to applying it to the imaging examination task of chest radiography and finally comparing this LRP algorithm with the Grad-CAM method. The results show that reproducing LRP on MNIST provides us a balanced understanding of the relevant pixel areas to identify the digits. LRP heatmaps provide us granular heatmaps comparing with Grad-CAM heatmaps when applied to the Chexpert dataset classification model. We posit that this is due to the inherent construction difference of these algorithms (LRP is layer-wise accumulation, whereas Grad-CAM focuses mostly on the final parts in the model's architecture), and both can be useful for understanding the classification from a micro or macro level. The Github repo for the codes used can be found here: Github.com/Jonrunar95/DD2412-Project

1 Introduction

Deep learning methods have increasingly been studied and applied in various fields, including medical diagnosis. However, these methods usually work as a black-box, due to the complexity and non-transparency of the models, making it difficult to give reliable and satisfactory justifications for the results. Deep network explanation methods aim to produce explanations that make it easier to understand how deep networks came to its decision. Many methods have been proposed in recent years.

According to [1], the current methods for explainable deep networks can be classified into three categories. (a) Visualization methods, by highlighting the characteristics of input data which have a bigger influence on the output. (b) Model distillation, which develops a simper model to mimic the input-output behavior of a Deep Neural Network (DNN). (c) Intrinsic methods, which create special DNNs to render an explanation along with its prediction. According to [2], the explanation models can also be classified into local or global ones depending on if it explains only a single prediction (local) or all the outputs (global). Another way to classify the models is model-specific or model-agnostic regarding the application.

Chest radiography plays a very important role in disease diagnosis and medical care, and currently, there are huge amounts of images produced each year. Therefore, how to interpret these images effectively with deep learning models is of high importance and at the same time a challenging task.

This project aims to explore the algorithm of Layerwise Relevance Propagation (LRP) [3], a visualization method, and apply it to the imaging examination task of chest radiography. Specifically, A large chest-radiography dataset with uncertainty labels - CheXpert [4] is applied to do our experiment based on deep Convolutional Neural Network (CNN) architectures, and the results are analyzed and visualized using the LRP algorithm. Additionally, the Gradient-weighted Class Activation Mapping

(Grad-CAM)[5] algorithm , another visualization method, is used for the same task, and the outputs of the two algorithms are compared.

2 Related work

In recent years, deep learning has been successfully applied to multiple medical imaging tasks, including skin cancer classification, lymph node metastases detection and pulmonary tuberculosis classification[6]. For chest radiography with explainable deep learning methods, many of current works are based on the explanation methods of class activation mappings (CAM) [7] or Grad-CAM. P. Rajpurkar et al. [8] proposed a 121-layer convolutional neural network, trained on the ChestX-ray14 dataset, called CheXNet. CheXNet can detect pneumonia from chest X-rays at a level exceeding practicing radiologists. Further in [6], the model of CheXNet is developed to detect 14 pathologies, which focuses on multi-pathology detection. Both papers applied CAM heatmaps to interpret the algorithm. A chest X-ray assisted classification model using a deep neural network is presented in [9], which interprets the abnormality of the chest X-ray film using Grad-CAM. In [10], both saliency maps and Grad-CAMs are used as tuberculosis visualization methods.

In [11], Layer-wise Relevance Propagation (LRP) is used to visualize convolutional neural network decisions for Alzheimer's disease (AD) based on structural magnetic resonance imaging (MRI) data. This argues that LRP may have good potential to assist clinicians in explaining neural network decisions. Chlebus et al. [12] proposes an approach to explain semantic segmentation networks, utilizing layer-wise relevance propagation, which investigates the importance of input MRI sequences for the task of automatic liver tumor segmentation. While the LRP algorithm explains very well on other image classification tasks [3], we would like to explore how it performs on examining chest radiography images.

3 Methods

3.1 LRP algorithm

LRP algorithm is introduced in [3], which calculates the relevance of each neuron, from the output to the input, in a layer-wise manner, based on the backpropagation algorithm. This method belongs to the visualization methods, and it assumes that the models can be decomposed into several layers of computation. For neural networks, LRP can be defined as

$$R_i^l = \sum_j R_{i\leftarrow j}^{(l,l+1)} = \sum_j \frac{z_{ij}}{z_j} R_j^{l+1} \tag{1}$$

We denote this equation as LRP_z, in which a lower layer relevance for the i_{th} neuron R_i^l is computed based on the upper layer relevance R_j^{l+1}, here $z_{ij} = x_i w_{ij}$ denotes the localized pre-activations and $z_j = \sum_i z_{ij} + b_j, b_j$ is a bias term. Usually, there are different rules which satisfy this formulation, in this reference paper the two rules LRP_ϵ and $LRP_{\alpha,\beta}$ are presented separately:

$$R_{i\leftarrow j}^{(l,l+1)} = \begin{cases} \frac{z_{ij}}{z_j+\epsilon} R_j^{l+1}, for z_{ij} \geq 0 \\ \frac{z_{ij}}{z_j-\epsilon} R_j^{l+1}, for z_{ij} \leq 0 \end{cases} \tag{2}$$

$$R_{i\leftarrow j}^{(l,l+1)} = R_j^{l+1} (\alpha \frac{z_{ij}^+}{z_j^+} + \beta \frac{z_{ij}^-}{z_j^-}) \tag{3}$$

Using these rules, the overall relevance can be obtained for each neuron in the lower layer. When it is applied to deep ReLU (Rectified Linear Unit) networks, LRP can be understood as a Deep Taylor decomposition [13] of the prediction.

In this project, the explanation methods of LRP_z, LRP_ϵ, $LRP_{\alpha,\beta}$ and Deep Taylor decomposition are implemented based on [3].

3.2 Grad-CAM

Grad-CAM [5] (Gradient-weighted Class Activation Mapping), also belongs to the visualization methods. It evaluates the importance of each neuron in the input, by using a feature map which uses the gradient information propagated to the last convolutional layer of a CNN network to produce a heatmap. Given a class score y_c, for a class a c, the importance of each neuron is evaluated by calculating the gradient of y_c with regards to the activation A. The gradient is used to calculate an importance weight α_k^c for each feature map k, in the convolutional layer, for class c.

$$\alpha_k^c = \frac{1}{z} \sum_i \sum_j \frac{\partial y_c}{\partial A_{ij}^k} \tag{4}$$

The heatmap is then produced by a weighted combination of the activation maps. A ReLU function is used to remove the negative values, since we are only interested in the positive ones.

$$L_{Grad-CAM}^c = ReLU(\sum_k \alpha_k^c A^k) \tag{5}$$

4 Dataset

In this project, two datasets are selected, one is the MNIST (Modified National Institute of Standards and Technology) database [14] containing images of hand-written digits. MNIST database is used to reproduce the experiments in the original LRP paper [3].

The second daset is Chexpert (**Ch**est e**Xpert**) [4]. CheXpert is a large dataset of chest X-rays, which comes from Stanford University Medical Center between October 2002 and July 2017. It consists of 224,316 chest radiographs of 65,240 patients, which are labeled as positive, negative, or uncertain. Based on the observations of pathology, this dataset is also divided into 14 subgroups, as shown in Table 1. The original high-resolution image containing dataset is about 439 gigabytes in size which was not feasible for us to use. Therefore, we used a downsampled version of the same datest provided by Stanford machine learning group [4]. The size is about 11 gigabytes.

5 Experiments and findings

In this project, the work is done by the following three steps: 1) reproducing the LRP algorithm based on MNIST dataset as described in [3]; 2) exploring LRP explanation method on CheXpert dataset; and 3) comparing the performance of LRP to Grad-CAM method.

5.1 Reproducing LRP on MNIST

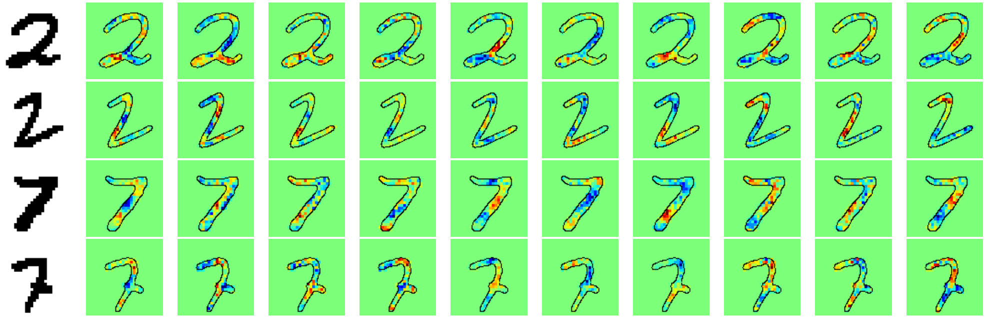


Figure 1: Heatmaps for all the ten predicted classes (zero to nine from left) for the input image two and seven

The implementation works for the *MNIST experiments I and II* as described in [3]. *MNIST experiments I* uses a multi-layer perceptron (MLP) architecture or fully connected layers, with three hidden layers having 400 neurons each. The 784-dimensional input layer represents the MNIST image with a size of 28*28 pixels for each image. The activation function is tanh, which takes a real-valued input and squashes it to the range $[-1, 1]$. The standard error-back propagation algorithm is used for training the network. The input data is normalized as zero mean and with unit average. The mini-batch size is selected as 25, with a total of 50000 training iterations. In each iteration, the randomly chosen training samples per batch were added with an additional Gaussian noise layer. Figure 1 is showing the results obtained using equation 1. Here, the prediction for all ten digits (from zero to nine, incrementally) is shown beside the digit's original input image. The accuracy obtained from the test data is 98.25%.

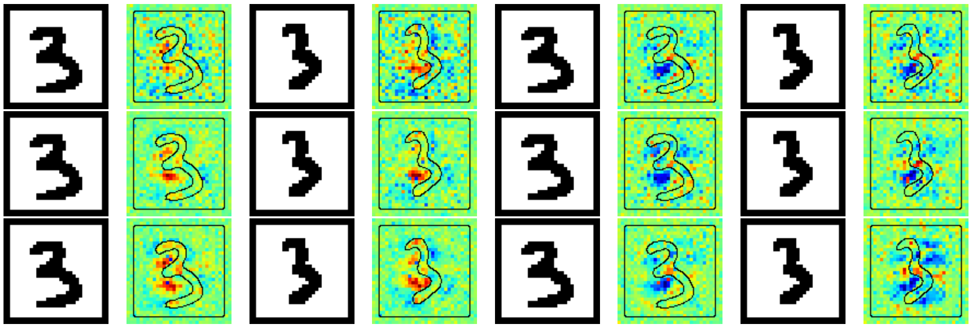


Figure 2: Heatmaps for three different activations, tanh (top row), ReLU (middle row), and ReLU* (bottom row). Every odd positioned image is the original input image of digit three. Heatmaps in position two, and four represent the prediction against class three. Heatmaps in position six, and eight represent the prediction against class eight.

MNIST experiments II uses similar three-layered feed-forward network architecture where each layer contains 1296 neurons. Both the tanh and ReLU (rectified linear unit) activation functions were used. ReLU takes a real-valued input and thresholds it at zero (replaces negative values with zero). The training data was augmented with ±4 pixels in vertical and horizontal directions. Input data was also augmented with 10% salt-and-pepper noise. Finally, it was also scaled between [-1,1]. The mini-batch size is similar to the previous experiment as 25. Two total training iteration numbers were used: 1000000 and 10000, with the test accuracy as 99.1% and 96.03% (with the ReLU activation layer, with tanh activation layer and 1000000 iterations it is 99.01%) respectively. Figure 2 is showing the results for three cases, (i) tanh, and (ii) ReLU with 1000000 iterations, and (iii) ReLU with 10000 iterations (denoted as ReLU*). The digit three is shown with the heatmap representation of the predicted classes three and eight. Similarly, in Figure 9 the input image four is shown along with the heatmaps of predicted classes four and nine. Figure 10 is showing ten randomly sampled digits and their all class-based heatmap representations in sorted order.

We used RGB (read-green-blue) heatmap representation to demonstrate the relevance where high or positive scores are denoted with hot (orange to red) colors. Neutrality is represented by green color, and low or negative scores are shown with blue colors. These colors represent the relevance of the class of input among the different pixels.

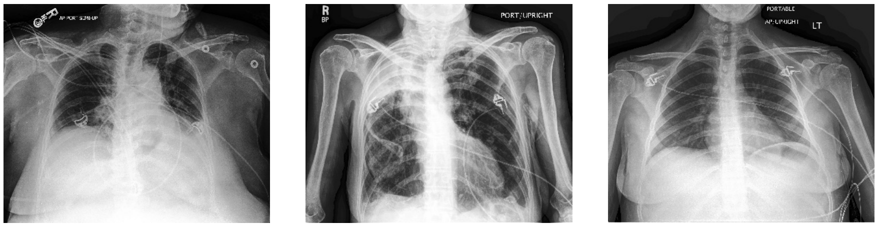
Figure 1 provides a glimpse of the inner mechanism of the model as we can see precisely which area (surrounding pixels) is responsible for identifying a particular digit (two or seven, which can be quite similar). We can observe that it is because different regions are selected for these two cases. Likewise, in Figure 2 and 9, we can see some other examples of confusing digits (four or nine, and three or eight), and we can observe that the bigger model performs better in identifying the correct digits. It is interesting to note in Figure 2 that the same area that is positive to identify three is negative to identify eight. Finally, in Figure 10, we can see a random list of examples to observe a generalized result. For some digits (for example, zero and one), the result is not entirely conclusive. In other cases, it is quite conclusive. Overall, MNIST experiments provide us with a balanced overview of the model's classification process and general idea about the LRP algorithm in an intuitive manner.

5.2 Exploring LRP on CheXpert

We considered the classification task as a multi-label classification and viewed all the uncertain labels as negative. To visualize the classification results in the CheXpert dataset, first, we trained the model with the data using a pre-trained DenseNet-121 (Dense Convolutional Network) [15] architecture. The input images were scaled to 224*224 pixels to suit the DenseNet-121 pre-trained architecture. The input pixel values were scaled by deviding them by 255 and then casting to RGB values. We used the ImageNet [16] weight for the pretraining. Adam optimizer is used for backpropagation with a batch size of 16. The total number of epochs used is 100.

Table 2 shows the evaluation of the model using AUROC (area under the receiver operating characteristic), which demonstrates the trade-off between the true-positive rate (TPR) and false-positive rate (FPR) across different decision thresholds. Given the high-imbalance in instances among different classes, as shown in Table 1, the result is sufficient for the visualization task.

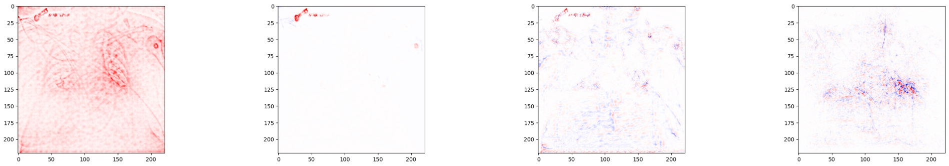
As these pathology or classes are restricted mostly to some distinct part of the radiology (chest-Xray) images, we selected some classes to provide the medical terminology with examples. Later we will demonstrate how the heatmaps are resembling the diagnosis. Based on the results in Table 2, we selected three classes. The first one is selected as the better-predicted class in terms of evaluation, *Pleural Effusion*. The second one is from the worse prediction range, *Pneumothorax*. The third one is based on the ambiguity to check how well LRP performs in such a case, *No Finding*.



(a) *Pleural Effusion* (b) *Pneumothorax* (c) *No Finding*

Figure 3: Example of the selected classes

Pleural Effusion can be denoted as the excess fluid accumulating in the surrounding fluid-filled space of the lungs (this space is denoted as a pleural cavity). Figure 3a is one such example where the left lung is affected, as can be seen in the chest X-ray image. *Pneumothorax* indicates that the lung is collapsed when there is an air leakage inside the space between the lung and chest wall. It can be a complete collapse or a portion of the lung. Figure 3b shows an example of it (left lung). *No Finding* indicates that no associated pathology has been identified in the chest X-ray image. We can think of it as a healthy person's chest X-ray to some degree. Figure 3c is showing one such example.

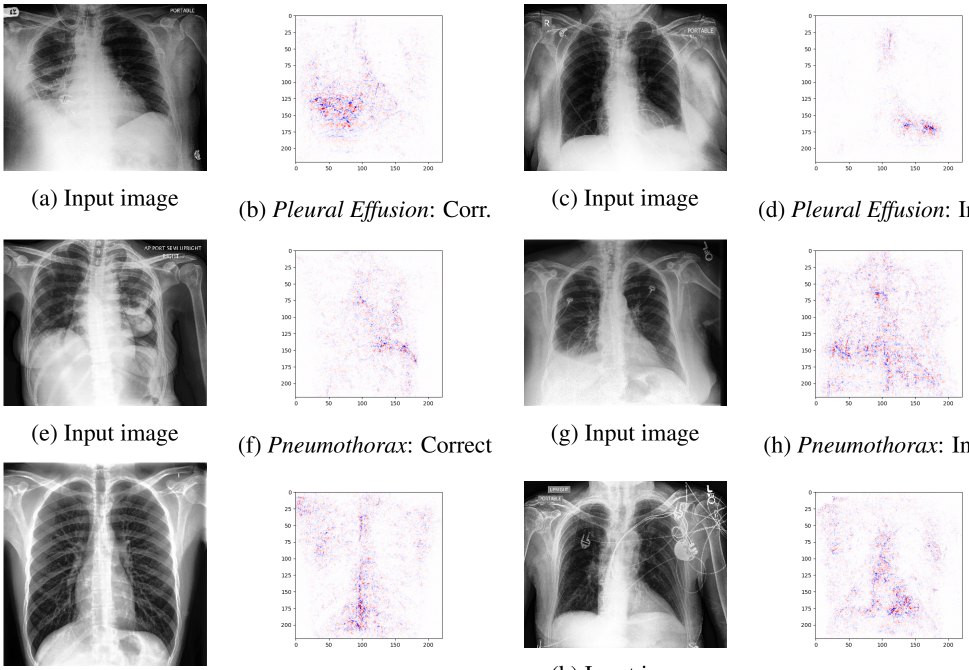


(a) *Deep Taylor* (b) $LRP_{\alpha,\beta}$ (c) LRP_{ϵ} (d) LRP_{z}

Figure 4: Different LRP heatmaps for the correct prediction of *Pleural Effusion* class generated from the image 3a

We have applied LRP_z, LRP_ϵ, $LRP_{\alpha,\beta}$ and deep Taylor decomposition algorithms to generate heatmaps. It has been observed that LRP_z provides better quality heatmaps in terms of understanding. Figure 4, shows such examples where the prediction class is *Pleural Effusion* based on the original image (under the left lung) shown in Figure 3a.

Figure 5 shows two examples of the heatmaps for the three selected predicted classes along with their original input images, one correct prediction, and another incorrect prediction for each prediction class. We can observe that for the *Pleural Effusion* class the correct heatmap (5b) rightly identifies the pleural cavity area under the right lung (5a). The incorrect prediction (5d) focuses wrongly on the left lung's pleural cavity area (5c). Figures in the row from 5e to 5h are showing the results of the classification heatmaps based on *Pneumothorax* class. For the correct case, we can observe that (5f) the heatmap focuses on the punctured left lung. For the incorrect prediction (5h), it was unable to show the right lung area. Figures in the row from 5i to 5l are showing the results of the classification heatmaps based on *No Finding* class. It is a bit tricky as there are no specific parts in general that



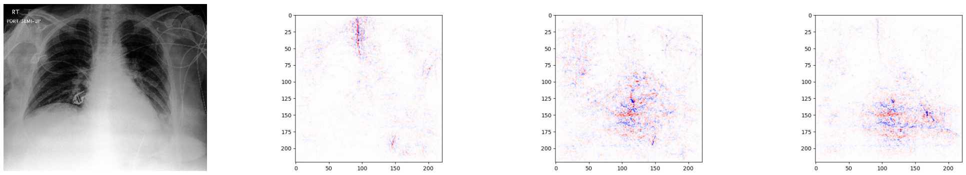
(a) Input image (b) *Pleural Effusion*: Corr. (c) Input image (d) *Pleural Effusion*: Incor.

(e) Input image (f) *Pneumothorax*: Correct (g) Input image (h) *Pneumothorax*: Incor.

(i) Input image (j) *No Finding*: Correct (k) Input image (l) *No Finding*: Incorrect

Figure 5: Example heatmaps of the correct and incorrect predictions for three different classes along with original images using LRP_z algorithm

could be selected for the positive classes. We can observe that for the correct case (5j), the heatmaps are pointing to the lower area of the chest, and for the incorrect prediction (5l), it focuses on the lower right area. It is a bit inconclusive, intuitively summarizing the findings.



(a) Input image (b) *Support Devices* (c) *Lung Opacity* (d) *Pleural Effusion*

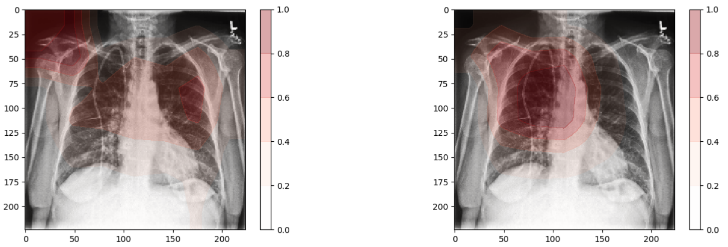
Figure 6: Example heatmaps of the correct multi-label predictions along with original images using LRP_z algorithm

Chexpert dataset is multi-labeled, which means there could be multiple correct predictions for a particular input image. Previous examples were showing the top predicted class only. Figure 6 is showing one instance where all the top-three predicted classes are correct. We can observe that LRP is providing different heatmaps for all these three classifications. *Support Devices* indicates any additional presence of devices. We can observe (b) that the wire on the neck area is shown in the class's heatmap. *Lung Opacity* indicates the decrease in the ratio of gas to soft tissue in the lung. We can see the heatmap (c) is pointing correctly to the left lung. Finally, *Pleural Effusion* is correctly shown on the (d) heatmap pointing to the left lung.

Overall, LRP provides a balanced overview explaining the classifications using the Chexpert dataset. It is worth mentioning that, unlike the image classification task, where edge detection can be a crucial factor in identifying correct classes or objects, the classification process is much more nuanced. For example, several classes are primarily specific to the lungs part. Therefore it gets difficult to pinpoint the particular region which is not entirely distinct from each other. Moreover, it is a multi-label, multi-class classification task, making the heatmap generation process more complicated.

The denseNet-121 architecture consists of a lot of layers. Therefore it was interesting to see the performance of LRP there, which is sufficient.

5.3 Performance comparing of LRP and Grad-CAM

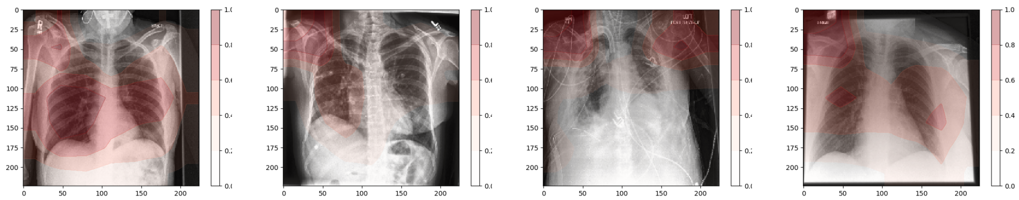


(a) Lung Opacity (b) Support Devices

Figure 7: Different Grad-CAM heatmaps of a correct multi-label classification

The same DenseNet-121 model was used to perform the Grad-CAM algorithm. The Grad-CAM heatmaps are larger and contain less details on which pixels motivate the classification. An example of different Grad-CAM generated heatmaps for the same image can be seen in Figure 7. The heatmap in Figure 7b has a wide circle around the support device while the one in Figure 6b, produced using LRP, shows mostly just the pixels that show the device. Same goes for the lung tissue in Figures 7a and 6c.

An example of correct and incorrect classification for the *No Finding* class can be seen in Figure 8a and 8b. For the correct classification, the motivation behind the prediction is the whole image, with the most emphasis on the lungs. That makes sense since the whole image needs to be considered in order to motivate the classification, and most of the diseases in this dataset are in the lungs. For the incorrect prediction, the motivation behind it is just a small section of the image and does not put emphasis on the lungs. The opposite is true for Figure 8c and 8d. The correct classification bases the prediction on areas that do not motivate the output. On the other hand the incorrect classification is basing the decision on the correct area, but finds evidence for an incorrect prediction. A reason for the incorrectly motivated correct prediction could be that Figure 8c is different from most other images. The cause could be that the image is of lower quality, different between X-Ray settings, or the anatomy of the individual produced an unusual image. This example of a badly motivated prediction should motivate further investigation. All of the examples in Figure 8 give insight into the capabilities of the DenseNet-121 model.



(a) No Findings - Corr. (b) No Findings - Incor. (c) Pneumothorax - Corr. (d) Pneumothorax - Incor.

Figure 8: Example heatmaps of the correct (corr.) and incorrect (incor.) predictions for *No Findings* class and the *Pneumothorax* imposed with original images using Grad-CAM algorithm

The difference between the heatmaps produced by LRP and Grad-CAM in this paper is big, but both can give value to the investigation of a model. The LRP heatmaps have a more detailed explanation of what pixels motivate the output. For that reason, it could be considered better for a more detailed analysis and explanation of what motivated the output and where the model went wrong. The Grad-CAM heatmaps present a bigger area, which can give a quick explanation of whether the model is focusing on the correct areas of the image. For a well-defined problem like this, where the presence of a disease is usually in a specific area, that can be extremely helpful in finding where the model is doing well and where it goes wrong.

6 Challenges

The first challenge we faced is on deciding the dataset. Naturally, our first choice was to use the standard dataset; unfortunately, due to its sheer size (approximately 539 gigabytes) and our computational resource limitation, we had to settle with the low-resolution one. As the classification is quite nuanced, higher resolution in input images would have provided better heatmaps. The same constraint is also impacted by deciding the training architecture. Although training based on the transfer learning approach, in this case, provides us entirely satisfactory results, the choice of a suitable pre-trained model, and appropriate architecture could have impacted the overall performance of the heatmap visualizations. We could use a simpler architecture and train it from scratch, but it could compromise the performance issue. Moreover, due to the time and resource constraint, it was not feasible.

The different LRP algorithms provided very different heatmaps for the same inputs and classes, as shown in Figure 4. LRP_ϵ algorithm provides almost non-existing heatmaps, and the Deep Taylor decomposition algorithm indicates the opposite. Only LRP_z performed consistently. The choice of correct hyper-parameters (for example, tuning of ϵ, α, and β values) could be a reason here. The main challenge is to have the appropriate amount of fusion of in domain (in our case, medical) knowledge and LRP algorithm(s) suitability for that particular case. To our knowledge, it is highly empirical; therefore, it requires a significant amount of fine-tuning, which is quite challenging.

Another challenge we faced on deciding how to represent the multi-label classification in terms of heatmaps. We showed example (Figure 6) of it. There is a trade-off between the best classification and multiple classifications. The heatmaps would have been more intuitive if we used a single-label classification. We wanted to explore the more complex situation to analyze how the LRP and Grad-CAM would perform in such a situation.

One could notice that there is no negative indication in Grad-CAM-based heatmaps, unlike LRP (blue is indicated as negative). It could be a bit challenging to interpret the images in some cases because of it. However, as it is mentioned earlier, we could view these two algorithms as a collaborative visualization process. Grad-CAM is relatively better in identifying the general relevant area, and then we can utilize LRP to get a more fine-grained overview.

7 Conclusion

In this work, we focus on the LRP method based on deep neural networks. Three parts of work have been done. We have reproduced the first two MNIST experiment, done in [3], getting comparable results. We implemented LRP on a DenseNet-121 architechture, trained on the CheXpert dataset. The results were positive and gave good insight into the DensNet-121. We also did a comparison between LRP and Grad-CAM, providing analysis on the difference between the methods and the heatmaps they produced.

Through this work, we make a full comprehension on the LRP algorithm and get a full picture of explainable deep learning methods. The experimental results of these methods applied to medical diagnosis give us a prospective direction for further study. Future work could be done on the explanation of other image classification tasks in medical services, or in other fields by using this LRP algorithm. Likely, other explanation methods can be explored on this CheXpert dataset.

References

[1] N. Xie, G. Ras, M. van Gerven, and D. Doran, "Explainable deep learning: A field guide for the uninitiated," *arXiv preprint arXiv:2004.14545*, 2020.

[2] V. Buhrmester, D. Münch, and M. Arens, "Analysis of explainers of black box deep neural networks for computer vision: A survey," *arXiv preprint arXiv:1911.12116*, 2019.

[3] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PloS one*, vol. 10, no. 7, p. e0130140, 2015.

[4] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, *et al.*, "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 590–597, 2019.

[5] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

[6] P. Rajpurkar, J. Irvin, R. L. Ball, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. P. Langlotz, *et al.*, "Deep learning for chest radiograph diagnosis: A retrospective comparison of the chexnext algorithm to practicing radiologists," *PLoS medicine*, vol. 15, no. 11, p. e1002686, 2018.

[7] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2921–2929, 2016.

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

[9] W. Hsu, F. Tsai, G. Zhang, C. Chang, P. Hsieh, S. Yang, S. Sun, K. Liao, and E. T. Huang, "Development of a deep learning model for chest x-ray screening," *MEDICAL PHYSICS INTERNATIONAL*, vol. 7, no. 3, p. 314, 2019.

[10] F. Pasa, V. Golkov, F. Pfeiffer, D. Cremers, and D. Pfeiffer, "Efficient deep network architectures for fast chest x-ray tuberculosis screening and visualization," *Scientific reports*, vol. 9, no. 1, pp. 1–9, 2019.

[11] M. Böhle, F. Eitel, M. Weygandt, and K. Ritter, "Layer-wise relevance propagation for explaining deep neural network decisions in mri-based alzheimer's disease classification," *Frontiers in aging neuroscience*, vol. 11, p. 194, 2019.

[12] G. Chlebus, N. Abolmaali, A. Schenk, and H. Meine, "Relevance analysis of mri sequences for automatic liver tumor segmentation," *arXiv preprint arXiv:1907.11773*, 2019.

[13] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K.-R. Müller, "Explaining nonlinear classification decisions with deep taylor decomposition," *Pattern Recognition*, vol. 65, pp. 211–222, 2017.

[14] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[15] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105, 2012.

Appendix

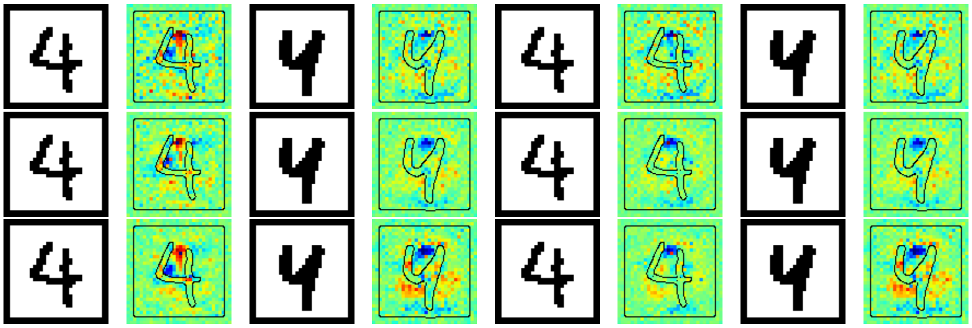


Figure 9: Heatmaps for three different activations, tanh (top row), ReLU (middle row), and ReLU* (bottom row). Every odd positioned image is the original input image of digit four. Heatmaps in position two, and four represent the prediction against class four. Heatmaps in position six, and eight represent the prediction against class nine.

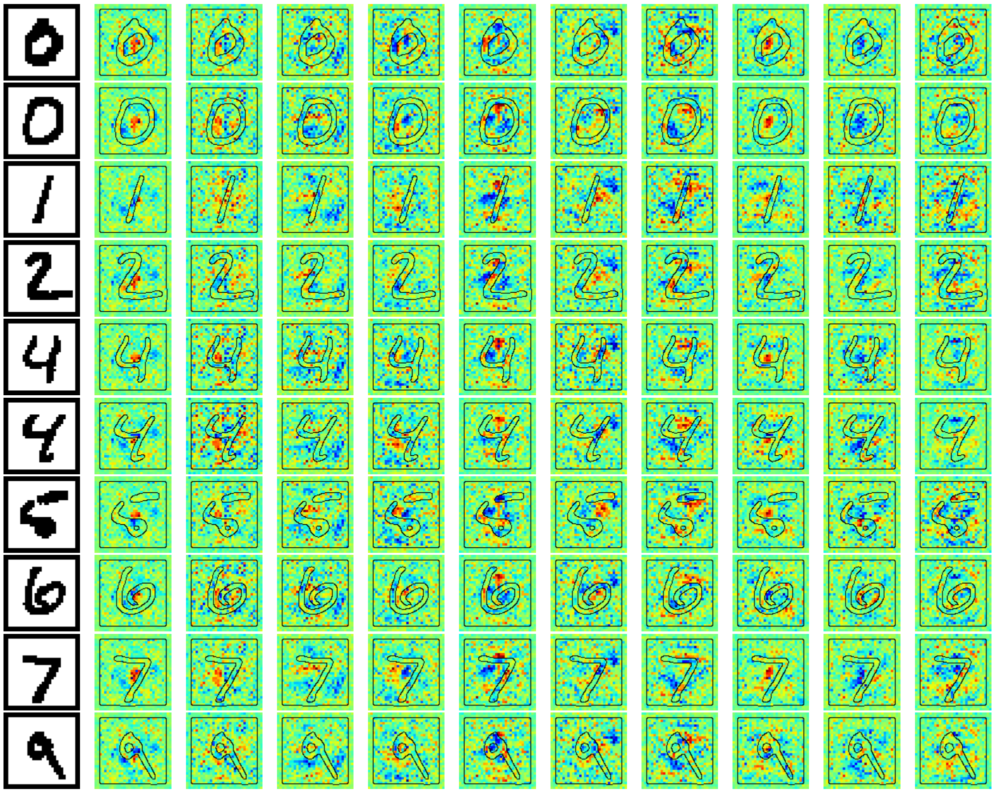


Figure 10: Heatmaps for all classes for 10 randomly drawn digits from the MNIST test set

Table 1: The CheXpert dataset consists of 14 labeled observations.

Pathology	Positive(%)	Uncertain(%)	Negative(%)
No Finding	16627 (8.86)	0 (0.0)	171014 (91.14)
Enlarged Cardiomediastinum	9020 (4.81)	10148 (5.41)	168473 (89.78)
Cardiomegaly	23002 (12.26)	6597 (3.52)	158042 (84.23)
Lung Lesion	6856 (3.65)	1071 (0.57)	179714 (95.78)
Lung Opacity	92669 (49.39)	4341 (2.31)	90631 (48.3)
Edema	48905 (26.06)	11571 (6.17)	127165 (67.77)
Consolidation	12730 (6.78)	23976 (12.78)	150935 (80.44)
Pneumonia	4576 (2.44)	15658 (8.34)	167407 (89.22)
Atelectasis	29333 (15.63)	29377 (15.66)	128931 (68.71)
Pneumothorax	17313 (9.23)	2663 (1.42)	167665 (89.35)
Pleural Effusion	75696 (40.34)	9419 (5.02)	102526 (54.64)
Pleural Other	2441 (1.3)	1771 (0.94)	183429 (97.76)
Fracture	7270 (3.87)	484 (0.26)	179887 (95.87)
Support Devices	105831 (56.4)	898 (0.48)	80912 (43.12)

Table 2: Chexpert data classification results

Class	**AUROC**	**Class**	**AUROC**
No Finding	0.79	Enlarged Cardiomediastinum	0.60
Cardiomegaly	0.80	Lung Opacity	0.87
Lung Lesion	0.53	Edema	0.87
Consolidation	0.51	Pneumonia	0.50
Atelectasis	0.53	Pneumothorax	0.68
Pleural Effusion	0.91	Pleural Other	1.00
Fracture	0.00	Support Devices	0.89
		Mean AUROC	0.73