Multi-label classification of fourteen thorax diseases on chest X-ray images

Saul Gómez
Universidad de Los Andes
sc.gomez11@uniandes.edu.co

Mateo Rueda
Universidad de Los Andes
ms.rueda10@uniandes.edu.co

Diego Valderrama
Universidad de Los Andes
df.valderrama@uniandes.edu.co

Abstract

The field of medical diagnosis contains a wealth of challenges which closely resemble classical machine learning problems. For instance, many tasks in the field of radiology are based on problems of multi-label classification wherein medical images are interpreted to indicate multiple present or suspected pathologies. Development of computer-aided techniques may lead to more accurate and more accessible diagnosis of radiology images such as thorax diseases on chest radiography. Despite the success of deep learning-based solutions, this task remains a major challenge in smart healthcare, since it is intrinsically a weakly supervised learning problem [17]. In this paper, we develop an algorithm based on ResNet-50 architecture trained end-to-end on ChestX-ray14 database, currently the largest publicly available chest Xray dataset, containing over 100,000 frontalview X-ray images with 14 diseases. Additionally, we evaluated our model against three state-of-the-art deep learning models on the Chest X-ray 14 dataset using AUROC metric.

1. Introduction

As Daffner mentioned some years ago, chest is "the mirror of health and disease" [3]. There is an enormous amount of information about the condition of the patient that can be extracted from a chest film, which provides an insight of the importance of thorax diseases diagnosis by analyzing chest images i.e. x-rays, computerized tomography images [16]. Chest X-ray exam is one of the most frequent and cost-effective medical imaging examination [18]. They are essential for the management of various diseases associated with high mortality and morbidity, such as lung diseases, and display a wide range of findings, many of them subtle [15]. The traditional chest radiograph is still ubiquitous in clinical practice, and will likely remain so for quite some time [16]. However, clinical diagnosis of chest X-ray can be challenging, and sometimes believed to be harder than diagnosis via chest CT imaging because its interpretation, frequently, is notoriously more difficult [18], [16].

The Chest X-ray (CXR) is the most frequently requested radiologic examination. due to its cheap, fast, reliable and can be used to diagnose many lungs and heart failure. However, reporting thorax diseases using chest X-rays is often an entry-level task for radiologist trainees [19]. Even experienced radiologists have trouble distinguishing infiltrates from the normal pattern of branching blood vessels in the lung fields, or detecting subtle nodules that indicate lung cancer [16]. Likewise, superimposed anatomical structures in the images joint with the fact that radiologists routinely must to observe multiple fine structures and subtle findings when they read medical images and compare it with radiological reports, demonstrate how complicated the task of making a diagnosis from X-ray images could be, in addition to exposing a field characterized by poorly accurate or, failing that, inefficient diagnostics [16], [19]. This explains the continued interest in computer-aided diagnosis for chest radiography and why reading a chest X-ray image remains a challenging job for learning-oriented machine intelligence, mainly due to shortage of large-scale machine-learnable medical image datasets, and lack of techniques that can mimic the high-level reasoning of human radiologists that requires years of knowledge accumulation and professional training [16], [19]. All these factors combined with the complicated nature of chest X-ray images, mentioned above, and its clinical importance, justifies the interest to develop computer algorithms to assist radiologists in reading chest images.

Given that, automatic processing and recognition of biomedical images has become one of the important branches of images processing, and much research has been reported in this field, including analysis of cell images, chromosome characterization, analysis of scinti-

grams, chest x-ray, CT images, etc [14]. In the case of chest radiographs, it is necessary to take into account that inherently display a wide dynamic range of X-ray intensities. Thus, it is often hard to "see through" the mediastinum and contrast in the lung fields is limited. Due to this, a classical solution in image processing is the use of local histogram equalization techniques in order to get an enhancement of high-frequency details [16]. On the other hand, it should be noted that instead of developing an overall computer system that can name and find any abnormality on a chest radiograph, many attempts have been made to automate just one of the many aspects involved in the evaluation of chest X-rays [15], which indicates the need to develop an algorithm that allows for remarkable acceleration of the diagnosis of Thorax's diseases, so that disease identification is more general and covers a wider range of diseases and reduces the error associated with manual diagnoses.

Furthermore, most clinical settings will drive a need for models which can accurately predict a large number of diagnostic features. This essentially turns medical problems into multi-label classification tasks with a huge number of targets, many of which may be poorly defined or are likely to be inconsistently labeled. Additionally, addressing this context involves considering predicting the absence of each label is significantly important as much as predicting its presence [11]. For this reason, deep learning techniques have achieved profound breakthroughs in many computer vision applications, including the classification of natural and medical images successfully in the last years, which has lead many investigators to adopt deep convolutional neural networks for automated diagnosis of thorax diseases on chest radiography [11].

2. Related work

In 2017, researchers Wang *et al*., presented a new ChestX-ray8 database which compresses 108,948 frontal view x-ray images and 32,717 unique patients examined with 8 text mined labels (where each image can have multi-labels), from the radiological reports using natural language processing (NLP). After the release of the ChestX-ray8 dataset, same researchers Wang *et al*. expanded the disease categories in this dataset including 6 more common thorax diseases (i.e. Consolidation, Edema, Emphysema, Fibrosis, Pleural Thickening and Hernia) and updated the NLP mined labels. This new database is called ChestX-ray14 and has a total number of 112,120 frontal-view X-ray images which surpasses any chest x-ray dataset in number of images and diseases labeled [17]. The release of these database increased the interest of researchers in the development of methods focused on classification and detection for this dataset. Some algorithms proposed were more general for the 14 types of diseases and others had promising results in a specific disease.

The first algorithm developed was created by the same database developers. The method roughly consisted of weakly-supervised multi-label image classification and pathology localization framework, which can detect the presence of multiple pathologies and subsequently generate bounding boxes around the corresponding pathologies. They tackle the problem with a training multi-label DCNN (Deep Convolutional Network) classification model. The model was pre-trained using ImageNet with different architectures such as AlexNet, GoogLeNet and ResNet50 leaving the fully connected and final classification layers. Due to the large variety of pre-trained DCNN architectures researchers insert a transition later to transform the activations from previous layers in a uniform dimension. Wang *et al*. also modify the basic DCNN architecture by adding a global pooling layer and a prediction layer not only designed for the multi-label classification but also to generate a heatmap of pathologies correspond to the presence of disease pattern with high probability. Different loss functions were experimented instead of using softmax loss such as Hinge Loss (HL), Euclidean Loss (EL) and Cross Entropy Loss (CEL). Finally, the evaluation proposed by these authors was the ROC curves for each disease and the performance was measured using the AUROC (area under the ROC curve) [17].

Later in 2017, Yiao *et al*., highlited a problem of the algorithm presented above which is the assumption that the labels are independent. For medical diagnostic application, there are significant dependencies between labels that must be modeled appropriately in order to maximize the performance of the classifier. In order to ovecome this problem, Yiao *et al*. used a model based on recurrent networks (RNN) adopting Long-short Term Memory Networks (LSTM) presented by Hochreiter & Schmidhuber in 1997 and treated the multi-label classification as a sequence prediction of fixed length. The formulation of our LSTM is particularly similar to those used in image and video captioning (Xu et al., 2015), but without the use of an attention mechanism and without the need of learning when to stop allowing to preserve the error that can be backpropagated through time and layers. By maintaining a more constant error, they allow LSTM nets to continue to learn over many time steps (over 1000). Additionally, the models were trained with MLE weighted cross-entropy loss to overcome the disbalance between classes and also were trained end-to-end from scratch without any pre-training on ImageNet data [20]. This method had promising results with an average AUROC of 0.798.

Subsequently, methods for solving this problem arose through conventional networks such as Resnet [2], Densenet [12] [8] and combinations between them [4] varying different parameters such as the loss function, the activation function and some final layers. Among these ap-

proaches, the most outstanding and with the best results corresponds to the method developed by Rajpurkar *et al*. called CheXNet. CheXNet is a 121-layer Dense Convolutional Network (DenseNet) trained on the ChestX-ray 14 dataset. DenseNets improve flow of information and gradients through the network, making the optimization of very deep networks tractable. To address the multi-label classification for multiple thoracic pathologies they made three important changes in the DenseNet architecture. First, instead of outputting one binary label, ChexNet outputs a vector t of binary labels indicating the absence or presence of each of the following 14 pathology classes: Atelectasis, Cardiomegaly, Consolidation, Edema, Effusion,Emphysema, Fibrosis, Hernia, Infiltration, Mass, Nodule, Pleural Thickening, Pneumonia, and Pneumothorax. Second, they replaced the final fully connected layer in CheXNet with a fully connected layer producing a 14-dimensional output, after which they applied an elementwise sigmoid nonlinearity. The final output is the predicted probability of the presence of each pathology class. Third, the authors modified the loss function to optimize the sum of unweighted binary cross entropy losses to take into account the imbalance between classes (eq. 1).

$$L(X, y) = \sum_{c=1}^{14} [-y_c \log p(Y_c = 1 \mid X) - (1 - y_c) \log p(Y_c = 0 \mid X)] \quad (1)$$

where $p(Y_c = 1|X)$ is the predicted probability that the image contains the pathology c and $p(Y_c = 0|X)$ is the predicted probability that the image does not contain the pathology c. The authors compared the performance of the algorithm on a test set which was annotated by four practicing academic radiologist and they also noted that this algorithm achieved state of the art results on all 14 pathology classes in Chest-Xray14 dataset with 0.84 of AUROC [12].

3. Approach

3.1. Baseline

The first approach to address this problem was a modification of the Pyramid of Histograms of Visual Words (PHOW) method proposed by [10]. The first modification. To train this method 10000 random images from the train set was selected. Subsequently, a vocabulary of 600 words was obtained and the histograms of each image were computed. Then, due to our problem consist in classify 14 diseases as a multitarget problem, we trained 14 SVM one vs all. Finally, the performance of our implementation was evaluated using the AUROC score for each class.

3.2. Proposed Method

Our method is a 50-layer Residual Network (ResNet50) trained on ChestX-ray 14 dataset. The network was initialized with the weights obtained from the model pretrained on ImageNet. Due to ResNet50 model has the final fully connected layer with 1000 outputs, we replace this layer with one that has 14 outputs. The network was trained using Adam optimizer with default parameters (learning rate = 0.001, β_1 = 0.9, β_1 = 0.99 and weight decay = 1e-8). We also decayed the learning rate in a factor of 10 each time the validation loss plateaus after 3 epochs and pick the model with the best average of auroc score on validation set.

The input images of our method are a downscale images of 256x256 with a normalization based on the mean and standard deviation of images in the ImageNet training set. With train set we also made a random crop of 224x224 and horizontal flipping in order to augment the data. Finally, we used a sigmoid activation function followed by a Binary Cross Entropy Logistic loss function because in literature we find that these functions improve performance in multi-label classification problems [9].

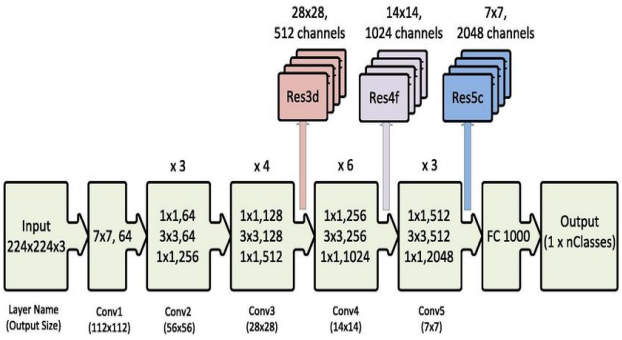


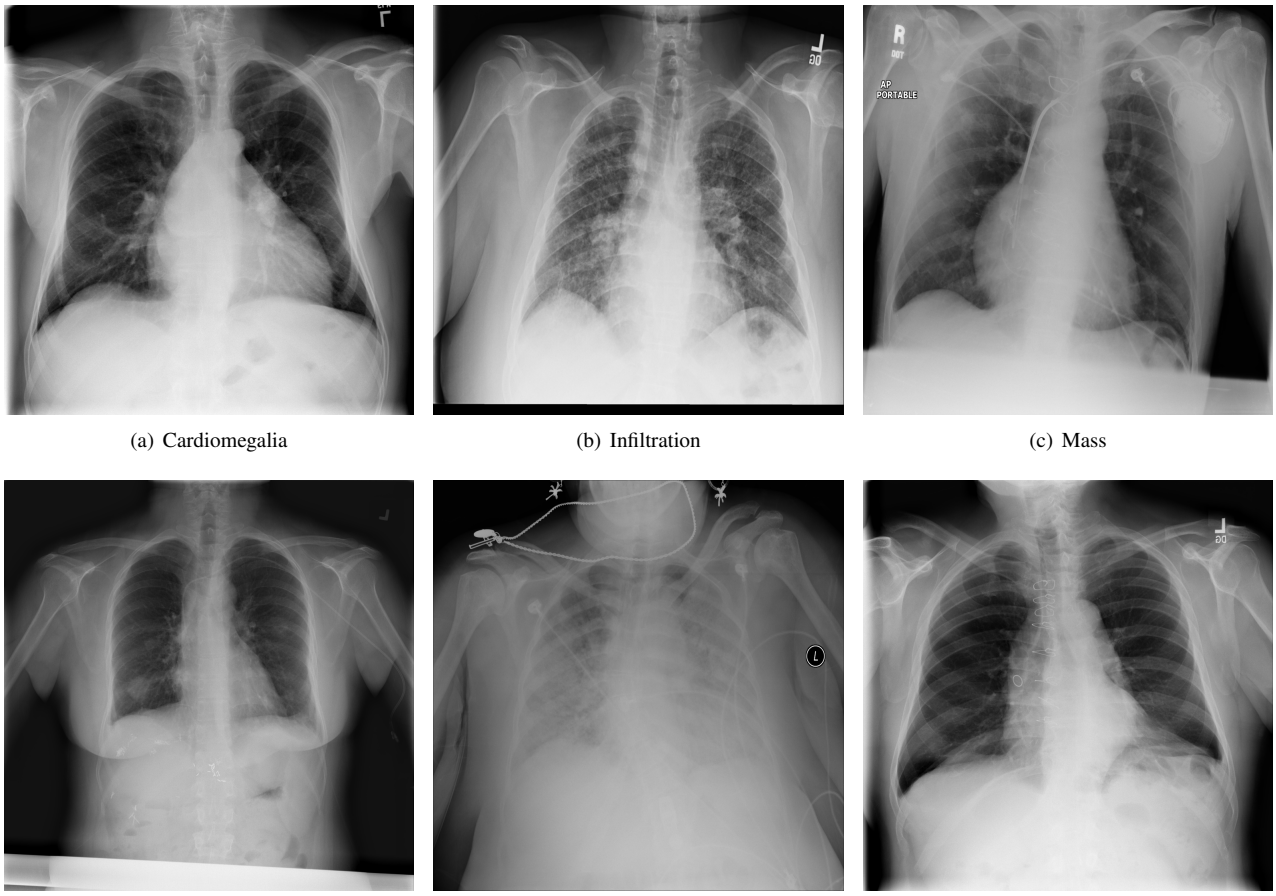
Figure 1. ResNet50 architecture used in our proposed method. Image obtained from [13]

4. Experiments

4.1. Database

4.1.1 Data

The database (Chest X-ray 14) was presented in 2017 by [18]. This dataset consist in 112119 chest X-ray images of 1024x1024 pixels. The images were obtained from 32717 unique patients with fourteen common thorax diseases, i.e, Atelectasis, Cardiomegaly, Effusion, Infiltration, Mass, Nodule, Pneumonia, Pneumothorax, Consolidation, Edema, Emphysema, Fibrosis, Pleural Thickening and Hernia (Fig 2 shows some chest X-ray image examples). The dataset is divided in 86523 images for training and validation, and 25596 images for the algorithm test [18]. It is important to clarify that the original figures for train and validated our algorithm were not classified in different stages.



(a) Cardiomegalia (b) Infiltration (c) Mass

(d) Nodule (e) Edema (f) Emphysema

Figure 2. Six thoracic diseases observed in chest X-ray dataset.

Therefore, it was a decision of the group to use 78484 and 8039 images in training and validation process, respectively.

The labels of this database present information about the gender, age, view position and thorax disease for all chest X-ray images in which some images have 2 or more diseases (multi-labels), i.e Cardiomegaly—Edema [18]. Likewise, the labels provide an approximation of the location for eight thorax diseases using a square to mark off the region in which the patterns of the diseases were found [18]. In this paper we only used the diseases of the images due to this represent the major information for our method.

4.1.2 Performance evaluation

To evaluated the performance of our algorithm, we select the best model obtained after training and validating our algorithm. Then, each image of the test set was evaluated and finally the AUROC score was obtained for all categories.

4.2. Validation Experiments

First we evaluated the proposed method as baseline obtaining an average of 31.37% in the AUROC score. The best performance reached with this methos was 39.37% for Edema classification. Once we have the performance using PHOW method, we change the representation space and the classifier through a neural network. We implemented our own architecture that consists of six convolutional layers and three fully connected layers. In all layers rectified linear unit (ReLu) was used as function of activation. These layers also have a dropout of 0.25 to reduce the over-fitting. A batch normalization with max pooling was also applied to the first three convolutional layers. However, this architecture was discarded because the results obtained were worse than those obtained with the PHOW method. Then, we used the architecture of VGG-11 with batch normalization (VGG-11 bn) model in order to improve the performance of our method. Although, the average of AUROC score increased to 35 % using this architecture, the performance was similar to that obtained with PHOW method.

To improve these results we used the pre-trained models on ImageNet of VGG 11 with batch normalization, ResNet 50 and ResNet 101. We also made a shuffle between train and val set images and changed the hyperparameters of the optimizer and loss function. we found that the performance obtain using ResNet50 and ResNet101 are similar because the performance differ in 1% being higher with ResNet101. Nevertheless, the train time is reduced in 20% with ResNet50 so we are implementing this architecture in our method. Table 1 shows the AUROC score for all classes using some validation experiments.

Table 1. Auroc Score for using our implementation of PHOW method and some mehotds pretrained on ImageNet

Class	PHOW	VGG 11 bn	ResNet101
Atelectasis	30.37 %	47.9%	51.3%
Cardiomegaly	24.40%	50.5%	57.4%
Effusion	31.87%	47.8%	50.3 %
Infiltration	32.69%	45.7%	49.9%
Mass	36.88%	45.6%	50%
Nodule	26.05%	50%	46.7%
Pneumonia	26.07%	44%	48.4%
Pneumothorax	24.62%	53.8%	45.7%
Consolidation	36.14%	44.7%	49.6%
Edema	39.67%	39.8%	57.6%
Emphysema	37.72%	50.5%	53.9%
Fibrosis	24.70%	50.8%	44.6%
Pleural thickening	33.48%	52.1%	49.1%
Hernia	37.57%	55.3%	55.7%
Average	31.37%	48.48%	50.79%

4.3. Evaluation Experiments

The AUROC measurement of our best model based on the ResNet50 network is compared with the current state of the art results and the most important approaches presented in the state of the art for the Chest-X-ray 14 database (Table 2). Additionally, the ROC curve is shown to observe the accuracy and recall of our algorithm for the 14 diseases classification (Figure 3).

5. Discussion

According to the results obtained presented in the previous section, it was achieved an acceptable performance and it is comparable with the first methods of the state of art. The efficiency of our method is roughly 50% for each of the diseases assessed, which is considerably high considering that the methods of the actual state of the art had a training close to 100 epochs, while our algorithm was only trained for 10 epochs. However, it is clear that the performance of our method is well below the current methods, indicating that it is necessary to modify the network architecture used as backbone, with the aim of improving the results obtained.

It is possible that the architecture of a dense neural network (DCNN approach) may have a better performance in this database as in those proposed for classification problems (CIFAR-10, ImageNet) since DenseNet connect all layers (with matching feature-map sizes) directly with each other. Also, to preserve the feed-forward nature, each layer of DenseNet obtains additional inputs from all preceding layers and passes on its own feature-maps to all subsequent layer. Our approach based on ResNet networks just connect the output of the lth layer as input to the $(l+1)th$ layer missing calculation of weights for different layer interactions [7]. Likewise, researchers like Rajpurkar *et al*. developed their algorithm with a large number of layers (121) by dramatically increasing the parameters with respect to our 50-layer recurrent network. For future experimentation should be taken into account the large amount of memory consumed by this type of DCNN approach. Furthermore, the problem can also be proposed as Gündel *et al*., who first carries out detections of the areas of interest for the classification of diseases such as the lungs and subsequently these

Table 2. AUROC score comparing our developed method (ResNet50) compared to the most influential methods for the Chest-xray14 database

Class	Wang *et al*.	Yao *et al*.	Rajpurkar *et al*.	ResNet50 (ours)
Atelectasis	71.6%	77.2%	80.9%	49.0%
Cardiomegaly	80.7%	90.4%	92.5%	54.1%
Effusion	78.4%	85.9%	86.4 %	55.1%
Infiltration	60.9%	69.5%	73.4%	50.0%
Mass	70.6%	79.2%	86.8%	52.8%
Nodule	67.1%	71.7%	78.0%	51.0%
Pneumonia	63.3%	71.3%	76.8%	50.7%
Pneumothorax	80.6%	84.1%	88.9	49.0%
Consolidation	70.8%	78.8%	79.0	50.0%
Edema	83.5%	88.2%	88.8%	47.6%
Emphysema	81.5%	82.9%	93.7%	54.6%
Fibrosis	76.9%	76.7%	80.5%	39.0%
Pleural thickening	70.8%	76.5%	80.6%	57.9%
Hernia	76.7%	91.4%	91.6%	48.8%
Average	73.81%	80.27%	86.57%	50.68%

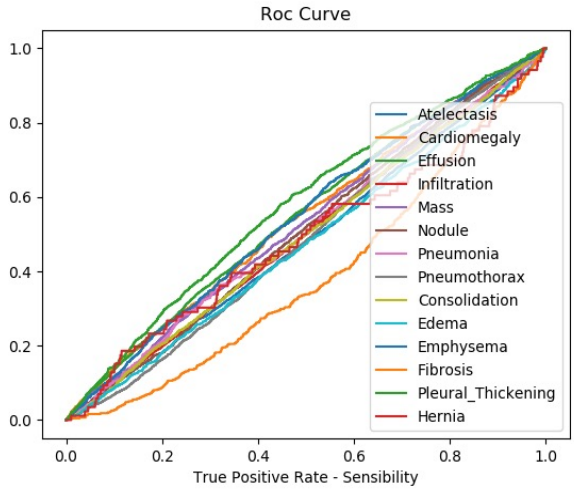
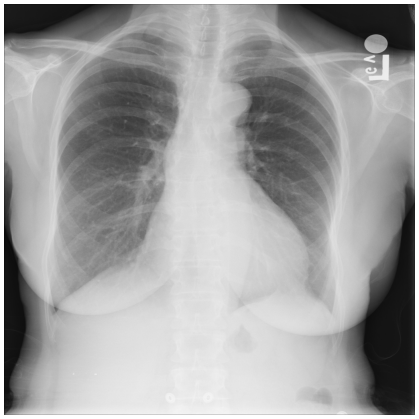


Figure 3. ROC curve of our method for each of the 14 diseases

regions work as input for the network in the classification problem for the 14 diseases [6].

In addition, by further analyzing the results presented in table 2, it is possible to identify that the class that makes it more difficult to learn for the model corresponds to the fibrosis disease. People with cystic fibrosis experience a build-up of thick sticky mucus in the lungs, which can be evidenced in figure 4(a) as small white structures. Given its small size compared to the dimensions of the image and due to the fact that you are not presenting a characteristic feature or pattern, it can hinder the learning of the implemented neural network, reflected in 39% AUROC performance. On the other hand, the results presented in table two indicate that the best performing class corresponds to pleural thickening. This may be because this disease can be seen in a change in the size of the space in the lungs due to the increased thickness of the lining of the lungs resulting from extensive scarring.



(a) Fibrosis

(b) Pleural Thickening

Figure 4. Examples of hard and easy classes. Fibrosis is the hard class with a 39% performance and Pleural Thickening is the easy with presenting a performance equals to 57.9%

Now, it is necessary to modify the architecture of our method in order to take into account the aspects identified in the most difficult classes, In other words, focus the model's attention on the areas where the disease is found, preventing it from having to process a lot of noise. To achieve this, it is possible to implement an architecture like the one presented in [5], where you have two branches, one global and one local, which are classification networks that predict whether the pathologies are in the images or not. For this reason, the global branch is first fine-tuned from a classification CNN using the global image and then they crop an attended region from the global image and train it for classification on the local branch. Finally, the last pooling layers of both the global and local branches are concatenated for fine-tuning the fusion branch, allowing the problem to be addressed in a similar way to how an expert radiologist would, analyzing those pathologies such as fibrosis in a local and detailed way, and diseases with more eye-catching features such as pleural thickening in a global way. Moreover, since attention-based models have demonstrated their proven ability to learn the localization and recognition of multiple objects despite being given only class labels during training, it is possible to implement a segmentation of the lungs with the aim of eliminating possible sources of distraction and forcing the model to learn only the characteristics of each disease, increasing the performance of the proposed method [1].

References

[1] J. Ba, V. Mnih, and K. Kavukcuoglu. Multiple object recognition with visual attention. *CoRR*, abs/1412.7755, 2015.

[2] I. M. Baltruschat, H. Nickisch, M. Grass, T. Knopp, and A. Saalbach. Comparison of deep learning approaches for multi-label chest x-ray classification. *CoRR*, abs/1803.02315, 2018.

[3] R. Daffner. Clinical radiology: The essentials, 3rd ed. *Radiology*, 250(3):630–630, 2009.

[4] Z. Ge, D. Mahapatra, S. Sedai, R. Garnavi, and R. Chakravorty. Chest x-rays classification: A multi-label and fine-grained problem. *CoRR*, abs/1807.07247, 2018.

[5] Q. Guan, Y. Huang, Z. Zhong, Z. Zheng, L. Zheng, and Y. Yang. Diagnose like a radiologist: Attention guided convolutional neural network for thorax disease classification. *CoRR*, abs/1801.09927, 2018.

[6] S. Gündel, S. Grbic, B. Georgescu, S. K. Zhou, L. Ritschl, A. Meier, and D. Comaniciu. Learning to recognize abnormalities in chest x-rays with location-aware dense networks. *CoRR*, abs/1803.04565, 2018.

[7] G. Huang, Z. Liu, and K. Q. Weinberger. Densely connected convolutional networks. *CoRR*, abs/1608.06993, 2016.

[8] P. Kumar, M. Grewal, and M. M. Srivastava. Boosted cascaded convnets for multilabel classification of thoracic diseases in chest radiographs. *CoRR*, abs/1711.08760, 2017.

[9] Y. Li, Y. Song, and J. Luo. Improving pairwise ranking for multi-label image classification. *CoRR*, abs/1704.03135, 2017.

[10] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[11] D. Quang Tran. Radiologist-level computer-aided detection to detect 14 common thoracic diseases. Master's thesis, Vietnam National University, 2019.

[12] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, R. Ball, C. Langlotz, K. Shpanskaya, M. Lungren, and Y. Ng. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv:1711.05225v3*, 2017.

[13] A. Singh and D. R. Kisku. Detection of rare genetic diseases using facial 2d images with transfer learning. In *2018 8th International Symposium on Embedded Computing and System Design (ISED)*, pages 26–30, Dec 2018.

[14] J.-I. Toriwaki, Y. Suenaga, T. Negoro, and T. Fukumura. Pattern recognition of chest x-ray images. *Computer Graphics and Image Processing*, 2(3):252 – 271, 1973.

[15] B. van Ginneken, L. Hogeweg, and M. Prokop. Computer-aided diagnosis in chest radiography: Beyond nodules. *European Journal of Radiology*, 72(2):226 – 230, 2009. Digital Radiography.

[16] B. Van Ginneken, B. M. Ter Haar Romeny, and M. A. Viergever. Computer-aided diagnosis in chest radiography: a survey. *IEEE Transactions on Medical Imaging*, 20(12):1228–1241, Dec 2001.

[17] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. Summers. 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)*, pages 3462–3471, 2017.

[18] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. *CoRR*, abs/1705.02315, 2017.

[19] X. Wang, Y. Peng, L. Lu, Z. Lu, and R. M. Summers. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. *CoRR*, abs/1801.04334, 2018.

[20] L. Yao, E. Poblenz, D. Dagunts, B. Covington, D. Devon Bernard, and K. Lyman. Learning to diagnose from scratch by exploit- ing dependencies among labels. *arXiv:1710.10501*, 2017.

Credits

All authors have read and approved the final manuscript. Baseline was developed in conjunction. Rueda made its contribution by implementing the VGG architecture. Gomez contributed to the development of the model based on the architecture of the ResNet network. Valderrama was responsible for the implementation of the evaluation methodology used to measure the performance of each of our models as well as the integration of data augmentation to improve the learning of the model. Finally, we certify that every member of this group had an equitable contribution in this work, resulting in 33.33% of the total for each one.