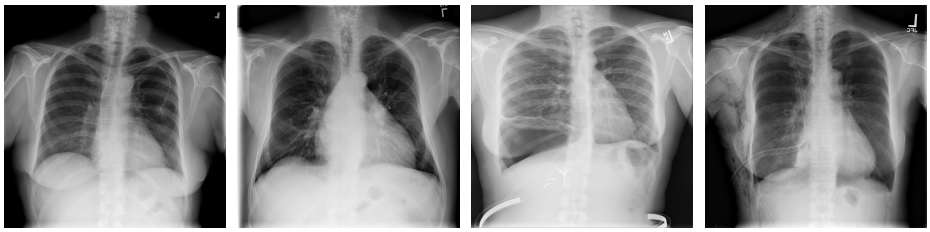
Comparison of Deep Learning Approaches for Multi-Label Chest X-Ray Classification

Ivo M. Baltruschat[1,2], Hannes Nickisch[3], Michael Grass[3], Tobias Knopp[1,2], Axel Saalbach[3]

[1] Section for Biomedical Imaging, University Medical Center Hamburg-Eppendorf
[2] Institute for Biomedical Imaging, Hamburg University of Technology
[3] Philips Research, Hamburg, Germany
i.baltruschat@uke.de, https://www.tuhh.de/ibi

Abstract. The increased availability of X-ray image archives (e.g. the ChestX-ray14 dataset from the NIH Clinical Center) has triggered a growing interest in deep learning techniques. To provide better insight into the different approaches, and their applications to chest X-ray classification, we investigate a powerful network architecture in detail: the ResNet-50. Building on prior work in this domain, we consider transfer learning with and without fine-tuning as well as the training of a dedicated X-ray network from scratch. To leverage the high spatial resolutions of X-ray data, we also include an extended ResNet-50 architecture, and a network integrating non-image data (patient age, gender and acquisition type) in the classification process.
In a systematic evaluation, using 5-fold re-sampling and a multi-label loss function, we evaluate the performance of the different approaches for pathology classification by ROC statistics and analyze differences between the classifiers using rank correlation. We observe a considerable spread in the achieved performance and conclude that the X-ray-specific ResNet-50, integrating non-image data yields the best overall results.

Keywords: Chest X-ray, Deep Learning, Convolutional Neural Networks, Transfer Learning, ChestX-ray14



(a) "No Finding" (b) "Cardiomegaly" (c) "Pneumothorax" (d) "Pneumothorax"

Fig. 1: Four examples of the ChestX-ray14 dataset which consists of 112,120 frontal chest X-rays from 30,805 patients. All images are labeled with up to 14 pathologies or "No Finding". The dataset does not only include acute findings, as the pneumothorax in figure (c), but also treated patients with a drain (d).

1 Introduction

In the United Kingdom, the care quality commission recently reported that – over the preceding 12 months – a total of 23,000 chest X-rays (CXRs) were not formally reviewed by a radiologist or clinician at Queen Alexandra Hospital alone. Furthermore, three patients with lung cancer suffered significant harm because their CXRs had not been properly assessed [1]. The Queen Alexandra Hospital is probably not the only hospital having problems with providing expert readings for every CXR. Increasing populations and life expectancies, is expected to drive an increase in demand for CXR readings.

In computer vision, deep learning has already shown its power for image classification with superhuman accuracy [8,18,16,3]. In addition, the medical image processing field is vividly exploring deep learning. However, one major problem in the medical domain is the availability of large datasets with reliable ground-truth annotation.

Two larger X-ray datasets have recently become available: the CXR dataset from Open-i [2] and ChestX-ray14 from the National Institutes of Health (NIH) Clinical Center [19]. Figure 1 illustrates four selected examples from ChestX-ray14. Due to its size, the ChestX-ray14, consisting of 112,120 frontal CXR images from 30,805 unique patients attracted considerable attention in the deep learning community. Triggered by the work of Wang et al. [19] using convolution neural networks (CNNs) from the computer vision domain, several research groups have begun to address the application of CNNs for CXR classification. In [20], Yao et al. presented a combination of a CNN and a recurrent neural network to exploit label dependencies. As a CNN backbone, they used a DenseNet [5] model which was adapted and trained entirely on X-ray data. Li et al. [9] presented a framework for pathology classification and localization using CNNs. More recently, Rajpurkar et al. [13] proposed transfer-learning with fine tuning, using a DenseNet-121 [5] and raised the AUC results on ChestX-ray14 for multi-label classification even higher. Unfortunately comparison of approaches remains difficult. Most reported results were obtained with differing experimental setups. This includes (among others) the employed network architecture, loss function and data augmentation. In addition, differing dataset splits were used and only [9] reported 5-fold cross-validated results. Contrary to these results, our experiments (Sec. 3) demonstrate that performance of a network depends significantly on the selected split.

Henceforth, to provide better insights into the effects of distinct design decisions for deep learning, we perform a systematic evaluation using a 5-fold re-sampling scheme[4]. We empirically analyze three major topics:

1. weight initialization, pre-training and transfer learning (Section 2.1)
2. network architectures such as ResNet-50 with large input size (Section 2.2)
3. non-image features such as age, gender, and view position (Section 2.3)

Prior work on ChestX-ray14 has been limited to the analysis of image data. In clinical practice however, radiologists employ a broad range of additional features

[4] Our training results will be made available upon acceptance. `https://CXR14Results`

during there diagnosis. To leverage the complete information of the dataset (i.e. age, gender, and view position), we propose in Section 2.3 a novel architecture integrating this information in addition to the learned image representation.

2 Methods

In the following, we consider pathology detection as a multi-label classification problem. All images $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}, \boldsymbol{x}_i \in \mathcal{X}$ are associated with a ground truth label $\boldsymbol{y}_i$, while we seek a classification function $\boldsymbol{f} : \mathcal{X} \to \mathcal{Y}$ that minimizes a specific loss function ℓ using N training sample-label pairs $(\boldsymbol{x}_i, \boldsymbol{y}_i), i = 1 \ldots N$. Here, we encode the label for each image as a binary vector $\boldsymbol{y} \in \{0, 1\}^M = \mathcal{Y}$ (with M labels). We encode "No Finding" as an explicit additional label and hence have $M = 15$ labels. After initial investigation of weighting loss functions such as positive/negative balancing [19] and class balancing, we noticed no significant differences and decided to employ the class-averaged binary cross entropy as a loss function:

$$\ell(\boldsymbol{y}, \boldsymbol{f}) = \frac{1}{M} \sum_{m=1}^{M} H[y_m, f_m], \text{ with } H[y, f] = -y \log f - (1 - y) \log(1 - f). \quad (1)$$

Prior work on the ChestX-ray14 dataset focused primarily on ResNet-50 and DenseNet-121 architectures. Due to its outstanding performance in the computer vision domain [6], we focus in our experiments on the ResNet-50 architecture [4]. To adapt the network to the new task, we replace the last dense layer of the original architecture with a new dense layer matching the number of labels and add a sigmoid activation function.

2.1 Weight Initialization and Transfer Learning

We investigate two distinct initialization strategies for the ResNet-50. First, we follow the scheme described in [3], where the network parameters are initialized with random values and thus the model is trained from scratch. Second, we initialize the network with pre-trained weights, where knowledge is gained in a different domain and task. Furthermore, we distinguish between *off-the-shelf* (OTS) and *fine-tuning* (FT) in the transfer-learning approach.

A major drawback in medical image processing with deep learning is the limited size of datasets compared to the computer vision domain. Hence, training a CNN from scratch is often not feasible. One solution is transfer-learning. Following the notation in [12], a source domain $\mathcal{D}_s = \{\mathcal{X}_s, P_s(X_s)\}$ with task $\mathcal{T}_s = \{\mathcal{Y}_s, f_s(\cdot)\}$ and a target domain $\mathcal{D}_t = \{\mathcal{X}_t, P_t(X_t)\}$ with task $\mathcal{T}_t = \{\mathcal{Y}_t, f_t(\cdot)\}$ are given with $\mathcal{D}_s \neq \mathcal{D}_t$ and/or $\mathcal{T}_s \neq \mathcal{T}_t$. In transfer-learning, the knowledge gained in $\mathcal{D}_s$ and $\mathcal{T}_s$ is used to help learning of the prediction function $f_t(\cdot)$ in $\mathcal{D}_t$.

Employing an off-the-shelf approach [21,14], the pre-trained network is used as a feature extractor, and only the weights of the last (classifier) layer are adapted. In fine-tuning, one chooses to re-train one or more layers with samples from the new domain. For both approaches, we use the weights of a ResNet-50 network trained on ImageNet as a starting point [15].

2.2 Architectures

In addition to the original ResNet-50 architecture, we employ two variants. First, we reduce the number of input channels to one (the ResNet-50 is designed for the processing of RGB images from the ImageNet dataset), which should facilitate the training of a X-ray specific CNN. Second, we increase the input size by a factor of two (i.e. 448×448). To keep the model architectures similar, we only add a new pooling layer after the first bottleneck block. In particular for the detection of small structures, which could be indicative of a pathology (e.g. masses and nodules), a higher effective resolution could be beneficial.

2.3 Non-Image Features

ChestX-ray14 contains information about the patient age, gender, and view position (i.e. if the X-ray image is acquired posterior-anterior (PA) or anterior-posterior (AP)). Radiologists use information beyond the image to conclude which pathologies are present or not. The view position changes the expected position of organs in the X-ray images (i.e. PA images are horizontally flipped compared to AP). In addition, organs (e.g. the heart) are magnified in an AP projection as the distance to the detector is increased.

We concatenate the image feature vector (i.e. output of the last pooling layer with dimension 2024×1) with the new non-image feature vector (with dimension 3×1). Therefore, view position and gender is encoded as $\{0, 1\}$ and the age is linearly scaled $[\min(X_{\text{pa}}), \max(X_{\text{pa}})] \mapsto [0, 1]$, in order to avoid a bias towards features with a large range of values.

3 Experiments and Results

To evaluate our approaches for multi-label pathology classification, the entire corpus of ChestX-ray14 (Fig. 1) is employed. The dataset does not include the original DICOM images but [19] performed a simple preprocessing where they rescaled the intensity range from a higher bit-depth down to 8-bit. In addition, they resized each image to 1024×1024 pixel without preserving the aspect ratio.

For an assessment of the generalization performance, we perform a 5 times re-sampling scheme [10]. Within each split, the data is divided into 70% training, 10% validation and 20 % testing. When working with deep learning, hyper-parameters and tuning without a validation set and/or cross-validation can easily result in over-fitting. Since individual patient have multiple follow-up acquisitions, all data from a patient is assigned to a subset only. This leads to a large patient number diversity (e.g. split two has 5,817 patients and 22,420 images whereas split 5 has 6,245 patients and the same number of images). We estimate the average validation loss over all re-samples to determine the best models. Finally, our results are calculated for each fold on the test set and averaged afterwards.

Implementation: In all experiments, we use a fixed setup. To extend ChestX-ray14, we use the same geometric data augmentation as in [18]. At training, we

sample various sized patches of the image with sizes between 70% and 100% of the image area. The aspect ratio is distributed evenly between $3:4$ and $4:3$. In addition, we employ random rotation between $\pm 7°$ and horizontal flipping. For validation and testing, we rescale images to 256×256 and 480×480 for small and large spatial size, respectively. Afterwards, we use the center crop as input image. As in [3], dropout is not employed [17]. As optimizer, we use ADAM[7] with default parameters for $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning rate lr is set to $lr = 0.001$ and $lr = 0.01$ for transfer-learning and from scratch, respectively. While training, we reduce the learning rate by a factor of 2 when the validation loss is not improved. Due to model architecture variations, we use batch sizes of 16 and 8 for transfer-learning and from scratch with a large input size, respectively. The models are implemented in CNTK and trained on GTX 1080 GPUs yielding a processing time of around 10ms per image.

Results: Table 2 summarizes the outcome of our evaluation and we show state-of-the-art reference results in Fig. 2. In total, we evaluate eight different experimental setups, with varying weight initialization schemes and network architectures as well as with and without non-image features. We compare the classifier scores by Spearman's pairwise rank correlation coefficient and we perform a ROC analysis using the area under the curve (AUC) for all pathologies.

In Tab. 1, the correlation matrix between our models of Tab. 2 shows a clustering of the models w.r.t. their pairwise rank correlation into three groups. First, we note that the "from scratch models" (i.e. "1channel" and "large) without non-image features have the highest correlation of 0.93 amongst each other, followed by the fine-tuned models with 0.81 and 0.80 for "1channel" and "large", respectively. Second, the OTS model surprisingly has higher correlation with the from scratch models than the fine-tuned model. Third, for models with non-image feature, no such correlation is observed and their value is between 0.32 to 0.47. The results indicate a high variability of the outcome with respect to the selected dataset split. Especially for "Hernia", which is the class with the smallest number of positive samples, we observe a standard deviation of up to 0.05. As a result, an assessment of existing approaches, and a comparison of their performance is difficult, as prior work focused mostly on a single (random) split.

With respect to the different initialization schemes we observe already reasonable results for OTS networks that are optimized on natural images. Using fine-tuning techniques, the results are improved considerably, from 0.730 to 0.819 AUC on average. A complete training of the ResNet-50 using CXRs results in a rather comparable performance. Only the high-resolution variant of the ResNet-50 outperforms the FT approach by 0.002 on average AUC. In particular, for smaller pathologies like masses and nodules an improvement is observed (i.e. 0.017 and 0.006 AUC increase, respectively), while for other pathologies a similar, or slightly lower performance is estimated.

Finally, all our experiments with non-image features slightly increase the AUC on average to its counterpart (i.e. without non-image feature). Our from scratch trained ResNet-50 with an enlarged input size and integrated non-image data yields the best overall performance with 0.822 average AUC.

Table 1: Spearman's rank correlation coefficient is calculated between all model pairs and is averaged over all 5 splits. Our experiments are grouped into three categories. First, "Without" and "With" non-image features. Second, transfer-learning with off-the-shelf (OTS) and fine-tuned (FT) models. Third, from scratch where "1channel" refers to same input size as in transfer-learning but changed number of channels. "large" means we changed the input dimensions to $448 \times 448 \times 1$. We identify three clusters: all models under "With", models trained from scratch and "Without", and the "OTS" model.

		Without				With			
		OTS	FT	1channel	large	OTS	FT	1channel	large
Without	OTS	0.00	0.65	0.74	0.73	0.46	0.38	0.40	0.59
	FT	0.65	0.00	0.81	0.80	0.38	0.42	0.43	0.64
	1channel	0.74	0.81	0.00	0.93	0.41	0.43	0.47	0.71
	large	0.73	0.80	0.93	0.00	0.40	0.43	0.47	0.71
With	OTS	0.46	0.38	0.41	0.40	0.00	0.32	0.33	0.39
	FT	0.38	0.42	0.43	0.43	0.32	0.00	0.35	0.42
	1channel	0.40	0.43	0.47	0.47	0.33	0.35	0.00	0.45
	large	0.59	0.64	0.71	0.71	0.39	0.42	0.45	0.00

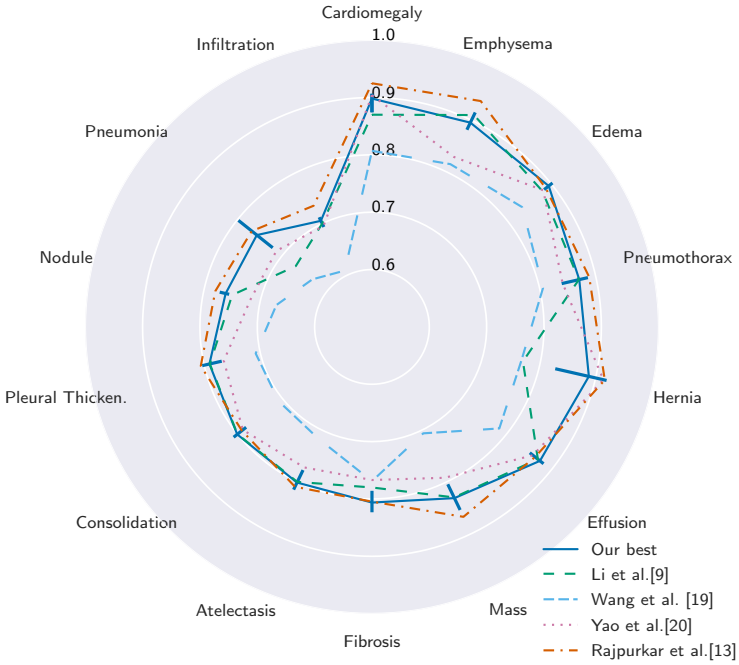


Fig. 2: Comparison of our best model to other groups. We sort the pathologies with increasing average AUC over all groups. For our model, we report the min and max over all folds as error bar to illustrate the effect of splitting.

Table 2: AUC result overview for all our experiments. In this Table, we present averaged results over all 5 splits and the calculated standard deviation (std) for each pathology. We divide our experiments into three categories. First, without and with non-image features. Second, transfer-learning with off-the-shelf (OTS) and fine-tuned (FT) models. Third, from scratch where "1channel" refers to same input size as in transfer-learning but changed number of channels. "large" means we changed the input dimensions to $448 \times 448 \times 1$. For better comparison, we present the average AUC and the standard deviation over all pathologies in the last row. Bold text emphasizes the overall highest AUC value. Values are scaled by 100 for convenience.

	Without non-image features				With non-image features			
Pathology	OTS	FT	1channel	large	OTS	FT	1channel	large
Cardiomegaly	72.7 ± 1.8	88.5 ± 0.7	88.9 ± 0.5	89.7 ± 0.3	75.9 ± 1.4	88.4 ± 0.8	$\mathbf{90.2 \pm 0.4}$	89.8 ± 0.8
Emphysema	77.8 ± 2.1	89.2 ± 1.0	87.0 ± 0.8	88.3 ± 1.3	79.8 ± 1.9	$\mathbf{89.4 \pm 1.2}$	87.4 ± 1.3	89.1 ± 1.2
Edema	84.4 ± 0.6	$\mathbf{89.1 \pm 0.4}$	$\mathbf{89.1 \pm 0.6}$	88.8 ± 0.5	85.7 ± 0.5	$\mathbf{89.1 \pm 0.7}$	89.0 ± 0.6	88.9 ± 0.3
Hernia	78.8 ± 1.4	85.5 ± 3.8	88.1 ± 4.2	87.5 ± 4.5	81.9 ± 2.5	88.2 ± 3.2	89.3 ± 4.4	$\mathbf{89.6 \pm 4.4}$
Pneumothorax	77.3 ± 1.3	$\mathbf{87.0 \pm 0.8}$	85.7 ± 0.9	85.9 ± 0.9	79.1 ± 1.2	86.5 ± 0.6	85.4 ± 0.7	85.9 ± 1.1
Effusion	79.4 ± 0.4	87.1 ± 0.2	$\mathbf{87.6 \pm 0.2}$	$\mathbf{87.6 \pm 0.2}$	80.6 ± 0.4	87.2 ± 0.3	$\mathbf{87.6 \pm 0.2}$	87.3 ± 0.3
Mass	66.8 ± 0.6	82.2 ± 1.0	83.3 ± 0.6	$\mathbf{83.9 \pm 0.9}$	68.6 ± 0.6	82.2 ± 1.0	83.3 ± 0.7	83.2 ± 0.3
Fibrosis	72.0 ± 0.9	$\mathbf{80.0 \pm 0.9}$	79.9 ± 0.8	79.2 ± 1.6	73.9 ± 0.8	$\mathbf{80.0 \pm 0.9}$	79.6 ± 0.5	78.9 ± 0.5
Atelectasis	71.8 ± 0.6	$\mathbf{80.3 \pm 0.7}$	79.9 ± 0.4	79.2 ± 0.7	73.2 ± 0.7	80.1 ± 0.6	79.3 ± 0.6	79.1 ± 0.4
Consolidation	74.3 ± 0.3	79.5 ± 0.5	$\mathbf{80.6 \pm 0.4}$	80.0 ± 0.3	75.3 ± 0.3	79.6 ± 0.5	80.4 ± 0.5	80.0 ± 0.7
Pleural Thicken.	68.8 ± 1.0	$\mathbf{79.0 \pm 0.7}$	78.4 ± 0.9	78.0 ± 1.1	70.8 ± 1.1	78.6 ± 1.1	78.2 ± 1.3	77.1 ± 1.3
Nodule	65.0 ± 0.8	72.6 ± 0.9	73.3 ± 0.8	75.1 ± 1.3	66.5 ± 0.7	74.7 ± 0.6	74.0 ± 0.7	$\mathbf{75.8 \pm 1.4}$
Pneumonia	66.4 ± 2.7	74.4 ± 1.6	74.3 ± 1.5	75.3 ± 2.2	68.3 ± 2.3	73.3 ± 1.3	74.8 ± 1.5	$\mathbf{76.7 \pm 1.5}$
Infiltration	65.9 ± 0.2	69.9 ± 0.6	$\mathbf{70.2 \pm 0.3}$	$\mathbf{70.2 \pm 0.5}$	67.0 ± 0.4	$\mathbf{70.2 \pm 0.2}$	70.1 ± 0.5	70.0 ± 0.7
Average	73.0 ± 1.1	81.7 ± 1.0	81.9 ± 0.9	82.1 ± 1.2	74.8 ± 1.1	82.0 ± 0.9	82.0 ± 1.0	$\mathbf{82.2 \pm 1.1}$
No Findings	71.6 ± 0.3	76.9 ± 0.5	$\mathbf{77.3 \pm 0.3}$	77.1 ± 0.4	72.5 ± 0.3	76.8 ± 0.4	77.1 ± 0.4	77.1 ± 0.3

4 Conclusion and Discussion

We present a systematic evaluation of different approaches for CNN-based X-ray classification on ChestX-ray14 and gain a better understanding of medical image processing with deep learning. While surprisingly satisfactory results are obtained with networks optimized on the ImageNet dataset, the best overall results can be reported for the model that is exclusively trained with CXRs and that incorporate non-image data (i.e. view position, patient age, and gender). However, our fine-tuned ResNet-50 model achieves state-of-the-art results in four out of fourteen classes compared to [13] (who had state-of-the-art results in all fourteen classes). At the same time, a substantial variability in the results can be observed when different splits are considered. While this suggests that the training of deep neural networks in the medical domain becomes a viable option as more and more public datasets are available, the practical use of deep learning in clinical practice is still an open issue. As discussed by [11] the quality of the (automatically generated) labels, and their precise medical interpretation may be a limiting factor addition to the presence of treated findings. Future work, will include investigation of other model architectures, new architectures for leveraging label dependencies and incorporating segmentation information.

References

1. Care Quality Commission: Queen Alexandra hospital quality report (2017)
2. Demner-Fushman, D., Kohli, M.D., Rosenman, M.B., Shooshan, S.E., Rodriguez, L., Antani, S.K., Thoma, G.R., McDonald, C.J.: Preparing a collection of radiology examinations for distribution and retrieval. In: JAMIA (2016)
3. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
4. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: ECCV (2016)
5. Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: CVPR (2017)
6. Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. pp. 2261–2269 (2017), https://doi.org/10.1109/CVPR.2017.243
7. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2015)
8. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS (2012)
9. Li, Z., Wang, C., Han, M., Xue, Y., Wei, W., Li, L., Li, F.: Thoracic disease identification and localization with limited supervision. In: CoRR (2017)
10. Molinaro, A.M., Simon, R., Pfeiffer, R.M.: Prediction error estimation: a comparison of resampling methods. In: Bioinformatics (2005)
11. Oakden-Rayner, L.: Exploring the ChestXray14 dataset: problems (2017), https://lukeoakdenrayner.wordpress.com/2017/12/18/
12. Pan, S.J., Yang, Q.: A survey on transfer learning. In: IEEE Trans. Knowl. Data Eng. (2010)

13. Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., Lungren, M.P., Ng, A.Y.: Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. In: CoRR (2017)
14. Razavian, A.S., Azizpour, H., Sullivan, J., Carlsson, S.: CNN features off-the-shelf: An astounding baseline for recognition. In: CVPR (2014)
15. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M.S., Berg, A.C., Li, F.: Imagenet large scale visual recognition challenge. In: International Journal of Computer Vision (2015)
16. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: CoRR (2014)
17. Srivastava, N., Hinton, G.E., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. In: Journal of Machine Learning Research (2014)
18. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S.E., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: CVPR (2015)
19. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: CVPR (2017)
20. Yao, L., Poblenz, E., Dagunts, D., Covington, B., Bernard, D., Lyman, K.: Learning to diagnose from scratch by exploiting dependencies among labels. In: CoRR (2017)
21. Yosinski, J., Clune, J., Bengio, Y., Lipson, H.: How transferable are features in deep neural networks? In: NIPS (2014)