CXNet-m2: A Deep Model with Visual and Clinical Contexts for Image-Based Detection of Multiple Lesions

Shuaijing Xu[1], Guangzhi Zhang[1], Rongfang Bie[1](✉), and Anton Kos[2]

[1] Beijing Normal University, Beijing 100875, China
rfbie@bnu.edu.cn
[2] University of Ljubljana, Ljubljana, Slovenia

Abstract. Diagnosing multiple lesions on images is facing with challenges of incomplete and incorrect disease detection. In this paper, we propose a deep model called CXNet-m2 for the detection of multiple lesions on chest X-ray images. In our model, there is a convolutional neural network (CNN) for encoding the images, a recurrent neural network (RNN) for generating the next word (the name of lesion) and an attention mechanism to align the visual contexts with the prediction of words. There are two main contributions of CXNet-m2 to improve the work efficiency and increase the diagnosis accuracy. (1) Inspired by image captioning, CXNet-m2 adapts the classification system to a language model, where Bi-LSTM is used to learn the clinical relationship between lesions. (2) Inspired by attention mechanism, the prediction of possible lesions is guided by visual contexts, where the visual contexts are selected by the previously generated words and chosen visual regions.

The experimental results on Chestx-ray14 show that CXNet-m2 achieves better AUC and the different versions of CXNet-m2 illustrate the importance of pre-training and clinical contexts.

Keywords: Chest X-Rays image · Multi-label classification · Neural network

1 Introduction

Advanced technologies and automated algorithms of image analysis is becoming increasingly important and urgent to support clinical diagnosis and treatment. It is usually formulated as a classification problem and then to identify the abnormalities of images using representation-learning methods such as support vector machine (SVM), random forests (RF) and deep neural networks [1,2]. The traditional shallow methodologies including SVM and RF should build upon hand-crafted image features such as local binary patterns (LBP) and histogram of oriented gradient (HOG), which makes the process complex and difficult [3]. Compared with them, deep neural networks learn more representative image

E. S. Biagioni et al. (Eds.): WASA 2019, LNCS 11604, pp. 407–418, 2019.
https://doi.org/10.1007/978-3-030-23597-0_33

features through automatic back propagation and yield better performances in many fields including medical image analysis [4].

In this paper, we focus on chest X-ray image analysis, one of the most commonly accessible examinations for screening and diagnosis of many lung diseases. Our goal is to detect the multiple lesions of chest X-ray images using the most promising deep neural network methods. In the past decades, the lack of large-scale dataset literally stalls the advancement of chest X-ray image analysis. We therefore choose Chestx-ray14 as our dataset, which was released by reference [5] in 2017. Chestx-ray14 is one of the largest publicly available chest x-ray data set, containing 14 diseases, more than 30,000 patients and 112,120 labeled chest x-ray images. As analysed in reference [4], there are 60361 normal images, 30963 single-lesion images and 20795 muiti-lesion images. Most papers concentrate on multi-class classification problem, where all the abnormal images are classified into 14 categories. Reference [5] fine-tuned four standard CNN architectures (AlexNet, VGGNet, GoogLeNet and ResNet) and ResNet achieved the best result [6–9]. Reference [10] utilized a 121-layer DenseNet architecture and made little modification on it [11]. Reference [12] presented a model that simultaneously performed disease classification and localization based on Resnet and a simple recognition network. All of them made use of fine-tuning pre-trained networks, a good choice when the number of training examples is limited. However, they mixed multi-lesion images and single-lesion images together when training and testing, and they did not focus on multi-lesion detection specially. Reference [13] is the only work to classify multi-label images on this data set, found on arXiv. they combined DenseNet and Longshort Term Memory Networks (LSTM), and illustrated the necessaries of exploiting the dependencies between abnormalities. Although their method is just a simple combination of CNN and RNN, they offer a thought of using the clinical contexts.

Therefore, we present CXNet-m2 taking advantage of fine-tuning existing deep convolutional neural networks and learning interdependencies between lesions, and made some improvements on it. (1) In the first improvement, we adapt the multi-label classification problem to an image-captioning problem, where the names of lesions are generated from a Bi-LSTM decoder. Compared with LSTM, Bi-LSTM makes the prediction of the lesion under the information of both previously and lately generated lesions. Bi-LSTM can make better use of clinical contexts because there are not exact causal relationships between lesions. For example, studies have suggested that a chest x-ray image containing cardiomegaly is more likely to contain pulmonary edema because of the left ventricular failure and chronic nasopharyngeal obstruction [14,15]. In this case, the sequential order of prediction is not fixed and the lately predicted word may give some information for the generation of the previous words. (2) In the second improvement, we add visual contexts for generating the possible lesions by attention mechanism, which means we can make better use of the encoded image features. Different from the common attention mechanism, the visual contexts are selected according to not only the weight function, but also the previously generated words and previously chosen visual regions.

2 Related Work

Convolutional Neural Networks. As a deep-learning method, deep convolutional neural networks (CNN) is wisely used to image analysis because of the local connectivity and shared weights. These two features not only keep the affine invariance of CNN, but also reduce the number of parameters, which ensures that CNN can be widely used in the learning and processing of complex data. The basic architecture of CNN contains convolutional layers, pooling layers and fully connected layers. Convolutional layers are stacked to detect local conjunctions of features from the previous layer, pooling layers behind are designed to reduce computational complexity, and fully connected layers in the end are used to output the classification result. Some researchers replace the fully connected layers with convolutional layers to input test images on different scales. Many robust CNN frameworks have been designed including VGGnet, Resnet and Densenet [7,9,11].

Recurrent Neural Networks. In the traditional neural network model, information is passed from the input layer to the hidden layer and then to the output layer. The layers are connected to each other, the nodes between each layer are not connected. This structure can not model data with timing sequence. For example, it is generally necessary to use the information of the previous and subsequent words to predict the next word of a sentence, because the words before and after in a sentence are semantically linked. Recurrent Neural Networks models sequence data, where the current output of a sequence is also related to the previous output [16]. All RNNs have a chain of a repetitive neural network module. In a standard RNN, this duplicate module has only a very simple structure, such as a tanh layer, while duplicate module of LSTM has a complex structure. The structure of LSTM solves the problem of long-term dependencies, which RNN loses the ability to learn [17]. GRU, a variant of LSTM, replaces the forgetting gate and inputting gate the updating gate, reducing the computational complexity of LSTM [18]. In some problems, the output of the current moment is not only related to the previous state, but also related to the state after it. At this time, a two-way RNN (Bi-RNN) is needed, where there is a combination of two unidirectional RNNs [19].

Attention Mechanism. Attention is a brain signal processing mechanism unique to human vision. Human vision captures the target area that needs to be focused on by quickly scanning the global image, suppressing other useless information [20]. In deep learning method, most of the current attention models are attached to the Encoder-Decoder framework. There are two common attention-based models, global attentional model and local attentional model [22]. The idea of a global attentional model is to consider all the hidden states of the encoder when deriving the context vector, where it has to attend to all words on the source side for each target word. A local attentional mechanism can choose to focus only on a small subset of the source positions per target word.

Image Captioning. The recent progress on image captioning has greatly proved that it is possible to describe the images with accurate and meaningful sentences or words. In most cases, there are a CNN and a RNN or other advanced versions of them to understand images. CNN is used to encode images and RNN are used to decode and output the words. Several methods used lexical representations instead of visual representations. Reference [23] first used multiple instance learning to train a word detector. Reference [24] minimized a joint objective learning from these diverse data sources and leverage distributional semantic embeddings, enabling the model to describe novel objects outside of training data sets. Other methods used visual representations instead of lexical representations.Reference [25] proposed a multimodal Recurrent Neural Network (m-RNN) architecture to fuse text information and visual features extracted on the whole image. Reference [26] formulate a discriminative bimodal neural network, which can be trained by a classifier with extensive use of negative samples.

3 CheXray-m2

3.1 Image-Caption Model

Our model for detection of multiple lesions on chest X-ray images is composed of the following components: a pre-trained CNN encoder that extracts visual feature representation and an LSTM-based neural network that models the attention dynamics of focusing on those regions and generating sequentially labels. We describe in detail each of the components.

Encoder: Image Representation. Our model takes a single raw image and generates a sequence of encoded words which denote different lesions. Convolutional neural networks (CNN) has gained popularity in recent years due to their ability to learn representative image features through automatic back propagation. Many robust convolutional neural network frameworks have been designed including VGGnet, Resnet, Inception-Resnet and Densenet [6–9,11]. They have been trained on ImageNet, containing 1.3 million natural images [6]. It is promising to fine tune these existing deep networks due to the limit of data size, labeling and computer hardware. However, it may lead to low transfer efficiency, over-fitting and other problems when medical image data set is totally different from the source dataset and the number of training examples are quite limited. We therefore propose an encoder based on the advantage of CNN while taking into account the peculiarity of the chest X-ray image at hand. There are three main improvements.

Firstly, the proposed encoder should be much thinner in network depth and smaller in parameter number. Models using hundreds of layers, such as DenseNet, typically require hundreds of thousands to millions of examples to train and they are more likely to over-fitting with one tenth the training data. Based on VGGNet and ResNet, we reduce some repetitive convolutional layers and

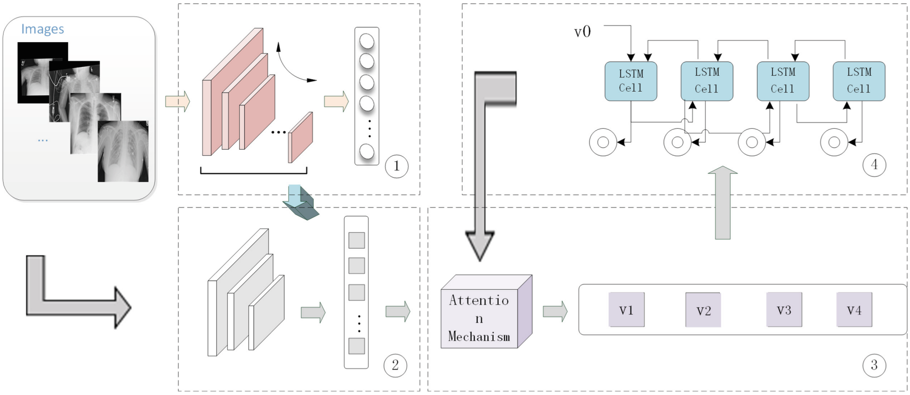


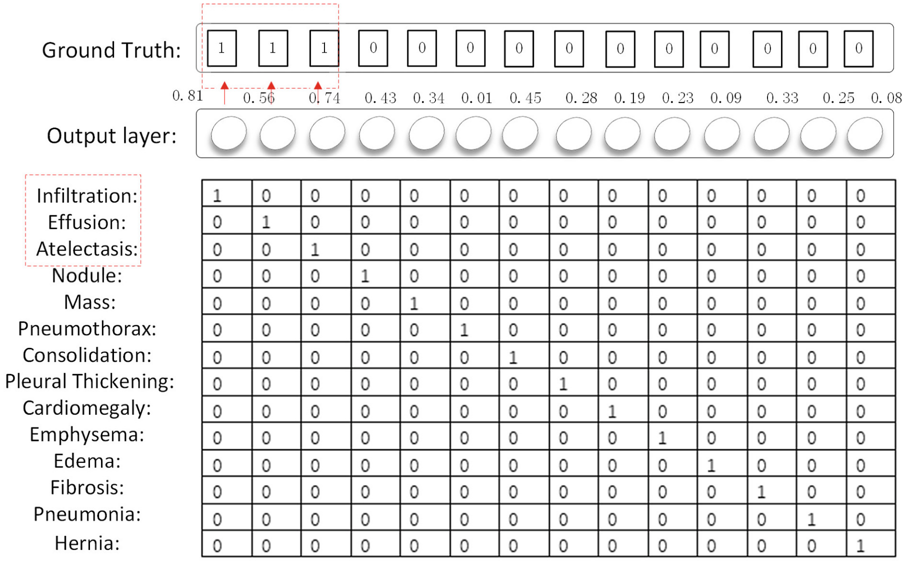
Fig. 1. The architectural diagram of our CheXray-m2 model. Parameters of pre-trained CNN model (modified from Vggnet or Resnet) are first trained on images as a 14-label classification problem, as shown in ①. Image Features are then extracted from a relatively low layer of the CNN model trained in ①, as shown in ②. The visual feature vectors are then fed into a Bi-LSTM network which predicts both the sequence of concentrated visual contexts and the sequence of generating words based on attention mechanism, shown in ③ and ④.

pooling layers behind randomly. Our framework can be easily extended to any other advanced CNN models and modified to be a proper encoder.

Secondly, parameters of the proposed encoder should not be the same as those pre-trained on ImageNet. There are ample evidences suggesting the transfer learning from natural images to chest x-ray images without training again is not a good choice [4,5]. As shown in Fig. 1, we therefore train the modified convolutional neural network by changing to a new classification layer and minimizing the multi-label loss function. The last layer has 14 output ports and each of them denotes a kind of lesion. Sigmoid layer turn the input into probability value between 0 and 1. If an output of this layer is larger than the threshold (here is 0.5), the model classify the image into this category. Figure 2 shows the coding rule of the ground truth and an example. If the label of a training image is infiltration, effusion and atelectasis, the label should be encoded as 11100000000000.

Thirdly, features should be extracted from a lower convolutional layer of the proposed encoder, which is unlike previous work. Extracting relatively low-level feature vectors allows the decoder to selectively focus on certain parts of an image by using the attention model to weight the vectors. As shown in (1), the encoder extracts L vectors, each of which is a D-dimensional representation corresponding to a part of the image.

$$A = \{a_1, ..., a_M\}, a_i \in R^N \tag{1}$$



Infiltration:	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Effusion:	0	1	0	0	0	0	0	0	0	0	0	0	0	0
Atelectasis:	0	0	1	0	0	0	0	0	0	0	0	0	0	0
Nodule:	0	0	0	1	0	0	0	0	0	0	0	0	0	0
Mass:	0	0	0	0	1	0	0	0	0	0	0	0	0	0
Pneumothorax:	0	0	0	0	0	1	0	0	0	0	0	0	0	0
Consolidation:	0	0	0	0	0	0	1	0	0	0	0	0	0	0
Pleural Thickening:	0	0	0	0	0	0	0	1	0	0	0	0	0	0
Cardiomegaly:	0	0	0	0	0	0	0	0	1	0	0	0	0	0
Emphysema:	0	0	0	0	0	0	0	0	0	1	0	0	0	0
Edema:	0	0	0	0	0	0	0	0	0	0	1	0	0	0
Fibrosis:	0	0	0	0	0	0	0	0	0	0	0	1	0	0
Pneumonia:	0	0	0	0	0	0	0	0	0	0	0	0	1	0
Hernia:	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Fig. 2. The coding rule of the ground truth. If an image contains infiltration, effusion and atelectasis, the label should be encoded as 11100000000000.

Decoder: Attention-Based Bi-LSTM. Recurrent Neural Networks (RNN) is a kind of neural network that models the dynamic temporal behavior of sequences through connections between the units. The t in the lower right corner of these elements represents the state at time t. Current hidden states h_t is updated as (2), where x is the input, h is the hidden layer unit, wo is the output word, and V, W, and U are weights.

$$h_t = tanh(Ww_t + Uh_{t-1} + b) \tag{2}$$

Current output o_t is updated as (3):

$$wo_t = Vh_t + c \tag{3}$$

Current hidden states h_t is generated by the previous hidden state h_{t-1} and the current word w_t. The recurrent transition makes h_t also contain information of the previously generated words and states h_{t-2}, h_{t-3}, h_{t-4}, ...

LSTM extends RNN by adding three gates to a RNN neuron: a forget gate f to control whether to forget the current state; an input gate i to indicate if it should read the input; an output gate o to control whether to output the state. We use a long short-term memory (LSTM) network that inspired by reference [27] where i_t^1, f_t^1, o_t^1, c_t^1 and h_t^1 represent the outputs of the input, forget, output gates, memory cell and hidden state of the LSTM respectively. (4), (5) and (6) show how those variables are related, where T is a properly defined affine transformation, σ is the logistic sigmoid function, $tanh$ is the hyperbolic tangent function and $\odot$ is element-wise multiplication.

$$\begin{pmatrix} i_t^1 \\ f_t^1 \\ o_t^1 \\ g_t^1 \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ tanh \end{pmatrix} T \begin{pmatrix} w_{t-1} \\ h_{t-1} \\ v_t \end{pmatrix} \tag{4}$$

$$c_t^1 = f_t^1 \odot c_{t-1}^1 + i_t^1 \odot g_t^1 \tag{5}$$

$$h_t^1 = o_t^1 \odot tanh(c_t^1) \tag{6}$$

Different from common LSTM, there is a context vector v_t, a dynamic representation of the relevant part of the image input at time t. There are two steps to computes v_t from the vectors $A = \{a_1, ..., a_M\}, a_i \in R^N$. Firstly, we compute the probability α_i of focusing on the ith location by an attention model f_{att}, as shown in (7). The inputs to the model is the extracted features from the ith visual element, the hidden state h_{t-1} and the previously generated word w_{t-1}. We believe that they can help us to find a proper v_t because of the association of lung lesions. Then, we use weighted sum to update the visual context vector v_t, as shown in (8).

$$\alpha_{it} = \frac{exp(f_{att}(a_i, h_{t-1}, w_{t-1}))}{\sum_{k=1}^{M} exp(f_{att}(a_k, h_{t-1}, w_{t-1}))} \tag{7}$$

$$v_t = \sum_i \alpha_{it} a_i \tag{8}$$

The next word w_t can be updated conditioning on the previously generated word w_{t-1}, the context vector v_t, and the decoder state h_t, as shown in (9) and (10):

$$p_{wt} = f_{softmax}(w_{t-1}, h_t^1, vt) \tag{9}$$

$$w_t = f_w(p_{wt}) \tag{10}$$

In order to make better use of the interdependency of lesions, we believe that w_{t+1} and h_{t+1} can also contribute to the update of h_t and the generation of the current word. We therefore define i_t^2, f_t^2, o_t^2, c_t^2 and h_t^2 in (11), (12) and (13):

$$\begin{pmatrix} i_t^2 \\ f_t^2 \\ o_t^2 \\ g_t^2 \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ tanh \end{pmatrix} T \begin{pmatrix} w_{t+1} \\ h_{t+1} \\ v_t \end{pmatrix} \tag{11}$$

$$c_t^2 = f_t^2 \odot c_{t+1}^2 + i_t^2 \odot g_t^2 \tag{12}$$

$$h_t^2 = o_t^2 \odot tanh(c_t^2) \tag{13}$$

The word w_t can be updated as (14) and (15):

$$p_{wt}^{new} = f_{softmax}(w_{t-1}, w_{t+1}, [h_t^1, h_t^2], vt) \tag{14}$$

$$w_t^{new} = f_w(p_{wt}^{new}) \tag{15}$$

3.2 Training

There are two steps of training. Firstly, we see the modified CNN as a model for multi-label classification and roughly train it on the whole Chest x-ray14 data set. The purpose of this step is to increase the transfer efficiency as medical images are quite different from natural images, which were used to train Vggnet and Resnet. We use binary cross entropy loss function to learn, defined as (16):

$$loss_1 = -\frac{1}{n}\sum_{i}[y_i lnp_i + (1-y_i)ln(1-p_i)] \tag{16}$$

Where y_i is the ground truth and $p_i \in [0,1]$ is the output of the sigmoid layer.

After that, we extract image features from a lower-level layer and feed them into decoder as our purpose is to train the decoder under the help of clinical contexts and visual contexts. We train our network on 16637 of 20796 training images (setting aside 2079 for validation and 2080 for test) on Chest x-ray14 data set. The length of Bi-LSTM is $T = 4$ as most images contain less than 4 lesions. The scale of the vocabulary is $C = 14$ as the number of lesions is 14 in this data set. The loss function is also a binary cross entropy loss function, shown as (17):

$$loss_2 = -\frac{1}{n}\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{|C|}[y_{ti} lnp_{ti} + (1-y_{ti})ln(1-p_{ti})] \tag{17}$$

As for some setups, the experimental environment is an ubuntu linux server with 2 GeForce GTX 1080 Ti GPUs and the model is implemented using Tensorflow. Due to the limit of GPU memory, the batch size = 8 is set as a constant. According to experience and the validation results, we finally chose stochastic gradient descent (SGD) and the learning rate is decayed from 0.01.

4 Experiment

4.1 Dataset

To verify the efficacy of CXNet-m2 in medical diagnosis on chest x-ray images, we conduct experiments on Chestx-ray14 introduced in reference [5]. It contains more than 30,000 patients, 112,000 labeled chest x-ray images and 14 kinds abnormal images including Infiltration, Effusion, Atelectasis, Nodule, Mass, Pneumothorax, Consolidation, Pleural Thickening, Cardiomegaly, Emphysema, Edema, Fibrosis, Pneumonia and Hernia. Among them, there are 20795 muitilesion images where the number of labels ranges from 2 to 14.

4.2 Metrics

AUC[28]**.** There are 4 quantities and the specific definitions shown in Table 1. The ROC curve has typically horizontal axis as specificity and vertical axis as sensitivity, where sensitivity is computed as $TP/(TP + FN)$ and specificity is defined as $TN/(TN + FP)$. AUC is the Area Under the ROC Curve, and using AUC to evaluate the result is more clear and direct than ROC. In order to compare our model with other models, we use AUC as one of the metric because it is widely used in many references to measure experimental results. We did not consider the clinical relevance to compute AUC as it is intractable to compute with equation (17).

BLEU[29]**.** BLEU is a popular metric of machine translation including image captioning that analyzes the co-occurrences of n-grams between the result and ground truth. It computes a corpus-level clipped n-gram precision between sentences.

ROUGE-L[30]**.** ROUGE is a set of evaluation metrics designed to evaluate text summarization algorithms. Given the length of Longest Common Subsequence (LCS) between a pair of sentences, ROUGE-L is found by computing recall and precision of LCS.

Table 1. 4 Evaluation results and corresponding symbols

Quantities	Descriptions
TP	The prediction is 1 with ground truth 1
TN	The prediction is 0 with ground truth 0
FP	The prediction is 1 with ground truth 0
FN	The prediction is 0 with ground truth 1

4.3 Results

The AUC per abnormality is shown in Table 2, compared with the result of Reference [5] and Reference [13] including average AUC. It can be found that the total AUC of our model is better, despite of the lower value of cardiomegaly, edema and hernia than those of Reference [13]. In fact, the structure of our CNN encoder is similar with that in Reference [5]. The main reason of higher AUC may be that we use larger training set and there is patient-wise overlap between training and test sets for learning more distinguishable features. The basic structure of Reference [13] is Densenet, which we used but did not get the best average result. The BLEU and ROUGE-L are compared in Table 3, where there are three versions of CXNet-m2. CXNet-m2-a is the basic model with considering the visual and clinical contexts by LSTM, CXNet-m2-b is trained by attention-based LSTM with pre-training in Fig. 1① and CXNet-m2-c is trained by attention-based Bi-LSTM with pre-training in Fig. 1①. It can be found that the pre-training step in Fig. 1① is important to improve the performance of the model and using more clinical information helps learn the pattern accurately.

Table 2. Result evaluation by AUC

Abnormality	Reference [5]	Reference [13]	CXNet-m2
Atelectasis	0.716	0.772	**0.788**
Cardiomegaly	0.807	**0.904**	0.848
Effusion	0.784	0.859	**0.865**
Infiltration	0.609	0.695	**0.702**
Mass	0.706	0.792	**0.793**
Nodule	0.671	0.717	**0.732**
Pneumonia	0.633	0.713	**0.719**
Pneumothorax	0.806	0.841	**0.861**
Consolidation	0.708	0.788	**0.792**
Edema	0.835	**0.882**	0.869
Emphysema	0.815	0.829	**0.853**
Fibrosis	0.769	0.767	**0.781**
PT	0.708	0.765	**0.774**
Hernia	0.767	**0.914**	0.853
A.V.G	0.738	0.798	**0.802**

Table 3. Result evaluation by BLEU and ROUGE-L

Models	BLEU-1	ROUGE-L
CXNet-m2-a	0.632	0.636
CXNet-m2-b	0.724	0.713
CXNet-m2-c	0.739	0.741

5 Conclusion

Chest X-ray is the most popular mean to detect lung lesion and deep learning is a good tool to assist the diagnosis. For detecting multiple lesions on a chest x-ray image, the common approach is to see it as a multi-label classification problem and break the multi-label classification problem into independent binary classification problems. It takes advantage of a rich body of work on binary classification but suffers from ignoring the interdependencies between labels. To avoid this problem, we propose an unified model that extracts the image features with a modified CNN encoder and predicts multiple lesions with an attention-based Bi-LSTM decoder by making use of visual and clinical contexts. Inspired by image captioning and attention mechanism, we adapt the multi-label classification problem into an image understanding problem as researches have proved there are relationships between lesions. Quantitative and qualitative results demonstrate that our method significantly outperforms the state-of-the-art algorithm.

In the future, we will continue to explore the dependencies between lesions on chest x-ray images from a perspective of association mining. How to take advantage of clinical contexts to predict multiple lesions on chest x-ray images is still the emphasis of our further research.

Acknowledgements. This research is sponsored by Fundamental Research Funds for the Central Universities (No. 2016NT14), National Natural Science Foundation of China (No. 61601033, No. 61571049) and Inter-discipline Research Funds of Beijing Normal University (BNUXKJC1825).

References

1. Ari, S., Hembram, K., Saha, G.: Detection of cardiac abnormality from PCG signal using LMS based least square SVM classifier. Expert Syst. Appl. **37**(12), 8019–8026 (2010)
2. Niu, D., Li, Y., Dai, S., et al.: Sustainability evaluation of power grid construction projects using improved TOPSIS and least square support vector machine with modified fly optimization algorithm. Sustainability **10**(1), 231 (2018)
3. Hu, Z., Tang, J., Zhang, P., et al.: Identification of bruised apples using a 3-D multi-order local binary patterns based feature extraction algorithm. IEEE Access **6**, 34846–34862 (2018)
4. Xu, S., Hao, W., Bie, R.: CXNet-m1: anomaly detection on chest X-Rays with image-based deep learning. IEEE Access **7**, 4466–4477 (2019)
5. Wang, X., Peng, Y., Lu, L., et al.: 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)
6. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
7. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
8. Szegedy, C., Liu, W., Jia, Y., et al.: Going deeper with convolutions. In: CVPR (2015)
9. He, K., Zhang, X., Ren, S., et al.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
10. Rajpurkar, P., Irvin, J., Zhu, K., et al.: CheXNet: radiologist-level pneumonia detection on chest X-Rays with deep learning. arXiv preprint arXiv: 1711.05225 (2017)
11. Huang, G., Liu, Z., Weinberger, K.Q., et al.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, no. 2, p. 3 (2017)
12. Li, Z., et al.: Thoracic disease identification and localization with limited supervision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018)
13. Yao, L., Poblenz, E., Dagunts, D., et al.: Learning to diagnose from scratch by exploiting dependencies among labels. arXiv preprint arXiv:1710.10501 (2017)

14. Luke, M.J., et al.: Chronic nasopharyngeal obstruction as a cause of cardiomegaly, cor pulmonale, and pulmonary edema. Pediatrics **37**(5), 762–768 (1966)
15. Dodek, A., Kassebaum, D.G., Bristow, J.D.: Pulmonary edema in coronary-artery disease without cardiomegaly: paradox of the stiff heart. N. Engl. J. Med. **286**(25), 1347–1350 (1972)
16. Castrejon, L., Kundu, K., Urtasun, R., et al.: Annotating object instances with a polygon-RNN. In: CVPR, vol. 1, p. 2 (2017)
17. Williams, R.J., Zipser, D.: A learning algorithm for continually running fully recurrent neural networks. Neural Comput. **1**(2), 270–280 (1989)
18. Chung, J., Gulcehre, C., Cho, K., et al.: Gated feedback recurrent neural networks. In: International Conference on Machine Learning, pp. 2067–2075 (2015)
19. Schuster, M., Paliwal, K.K.: Bidirectional recurrent neural networks. IEEE Trans. Signal Process. **45**(11), 2673–2681 (1997)
20. Yantis, S.: Control of visual attention. Attention **1**(1), 223–256 (1998)
21. Ha, T.L., Niehues, J., Waibel, A.: Effective strategies in zero-shot neural machine translation. arXiv preprint arXiv:1711.07893 (2017)
22. Rush, A.M., Chopra, S., Weston, J.: A neural attention model for abstractive sentence summarization. arXiv preprint arXiv:1509.00685 (2015)
23. Fang, H., et al.: From captions to visual concepts and back. In: Proceedings of IEEE Computer Vision and Pattern Recognition, pp. 1473–1482 (2015)
24. Venugopalan, S., Hendricks, L.A., Rohrbach, M., et al.: Captioning images with diverse objects. arXiv preprint arXiv:1606.07770, vol. 1, no. 3 (2016)
25. Mao, J., Xu, W., Yang, Y., et al.: Explain images with multimodal recurrent neural networks. arXiv preprint arXiv:1410.1090 (2014)
26. Zhang, Y., Yuan, L., Guo, Y., et al.: Discriminative bimodal networks for visual localization and detection with natural language queries. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
27. Fu, K., et al.: Aligning where to see and what to tell: image captioning with region-based attention and scene-specific contexts. IEEE Trans. Pattern Anal. Mach. Intell. **39**(12), 2321–2334 (2017)
28. Lobo, J.M., Jiménez-Valverde, A., Real, R.: AUC: a misleading measure of the performance of predictive distribution models. Glob. Ecol. Biogeogr. **17**(2), 145–151 (2008)
29. Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: BLEU: a method for automatic evaluation of machine translation. In: ACL (2002)
30. Lin, C.-Y.: Rouge: a package for automatic evaluation of summaries. In: ACL Workshop (2004)