Automated Radiological Impression Generation for Plain Chest X-rays with End to End Deep Learning

Authors: Zhang, Shuai, Xin, Xiaoyan, Shen, Jingtao, Guo, Yachong, Wang, Yang, Yang, Xianfeng, Wang, Jun, Zhang, Jian, Zhang, Bing, Zhang, Jian, Zhang, Bing

Date: 2020-06-09T00:00:00+00:00

Abstract

Chest X-ray (CXR) is one of the most common clinical examinations used to diagnose thoracic diseases and abnormalities. The volume of CXR scans generated daily in hospitals is substantial. Therefore, an automated diagnosis system capable of reducing the workload of physicians would be of significant value. Currently, applications of artificial intelligence in CXR diagnosis typically employ pattern recognition to classify scans. However, such methods rely on labeled databases, which are costly to produce and often exhibit high error rates. In this work, we constructed a database comprising over 12,000 CXR scans and radiological reports, and developed a model based on deep convolutional neural networks and recurrent networks with attention mechanisms. The model learns features directly from CXR scans and associated raw radiological reports without requiring additional labeling. The model provides automated recognition of input scans and generation of impressions. The quality of generated impressions was evaluated using both CIDEr metrics and radiologist assessment. CIDEr scores averaged approximately 5.8 on the test dataset. Further blind evaluation demonstrated performance comparable to that of radiologists.

Full Text

Preamble

Automated Radiological Impression Generation for Plain Chest X-rays with End-to-End Deep Learning

Shuai Zhang[1],[2]†, Xiaoyan Xin[1]†, Jingtao Shen[3], Yachong Guo[2],[4], Yang Wang[1], Xianfeng Yang[1], Jun Wang[2],[4],[5], Jian Zhang[2],[4],[5], *Bing Zhang*[1]

[1] Department of Radiology, Nanjing Drum Tower Hospital, the Affiliated Hospital of Nanjing University Medical School, Nanjing 210008, P.R. China
[2] School of Physics, Collaborative Innovation Center of Advanced Microstructures, Nanjing University, Nanjing, China
[3] Department of Nuclear Medicine, Nanjing Drum Tower Hospital, the Affiliated Hospital of Nanjing University Medical School, Nanjing 210008, P.R. China
[4] State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China

† Shuai Zhang and Xiaoyan Xin contributed equally to this work and are considered co-first authors.
*Correspondence authors: Jian Zhang, jzhang@nju.edu.cn; Bing Zhang, zhangbing_{nanjing}@nju.edu.cn

Abstract

The chest X-ray (CXR) is one of the most common clinical examinations used to diagnose thoracic diseases and abnormalities. The volume of CXR scans generated daily in hospitals is enormous, making an automated diagnosis system that can reduce physician workload extremely valuable. Current applications of artificial intelligence in CXR diagnosis typically employ pattern recognition to classify scans; however, such methods rely on labeled databases that are costly to create and often exhibit high error rates. In this work, we constructed a database containing more than 12,000 CXR scans with corresponding radiological reports and developed a model based on deep convolutional neural networks and recurrent networks with attention mechanisms. Our model learns features directly from CXR scans and associated raw radiological reports without requiring additional labeling. The system provides both automated recognition of given scans and generation of radiological impressions. The quality of generated impressions was evaluated using CIDEr scores and by radiologist assessment, achieving an average CIDEr score of approximately 5.8 on the testing dataset. Further blind evaluation demonstrated performance comparable to radiologists.

Keywords: Chest X-ray, Machine learning, Deep learning, Neural networks, Attention mechanism

Introduction

The chest X-ray (CXR) is one of the most common diagnostic techniques for respiratory system evaluation. It is rapid, inexpensive, and involves low radiation exposure. Hospitals generate enormous volumes of CXR scans daily, and their examination and interpretation consume substantial time and effort from radiologists. Consequently, developing an automated system capable of examining and interpreting CXR radiographs is highly desirable. Moreover, such a system may help reduce inter-observer variations arising from factors including individual experience, radiograph quality, time constraints, and personality differences [1]. Adoption of an automated system would lead to more standardized

terminology and treatment protocols, benefiting collaborative efforts across different institutions and enabling new applications such as remote diagnosis and self-service screening.

Previous research has focused primarily on automated classification of CXR scans, typically employing variants of Convolutional Neural Networks (CNNs) with supervised learning [2-9]. However, three major problems hinder the practical hospital deployment of these methods. First, labels for chest films are usually extracted from reports, and their accuracy is not guaranteed. Second, the sensitivity and false positive rates of these classification approaches have plateaued, as many diseases cannot be distinguished by experts based solely on chest film examination. Third, the decision-making strategies underlying these systems remain poorly understood, making it difficult to track errors and gain the trust of physicians and patients.

In this work, we developed a model based on deep convolutional neural networks and recurrent neural networks with attention mechanisms that simultaneously learns from CXR images and raw radiological reports. Deep neural networks have demonstrated great potential for characterizing and classifying complex data across numerous fields [10-12]. After training on our database, the network can automatically generate radiological impressions for given scans. Our work offers three novel features. First, our model requires only weak supervision; it learns directly from images and raw radiological reports stored in hospital databases without requiring further human classification or labeling. This feature greatly facilitates data acquisition and large-scale model training. Second, rather than simply classifying cases into one or several disease categories, our model outputs descriptive reports regarding various chest conditions that are directly readable by physicians. Third, the attention mechanism provides additional insight into how the model operates, facilitating debugging and optimization. In the following sections, we describe the model architecture, training and testing procedures, and performance evaluation using CIDEr scores and human radiologist assessment.

Methods

Network Architecture

[Figure 1: see original paper] shows the overall architecture of our neural network. During training, the network reads both CXR images and raw radiological reports, outputting human-readable text. The generated output is compared with ground truth to compute the loss function, which is minimized using gradient back-propagation. After training, the model can automatically generate impressions for given CXR images. Our architecture was inspired by the pioneering work of Xu et al. [13], who developed an RNN to generate captions for everyday images from Flickr and MS COCO databases. Our model shares similar goals with those of Zhang et al. [14] and Wang et al. [15] for automatic medical report generation, though we redesigned the architecture to better ac-

commodate our database structure.

The model incorporates a 121-layer Densely Connected Convolutional Network (DenseNet) [16] as a visual information encoder to extract features from input images. The encoder comprises four blocks, each containing several convolutional layers that take all preceding feature maps as inputs. Transition layers connect these blocks. According to Huang et al., DenseNets alleviate the vanishing gradient problem, strengthen feature propagation, encourage feature reuse, and substantially reduce parameter count [16]. Compared to many other CNNs, they converge faster and perform well on smaller datasets, making them particularly suitable for medical images. The output from DenseNet' s final layer is fed into a Long Short-Term Memory (LSTM) network to generate descriptions for the given CXR image.

The LSTM network [17] generates text word-by-word for each CXR image. At each step, it reads the output from DenseNet' s final layer and the previously generated word, then outputs the next word. Our LSTM implementation follows [18]:

tanh(where i_t, f_t, c_t, o_t, h_t are the input, forget, memory, output and hidden gates of the LSTM, respectively.

Attention mechanisms have been widely adopted in visual image processing as they improve model performance and provide insight into model operation. They mimic human visual attention by learning to focus on specific image regions. We implemented a soft attention mechanism that calculates a set of weights conditioned on the image representation and hidden state. These weights are multiplied with DenseNet output vectors to obtain a weighted image representation, which the recurrent neural network then uses to generate descriptions. The corresponding equations are:

softmax($p_{ht}CV\alpha = \exp Wh{+}WC$ softmax $Wh{+}WC$ where α_t is the attention weight, V is the DenseNet out

The loss function is the cross-entropy between ground truth and predicted text distributions:

$(,) = -\sum$ where p_j is the probability distribution predicted at the j-th step, y_j is the index of the j-th word

Datasets

All chest X-ray scans and associated radiological reports were provided by Nanjing Drum Tower Hospital. The dataset comprises 12,219 images with an equal number of Chinese-language reports. Among these, 7,516 cases were from the outpatient department and 5,063 from physical examinations. Based on the

corresponding reports, 7,547 cases were normal and 4,672 were abnormal. All reports were reviewed by one expert (attending physician level or above) and double-checked by another expert (associate chief physician level or above). The dataset was randomly split into three subsets: 80% for training, 10% for validation, and 10% for testing. [Figure 2: see original paper] shows randomly selected examples from the dataset.

Since Chinese text lacks spaces between words, we used the Python module jieba [19] for text segmentation. Processing all radiological reports yielded a vocabulary of 424 words, represented as one-hot vectors. Words appearing fewer than three times were replaced with a special token `<nou>`. Two additional special tokens, `<start>` and `<end>`, marked report boundaries.

Training Procedures

We employed transfer learning to accelerate training convergence. Specifically, the 121-layer DenseNet was pre-trained on the ChestX-ray8 dataset released in September 2017 [20] for a supervised classification task, following the procedure in [2]. The ChestX-ray8 dataset contains 110,000 chest X-ray images with 14 disease type labels. These trained weights were transferred to our model' s encoder module. During subsequent training, parameters in the first two dense blocks were fixed while remaining parameters were fine-tuned via gradient back-propagation.

During training, original X-ray images were resized to 256$×256*pixelsandprocessedwithhistogramequalization*
4}$ and LSTM learning rate to 5.0$×10^{-4}$.

Evaluation Metrics

To evaluate how well a generated sentence c_i matches the consensus of reference descriptions $s_{i1}, ..., s_{im}$ for image I_i, we used the Consensus-based Image Description Evaluation (CIDEr) score [21]. CIDEr calculation first computes Term Frequency-Inverse Document Frequency (TF-IDF) weighting for each n-gram ω_k in sentence s_{ij}:

min(1, where is the number of times n-gram ω_k occurs in sentence s_{ij}, Ω is the vocabulary of all n-grams and

Then the CIDEr$_n$ score for n-grams of length n is calculated as:

CIDEr(where is a vector formed by corresponding to all n-grams of length n. similarly defined for the genera

Finally, the CIDEr score is computed as the average over all n-grams:

CIDEr(CIDEr(

Results

Figure 3: see original paper shows training and validation losses as functions of training epoch. The training loss continues decreasing while validation loss saturates around the 10th epoch, indicating maximum model generalization. Therefore, parameters from the 10th epoch were used for all subsequent results.

Figure 3: see original paper shows CIDEr values for the testing dataset across epochs. Note that ground truth sentences were not used during description generation for the test set; they were only used for post-generation evaluation. We employed Beam Search [22] to generate multiple sentences per CXR image, each assigned a preference probability. The top three sentences with highest probabilities were recorded, their CIDEr values calculated against ground truth, and the highest value used for the curve in Fig. 3(b). The average CIDEr value for the testing set increases with epoch and saturates around 5.8 at the 10th epoch.

[Figure 4: see original paper] shows several examples of generated descriptions. For each scan, we present the top three predictions (Pd1, Pd2, Pd3) in decreasing order of preference probability. Additional examples are provided in supplemental materials. Fig. 4(a) shows a normal case with increased lung markings in both lungs. The model correctly recognizes this condition, generating descriptions containing "increased lung markings in both lungs" in Pd1 and Pd3, while Pd2 states "no obvious abnormalities." Fig. 4(b) shows a patient with chronic bronchitis and inflammation, diagnosed based on both imaging and medical history. The model reports "increased lung markings" in Pd1 and Pd2, and directly provides "bronchitis"in Pd3—remarkable given the model lacks medical history information. Fig. 4(c) demonstrates a case with right-sided pleural effusion, which the model correctly identifies. In Fig. 4(d), the model correctly recognizes cardiomegaly and infers a postoperative view, possibly based on thin, bright strips near the cardiac region.

[Figure 5: see original paper] illustrates the alignment between generated words and relevant CXR image regions. These alignments generally align with human intuition, enabled by the attention mechanism to provide deeper understanding of network operation and facilitate result debugging.

We also evaluated report quality through radiologist assessment. We randomly extracted 100 CXR scans from the testing dataset to generate impressions with our model, and separately extracted another 100 CXR scans with corresponding human-written reports from the same dataset. These 200 scans and reports were combined, shuffled, and sent to experts for blind evaluation. Two associate chief physician-level radiologists examined the images and assessed report quality without knowing whether reports originated from humans or machines, preventing bias. Radiologists scored each report from 1 to 5 based on the following criteria: 5—all conditions identified and accurately described; 4—major conditions correctly identified but minor extra-thoracic problems missed (e.g., scoliosis, foreign objects); 3—major conditions correctly identified but minor

intra-thoracic problems missed (e.g., old lesions, fibrous stripes, post-thoracic surgery, aortic calcification); 2—major conditions identified but described inaccurately; 1—major conditions missed or identified incorrectly.

[Figure 6: see original paper] shows score distributions for both report groups. The majority received score 5 in both groups: 77% for human-written reports and 72% for model-generated reports. At score 4, the percentages were 9% and 14%, respectively. Considering scores of 4 or above as acceptable, both groups achieved 86% in this range, demonstrating that our neural network generates reports with quality comparable to radiologists.

Discussion

In summary, we developed a scheme for automatic radiological description generation from CXR images using deep convolutional neural networks and recurrent neural networks with attention mechanisms. We constructed a database of over 12,000 CXR scans, trained our model, and evaluated description quality. Comparison with ground truth yielded a CIDEr value of 5.8. We also blended model-generated descriptions with radiologist-written reports and invited other radiologists to score them blindly. Human-written reports received the highest score (5) in 77% of cases, while model-generated reports achieved this in 72% of cases. For score 4, the percentages were 9% and 14%, respectively. Thus, our model generates high-quality reports comparable to radiologists and shows potential for significant improvement with additional training data.

Our model possesses several distinctive features. First, it learns from raw radiological reports and can directly utilize the vast volume of CXR data generated in hospitals without requiring additional labeling efforts—particularly valuable since medical annotation acquisition is extremely expensive. Second, the model outputs descriptive reports for given CXR images rather than simple disease classifications. This design reflects clinical reality, where solid conclusions cannot always be drawn from CXR images alone. For instance, prominent lung markings may indicate infection, chronic bronchitis, interstitial lung disease, heart failure, or normal aging. When such symptoms are observed, describing the finding is more appropriate than classifying a specific disease. Our model follows this strategy, mimicking radiologists' daily practice.

However, the model requires further improvement. As an end-to-end architecture that directly learns from and generates reports, it does not explicitly provide classification results, making quantitative performance evaluation challenging. We currently rely on human inspection for evaluation and are addressing this limitation by adding a classification module to the neural network.

We believe the automated AI system developed in this work is valuable and will substantially reduce physician workload in the near future.

Acknowledgments

This work was funded by the National Natural Science Foundation of China (Grant No. 11774158 to JZ, 11774157 to JW, 81720108022, 91649116 and 81571040 to BZ, and 11334004 to WW); by the Social Development Project of Science and Technology in Jiangsu Province (BE2016605 and BE201707 to BZ); in part by the "Six Big Talent Peak" High-Level Talent Project (2016-WSN-160) from Jiangsu Provincial Department of Human Resources; by the Key Project supported by Medical Science and Technology Development Foundation, Nanjing Department of Health (YKK18062, YKK17058); and by the 66th batch of China Postdoctoral Science Foundation surface projects (2019M661805). The authors acknowledge HPCC of Nanjing University for computational support.

Conflict of Interests: The authors declare no conflict of interests.

References

[1] Tudor G, Finlay D, and Taub N. An assessment of inter-observer agreement and accuracy when reporting plain radiographs. *Clinical Radiology* 1997;52(3):235-238. doi: 10.1016/S0009-9260(97)80280-2

[2] Rajpurkar P, Irvin J, Zhu K, Yang B, Mehta H, Duan T et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv preprint 2017;arXiv:1711.05225.

[3] Ypsilantis PP and Montana G. Learning what to look in chest x-rays with a recurrent visual attention model. arXiv preprint 2017;arXiv:1701.06452.

[4] Pesce E, Ypsilantis PP, Withey S, Bakewell R, Goh V, and Montana G. Learning to detect chest radiographs containing lung nodules using visual attention networks. arXiv preprint 2017;arXiv:1712.00996.

[5] Islam MT, Aowal MA, Minhaz AT, and Ashraf K. Abnormality detection and localization in chest x-rays using deep convolutional neural networks. arXiv preprint 2017;arXiv:1705.09850.

[6] Yao L, Poblenz E, Dagunts D, Covington B, Bernard D, and Lyman K. Learning to diagnose from scratch by exploiting dependencies among labels. arXiv preprint 2017;arXiv:1710.10501.

[7] Yan C, Yao J, Li R, Xu Z, and Huang J. Weakly supervised deep learning for thoracic disease classification and localization on chest x-rays. In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics* 2018;103-110.

[8] Guan Q, Huang Y, Zhong Z, Zheng Z, Zheng L, and Yang Y. Diagnose like a radiologist: Attention guided convolutional neural network thorax disease classification. arXiv preprint 2018;arXiv:1801.09927.

[9] Rubin J, Sanghavi D, Zhao C, Lee K, Qadir A and Xuwilson M. Large scale automated reading of frontal and lateral chest x-rays using dual convolutional neural networks. arXiv preprint 2018;arXiv:1804.07839.

[10] Goodfellow I, Bengio Y, and Courville A. *Deep Learning* (Adaptive Computation and Machine Learning series). Adaptive Computation and

Machine Learning series 2016;800-800.
[11] Silver D, Huang A, Maddison CJ, Guez A, Sifre L, Van DJ et al. Mastering the game of go with deep neural networks and tree search. *Nature* 2016;529:484-489. doi: 10.1038/nature16961
[12] Xiong HY, Alipanahi B, Lee LJ, Bretschneider H, Merico D, Yuen RK et al. The human splicing code reveals new insights into the genetic determinants of disease[J]. *Science* 2015;347(6218):1254806-1254806. doi: 10.1126/science.1254806
[13] Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhudinov R et al. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning* 2015;2048-2057.
[14] Zhang Z, Xie Y, Xing F, McGough M, and Yang L. Mdnet: A semantically and visually interpretable medical image diagnosis network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2017;6428-6436. doi: 10.1109/CVPR.2017.378
[15] Wang X, Peng Y, Lu L, Lu Z, and Summers RM. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In *Computer Vision and Pattern Recognition* 2018;9049-9058.
[16] Huang G, Liu Z, Maaten LVD, and Weinberger KQ. Densely connected convolutional networks. In *Computer Vision and Pattern Recognition* 2017;2261-2269. doi:10.1109/CVPR.2017.243
[17] Hochreiter S and Schmidhuber J. Long short-term memory. *Neural Computation* 1997;9:1735-1780. doi:10.1162/neco.1997.9.8.1735
[18] Zaremba W, Sutskever I, and Vinyals O. Recurrent neural network regularization. arXiv preprint 2014;arXiv:1409.2329.
[19] Sun J. Chinese word segmentation module. https://github.com/fxsjy/jieba.
[20] Wang X, Peng Y, Lu L, Lu Z, Bagheri M, and Summers RM. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Computer Vision and Pattern Recognition* 2017;3462-3471. doi: 10.1109/CVPR.2017.369.
[21] Vedantam R, Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Computer Vision and Pattern Recognition* 2015;4566-4575. doi: 10.1109/CVPR.2015.7299087
[22] Sutskever I, Vinyals O, and Le QV. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems* 2014;3104-3112.

Note: Figure translations are in progress. See original paper for figures.