

UNIVERSITY OF TECHNOLOGY SYDNEY
Faculty of Engineering and Information Technology

**CHEST X-RAY IMAGE CLASSIFICATION WITH
DEEP LEARNING**

by

Qingji Guan

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

Doctor of Philosophy

Sydney, Australia

2021

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, Qingji Guan, declare that this thesis, is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the Faculty of Engineering and Information Technology at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise reference or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Production Note:

Signature: Signature removed
prior to publication.

Date: 1 Sep 2021

ABSTRACT

CHEST X-RAY IMAGE CLASSIFICATION WITH DEEP LEARNING

by

Qingji Guan

Computer-aided diagnosis (CAD) systems have been successfully helped to clinical diagnosis. This dissertation considers one essential task in CAD, the chest X-ray (CXR) image classification problem, with the deep learning technologies from the following three aspects.

First, considering most diseases existing in CXRs usually happen in small localized areas, we propose to localize the local discriminative regions and integrate the global and local cues into an attention guided convolution neural network (AG-CNN) to identify thorax diseases. AG-CNN consists of three branches (global, local, and fusion branches). The global branch learns the global features for classification. The local branch localizes the discriminative regions, which avoids noise and improves misalignment in the global branch. AG-CNN fuses the global and local features for diagnosis in a fusion branch.

Second, due to the common and complex relationships of multiple diseases in CXRs, it is worth exploiting their correlations to help the diagnosis. This thesis will present a category-wise residual attention learning method to concentrate on learning the correlations of multiple diseases. It is expected to suppress the obstacles of irrelevant categories and strengthen the relevant features at the same time.

Last, a robust and stable CXR image analysis system should be able to: 1) automatically focus on the disease-critical regions, which usually are of small sizes; 2) adaptively capture the intrinsic relationships among different disease features and utilize them to boost the multi-label disease recognition rates jointly. We introduce a discriminative feature learning framework, ConsultNet, to achieve those two purposes simultaneously. ConsultNet consists of a variational selective information bottleneck branch and a spatial-

and-channel encoding branch. These two branches learn discriminative features collaboratively.

In addition, each of the proposed methods is comprehensively verified and analyzed by conducting various experiments.

Dissertation directed by Professor Yi Yang

Australian Artificial Intelligence Institute (AII), School of Computer Science

Acknowledgements

I would like to dedicate this dissertation to those who have offered assistance and support while pursuing the Ph.D. degree.

First and foremost, I would like to thank my supervisor, Prof. Yi Yang, for his guidance, encouragement, and most of all, his support during my preparation for the thesis. Prof. Yang taught me to insist on being myself and always do the right things in research work. He has also given me many valuable suggestions in lifetime. He has my deepest gratitude.

Besides, I would like to thank my co-supervisor, Dr. Liang Zheng, who has given me valuable suggestions and guidance in my research. His profound thinking and painstaking inspire and encourage me to do my best to do perfect work.

Thanks to all my colleagues in the ReLER Lab at the University of Technology Sydney. Thanks to Qianyu Feng, Peike Li, Pingbo Pan, Zhedong Zheng, Fan Ma, Xiaohan Wang, Yu Wu, Yutian Lin, Hehe Fan, Yanbin Liu, Jiayu Miao, Hu Zhang, Guang Li, Xiaolin Zhang, Ruoyu Liu, Yang He, Zongxin Yang. Special thanks to Zhun Zhong, Yawei Luo and Ping Liu, who have many discussions with my research together. I am grateful and delighted to meet these lovely, passionate people.

I would like to thank my parents Fengxian Sheng and Chengwen Guan, my brother Qingli Guan for their love over the years.

Finally and most importantly, I would like to thank my husband, Liming Yang, for his selfless love. Without his tremendous support, I could not complete my study. I thank my son, Yichen Yang, who always cheers me up.

Qingji Guan

March 2021 at Beijing.

List of Publications

Journal Papers

- J-1. **Q. Guan**, Y. Huang, Y. Luo, P. Liu, M. Xu and Y. Yang, “Discriminative feature learning for thorax disease classification in chest X-ray images,” *IEEE Transactions on Image Processing*, vol. 30, 2021.
- J-2. **Q. Guan**, Y. Huang, Z. Zhong, Z. Zheng, L. Zheng and Y. Yang. “Thorax disease classification with attention guided convolutional neural network ,” *Pattern Recognition Letters*, vol. 131, 38-45, 2020.
- J-3. **Q. Guan** and Y. Huang. “Multi-label chest X-ray image classification via category-wise residual attention learning ,” *Pattern Recognition Letters*, vol. 130, 259-266, 2020.

Contents

Certificate	ii
Abstract	iii
Acknowledgments	v
List of Publications	vi
List of Figures	x
1 Introduction	1
1.1 Background	1
1.1.1 Combining the global and local cues	2
1.1.2 Exploiting the correlations of multiple diseases	2
1.1.3 Learning discriminative features for CXR classification	3
1.2 Thesis Organization	3
2 Literature Review	5
2.1 Chest X-ray image classification with deep learning	5
2.1.1 Lesion areas related methods	6
2.1.2 Multi-label learning	8
3 Attention Guided Convolution Neural network	11
3.1 Motivation	11
3.2 Method	14
3.2.1 Structure of AG-CNN	14

3.2.2	Attention Guided Mask Inference	16
3.2.3	Training Strategy of AG-CNN	18
3.3	Experiment	19
3.3.1	Experimental details	19
3.3.2	Evaluation	20
3.3.3	Parameter Analysis	31
3.4	Conclusion	33
4	Category-wise Residual Attention Learning	34
4.1	Introduction	34
4.2	The proposed method	36
4.2.1	Architecture of CRAL	36
4.2.2	Feature Embedding	36
4.2.3	Category-wise Residual Attention Learning	38
4.2.4	Optimization	42
4.3	Experiment	42
4.3.1	Dataset and Evaluation Metric	43
4.3.2	Experimental Settings	44
4.3.3	Evaluation	44
4.3.4	Ablation Study	49
4.3.5	Qualitative results	49
4.4	Conclusion	52
5	Discriminative Feature Learning	53
5.1	Introduction	53
5.2	Methodology	57

5.2.1	Problem Settings and Motivation	57
5.2.2	Variational Selective Information Bottleneck	59
5.2.3	Spatial-and-Channel Encoding	62
5.2.4	Optimization with Pairwise Confusion	65
5.3	Experiment	67
5.3.1	Datasets	67
5.3.2	Implementation Details	68
5.3.3	Comparative Studies	71
5.3.4	Effectiveness of <i>ConsultNet</i>	78
5.4	Conclusion	82
6	Conclusions and Future Work	83
6.1	Summary of Contributions	83
6.2	Future Directions	84
	Bibliography	85

List of Figures

3.1	Two images from the ChestX-ray14 dataset. (a) The global images. (b) heatmaps extracted from a specific convolutional layer. (c) The cropped images from (a) guided by (b).	12
3.2	Overall framework of the attention guided convolutional neural network (AG-CNN, showing ResNet-50 as backbone). "BCE" represents binary cross entropy loss. The spatial resolution of heatmap generated from the last convolutional layer of the global branch is 7×7 . Then we resize the heatmap to 224×224 by bilinear interpolation. The input image is added to the heatmap for visualization.	15
3.3	The process of lesion area generation. (Top:) global CXR images of various thorax diseases for the global branch. Note that we do not use the bounding boxes for training or testing. (Middle:) corresponding visual examples of the output of the mask inference process. Higher/lower response is denoted with red/blue. (Bottom:) cropped and resized images from the green bounding boxes which are fed to the local branch.	17
3.4	The localization accuracy of different threshold of τ . Each sub-figure is the accuracy for different τ . And in each sub-figure, different color represents the threshold of IoU (T(IoU)) when measuring the accuracy of the predicted bounding box. Better view as zoomed.	23
3.5	Examples of heatmaps for "no finding" images. The cropped regions are denoted by green bounding boxes.	25

3.6	ROC curves of the global, local and fusion branches (DenseNet-121 as backbone) over the 14 pathologies. The corresponding AUC values are given in Table. 3.1. We observe that fusing global and local information yields clear improvement.	26
3.7	The visualized cropped regions and the lesion areas. The red bounding boxes are the ground truths of lesion areas, and the green bounding boxes are the cropped regions in AG-CNN.	27
3.8	ROC curves of AG-CNN on the 14 diseases (ResNet-50 and DenseNet-121 as backbones, respectively).	29
3.9	Average AUCs for different settings of τ on the test set (ResNet-50 as backbone). Note that the results from global branch are our baseline. . . .	31
3.10	Average AUC scores of AG-CNN with different settings of τ on the validation set (ResNet-50 as backbone).	32
4.1	Overview of the framework. There are two different attention mechanisms investigated in Section 4.2. Here, we take the first one <i>att1</i> as an example to illustrate the proposed framework. <i>CRAL</i> consists of two main modules. The feature embedding module is a CNN network which can be replaced by any network. In our experiment, we use ResNet-50 or Densenet-121 as the backbone. The normalized attention scores are obtained from the attention module. Attention scores contain C channels, and each channel corresponds to one category (highlighted with blue or red). By combining the channel-wise Hadamard product and element-wise sum operations, the high-level features and the attention scores are integrated into a residual attention block to classify the input image. Each class/disease is classified by a binary classifier in our model. “Pooling” represents a global average pooling layer. “FC” and “BCE” represent the fully connected layer and the binary cross entropy loss function, respectively.	37

4.2	Architecture of residual attention module (with <i>att1</i> and <i>att2</i>). <i>att1</i> consists of two 3×3 convolutional layers followed by ReLU, one 1×1 convolutional layer and one non-linear activation layer (Sigmoid). For <i>att2</i> , the input CNN features F are fed into the “hourglass” attention branch and a convolutional branch, respectively. Through the channel-wise Hadamard product and element-wise sum operations, a residual feature is formed by the learned features \tilde{F} and its weighted version $A \odot \tilde{F}$	39
4.3	Example images and the corresponding labels in the ChestX-ray14 dataset. Each image is labeled with one or more pathologies.	43
4.4	ROC curves of four combinations of CNN backbones and attention mechanisms (ResNet-50-att1, ResNet-50-att2, DenseNet-121-att1, and DenseNet-121-att2) over the 14 pathologies. The corresponding AUC scores are given in Table. 4.1.	45
4.5	Examples of heatmaps generated from the learned features (from ResNet-50). The ground truth bounding boxes provided by (Wang et al., 2017b) are annotated on the original images. Note that the heatmaps are zoomed to the same size as the input images, and the heatmaps may be a few difference due to the usage of random cropping in testing.	51
4.6	Examples of classification results. We present the top-8 predicted categories and the corresponding probability scores. The ground truth labels are highlighted in red or blue.	51
5.1	Examples of lesion areas on the ChestX-ray14 dataset. The first row presents some chest X-ray images with lesion areas, which are small compared to the global ones. The second row shows multiple pathologies existing in an image, which means the corresponding patient suffers from various diseases in a period. The disease existing in each bounding box corresponds to the pathology name with same color in the middle row.	54

5.2	Overview of the proposed <i>ConsultNet</i> . The <i>ConsultNet</i> consists of an Encoder, a Feature Selector, a Feature Integrator, and a Decoder. Given an image, we first feed it into the Encoder and obtain a mid-level feature representation. Then we learn the disease-specific and disease-correlated features by a VSIB based Feature Selector and an SCE based Feature Integrator, respectively. At last, both of them are concentrated together to classify the input image. Note that the “Conv”, “VIB”, “VSIB”, “SCE”, “GMP” and “FC” represent the convolutional layer, variational information bottleneck, spatial-channel encoding, global max pooling layer and fully connected layer respectively.	58
5.3	The architectures of <i>S_module</i> . (a) and (b) represent the S_s and S_c submodules, respectively.	61
5.4	Visualized heatmaps generated by VIB and VSIB. (a) is the input image with the lesion area bounding box annotated by (Wang et al., 2017b). (b) and (c) are the heatmaps generated by the VIB and VSIB constraint, respectively. The large/small response trends to be red/blue in the heatmaps. The larger responses that locate at the position of the corresponding bounding box would be expected.	63
5.5	Examples in the ChestX-ray14 dataset. The second row shows some cases captured with abnormal conditions, which introduce noises at the edges of images.	67

- 5.6 Examples of heatmaps generated from VSIB (the first row) and SCE (the second row). (a) The VIB and SCE collaboratively focus on different lesion areas. (b) The VSIB module learns much more accurate lesion positions while SCE misjudges some healthy tissues. (c) The VSIB module misses recognizing the disease existing in the image while the SCE module complements the VSIB module in the first column. While in the second column, VSIB module successes to localize the accurate disease region, but the region localized by SCE module drifts out of the truly lesion area. The different color of bounding box presents the different pathology existing in the image. Same as Fig. 5.4, the position of bounding box is the ground truth provided in (Wang et al., 2017b). The larger responses locate at the position of the corresponding bounding box are expected. 79