2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 978-1-6654-6819-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/BIBM55620.2022.9995346

Learning Group-Disentangled Representation for Interpretable Thoracic Pathologic Prediction

Hao Li[1,2], Yirui Wu[1,2,3,*], Hexuan Hu[1,2], Hu Lu[4], Yong Lai[3], Shaohua Wan[5]
[1]Key Laboratory of Water Big Data Technology of Ministry of Water Resources, Hohai University
[2]College of Computer and Information, Hohai University
[3]Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University
[4]School of Computer Science and Communication Engineering, Jiangsu University
[5]Shenzhen Institute for Advanced Study, University of Electronic Science and Technology of China
Email:{lihao1998h@163.com, wuyirui@hhu.edu.cn, hexuan_hu@hhu.edu.cn,luhu@ujs.edu.cn, laiy@jlu.edu.cn, shaohua.wan@uestc.edu.cn}

***Abstract*—Deep learning methods have shown significant performance in medical image analysis tasks. However, they generally act like "black box" without explanations in both feature extraction and decision processes, leading to lack of clinical insights and high risk assessments. To aid deep learning in envisioning diseases with visual clues, we propose Representation Group-Disentangling Network (RGD-Net), which can completely disentangle feature space of input X-ray images into several independent feature groups, each corresponding to a specific disease. Taking several semantically related and labeled X-ray images as input, RGD-Net firstly extracts completely group-disentangled representations of diseases through Group-Disentangle Module, which applies group-swap and linking operations to construct latent space by enforcing semantic consistency of attributes. To prevent learning degenerate representations defined as shortcut problem, we further introduce adversarial constricts on mapping from features to diseases, thus avoiding model collapse with former free-form disentanglement. Experiments on chestxray-14 and ChestXpert datasets demonstrate that RGD-Net are effective in predicting diseases with remarkable advantages, which leverage potential factors contributing to different diseases, thus enhancing interpretability in working patterns of deep learning methods.**

***Index Terms*—Interpretable Deep Learning, Group-Disentangled Representation Learning, Thoracic Pathologic Prediction, Adversarial Constricts**

I. Introduction

Despite deep learning methods have achieved remarkable progress in medical image analysis [1], [2], most methods work as mappings from input factors to output classification results without explicit explanations. Most attempts [3]–[5] to explain deep learning focus on 'post-hoc' analysis by proving the importance of low-level visual features in producing accurate predictions. However, they couldn't directly link low-level visual features with high-level semantical diseases, and visually explain the decision making process.

As an alternative way, interpretable deep learning [6] considers the inherent requirement of interpretation to embed clues based explanations in their neural network design. Most of them built their framework on variational auto-encoder (VAE), which achieve significant process towards explainable deep learning by performing linking and explaining steps with help of visual clues represented as feature groups. However, they generally ignore independence of learned clues, where they map visual samples onto a latent space that overlapped separates the information belonging to different attributes. Therefore, they only achieve partly disentangled effects with overlapping and coarse-grained low-level features, resulting in confused explanations and low accuracy classification results.

In this paper, we propose RGD-Net for interpretable thoracic pathologic prediction. We firstly achieve completely group-disentangled representations of diseases through the proposed Group-Disentangle Module. Such module is designed with group-swap and linking operations to leverage semantic links between input X-ray images and diseases, enforcing semantic consistency of attributes. To mitigate shortcut problem, we further propose adversarial constricts, which borrows the idea of GAN to retain informative features during iteratively updating via group-swap and linking operations. Such constricts guarantee the model to seek for global minimum by forcing nash equilibrium between free-form grouping and convinced diagnosis, thus preventing model collapse.

To sum up, our contributions are as follows:

- We propose *Representation Group-Disentangling Network* (RGD-Net), which completely extracts group-disentangled disease representations with fine-grained and non-overlapping features, thus promoting both interpretability and prediction accuracy.
- To resist shortcut problem caused by trapping in local minimum, an adversarial constraint is proposed to retain informative features during iteratively updating, thus forcing global minimum and avoiding model collapse.
- We experimentally demonstrate that RGD-Net can significantly improve classification accuracy, and showcase the potential of RGD-Net to disentangle information.

II. METHODOLOGY

A. Network Overview

As shown in Fig. 1, RGD-Net firstly takes a group of semantically-related X-ray images as inputs. Then,

* indicates Corresponding author

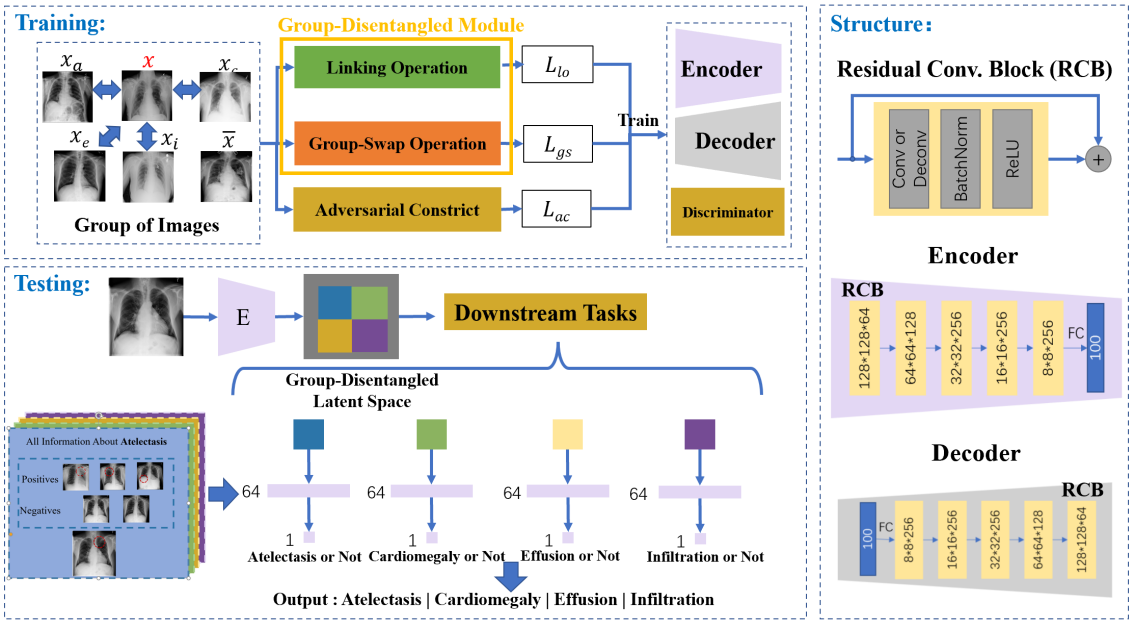

Fig. 1. The overall structure of RGD-Net, which extracts group-disentangled representations of disease through the Group-Disentangled Module and Adversarial Constrict. During testing, we use them to accurately predict corresponding disease labels.

it trains its encoder and decoder through the proposed Group-Disentangled Module (contains Linking Operation and Group-Swap Operation.) and Adversarial Constrict. Group-Disentangled Module enforces semantic consistency between disease concepts and the group-disentangled latent spaces. Adversarial Constrict builds on the idea of GAN by involving adversarial loss to solve the model collapse, that may encounter in the process of group-disentanglement and is generally defined as shortcut problem.

During training, we combine three kinds of losses as a total loss L:

$$L = \min_{\mathrm{D,E}} \max_{Dis} L_{lo} + \lambda_{gs} L_{gs} + \lambda_{ac} L_{ac}, \tag{1}$$

where L_{lo}, L_{gs} and L_{ac} refer to losses of linking operation, group-swap operation and adversarial constrict part respectively, and scalar coefficients λ_{gs}, λ_{ac} represent the importance factor of different loss terms.

After training, we demonstrate the effectiveness of RGD-Net to predict four categories of diseases based on chest X-ray images. Guided by the idea to apply on automatical medical application, we use a trained encoder to convert the input image into a group-disentangled latent space during testing phase. Afterwards, we predict thoracic pathologies disease concepts based on the new input X-ray images with an additional classification module with 3 layers of MLPs.

B. Group-Disentangled Module

To retain the information of images in the latent space, we propose an auto-encoder based Linking Operation, which links relationship between semantical concepts of disease and low-level visual features as shown in Fig. 2 (b). Specifically, for each input x, we embed data in a low-dimensional vector by the encoder. Then we link part of units of the vector to a specific disease concept. Finally, we input this latent vector into the decoder and calculate the reconstruction loss L_{lo} for each image.

To enforce semantic consistency of disease concepts, we propose the Group-Swap Operation (Fig. 2 (c)), which extracts features of disease concepts by leveraging semantic links between input image pairs. Taking an image pair sharing a disease as input, the Group-Swap Operation exchanges the corresponding part of the disease in pair's latent space, and expects to get same result as the input through the decoder.

The Group-Swap Operation is subject to the after-swap reconstruction loss:

$$L_{gs} = \sum_{concepts} (||D(z_s) - x||_2^2 + ||D(z_s^o) - x^o||_2^2), \tag{2}$$

where x^o is the paird image, z_s is the after-swap latent space, $\sum_{concepts}$ represents the sum of after-swap reconstruction loss for each concept in a group of images.

C. Adversarial Constrict

Ideally, if there exists sufficient sample pairs sharing no duplicate concepts, loss of group-swap operation L_{gs} will be zero, so that complete group-disentanglement being logically obtained. However, due to free-form group-swap operation in former group-disentangled module, shortcut problem can occasionally occur with local minimum trap, where RGD-Net may learn degenerate encodings that all information of input images are retained in the group of background features.

We propose an adversarial constrict to solve the shortcut problem. As shown in Fig. 2 (d), we take triplet images,

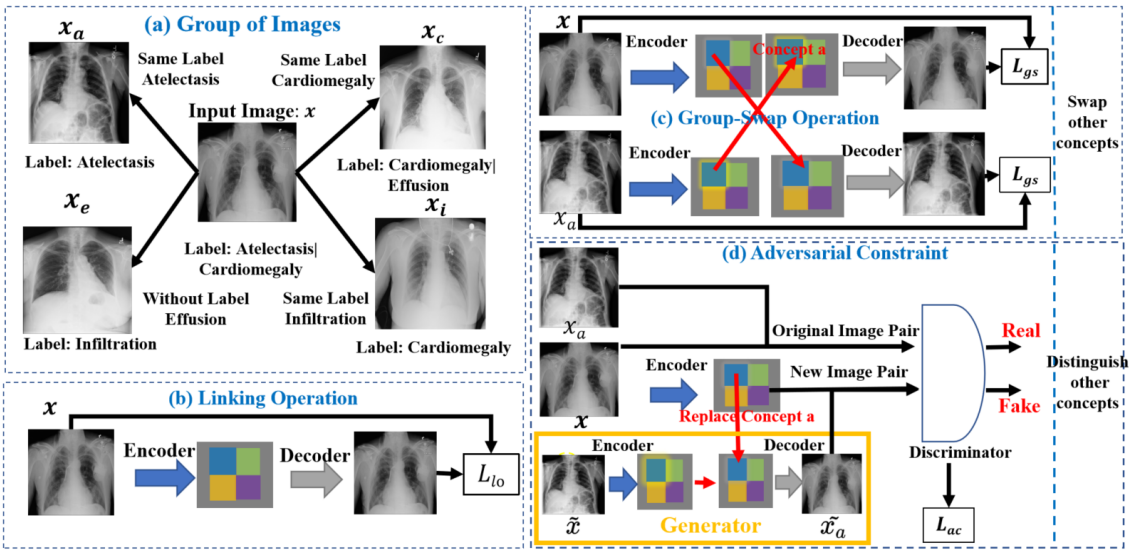

Fig. 2. Four steps in training RGD-Net to learn group-disentangled latent space: (a) **Group of Images**, where we input a group of semantically related images to learn their common properties; (b) **Linking Operation**, where we link relationship between semantical concepts of disease and low-level visual features, calculating self-reconstruction loss for each image; (c) **Group-Swap Operation**: where we swap part of the latent representations of their shared concepts to enforce semantic consistency of disease concepts. (d) **Adversarial Constraint** takes triplet images as input to solve the *shortcut problem* caused by trapping in local minimal.

i.e., x_a, x and $\tilde{x_a}$, as input and introduce an adversarial training style. Specifically, the generator uses encoder-decoder structure to replace one specific feature group (represented as concept a) from x to $\tilde{x}$, thus generating new image $\tilde{x_a}$. Meanwhile, the discriminator is designed as neural network to distinguish between original/real image pair $[x, x_a]$ and new/fake image pair $[x, \tilde{x_a}]$.

In Fig. 2 (d), we show an example in adversarial training style by swapping the first feature group. Similarly, we construct image triples with different disease concepts, and calculate total adversarial losses:

$$L_{ac} = \sum_{a=1,\ldots,5} log(Dis(x, x_a)) + log(1 - Dis(x, \bar{x_a})). \quad (3)$$

where the total number of disease concepts is 5 in our medical diagnosis application, and function $Dis()$ represents the discriminator to judge real or fake pair.

III. Experiments

A. Datasets

ChestXray-14 is a wildly used Chest abnormality detection dataset, which contains 112120 front X-ray image from 30805 individuals, including 14 categories of chest pathology. ChestXpert is a large chest x-ray dataset containing 224316 front X-ray image from 65240 patients, also including 14 categories of chest pathology.

We select a subset of ChestXray-14 for experiments, which contains 36764 training images and 7353 testing images with 4 pathology labels (Atelectasis, Cardiomegaly, Effusion and Infiltration). We also select a subset of ChestXpert, which contains 162188 training images and 32437 testing images with 3 pathology labels(Pleural Effusion, Edema and Cardiomegaly).

TABLE I
COMPARISON EXPERIMENTS ON CHESTXRAY-14 DATASET. FOR EACH PATHOLOGY, THE HIGHEST AUROC SCORES ARE BOLDED.

Methods	Atel	Card	Effu	Infi
RGD-Net (ours)	**0.8630**	0.8980	**0.9269**	**0.8653**
CheXNet [7]	0.8094	**0.9248**	0.8638	0.7345
Yao et al. [8]	0.7720	0.9040	0.8590	0.6950
Wang et al. [9]	0.7160	0.8070	0.7840	0.6090
ChestNet [10]	0.7433	0.8748	0.8114	0.6772
Li et al. [1]	0.8000	0.8700	0.8700	0.7000
Zhou et al. [11]	0.8121	0.9066	0.8786	0.7065

TABLE II
COMPARISON EXPERIMENTS ON CHESTXPERT DATASET. FOR EACH PATHOLOGY, THE HIGHEST AUROC SCORES ARE BOLDED.

Methods	Effu	Edema	Card
RGD-Net (ours)	0.900	0.9023	0.8871
Ye et al. [12]	0.9166	0.9436	0.8703
Pham et al. [13]	**0.9640**	**0.9580**	**0.910**
Irvin et al. [14]	0.9360	0.9280	0.8540

B. Accuracy of Thoracic Pathologic Prediction

We evaluate the performance of RGD-Net on thoracic pathologic prediction and compare it with other non-disentangled DL methods. Table. I shows that our RGD-Net, which has significantly improved on ChestXray-14 dataset by prediction with group-disentangled latent representation compared with the existing methods methods.

The AUROC values of RGD-Net on Atelectasis, Cardiomegaly, Effusion and Infiltration reached 86.30%, 89.80%, 92.69%, 86.53% respectively, being 5.36% , -2.68%, 6.31% and 13.08% higher than the second-highest achieved by

TABLE III

GROUP-DISENTANGLED REPRESENTATION ANALYSIS. WE USE THE ROW DISEASE FEATURES TO PREDICT THE AUROC OF COLUMN DISEASES ON THE TEST SET BY A SIMPLE 3-MLP. DIAGONALS ARE BOLDED AND '-' MEANS THAT ESTHER ET AL. [6] FAIL TO DISENTANGLE CONCEPT OF BACKGROUND.

	RGD-Net (ours) (completely disentangled)				Esther et al. [6] (partly disentangled)				AutoEncoder (without disentangled)			
Disease	Atel	Card	Effu	Infi	Atel	Card	Effu	Infi	Atel	Card	Effu	Infi
Atelectasis	**0.8630**	0.4855	0.5094	0.5005	**0.6136**	0.4960	0.4816	0.5050	**0.6076**	0.4990	0.4802	0.5297
Cardiomegaly	0.4822	**0.8980**	0.4836	0.5063	0.5062	**0.6610**	0.4968	0.4758	0.5067	**0.7048**	0.5183	0.4864
Effusion	0.4893	0.5061	**0.9269**	0.5229	0.5153	0.5038	**0.6688**	0.5099	0.4884	0.4985	**0.7444**	0.5292
Infiltration	0.4986	0.4900	0.4985	**0.8653**	0.4863	0.5230	0.5315	**0.5910**	0.4996	0.4955	0.4911	**0.6332**
Background	0.4983	0.5200	0.4926	0.4926	-	-	-	-	0.5045	0.5029	0.5087	0.4887

CheXNet.

To prove the performance of the proposed method on large medical datasets, we test the prediction performance of the proposed model on ChestXpert, one of the largest datasets currently available. As shown in Table. II, the accuracy of the proposed network is slightly lower on ChestXpert than the two latest networks, that is because our method considers not only the categories of predicted pathology, but also the interpretability of the network.

C. Group-Disentangled Representation Analysis

To prove the effect of group-disentanglement of our RGD-Net, we use the subspaces of disease concepts to predict four thoracic pathologies through a simple 3-MLP. If the hidden subspace contains all the information about the disease, the predicted result should be a matrix with 1 on the diagonal and 0.5 on the rest.

We use Esther et al. [6] and standard auto-encoder with classification head as comparison methods. The former partly disentangles the latent space, and the latter is not a disentangled method. Table. III shows that RGD-Net successfully decomposes the image into a group-disentangled latent space and uses each subspace to accurately predict the corresponding concept, but not to predict other concepts. The results of two comparison methods, whose latent space is not completely group-disentangled, show that each subspace doesn't know what it corresponds to, so their AUROCs are nearly 0.5.

IV. CONCLUSION

This paper proposes a Representation Group-Disentangling Network (RGD-Net), which completely extracts group-disentangled disease representations with fine-grained and non-overlapping features, thus promoting both interpretability and prediction accuracy. Further, we found the possible model collapse problem in the training process, and proposed an adversarial constrict to solve it. Finally, we experimentally demonstrate that RGD-Net can significantly improve classification accuracy compared with partly disentangled methods or other DL methods, and showcase the potential of RGD-Net to disentangle information.

V. ACKNOWLEDGMENTS

This work was supported in part by a grant from National Key R&D Program of China under Grant No. 2021YFB3900601, the Fundamental Research Funds for the Central Universities under Grant B220202074, the Fundamental Research Funds for the Central Universities, JLU, the Natural Science Foundation of China under Grant 61702160, and the National Natural Science Foundation of China under Grant 61976050.

REFERENCES

[1] Zhe Li, Chong Wang, Mei Han, Yuan Xue, Wei Wei, Li-Jia Li, and Li Fei-Fei, "Thoracic disease identification and localization with limited supervision," in *Proceedings of CVPR*, 2018, pp. 8290–8299.

[2] Qiran Kong, Yirui Wu, Chi Yuan, and Yongli Wang, "CT-CAD: context-aware transformers for end-to-end chest abnormality detection on x-rays," in *Proceedings of BIBM*, 2021, pp. 1385–1388.

[3] Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba, "Learning deep features for discriminative localization," in *Proceedings of CVPR*, 2016, pp. 2921–2929.

[4] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of ICCV*, 2017, pp. 618–626.

[5] Guangchen Shi, Yirui Wu, Jun Liu, Shaohua Wan, Wenhai Wang, and Tong Lu, ," .

[6] Puyol-Antón Esther, Chen Chen, and James R. Clough, "Interpretable deep models for cardiac resynchronisation therapy response prediction," in *Proceedings of MICCAI*, 2020, vol. 12261, pp. 284–293.

[7] Pranav Rajpurkar, Jeremy Irvin, et al., "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning," *CoRR*, vol. abs/1711.05225, 2017.

[8] Li Yao, Eric Poblenz, Dmitry Dagunts, Ben Covington, Devon Bernard, and Kevin Lyman, "Learning to diagnose from scratch by exploiting dependencies among labels," *CoRR*, vol. abs/1710.10501, 2017.

[9] Xiaosong Wang, Yifan Peng, et al., "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proceedings of CVPR*, 2017, pp. 3462–3471.

[10] Hongyu Wang and Yong Xia, "Chestnet: A deep neural network for classification of thoracic diseases on chest radiography," *CoRR*, vol. abs/1807.03058, 2018.

[11] Bo Zhou, Yuemeng Li, and Jiangcong Wang, "A weakly supervised adaptive densenet for classifying thoracic diseases and identifying abnormalities," *CoRR*, vol. abs/1807.01257, 2018.

[12] Wenwu Ye, Jin Yao, Hui Xue, and Yi Li, "Weakly supervised lesion localization with probabilistic-cam pooling," *CoRR*, vol. abs/2005.14480, 2020.

[13] Hieu H. Pham, Tung T. Le, Dat Q. Tran, Dat T. Ngo, and Ha Q. Nguyen, "Interpreting chest x-rays via cnns that exploit hierarchical disease dependencies and uncertainty labels," *Neurocomputing*, vol. 437, pp. 186–194, 2021.

[14] Jeremy Irvin, Pranav Rajpurkar, et al., "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proceedings of AAAI*, 2019, pp. 590–597.