Anomaly Detection in Chest X-Rays

Tarun Saxena and Kush Suryavanshi

Abstract

Classifying chest X-ray abnormalities is an important task in medical diagnosis, especially for respiratory and cardiovascular disorders. In this study, we present an innovative method for detecting anomalies in chest X-rays utilizing Dual-Distribution Discrepancy. This approach compares the distributions of normal and abnormal chest X-rays in a low-dimensional feature space, looking for disparities that may suggest the existence of abnormalities. We will evaluate our proposed method on the VinBigData Chest X-ray dataset and the RSNA Pneumonia Detection Challenge dataset. Anomaly classification can assist discover probable illnesses or problems at an early stage before they become serious. This can result in early treatment, improved results, and lower healthcare expenditures.

Keywords: Anomaly Detection, Chest X-ray , Deep Learning.

1 Introduction

Chest X-ray (CXR) is the most typical radiological exam for the diagnosis of various diseases. However, obtaining annotations for CXRs can be difficult and time-consuming, leading to the development of anomaly detection methods that require minimal or no annotations. With the increasing popularity of deep learning, automatic CXR analysis has become more common, aiming to improve diagnosis efficiency and alleviate the reading burden on radiologists. Unlike existing methods that consider anomaly detection as a one-class classification problem and only model the distribution of known normal images during training. In this project, we use a novel strategy, Dual-distribution Discrepancy for Anomaly Detection (DDAD), for detecting anomalies in chest X-rays (CXRs) in an unsupervised fashion. DDAD utilizes both known normal and unlabeled images to capture anomalous features from the latter. The proposed method consists of two modules, A and B, with A taking both known normal and unlabeled images as inputs and B modeling the distribution of only known normal images.

1.1 Related Work

In the past, the majority of methods for anomaly detection have treated it as a one-class classification problem [1], where only normal images are used for training, and any samples that don't match the normal profile are flagged as anomalies during testing. This approach has led to the development of various techniques that rely on either reconstruction [2] or self-supervised learning [3] for anomaly detection. The methods using reconstruction train reconstruction networks, which are similar to autoencoders, on normal images. The goal is to minimize the reconstruction error for these normal images. Unseen abnormal images, which are assumed to be unable to be reconstructed, will result in larger reconstruction errors. These Reconstruction-based methods have been shown to be effective in anomaly detection [4,5,6]. To prevent anomalies from being reconstructed and reduce the possibility of missed detections, some approaches [7,8] utilize Variational AE (VAE) [9] to estimate the normal distribution and iteratively restore

images to ensure that the output is free of anomalies, resulting in higher discrepancies with the abnormal input. Nevertheless, this iterative restoring process is computationally demanding and time-consuming. In recent times, self-supervised techniques [10,11] have been developed that manually generate defects to train models for detecting abnormalities.

In most existing methods for anomaly detection, either only normal images are used or synthetic abnormal images are added to the training set, which limits their discriminative capability due to the lack of training on real abnormal images.

We use the dual-distribution discrepancy method for anomaly detection approach that utilizes both known normal and unlabeled images to improve anomaly detection performance. The proposed approach consists of two modules, A and B, each of which is an ensemble of several reconstruction networks. During training, Module A takes both known normal and unlabeled images as inputs to capture anomalous features, while Module B models the distribution of only known normal images. The discrepancy between these two modules is used as an anomaly score. The proposed method achieves state-of-the-art performance on RSNA datasets [12] and Vin Big Data chest X-ray datasets [13].

2 Methodology

In this section, we present the dual discrepancy method, reconstruction networks which differ from previous approaches that only use normal images or synthetic abnormal images for anomaly detection.

Most one-class classification anomaly detection methods use an annotated normal image dataset (D_n) and an annotated test dataset (D_t) containing both normal and abnormal samples. In these methods, the model is trained on Dn and then it is tested on D_t to identify anomalies. In contrast, we also use the unlabeled dataset D_u including both normal and abnormal images to improve the performance of anomaly detection.

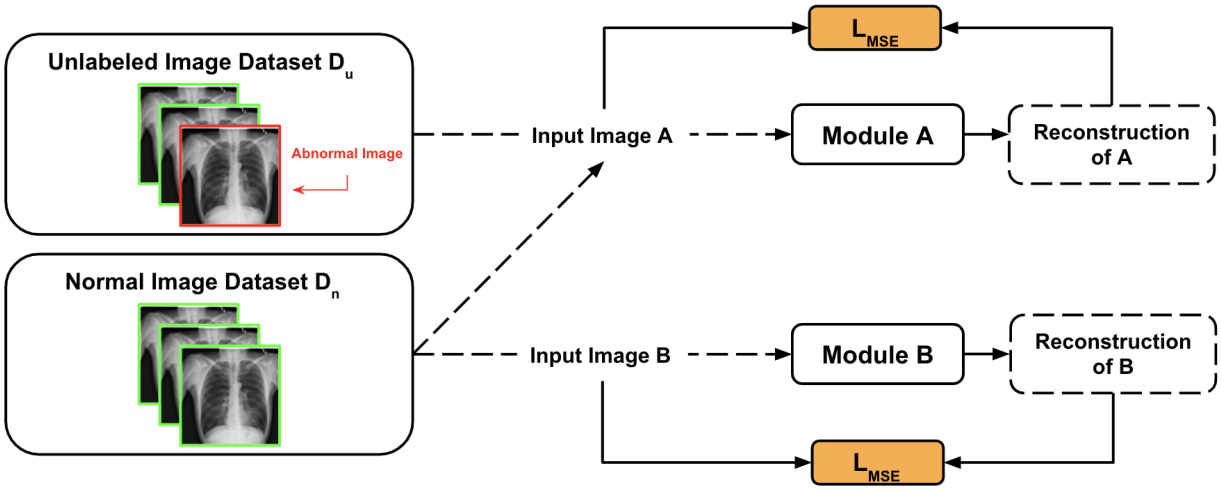

Fig. 1. *Illustration of the proposed method in training stage*

2.1 **Dual Distribution Modeling**

As shown in Fig. 1, we use two modules, denoted A and B, trained on different datasets to model the dual distribution. Module A is trained to capture anomalous features from unlabeled images while still being able to reconstruct normal images. On the other hand, Module B is trained only on the normal image dataset D_n and models the distribution of normal images.

Both modules consist of an ensemble of K reconstruction networks with the same architecture but different random initialization of parameters and random shuffling of training samples. Module A is trained on both normal image dataset D_n and unlabeled image dataset D_u by the Mean Squared Error (MSE) Loss to minimize reconstruction errors on the training set. Similarly, Module B also uses the same loss function but it is trained on normal image dataset D_n. This will lead Module A to capture effective features and model anomalies as a complement of the normal images which will show a higher difference score between Module A and Module B as Module B is only trained on normal images.

2.2 **Anomaly scores**

Given a test image x, we calculate the inter discrepancy by finding pixel-wise error in the average reconstructed vectors from Module A and Module B by,

$$\mathbf{A^p_{inter}} = | \boldsymbol{m_A^p} - \boldsymbol{m_B^p} |$$

here m_A^p , m_B^p are the average (averaged over K networks) reconstructed values of pixel p obtained from Module A and Module B. Whereas intra-discrepancy score is calculating the pixel-wise root mean (over K networks) squared error between m_B^p and x_{Bi}^p (reconstructed value of pixel p from Module B's i^{th} reconstruction network). The discrepancy maps can indicate potential abnormal regions based on pixel-wise anomaly scores. The image-level anomaly score is obtained by averaging the pixel-level scores in each image. These anomaly scores lead to better discriminative capabilities as they consider the difference between both the distributions modeled by Module A and Module B.

$$\boldsymbol{A^p_{intra}} = \sqrt{\frac{1}{K}\sum_{i=1}^{K}(m_B^p - x_{Bi}^p)^2}$$

3 Experiments

3.1 **Datasets**

We conducted experiments on two chest X-ray datasets, the RSNA Pneumonia Detection Challenge dataset, and the VinBigData Chest X-ray Abnormalities Detection dataset. The performance of our models was evaluated using the area under the ROC curve (AUC). The RSNA dataset consisted of 8,851 normal and 6,012 lung opacity images. For the experiments, we used 3,851 normal images from the RSNA dataset as the normal dataset Dn, 4,000 images with varying anomaly rates as the unlabeled dataset Du, and 1,000 normal and 1,000 lung opacity

images from the RSNA dataset as the test dataset Dt. The VinBigData dataset had 10,606 normal and 4,394 abnormal images, covering a total of 14 categories of anomalies. For the experiments, we used 4,000 normal images from the VinBigData dataset as Dn, 4,000 images as D_u, and 1,000 normal and 1,000 abnormal images from the VinBigData dataset as D_t.

3.2 **Implementation Details**

In our experiments, the autoencoder used consists of two main parts - an encoder and a decoder. The encoder has four convolutional layers with a kernel size of 4 and a stride of 2. The channel sizes of the layers are 16, 32, 64, and 64 respectively. The decoder, on the other hand, has four deconvolutional layers with the same kernel size and stride as the encoder. The channel sizes of these layers are 64, 32, 16, and 1 respectively. The encoder and decoder are connected by three fully connected layers. All layers except the output layer are followed by batch normalization and ReLU. In order to make a fair comparison with our proposed DDAD, we modify the MemAE [4] and AE-U [14] models based on the same Autoencoder (AE) architecture used in our approach. To ensure consistency, all input images in our experiments are resized to 64x64. The number of ensembles (K) is set to 3 for all models. Each model is trained for 250 epochs using the Adam optimizer with a learning rate of $5e^{-4}$

Table 1. Performance of different methods built on three backbones approaches

Method	**Anomaly Score**	**AUC**					
		RSNA			VinBigData		
		AE	MemAE	AE-U	AE	MemAE	AE-U
Reconstruction Baselines	*Arec*	0.669	0.68	0.867	0.559	0.558	0.738
Reconstructed Baselines (ensemble)		0.669	0.67	0.866	0.555	0.553	0.731
DDAD	*Aintra*	0.694	0.729	0.873	0.601	0.595	0.743
	Ainter	**0.8145**	**0.788**	**0.91**	**0.71**	**0.69**	**0.859**

To explore the advantages, we compared DDAD models built on different backbones with corresponding reconstruction baselines which use pixel-wise reconstruction error A_{rec}, The results are shown in Table 1 to further investigate the advantages. The AR (anomaly rate) of Du is set to 60% for all DDAD methods. The experimental results indicate that DDAD, which is based on AE, MemAE, and AE-U, can significantly outperform their corresponding baselines on all three datasets. DDAD-Ainter, in terms of AUC, exhibits a performance improvement of 14.6%, 10.8%, and 4.3% over the AE, MemAE, and AE-U baselines, respectively, on the RSNA dataset. On the VinBigData dataset, DDAD-Ainter shows a performance improvement of 15.1%, 13.2%, and 12.1% over the AE, MemAE, and AE-U baselines, respectively. DDAD-Aintra also shows improvement over the baselines, with gains of 2.5%, 4.9%, and 0.6% on the RSNA dataset and 4.2%, 3.7%, and 0.5% on the VinBigData dataset, for the AE, MemAE, and AE-U baselines, respectively. Additionally, the study evaluated the performance of a simple ensemble of K reconstruction models using the *Arec* anomaly score, which showed no significant improvement.

4 Conclusion

This project uses a method for anomaly detection in medical images, called dual-distribution discrepancy for anomaly detection, which makes use of both labeled and unlabeled images. We used two anomaly scores, intra-, and inter-discrepancy, are used to identify abnormalities. The experiments are conducted on two chest X-ray datasets, and the proposed method outperforms the reconstruction method, achieving state-of-the-art performance on the AE-U backbone. The study demonstrates that the proposed method can consistently improve the performance with the increasing of unlabeled dataset size, and it can also achieve superior performance even with no unlabeled data. Overall, this work provides a promising approach to anomaly detection.

References

1. Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In International conference on machine learning, pages 4393–4402.PMLR, 2018.
2. Christoph Baur, Stefan Denner, Benedikt Wiestler, Nassir Navab, and Shadi Albarqouni. Autoencoders for unsupervised anomaly segmentation in brain mr images: a comparative study. Medical Image Analysis, 69:101952, 2021.
3. Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. IEEE Transactions on pattern analysis and machine intelligence, 43(11):4037–4058, 2020.
4. Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In Proceedings of the IEEE/
5. Yifan Mao, Fei-Fei Xue, RuixuanWang, Jianguo Zhang,Wei-Shi Zheng, and Hongmei Liu. Abnormality detection in chest x-ray images using uncertainty prediction autoencoders. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 529–538. Springer, 2020.

6. Thomas Schlegl, Philipp Seeböck, Sebastian MWaldstein, Georg Langs, and Ursula Schmidt-Erfurth. f-logan: Fast unsupervised anomaly detection with generative adversarial networks. Medical image analysis, 54:30–44, 2019.
7. Xiaoran Chen, Nick Pawlowski, Ben Glocker, and Ender Konukoglu. Normative ascent with local Gaussians for unsupervised lesion detection. Medical Image Analysis, 74:102208, 2021.
8. Sergio Naval Marimont and Giacomo Tarroni. Anomaly detection through latent space restoration using vector quantized variational autoencoders. In 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), pages 1764–1767. IEEE, 2021.
9. Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. arXiv preprint arXiv:1312.6114, 2013
10. Jeremy Tan, Benjamin Hou, James Batten, Huaqi Qiu, and Bernhard Kainz. Detecting outliers with foreign patch interpolation. arXiv preprint arXiv:2011.04197, 2020.
11. Jeremy Tan, Benjamin Hou, Thomas Day, John Simpson, Daniel Rueckert, and
12. Bernhard Kainz. Detecting outliers with Poisson image interpolation. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 581–591. Springer, 2021.
13. https://www.kaggle.com/c/rsna-pneumonia-detection-challenge
14. https://www.kaggle.com/c/vinbigdata-chest-xray-abnormalities-detection
15. Yifan Mao, Fei-Fei Xue, RuixuanWang, Jianguo Zhang,Wei-Shi Zheng, and Hongmei Liu. Abnormality detection in chest x-ray images using uncertainty prediction autoencoders. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 529–538. Springer, 2020.