Towards Practical Unsupervised Anomaly Detection on Retinal Images

Khalil Ouardini[1,2], Huijuan Yang[2], Balagopal Unnikrishnan[2], Manon Romain[2,3], Camille Garcin[1,2], Houssam Zenati[1,2], J. Peter Campbell[4], Michael F. Chiang[4], Jayashree Kalpathy-Cramer[5], Vijay Chandrasekhar[2], Pavitra Krishnaswamy[2], and Chuan-Sheng Foo[2(✉)]

[1] CentraleSupelec, Gif-sur-Yvette, France
khalil.ouardini@student.ecp.fr
[2] Institute for Infocomm Research, A*STAR, Singapore, Singapore
{pavitrak,foo_chuan_sheng}@i2r.a-star.edu.sg
[3] École Polytechnique, Palaiseau, France
[4] Oregon Health & Science University, Portland, USA
[5] Massachusetts General Hospital, Harvard Medical School, Boston, USA

Abstract. Supervised deep learning approaches provide state-of-the-art performance on medical image classification tasks for disease screening. However, these methods require large labeled datasets that involve resource-intensive expert annotation. Further, disease screening applications have low prevalence of abnormal samples; this class imbalance makes the task more akin to anomaly detection. While the machine learning community has proposed unsupervised deep learning methods for anomaly detection, they have yet to be characterized on medical images where normal vs. anomaly distinctions may be more subtle and variable. In this work, we characterize existing unsupervised anomaly detection methods on retinal fundus images, and find that they require significant fine tuning and offer unsatisfactory performance. We thus propose an efficient and effective transfer-learning based approach for unsupervised anomaly detection. Our method employs a deep convolutional neural network trained on ImageNet as a feature extractor, and subsequently feeds the learned feature representations into an existing unsupervised anomaly detection method. We show that our approach significantly outperforms baselines on two natural image datasets and two retinal fundus image datasets, all with minimal fine-tuning. We further show the ability to leverage very small numbers of labelled anomalies to improve performance. Our work establishes a strong unsupervised baseline for image-based anomaly detection, alongside a flexible and scalable approach for screening applications.

Keywords: Unsupervised deep learning · Transfer learning · Anomaly detection · Retinal images

K. Ouardini and H. Yang—Equal contribution.
P. Krishnaswamy and C.-S. Foo—Equal contribution.
Supplementary Material: http://s000.tinyupload.com/?file_id=50006502228459557624.

Q. Wang et al. (Eds.): DART 2019/MIL3ID 2019, LNCS 11795, pp. 225–234, 2019.
https://doi.org/10.1007/978-3-030-33391-1_26

1 Introduction

Deep learning approaches offer state-of-the-art performance for a variety of medical image classification tasks. However, a major challenge in practical translation of these methods is that model training and/or fine-tuning requires thousands of images labelled by domain experts or clinical specialists. Such labelling is laborious, expensive, inefficient, and difficult to scale across diverse settings and applications. Moreover, expert raters can have discordant opinions [5,9], resulting in noisy or biased labels. Accordingly, there has been increasing interest in semi-supervised approaches for medical image classification [10,13], but less work on unsupervised learning.

For disease screening, normal samples usually have higher prevalence than abnormal samples. Thus, the classification task is akin to a rare anomaly detection task. We focus on unsupervised methods to detect anomalies for medical image-based screening. The machine learning community has developed many methods for unsupervised anomaly detection on natural image datasets like CIFAR-10 and SVHN [3,16,19,20]. However, the tasks of detecting anomalies on natural vs. medical images are distinct. Medical images exhibit greater variability due to the heterogeneity in abnormality presentation across patients or cohorts, and differences in acquisition devices or parameters. Further, anomalies in medical images tend to have finer resolution or more localized features. Yet, there has been limited focus on unsupervised anomaly detection methods for medical image datasets.

In this work, we characterize a range of unsupervised anomaly detection methods on natural image benchmarks (CIFAR-10, SVHN) and medical image datasets. For the latter, we employ fundus images obtained to screen for Diabetic Retinopathy (DR) and Retinopathy of Prematurity (ROP). We compare and contrast performance to find that existing methods have relatively unsatisfactory performance on medical image datasets. We further find that the unsupervised methods often require significant fine tuning and intensive computational resources, and therefore have limited practical applicability.

To overcome these challenges, we introduce a simple yet effective transfer learning method for unsupervised anomaly detection on medical images. Our method leverages the expressive representations learned by deep learning based classifiers trained on large image collections (like ImageNet). We extract features learned with these models and feed them into Isolation Forests [11], which offer efficient and robust anomaly detection for high-dimensional data with minimal tuning requirements. We perform extensive experiments and show that this approach outperforms baselines on both data types, with more significant gains on medical image datasets. We further show how to use a small collection of labeled anomalous samples, akin to a "validation set" to improve performance by selecting the best feature representation. As such, our work provides a strong baseline for unsupervised image-based anomaly detection, and a flexible and scalable approach for screening applications.

2 Methods

2.1 Task Definition

We consider two experimental settings. First, we assume the training data comprises only of normal images and focus on identifying images that fall out-of-distribution. We term this as Novelty Detection. Second, we relax this assumption, and consider the fully unsupervised scenario where the training set contains normal images alongside a small number of anomalies. We term this as Anomaly Detection. We now describe our method and the baselines.

2.2 Transfer Learning for Anomaly Detection

Figure 1 illustrates our method, which leverages the feature representations learned by networks trained on large, diverse image collections. The basic approach consists of (1) computing feature representations with a pre-trained network, and (2) training an anomaly detection algorithm on top of the computed representations. Implementing this general approach requires choosing (1) the pre-trained network, (2) how representations are derived (e.g., choice of layer), (3) an anomaly detection algorithm, and (4) tuning hyperparameters of the anomaly detection algorithm. We detail these choices below.

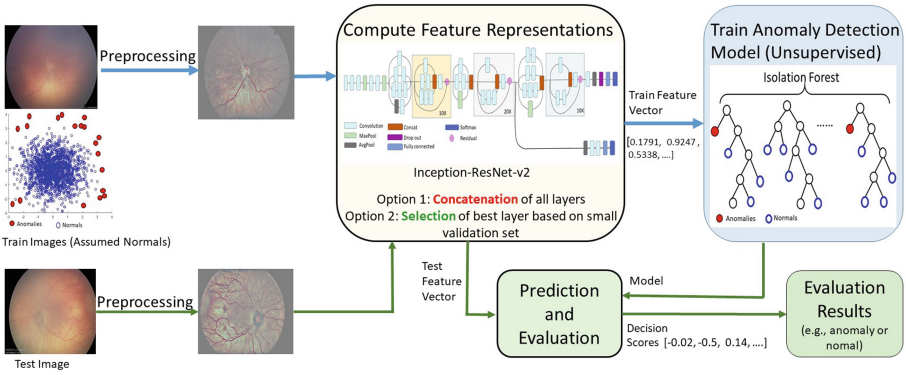

Fig. 1. Overview of proposed transfer learning-based anomaly detection method. Training images are assumed to come from a distribution of primarily normal images. We compute feature representations using a pre-trained deep learning model, and use the resulting feature vectors to train an unsupervised anomaly detection model. During testing, we transform images into feature space and use the trained model for anomaly detection.

Choice of Network: In our experiments, we used the Inception-ResNet-v2 network [18] trained on the ImageNet ILSVRC-2012-CLS dataset [15], as it is

one of the best performing networks for such tasks; our results were similar when using the Inception-v3 network (Supplementary Table 17).

Deriving Representations: We evaluated two strategies for deriving the representations.

1. *Computing a representation from all layers:* Previous work suggests that features from convolutional layers earlier in the network can contain very discriminative features [12]. To harness the power of these features, we derive representations including these earlier convolutional layers as well. For computational tractability, the outputs of these layers are first spatially averaged and then concatenated to the output of the other fully connected layers to produce a representation from the whole network.
2. *Picking the best representation using a validation set:* If annotated anomalies are available, they could be used to pick the best performing representation (from a single network layer/module) by evaluating model performance on a constructed validation set including these anomalies.

Anomaly Detection Algorithm: We chose the Isolation Forest method as it is fast, handles high-dimensional data well, does not require much tuning, and works well whether the training set consists only of normal data, or is mixed with some anomalies.

Hyperparmeters for Anomaly Detection Algorithm: We used the scikit-learn implementation of Isolation Forests with default parameters, in line with our goal of proposing a method requiring minimal fine-tuning.

Supplementary Sects. 8 and 9 describe the impact of the choices of network and feature representations. We note that the utility of transfer-learned representations has been demonstrated across a wide range of *supervised* computer vision tasks [2]. However, such approaches remain largely unexplored for unsupervised anomaly detection. To our knowledge, [1] is the only work exploring transfer learning for unsupervised anomaly detection. However, they focused solely on non-medical images in the novelty detection setting, and did not comprehensively benchmark against other competing methods. In contrast, we provide extensive comparisons to recent approaches on retinal fundus images and offer ways to select and improve feature representations for transfer.

2.3 Baselines

We evaluate a range of methods including shallow models (one-class SVM [17], Isolation Forest (IF) [11]), deep anomaly detection methods based on autoencoders (DAGMM [3], DSEBM [20]) and generative adversarial networks (AnoGAN [16]), as well as recently emerging unsupervised methods based on geometric transformations (DeepGEO, [6]) and SVDD based representations (DeepSVDD, [14]). For medical datasets, we include a supervised baseline: we finetune an Inception Resnet V2 [18] network that is initialized with ImageNet weights for "normal" vs. "abnormal" classification. Supplementary Sect. 1 details the baselines and associated hyperparameters.

3 Experiments

Here, we detail datasets with definition of the anomalies in each case, and provides evaluation results across the datasets and methods for the two settings.

3.1 Datasets

Figure 2 shows an overview of data types and illustrates example normal vs. anomalous images in the different datasets. Supplementary Fig. 1 provides more examples highlighting the variations. Supplementary Table 10 breaks down the statistics.

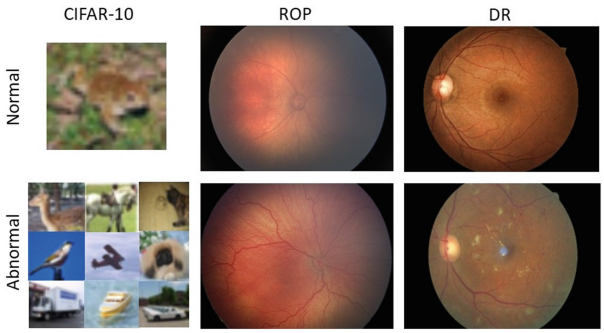

Fig. 2. Overview of datasets. The upper panel shows normal images while the lower panel shows abnormal images. These examples show the differences between natural and medical images, and highlight the nuanced nature of medical image anomaly detection.

Natural Image Datasets (SVHN, CIFAR-10): We used the official training and testing sets for SVHN and CIFAR-10. K denotes the number of classes in the dataset. Following previous works [6,14,19], we design K different experiments where samples from each label are alternately considered as "normal" and are used for training. We use 25% of the training set as a validation set and evaluate each model on the official test set containing anomalies at a ratio of $(K-1)/K$ (i.e, 90% for CIFAR-10 and SVHN). The only preprocessing was rescaling the images to [0, 1].

Retinopathy of Prematurity (ROP): ROP is an eye disease affecting premature babies, and is graded as "pre-plus" and "plus" based on the extent of retinal arterial tortuosity and venous dilation at the posterior pole [4]. As ROP is a leading cause of childhood blindness, there is a need for automated systems to regularly screen for "plus" disease, a key determinant for treatment. We obtained posterior pole retinal RGB photographs as part of the ongoing 'Imaging & Informatics in ROP' (i-ROP) cohort study. Each image was annotated

as "normal", "pre-plus" or "plus" by at least three independent experts, and a consensus reference standard label was assigned [4]. We squared cropped to cut the neutral background and resized images to 256 pixels, before subtracting the local average color to reduce differences in lighting. Pixel values were rescaled to [0,1]. Our experiments consider two scenarios: (1) "normal" vs. "plus" anomalies (total 4707 images, denoted as ROP by default), and (2) "normal" vs. "pre-plus" and "plus" anomalies (total 5511 images, denoted as ROP (All Grades)).

Diabetic Retinopathy (DR): DR is diagnosed based on the presence of microaneurysms, hemorrhages, hard exudates, microvascular abnormalities and neovascularization in retinal fundus photographs [7]. Due to the high prevalence of diabetes, there is a need to screen patients regularly. We obtained color retinal fundus photographs annotated with severity ratings from licensed clinicians as part of the Kaggle Diabetic Retinopathy challenge [8]. This is a large dataset from multiple sites with diverse patient demographics and varying acquisition conditions. It includes several poor quality images with over-exposed, out-of-focus and artefactual images, hence poses significant challenges for anomaly detection. For our experiments, we denoted images with severity rating of 0 (healthy) as normal and images with severity rating of 4 (advanced symptoms) as anomalous. We randomly sampled subsets of 3912 training and 7829 testing images from the official dataset, and preprocessed in the same way as for ROP.

3.2 Training and Evaluation

Except the DR dataset, we ran all experiments using five-fold cross-validation and quantify performance using the cross-validated area under the ROC curve (averaged across 5 seeds) and the corresponding standard deviation. AUROC is the common metric of choice for both anomaly detection papers [6,14,19] and medical applications [4,7]. We present the area under the precision-recall curve and the recall in Supplementary Tables 12–14.

3.3 Novelty Detection Setting

Results in the novelty detection setting are presented in Table 1. Our model outperforms all the baselines on CIFAR-10. We present results on a more challenging CIFAR-100 dataset in Supplementary Table 11. On SVHN, all methods perform only slightly better than random guessing, with a small advantage to DeepSVDD. The slightly lower performance for our method is likely due to a domain shift between the source (ImageNet) and target (SVHN) datasets. This is consistent with the fact that transfer learning performance can drop with dissimilarity between source and target datasets [2]).

On both the medical imaging datasets, our method outperforms every unsupervised baseline by a wide margin of around 20%. In all cases, however, the supervised classifier has better performance than the best unsupervised method. We repeated the ROP experiments on the more challenging setting that includes all grades of anomalies ("pre-plus" and "plus"). The results, in Supplementary Table 15, show similar trends.

Table 1. Area under the ROC curve in % with standard deviation in novelty detection setting. Results are averaged over the number of classes for natural images (see Sect. 3.1) and over 5 runs for medical images.

Natural images							
	IF	DAGMM	AnoGAN	DSEBM	DeepSVDD	DeepGEO	Ours
CIFAR-10	59.4 ± 11	57.5 ± 10	57.6 ± 12	58.8 ± 11	64.8	86.0	$\mathbf{88.2 \pm 6.6}$
SVHN	51.4 ± 0.9	51.8 ± 1.2	53.3 ± 3.1	57.1 ± 2.8	$\mathbf{57.3 \pm 3.3}$	–	55.4 ± 4.1
Medical images							
	IF	AnoGAN	DSEBM	DeepSVDD	DAGMM	Ours	Supervised
ROP	55.1 ± 5.0	49.5 ± 4.4	49.6 ± 3.9	57.5 ± 2.4	58.1 ± 6.2	$\mathbf{77.0 \pm 3.8}$	$\mathbf{97.3 \pm 2.0}$
DR	44.0 ± 0.5	44.2 ± 1.1	43.1 ± 0.2	46.4 ± 1.3	52.0 ± 0.1	$\mathbf{74.5 \pm 1.7}$	$\mathbf{94.5 \pm 2.7}$

3.4 Utilizing Small Numbers of Labeled Anomalies to Improve Performance

While it is difficult to curate large labeled datasets with sizeable numbers of anomalous samples for supervised learning, it is often feasible to obtain small numbers of labeled anomalies. We therefore explored whether it is possible to use such small "validation" sets to improve the choice of feature representation and anomaly detection performance. These experiments are done in the Novelty Detection setting. For the CIFAR-10 and ROP datasets, we compiled a small collection of N annotated anomalous samples, with N set as 3% of the total dataset size. We then evaluated representations from each of the blocks in the pre-trained network on this validation set, and chose the representation with best validation AUC. We employ these chosen representations to obtain evaluation results on the test set (Table 2). This strategy is especially useful for the medical image datasets (unlike for CIFAR-10 where gains are limited). In particular, we observed 6% AUC gain on the ROP dataset with just 4 annotated anomalies. Supplementary Sect. 10 provides further detailed results from individual blocks for varying sizes of the validation sets and for the complex ROP (AllGrades) task. Overall, these results suggest that our method could offer significant gains for medical domain end-users who are able to invest in limited resources to label a few examples.

Table 2. Averaged area under the ROC curve (over 5 runs) in % with standard deviation for different representations.

Representation	Concatenation	Best (picked) representation
CIFAR-10	$\mathbf{88.2 \pm 6.6}$	$\mathbf{88.2 \pm 6.6}$
ROP	77.0 ± 3.8	$\mathbf{82.6 \pm 6.5}$

3.5 Anomaly Detection Setting

We now consider how robust our method is to inclusion of varying proportions of anomalies in the training data. These evaluations correspond to the fully unsupervised Anomaly Detection setting. Table 3 illustrates the robustness of our method to varying numbers of anomalies mixed in to the training set for the CIFAR-10 and ROP datasets. We see that on CIFAR-10, test AUC decreases gradually as the proportion of anomalous samples in the training set increases. As CIFAR-10 has a high 90% proportion of anomalies, it provides an opportunity to understand how the performance of our method changes with varying anomaly proportions. For ROP, we expanded the training set to include up to 3% anomalies, to mimic the prevalence of disease in screening applications. Our results show that the performance is robust to inclusion of anomalous samples. We include evaluation against other baselines in Supplementary Table 16, and show that our method exhibits robust performance gains over competing methods even in the fully unsupervised setting.

Table 3. Area under the ROC curve (over 5 runs) for different anomaly ratios ρ in training set

ρ	1%	2%	3%	5%	10%	15%	20%	25%
CIFAR10	–	–	–	**86.8**	85.3	84.2	82.8	81.1
ROP	75.4	**76.9**	75.5	–	–	–	–	

4 Discussion and Conclusion

In this work, we characterized a range of unsupervised anomaly detection methods from the machine learning literature on medical images, and proposed a simple, efficient and effective transfer learning method to overcome prevailing limitations in this area. Our proposed method significantly outperforms competing methods on two computer vision benchmarks and two medical imaging datasets. Importantly, our method is flexible, and can effectively leverage very small numbers of labelled anomalies to improve performance. While our work offers a step towards closing the performance gap between unsupervised and supervised anomaly detection methods, we recognize the need for further performance improvements before they become suitable for clinical use. We anticipate that the first applications could lie in processes for more efficient labeling before diagnostic decision support applications can take shape.[1]

Acknowledgement. This project was supported by funding from the Deep Learning 2.0 program at the Institute for Infocomm Research (I2R), A*STAR, Singapore; and

[1] Link to code: https://github.com/khalilouardini/towards-practical-unsupervised-AD.

partially supported by SERC Strategic Funding (A1718g0045) research grants from the US National Institutes of Health (NIH grants R01EY19474, P30EY010572, and K12EY027720) and the US National Science Foundation (NSF grants SCH-1622679 and SCH-1622542); unrestricted departmental funding from the Oregon Health Sciences University, and a Career Development Award from Research to Prevent Blindness (New York, NY). We acknowledge helpful discussions with James M. Brown and Ken Chang (MGH) on datasets and experiment planning.

References

1. Andrews, J.T.A., Tanay, T., Morton, E.J., Griffin, L.D.: Transfer representation-learning for anomaly detection. In: ICML Anomaly Detection Workshop (2016)
2. Azizpour, H., Razavian, A.S., Sullivan, J., Maki, A., Carlsson, S.: Factors of transferability for a generic convnet representation. IEEE Trans. Pattern Anal. Mach. Intell. **38**(9), 1790–1802 (2016)
3. Bo, Z., et al.: Deep autoencoding Gaussian mixture model for unsupervised anomaly detection. In: International Conference on Learning Representations (2018)
4. Brown, J.M., et al.: Automated diagnosis of plus disease in retinopathy of prematurity using deep convolutional neural networks. JAMA Ophthalmol. **136**(7), 803–810 (2018)
5. Campbell, J.P., Kalpathy-Cramer, J., Dulanto-Reinoso, C.M., Montero-Mendoza, C., et al.: Plus disease in retinopathy of prematurity: a continuous spectrum of vascular abnormality as a basis of diagnostic variability. Ophthalmology **123**(11), 2338–2344 (2016)
6. Golan, I., El-Yaniv, R.: Deep anomaly detection using geometric transformations. In: Advances in Neural Information Processing Systems 31, pp. 9758–9769 (2018)
7. Gulshan, V., Peng, L., Mega, J.L., Webster, D.R., et al.: Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA **316**(22), 2402–2410 (2016)
8. Kaggle: Diabetic Retinopathy Detection (2015)
9. Krause, J., et al.: Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. Ophthalmology **125**(8), 1264–1272 (2018)
10. Lecouat, B., Chang, K., Kalpathy-Cramer, J., Krishnaswamy, P., et al.: Semi-supervised deep learning for abnormality classification in retinal images. CoRR abs/1812.07832 (2018)
11. Liu, F.T., Ting, K.M., Zhou, Z.H.: Isolation forest. In: Proceedings of the 2008 Eighth IEEE International Conference on Data Mining, pp. 413–422 (2008)
12. Liu, L., Shen, C., van den Hengel, A.: The treasure beneath convolutional layers: cross-convolutional-layer pooling for image classification. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4749–4757, June 2015
13. Madani, A., Ong, J.R., Tibrewal, A., Mofrad, M.R.K.: Deep echocardiography: data-efficient supervised and semi-supervised deep learning towards automated diagnosis of cardiac disease. npj Digit. Med. **1**(1), 59 (2018)
14. Ruff, L., Vandermeulen, R., Müller, E., Kloft, M., et al.: Deep one-class classification. In: Proceedings of the 35th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 80, pp. 4393–4402 (2018)
15. Russakovsky, O., et al.: Imagenet large scale visual recognition challenge. CoRR abs/1409.0575 (2014)

16. Schlegl, T., Seeböck, P., Waldstein, S.M., Schmidt-Erfurth, U., Langs, G.: Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In: Niethammer, M., et al. (eds.) IPMI 2017. LNCS, vol. 10265, pp. 146–157. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-59050-9_12
17. Schölkopf, B., Williamson, R., Smola, A., Shawe-Taylor, J., Platt, J.: Support vector method for novelty detection. In: Proceedings of the 12th International Conference on Neural Information Processing Systems, pp. 582–588 (1999)
18. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.: Inception-v4, inception-resnet and the impact of residual connections on learning. CoRR abs/1602.07261 (2016)
19. Zenati, H., Romain, M., Foo, C., Lecouat, B., Chandrasekhar, V.: Adversarially learned anomaly detection. In: 2018 IEEE International Conference on Data Mining (ICDM), pp. 727–736 (2018)
20. Zhai, S., Cheng, Y., Lu, W., Zhang, Z.: Deep structured energy based models for anomaly detection. In: International Conference on Machine Learning, pp. 1100–1109 (2016)