MethodsX 13 (2024) 102959

Contents lists available at ScienceDirect

MethodsX

journal homepage: www.elsevier.com/locate/methodsx

Explainable rotation-invariant self-supervised representation learning ☆

Devansh Singh [a], Aboli Marathe [b], Sidharth Roy [c], Rahee Walambe [a,*], Ketan Kotecha [a]

[a] Symbiosis Centre for Applied Artificial Intelligence, Symbiosis Institute of Technology, India
[b] Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA, USA
[c] Department of Computer and Information Science, University of Pennsylvania, USA

ARTICLE INFO

Method name:
RISC - rotation invariant self-supervised vision framework

Keywords:
Computer vision
Self-supervised learning
Robustness
Medical imaging data
Rotation invariance

ABSTRACT

This paper describes a method that can perform robust detection and classification in out-of-distribution rotated images in the medical domain. In real-world medical imaging tools, noise due to the rotation of the body part is frequently observed. This noise reduces the accuracy of AI-based classification and prediction models. Hence, it is important to develop models which are rotation invariant. To that end, the proposed method - RISC (rotation invariant self-supervised vision framework) addresses this issue of rotational corruption. We present state-of-the-art rotation-invariant classification results and provide explainability for the performance in the domain. The evaluation of the proposed method is carried out on real-world adversarial examples in Medical Imagery-OrganAMNIST, RetinaMNIST and PneumoniaMNIST. It is observed that RISC outperforms the rotation-affected benchmark methods by obtaining 22\%, 17\% and 2\% accuracy boost on OrganAMNIST, PneumoniaMNIST and RetinaMNIST rotated baselines respectively. Further, explainability results are demonstrated.

This methods paper describes:

- a representation learning approach that can perform robust detection and classification in out-of-distribution rotated images in the medical domain.
- It presents a method that incorporates self-supervised rotation invariance for correcting rotational corruptions.
- GradCAM-based explainability for the rotational SSL pretext task and the downstream classification outcomes for the three benchmark datasets are presented

Specifications table

Subject area:	Computer Science
More specific subject area:	Computer Vision, Medical Data, Self - Supervised learning
Name of your method:	RISC - rotation invariant self-supervised vision framework
Name and reference of original method:	None
Resource availability:	Datasets: https://medmnist.com/. The code will be made available upon request

☆ **Related research article:** None

* Corresponding author.
E-mail address: rahee.walambe@sitpune.edu.in (R. Walambe).

https://doi.org/10.1016/j.mex.2024.102959
Received 27 June 2024; Accepted 11 September 2024
Available online 14 September 2024

Background

Imagery corrupted by adversarial attacks or errant data collection arises commonly in medical [1,2] settings. As we brace frameworks for achieving robustness in the real-world deployment scenarios leveraging self-supervised methods can be fruitful as seen in prior art [3–8]. CNN-based methods are usually trained on well-balanced good-quality data captured in a controlled environment. Such models do not generalize well to rotated images and their performance degrades during testing time if the testing data contains rotated images. In the medical domain, rotated medical images pose a significant challenge and are common due to the movement of the patient during the imaging procedure. Although studies on rotation-invariance and more efficient object detection have been pursued [9–11], a singular approach to tackle rotational-invariant image classification in the real world with limited data has not been satisfactorily achieved. In this paper, a rotation invariant self-supervised learning (SSL) representation for image classification framework is presented for solving the aforementioned problems. We call this framework RISC. The main contributions include:

1. Proposal and demonstration of a novel rotation invariant self-supervised vision framework (RISC) for robust image classification in standard medical datasets.
2. Visual explanation of classification models applied to medical image datasets for an insight into algorithmic decision-making in the healthcare domain using GradCAM [12] for both the Rotation-Invariant SSL pre-text task and the downstream classification task.

This is a generic approach and can be extended to a multitude of real-world applications where rotations cause a reduction in the performance of the trained model. The paper is organized in four sections. Section 2 presents the methodology, followed by Section 3 which discusses the results of the RISC framework. The paper concludes in Section 4.

Method details

Datasets

As shown in Fig. 1, the MedMNIST datasets [13,14] are standard and diverse sets of labelled medical images where standard train-validation-test splits and baseline performance on various frameworks are provided. It is one of the most standardised medical image datasets to evaluate AI models in its domain.

RetinaMNIST [15] is a dataset of 1600 (3D) retina fundus images. The task is a 5-level grading of diabetic retinopathy severity. The PneumoniaMNIST [16,17] contains 5856 (2D) pediatric chest X-Ray images. It has two classes - pneumonia and normal. The OrganAMNIST [18,19] is based on 3D computed tomography (CT) images. It contains 58,850 (2D) images. Each image corresponds to one of 11 organs namely bladder, femur-left, femur-right, heart, kidney-left, kidney-right, liver, lung-left, lung-right, pancreas and spleen. All 3-D and 2-D images were originally of shape $28 \times 28 \times 3$ and 28×28 respectively and were later reshaped according to the model input shapes.

Self-supervised learning - RotNet

The RotNet is a ConvNet proposed by Gidaris et al. [20] which predicts the orientation of images in a self-supervised manner. A random 2D rotation transformation of 0, 90, 180, or 270° is applied to an image, and the ConvNet is trained to estimate the 2D rotation applied to obtain the rotated image. In order to succeed at the task, the model is forced to identify salient features in the image along with their orientations. The generic pipeline for the RotNet [21] model is shown below in Fig. 2. The learned representations can then be fine-tuned for other downstream tasks including object detection and image classification. In the RISC framework, the RotNet model with a ResNet50 backbone to correct for rotational corruptions is employed to improve the robustness of the system.

RISC SSL framework

The self-supervised learning (SSL) paradigm, pivotal in this research methodology, relies on the inherent similarity between images of the dataset, facilitated by the RotNet [21] for predictive modelling of rotation transformations within input images. The model subsequently assigns a label to denote the inferred rotation. This is the pretext task of learning rotation representations and serves as the preliminary step in the overall research objective. Subsequently, in the downstream task of classifying the images, the classifier leverages knowledge acquired by the RotNet to enhance its classification accuracy.

The self-supervised learning paradigm as shown in Fig. 3 for image classification has been implemented in the following stages:

Stage 1: Self-Supervised Pre-training: Utilizing the dataset itself, rotated images and corresponding rotation labels are generated which are subsequently used to train the RotNet. Within this stage, the RotNet of the self-supervised RISC framework learns feature representations corresponding to rotations in the images. This knowledge helps us to generate pseudo-rotation labels for input image which are used to fix the orientation of images before classification. The RotNet learns to effectively predict the correct rotation labels using the cross-entropy loss function as shown in Eq. (1):

$$\mathcal{L}_{CE} = -\sum_{i=1}^{n} y_i \log\left(p_i\right) \tag{1}$$

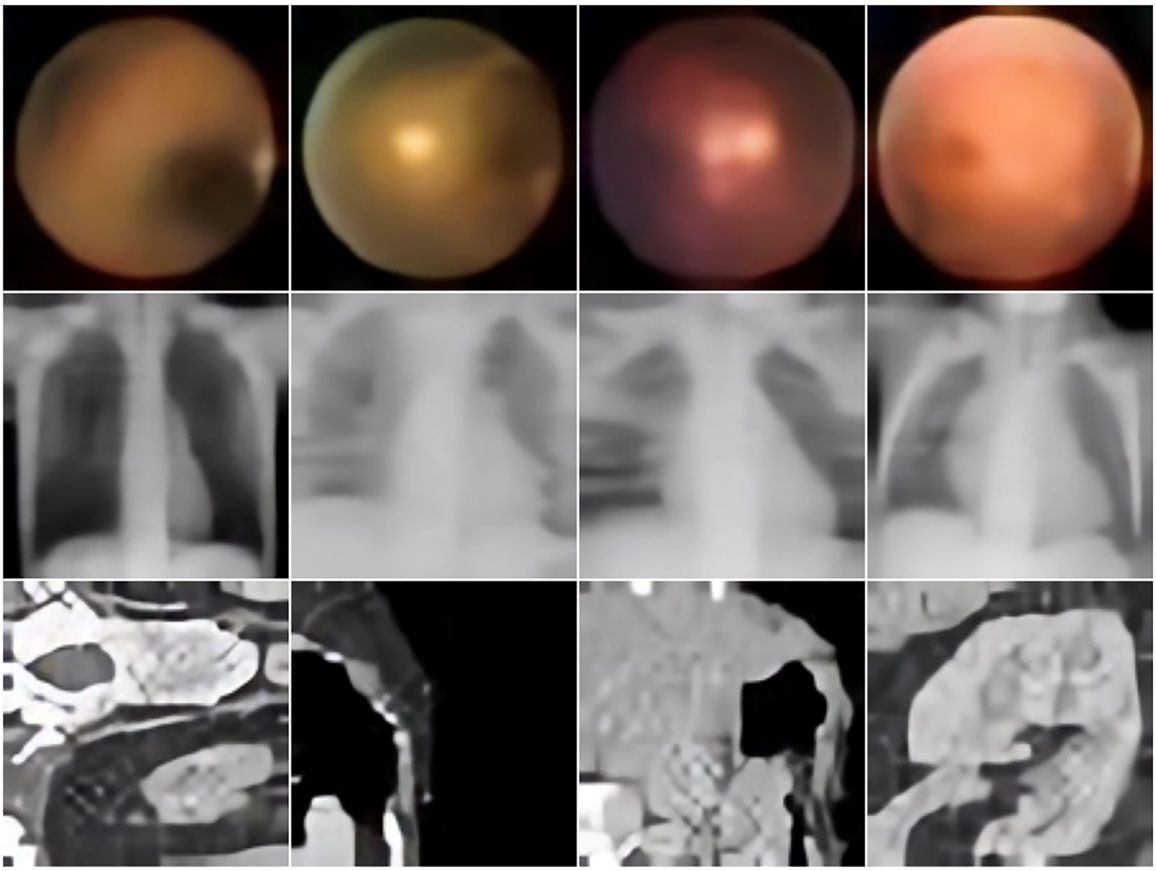

Fig. 1. MedMNIST samples for RetinaMNIST, PneumoniaMNIST and OrganAMNIST sub-datasets [13]. (Images upscaled to 200×200 for better visibility).

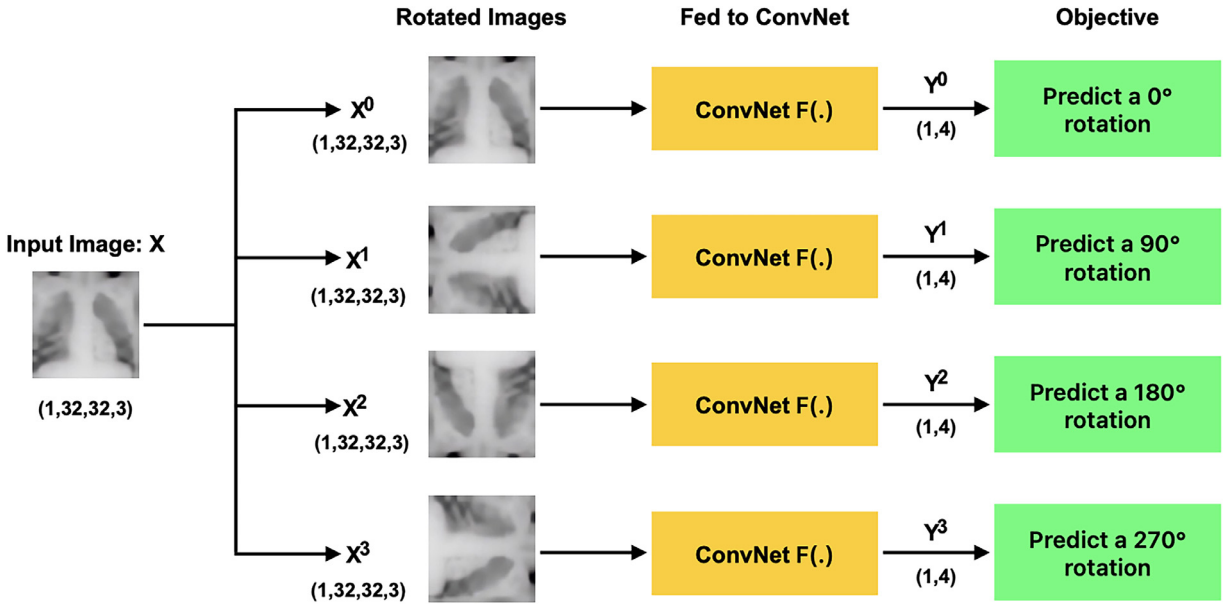

Fig. 2. RotNet Working Diagra. The ConvNet is a modified ResNet50 [22] having 3 fully- connected layers at the end instead of the usual top. Its input shape is (1,32,32,32) while the output is a one-hot encoded vector of shape (1,4). The first dimension represents the batch-size. The output vector corresponds to a predicted angle of rotation.

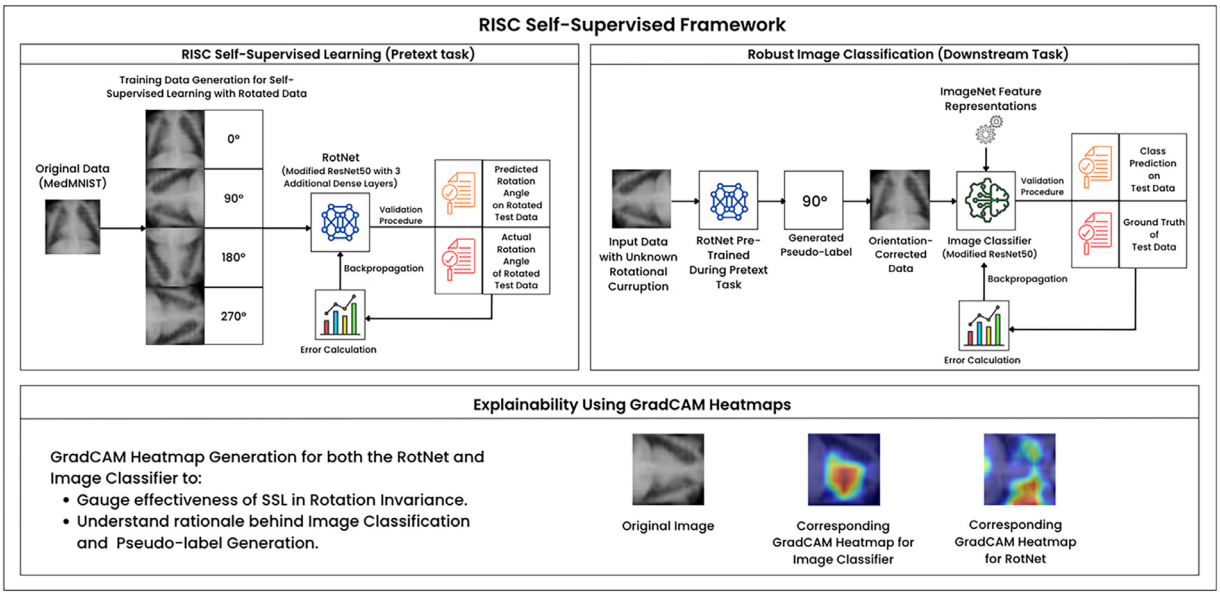

Fig. 3. RISC architecture diagram. (a sample image from the MedMNIST dataset has been used to demonstrate various stages of the RISC Framework).

Where $\mathcal{L}_{CE}$ represents the cross-entropy loss, n is the number of different rotation values, y_i is the true rotation value (binary vector with 1 for the correct class and 0 for others) which is the artificial rotation applied to the image before forward propagation through RotNet as shown in Eq. (2)., and p_i is the predicted probability of the image having rotation label i according to RotNet.

$$y_i = \begin{cases} 1, & if\ image\ has\ no\ rotation\ label\ i \\ 0, & otherwise \end{cases} \tag{2}$$

Stage 2: Classifier Training: A ResNet50 [22] pre-trained on ImageNet [23] with an input shape of $32 \times 32 \times 3$ for all datasets is used to classify the images. Since OrganAMNIST and PneumoniaMNIST have images of shape 32×32, their single channel is stacked in order to match the input shape of the classifier. The classifier is then trained on correctly oriented images along with ground truth labels sourced from the original datasets. This constitutes the downstream task of the RISC framework.

Stage 3: Evaluation of Classification Model: The classification model is evaluated on the test set of the original datasets to assess its performance for correctly- oriented images.

Stage 4: Application of Random Rotations: Random rotations (multiples of 90° for PneumoniaMNIST and OrganAMNIST and multiples of 45° for RetinaMNIST) are applied to the images within the test set, subsequently evaluating the classification model's robustness on these rotated images, using the original class labels.

Stage 5: Evaluation of SSL Framework: The RISC SSL framework, comprising RotNet and the classification model is implemented and tested, on randomly rotated images. In contrast to Stage 4, this time the system leverages the pseudo-label generation and orientation correction capabilities inherent in self-supervised learning. Importantly, rotation labels are deliberately withheld from the models during Stages 4 and 5.

Generating explanations using GradCAM

Generating explanations of the models' outcome is specifically important in domains such as healthcare to increase trustworthiness and reliability. This also allows the responsible and ethical implementation of AI models. GradCAM (gradient-weighted Class Activation Mapping) [12] is employed for visual explainability and to explain the decision-making process of the models. Gradients were computed with respect to the outputs of various convolution layers of the model, generating heatmaps that depict the critical regions that contributed the most to the prediction. These heatmaps were overlaid onto the original images to visually highlight the relevant parts of the images. This process provides a deeper look into the neural net- work's decision rationale and enhances its transparency. Notably, a different choice of the convolutional layer produces different heatmaps which vary across datasets too. This research experiments with multiple layers to find out the most suitable ones which help to better explain the model's working. Open-source libraries were utilized for this process and the explanations were generated by following the same procedure for both the RotNet predictions and the image classifiers of the downstream task.

Method validation

RISC SSL framework performance

The RISC framework aims to account for the calibration of medical instruments and accommodate the rotational artefacts while a classification task is carried out on medical images. The SSL model is able to capture the inherent properties of the images for rotational correction as seen in Table 1 which illustrates the performance of the RISC framework and its components. The dataset sizes, comprising training, validation, and test sets, along with the corresponding number of classes, are detailed. Rotated images and their corresponding rotation labels were generated from each original dataset itself, facilitating the RotNet training to learn the features indicative of the correct rotation. It was trained and tested on each of the three MedMNIST datasets independently and the resulting accuracies, delineated in the 'RotNet Accuracy' row, show the model's proficiency in rotational corrections across a diverse set of medical image datasets. The RotNet shows the most superior performance on the PneumoniaMNIST dataset with a 99% accuracy in predicting rotations on unseen data.

This research trained image classifiers on the original data with similar architectures as used in the benchmark classifiers of [13]. The state-of-the-art benchmark accuracies and the accuracies of classifiers used in this research on the test sets of original correctly oriented images are reported in Table 1. The same trained classifiers are tested on randomly rotated images and show that the classification accuracy drops quite significantly. This proves that rotational corruptions indeed have an adverse effect on the performance of such classifiers.

Finally, the RISC framework, which incorporates pseudo-label generation and rotation correction before classification is evaluated. The RISC framework demonstrates notably superior performance across all three datasets when compared to original classifiers when fed with randomly-rotated images, with a 22% accuracy boost on OrganAMNIST, 17% increase on PneumoniaMNIST and 2% increase on RetinaMNIST, thereby proving its efficacy in robust image classification of images affected with rotational corruptions. This difference can be seen in Table 1 in the 'Classifier Accuracy (Randomly-Rotated Images)' and 'Classifier Accuracy (RISC Framework)' rows.

Table 2 lists the inference time per sample, cross-entropy loss of the RISC framework and RotNet and Classifier training time for each dataset as tested on a MacBook Air with Apple M1 chip and 8GB RAM.

Enhancing model explainability using GradCAM heatmaps

To fully trust a model's predictions, it's crucial to ensure it focuses on the right features, not just its test performance. This issue is demonstrated in Fig. 4 which shows GradCAM [12] heatmaps for the image classification of one image from each dataset, rotated at different angles. The PneumoniaMNIST classifier correctly classifies all four images as pneumonia-infected, but it gives weightage to different regions in each image. A cloudy area in the lung region indicates a pneumonia infection. The model correctly focuses on that cloudy region in the correctly- oriented image but fails to do so in the rotated ones. This helps us visualize that even when the model makes a correct prediction on rotation-affected images, it might not be doing so based on the right principles. The RetinaMNIST and

Table 1
RISC and novel performance evaluation.

	Train Data	Test Data	Organ MNIST	Pneumonia MNIST	Retina MNIST
Dataset side Train			34,581	4708	1080
Val			6491	524	120
Test			17,778	624	400
No of Classes			11	2	5
RotNet Accuracy	Rotated Training Data with Pseudo Labels	Rotated Testing Data with Pseudo Labels	0.66	0.99	0.73
Benchmark Accuracy (SOTA)			0.94	0.88	0.51
Classifier Accuracy (Original Images)	Original Training Data with Real Labels	Testing Data with Real Labels	0.93	0.88	0.51
Classifier Accuracy (Randomly Rotated Images)	Original Training Data with Real Labels	Randomly Rotated Images with Real Labels	0.53	0.7	0.48
Classifier Accuracy (RISC Framework)			0.75	0.87	0.5

Table 2
Details of RISC framework.

Dataset	Inference time per sample (s)	Cross-Entropy Loss	RotNet Training Time (Min)	Classifier Training Time (Min)
RetinaMNIST	0.0125	1.78	20	7
PneuomoniaMNIST	0.0096	0.53	16	41
OrganMNIST	0.011	1.367	11	24

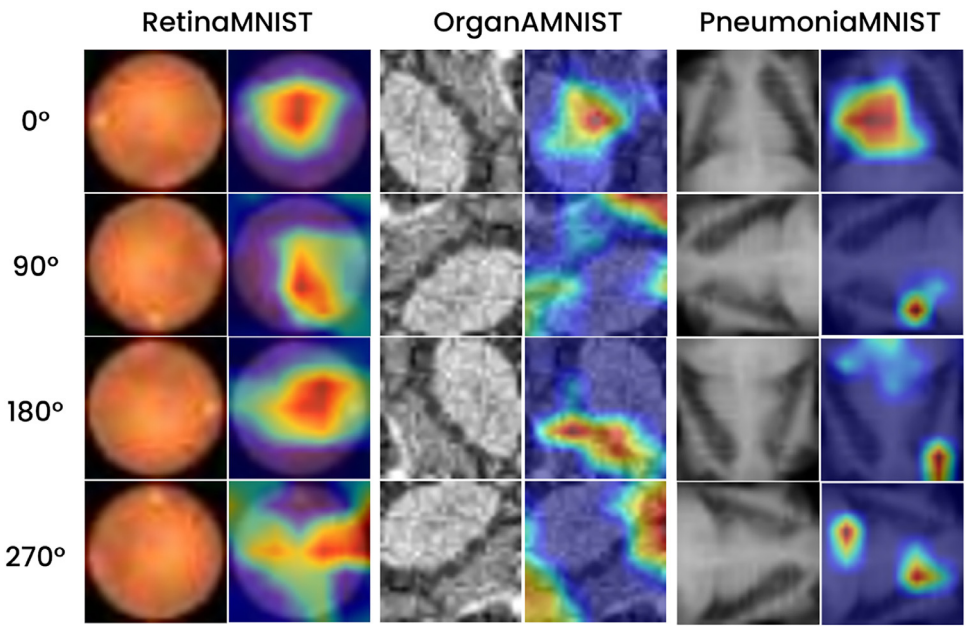

Fig. 4. GradCAM Heatmaps for Image Classification in MedMNIST. Rotated Image and its corresponding GradCAM heatmap for the downstream image classification task are shown in adjacent columns for all three datasets. (Images upscaled to 200×200 for better visibility).

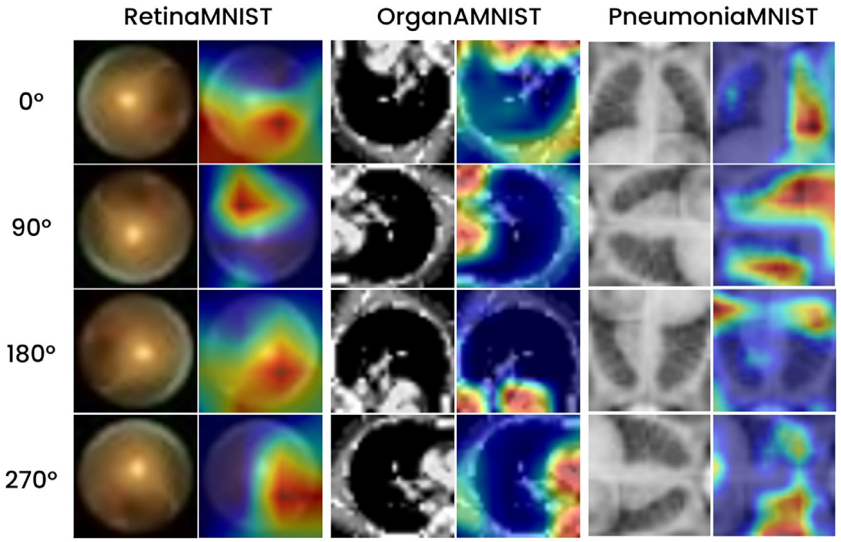

Fig. 5. GradCAM Heatmaps for RotNet. The rotated Image and its corresponding GradCAM heatmap for the RotNet (rotation label) in the pretext task are shown in adjacent columns for all three datasets.

OrganAMIST classifiers could only classify the depicted original image correctly while failing to do so for all rotations. The heatmaps show the inconsistency in attention for the same image but with applied rotation.

A similar methodology is used to obtain the GradCAM heatmaps shown in Fig. 5 which provide an insight into the pseudo-label generation by the RotNet. The RotNet was able to correctly predict all the rotation angles for the three samples and their corresponding rotated images. In the analysis of PneumoniaMNIST heatmaps, we note that the RotNet seems to predominantly emphasize the lung region during orientation prediction which is a strategy consistent with human perception and understanding. The OrganAMNIST shows a left kidney, where the RotNet seems to analyze the boundary of the kidney to determine its orientation. The RotNet heatmaps obtained from RetinaMNIST are difficult to explain and understand since all the images present a broadly uniform, circular image of the retina but as visualized, the model emphasizes on different regions in different rotations of the same sample.

We tackle multiple vision problems in medical image classification through the proposed SSL framework RISC. We were successful in firstly, crossing the initial results on three medical image datasets with a well-defined ensemble and fitting of pseudo-labels. In

the self-supervised learning task, this paper presents strong results on OrganAMNIST, RetinaMNIST and PneumoniaMNIST datasets, providing results with low-compute and high precision. The generalization capability of the self-supervised learning framework in the presence of rotations provides promising results which can be extended to larger benchmarks and tasks in the future including 3D Detection.

The impact of rotational corruptions has been highlighted as an important challenge, especially in various real-world problems such as classification tasks for medical images, object detection and segmentation tasks for autonomous driving etc. These should be corrected timely, without any assumptions of orientation and scaling as in unlabeled data. The rotation-invariant module is in fact able to outperform source data performance, which is another success of self-supervision as an intermediate outcome in this process. One significant contribution that we present is the ability of the framework to detect robustly without any access to rotation labels, being fully self-supervised and thus ideal for growing systems in medical imaging. This RISC framework mitigates rotational corruptions in medical imaging scenarios, however, with relevant data, it can be applied to adversarial attacks and other perturbations. We additionally present GradCAM-based mapping visualizations in Fig. 4 to highlight the effectiveness of rotational corrections in this study. We aim to expand this work to various real-world applications including multi-weather models for autonomous vehicles and multi-instrument datasets for medical imagery classification in the future.

Limitations

Currently, our method focuses on only four specific angle rotations, but more granularity in terms of rotational angles needs to be included to be able to generate even better outcomes

Ethics statements

Not Applicable

CRediT author statement

Devansh Singh: Methodology, Software, Writing- Original draft preparation. **Aboli Marathe:** Methodology, Software. **Sidharth Roy:** Methodology, Software, **Rahee Walambe:** Supervision, Validation, Writing- Reviewing and Editing. **Ketan Kotecha:** Validation, Reviewing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data is from open domain. Links are provided in reference list.

Acknowledgements

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

[1] K. Suzuki, Overview of deep learning in medical imaging, Radiol. Phys. Technol. 10 (3) (2017) 257–273 1.
[2] X. He, X. Yang, S. Zhang, J. Zhao, Y. Zhang, E. Xing, P. Xie, Sample- efficient deep learning for covid-19 diagnosis based on CT scans, medrxiv (2020) 2020–04. 1
[3] R. Krishnan, P. Rajpurkar, E.J. Topol, Self-supervised learning in medicine and healthcare, Nat. Biomed. Eng. 6 (12) (2022) 1346–1352 2.
[4] I. Misra, L.v.d. Maaten, Self-supervised learning of pretext-invariant representations, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2, 2020, pp. 6707–6717.
[5] T. Han, W. Xie, A. Zisserman, Self-supervised co-training for video representation learning, Adv. Neural Inf. Process. Syst. 33 (2020) 5679–5690 2.
[6] P. Goyal, M. Caron, B. Lefaudeux, M. Xu, P. Wang, V. Pai, M. Singh, V. Liptchinsky, I. Misra, A. Joulin, et al., Self-supervised pretraining of visual features in the wild, arXiv preprint arXiv:2103.01988 (2021). 2
[7] B. Chen, A. Rouditchenko, K. Duarte, H. Kuehne, S. Thomas, A. Boggust, R. Panda, B. Kingsbury, R. Feris, D. Harwath, et al., Multimodal clustering networks for self-supervised learning from unlabeled videos, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 8012–8021. 2.
[8] I. Misra, L. van der Maaten, Self-supervised learning of pretext-invariant representations (2019). arXiv:1912.01991.URL https://arxiv.org/abs/1912.01991 2
[9] G. Cheng, J. Han, P. Zhou, D. Xu, Learning rotation-invariant and fisher discriminative convolutional neural networks for object detection, IEEE Trans. Image Process. 28 (1) (2019) 265–278, doi:10.1109/TIP.2018.2867198.2.
[10] Y. Yao, G. Cheng, G. Wang, S. Li, P. Zhou, X. Xie, J. Han, On improving bounding box representations for oriented object detection, IEEE Trans. Geosci. Remote Sens. 61 (2023) 1–11, doi:10.1109/TGRS.2022.3231340.2.
[11] X. Xie, G. Cheng, Q. Li, S. Miao, K. Li, J. Han, Fewer is more: efficient object detection in large aerial images, Sci. China Inf. Sci. 67 (1) (2023) Dec10.1007/s11432-022-3718-5.URL2, doi:10.1007/s11432-022-3718-5.
[12] R.R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, D. Batra, Grad-CAM: why did you say that? Visual explanations from deep networks via gradient-based localization, CoRR abs/1610.02391 (2016). URL https://arxiv.org/abs/1610.02391 2, 6, 8
[13] J. Yang, R. Shi, D. Wei, Z. Liu, L. Zhao, B. Ke, H. Pfister, B. Ni, Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification, Sci. Data 10 (1) (2023) 41 2, 3, 7.

[14] J. Yang, R. Shi, B. Ni, Medmnist classification decathlon: a lightweight auto ml benchmark for medical image analysis, in: IEEE 18th International Symposium on Biomedical Imaging (ISBI), 2021, pp. 191–195. 2.
[15] R. Liu, X. Wang, Q. Wu, L. Dai, X. Fang, T. Yan, J. Son, S. Tang, J. Li, Z. Gao, et al., Deepdrid: diabetic retinopathy—grading and image quality estimation challenge, Patterns 3 (6) (2022) 2.
[16] D.S. Kermany, M. Goldbaum, W. Cai, C.C. Valentim, H. Liang, S.L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan, et al., Identifying medical diagnoses and treatable diseases by image-based deep learning, Cell 172 (5) (2018) 1122–1131 2.
[17] D. Kermany, K. Zhang, M. Goldbaum, Large dataset of labelled optical coherence tomography (oct) and chest x-ray images, Mendeley Data 3 (10.17632) (2018) 2.
[18] P. Bilic, P. Christ, H.B. Li, E. Vorontsov, A. Ben-Cohen, G. Kaissis, A. Sze- skin, C. Jacobs, G.E.H. Mamani, G. Chartrand, et al., The liver tumour segmentation benchmark (lits), Med. Image Anal. 84 (2) (2023) 102680.
[19] X. Xu, F. Zhou, B. Liu, D. Fu, X. Bai, Efficient multiple organ localization in ct image using 3d region proposal network, IEEE Trans. Med. Imaging 38 (8) (2019) 1885–1898 2.
[20] S. Gidaris, P. Singh, N. Komodakis, Unsupervised representation learning by predicting image rotations, arXiv preprint arXiv:1803.07728 (2018). 3
[21] S. Roy, A. Marathe, R. Walambe, K. Kotecha, Self-supervised learning for classifying the rotated images, in: 12th International Advanced Computing Conference 2022, 2022, p. 3.
[22] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, CoRR abs/1512.03385 (2015). arXiv:1512.03385. URL http://arxiv.org/abs/1512.033855.
[23] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M.S. Bernstein, A.C. Berg, L. Fei-Fei, Ima- genet large scale visual recognition challenge, CoRR abs/1409.0575 (2014). arXiv:1409.0575. URL http://arxiv.org/abs/1409.05755.