Can We Trust Saliency Maps to "Explain" Deep Learning Algorithms for Musculoskeletal Radiograph Abnormality Detection?

Kesavan Venkatesh, BSE Candidate, Dual-Affiliate, Johns Hopkins Biomedical Engineering and University of Maryland Medical Intelligent Imaging (UM2ii) Center

Simukayi Mutasa, MD; Fletcher Moore, MD; Jeremias Sulam, PhD; Paul H Yi, MD

Introduction

Saliency maps are commonly used to "explain" the decision-making of deep convolutional neural networks (DCNN) via heatmaps showing important image features. However, the utility and robustness of these saliency maps has not been rigorously evaluated for musculoskeletal imaging. The purpose of this study was to systematically evaluate the trustworthiness of saliency maps for identifying abnormalities on upper extremity (UE) radiographs.

Hypothesis

Saliency maps for extremity radiographs will localize reasonably to areas of abnormality but will fail other trustworthiness criteria.

Methods

We used Stanford's MURA dataset of 40,561 UE radiographs to train, validate, and test InceptionV3 and DenseNet-121 DCNNs to identify abnormal radiographs. We held-out a testset of 1276 images (638 abnormal), and used the remainder for DCNN training/validation. Three fellowship-trained musculoskeletal radiologists placed bounding boxes around abnormality(s) on the 638 positive test images with groundtruth defined as majority vote of annotated pixels.

We evaluated 6 saliency methods: Grad-CAM, Gradient Explanation (GRAD), Integrated Gradients (IG), Smoothgrad (SG), Smooth IG (SIG), and XRAI. We applied four trustworthiness criteria according to Arun N et al.'s framework [1] (Fig.1): 1) localization of abnormalities, 2) sensitivity to weight

randomization, 3) repeatability, and 4) reproducibility. We quantified performance of these saliency methods for the above criteria using relevant measures, such as AUROC.

Results

All saliency methods showed reasonable localization with AUROCs of 0.755 (Grad-CAM) to 0.863 (XRAI) for a range of abnormalities from fracture (Fig.2A) to orthopedic hardware (Fig.2B). Only 3 methods passed the test for sensitivity to weight randomization test (GRAD, IG, SIG). Four passed repeatability (Grad-CAM, IG, SIG, XRAI) and 3 passed reproducibility tests (GRAD-CAM, IG, XRAI). None of the saliency methods met all four criteria for trustworthiness.

Conclusion

Although saliency maps appear to have reasonable localization ability for UE radiograph abnormalities, few meet the other trustworthiness criteria. We recommend caution when interpreting these saliency maps, which should be further scrutinized and evaluated before being widely accepted for clinical use.

Figures

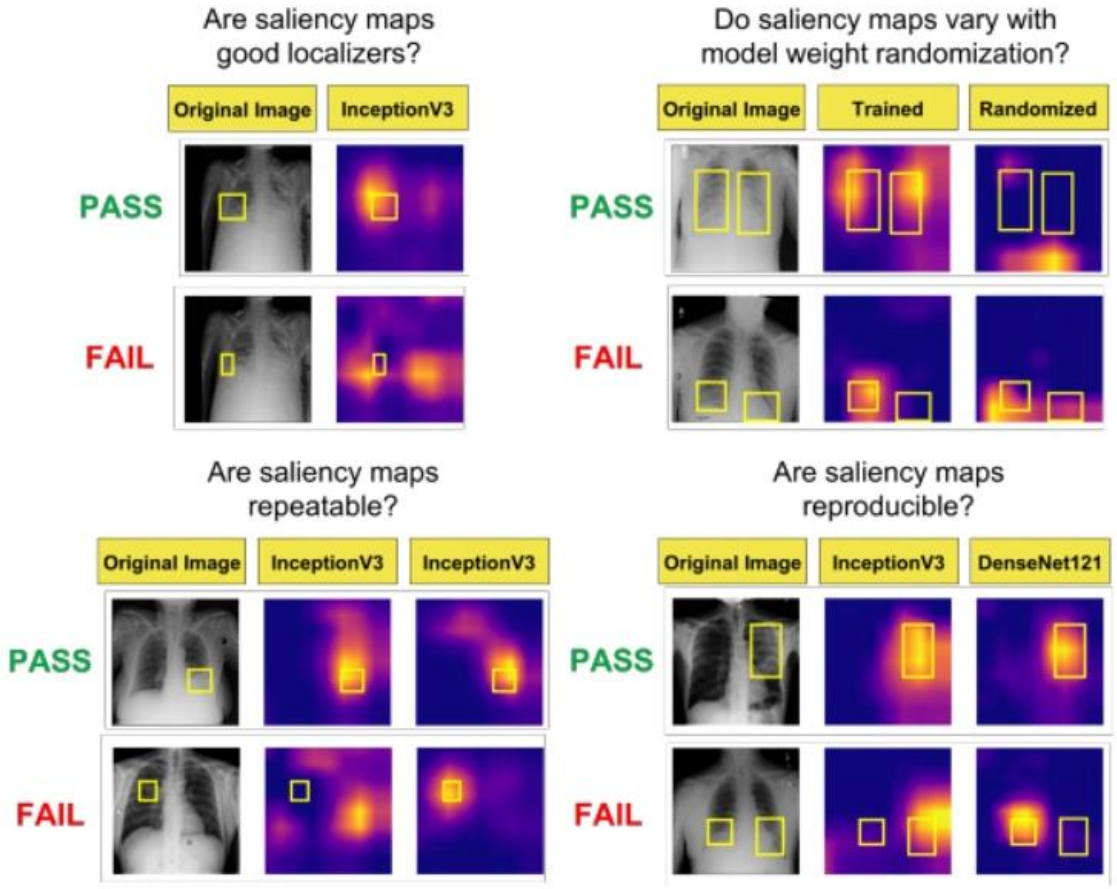

Figure 1. Framework for evaluating saliency map trustworthiness described by [Arun N et al. ArXiv 2020]. Please note that while this framework depicts chest radiographs, it is applicable to other medical images, i.e., the upper extremity radiographs used in our study.

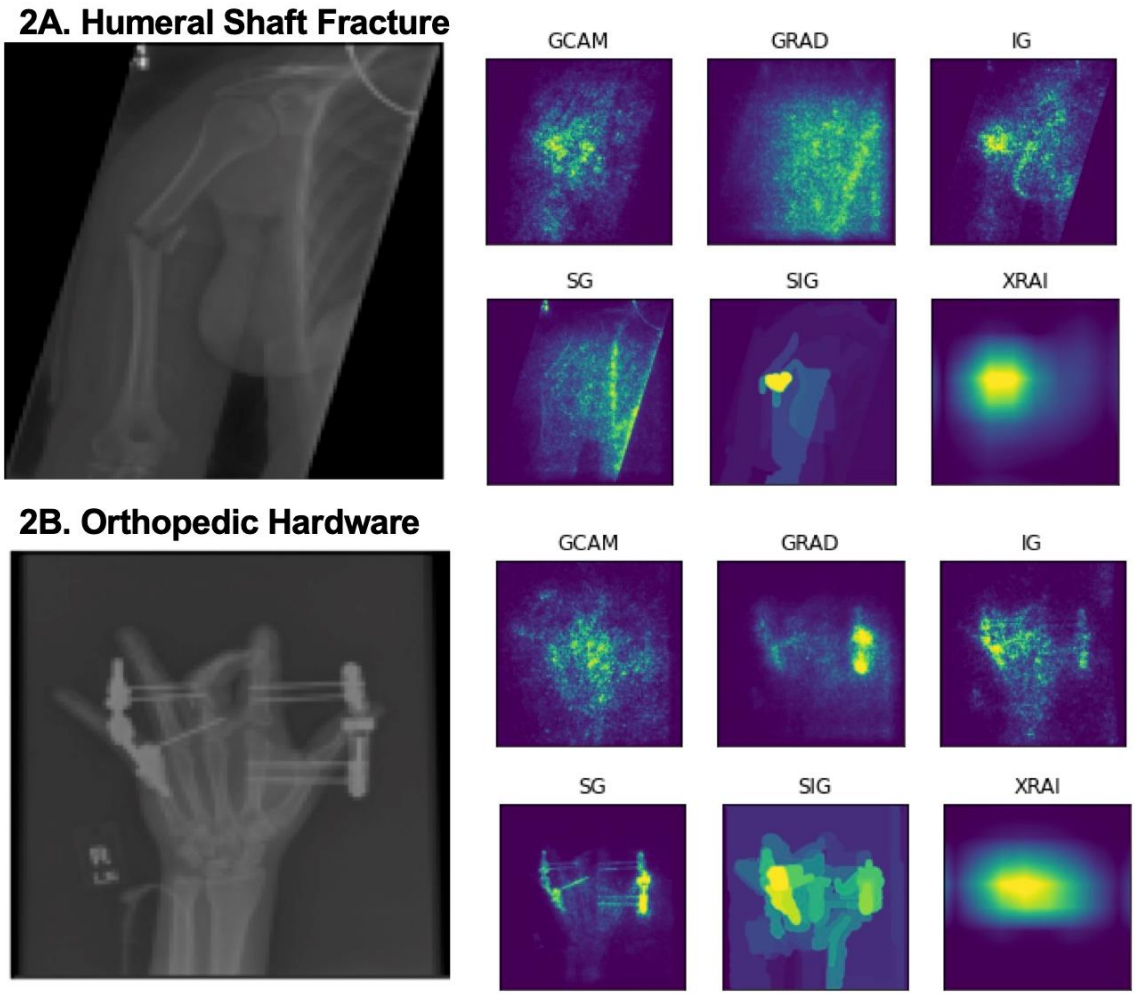

Figure 2. Two examples of abnormal images in the MURA test set showing A) humeral shaft fracture and B) Orthopedic hardware (external fixators in the hand) with accompanying saliency heatmaps. GCAM = Grad-CAM, GRAD = Gradient Explanation, IG = Integrated Gradients, SG = Smoothgrad, SIG = Smooth IG.

Keywords

Artificial Intelligence; Imaging Research

SIIM22 Presentation

Pediatric & MSK Machine Learning Research Abstracts | 3004 | In-Person