Vision Transformers are More Robust to Real-World Medical Image Variations than Convolutional Neural Networks

Skylar Chan, Research Assistant, University of Maryland School of Medicine

Introduction

Vision transformers (ViTs) may have advantages over convolutional neural networks (CNNs) in medical diagnosis, including improved explainability. However, it is unclear if ViTs are most robust to real-world variations in images. We compared the robustness of ViTs and CNN chest x-ray (CXR) classification models to real-world variations in images by applying computational "stress tests."

Hypothesis

ViTs will be more robust towards common imaging variations than CNNs.

Methods

We trained and evaluated ImageNet-pretrained DenseNet121 CNN and DeiT-B ViT classification models for 14 disease labels using the NIH CXR14 dataset (N&#3f112,120) split into training (n=77,358), validation (n=9,166), and testing (n=25,596) sets; hyperparameter optimization was performed using grid search across 128 hyperparameter combinations. We performed "stress tests" by evaluating model performance on the test set after applying transformations to each test image: pixel inversion, vertical flip, horizontal flip, rotation, brightness scaling, contrast scaling, and resizing (Fig.1). Robustness was defined as model prediction consistency between baseline (original) images and transformed images and was measured by comparing weighted AUC (wAUC) and Jaccard (intersection over union) loss between baseline and transformed test images using paired t-tests and Wilcoxon signed-rank tests, respectively. Predictions were visualized in latent space using Uniform Aligned Manifold Approximation (UMAP).

Results

ViT and CNN models had similar baseline performance with wAUCs of 0.78-0.79. On computational stress testing, however, the ViT had significantly higher wAUC than the CNN ($p < 0.01$) for most image transformations (Fig.2A) and lower Jaccard loss (fewer prediction deviations from baseline) for all transformation groups ($p < 0.0001$, all) [Fig.2B]. Uniformity of ViT and CNN predictions directly correlated with increasing Jaccard loss, indicating similar model behavior on image transformations at the prediction level (Fig.2C).

Conclusion

Despite similar baseline performance, ViTs are more robust than CNNs to clinically encountered variations in CXR processing, suggesting that ViTs may be more durable when deployed in real-world settings.

Synopsis

We compare the robustness of vision transformers (ViTs) to convolutional neural networks (CNNs) on real-world image variations. We show ViTs are more robust than CNNs, with higher performance and lower errors in Chest X-Ray classification.

Figure(s)

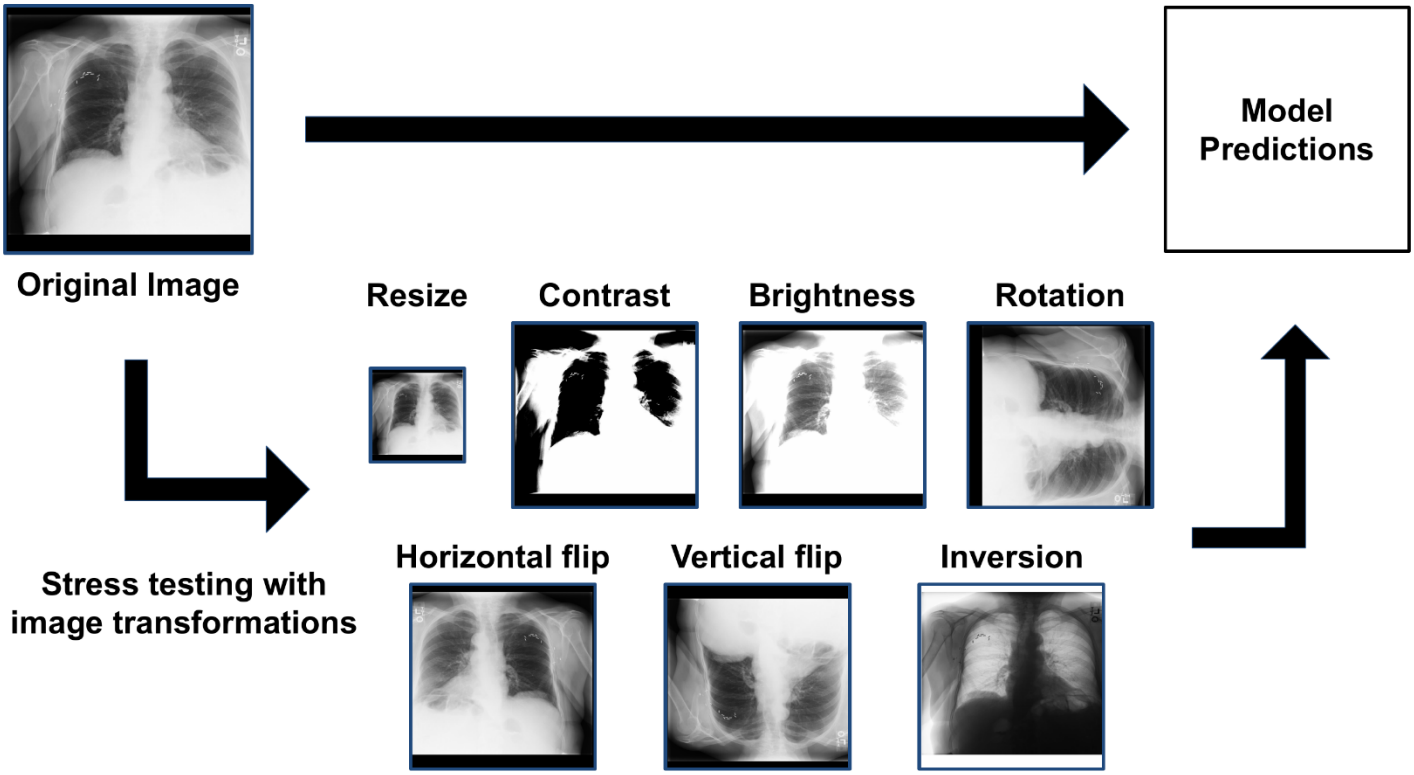

Figure 1. Computational stress test pipeline for evaluating robustness of DL models. (Top) Untransformed images undergo standard processing before obtaining disease label predictions. (Bottom) Transformed images undergo resizing, contrast scaling, brightness scaling, rotation, horizontal flip, vertical flip, or inversion operations before standard processing. Resulting images are fed to models to obtain disease label predictions.

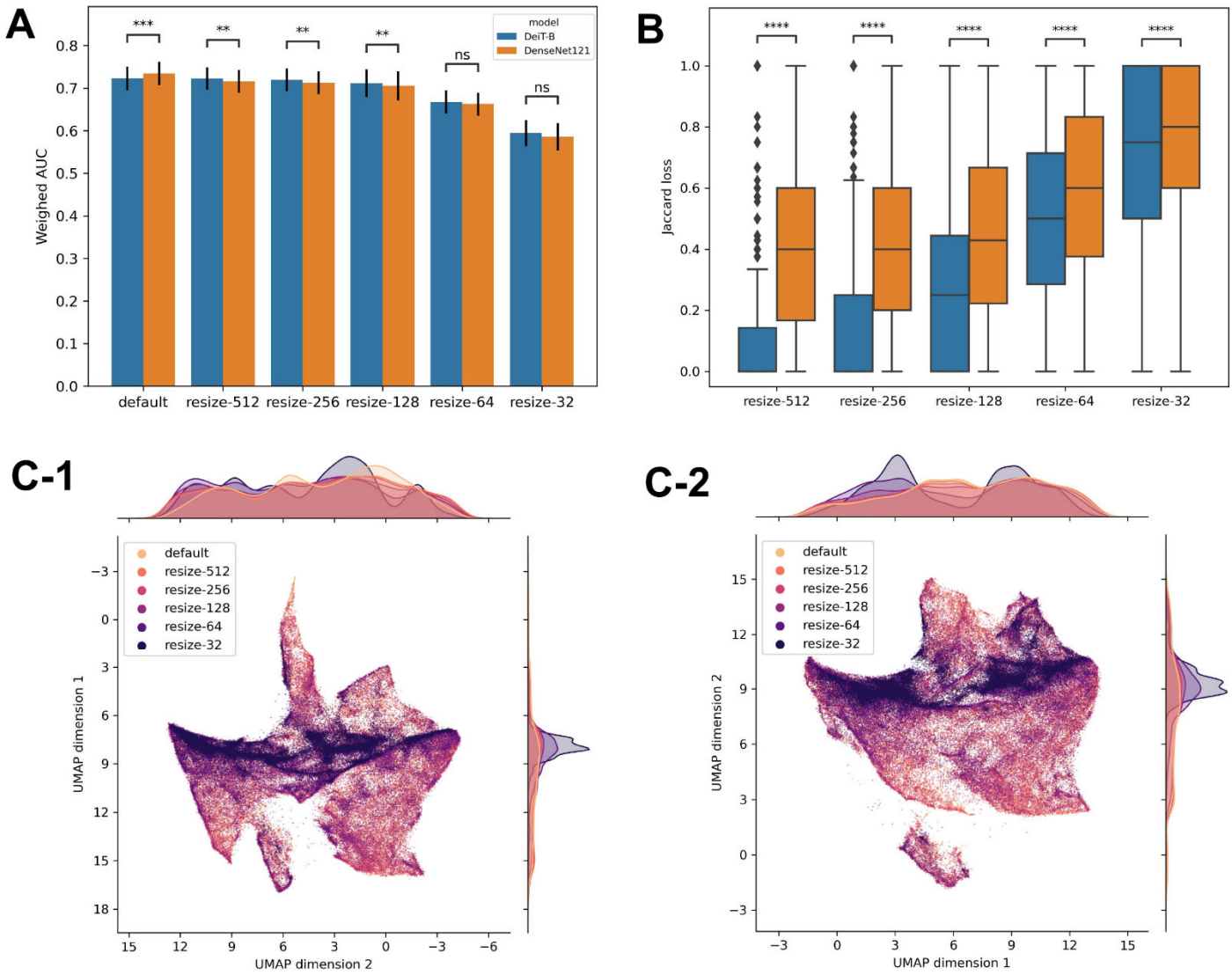

Figure 2. Model evaluations of resize transformations using A) weighted AUC, B) Jaccard loss, C) UMAP visualizations. Weighted AUC and Jaccard loss were compared with paired T-test of per-label AUCs and Wilcoxon signed-rank test respectively (ns: p>0.05, *: p<0.05, **: p<0.01, ***: p<0.001, ****: p<0.0001). Similar trends were observed for other transformations.

Keywords

Artificial Intelligence / Machine Learning; Imaging Research