Benchmarking and Boosting Transformers for Medical Image Classification

DongAo Ma[1], Mohammad Reza Hosseinzadeh Taher[1], Jiaxuan Pang[1], Nahid Ul Islam[1], Fatemeh Haghighi[1], Michael B. Gotway[2], Jianming Liang[1]

[1] Arizona State University [2] Mayo Clinic

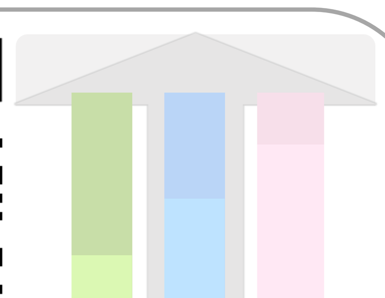
ASU College of Health Solutions, Arizona State University

MICCAI 2022 Singapore — DART International Workshop on Domain Adaptation & Representation Transfer

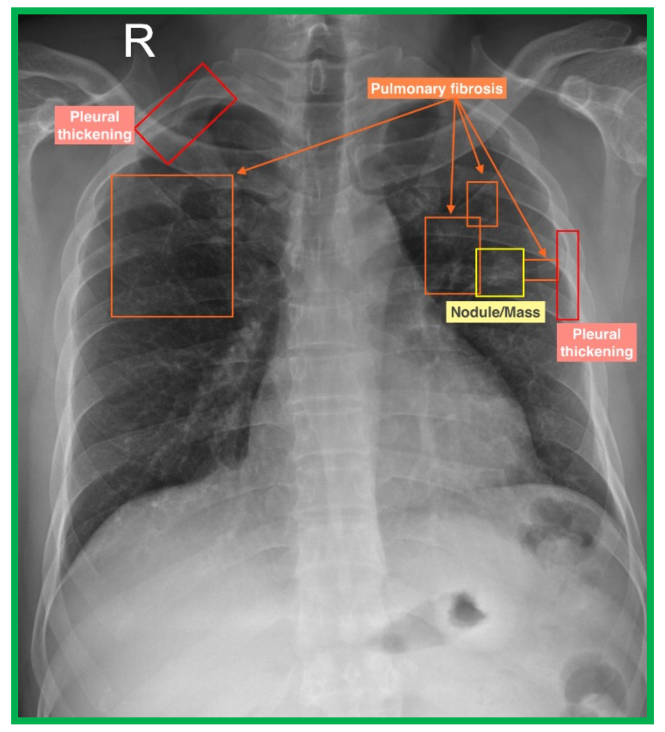
MAYO CLINIC

MOTIVATION: Transformers have good properties for medical images

- Pay attention to whole image to model global context of a body region
- Capture intra-image relations to detect co-occurrence of pathologies
- Build hierarchical feature maps to identify lesions at different scales



This research has been supported in part by ASU and Mayo Clinic through a Seed Grant and an Innovation Grant, and in part by the NIH under Award Number R01HL128785. The content of this paper is covered by patents pending.



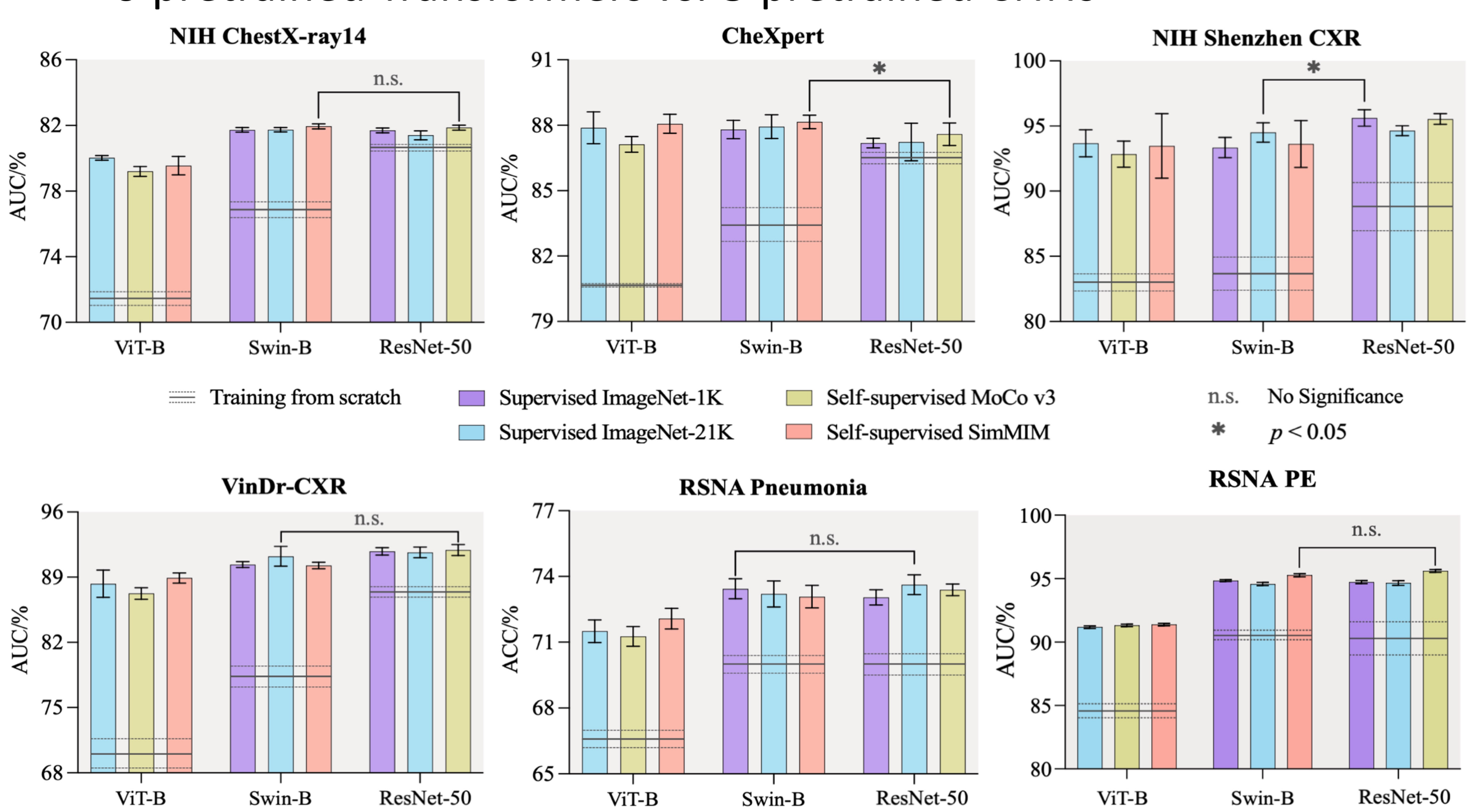
Q1
How do visual transformers perform on medical images?

Transformers need pretraining to achieve competitive performance on medical images

Benchmarking:

2 Transformer architectures *vs.* **1** CNN architecture

6 pretrained Transformers *vs.* **3** pretrained CNNs



Result I : Transformers significantly underperform CNNs when training from scratch

Result II: Transformers can offer similar results as CNNs with ImageNet pretraining

Q2
What pretraining method for transformers performs the best?

Self-supervised learning based on masked image modeling can learn preferable features for medical tasks

Comparing:

Supervised pretraining using both ImageNet images and labels *vs.* Self-supervised pretraining using only ImageNet images

Method		ChestX-ray14	CheXpert	Shenzhen	VinDr-CXR	RSNA Pneumonia	RSNA PE
Supervised	ViT-B	80.05±0.17	87.88±0.50	93.67±1.03	88.30±1.45	71.50±0.52	91.19±0.11
	Swin-B	81.73±0.14	87.80±0.42	93.35±0.77	**90.35±0.31**	73.44±0.46	94.85±0.07
SimMIM	ViT-B	79.55±0.56	88.07±0.43	93.47±2.48	88.91±0.55	72.08±0.47	91.39±0.10
	Swin-B	**81.95±0.15**	**88.16±0.31**	**94.12±0.96**	90.24±0.35	**73.66±0.34**	**95.27±0.12**
Swin-B vs. Swin-B		*	n.s.	*	n.s.	n.s.	*

*The best methods are bolded while the others are highlighted in green if they achieve equivalent performance compared with the best one (i.e., p > 0.05).

Result III: Self-supervised SimMIM model with the Swin-B backbone outperforms fully-supervised baselines

Q3
Can we further boost transformers' performance for medical images?

Domain-adaptive pretraining using a large-scale in-domain dataset can further boost transformers' performance

Boosting:

Create a large-scale in-domain dataset by assembling 13 datasets to satisfy transformer's data hunger → X-rays(926K)

Adopt self-supervised learning to overcome heterogeneity of expert labels

Model (SimMIM+Swin-B)	ChestX-ray14	CheXpert	Shenzhen	VinDr-CXR	RSNA Pneumonia
Scratch	77.04±0.34	83.39±0.84	92.52±4.98	78.49±1.00	70.02±0.42
ImageNet	81.95±0.15	88.16±0.31	94.12±0.96	90.24±0.35	73.66±0.34
ChestX-ray14	78.87±0.69	86.75±0.96	93.03±0.48	79.86±1.82	71.99±0.55
X-rays(926K)	82.72±0.17	87.83±0.23	95.21±1.44	90.60±1.95	73.57±0.27
ImageNet→ChestX-ray14	82.45±0.15	87.74±0.31	94.83±0.90	90.33±0.88	73.85±0.72
ImageNet→X-rays(926K)	**83.04±0.15**	**88.37±0.40**	**95.76±1.79**	**91.71±1.04**	**74.09±0.39**
ImageNet vs. ImageNet→X-rays(926K)	>***	n.s.	*	>***	*

*The best methods are bolded while the others are highlighted in green if they achieve equivalent performance compared with the best one (i.e., p > 0.05).

Result IV: The domain-adapted model with learning experience on both ImageNet and a large-scale in-domain data (X-rays(926K)) achieves the highest performance