TOWARD TRUSTWORTHY HEALTHCARE AI: ATTENTION-BASED FEATURE LEARNING FOR COVID-19 WITH CHEST RADIOGRAPHY

7/22/2022

Kai Ma, Pengcheng Xi, Karim Habashy, Ashkan Ebadi, Stéphane Tremblay, Alexander Wong

Faculty of Engineering, University of Waterloo

Digital Technologies Research Centre, National Research Council Canada

UNIVERSITY OF WATERLOO | FACULTY OF ENGINEERING

NRC·CNRC Canada

OVERVIEW

- Evaluation of CNNs and Transformer deep learning architectures
  - Traditional CNN approach vs. Attention mechanism
  - Representation learning capability for chest X-ray (CXR) classification
- Model performance — Precision, Sensitivity
- Model trustworthiness
  - Trust Score — Wong et. al., 2020
  - Visualization — AblationCAM

MOTIVATION

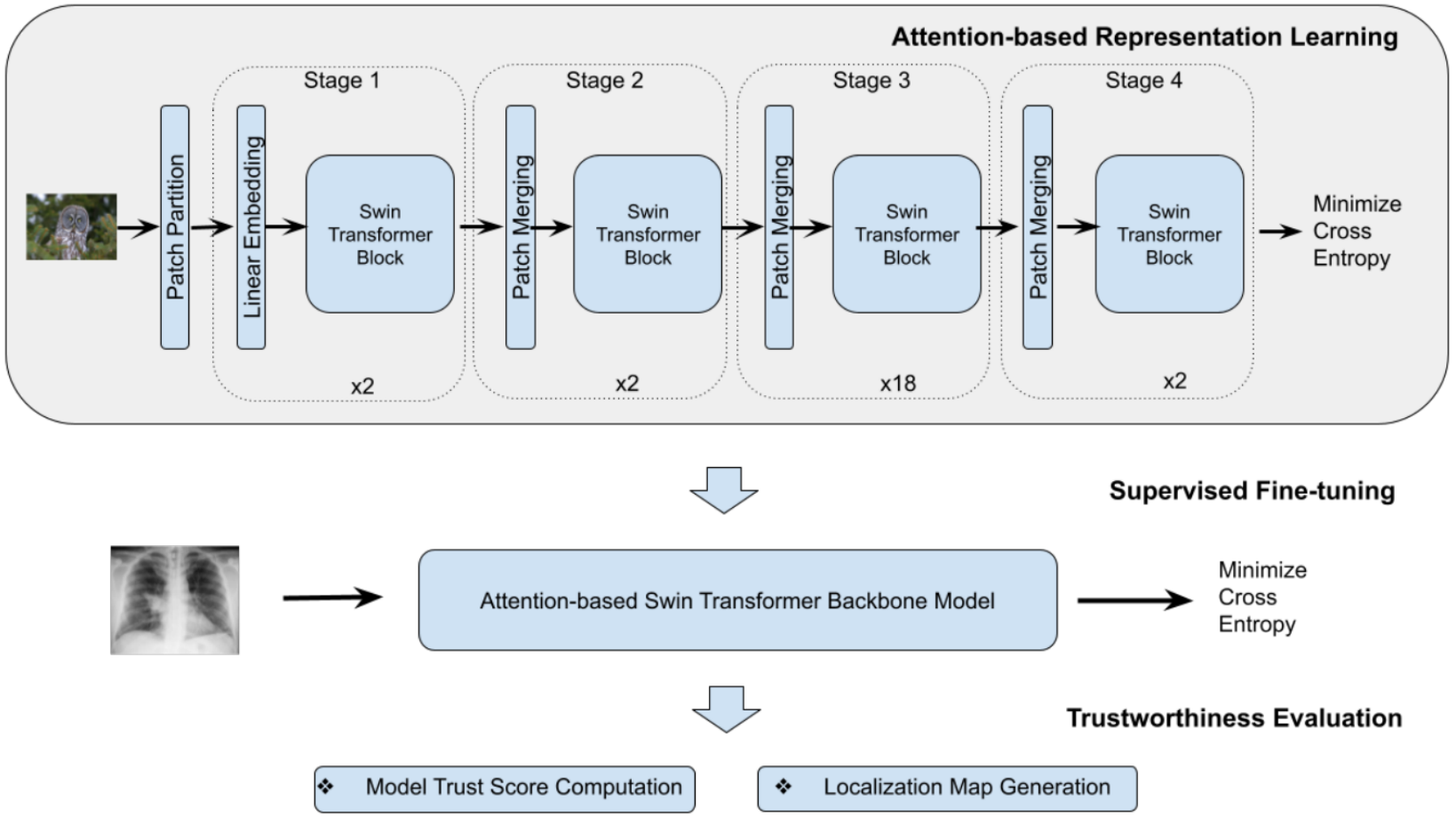
- Importance of trust in medical AI
  - Explainability of results and model confidence are crucial
  - Quantifiable metric

- Transformer's potential architectural advantages
  - CNNs have inductive biases
  - Attention mechanism — analogous to imaging evaluation by doctors

- Fair evaluation of localization maps
  - Existing literature uses visualization techniques that are specific to architecture

METHODOLOGY – MODEL ARCHITECTURE

- Models:
  - ResNet-50
  - DenseNet-121
  - Swin Transformer



METHODOLOGY – DATASET, TRUST QUANTIFICATION

- Dataset: COVIDx V9B
  - Combination of data repositories (RSNA, etc)
  - 10% of training set sampled for validation

Table 1. Data split for COVIDx V9B

SPLIT	NEGATIVE	POSITIVE	TOTAL
TRAIN	13,992	15,950	30,482
TEST	200	200	400

- Trust score computation:
  - Penalize undeserved confidence, reward well-placed confidence

$$Q_z(x,y) = \begin{cases} C(y \mid x)^{\alpha}, & \text{if } x \in R_{y=z|M} \\ (1 - C(y \mid x))^{\beta}, & \text{if } x \in R_{y\neq z|M}, \end{cases}$$

- AblationCAM
  - Localization maps through ablation analysis

RESULTS – PERFORMANCE & TRUST SCORE

Table 3. Precision scores on the unseen COVIDx V9B test split. The best results in each class are bolded.

MODEL	NEGATIVE	POSITIVE
RESNET (200 EPOCHS)	**0.952**	**1.000**
DENSENET (200 EPOCHS)	0.948	0.995
SWIN-B (30 EPOCHS)	0.926	**1.000**
SWIN-B (50 EPOCHS)	0.935	**1.000**
SWIN-B (100 EPOCHS)	0.930	**1.000**
SWIN-B (200 EPOCHS)	**0.952**	**1.000**

Table 4. Sensitivity scores on the unseen COVIDx V9B test split. The best results in each class are bolded.

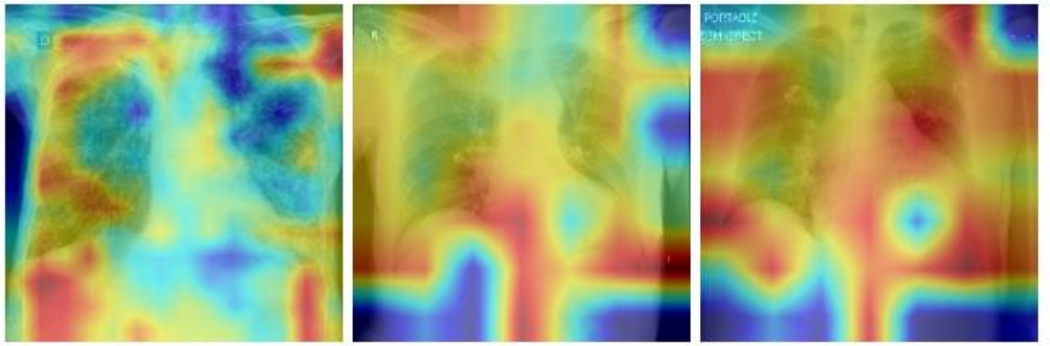
MODEL	NEGATIVE	POSITIVE
RESNET (200 EPOCHS)	**1.000**	**0.950**
DENSENET (200 EPOCHS)	0.995	0.945
SWIN-B (30 EPOCHS)	**1.000**	0.920
SWIN-B (50 EPOCHS)	**1.000**	0.930
SWIN-B (100 EPOCHS)	**1.000**	0.925
SWIN-B (200 EPOCHS)	**1.000**	**0.950**

Table 5. Trust scores calculated from each experiment on the positive class. The best result is bolded.

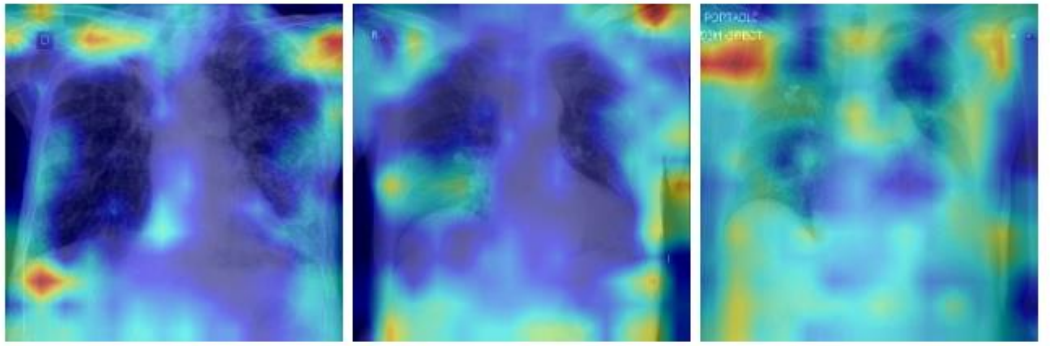
MODEL	TRUST SCORE
RESNET (200 EPOCHS)	0.923
DENSENET (200 EPOCHS)	0.922
SWIN-B (30 EPOCHS)	0.943
SWIN-B (50 EPOCHS)	0.959
SWIN-B (100 EPOCHS)	0.954
SWIN-B (200 EPOCHS)	**0.963**

RESULTS – VISUALIZATION

(a) Original chest radiographs for positive COVID-19 samples



(b) Swin-B 200-epoch Ablation-CAM



(c) ResNet-50 200-epoch Ablation-CAM

WHAT'S NEXT?

- Comparison against newer, leading CNNs and different Transformer models
  - ConvNeXt, NFNet
  - Swin-L, CoCa, MaxViT
- Further validation of results on other datasets
  - SIIM-FISABIO-RSNA, 3-class COVIDx
- Exploration of data-hungriness of Transformers
- Departure from model evaluation to introducing more novelty

THANKS FOR LISTENING!

Feel free to ask any questions!

UNIVERSITY OF WATERLOO | FACULTY OF ENGINEERING