New Microbes and New Infections 62 (2024) 101457

Contents lists available at ScienceDirect

New Microbes and New Infections

journal homepage: www.journals.elsevier.com/new-microbes-and-new-infections

Automated quantification of SARS-CoV-2 pneumonia with large vision model knowledge adaptation

Zhaohui Liang, Zhiyun Xue, Sivaramakrishnan Rajaraman, Sameer Antani [*,1]

Computational Health Research Branch, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA

ARTICLE INFO

Keywords:
SARS-CoV2 pneumonia
Artificial intelligence
Large vision model
Self-supervised learning
Vision transformer

ABSTRACT

Background: Large vision models (LVM) pretrained by large datasets have demonstrated their enormous capacity to understand visual patterns and capture semantic information from images. We proposed a novel method of knowledge domain adaptation with pretrained LVM for a low-cost artificial intelligence (AI) model to quantify the severity of SARS-CoV-2 pneumonia based on frontal chest X-ray (CXR) images.

Methods: Our method used the pretrained LVMs as the primary feature extractor and self-supervised contrastive learning for domain adaptation. An encoder with a 2048-dimensional feature vector output was first trained by self-supervised learning for knowledge domain adaptation. Then a multi-layer perceptron (MLP) was trained for the final severity prediction. A dataset with 2599 CXR images was used for model training and evaluation.

Results: The model based on the pretrained vision transformer (ViT) and self-supervised learning achieved the best performance in cross validation, with mean squared error (MSE) of 23.83 (95 % CI 22.67–25.00) and mean absolute error (MAE) of 3.64 (95 % CI 3.54–3.73). Its prediction correlation has the R^2 of 0.81 (95 % CI 0.79–0.82) and Spearman ρ of 0.80 (95 % CI 0.77–0.81), which are comparable to the current state-of-the-art (SOTA) methods trained by much larger CXR datasets.

Conclusion: The proposed new method has achieved the SOTA performance to quantify the severity of SARS-CoV-2 pneumonia at a significantly lower cost. The method can be extended to other infectious disease detection or quantification to expedite the application of AI in medical research.

1. Introduction

The ongoing coronavirus disease 2019 (COVID-19) global pandemic caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has had a significant impact on global health. It has infected over 775 million people and caused 7 million deaths worldwide as of July 2024 [1]. SARS-CoV-2 pneumonia is not only associated with critical cases leading to acute respiratory distress syndrome (ARDS), but also related to asymptomatic or presymptomatic COVID-19 infection as confirmed by radiographic findings [2]. Multiple clinical studies confirmed that SARS-CoV-2 pneumonia is a high-risk factor for the mortality of severe COVID-19 cases particularly associated with the elderly and comorbidities, or other immunocompromising conditions [3–5].

There have been several successful studies on applying artificial intelligence (AI) methods for COVID-19 detection and diagnosis since the outbreak of the pandemic. The major focus was on the use of deep learning (DL) with convolutional neural network (CNN) for detection of radiological patterns indicative of the disease on either chest X-ray (CXR) or computed tomography (CT) images for fast screening and initial diagnosis [6–8]. Although CT image detection has good sensitivity and specificity for SARS-CoV-2 pneumonia diagnosis, its high cost and high dosage of radiational exposure restricted its use as a primary screening tool. In contrast, CXR has many merits as a primary tool for large-volume screening for COVID-19 infection and continuous monitoring for SARS-CoV-2 pneumonia. The COVID-19 manifestations featured by ground-glass opacities and patchy consolidations on CXR are usually involved bilaterally at the peripheral and lower zone of the pulmonary area [9]. Thus, the success of capturing the CXR visual features of COVID-19 requires the AI model's capacity to integrate global pixel characteristics. Although convolutional neural network (CNN) models have demonstrated excellent performance in many computer vision tasks, the intrinsic reliance of these models on a small receptive

* Corresponding author.
E-mail addresses: zhaohui.liang@nih.gov (Z. Liang), zhiyun.xue@nih.gov (Z. Xue), sivaramakrishnan.rajaraman@nih.gov (S. Rajaraman), sameer.antani@nih.gov (S. Antani).
[1] Postal address: Building 38A, 8600 Rockville Pike MSC 6075 Bethesda, MD 20894, USA.

https://doi.org/10.1016/j.nmni.2024.101457
Received 30 December 2023; Received in revised form 10 July 2024; Accepted 12 August 2024
Available online 15 August 2024
2052-2975/© 2024 Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

field and locality of pixel dependencies through the convolution operations restrict their performance from integrating global pixel relations [10].

Inspired by recent advances in AI, we proposed a novel method to combine the merits of transformers and contrastive learning. Taking advantage of the availability of large vision models (LVM), e.g., ViT, DINOv2 [11], etc., we apply the pretrained LVMs as the backbone for primary feature extraction and an encoder trained by self-supervised learning to adapt the feature domain to CXR images. The encoded CXR domain features will be fed to a multi-layer perceptron (MLP) to predict the final quantitative estimate of the severity of SARS-CoV-2 pneumonia. We evaluate the performance of our proposed new method to demonstrate its effectiveness. We also expect that such an approach could significantly lower the cost of developing high-performance, disease-specific AI models for newly emerging infectious diseases.

2. Materials and methods

2.1. Quantification of SARS-CoV-2 pneumonia

The continuous monitoring of SARS-CoV-2 pneumonia is an effective measure of the severity, and a reliable predictor for the prognosis of COVID-19 patients. A multicenter study concluded that the progressive increase of radiographic assessment of the lung edema (RALE) score was associated with the mortality of ARDS cases by COVID-19 [12]. RALE is a quantitative tool to assess the extent and density of alveolar opacities of different regions of the lung on CXRs that reflects the degree of pulmonary edema [13]. A multicenter study on 350 CXR images from 139 COVID-19 ARDS patients revealed that the progressive increase of RALE score was associated with higher mortality and longer use time of ventilator [6]. The modified RALE scoring system (mRALE) [13] that will be used in this study is an improved version of RALE to quantify the overall pulmonary edema in a range from 0 to 24 on frontal CXR images. Applying artificial intelligence (AI) to automatically quantify the mRALE score directly from frontal CXR images can serve not only as an

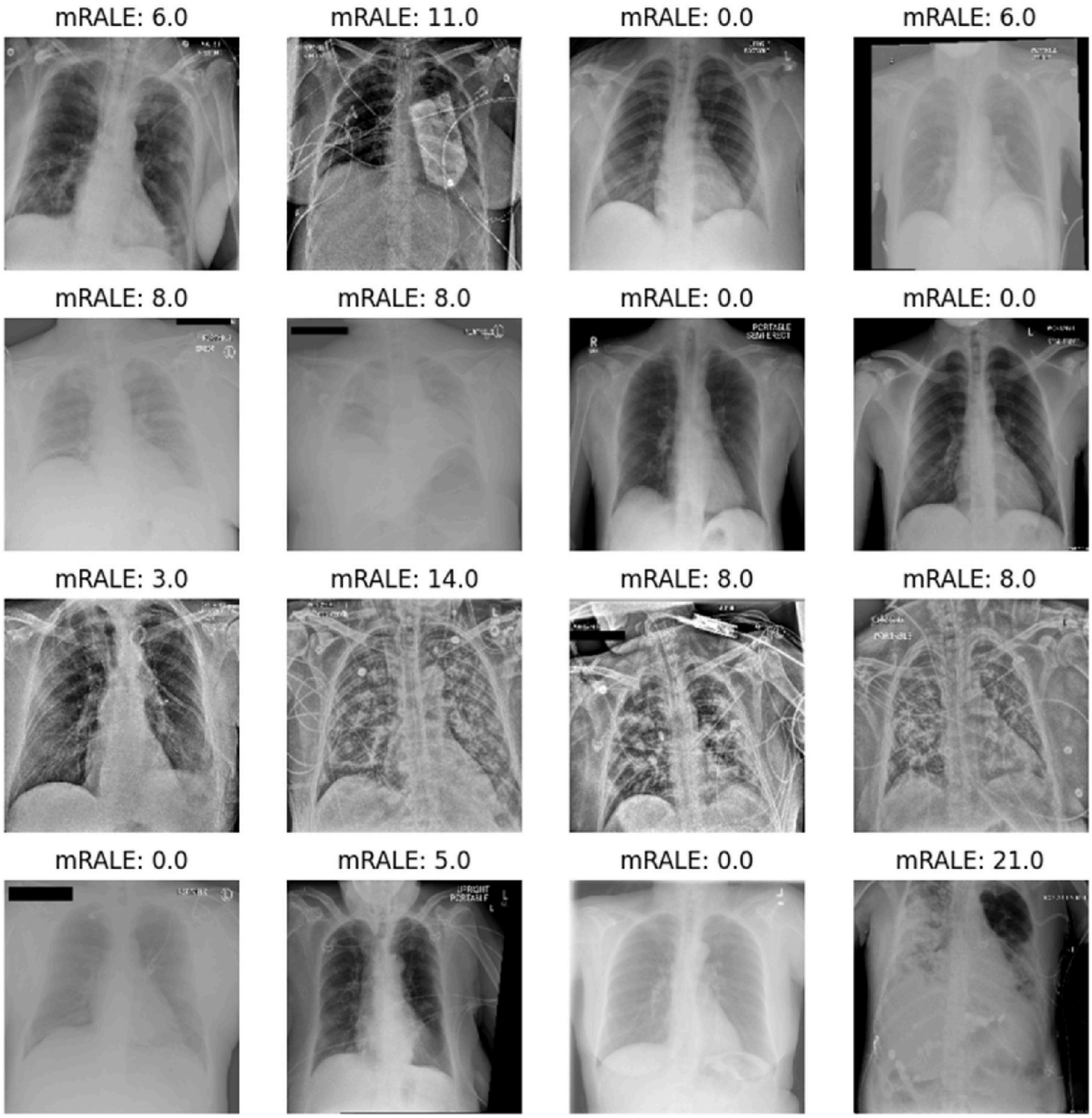

Fig. 1. CXR image with mRALE annotation.

effective and convenient method to assess the severity of SARS-CoV-2 pneumonia, but also as a reliable predictor of the prognosis of COVID-19 cases.

2.2. CXR image dataset

The dataset for this study was retrieved from the 2023 MIDRC mRALE Mastermind Challenge [14] with 2599 annotated CXR images. The mRALE scores were determined by the expert radiologists in the US that contributed their time and expertise to the challenge. The original CXR image dataset is in the DICOM format which is accessible at the MIDRC challenge repository (https://github.com/MIDRC/COVID19_Challenges). We converted the images into PNG format and resized them to $224 \times 224 \times 3$ to fit the input requirement of the pretrained models. Some CXR image examples are shown in Fig. 1.

2.3. Vision transformer for medical image processing

There have been many successful studies on applying AI for COVID-19 detection and diagnosis since the outbreak of the pandemic. The majority used deep learning (DL) with convolutional neural network (CNN) for radiological patterns detection on either chest X-ray (CXR) or computed tomography (CT) images for fast screening and initial diagnosis [6–8]. A survey covered more than 100 published studies of DL for COVID-19 concluded that using CNN on CXR images for COVID-19 detection has many advantages such as easy access, low radiational exposure and low cost compared to CT scans, and it has higher sensitivity compared to ultrasound images. However, the CNN-based AI methods had concerns about reliability and robustness due to the insufficient training without large and well-annotated image datasets [15].

To overcome the restrictions of the conventional CNN architecture, the Vision Transformer (ViT) based on the transformer architecture with the multi-head attention mechanism [16] was proposed by Dosovitskiy et al., in 2020 [17] to model long-range pixel dependency. The vision transformers have shown the state-of-the-art (SOTA) performance on image classification given sufficient training by large image datasets [17]. For example, Park et al. proposed a multi-task transformer model trained by several public and internal CXR image datasets to detect COVID-19-positive CXR images and to quantify the pneumonia severity [10]. Another study by Li et al. used the Siamese network trained by large CXR datasets to learn the CXR features, then they used the contrastive learning strategy to predict the severity of SARS-CoV-2 pneumonia by comparing the X-ray similarity [18].

2.4. Architecture of vision transformer

A vision transformer (ViT) is a deep neural network with the multi-head self-attention mechanism [16] to enhance the global integration of patterns. It was originally used in natural language processing (NLP) to process sequence data. The transformer architecture is later adapted for computer vision tasks to integrate the long-range dependency of pixel regions [17,19,20]. The ViT divides the input image into small patches (Fig. 2(a)) with the corresponding positional encoding to form the tokens (z_0). z_0 is fed to the transformer encoders. The transformer encoder layers receiving the sequence of embedded patches are composed of transformer units consisting of the multi-head self-attention (MSA) unit and the MLP block connected by layer normalization (LN) (Fig. 2(b)). In the MSA, the input sequence generates three values (Q-query, K-key, V-value) by multiplying the tokens with three learned matrices U_{QKV} to determine the relevance among different tokens (Fig. 2(c) & (d)). Taking advantage of the transformer architecture, ViT can effectively learn the global pixel dependency and integrate the patterns captured by different receptive fields with the patching and embedding operations, thus the limitation of CNN on learning only local pixel dependency will be alleviated. Based on this understanding, we used two vision transformer models pretrained by large image datasets as the LVM for primary feature extraction, and the extracted features will be used for domain adaptation through self-supervised contrastive learning.

2.5. Self-supervised contrastive learning for LVM knowledge domain adaptation

Self-supervised contrastive learning is a machine learning (ML) strategy to train the ML model in an unsupervised manner by comparing the similarity of training data as supervisory signals. We applied this method to train the feature encoders for knowledge domain adaptation. In our previous study, we used the SimSiam (Simple Siamese network) architecture [21] to train a ResNet-50 encoder using a 2048-dimensional feature vector to represent the CXR features for the downstream mRALE score predictor. The result showed it can reduce the mRALE prediction errors compared to the conventional CNN models [22]. Based on this finding, we propose a novel method to combine the strengths of large vision models (LVM) and self-supervised learning to further perform domain adaptation. The training has two phases. In the first phase, we used the pretrained LVM (ViT or DINOv2) to extract the pixel features.

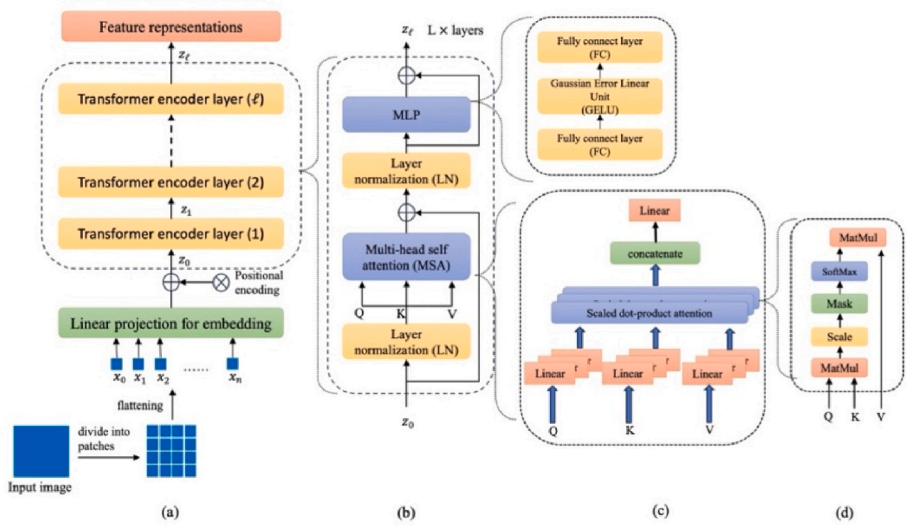

Fig. 2. Vision transformer with multi-head attention.

Next, we used the extracted features to optimize a feature encoder by SimSiam self-supervised learning for knowledge domain adaptation. In the second training phase, we froze the encoder trained in the first phase and used the LVM + encoder structure to transfer the pixel features to the medical domain features as inputs to an MLP to predict the mRALE score. The whole architecture is illustrated in Fig. 3.

2.6. Model training

We first randomly separated 200 images from the dataset as the holdout test set for evaluation. The remaining 2399 images were used for model training and for 5-fold cross-validation. All CXR images are independent without duplicates. Since our goal is to assess the LVMs capacity of extracting significant patterns for the quantitative analysis of SARS-CoV-2 pneumonia, we used the pretrained LVMs as primary feature extractors and respectively applied two training strategies for feature domain adaptation. In the first strategy, we connected a pretrained LVM as the primary pattern extractor to optimize the MLP encoder using SimSiam self-supervised learning. After optimization, the LVM + encoder will output a 2048-dimensional feature vector as the input to the regression model composed of two 512-dimensional fully connected dense layers with batch normalization for the mRALE score prediction (Fig. 4(a)). In the second strategy, we directly used the pretrained LVM as the primary pattern extractor, which will output a 2048-dimensional feature vector to the regression model. (Fig. 4(b)).

The LVMs we used for the experiments included ViT [17], DINOv2 [11], and Big Transfer (BiT) [23]. We also used the ResNet-50 model by self-supervised learning in our previous experiments [22] for comparison. Routine data augmentation techniques such as random image flipping, rotation, and color jittering were applied to the training procedure of SimSiam self-supervised learning (Fig. 4(a)). All models were trained with the Adam (adaptive moment estimation) optimizer [24] with an initial learning rate of 1×10^{-3}. The experiments were performed on AWS SageMaker Studio with a ml.g4dn instance with a Nvidia T4 GPU. The SimSiam self-supervised encoders were trained by 40 epochs using the negative cosine similarity loss objective. The final MLP regression models were trained for 50 epochs with the mean of squared error (MSE) objective loss. The mini-batch size was 32.

3. Results

3.1. Comparison of prediction performance

We respectively used the holdout test set and performed 5-fold cross-validation to evaluate the overall performance of the LVM-based models, which respectively used the pretrained ViT, DINOv2, and BiT as feature extractors. The pretrained ResNet-50 model from our previous study [22] was added as the baseline for comparison. The MSE and mean absolute error (MAE) of the predictions using different methods are presented in Table 1.

Table 1 indicates the model architecture using ViT as the primary feature extractor with SimSiam self-supervised learning for domain adaptation has the best performance compared to other methods ($p < 0.01$). The BiT optimized by self-supervised learning for domain adaptation has the second-best performance. This finding reflects that the

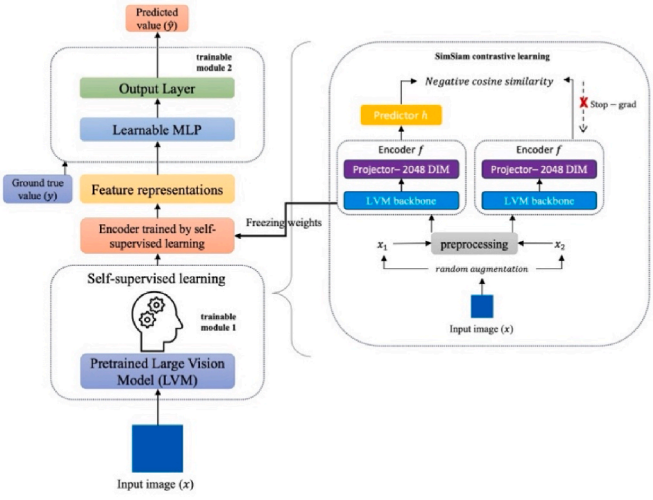

Fig. 3. Self-supervised learning of LVM for knowledge domain adaptation.

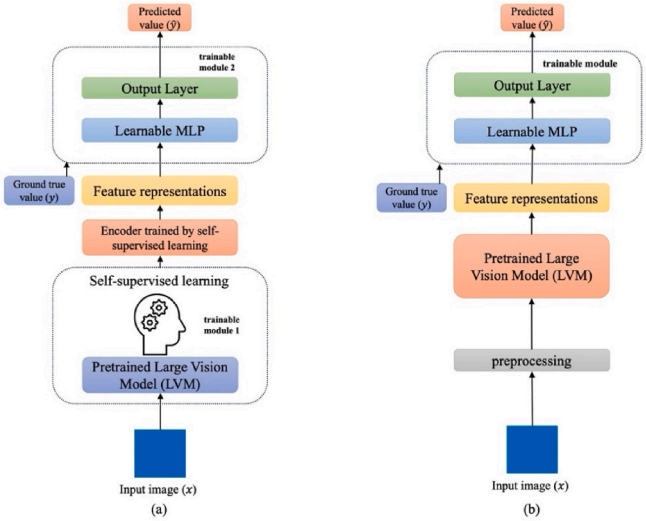

Fig. 4. LVM training strategies to quantify SARS-CoV-2 pneumonia severity.

Table 1
Comparison of mRALE score prediction errors.

Extraction Method	MSE (95 % CI)	MAE (95 % CI)
ViT		
Hold-out test set		
Direct feature extraction	**15.59 (15.17–16.01)** *	**2.97 (2.93–3.01)** *
Self-supervised Learning	20.91 (20.73–21.09) *	3.62 (3.61–3.64) *
5-fold cross-validation		
Direct feature extraction	24.43 (21.78–27.09)	3.83 (3.58–4.07)
Self-supervised Learning	**23.83 (22.67–25.00)**	**3.64 (3.54–3.73)**
DINOv2		
Hold-out test set		
Direct feature extraction	33.09 (30.82–35.36)	4.65 (4.49–4.81)
Self-supervised Learning	35.11 (34.83–35.39)	4.48 (4.46–4.50)
5-fold cross-validation		
Direct feature extraction	40.76 (26.36–55.17)#	5.11 (4.09–6.14)#
Self-supervised Learning	40.77 (38.98–42.56) ▼	5.07 (4.97–5.16) ▼
BiT		
Hold-out test set		
Direct feature extraction	46.24 (45.75–46.73) *	5.87 (5.84–5.90) *
Self-supervised Learning	23.60 (23.48–23.72) *	3.63 (3.61–3.65) *
5-fold cross-validation		
Direct feature extraction	51.02 (47.90–54.14) *▼	6.07 (5.86–6.27) *▼
Self-supervised Learning	26.44 (25.69–27.20) *▼	3.90 (3.75–4.04) *▼
ResNet-50 based self-supervised learning		
Hold-out test set		
Trained from scratch	71.76 (71.66–71.85) *	5.70 (5.69–5.71) *
ImageNet pretrained	68.12 (68.06–68.18) *	5.22 (5.21–5.23) *
5-fold cross-validation		
Trained from scratch	71.25 (70.66–71.84) *▼	5.71 (5.68–5.76) *▼
ImageNet pretrained	66.67 (66.29–67.06) *▼	5.05 (5.03–5.08) *▼

Internal comparison: * $p < 0.01$.
External comparison: # $p < 0.05$, ▼ $p < 0.01$.

LVM models outperform the baseline CNN model using ResNet-50 as a feature extractor. Another notable finding is that the performance of the DINOv2 model is stable and not affected by training methods.

3.2. Comparison of model correlation

The correlation of the model predictions to the expert annotations was evaluated with three metrics: explained variance, coefficient of determination (R^2) and Spearman's rank correlation coefficient (Spearman ρ). Higher values indicate better prediction consistency with the expert knowledge (Table 2).

The results in Table 2 indicates the predictions by LVMs are better correlated to the mRALE scores graded by human experts compared to the baseline ResNet-50 method. The ViT methods have the best performance. However, compared to the reduction of prediction errors (Table 1), using self-supervised learning cannot effectively improve the correlation of the model predictions to the grading by human experts.

Table 2
Comparison of model prediction correlation.

Extraction method	Explained Variance (95 % CI)	Coefficient of determination (R^2) (95 % CI)	Spearman ρ (95 % CI)
ViT			
Hold-out test set			
Direct feature extraction	0.75 (0.74–0.76) *	0.74 (0.73–0.74) *	0.75 (0.74–0.75) *
Self-supervised Learning	0.70 (0.69–0.70) *	0.68 (0.67–0.68) *	0.78 (0.77–0.78) *
5-fold cross-validation			
Direct feature extraction	**0.67 (0.63–0.70)**	0.64 (0.62–0.67)	**0.81 (0.79–0.82)**
Self-supervised Learning	0.66 (0.63–0.68)	**0.65 (0.63–0.67)**	0.80 (0.77–0.81)
DINOv2			
Hold-out test set			
Direct feature extraction	0.52 (0.50–0.53) *	0.49 (0.45–0.53) *	0.69 (0.68–0.70) *
Self-supervised Learning	0.34 (0.33–0.35) *	0.30 (0.29–0.31) *	0.56 (0.55–0.56) *
5-fold cross-validation			
Direct feature extraction	0.51 (0.42–0.60)#	0.37 (0.13–0.61)#	0.70 (0.64–0.77)#
Self-supervised Learning	0.41 (0.36–0.45) ▼	0.38 (0.33–0.44) ▼	0.64 (0.62–0.67) ▼
BiT			
Hold-out test set			
Direct feature extraction	0.29 (0.28–0.30) *	0.25 (0.24–0.27) *	0.56 (0.55–0.57) *
Self-supervised Learning	0.63 (0.62–0.63) *	0.63 (0.62–0.63) *	0.73 (0.72–0.73) *
5-fold cross-validation			
Direct feature extraction	0.27 (0.24–0.30) ▼	0.26 (0.22–0.29) ▼	0.53 (0.50–0.57) ▼
Self-supervised Learning	0.62 (0.60–0.64) *#	0.61 (0.59–0.63) *#	0.79 (0.76–0.81) *
ResNet-50 based self-supervised learning			
Hold-out test set			
Trained from scratch	−0.15 (−0.15 to −0.14) *	0.02 (0.01–0.02) *	0.43 (0.42–0.44) *
ImageNet pretrained	0.14 (0.13–0.14) *	−0.09 (−0.09 to −0.08) *	0.68 (0.67–0.68) *
5-fold cross-validation			
Trained from scratch	0.02 (0.01–0.03)* ▼	−0.13 (−0.14 to −0.12)* ▼	0.52 (0.45–0.59) *▼
ImageNet pretrained	0.21 (0.20–0.21) ▼	−0.07 (−0.07 to −0.06) ▼	0.77 (0.75–0.79)▼

Internal comparison: * $p < 0.01$.
External comparison: # $p < 0.05$,▼external comparison: $p < 0.01$.

3.3. Explainable AI visualization

Most of the LVMs such as ViT and DINO use the vision transformer architecture with multiple self-attention heads for learning both global and local visual patterns to improve the model inference performance. We use the attention heatmaps [25] to visualize the attention values of the attention heads at the last transformer block of the LVMs. In Fig. 5, we compared the attention distribution of the 12 attention heads of the ViT and DINO. It shows the LVMs distribute the focus of the attention heads globally on the images. This finding helps to explain why LVM models can outperform the conventional CNN models for medical image pattern detections. Note that BiT is a residual CNN architecture [23], thus it does not use the multi-head attention mechanism to capture visual patterns.

4. Discussions

The success of large language models (LLMs) in natural language processing (NLP) in the AI industry since 2023 has inspired the research interest of large vision models (LVMs) for multiple AI applications [26]. LVMs pretrained by large image datasets have possessed the ability to understand visual patterns and to capture semantic information from the images. The proposed knowledge adaptation by self-supervised learning succeeds in leveraging the LVMs' capacity to learn the visual representations of medical knowledge from images of radiology, histology, anatomy, etc., which becomes a promising field of medical AI research.

Compared to other SARS-CoV-2 pneumonia severity quantification AI methods such as the SiameseNet-based model by Li et al. [18], and to the transformer-based model by Park et al. [10], our LVM-based self--supervised knowledge adaptation technique acquired similar performance with lower training cost. The latter models were trained by multiple large datasets such as CheXpert containing 224,316 CXR images. Therefore, using pretrained LVMs provides a shortcut for AI development.

The strength of LVM has two folds. First, the LVM + self-supervised encoder architecture offers a novel data augmentation method for the fine-tuning of LVM, in which the newly acquired knowledge is stored in a shallow model apart from the original LVM. It helps keep the integrity of the LVM's original capacity for visual pattern capturing and prevent overfitting during model fine-tuning. Second, the self-supervised training strategy significantly reduces the amount of training examples for high performance as the new model only used a training set with 2399 images.

The limitation of the LVM knowledge adaption is that the model performance relies on the proper pre-training of the original LVM. The DINOv2 model is an efficient model for general-purposed visual pattern extraction, but its performance cannot be effectively improved by the self-supervised encoder. It can be attributed to the pretraining of DINOv2, which is also by self-supervised learning. It implies that there is a limitation for self-supervised knowledge adaptation method being further applied to secondary self-supervised learning. Another limitation is the lack of third-party validation, though our baseline method was tested externally by MIDRC with a prediction probability concordance of 0.811 and a quadratic weighted kappa of 0.739 [22], more third-party validations need to be performed in the future.

5. Conclusions

We proposed a novel knowledge domain adaptation method utilizing the pretrained large vision models (LVM) to transfer the generalized pixel features to the CXR domain-specific features by self-supervised learning. The experiment results showed the new method can achieved comparable SOTA performance to quantify the severity of SARS-CoV-2 pneumonia from CXR images using the mRALE score prediction.

The experiment showed the ViT model achieved the best performance at quantifying the severity of SARS-CoV-2 pneumonia from CXR

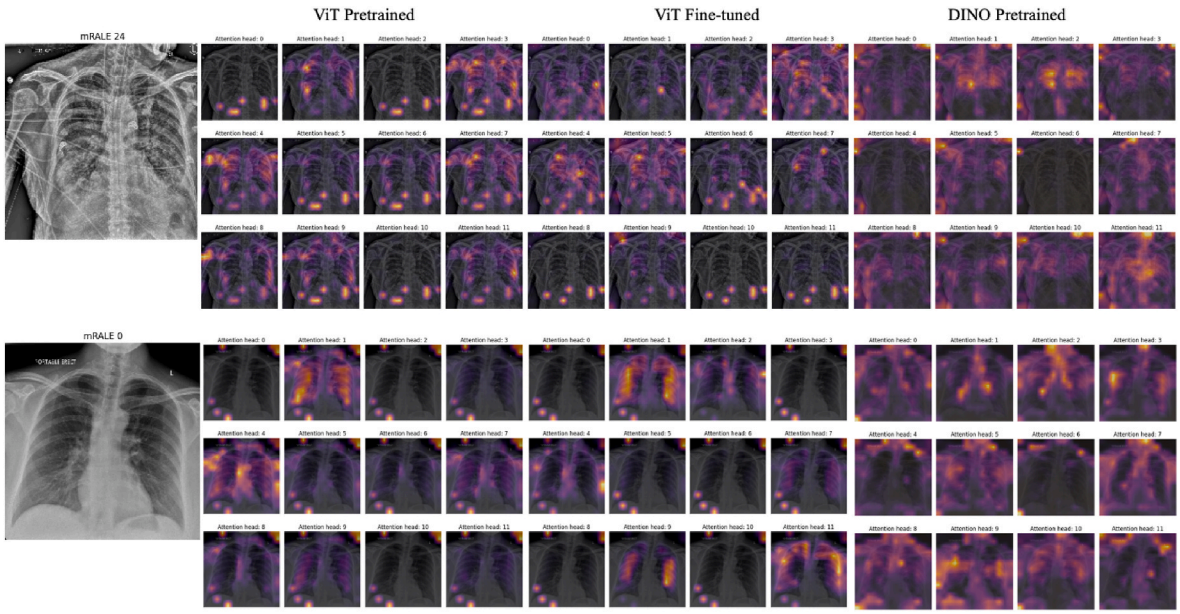

Fig. 5. Distribution of the attentions of LVM.

images based on the prediction of mRALE score. The comparison also indicates the proposed knowledge domain adaptation method by self-supervised learning can further improve the ViT capacity for medical significant pattern extractions given the pretrained LVM was not pre-trained by self-supervised learning. However, the self-supervised-based knowledge domain adaptation method has less impact on enhancing the AI prediction correlation to human expertise. The findings provide a comprehensive guideline for the use of LVM in medical AI developments.

CRediT authorship contribution statement

Zhaohui Liang: Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Zhiyun Xue:** Writing – review & editing, Supervision. **Sivaramakrishnan Rajaraman:** Writing – review & editing, Resources. **Sameer Antani:** Writing – review & editing, Supervision, Resources, Funding acquisition.

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.

Acknowledgment

This work was supported by the Intramural Research Program of the National Library of Medicine, National Institutes of Health.

References

[1] WHO, WHO Coronavirus (COVID-19) Dashboard, https://covid19.who.int/, last accessed 2024/July/10.
[2] National Institutes of Health. NIH covid-19 treatment guidelines. Clinical spectrum of SARS-CoV-2 Infection. 2020. https://www.covid19treatmentguidelines.nih.gov/overview/clinical-spectrum, last accessed 2024/December/27.
[3] Yang X, Yu Y, Xu J, Shu H, Xia J, Liu H, et al. Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: a single-centered, retrospective, observational study. Lancet Respir Med 2020 May;8(5):475–81.
[4] Wu C, Chen X, Cai Y, Xia J, Zhou X, Xu S, et al. Risk factors associated with acute respiratory distress syndrome and death in patients with coronavirus disease 2019 pneumonia in wuhan, China. JAMA Intern Med 2020 Jul 1;180(7):934–43.
[5] Attaway AH, Scheraga RG, Bhimraj A, Biehl M, Hatipoglu U. Severe covid-19 pneumonia: pathogenesis and clinical management. BMJ 2021 Mar 10;372:n436.
[6] Valk CMA, Zimatore C, Mazzinari G, Pierrakos C, Sivakorn C, Dechsanga J, et al. The prognostic capacity of the radiographic assessment for lung edema score in patients with COVID-19 acute respiratory distress syndrome-an international multicenter observational study. Front Med 2022 Jan;8:772056.
[7] Aggarwal P, Mishra NK, Fatimah B, Singh P, Gupta A, Joshi SD. COVID-19 image classification using deep learning: advances, challenges and opportunities. Comput Biol Med 2022 May;144:105350.
[8] Bernheim A, Mei X, Huang M, Yang Y, Fayad ZA, Zhang N, et al. Chest CT findings in coronavirus disease-19 (COVID-19): relationship to duration of infection. Radiology 2020 Jun;295(3):200463.
[9] Cozzi D, Albanesi M, Cavigli E, Moroni C, Bindi A, Luvara S, et al. Chest X-ray in new Coronavirus Disease 2019 (COVID-19) infection: findings and correlation with clinical outcome. Radiol Med 2020 Aug;125(8):730–7.
[10] Park S, Kim G, Oh Y, Seo JB, Lee SM, Kim JH, et al. Multi-task vision transformer using low-level chest X-ray feature corpus for COVID-19 diagnosis and severity quantification. Med Image Anal 2022 Jan;75:102299.
[11] Oquab M, Darcet T, Moutakanni T, Vo H, Szafraniec M, Khalidov V, et al. Dinov2: learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 2024 Jan;01:1–32.
[12] Voigt I, Mighali M, Manda D, Aurich P, Bruder O. Radiographic assessment of lung edema (RALE) score is associated with clinical outcomes in patients with refractory cardiogenic shock and refractory cardiac arrest after percutaneous implantation of extracorporeal life support. Intern Emerg Med 2022 Aug;17(5):1463–70.
[13] Warren MA, Zhao Z, Koyama T, Bastarache JA, Shaver CM, Semler MW, et al. Severity scoring of lung oedema on the chest radiograph is associated with clinical outcomes in ARDS. Thorax 2018 Sep;73(9):840–6.
[14] MIDRC. MIDRC mRALE Mastermind Challenge: AI to predict COVID severity on chest radiographs. https://www.midrc.org/mrale-mastermind-2023, last accessed 2023/July/2.
[15] Khattab R, Abdelmaksoud IR, Abdelrazek S. Deep convolutional neural networks for detecting COVID-19 using medical images: a survey. New Generat Comput 2023 Jun;41(2):343–400.
[16] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. Adv Neural Inf Process Syst 2017;30.
[17] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J. An image is worth 16x16 words: transformers for image recognition at scale. 2020 Oct 22. arXiv preprint arXiv:2010.11929.
[18] Li MD, Arun NT, Gidwani M, Chang K, Deng F, Little BP, et al. Automated assessment and tracking of COVID-19 pulmonary disease severity on chest radiographs using convolutional siamese neural networks. Radiol Artif Intell 2020 Jul;2(4):e200079.
[19] Huang L, Zhu E, Chen L, Wang Z, Chai S, Zhang B. A transformer-based generative adversarial network for brain tumor segmentation. Front Neurosci 2022;16:1054948.
[20] Atabansi CC, Nie J, Liu H, Song Q, Yan L, Zhou X. A survey of Transformer applications for histopathological image analysis: new developments and future directions. Biomed Eng Online 2023 Sep;22(1):96.
[21] Chen X, He K. Exploring simple siamese representation learning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); Nashville, TN, USA; 2021. p. 15745–53.

[22] Liang Z, Xue Z, Rajaraman S, Feng Y, Antani S. Automatic quantification of COVID-19 pulmonary edema by self-supervised contrastive learning. In: Workshop on medical image learning with limited and noisy data. Cham: Springer Nature Switzerland; 2023 Oct 8. p. 128–37.
[23] Kolesnikov A, Beyer L, Zhai X, Puigcerver J, Yung J, Gelly S, Houlsby N. Big transfer (bit): general visual representation learning. In: Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, proceedings. vol. 16. Springer International Publishing; 2020. p. 491–507.
[24] Diederik P Kingma, Adam Jimmy Ba. A method for stochastic optimization. In: Proceedings of 3rd international conference on learning representations (ICLR 2015); May 7-9, 2015. San Diego, CA, USA.
[25] Chefer H, Gur S, Wolf L. Transformer interpretability beyond attention visualization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2021. p. 782–91.
[26] Wang J, Liu Z, Zhao L, Wu Z, Ma C, Yu S, et al. Review of large vision models and visual prompt engineering. Meta-Radiology; 2023, 100047.