Effect of pre-training scale on intra- and inter-domain transfer for natural and X-Ray chest images

Mehdi Cherti and Jenia Jitsev

{m.cherti,j.jitsev}@fz-juelich.de

Juelich Supercomputing Center, Research Center Juelich

Helmholtz AI

JÜLICH Forschungszentrum | JÜLICH SUPERCOMPUTING CENTRE

HELMHOLTZAI | ARTIFICIAL INTELLIGENCE COOPERATION UNIT

ArXiv:2106.00116

SLAMPAI/large-scale-pretraining-transfer

Motivation

- Neural scaling law studies [1, 3] showing **positive effect of larger scale on transfer learning** focus mostly on in-domain transfer, where source and target data are in close proximity
- Does the positive effect of larger pre-training model and data scale still uphold for transfer when **source and target are far apart**?
- In this work, we conduct a series of large-scale pre-training and transfer experiments where we vary not only model and dataset size during pre-training, but also the domain of the source and the target datasets, being either natural or medical X-Ray chest images

Large scale pre-training

In order to study the effect of model and data scale, we pre-trained **ResNet-50x1** (26M parameters) and **ResNet-152x4** (928M parameters) from [2] either on natural or X-Ray chest image data of various sizes.

Natural image domain

- We pre-trained the models on either **ImageNet-1k** ($\approx$ 1.4 Millions images) or **ImageNet-21k** ($\approx$ 14 Millions images)
- For both datasets, we used a standard supervised classification setup with softmax as an output activation and cross entropy as a loss

Medical image domain

- For the first time, we combine several large public X-Ray chest datasets (**CheXpert**, **MIMIC-CXR**, **NIH ChestX-ray14**, **PadChest**), spanning scales from small ($\approx$ 200k samples) to large ($\approx$ 870k samples) for the pre-training, closing up to ImageNet-1k scale.
- We used sigmoid as an output activation and binary cross entropy loss, in a multi-label classification setting

Fine-tuning and transfer evaluation

- We transfer on both **natural** (CIFAR-10, CIFAR-100, Flowers, Pets) and **medical** (*large:* CheXpert, MIMIC-CXR, NIH ChestX-ray14, PadChest; *small:* COVIDx, Tuberculosis) image datasets
- We consider both **full-shot** as well as **few-shot** transfer, where only few examples per class are used
- For fine-tuning, we use the **BiT-HyperRule** [2] to automatically select hyper-parameters based on target dataset. We measure performance using either **accuracy** or **ROC-AUC** using 5 independent runs with different seeds

Experimental results

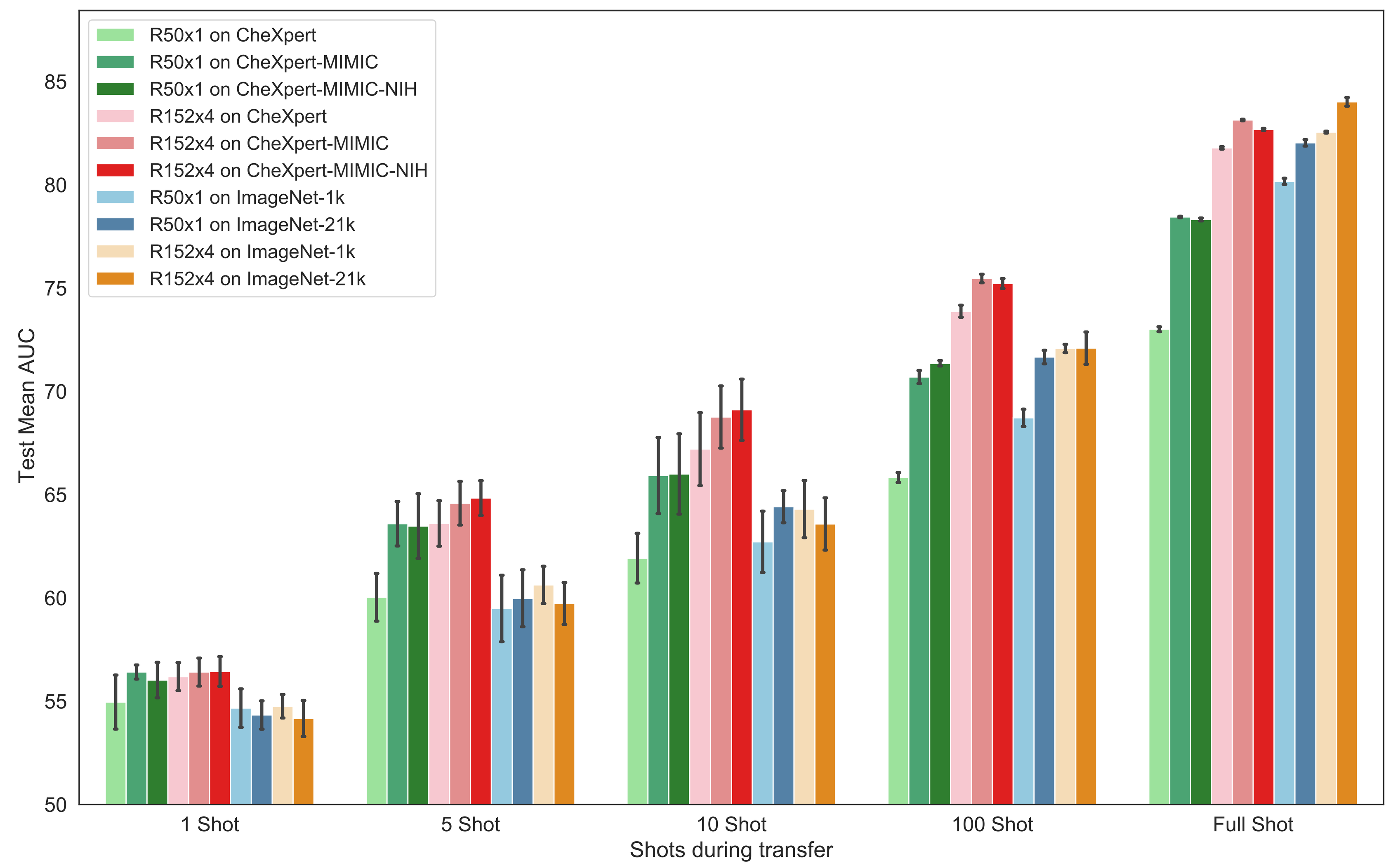

Figure 1: Few- and full-shot transfer performance on *large* X-Ray target PadChest-Cl when varying model and data scale in pre-training

Target	ResNet-50x1				ResNet-152x4			
	S-MED	L-MED	1K-NAT	21K-NAT	S-MED	L-MED	1K-NAT	21K-NAT
CIFAR-10(1)	56.07 ± 0.32	63.27 ± 0.30	94.26 ± 0.05	**95.78 ± 0.09**	74.26 ± 0.20	78.05 ± 0.18	96.93 ± 0.05	**97.82 ± 0.07**
CIFAR-100(1)	16.64 ± 0.21	18.71 ± 0.15	75.90 ± 0.05	**82.47 ± 0.21**	36.29 ± 0.29	37.94 ± 0.23	83.90 ± 0.09	**88.54 ± 0.14**
Flowers-102(1)	7.05 ± 0.59	6.96 ± 1.26	74.94 ± 0.99	**98.21 ± 0.22**	25.19 ± 0.78	23.91 ± 0.86	89.41 ± 0.25	**99.49 ± 0.08**
Pets(1)	7.06 ± 0.46	7.88 ± 0.42	85.21 ± 0.58	**87.23 ± 0.18**	15.07 ± 0.18	16.78 ± 0.35	*93.32 ± 0.30*	*93.21 ± 0.14*
COVIDx(1)	68.50 ± 0.18	76.05 ± 0.21	76.30 ± 1.30	**78.35 ± 1.63**	78.65 ± 0.84	**83.00 ± 1.16**	78.10 ± 0.95	78.90 ± 0.49
Tuberculosis(1)	79.83 ± 0.45	81.65 ± 0.91	79.83 ± 1.50	**83.47 ± 0.83**	79.01 ± 0.45	**90.91 ± 0.83**	81.49 ± 2.23	80.83 ± 2.51
MIMIC CXR(2)	84.17 ± 0.03	86.38 ± 0.03	85.41 ± 0.10	**86.82 ± 0.10**	87.63 ± 0.04	**88.00 ± 0.03**	86.85 ± 0.06	87.79 ± 0.13
CheXpert(2)	82.10 ± 0.07	*86.66 ± 0.05*	84.83 ± 0.14	*86.60 ± 0.14*	84.92 ± 0.07	*87.82 ± 0.03*	86.82 ± 0.06	*87.77 ± 0.07*
PadChest(2)	68.06 ± 0.24	68.14 ± 0.21	76.72 ± 0.27	**80.99 ± 0.22**	75.91 ± 0.12	75.23 ± 0.17	79.59 ± 0.17	**83.94 ± 0.19**
PadChest-Cl(2)	73.01 ± 0.13	78.33 ± 0.08	80.17 ± 0.17	**82.03 ± 0.17**	81.79 ± 0.07	82.68 ± 0.05	82.55 ± 0.05	**84.02 ± 0.24**
NIH CXR(2)	70.11 ± 0.15	74.21 ± 0.57	75.53 ± 0.47	**81.02 ± 0.57**	77.95 ± 0.13	78.95 ± 0.13	79.82 ± 0.38	**82.80 ± 0.41**

Table 1: Full shot intra- and inter-domain transfer **Bold** indicates best transfer performance for a fixed network size. *Italics* indicates transfer performance with no significant difference between data scale. Red indicates best overall performance for a given target

Effect of scale on intra-domain transfer (natural-natural, medical-medical)

- We observe consistent improvement with model scale
- For data scale, we observe improvement on almost all cases
- Few-shot transfer: strong improvement on natural-natural, but no improvement on medical-medical scenario

Effect of scale on inter-domain transfer (natural-medical)

- For large medical X-Ray targets, clear full-shot transfer improvement due to larger pre-training scale, both for model and data scale
- No improvement for small medical X-Ray targets
- Few-shot transfer: no improvement

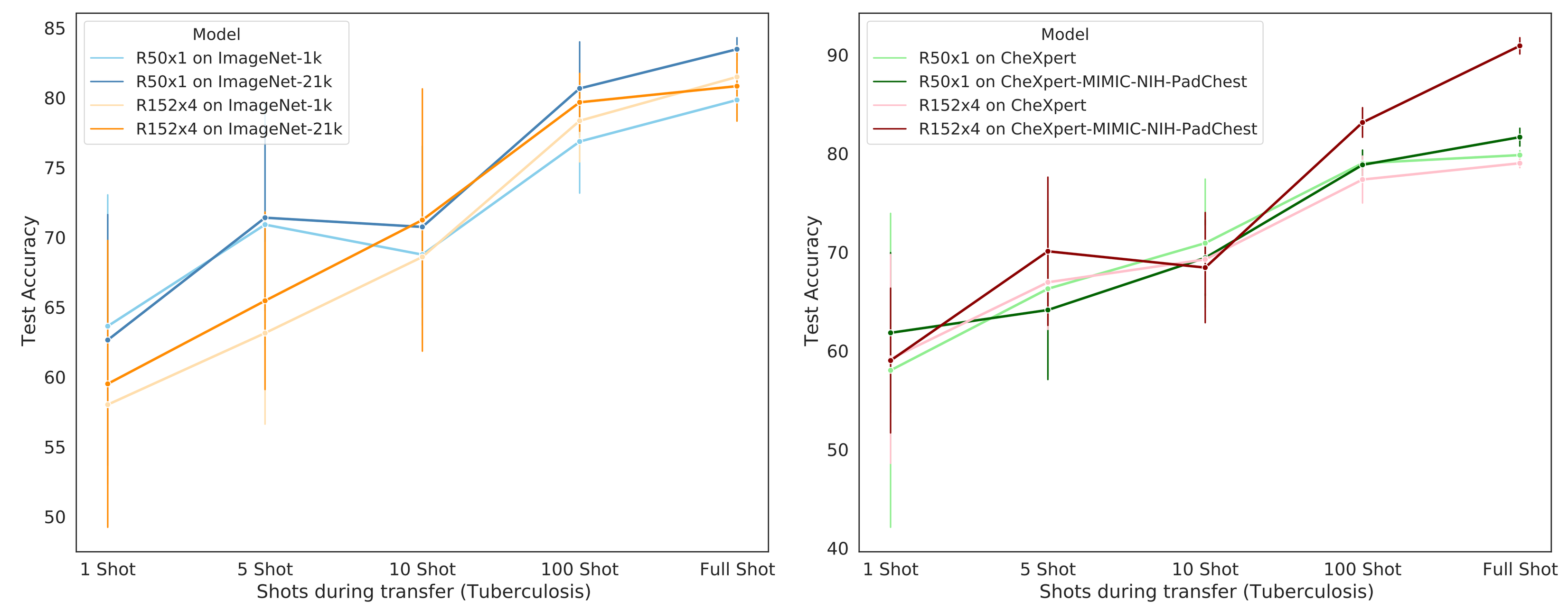

Figure 2: Few- and full-shot transfer performance on *small* X-Ray target Tuberculosis. **left**: pre-training on natural data, **right**: pre-training on medical data.

Conclusion & Outlook

- **Substantially increasing model and data scale in the pre-training** provides benefits for both intra- and inter-domain transfer. Effect of pre-training scale is differential depending on transfer setting
- Remarkably, when comparing **natural-medical** and **medical-medical** transfer on **large X-Ray targets**, we observe that the large ResNet-152x4 pre-trained on **generic** Imagenet-21k can be **as good or even better** than networks pre-trained on largest available **domain-specific** X-Ray data
- This indicates that high quality models for large X-Ray targets can be also obtained by substantially increasing generic natural image source data scale when using large networks. This is relevant for the practice, where large amount of medical domain-specific data is often not available for pre-training.
- In contrast to large X-Ray targets, in **small X-Ray targets** (COVIDx and Tuberculosis), we do not observe improvement with larger pre-training scale when using natural images as source data
- Scaling model and data size by going **beyond ImageNet-21k** may improve transfer further, and show also benefits for few-shot transfer and on small X-Ray targets
- **Combining natural and medical data** in the pre-training may offer further opportunity to amplify effect of scale.
- What may happen to scale effect on intra- or inter-domain transfer when employing **self-supervised learning** for pre-training?

References

[1] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[2] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 491–507, Cham, 2020. Springer International Publishing.

[3] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. *arXiv preprint arXiv:2106.04560*, 2021.