A Strong Baseline for Domain Adaptation and Generalization in Medical Imaging

Li Yao LI@ENLITIC.COM
Jordan Prosky PROSKY@ENLITIC.COM
Ben Covington BEN@ENLITIC.COM
Kevin Lyman KEVIN@ENLITIC.COM

Abstract

This work provides a strong baseline for the problem of multi-source multi-target domain adaptation and generalization in medical imaging. Using a diverse collection of ten chest X-ray datasets, we empirically demonstrate the benefits of training medical imaging deep learning models on varied patient populations for generalization to out-of-sample domains.

1. Introduction

Recent advancement in machine learning has created a surge in developing neural-network-based computer-aided diagnostic algorithms (Mazurowski et al., 2018). Despite widely acclaimed performance accompanied by rigorous regulatory assessments, most models rarely make their way into real world clinical environments. One major barrier that stops the successful technology transfer is that models do not generalize well to diverse patient populations. The situation is further aggravated by differences in acquisition parameters and manufacturing standards of medical devices. It is therefore not surprising that models trained on data from particular institutions perform well on in-domain validation sets while inevitably suffer from performance degradation when used in other domains. We refer to this as the problem of domain over-fitting.

Machine learning research has produced strategies and algorithms to mitigate domain over-fitting with the study of domain adaptation (DA) (Wang and Deng, 2018) and domain generalization (DG) (Li et al., 2017). Modeling in medical imaging, however, comes with unique challenges that are not faced in day-to-day computer vision tasks. For instance, medical images are typically of much higher resolution in 2D (and often are 3D or 4D), contain subtle artifacts, and have small regions of interest. Moreover, the interpretation of medical images can involve a high degree of uncertainty, even for highly-trained radiologists.

Reliable predictive modeling in medical imaging calls for remedies from DA and DG. This work illustrates the problem of domain over-fitting in the context of classifying chest X-rays (CXRs), the most commonly prescribed imaging exams worldwide. Experimental data are gathered from ten domains varied by their patient distributions, clinical environments, and global locations. We empirically show the phenomenon of performance degradation with inter-domain generalization. In this preliminary work, we suggest a simple solution which quantitatively shows its promise as a strong baseline for better generalization.

Related work. High performance in classification, detection and segmentation is regularly observed in retrospective clinical studies and publications. For instance, an AUC of 0.99 was reported in Lakhani and Sundaram (2017) for classifying pulmonary tuberculosis from CXRs. An average AUC of 0.96 was recorded in Dunnmon et al. (2018) in triaging normal and abnormal CXRs. A DICE score of 0.98 was shown in Weston et al. (2018) in segmenting body parts in abdominal CTs. A sensitivity of 0.96 was shown in Thian et al. (2019) in detecting fracture in wrist X-rays. Ueda et al. (2018) reported a sensitivity of 0.93 in detecting cerebral aneurysms in head MR angiography. As Kim et al. (2019) pointed out, however, most clinical publications do not contain a sufficient amount of external validation that is beyond the source domain, whose data is used to train the models. Among those that did, the recent work of Zech et al. (2018) empirically showed drastic performance gaps of models across three medical institutions in classifying pneumonia in CXRs. The work of Prevedello et al. (2019) also discussed a similar issue of coping with data heterogeneity, but offered no practical recommendations. Unlike previous work, we conduct an unprecedented study with ten datasets collected internationally, measuring the ability of state-of-the-art machine learning models to perform domain adaptation and domain generalization in the context of medical imaging. We establish baseline solutions that are intuitive and practical, and lead to better generalization performance in experiments.

2. Experiments

Data. We utilize ten datasets from diverse sources to empirically show the benefits of training with data from multiple domains for model generalization. For training, we use four publicly available datasets: ChestX-ray14 (Wang et al., 2017) (NIH), CheXpert (Irvin et al., 2019) (CHX), PadChest (Bustos et al., 2019) (PAD), Mimic-CXR (Johnson et al., 2019) (MIM), and one private data set from Australia (AUS). In addition, to evaluate generalization, we use one public data set, Open-i (Demner-Fushman et al., 2015) (OPI), and four private sets - one from Canada (CAN) and three from different sources in China (CHN_1, CHN_2, CHN_3). Table 1 below summarizes the data used in our experiments.

Table 1: Summary of our CXR data. We use a random 80/20 patient split when applicable.

Dataset	Origin	# Patients	# Train Scans	# Test Scans
NIH	Bethesda, MD, USA	30,806	89,322	22,798
CHX	Stanford, CA, USA	64,534	152,938	38,089
PAD	Alicante, Spain	67,216	88,207	22,347
MIM	Boston, MA, USA	62,592	200,874	49,170
AUS	Australia	125,000	100,000	25,000
OPI	Bloomington, IN, USA	3,670	-	3,670
CAN	Canada	18,000	-	19,000
CHN_1	China	7,000	-	7,000
CHN_2	China	3,000	-	3,000
CHN_3	China	2,500	-	2,500

For all of the following experiments, we use a DenseNet-121 pretrained on ImageNet, and we are concerned with classifying CXRs as normal or abnormal. In the first set of

experiments, we train a model on each of the five training datasets and test on all ten sets. Table 2 shows AUCs from training on each source domain and evaluating on all target domains. We notice a couple of nice effects of training on all source domains. First, when aggregating all source domain data for training, test performance on those domains is essentially as good as training on any single source. Moreover, training on all sources simultaneously results in consistent improvement and yields the best performance on each of the five target domains, on which models are never trained on.

Table 2: AUCs on target domains when trained on different source domains.

		Source Domain					
		NIH	CHX	PAD	MIM	AUS	ALL 5
Target Domain	NIH	**0.769**	0.732	0.751	0.756	0.744	**0.769**
	CHX	0.823	**0.866**	0.816	0.851	0.782	**0.862**
	PAD	0.811	0.807	**0.853**	0.803	0.832	**0.850**
	MIM	0.814	0.825	0.793	**0.854**	0.783	**0.853**
	AUS	0.795	0.776	0.808	0.769	**0.848**	**0.841**
	OPI	0.758	0.744	0.783	0.723	**0.791**	**0.786**
	CAN	0.772	**0.789**	0.783	0.783	**0.787**	**0.788**
	CHN_1	0.749	0.744	0.773	0.754	0.781	**0.835**
	CHN_2	0.771	0.725	0.770	0.716	0.786	**0.805**
	CHN_3	0.736	0.694	0.762	0.710	**0.772**	**0.772**

Figure 1 shows AUCs for experiments where one of the five source domains is left out during training. We observe that for NIH and CHX, leaving them out has a negligible impact on the model's performance. For AUS, PAD, and MIM, however, we notice what we expect: a moderate decrease in performance when the source is held out during training. There are many possible reasons to explain why performance is hurt more for some domains than for others, which we leave for further research.

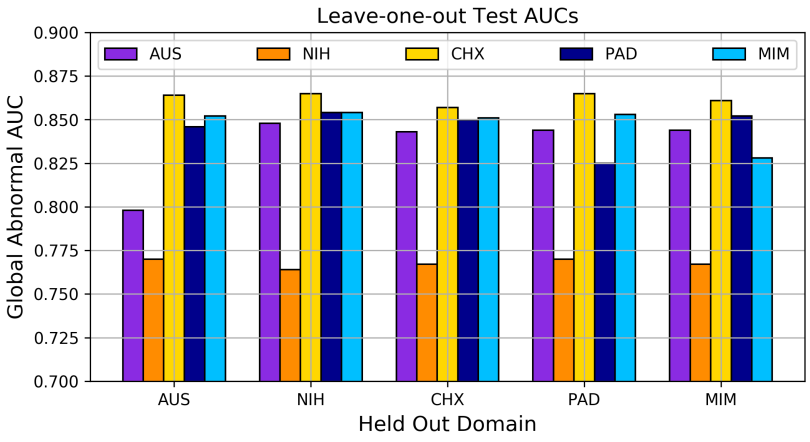

Figure 1: Varied performance of DG with leave-one-domain-out training.

References

Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria de la Iglesia-Vay. Padchest: A large chest x-ray image dataset with multi-label annotated reports, 2019.

Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza Rodriguez, Sameer Antani, George R. Thoma, and Clement J. Mcdonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304310, 2015. doi: 10.1093/jamia/ocv080.

Jared A Dunnmon, Darvin Yi, Curtis P Langlotz, Christopher Ré, Daniel L Rubin, and Matthew P Lungren. Assessment of convolutional neural networks for automated classification of chest radiographs. *Radiology*, page 181422, 2018.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison, 2019.

Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih ying Deng, Roger G. Mark, and Steven Horng. Mimic-cxr: A large publicly available database of labeled chest radiographs, 2019.

Dong Wook Kim, Hye Young Jang, Kyung Won Kim, Youngbin Shin, and Seong Ho Park. Design characteristics of studies reporting the performance of artificial intelligence algorithms for diagnostic analysis of medical images: Results from recently published papers. *Korean journal of radiology*, 20(3):405–410, 2019.

Paras Lakhani and Baskaran Sundaram. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology*, 284(2):574–582, 2017.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5542–5550, 2017.

Maciej A Mazurowski, Mateusz Buda, Ashirbani Saha, and Mustafa R Bashir. Deep learning in radiology: an overview of the concepts and a survey of the state of the art. *arXiv preprint arXiv:1802.08717*, 2018.

Luciano M Prevedello, Safwan S Halabi, George Shih, Carol C Wu, Marc D Kohli, Falgun H Chokshi, Bradley J Erickson, Jayashree Kalpathy-Cramer, Katherine P Andriole, and Adam E Flanders. Challenges related to artificial intelligence research in medical imaging and the importance of image analysis competitions. *Radiology: Artificial Intelligence*, 1 (1):e180031, 2019.

Yee Liang Thian, Yiting Li, Pooja Jagmohan, David Sia, Vincent Ern Yao Chan, and Robby T Tan. Convolutional neural networks for automated fracture detection and localization on wrist radiographs. *Radiology: Artificial Intelligence*, 1(1):e180001, 2019.

Daiju Ueda, Akira Yamamoto, Masataka Nishimori, Taro Shimono, Satoshi Doishita, Akitoshi Shimazaki, Yutaka Katayama, Shinya Fukumoto, Antoine Choppin, Yuki Shimahara, and Yukio Miki. Deep learning for mr angiography: Automated detection of cerebral aneurysms. *Radiology*, 290:180901, 10 2018. doi: 10.1148/radiol.2018180901.

Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, 2018.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. doi: 10.1109/cvpr.2017.369.

Alexander D Weston, Panagiotis Korfiatis, Timothy L Kline, Kenneth A Philbrick, Petro Kostandy, Tomas Sakinis, Motokazu Sugimoto, Naoki Takahashi, and Bradley J Erickson. Automated abdominal segmentation of ct scans for body composition analysis using deep learning. *Radiology*, page 181432, 2018.

John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLoS medicine*, 15(11): e1002683, 2018.