Generalizability Issue of Deep Learning Models in Medicine and Its Potential Solutions: Illustrated with CBCT to CT Image Conversion

JULY 12–16 2020 VIRTUAL
JOINT AAPM | COMP MEETING
EASTERN TIME [GMT-4]

Xiao Liang, Dan Nguyen, Steve Jiang
Medical Artificial Intelligence and Automation (MAIA) Laboratory,
Department of Radiation Oncology,
UT Southwestern Medical Center, Dallas, Texas
Xiao.Liang@UTSouthwestern.edu, Steve.Jiang@UTSouthwestern.edu

MAIA Laboratory
UT Southwestern Medical Center Radiation Oncology

INTRODUCTION

Generalizability is a concern when applying a deep learning (DL) model trained on one dataset to other datasets. It is challenging to demonstrate a DL model's generalizability efficiently and sufficiently before implementing the model in clinical practice. Training a universal model that works anywhere, anytime, for anybody is unrealistic.

AIM

In this work, we demonstrate the generalizability problem, then explore potential solutions based on transfer learning by using the CBCT to CT image conversion task as the testbed.

DATASETS

We split the 7 datasets into source dataset (H&N1 and H&N2) and target dataset (Prostate1, Prostate2, Prostate3, Cervix and Pancreas) to mimic a situation where CBCT scans come from different clinical environments. The numbers of patients and 2D CBCT/CT images in each dataset for training, validation and testing are shown in Table 1.

Table 1. CBCT image datasets used for experiments.

	Vender	Scanner	Scanning Protocol (kVp/mAs)	No. of patients for train/validate/test	No. of images for train/validate/test
H&N1	Varian	OBI	100/150	83/9/23	6640/720/1840
H&N2	Varian	OBI	125/750	11/1/10	880/80/800
Prostate1	Varian	OBI	125/1070	15/3/11	1200/240/880
Prostate2	Elekta	XVI (Versa)	120/1600	15/3/11	1050/210/770
Prostate3	Elekta	XVI (Agility)	120/1600	15/2/10	1035/138/690
Cervix	Elekta	XVI (Agility)	120/1600	15/3/10	1035/207/690
Pancreas	Elekta	XVI (Versa)	120/1600	15/3/10	1050/210/700

METHODS

We first demonstrate the problem of model generalizability, taking CycleGAN as an example (Figure 1).

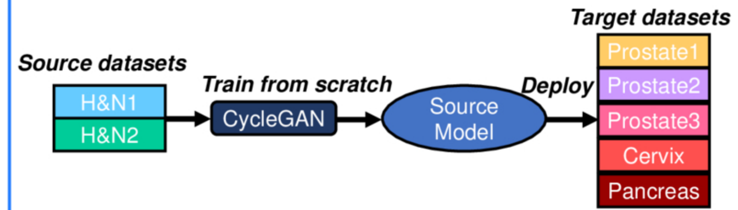

Figure1. Illustration of DL model generalization problem by directly applying a source-dataset-trained DL model to target datasets.

Second we explore different methods to solve this problem (Figure 2). We evaluated the model performance by using mean absolute error (MAE) for measuring the similarity between generated synthetic CT (sCT) and the deformed CT (dCT) images.

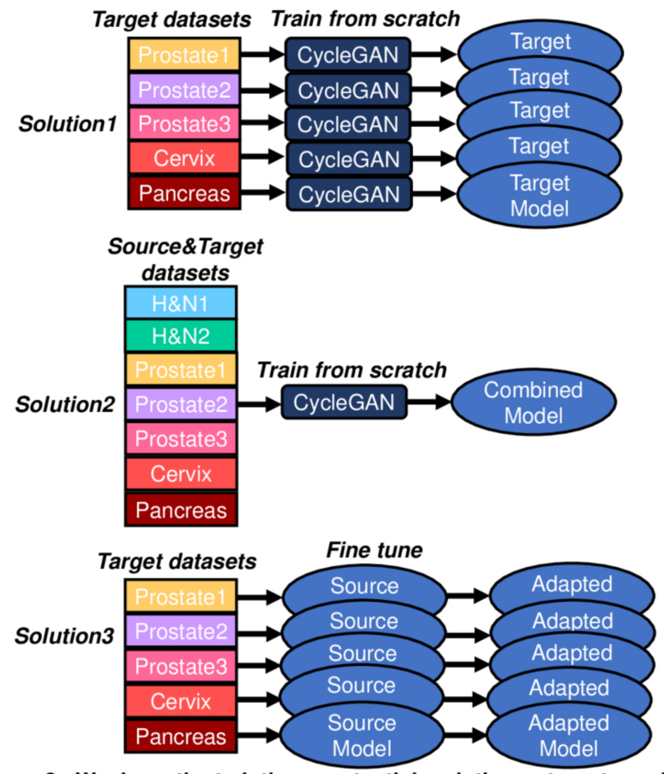

Figure 2. We investigated three potential solutions: target model, combined model and adapted model.

RESULTS

For the Prostate1 dataset, which comes from Varian OBI scanners. Figure 3.a shows similar MAE scores for the source, target, combined, and adapted models. Thus, when applying the source model to a dataset which it has never seen before, but coming from the same vendor's scanners, the source model generates good quality sCT images from CBCT, and the three updated models slightly improve upon this performance.

For the Prostate2, Prostate3, Cervix, and Pancreas datasets, which come from Elekta XVI (Versa) or XVI (Agility) scanners, the source model performed much worse in these target datasets (Figure 3.b-e). Thus, when applying the source model to datasets which it has never seen before and been collected from different anatomical sites and different vendors' scanners, the source model fails to generate good quality sCT images. All three updated models greatly outperform the source model, and the adapted model always performs best.

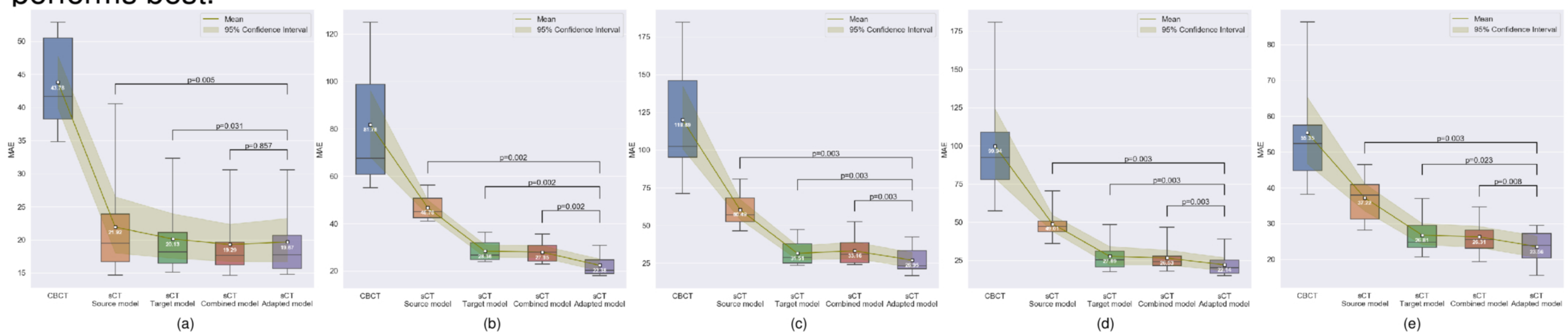

Figure 3. Evaluation of similarity measure of CBCT, sCT (source model), sCT (target model), sCT (combined model), and sCT (adapted model) against dCT on target testing datasets (a) Prostate1, (b) Prostate2, (c) Prostate3, (d) Cervix, and (e) Pancreas. One-tailed Wilcoxon signed rank tests and 95% CIs of the means were calculated.

CONCLUSIONS

In our application, disease site was a minor influence on the source model's performance, but vendor's scanner was a major influence that could dramatically decrease the accuracy of the source model. We found that the adapted model works the best among the three updated models.

REFERENCES:

Liang X, Nguyen D, Jiang S. Generalizability issues with deep learning models in medicine and their potential solutions: illustrated with Cone-Beam Computed Tomography (CBCT) to Computed Tomography (CT) image conversion[J]. arXiv preprint arXiv:2004.07700, 2020.