SIIM20
ANNUAL MEETING
June 24-26, 2020
Austin, Texas
Reimagining the Future.
CELEBRATING 40 YEARS

Title

“Name that manufacturer!”: a simple experiment to show image acquisition bias when training deep learning models

Introduction

Interest in applying machine learning techniques for medical images is growing rapidly and models are starting to be used in clinical practice. Machine learning scientists and clinicians have the vital role of collecting data and recognizing bias and shifts in datasets. It is still unclear how to account for all of the different types of biases that can occur, and what methods can be used to help mitigate them. In binary classification tasks, for example, it is common practice to balance classes across patient gender and age (for anatomy reasons). Other criteria related to image quality[1] (acquisition parameters, manufacturer, software versions) are rarely taken into account.

When performing retrospective studies, data is frequently collected from different sources, in particular for the positives (findings/disease) and negatives (e.g. positives have been acquired in the emergency department which have a different scanner than for outpatients). Can this bias deep learning models and by how much? Following the similar idea as work by Torralba and Efros[2], we propose the game of “Name that manufacturer!” and show that image quality, and in particular how images are acquired can strongly impact and bias deep learning models.

Hypothesis

We show how a deep learning model can be biased by the image acquisition process.

Methods

We trained a Convolutional Neural Network (CNN) to distinguish between two manufacturers on a dataset of Non-Contrast CT (NCCT) heads. Our dataset was composed of 684 NCCT head exams, 342 from GE and 342 from Siemens scanners.

To avoid confounding biases, we first balanced our dataset to ensure that the age and sex distribution was similar across manufacturers. Then we randomly sampled our dataset and split into training 60% (414), validation 30% (204), and test 10% (66) sets.

We preprocessed the data by resizing all series slices to 256x256, windowing (center 40, width 400) and rescaling pixel values to [-1,1].

To verify our hypothesis, we used a simple 3-block CNN: 3 blocks of 3x3 convolution + batch-normalization + max-pooling layers, and at the end of the network a dense layer to classify scanner type.

The input data consisted of 3 axial slices (that the CNN sees as channels) taken from the NCCT series. They were randomly selected and randomly flipped for data augmentation at the end of each epoch of training.

Results

The CNN shows an average accuracy on the test set of $95 \pm 4\,\%$ in being able to determine scanner manufacturer. Errors are equally distributed between classes, age and sex. Table 1 shows some statistics of the dataset, table 2 shows some of our model results testing them on 3 randomly picked slices from the test set.

[1] Kruggel, Frithjof, et al. “Impact of Scanner Hardware and Imaging Protocol on Image Quality and Compartment Volume Precision in the ADNI Cohort.” *NeuroImage*, vol. 49, no. 3, 2010, pp. 2123–2133., doi:10.1016/j.neuroimage.2009.11.006.

[2] Torralba, Antonio, and Alexei A. Efros. “Unbiased Look at Dataset Bias.” *Cvpr 2011*, 2011, doi:10.1109/cvpr.2011.5995347.

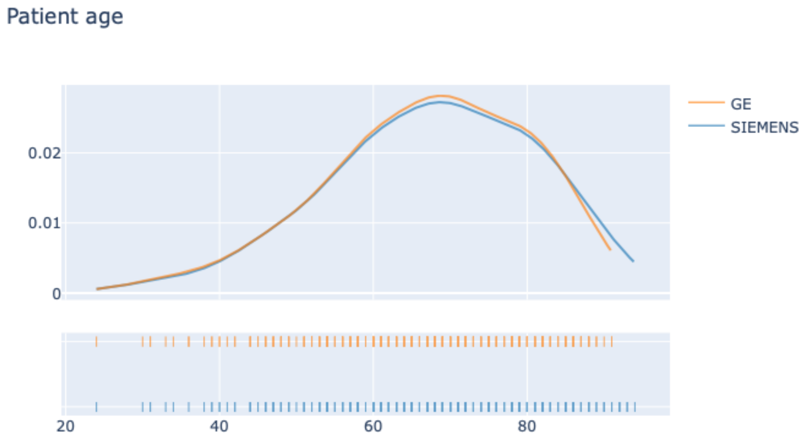

Figure1: age distribution.

	SIEMENS	GE	TOTAL
M	52.6%	57.9%	55..3%
F	47.4%	42.1%	44.7%
NCCT SERIES	50%	50%	100%

Table1: percentage of sex and NCCT images for each class.

Best 2 models	Test accuracy	AUROC	Siemens	GE
Model 1	$98.11 \pm 0.65\%$	0.99 ± 0.01	100%	$96.21 \pm 1.31\%$
Model 2	$98.86 \pm 1.25\%$	0.99 ± 0.01	100%	$97.72 \pm 2.21\%$

Table2: Mean and variance of test accuracy and AUROC testing on 4 different packages of 3 slices (threshold = 0.5).

Conclusion

We demonstrate a significant difference between scanners and their settings by a simple CNN, which implies that it is remarkably easy for any model to be biased if the distribution of targets classes is not equal across manufacturers.

Statement of Impact

Clinicians and machine learning scientists interested in building an image dataset for classification tasks should take into account scanner manufacturer bias in addition to other dataset balance considerations.

Keywords
Distribution shift, Dataset bias, CNN, Classification, CT.