Physics-based data augmentation to improve the generalizability of deep neural networks

Akinyinka O. Omigbodun\textsuperscript{1} \hspace{1cm} AOMIGBODUN@MEDNET.UCLA.EDU
\textsuperscript{1} Department of Radiological Sciences, David Geffen School of Medicine, University of California Los Angeles, 924 Westwood Boulevard, Los Angeles, California 90024, United States of America

Frederic Noo\textsuperscript{2} \hspace{1cm} NOO@UCAIR.MED.UTAH.EDU
\textsuperscript{2} Department of Radiology and Imaging Sciences, The University of Utah, Salt Lake City, Utah 84108, United States of America

Michael McNitt-Gray\textsuperscript{1} \hspace{1cm} MMCNITTGRAY@MEDNET.UCLA.EDU
William Hsu\textsuperscript{1} \hspace{1cm} WHSU@MEDNET.UCLA.EDU
Scott S. Hsieh\textsuperscript{1} \hspace{1cm} SHSIEH@MEDNET.UCLA.EDU

Editors: Under Review for MIDL 2019

Keywords: Augmentation, protocol variation, lung nodule detection

1. Introduction

An important challenge for deep learning models is their ability to generalize to new datasets that may be acquired with an acquisition protocol that differs from the training set. Geometric augmentation (shift, rotation) has become standard in deep learning problems and is known to improve performance by emulating variation in camera position. The purpose of this work is to introduce a new technique, physics-based data augmentation (PBDA), that emulates differences in CT data acquisition and can hence improve the generalizability of neural networks to new datasets. We demonstrate one form of PBDA, emulating increases in slice thickness, on the specific problem of false positive reduction in automatic detection of lung nodules (Nishio et al., 2018; Zhao et al., 2018; Khosravan and Bagci, 2018).

2. Materials and Methods

2.1. Data Preprocessing

The data preprocessing was adapted from the work of Shen (2018). We used 1010 CT scans from the LIDC-IDRI database (Armato III et al., 2011). Only nodules that were annotated by at least 3 (of 4) experienced radiologists were included. We generated false positive (or non-nodule) candidates that did not overlap with true nodules and were within the segmented lungs. Axial, sagittal, and coronal planes through true nodules and false positives were input to the neural network, thereby supplying limited 3D information. Well-established data augmentation approaches such as random rotations, random shifts, and random zoom variation were used with all deep learning models (those with and without PBDA).

2.2. Deep Neural Network Models

The hybrid ensemble CNN (Shen, 2018) comprised a VGG-11 module (Simonyan and Zisserman, 2014), a ResNet-50 module (He et al., 2016), and a DenseNet-40 module (Huang et al., 2017). The ensemble decision was derived by combining output probabilities from the modules; this was used as a mechanism to improve the performance beyond what could be achieved by any component module.

2.3. Physics-based Data Augmentation (PBDA)

To emulate variations in training set data acquisition slice thickness, we simulated thick slice CT scans from thin slice CT scans by grouping contiguous CT slices together. These slices were then averaged together, and each slice was then replaced by the average of the set. While one may also be interested in simulating thinner slice acquisition, this is not easy to perform well without raw data from the scans. Other forms of PBDA not explored here include emulation of lower radiation dose images by injecting colored noise, or emulation of different reconstruction kernels by successive blurring or sharpening of reconstructed images. Figure 1 shows examples of slice thickness PBDA.

2.4. Experiment

Often, our best data is derived from clinical trials with with minimal variation in acquisition protocol, and after training, the deep neural network must then be applied on a broader scale with departures from the original acquisition protocol. To emulate this scenario, we sorted the CT scans according to slice thickness and selected the thinnest 10% of slice thicknesses to constrain the parameter to a small range for training our deep neural network models; we evaluated on the remaining data. For the training data, the number of nodules, number of non-nodules, minimum, maximum, mean, and median slice thicknesses were 143, 2224, 0.6 mm, 1.25 mm, 0.91 mm, and 1 mm respectively, and 1169, 37949, 1.25 mm, 5 mm, 2.02 mm, and 2 mm for the evaluation data. Data used for training and validation was augmented with data with doubled as well as tripled simulated slice thicknesses in two parallel analyses. Because our networks were trained on a small subset of the data, our performance is naturally less than networks trained on the entire LIDC database.

Figure 1: In PBDA, (a, b) two adjacent slices are combined in (c). A CNN input axial plane (d) before and (e) after slice thickening. (f) Free-response ROC curves for the ensemble network. Color fill shows 95% confidence intervals.
3. Results

Following the LUNA16 Challenge, our figure of merit ("Score") was defined as the average sensitivity across a clinically relevant range of the seven false-positive rates shown in Table 1. We observed performance increases with slice thickening PBDA across all three network modules as well as the ensemble. The numbers provided in these tables represent the average of 10 different training runs with 95% confidence intervals for the scores. Figure 1 shows the ensemble network performance.

Table 1: Sensitivity results of the models. The format of each model name is as follows: the neural network name, followed by ‘s’ (if slice thickness PBDA is used), and then an integer to denote the slice thickening factor. The models are scored in a manner similar to the LUNA16 Challenge.

<table>
<thead>
<tr>
<th>FPs/scan</th>
<th>0.125</th>
<th>0.25</th>
<th>0.5</th>
<th>1</th>
<th>2</th>
<th>4</th>
<th>8</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>VGG-11</td>
<td>0.647</td>
<td>0.743</td>
<td>0.813</td>
<td>0.864</td>
<td>0.907</td>
<td>0.940</td>
<td>0.965</td>
<td>0.840±0.013</td>
</tr>
<tr>
<td>VGG-11-s-2</td>
<td>0.727</td>
<td>0.792</td>
<td>0.848</td>
<td>0.888</td>
<td>0.918</td>
<td>0.944</td>
<td>0.968</td>
<td>0.869±0.008</td>
</tr>
<tr>
<td>VGG-11-s-3</td>
<td>0.724</td>
<td>0.795</td>
<td>0.844</td>
<td>0.886</td>
<td>0.917</td>
<td>0.945</td>
<td>0.969</td>
<td>0.869±0.007</td>
</tr>
<tr>
<td>ResNet-50</td>
<td>0.639</td>
<td>0.712</td>
<td>0.776</td>
<td>0.832</td>
<td>0.880</td>
<td>0.921</td>
<td>0.958</td>
<td>0.817±0.015</td>
</tr>
<tr>
<td>ResNet-50-s-2</td>
<td>0.679</td>
<td>0.737</td>
<td>0.794</td>
<td>0.847</td>
<td>0.892</td>
<td>0.931</td>
<td>0.962</td>
<td>0.835±0.008</td>
</tr>
<tr>
<td>ResNet-50-s-3</td>
<td>0.663</td>
<td>0.730</td>
<td>0.788</td>
<td>0.842</td>
<td>0.889</td>
<td>0.929</td>
<td>0.962</td>
<td>0.829±0.007</td>
</tr>
<tr>
<td>DenseNet-40</td>
<td>0.677</td>
<td>0.743</td>
<td>0.806</td>
<td>0.855</td>
<td>0.903</td>
<td>0.940</td>
<td>0.967</td>
<td>0.842±0.015</td>
</tr>
<tr>
<td>DenseNet-40-s-2</td>
<td>0.722</td>
<td>0.781</td>
<td>0.833</td>
<td>0.878</td>
<td>0.915</td>
<td>0.947</td>
<td>0.969</td>
<td>0.864±0.009</td>
</tr>
<tr>
<td>DenseNet-40-s-3</td>
<td>0.718</td>
<td>0.780</td>
<td>0.836</td>
<td>0.883</td>
<td>0.924</td>
<td>0.952</td>
<td>0.975</td>
<td>0.867±0.008</td>
</tr>
<tr>
<td>Ensemble</td>
<td>0.726</td>
<td>0.780</td>
<td>0.837</td>
<td>0.888</td>
<td>0.924</td>
<td>0.950</td>
<td>0.972</td>
<td>0.868±0.007</td>
</tr>
<tr>
<td>Ensemble-s-2</td>
<td>0.751</td>
<td>0.806</td>
<td>0.854</td>
<td>0.898</td>
<td>0.932</td>
<td>0.953</td>
<td>0.973</td>
<td>0.881±0.006</td>
</tr>
<tr>
<td>Ensemble-s-3</td>
<td>0.757</td>
<td>0.807</td>
<td>0.854</td>
<td>0.898</td>
<td>0.932</td>
<td>0.956</td>
<td>0.977</td>
<td>0.883±0.004</td>
</tr>
</tbody>
</table>

Slice thickness PBDA enabled moderate improvements in most tested cases. The ensemble network showed an improvement in average sensitivity of 1.4%. Doubling and tripling the slice thickness yielded similar results.

4. Discussion

In the same way that conventional augmentation improves performance by emulating variation in camera pose, PBDA can improve performance by emulating variations in CT scanner settings. The improvements seen here, between 1 to 2 percent, are modest but widely applicable, as PBDA can be applied on a range of problems using CT data. We chose false positive reduction because it has been well studied and characterized by past researchers. The improvement seen here is comparable to fine-tuned implementations of augmentation in other computer vision problems (Cubuk et al., 2018).
Acknowledgments

The authors would like to thank John Hoffman, Nastaran Emaminejad, Angela Sultan, and Muhammad Wahi-Anwar for their assistance in accessing CT data used in preliminary analyses. The authors would also like to thank Professor Matt Brown for helpful feedback in the course of this project.

References


