The Impact of an Inter-rater Bias on Neural Network Training

Or Shwartzman\textsuperscript{1,2}, Harel Gazit\textsuperscript{1,2}, Ilan Shelef\textsuperscript{2,3,4}, and Tammy Riklin-Raviv\textsuperscript{1,2}

\textsuperscript{1} Electrical and Computer Engineering Department, Ben-Gurion University
\textsuperscript{2} The Zlotowski Center for Neuroscience, Ben-Gurion University
\textsuperscript{3} Department of Radiology, Soroka Medical Center
\textsuperscript{4} Department of Health Sciences, Ben-Gurion University

Abstract. The problem of inter-rater variability is often discussed in the context of manual labeling of medical images. It is assumed to be bypassed by automatic model-based approaches for image segmentation which are considered ‘objective’, providing single, deterministic solutions. However, the emergence of data-driven approaches such as Deep Neural Networks (DNNs) and their application to supervised semantic segmentation - brought this issue of raters’ disagreement back to the front-stage. In this paper, we highlight the issue of inter-rater bias as opposed to random inter-observer variability and demonstrate its influence on DNN training, leading to different segmentation results for the same input images. In fact, lower Dice scores are calculated if training and test segmentations are of different raters. Moreover, we demonstrate that inter-rater bias in the training examples is amplified when considering the segmentation predictions for the test data. We support our findings by showing that a classifier-DNN trained to distinguish between raters based on their manual annotations performs better when the automatic segmentation predictions rather than the raters’ annotations were tested.

For this study, we used the ISBI 2015 Multiple Sclerosis (MS) challenge dataset, which includes annotations by two raters with different levels of expertise\textsuperscript{2}. The results obtained allow us to underline a worrisome clinical implication of a \textit{DNN bias induced by an inter-rater bias} during training. Specifically, we show that the differences in MS-lesion load estimates increase when the volume calculations are done based on the DNNs’ segmentation predictions instead of the manual annotations used for training.

1 Introduction

Semantic image segmentation plays an essential role in biomedical imaging analysis. It is considered a challenging task not necessarily due to moderate image quality but mainly since it is not well defined. Different human raters and even the same rater at different time points may draw the boundary of a particular region of interest (ROI) in different manners, see e.g., \textsuperscript{7,8}. To address this well-known issue, an Expectation Maximization (EM) framework has been suggested.
for simultaneous evaluation of the raters’ performance level and the consensus segmentation \cite{16}. In \cite{6} the variability of experts annotations was discussed in the context of segmentation evaluation. Standardizing measures to assess human observer variability for clinical studies was suggested in \cite{12}.

Machine vision algorithms are often praised for being repeatable and objective. This is, however, true only when considering deterministic, model-based approaches and obviously, the resulting segmentation is model-dependent. The emergence of machine learning and deep learning, in particular, have made data-driven approaches dominant. A supervised deep neural network (DNN) for image segmentation is trained by input images and their corresponding manual annotations that are used for calculating the network’s loss. Backpropagation guided by the loss enables implicit modeling, learned indirectly from the conditional distribution of the data. This process has shown to be very powerful, outperforming model-based segmentation approaches, as it allows to generalize from seen to unseen data. Nevertheless, in most benchmark’s datasets, while the images to segment vary, the annotation is often done by a single annotator. Therefore, it is not unlikely that it is the annotator’s subjective outlook or bias that actually shapes the segmentation model and consequently influences the resulting image labels.

While the issue of inter-rater bias has significant implications, in particular nowadays, when an increasing number of deep learning systems are utilized for the analysis of clinical data \cite{9,14}, it often seems to be neglected. In \cite{1} deep learning is used to analyze the effect of common label fusion techniques on the estimate of segmentation uncertainty among observers. Nevertheless, the problem addressed there is completely different and the distinction between random versus consistent (i.e., bias) inter-observer variability is not made. Consistent differences in medical imaging data have been addressed for a related problem of inter-site variability by \cite{15} for structural MRI and by \cite{10} for diffusion MRI. In addition, inter-site variability was also investigated in the context of deep-learning in \cite{5}. Yet, inter-site variability refers mainly to the differences in imaging data acquired by different machines and not the raters.

The main objective of this work is to quantify the impact of the observer’s conception and competence on segmentation predictions by a supervised deep learning framework. In a sequence of experiments utilizing both classifier- and segmenter-DNNs we show that inter-rater bias affects significantly the DNN training process, and therefore leads to different segmentation results for the same test data, despite carefully using an identical DNN architecture and training regime. In fact, consistently lower Dice scores are calculated if training and test segmentations are of different raters. We support our findings by training a classifier DNN to distinguish between different raters based on their segmentations. Surprisingly, much more significant rater-classification results were obtained when the segmentation predictions (the outputs of the segmentation DNNs) rather then the manual annotations (used for training) were considered. We then suggest a compromise training regime, that incorporates the segmentations of both raters.
Fig. 1. MS MRI brain scans with two different annotations. (a) 2D slices of 3D MRI brain scans (FLAIR) (b-c) MS segmentation of rater #1 and rater #2, respectively.

For the purpose of this study, we used the publicly available ISBI 2015 Multiple Sclerosis (MS) lesion dataset, which includes annotations by two different raters with different levels of expertise. The results obtained allow us to highlight a worrisome clinical implication of DNN bias induced by inter-rater bias during training. Specifically, we show a significant and consistent difference in MS lesion loads calculated based on DNNs’ segmentation predictions. This difference was much larger than the difference in lesion load estimates directly based on the raters’ manual segmentations.

The rest of the paper is organized as follows. In Section 2 we describe the data; the segmenter and classifier DNNs as well as the evaluation measures we used. In Section 3 we present the experimental results. We conclude in Section 4.

2 Methods

2.1 Data: Multiple sclerosis (MS) Lesion MRI scans

For the purpose of this study, we used 21 brain MRI scans of MS patients from the ISBI 2015 MS segmentation challenge dataset [2]. Scans were acquired on 3T Philips scanner and include T1-w, T2-w, PD-w and FLAIR sequences. MS-lesions were annotated by two different raters with four (rater #1) and ten (rater #2) years of experience in delineating lesions. The second rater has overall 17 years of experience in structural MRI analysis. More details can be found in [2]. Fig. 1 presents two MRI slices along with the labels of both raters from the MS-lesion dataset.
2.2 Neural Networks: Architectures and Training

The architectures of both the segmentation-DNN and the classifier-DNNs used in this work are illustrated in Figure 2.

**3D U-Net.** The U-Net is a symmetrical fully convolutional DNN with skip connections between its down-sampling and the up-sampling paths [13]. For this study, we implemented a 3D U-Net [3] with Tensorflow and trained it for MS-lesion segmentation of multi-modal 3D brain MRI scans. We used a cross-validation scheme with 11, 5 and 5 brain scans for training, validation and test, respectively. Training was based on cross-entropy loss. Due to GPU memory limitations, the input to the U-Net consists of 100 × 100 × 100 × 4 (all 4 modalities are used) patches rather than the full 4D scans. As in the original U-Net paper [13], we favored large input patches over a large batch size and hence set the batch size to two with a learning rate of 0.0005. We also used batch normalization and dropout of 0.5.

**CNN Classifier.** The CNN classifier consists of 4 consecutive blocks, followed by a dense layer and a fully connected layer. Each block is comprised of 2 convolutions followed by one max pooling (besides the last one). We used a cross-validation scheme with 13, 1 and 7 scans for training, validation, and test, respectively. Image patches were used for both training and evaluation. We used cross entropy loss, a batch size of only 1, dropout of 0.5, batch normalization and set the learning rate to 0.001.

2.3 Quantitative evaluation measure

We use the Dice score [4] to quantify the compatibility between the manual segmentations of the two raters and for comparison between raters’ annotations and the segmentations generated by two 3D U-Nets, where each was exclusively trained on the manual segmentations of either of them. We also present the accuracy of the classifier-CNN defined as the number of correctly classified segmentation patches per-brain (rater #1 or rater #2) with respect to the total number of patches.

3 Experiments

3.1 MS-lesion segmentation for cross evaluation

**Experimental setup:** The mean Dice score between the manual segmentations of rater #1 and rater #2 is 0.7341 ± 0.0967 indicating inter-rater variability. To demonstrate the impact of the different raters’ segmentations on the DNN training, we trained a 3D U-Net twice, each time with the segmentations of either of the raters. For the sake of convenience, we term the U-Nets trained on the segmentations of rater #1 and rater #2 by network #1 and network #2, respectively, although the very same architecture and training regime were used. We then tested the U-Net performances by comparing the manual segmentations of each of the raters with the output segmentations obtained for each of the training sessions.
Results: Figure 3 presents three example MR slices of the MS lesion data with the manual segmentations of the two different raters and the respective output segmentations of network #1 and network #2. Note the relative visual similarity between the manual segmentations of rater #1 (#2) and the automatic segmentations of network #1 (#2). Cross-evaluation Dice scores obtained by training based on rater #1 and testing based on rater #2 and vice versa are presented in rows 1-2 of Table 1. For comparison, the Dice scores obtained for training and testing with the segmentations of the same rater are presented as well. Two-sample t-test with a p-value of $9.9029 \times 10^{-4}$ indicates a statistically significant difference between same-rater and cross-rater Dice scores results.

3.2 Multi-Rater Training

Experimental design. In this experiment, we trained a 3D U-Net, based on the annotations of both raters by randomly choosing in each training iteration the segmentation of either of the raters to calculate the loss.

Results. The results, as can be seen in row 3 of Table 1, show that the multi-rater trained DNN achieves on the average higher Dice scores when evaluated with respect to both raters.
3.3 MS-lesion load

An important clinical measure is the MS lesion load which indicates the severity of the disease and may predict long-term cognitive dysfunction in MS patients [11]. The lesion load is estimated by the sum of voxels labeled as lesion. Figure 4 presents a bar plot of the lesion loads of 21 brain MR volumes as calculated from the segmentations of the raters and the networks. For clarity, brain indices (x-axis) were arranged in an increasing order of the estimated MS-lesion load. The plot shows that the lesion loads calculated from the manual segmentations of rater #1 (yellow) are, for most brains, lower than the lesion loads calculated from the manual segmentations of rater #2 (dark blue). However, the more important results refer to the network segmentations. MS-load estimations based on network #1 (red) segmentations are consistently much smaller than the load estimations based on the manual segmentations it was trained on. Comparing load estimations for network #2 (light blue) and network #1 (red) shows that the inter-rater bias induces an increased bias between the networks.

![Fig. 3. Three MS-lesion brain MRI scans along with the manual and the U-Net’s segmentation contours. (a) Manual segmentations of rater #1 (b) Manual segmentation of rater #2 (c) Automatic segmentations of the U-Net, trained with the data labeled by the rater #1 (d) Automatic segmentation of the U-Net trained with the data labeled by rater #2.](image)
3.4 Rater and network classification

**Experimental design.** To support our finding we trained a classifier DNN to differentiate between the two different raters, based on their annotations. The input to the classifier includes an MR image and the corresponding segmentation of one of the raters (randomly selected). The classifier was also used to distinguish between segmentation predictions of network #1 and network #2 which were trained based on the segmentations of rater #1 and rater #2 respectively.

**Results.** The first row of Table 2 presents rater classification results based on their manual annotations. The successful classification (0.87) indicates a raters’ segmentation bias. The second row of Table 2 presents much better classification results (0.9) when a classifier CNN receives as input the segmentation predictions of the U-Nets (network #1 and network #2), rather than the manual segmentations themselves.

4 Conclusions and Discussion

In this paper we used the MS-lesion challenge dataset that was annotated by raters with different level of expertise to present the impact of inter-rater bias on DNN training. In our experiments, we demonstrate an amplification of the bias in segmentation predictions of DNNs trained on the manual segmentations of different raters. Moreover, we show that while the less experienced observer

| Trained/Evaluated | rater 1     | rater 2     |
|-------------------|-------------|-------------|
| rater 1           | 0.80 ± 0.05 | 0.74 ± 0.08 |
| rater 2           | 0.70 ± 0.10 | 0.81 ± 0.07 |
| raters1+2         | 0.78 ± 0.08 | 0.77 ± 0.09 |
Table 2. Raters’ and networks’ classification results

|            | # of patches | Mean accuracy ± std |
|------------|--------------|---------------------|
| Raters     | 6812         | 0.87 ± 0.06         |
| Networks   | 3532         | 0.90 ± 0.05         |

tended to underestimate the MS-lesion load, a DNN trained on this rater’s segmentations - provided predictions with a consistent and larger reduction in the lesions load. In the future, further experiments on different datasets and multiple rater annotations should be conducted to assess these findings. Yet, we believe that the phenomena of DNN induced bias is a general one, and as such has clinical implications that the biomedical imaging community should be aware of. Nowadays, when much effort is made to improve DNNs performances, the clinical training of the human annotator, who provides the ‘ground truth’ segmentations to the network, should be also considered.

References

1. Alain, J., Raphael, M., Ekin, E., et al.: On the effect of inter-observer variability for a reliable estimation of uncertainty of medical image segmentation. In: MICCAI. pp. 682–690 (2018)
2. Carass, A., Roy, S., Jog, A., et al.: Longitudinal multiple sclerosis lesion segmentation: resource and challenge. NeuroImage 148, 77–102 (2017)
3. Çiçek, O., Abdulkadir, A., Lienkamp, S., et al.: 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: MICCAI. pp. 424–432 (2016)
4. Dice, L.R.: Measures of the amount of ecologic association between species. Ecology 26(3), 297–302 (1945)
5. Gibson, E., Hu, Y., Ghavami, N., et al.: Inter-site variability in prostate segmentation accuracy using deep learning. In: MICCAI. pp. 506–514. Springer (2018)
6. Gordon, S., Lotenberg, S., Long, R., et al.: Evaluation of uterine cervix segmentations using ground truth from multiple experts. Computerized Medical Imaging and Graphics 33(3), 205–216 (2009)
7. Hayward, R., Patronas, N., et al.: Inter-observer variability in the measurement of diffuse intrinsic pontine gliomas. J. of Neuro-Oncology 90(1), 5761 (2008)
8. Joskowicz, L., Cohen, D., Caplan, N., Kosna, J.: Inter-observer variability of manual contour delineation of structures in CT. Eur Radiol. 29(3), 1391–1399 (2019)
9. Litjens, G., Kooi, T., Bejnordi, B., et al.: A survey on deep learning in medical image analysis. Medical Image Analysis 42, 60–88 (2017)
10. Mirzaalian, H., De Pierre, A., Savadjiev, P., et al.: Harmonizing diffusion MRI data across multiple sites and scanners. In: MICCAI. pp. 12–19 (2015)
11. Patti, F., De Stefano, M., Lavorgna, L., Messina, S.: Lesion load may predict long-term cognitive dysfunction in multiple sclerosis patients. PLoS One 10(3) (2015)
12. Popovic, Z., Thomas, J.: Assessing observer variability: a user guide. Cardiovasc Diagn Ther 7(3), 317–324 (2017)
13. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: MICCAI. pp. 234–241 (2015)
14. Shen, D., Wu, G., Suk, H.: Deep learning in medical image analysis. Annu. Rev. Biomed. Eng. 19, 221248 (2017)
15. Styner, M.and Charles, H., Park, J., Gerig, G.: Multisite validation of image analysis methods: assessing intra-and intersite variability. In: SPIE Medical Imaging, vol. 4684, pp. 278–286 (2002)
16. Warfield, S.K., Zou, K., Wells, W.: Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE transactions on medical imaging 23(7), 903–921 (2004)