Published in final edited form as:

Proc IEEE Int Symp Biomed Imaging. 2019 April ; 2019: 679–682. doi:10.1109/isbi.2019.8759474.

LOCALIZING IMAGE-BASED BIOMARKER REGRESSION WITHOUT TRAINING MASKS: A NEW APPROACH TO BIOMARKER DISCOVERY

Carlos Cano-Espinosa¹, Germán González², George R. Washko³, Miguel Cazorla¹, Raúl San José Estépar³

¹University of Alicante. Spain.

²Sierra Research SL. Spain.

³Applied Chest Imaging Laboratory, Department of Radiology, Brigham and Women's Hospital, Harvard Medical School. USA

Abstract

Biomarker inference from biomedical images is one of the main tasks of medical image analysis. Standard techniques follow a segmentation-and-measure strategy, where the structure is first segmented and then the measurement is performed. Recent work has shown that such strategy could be replaced by a direct regression of the biomarker value in using regression networks. While achieving high correlation coefficients, such techniques operate as a 'black-box', not offering quality-control images. We present a methodology to regress the biomarker from the image while simultaneously computing the quality control image. Our proposed methodology does not require segmentation masks for training, but infers the segmentations directly from the pixels that used to compute the biomarker value. The network proposed consists of two steps: a segmentation method to an unknown reference and a summation method for the biomarker estimation. The network is optimized using a dual loss function, L2 for the biomarkers and an L1 to enforce sparsity.

We showcase our methodology in the problem of pectoralis muscle area (PMA) and subcutaneous fat area (SFA) inference in a single slice from chest-CT images. We use a database of 7000 cases to which only the value of the biomarker is known for training and a test set of 3000 cases with both, biomarkers and segmentations. We achieve a correlation coefficient of 0.97 for PMA and 0.98 for SFA with respect to the reference standard. The average DICE coefficient is of 0.88 (PMA) and 0.89 (SFA). Comparing with standard segment-and-measure techniques, we achieve the same correlation for the biomarkers but smaller DICE coefficients in segmentation. Such is of little surprise, since segmentation networks are the upper limit of performance achievable, and we are not using segmentation masks for training. We can conclude that it is possible to infer segmentation masks from biomarker regression networks.

Index Terms—

Segmentation; deep learning; biomarker regression

1. INTRODUCTION

The advent of deep learning has opened different manners of image analysis. The standard paradigm of segment-and-measure is challenged by networks that directly infer the value of the biomarkers [1, 2], achieving high correlation coefficients. One of the main criticisms of deep-learning methods is their operation as ‘black-boxes’, not providing enough evidence to the medical community of their performance. With our proposed method, the regression network still operates as a ‘black-box’, but a by-product of the biomarker regression is the segmentation mask of the organ from which the biomarker is inferred, letting the end-user know how the measurement was computed. Prior work has used weak labeled datasets to infer probabilistic low-resolution heatmaps of the presence of the structures of interest [3, 4]. The closest work to us is [5], where the authors use global maximum pooling layers to detect lesions and regress their number. Our work differs from theirs in that we infer simultaneously the biomarker and the segmentation mask, without changing the structure of the network between regression and detection.

The proposed methodology is applied to the problem of the estimation of body composition measurements from a single axial slice of a thoracic CT measured at the level of the upmost part the transversal aorta. The two biomarkers that are extracted from such image are the pectoralis muscle area (PMA) and the subcutaneous fat (SFA), as shown in Figure 1. Such measurements have been shown to be more predictive of important COPD-related traits than body mass index measurements [6]. Further than PMA and SFA, the proposed methodology could be extended to any area or volume measurement that is obtained from segment-and-measure paradigms. In our experience, it is easier to obtain measurements from research studies or the medical records, already present on databases, than to engage radiologists or experts in manual segmentation tasks. The presented approach, in difference to other biomarker regression methods, produces the segmentation of the structure, thus likely easing the acceptance of the measured biomarkers by researchers or the medical community.

2. MATERIALS AND METHODS

2.1. Database

The data used consist of 10000 axial plane slices acquired from CT scans of a large multi-site COPD observational study (COPDGene [7]). An expert identified location of the upmost part of the transversal aorta and annotated the area of the pectoralis muscles and the subcutaneous fat with a semi-automated segmentation interface based on the Chest Imaging Platform [8]. We divide the data randomly in three sub-sets: training (5000 cases), validation (2000) and test (3000). For the training and validation datasets we do not use the segmentations provided by the expert, we only keep the value of their area. We keep the segmentations on the test dataset for evaluation purposes.

2.2. Network structure

The network structure can be found in Fig. 2. This consist is a U-net-like structure [9], to which we have added batch normalization after each layer. The last layer of the network is a one channel 1×1 2D convolution layer with a sigmoid activation function to ensure that the

output is limited to the range $[0, 1]$. This layer is multiplied by a binary mask obtained by applying a threshold to the input to keep only image candidates that could represent the structure to be segmented given their intensities. This mask is used to reduce the problem space to the set of feasible solutions, focusing the network attention to the set of pixels that could be part of the structure. Specifically, muscle structures can only be found if their pixel values are in the range between $[-20, 100]$ HUs and fat voxels lie inside the range $[-140, -70]$ HUs. Finally, the output of the network is added to compute the area of the structure. After thresholding with a value of 0.5, a binary opening is applied to the masked network activation to reduce the noise and generate the output segmentation.

The training function used is:

$$\mathcal{L} = \mathcal{L}_2(y, \hat{y}) + \lambda \times |\hat{y}|, \quad (1)$$

where y represents the biomarker value, $\hat{y}$ represents the estimated biomarker, $\mathcal{L}_2$ is the mean squared error. The second term of the equation is created to enforce sparsity in the solution. Experimentally, we have set the value of λ to 0.001.

2.3. Training

Two networks were trained for muscle and fat separately, using 5000 samples for training and 2000 for validation. Training is performed for 100 epochs. Each epoch takes around 600 seconds on a NVIDIA GTX-1080Ti GPU. The model of the epoch with the lowest validation loss is chosen for testing. The weights are updated using the ADAM optimizer with default parameters.

2.4. Baseline algorithm

For comparison purpose we implemented a baseline U-Net network with the same structure as the network of Fig. 2, but removing the final convolution, multiplication and summation operations. The network is trained for 100 epochs with an early stopping condition on the validation loss. Training is halted if the validation loss does not improve in 5 epochs.

3. RESULTS

We evaluate the proposed method and the baseline U-net in the testing database of 3000 unseen samples, both for the PMA and the SFA. The figures of merit used to compare the methodologies are the Pearson correlation coefficient (ρ) and the DICE coefficient. Results can be found in Fig. 3 and Table. 1.

The Pearson correlation coefficients of the proposed method and that of the UNet are equivalent (PMA: 0.97 for the proposed and 0.97 for the baseline U-Net; SFA: 0.98 for the proposed and 0.98 for the UNet), showing comparable performance of the two methods when measuring the biomarker.

DICE coefficients differ significantly between the proposed method and the UNet, in favor of the second one (PMA: 0.88 for the proposed vs. 0.95 of the UNet; SFA: 0.89 for the proposed vs. 0.95 for the UNet). In addition, a blinded expert scored 40 randomly selected

segmentations (20 PMA and 20 SFA) from both the UNet and the proposed network. Each segmentation was rated with a scale in the range [0, 5], being 5 the best. The proposed method achieve an average score of 4.60 for PMA and 4.05 for SFA, while the reference is 4.85 for PMA and 4.75 for SFA. This indicates that the obtained segmentations are not subjectively far from the ones obtained with the baseline method.

The difference in segmentation performance should come as little surprise, since the baseline is optimizing the segmentation using the ground truth as the reference. Indeed, we view the UNet performance as the upper limit that the proposed methodology can achieve. As seen in Fig. 3, the proposed network produces similar segmentations to that of the UNet, but with some false negatives in the border of the structures and few false positives in distant candidates to the structures.

4. CONCLUSIONS

We have presented a network structure that enables direct biomarker regression while simultaneously estimating the segmentation of the structure of interest. The key innovation of the presented methods vs. state-of-the-art is that the proposed network does not need the segmentation masks for training. Instead, the segmentation is inferred from only the value of the biomarker.

We have shown that the proposed network achieves the same correlation than a baseline UNet when computing the area of the pectoralis muscles and the subcutaneous fat. However, when evaluating the estimated mask and manual segmentations, our method shows lower quantitative and qualitative performance than the baseline UNet. Such is of little surprise, since the UNet is using the reference masks during training, and thus constitutes the upper limit that can be achieved by the proposed methodology.

The main failure reasons are the detection of false positives far from the structures of interest, as seen in Fig. 3. We are exploring modifications of the cost function to include segmentation compactness measurements as regularizers that could deal with such distant false positives.

The proposed methodology is, to our knowledge, the first regression network that simultaneously infers the segmentation from which the biomarker is computed. This new paradigm enables the use of databases to which the segmentation reference standard is either unknown or not available.

Acknowledgments

This research was partially funded by the National Institutes of Health NHLBI awards R01HL116931, R01HL116473 and R21HL140422. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the GPU used for this research.

5. REFERENCES

- [1]. González Germán, Washko George R, and José Raúl San, "Deep learning for biomarker regression : application to osteoporosis and emphysema on chest CT scans," in Medical Imaging - Image Processing- Proceedings of SPIE, 2018, number 3.

- [2]. Cano-espinosa Carlos, Germán González George R. Washko, et al., “Automated Agatston score computation in non-ECG gated CT scans using deep learning,” in Medical Imaging - Image Processing- Proceedings of SPIE, 2018, number 3.
- [3]. Bortsova Gerda, Dubost Florian, Ørting Silas, et al., “Deep learning from label proportions for emphysema quantification,” in International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2018, pp. 768–776.
- [4]. Zhu Wentao, Lou Qi, Vang Yeeleng Scott, et al., “Deep multi-instance networks with sparse label assignment for whole mammogram classification,” in International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2017, pp. 603–611.
- [5]. Dubost Florian, Bortsova Gerda, Adams Hieab, et al., “Gp-unet: Lesion detection from weak labels with a 3d regression network,” in International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2017, pp. 214–221.
- [6]. McDonald Merry-Lynn N., Diaz Alejandro A., Ross James C., San Jose Estepar Raul, et al., “Quantitative Computed Tomography Measures of Pectoralis Muscle Area and Disease Severity in Chronic Obstructive Pulmonary Disease. A Cross-Sectional Study,” Annals of the American Thoracic Society, vol. 11, no. 3, pp. 326–334, 2014. [PubMed: 24558953]
- [7]. Regan Elizabeth a, Hokanson John E, Murphy James R, Lynch, et al., “Genetic Epidemiology of COPD (COPDGene) Study Design,” Epidemiology, vol. 7, no. 1, pp. 1–10, 2011.
- [8]. San Jose Estepar Raul, Ross James C, Harmouche Rola, et al., “Chest imaging platform: an open-source library and workstation for quantitative chest imaging,” in C66. LUNG IMAGING II: NEW PROBES AND EMERGING TECHNOLOGIES, pp. A4975–A4975. American Thoracic Society, 2015.
- [9]. Ronneberger Olaf, Fischer Philipp, and Brox Thomas, “U-net: Convolutional networks for biomedical image segmentation,” in International Conference on Medical image computing and computer-assisted intervention. Springer, 2015, pp. 234–241.

Fig. 1.

Left: example axial slice taken at the top of the transversal aorta. Right: reference standard segmentations of the pectoralis muscle (green) and subcutaneous fat (blue).

Fig. 2. Network structure. Blue and yellow operations means ReLU and sigmoid Activation respectively. Red arrows are 2×2 downsampling operation and green arrows perform a 2×2 upsampling.

Fig. 3.

Comparison of the proposed and baseline methods with the reference standard on the test dataset. Top row: pectoralis muscle area. Bottom row: subcutaneous fat. First column: correlation between the baseline methods and the reference standard. Second column: correlation between the proposed method and the reference standard. Third row: box plots of the dice metric baseline and proposed method. Please note that the correlation coefficients are equivalent for both methods. The Dice coefficient is lower for the proposed methods, of little surprise, since the segmentation masks are never used for training.

Fig. 4.

Qualitative Evaluation of the results. Reference columns: Candidate mask (purple) with reference standard segmentation overlapped (green). Proposed and Unet columns: segmentation results with True Positives (Green), False Negatives (Red) and False Positives (Yellow). Left three columns correspond to PMA measurements and right three columns correspond to SFA.

Table 1.

Evaluation of the proposed methods against the reference standard

		Pearson	DICE	Subjective
PMA	U-Net	0.973	0.95 (0.001)	4.85
	Proposed	0.973	0.88 (0.004)	4.60
SFA	U-Net	0.981	0.95 (0.003)	4.75
	Proposed	0.983	0.89 (0.020)	4.05