Medical Image Description Using Multi-task-loss CNN

Pavel Kisilev[1(✉)], Eli Sason[1], Ella Barkan[1], and Sharbell Hashoul[1,2]

[1] IBM Haifa Research Lab, Haifa, Israel
pavel.prvt@gmail.com
[2] Carmel Medical Center, Haifa, Israel

Abstract. Automatic detection and classification of lesions in medical images remains one of the most important and challenging problems. In this paper, we present a new multi-task convolutional neural network (CNN) approach for detection and semantic description of lesions in diagnostic images. The proposed CNN-based architecture is trained to generate and rank rectangular regions of interests (ROI's) surrounding suspicious areas. The highest score candidates are fed into the subsequent network layers. These layers are trained to generate semantic description of the remaining ROI's.

During the training stage, our approach uses rectangular ground truth boxes; it does not require accurately delineated lesion contours. It has a clear advantage for supervised training on large datasets. Our system learns discriminative features which are shared in the Detection and the Description stages. This eliminates the need for hand-crafted features, and allows application of the method to new modalities and organs with minimal overhead. The proposed approach generates medical report by estimating standard radiological lexicon descriptors which are a basis for diagnosis. The proposed approach should help radiologists to understand a diagnostic decision of a computer aided diagnosis (CADx) system. We test the proposed method on proprietary and publicly available breast databases, and show that our method outperforms the competing approaches.

Keywords: Deep learning · Mammography · Computer aided diagnosis · Semantic description · Lesion detection · Multi-task loss

1 Introduction

Automatic annotation and description of natural images became recently a very popular topic in computer vision. Various approaches are proposed in a number of papers dealing with the problems of automatic semantic tagging [1], and of automatic description generation of images [2–4]. However, in medical imaging domain, this topic is yet to gain popularity. Obviously, medical image description poses its own set of problems. In particular, specifics of medical images require a pragmatic choice of semantic descriptors. In contrast to natural images, it is important that such semantic description would be standardized. The need for standardized was already recognized in breast imaging in the late 1980s.

G. Carneiro et al. (Eds.): LABELS 2016/DLMIA 2016, LNCS 10008, pp. 121–129, 2016.
DOI: 10.1007/978-3-319-46976-8_13

The American College of Radiology developed the Breast Imaging Reporting and Data System (BI-RADS) [5] that standardizes the assessment and reporting of breast lesions. The BI-RADS system has proved to be efficient in quality assurance. It aided for the comprehension of a non-radiologist report reader, and standardized the communication between clinicians and radiologists. In the past decade, standardization has been implemented in other domains, such as the Pi-RADS in prostate lesions [6], Li-RAD in liver lesions [7]. Other domains (for example, brain tumors or lung diseases) are yet to have fixed standard lexicon, but use similar semi-standardized description systems.

Tumor lesions in different organs and modalities are described by radiologists in similar terms of high level semantic descriptors. The most important semantic descriptors include shape, boundary type, density and other characteristics that are organ or modality specific. Based on these characteristics, a radiologist makes the most vital diagnostic decision about malignancy or benignancy of a tumor. Therefore, automatic classification of lesions requires either explicit or implicit representation of the above semantic descriptors by corresponding image measurements used during the classification process. Typically, Computer Aided Detection (CADe) and Diagnosis (CADx) systems use hand-crafted features such as histograms of intensity values, shape-related features (s.a., aspect ratio), texture descriptors, and others (see [8] for an overview of such systems). Using such features, these systems are able to segment and characterize lesions, and to make a diagnosis (e.g., benign or malignant).

The existing methods can be categorized into two main groups. The first (e.g. [9, 10]) performs independent estimation of semantic descriptors using supervised classification methods. The second (e.g. [11]) is based on unsupervised clustering using k-Nearest Neighbours (KNN) approach. All the above methods assume either given lesion contour or a region of interest (ROI) around the lesion, provided by a radiologist. Recently, a structured learning approach to the problem of semantic description of lesions was proposed in [12]. Hand-crafted features were calculated from semi-automatically segmented lesion contours, and were used to predict semantic descriptors using Structured Support Vector Machine (SSVM) approach.

In this work, we propose a cardinally different approach. Our system is completely based on the Convolutional Neural Network (CNN) which is trained (1) to generate ROI candidates and to rank them, and (2) based on the best candidates, to generate semantic description of lesions inside of the ROIs. Our approach does not require accurately delineated lesions, which is a laborious work usually done by radiologists. We use rectangular ROI's instead, which has a clear advantage for supervised training on large datasets. Our system learns discriminative features shared for both detection and description tasks, eliminating the need for hand-crafted features. The deep network models individual representations and dependencies of the semantic descriptors using novel joint multi-loss training. The main mode of operation of our system is depicted in Fig. 1.

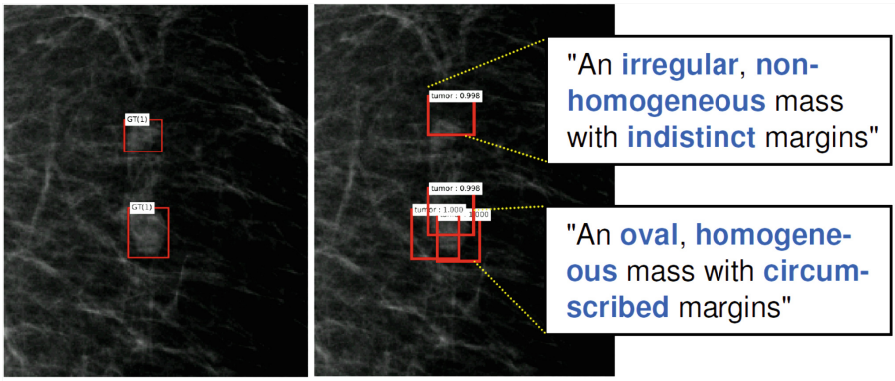

Fig. 1. An example of the output of the proposed multi-task-loss CNN based system. (a) cropped-out mammogram with 2 marked rectangular ground truth areas containing lesions, (b) corresponding top-4 automatically detected bounding box (BB) candidates, (c) automatically generated textual description of lesions in the BB's. The estimated semantic values (in blue) are embedded into predefined sentence templates. The three lower BB's have the same estimated description. (Color figure online)

2 Methodology

We define the problem of ROI detection and of semantic description of a lesion as learning of discriminative features which are partially shared in both detection and description steps. To achieve this, we use CNN-based architecture shown in Fig. 2, and described in details below.

The detection step finds ROI candidates by estimating their bounding box coordinates and the probability of an ROI being a valid lesion. The semantic description step solves a multi-attribute prediction problem; each valid candidate from the detection stage is described by multiple labels (semantic descriptors). In this stage, fully connected layers are trained to map the set of learned convolutional features, to a set of semantic descriptors. We deploy a multi-task loss and jointly train classifiers for all semantic descriptors.

Each ROI is described by a set of J semantic descriptors. The semantic description of the i-th ROI is an assignment: $\mathbf{y}_i = \{y_{i,j}\}, j = 1 \ldots J$ where each j-th semantic descriptor $y_{\mathrm{i,j}}$ can have one of the V_j possible discrete values, $Y_j \in \{1, \ldots, V_j\}$, corresponding to the categories in each one of the semantic labels of the radiological lexicon. For example, in mammography, there are $J = 3$ semantic descriptors: *shape*, *margin*, and *density*. For *shape* descriptor, $V_{1(shape)} = 3$ categories: {*oval*, *round*, *irregular*}.

Multi-task-loss CNN for Semantic Description of Medical Images The proposed system architecture is depicted in Fig. 2. Our system is based on the recently proposed Faster R-CNN architecture whose details can be found in [14]. We explain the main differences of our implementation below.

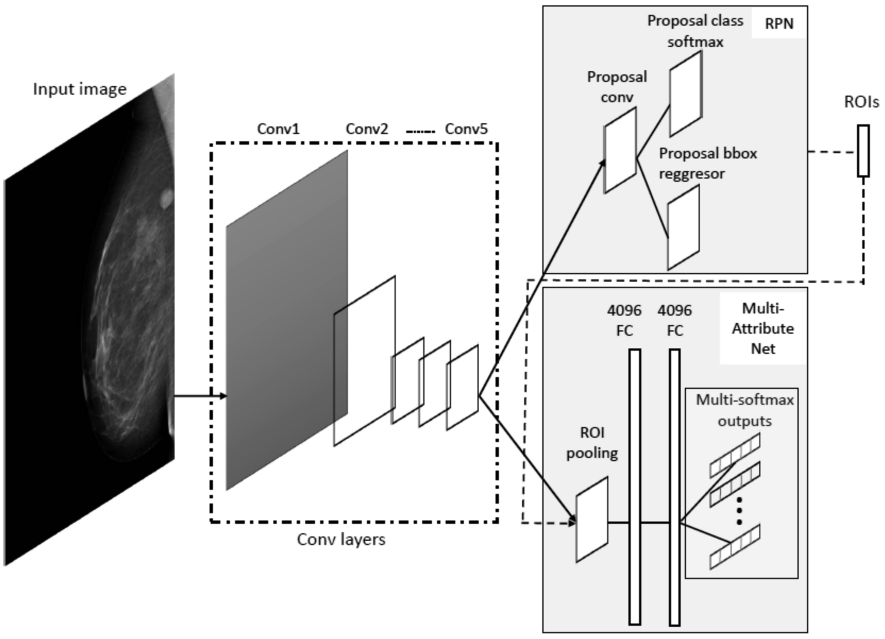

Fig. 2. The proposed multi-task loss CNN architecture for detection and description tasks

The first module (the left rectangle in Fig. 2) is a deep fully convolutional network that produces feature maps from the input image. We use this module as in [14] with a few minor changes. This module consists of 5 fully convolutional layers which are shared between the Detection and the Description stages.

The Region Proposal Network (RPN) module (the upper right rectangle in Fig. 2) generates candidates, and is trained to predict the ROI bounding box (BB) coordinates and its score. The 2 sibling sub-branches of RPN are responsible for the BB coordinates regression and for the 'objectness' score estimation. To accommodate the variety of lesion sizes, we generate BB's at several scales. The above two modules comprise the detection stage of our system.

The second stage of our system (the lower right rectangle in Fig. 2) accepts the candidate ROI's from the first stage as the inputs. In [14], the second stage is a *multi-class* classifier into one of the possible object categories. It uses single softmax loss layer. In contrast to [14], the second stage in our architecture is trained to jointly predict *multiple labels* that represent semantic descriptors. We call this branch the *Multi-Attribute Description Network* (MA-DN). It solves a *multi-class-multi-label* prediction problem. The learning is implemented in a multi-task manner, wherein our network has J sibling output layers as described below.

During the training of the network, we use a mini-batch of positive and negative ROI candidates, taken from 2 images, randomly chosen from the training set. The loss function is defined as the *multi-class-multi-label* loss, and is calculated for each mini-batch as:

$$L(\{p_{ij}\}) = \frac{1}{N}\sum_i \sum_j w_j l(p_{ij}, c_{ij}) \quad (1)$$

Here, i is the index of an ROI, N is the normalization constant according to the mini-batch size ($N \sim 128$ in most of the cases), w_j are the weights of the J terms corresponding to the different semantic descriptors. These weights are used to balance the contributions of different descriptors to the loss, and are chosen empirically in our current implementation.

The log loss for the true class c_{ij} is $l(p_{ij}, c_{ij}) = -\sum_{1\ldots V_j} t_{ij,c_{ij}} \log p_{ij,c_{ij}}$ where t_{ij} is 1 if j-th descriptor of i-th ROI is in the class c_{ij}, and 0 otherwise; $p_{ij,c_{ij}}$ is the predicted probability that the ROI is in the class c_{ij}. The probability p_{ij} for the sample -i and the label j is computed as the softmax over the $V_j + 1$ outputs of the fully connected layers. In order to implement (1), we create branches of fully connected layers for each one of the semantic descriptors, and sum the corresponding log loss terms.

During the joint training of the network branches, the proposed architecture imposes dependencies on the descriptors. In the Experiments section, we show that this architecture improves the accuracy of the descriptor estimation, as compared to the independent training of the separate branches responsible for each one of the descriptors.

Implementation details. The module of the shared convolutional network (the left branch in Fig. 2) processes the whole image with several convolutional (conv) and max pooling layers to produce conv feature maps. This branch follows the AlexNet architecture with five convolutional layers. In the RPN module, we use bounding boxes of the three aspect ratios of 1:1, 1:2 and 2:1, and of the three scales corresponding to the box sides of 32, 96, and 256 pixels. These parameters are chosen based on the statistics of the lesion sizes in the data bases that we use in our experiments.

The MA-DN network (the lower right branch in Fig. 2) accepts as an input the entire image and the RPN-generated bounding boxes (the proposals). The ROI max-pooling layer in the MA-DN branch converts the features inside of any valid ROI into a small feature map with a fixed spatial extent. As a result, each object proposal is represented by a fixed-length feature vector from the feature map of the last fully convolutional layer. Each feature vector is then fed into a sequence of 2 fully connected (fc) layers, each with 4096 neurons that finally branch into J-attribute softmax sibling output layers.

The MA-DN network is trained end-to-end by backpropagation and stochastic gradient descent (SGD) with momentum. Each SGD mini-batch is constructed from 2 images, chosen uniformly at random (we iterate over permutations of the dataset). We use mini-batches of size R = 128, sampling 64 ROI's from each image. During the training of MA-DN, we use object proposals (ROI's) that have intersection over union (IoU) with a ground truth BB of at least 0.5. These ROI's are the examples labelled as the positive class (lesion object). The remaining ROIs are sampled from object proposals that have a maximum IoU with a ground truth BB in the interval [0.05; 0.2]. These are the negative class examples (the background). During the training, we apply data augmentation by shifting the ROI's at random horizontally and vertically by up to 15 pixels. During the testing, we use the BB proposals whose score is greater than 0.85.

Optimization parameters. We use a learning rate of 0.001 for the first 12 K mini-batches, and 0.0001 for the next 16 K mini-batches generated from our datasets. We use a momentum of 0.9 and a weight decay of 0.0005.

We implemented the proposed architecture using the Caffe software framework [15]. Our system is trained on a TitanX GPU with 12 GB memory, and i7 Intel CPU with 64 GB RAM. Training times using around 400 images, each containing 256 ROI's (total about 100 K samples), are as follows. The candidate Detection stage training takes 6 h; the Description stage training takes 4 h.

3 Experiments

Our system is capable of performing end-to-end detection and semantic description of medical findings. However, the main goal of our experiments in this paper is to test thoroughly the proposed description rather than the detection stage. Also, because of the lower performance figures of CADe systems, detection is frequently performed in a semi-automatic manner. In this case, a radiologist marks the suspicious areas around a lesion.

We compare the proposed method, that uses a rectangular ROI's and their corresponding *learned* discriminative CNN-based features, to the methods based on accurately delineated lesion contours and hand-crafted features calculated from them. We also compare the performance of the proposed MA-DN architecture for joint estimation of semantic descriptors to the performance of independently trained classifiers per each descriptor. The independent classifier training was implemented using the same system but with a single semantic descriptor at a time. The results of these comparisons are summarized in Tables 1, 2 and 3, and explained in details below.

Datasets. We apply the proposed method to the breast mammography (MG) and the ultrasound (US) modalities. We used the public DDSM [13] and our proprietary data sets. In the DDSM dataset, we chose mass containing MG images with breast density of BI-RADS 1 and 2. The masses are annotated with semantic descriptors of shape and margin. The final set contains 974 images from 512 (232 benign, and 280 malignant) cases. Our proprietary dataset contains 408 US images from 330 cases, and 646 digital MG images from 281 cases. The proprietary datasets were processed by our trained radiologist who drew accurate lesion boundaries and annotated the lesions with their semantic descriptor values according to the BIRADS.

Experimental methodology. The following three approaches for the lesion description can be considered competing: (1) independent estimation of semantic descriptors (e.g., [10]); we used multiclass SVM classifiers with RBF kernel for each one of descriptors, (2) the KNN based approaches; we implemented the method from [11], and (3) the SSVM based approach of [12] which is easily implemented as well. The objective comparison of various methods for lesion detection and description is difficult. The papers conduct their experiments on different datasets or their subsets. In addition, there are very few publicly available datasets that are sufficiently large for training of deep neural networks. For that reason, we use DDSM, and our proprietary datasets.

Table 1. DDSM dataset: semantic descriptor estimation; mean performance (bold is the best result). The STDs of the metrics are all under 5 % of the mean values.

Semantic descriptor estimation method	Shape			Margin		
	ACC	PPV	TPR	ACC	PPV	TPR
Independent SVM's	0.64	0.64	0.66	0.62	0.63	0.63
k-NN based [11]	0.67	0.67	0.68	0.64	0.65	0.67
SSVM based [12]	0.71	0.71	0.72	0.69	0.68	0.69
Ours, independent	0.78	0.76	0.75	0.74	0.74	0.75
Ours, multi-task	**0.82**	**0.79**	**0.78**	**0.77**	**0.78**	**0.76**

Table 2. Proprietory mammography dataset: semantic descriptor estimation; mean performance (bold is the best result). The STDs of the metrics are all under 5.1 % of the mean values.

Semantic descriptor Estimation method	Shape ACC	Margin ACC	Density ACC
Independent SVM's	0.73	0.72	0.81
k-NN based [11]	0.74	0.76	0.80
SSVM based [12]	0.79	0.78	0.82
Ours, independent	0.84	0.82	0.81
Ours, multi-task	**0.88**	**0.86**	**0.84**

Table 3. Proprietary ultrasound dataset: semantic descriptor estimation; mean performance (bold is the best result). The STDs of the metrics are all under 7.5 % of the mean values.

Semantic descriptor Estimation method	Shape ACC	Orient. ACC	Margin ACC	Echo ACC	Transm. ACC	Boundary ACC
Independent SVM's	0.62	0.98	0.6	0.75	0.79	0.74
k-NN based [11]	0.64	0.92	0.63	0.76	0.78	0.76
SSVM based [12]	0.68	0.94	0.69	0.78	0.81	0.76
Ours, independent	0.77	**0.96**	0.75	0.77	**0.82**	0.78
Ours, multi-task	**0.82**	0.95	**0.81**	**0.78**	**0.82**	**0.8**

All the competing methods for lesion description, apart from ours, require accurate lesion contours. We therefore used a semi-automatic segmentation with a bounding box around a lesion chosen by the radiologist. We then used an active contour algorithm to extract the contours. In DDSM, we used the original annotated contours to define the ROI, and applied the active contour algorithm to refine the ground truth by extracting more accurate contours. We used the contours to compute standard image features usually deployed in detection and segmentation methods (see e.g. in [8, 11]). The groups of features include pixel intensity-, shape- and texture-related descriptors that are combined into a bag of words vector used during the training of classifiers. We used the same set of features in all the experiments. In contrast, our method did not require accurate lesion delineation, and only uses a rectangular ROI.

In all the experiments, we used the following methodology. The set of cases with corresponding images was divided (with stratification) into three equal parts (denoted

by segments A, B and C). Segment C was reserved as a testing set. Every algorithm was trained on segment A. The optimal values of parameters, evaluated on the segment B (the validation segment) were picked, and the algorithm was retrained on both segment A and B. Then, the algorithm was tested on segment C. This process was repeated with reversed roles for segments A and B, namely with segment B as the training set and segment A as the validation. This procedure was repeated five times (5 × 2 cross validation), and concluded with 10 trials. For each one of the semantic descriptors, we calculated the means and the standard deviations (STD) of the following performance metrics: (1) the accuracy, ACC = (TP + TN)/M, (2) the positive predictive value, PPV = TP/(TP + FP), and (3) the true positive rate, TPR = TP/(TP + FN). Here M is the total number of samples, TP, TN, FP, and FN are the number of true positives, true negatives, false positives, and false negatives, respectively, and we calculate these in a one-versus-all manner.

In DDSM experiments, we used the following descriptors and their corresponding values: ***shape*** {*round; oval; irregular*}, ***margin*** {*circumscribed; indistinct; spiculated; microlobulated; obscured*}. Because of the relatively small number of examples, we used a reduced set of semantic values. In particular, in US experiments, we used 3 classes for margin, shape, and echo, and 2 classes for the rest. In MG experiments, we used 3 classes for shape and margin, and 2 classes for density. We report only the accuracy for these experiments, since these numbers represent well the overall tendency.

As explained above, our main goal in this paper is to test the Description stage. However, we discuss briefly the Detection stage performance as well. In particular, we test the detection rate for the top ROI proposals. Applying the Detection stage of our system to the proprietary breast MG dataset, results in the following figures. The true lesion is detected: in 46 % of the top-1, in 64 % of the top-4, 74 % of the top-10, and 82 % of the top-50 candidates. We obtain similar figures on other datasets.

For the Description stage, the mean figures of the performance metrics for the DDSM dataset are given in Table 1. The STD's of the metrics were under 5 % of the mean values. The mean figures of the performance metrics for the proprietary breast US and MG datasets are summarized in Table 2 and Table 3 respectively. In this case, the STD's of the metrics were all under 5.1 % and 7.5 % of the mean values, respectively.

The proposed method outperforms all the competing methods in the accuracy of semantic description by up to 10 % margin. Furthermore, it is clear from the experimental results, that the proposed MA-DN architecture for joint estimation of semantic values has advantage over the independently trained classifiers per each descriptor.

4 Conclusions

This paper presents a new multi-task-loss CNN based approach for joint automatic detection and semantic description of lesions in diagnostic images. The proposed approach outperforms the competing methods by up to 10 % margin. We attribute this to the ability of deep network to learn good discriminative high level features from data. The learned features are shared in the Detection and the Description stages. The method accepts simple rectangular ground truth boxes, and, therefore, most suitable for supervised training on large datasets. The proposed approach generates standard radiological

lexicon description which should help radiologists in understanding of the decision making process of CADx systems. To that end, we plan to concentrate on improving the Detection stage performance, and making the proposed method sufficiently robust to be deployed as an end-to-end detection and description system. We also plan to extend the proposed framework and explore the use of recurrent neural networks in our system.

References

1. Guillaumin, M., Mensink, T., Verbeek, J.J., Schmid, C.: Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In: ICCV (2009)
2. Farhadi, A., Hejrati, M., Sadeghi, M.A., Young, P., Rashtchian, C., Hockenmaier, J., Forsyth, D.: Every picture tells a story: generating sentences from images. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6316, pp. 15–29. Springer, Heidelberg (2010). doi:10.1007/978-3-642-15561-1_2
3. V. Ordonez, G. Kulkarni and T. L. Berg. Im2Text: Describing Images Using 1 Million Captioned Photographs. In NIPS 2011, pages 1143–1151
4. Elliott, D., Keller, F.: Image description using visual dependency representations. EMNLP **13**, 1292–1302 (2013)
5. D'Orsi, C.J., Mendelson, E.B., Ikeda, D.M., et al.: Breast imaging reporting and data system: ACR BI-RADS - breast imaging atlas. American College of Radiology, Reston (2003)
6. Weinreb, J., et al.: PI-RADS prostate imaging - reporting and data system: 2015, Version 2. Eur. Urol. **69**(1), 16–40 (2016)
7. Mitchell, D., et al.: Li-RAD in liver lesions. Hepatology **61**(3), 1056–1065 (2015)
8. Oliver, A., Freixenet, J., Martí, J., Pérez, E., Pont, J., Denton, E.R., Zwiggelaar, R.: A review of automatic mass detection and segmentation in mammographic images. Med. Image Anal. **14**(2), 87–110 (2010)
9. Wei, C.-H., Li, Y., Huang, P.J.: Mammogram retrieval through machine learning within BI-RADS standards. J. Biomed. Inform. **44**(4), 607–614 (2011)
10. Rubin, D.L., Burnside, E.S., Shachter, R.: A bayesian network to assist mammography interpretation. In: Brandeau, M.L., Sainfort, F., Pierskalla, W.P. (eds.) Operations Research and Health Care. International Series in Operations Research & Management Science, vol. 70, pp. 695–720. Springer, New York (2004)
11. Narvaez, F., Diaz, G., Romero, E.: Automatic BI-RADS description of mammographic masses. In: Martí, J., Oliver, A., Freixenet, A., Martí, R. (eds.) Digital Mammography. Lecture Notes in Computer Science, vol. 6316, pp. 673–681. Springer, New York (2010)
12. Kisilev, P., Walach, E., Hashoul, S., Barkan, E., Ophir, B., Alpert, S.: Semantic description of medical image findings: structured learning approach. In: BMVC (2015)
13. Heath, M., Bowyer, K., Kopans, D., Moore, R., Philip Kegelmeyer, W.: The digital database for screening mammography. In: Yaffe, M.J. (ed.) Proceedings of the Fifth International Workshop on Digital Mammography, pp. 212–218. Medical Physics Publishing, Madison (2001)
14. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. In: NIPS (2015)
15. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding (2014). arXiv:1408.5093