Neural Comput & Applic
DOI 10.1007/s00521-017-3158-6

Medical image semantic segmentation based on deep learning

Feng Jiang[1] · Aleksei Grigorev[1] · Seungmin Rho[2] · Zhihong Tian[3,1] · YunSheng Fu[3] · Worku Jifara[1] · Khan Adil[1] · Shaohui Liu[1]

Received: 3 November 2016 / Accepted: 7 July 2017

Abstract The image semantic segmentation has been extensively studying. The modern methods rely on the deep convolutional neural networks, which can be trained to address this problem. A few years ago networks require the huge dataset to be trained. However, the recent advances in deep learning allow training networks on the small datasets, which is a critical issue for medical images, since the hospitals and research organizations usually do not provide the huge amount of data. In this paper, we address medical image semantic segmentation problem by applying the modern CNN model. Moreover, the recent achievements in deep learning allow processing the whole image per time by applying concepts of the fully convolutional neural network. Our qualitative and quantitate experiment results demonstrated that modern CNN can successfully tackle the medical image semantic segmentation problem.

Keywords Medical image · Semantic segmentation · Neural network · X-Ray

1 Introduction

Image segmentation refers to dividing image or classifying pixel with respect to their local features or/and texture features [1] which are very important to understand the image content (or what is where in an image).

In the field of the medical image, labeling medical image needs high care for the purpose of identifying the diagnosis of disease and clinical research [2]. Almost all of the existing medical image segmentations using convolutional neural networks were relying on the segmentation problems rather than semantic segmentation, due to different constraints. In this work, we were exploited the general image semantic segmentation to the task of medical image semantic segmentation by taking the desirable property of convolutional neural network.

Recently, deep convolutional networks are applied to segment image semantically [3–9]. For example, [4] converted fully connected layers of the convolutional neural network into convolutional layers, made accurate per-pixel classification likely using the current CNN architectures that were pre-trained on ImageNet, exploit for image semantic segmentation in a supervised way, and achieve a competitive result. However, this model suffers in modeling distant contextual region, i.e., the receptive field existing in the neuron fields of convolution layer is limited to a local area of the input image. Of course, this method was desirable for a high-level vision task such as image recognition [5] and object detection [6] but hinder the low-level task [10–12], like pose estimation, depth estimation, and semantic segmentation—where we need relatively accurate localization, rather than intangible or imperceptible spatial detail. More specifically, for the prediction and reasoning of semantic segmentation, contextual evidence from distance area is very necessary, especially for a

Feng Jiang and Aleksei Grigorev have contributed equally to this work.

✉ Feng Jiang
fjiang@hit.edu.cn

[1] Department Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China

[2] Department of Media Software, Sungkyul University, Anyang, Korea

[3] Institute of Computer Application, Chinese Academy of Engineering Physics, Mianyang 621900, China

Published online: 15 July 2017

medical image which needs a high attention to give the final decision on the medical image problem from the semantic label.

In order to tackle the problem of contextual evidence from distant, [7] proposed a spatially recurrent (ReNet) model in which the well-known recurrent neural network sweeps in two directions, horizontally and then sweeps the output of hidden unit vertically across the image, and is able to retrieve long-range dependence. Following this result, this paper also optimizes the result by integrating fully convolutional network (FCN) and ReNet. In this model, ReNet together with FCN (H-ReNet: Hybrid Spatially Recurrent Network and Fully convolutional layer), the ReNet fosters long-range dependence and local features are captured from the convolutional layer. Like this, H-ReNet capture global context, which is very important for feature representation, fosters long-range dependence and is able to train end to end. However, training ReNet inclines to be computationally severe.

Following ReNet [8, 10] model builds on the top of ReNet called ReSeg model: a recurrent neural network model for semantic segmentation which predicts structure architecture by employing local generic feature extracted by CNN and considers the ability to which RNN captures distant dependence. The model processes the input image in the first layer of the pre-trained VGG-16 model on pre-trained image Net, and the image was stated in such a way that the resolution becomes small. Then, the output or the feature map obtained from this layer was fed into ReNet layer that sweeps across the image and at the end, and upsampling layer is employed to resize the last feature map to make a same resolution as the input image. Different from ReNet model, the paper chooses gated recurrent unit instead of long short-term memory for the implementation of recurrent neural network, and finally, the model attempts to capture local and global pixel dependence.

In an isolated track, probabilistic graphical model like conditional random field (CRF) and Markov random field (MRF) has been developed as a successful approach to boost the precision of pixel-level label task in the field of computer vision. CRF can able to improve weak and coarse pixel-level label prediction to make sharp boundaries and fine-grained segmentation. Taking this into consideration, very recently, [3] introduce model which integrates CNN and conditional random field-based probabilistic model, and in this model, CRF is able to replace the role of recurrent neural network architecture proposed for scene parsing by [9], possible to train end to end and applied for semantic segmentation accurately.

Considering the limitation of pixel-level segmentation, such as expensive labeling effort for more training data, [13] replaces fully connected CRF with domain transform, which was first proposed by [14] and the amount of flattening is controlled with a reference edge map. It is seen that the domain transform can perform as equivalent as a recurrent neural network. Basically, this proposed model has three parts with a different purpose: The first part produces coarse semantic segmentation; the second part predicts edges by exploiting features from intermediate layers of Deep Lab proposed by [15]; and the last component contains domain transform, which is an edge-preserving filter.

The response at the final layer of the deep convolutional neural network in localizing image segmentation is not sufficiently accurate which is shown by [15]. To tackle this problem of localization property of the deep convolutional neural network, this work combines the output of the final deep CNN layer with a fully connected conditional random field and obtains a competitive result in terms of speed, accuracy, and simplicity.

Overall, the above-mentioned methods all rely on various kinds of the image dataset, different from medical image to address the problem of image semantic segmentation. Due to lack of the availability of medical image dataset for training, a very limited number of works have been done on medical image semantic. Our work tries to implement medical image semantic segmentation by taking the advantage of the fully convolutional layer and recurrent layer proposed by [7], and this is the first work shown for medical image semantic segmentation in this manner. We consider the online available Japan Society of Radiological Technology (JSRT) dataset to show the result of semantic segmentation of chest radiographs. Our contribution can be summarized as follows:

1. We proposed to employ specific convolutional neural network to tackle medical image semantic segmentation problem.
2. We trained neural network from scratch, did not use already trained model, and did not split train in “pre-train” and “fine-tune” stages.
3. We demonstrated that it is unnecessary to keep the full original VGG-16 architecture. Moreover, the network was trained without last block of convolutional layers, and ReNet layer was forced to play the role of the max-pooling layer.

The remainder of the paper is organized as follows. In Sect. 2, the related work is presented, Sect. 3 is composed of medical image preprocessing and network architecture description, Sect. 4 describes experimental part and discusses it, and finally, Sect. 5 gives the conclusion of the paper

2 Related works

In the field of medical images, recent efforts were mostly focused on image segmentation [16–19], which are presented different methods for lungs segmentation. For instance, [19] use the morphological operation such as region-growing and prediction-based segmentation in order to improve the result of the previous step. The method proposed in [18] contains several steps: First, it finds the most similar lung atlases to the input, then it warps them to the input by applying SIFT flow, and finally, it employs the graph-cut segmentation algorithm in order to detect lungs boundary.

Another approach [16] proposed a combination of deep learning method and distance regularized level set (DLRS) [20]. They use the deep belief network (DBN) to obtain the initial segmentation of the image, and then, they run DRLS algorithm until it converges to an optimal solution.

Another method [17] proposed a workflow for the semiautomatic segmentation of perfusion images. This workflow is composed of several steps: First, the lungs are delineated in HASTE images; next, the image is registered with the perfusion images; in the end, in order to align the lung segmentation from the morphological dataset with perfusion images, the transformation resulting from registration is used.

Several works have been done aiming to divide an image into the semantically meaningful section and classifying each section into one of the predefined classes for a different purpose.

By limiting the number of annotated samples used for training (i.e., taking image-level not pixel-level annotation), [21] consider social image consisting of noisy, to the semantical segment in a weakly supervised way. Distant dependence of the whole image was extracted from CNN, and another contexts like the high-level semantic concept and low-level visual appearance, inter-label correlation and label consistency between image level and pixel level are controlled by learning joint CRF from the weakly labeled social image. For each pixel, the system predicts an object label by making use of only image-level labels during training that represents a weak way of training an image. Similarly, by combining pixel-level and image-level annotated image for training, a work by [22] introduced weakly and semi-supervised learning model which learns from weakly annotated data and the combination of few strongly labeled images, sourced from one or multiple dataset.

A work by [23] employs multiple-task multiple-instance learning to segment image semantically that was weakly annotated in training set using geometric context estimation by considering it as a secondary task. Xu et al. [24] deploy weak annotation (no pixel-wise labeling even at the time of training) in order to reduce the label complexity, and then introduce the problem as the one of learning in a latent prediction framework, where there is the absence/presence of a class and projection of semantic label to a super-pixel generated by graphical model. This mode of annotation was desirable in reducing labeling cost, usually 15 times faster than the annotating image at pixel level.

In [25], neural network model with an iterative graphical optimization method is combined to approximate pixel-wise object segmentation. The fully connected conditional random field is exploited to regularize different contexts for segmentation task and the proposed model: DeepCut update training targets learned by convolutional neural network model.

Fully convolutional networks were exploited independently [4], and together with other methods [3–9] widely, having a certain deficit in covering local dependencies in some of these methods, and attempted to reduce the deficit in the other methods.

[7] exploits spatially recurrent layer (ReNet) to capture global context and later integrate groups of ReNet with the pre-trained fully convolutional networks like VGG-16 net [26] and Google net [27] and, as a result, obtains better performance over spatially recurrent layer. Following [7, 8], model ReSeg at the top of spatially recurrent layer is still used for image semantic segmentation. Considering domain transform proposed by [14], first introduced for high-quality edge-preserving filtering of images and videos in real time, [13] introduced semantic image segmentation with task-specific edge detection using CNNs and a discriminatively trained domain transform, where discriminative transform can equivalently perform what recurrent neural network can do with convolutional neural network.

All of the above-mentioned approaches of image semantic segmentation rely on various kinds of images dataset, different from medical image to address the problem of image semantic segmentation, and this affects the investigation of medical image segmentation used for various purpose. Therefore, this paper is looking at medical image semantic segmentation for the aforementioned purposes. Due to the lack of the availability of medical image dataset for training, a very limited number of works have been done on medical image segmentation in a weakly supervised manner. Our work tries to implement medical image segmentation by taking the advantage of the convolutional layer and recurrent layer proposed by [7], and this is the first work employed for medical image semantic segmentation. In the next section, we give the description of the model layer used for medical image segmentation independently (ReNet alone and ReNet with the fully convolutional network).

3 Model description

First, we describe a preprocessing method for medical images, and then following [8] and [10], we describe the how the combination of ReNet and the convolutional layer is employed for medical image segmentation.

3.1 Medical image preprocessing

Firstly, for an input medical image Y, the K-SVD method is used to get redundant dictionary D. Then, a novel method is proposed to get adaptive filters for FoE. In order to get better high-frequency components, DC components are removed from D by subtracting the average value of elements from each element in the atom. D keeps the high-frequency components. Then, an adaptive filter for FoE that is most relevant to current image using singular value decomposition (SVD) is learned as follows:

$$D = U\sum V = \sum_{i=1}^{d} \sigma_i u_i V_i^T,$$

where $\theta = \{\sigma_i\}_{i=1}^{d}$ are the singular value vector; $\sum = \mathrm{diag}(\theta)$ is a diagonal matrix; u_{i} and V_{i} are columns of U and V, respectively.

U has bigger projection values on the former atoms and smaller projection values on the latter atoms. Thirdly, the latter atoms of U are taken as adaptive filters $\tilde{\omega}_k$ for FoE.

By incorporating the proposed joint statistical modeling into the regularization-based framework, a new formulation for image restoration can be expressed as follows:

$$\mathrm{argmin}_x \frac{1}{2}\|X-Y\|_2^2 + \lambda \sum_{R_{ij}} \left\|J_{\mathrm{k-SVD}}^T R_i X\right\|_0,$$

where R_{ij} is the sample matrix and $J_{\mathrm{K-SVD}} = [J_1, \ldots, J_k]$, and each column J_i is an adaptive filter. The first term actually represents the observation constraint, and the second term represents the prior constraint. Let $x, y \in \mathbb{R}^N$, $\boldsymbol{G}_{\boldsymbol{x}_k}$, $\boldsymbol{G}_{\boldsymbol{y}_k} \in \mathbb{R}^{B_s \times c}$ and denote the error vector by $e = x - y$ and each element of e by $e(j)$, $j = 1, \ldots, N$. Assume that $e(j)$ is independent and comes from a distribution with zero mean and variance δ^2. Then, for any $\varepsilon > 0$, we have the following property to describe the relationship between $\|X-Y\|_2^2$ and $\sum_{k=1}^{n}\left\|\boldsymbol{G}_{\boldsymbol{x}_k} - \boldsymbol{G}_{\boldsymbol{y}_k}\right\|_2^2$, that is,

$$\lim_{N\to\infty, K\to\infty} P\left\{\left|\|X-Y\|_2^2 - \frac{N}{K}\sum_{k=1}^{n}\left\|\boldsymbol{G}_{\boldsymbol{x}_k} - \boldsymbol{G}_{\boldsymbol{y}_k}\right\|_2^2\right| < \varepsilon\right\} = 1,$$

where $P(.)$ represents the probability and $K = B_s \times c \times n$. Accordingly, there exits the following equation with very large probability:

$$\|X-Y\|_2^2 = \frac{N}{K}\sum_{k=1}^{n}\left\|\boldsymbol{G}_{\boldsymbol{x}_k} - \boldsymbol{G}_{\boldsymbol{ry}_k}\right\|_2^2.$$

It is worth emphasizing that the above assumption does not need to be Gaussian process, which is more general and reasonable and leads to

$$\sum_{R_i}\left\|J_{\mathrm{k-SVD}}^T R_{ij} X\right\|_0 = \frac{1}{c}\sum_{k=1}^{n}\left\|J_{\mathrm{k-SVD}}^T \boldsymbol{G}_{\boldsymbol{x}_k}\right\|_0.$$

Accordingly,

$$\begin{aligned}&\mathrm{argmin}_x \frac{1}{2}\|X-Y\|_2^2 + \lambda\sum_{R_i}\left\|J_{\mathrm{k-SVD}}^T R_{ij} X\right\|_0 \\ &\quad = \mathrm{argmin}_x \frac{1}{2}\left\|\boldsymbol{G}_{\boldsymbol{x}_k} - \boldsymbol{G}_{\boldsymbol{y}_k}\right\|_2^2 + \frac{\lambda \times B_s \times n}{c \times N}\sum_{k=1n}\left\|J_{\mathrm{k-SVD}}^T \boldsymbol{G}_{\boldsymbol{x}_k}\right\|_0,\end{aligned}$$

which is equivalent to

$$\mathrm{argmin}_x \frac{1}{2}\left\|\boldsymbol{G}_{\boldsymbol{x}_k} - \boldsymbol{G}_{\boldsymbol{y}_k}\right\|_2^2 + \tau\sum_{k=1}\left\|J_{\mathrm{k-SVD}}^T \boldsymbol{G}_{\boldsymbol{x}_k}\right\|_0.$$

Defining $(\overline{J}_{\mathrm{K-SVD}}, \overline{J}_{\mathrm{K-SVD}}) = U$, due to the construction of U and the unitary property of U_i, it yields

$$\begin{aligned}&\mathrm{argmin}_x \frac{1}{2}\left\|\boldsymbol{G}_{\boldsymbol{x}_k} - \boldsymbol{G}_{\boldsymbol{y}_k}\right\|_2^2 + \frac{\lambda \times B_s \times n}{c \times N}\sum_{k=1}\left\|J_{\mathrm{k-SVD}}^T \boldsymbol{G}_{\boldsymbol{x}_k}\right\|_0 \\ &\quad = \mathrm{argmin}_x \left\|\left(\overline{J}_{\mathrm{K-SVD}}, J_{\mathrm{K-SVD}}\right)^T\left(\boldsymbol{G}_{\boldsymbol{x}_k} - \boldsymbol{G}_{\boldsymbol{y}_k}\right)\right\|_2^2 \\ &\qquad + \tau\sum_{R_i}\left\|J_{\mathrm{k-SVD}}^T \boldsymbol{G}_{\boldsymbol{x}_k}\right\|_0 \\ &\quad = \mathrm{argmin}_x \sum_{k=1}^{n}\frac{1}{2}\left\|\overline{J}_{\mathrm{K-SVD}}^T\left(\boldsymbol{G}_{\boldsymbol{x}_k} - \boldsymbol{G}_{\boldsymbol{y}_k}\right)\right\|_2^2 \\ &\qquad + \frac{1}{2}\sum_{k=1}^{n}\left\|J_{\mathrm{K-SVD}}^T\left(\boldsymbol{G}_{\boldsymbol{x}_k} - \boldsymbol{G}_{\boldsymbol{y}_k}\right)\right\|_2^2 + \tau\sum_{k=1}\left\|J_{\mathrm{k-SVD}}^T \boldsymbol{G}_{\boldsymbol{x}_k}\right\|_0.\end{aligned}$$

Clearly, the minimization can be decoupled and solved separately. To sum up, the minimization is equivalent to solve the subproblems; namely, for each subproblem in this optimization, the subproblem is equivalent to

$$\begin{aligned}&\mathrm{argmin}_x \frac{1}{2}\left\|\overline{J}_{\mathrm{K-SVD}}^T\left(\boldsymbol{G}_{\boldsymbol{x}_k} - \boldsymbol{G}_{\boldsymbol{y}_k}\right)\right\|_2^2 \\ &\qquad + \frac{1}{2}\left\|J_{\mathrm{K-SVD}}^T\left(\boldsymbol{G}_{\boldsymbol{x}_k} - \boldsymbol{G}_{\boldsymbol{y}_k}\right)\right\|_2^2 + \tau\left\|J_{\mathrm{k-SVD}}^T \boldsymbol{G}_{\boldsymbol{x}_k}\right\|_0.\end{aligned}$$

We get

$$\overline{J}_{\mathrm{K-SVD}}^T \boldsymbol{G}_{\boldsymbol{x}_k} = \overline{J}_{\mathrm{K-SVD}}^T \boldsymbol{G}_{\boldsymbol{y}_k}$$

and

$$\begin{aligned}J_{\mathrm{K-SVD}}^T \boldsymbol{G}_{\boldsymbol{x}_k} &= \mathbf{hard}\left(J_{\mathrm{K-SVD}}^T \boldsymbol{G}_{\boldsymbol{y}_k}, \sqrt{2\tau}\right) \\ &= J_{\mathrm{K-SVD}}^T \boldsymbol{G}_{\boldsymbol{y}_k}.(\mathbf{abs}(J_{\mathrm{K-SVD}}^T \boldsymbol{G}_{\boldsymbol{y}_k}), \sqrt{2\tau}) = \boldsymbol{A},\end{aligned}$$

where hard(.) denotes the operator of hard thresholding and . stands for the elementwise product of two vectors. And we get $\begin{bmatrix} \overline{J}^T_{\mathrm{K-SVD}} \\ J^T_{\mathrm{K-SVD}} \end{bmatrix} \boldsymbol{G}_{\boldsymbol{x}_k} = \begin{bmatrix} \overline{J}^T_{\mathrm{K-SVD}} \boldsymbol{G}_{\boldsymbol{y}_k} \\ \boldsymbol{A} \end{bmatrix}$. And it yields $\boldsymbol{G}_{\boldsymbol{x}_k} = \boldsymbol{U} \begin{bmatrix} \overline{J}^T_{\mathrm{K-SVD}} \boldsymbol{G}_{\boldsymbol{r}_k} \\ \boldsymbol{A} \end{bmatrix}$.

This process is applied for all groups to achieve estimates $\hat{\boldsymbol{G}}_{\boldsymbol{x}_k}$, and all $\hat{\boldsymbol{G}}_{\boldsymbol{x}_k}$ are returned to their original positions and averaged at each pixel and obtain the enhanced image X.

3.2 Spatially recurrent layer

Spatially recurrent layer receives an enhanced input image (or feature map of the previous layer) I of size $H \times W$ where H and W are height and width of the image and then split it into set of non-overlapping path with size of $w_p \times h_p$. The spatially recurrent layer has two one-dimensional RNNs with independent weight sweeping across the grid in two directions (vertically and horizontally). In this work for the implementation of RNNs, LSTM described by [28] is chosen because it has a good property in overcoming the problem of the gradient vanishing.

More formally, the ReNet layer receives a 2D map as input, sweeps across the grid in opposite direction, and updates the cell memory $C_{x,y}$ and hidden state $h_{x,y}$ of its LSTM unit at a location (x, y) as

$$\left(h^{\rightarrow}_{x,y}, C^{\rightarrow}_{x,y}\right) = \mathrm{LSTM}^{\rightarrow}\left(I_{x,y}, h^{\rightarrow}_{x-1,y}, C^{\rightarrow}_{x-1,y}\right) \quad \text{for } x = 1, \ldots, H$$

$$\left(h^{\leftarrow}_{x,y}, C^{\leftarrow}_{x,y}\right) = \mathrm{LSTM}^{\leftarrow}\left(I_{x,y}, h^{\leftarrow}_{x-1,y}, C^{\leftarrow}_{x-1,y}\right) \quad \text{for } x = H, \ldots, 1,$$

where the arrow represents the forward and backward direction. After the first process in the first RNN, by concatenating the hidden state $C^{\rightarrow}_{x,y}$ and $C^{\leftarrow}_{x,y}$, each of which has d hidden units, we can composite feature map of size $h \times w \times 2d$. The output of future map covering the full image is obtained by stacking two ReNet layers with orthogonal sweeping direction. Example of features map processing is shown in Fig. 1.

On the top of each other, constructing a simple deep ReNet by stacking a multiple-layer group in which the first group receives the raw pixel as input and the output of the last group is passed through a softmax layer to produce dense prediction, but relatively less accurate.

With the number of convolutional kernels same to the number of labels, supplementary 1×1 convolution layer is appended on the top of ReNet, to align the channel number of output feature map with the number of semantic labels.

Then, this presented ReNet layer was combined with the fully convolutional network to achieve better result compared with ReNet layer-based semantic segmentation alone. Next, we briefly presented the integrated fully convolutional network with the ReNet layer which was used for medical image semantic segmentation for our case.

3.3 Recurrent layer with fully convolution layer

The basic model used for the segmentation of medical image relies on this model (ReNet with the fully

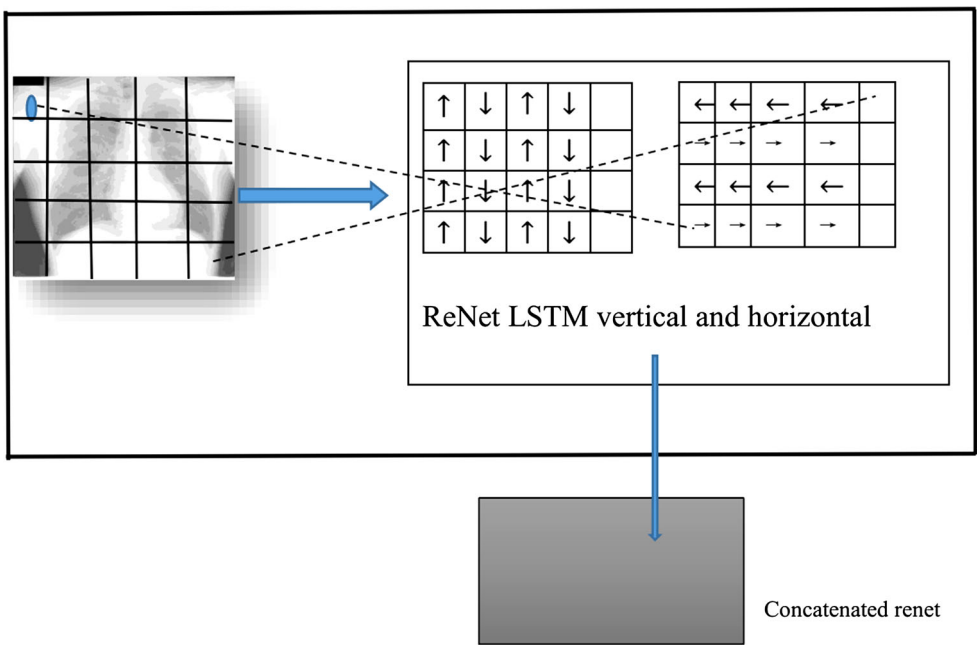

Fig. 1 ReNet layer

convolutional network), which is modified from [7]. First, the baseline for fully convolutional network architecture was adapted from the VGG-16 network. We used some of the first layers of VGG-16 (four convolution layers and four max-pooling layers).

As max-pooling layers reduce the size of the feature map, this reduction in feature map may not have visible effect for a high-level task like classification, but the high-resolution map is very important for semantic segmentation.

Second, the ReNet layer group discussed in Sect. 3.2 is inserted on the top of the layer of fully convolution network of selected VGG-16 layers, and then, this layer is followed by one convolutional layer and upsampling via transposed convolution or fractional stride convolution.

Since combining features from multiple layers results in more discriminative feature for semantic segmentation and object detection problems, we concatenate feature map from different selected layers of the VGG-16 where the selection of the features from each layer was carried out systematically. As the final combination may be dominated by one or two feature maps and the magnitude of feature map fluctuates substantially, we further normalize the feature map before concatenation.

More specifically, the architecture is described as follows: First, the image was processed in some of the first layers of VGG-16 (first four convolutional and four max-pooling layers). Moreover, by considering the advantage of multi-layer feature combination, feature map from selected VGG-16 layer was concatenated from pool 4 and conv7 with batch normalization and then one ReNet layer group with horizontal and vertical direction was put on the top of concatenation result, where first ReNet layer has receptive field with size of 2×2 which decreases computational complexity by reducing size of feature maps. Finally, the output of the model visualizes the semantic prediction of an image. The difference of this model with the model in [7] is that we dropped the conv5 layer from original VGG-16 model, concatenated only output of conv4 layer and conv7 layer, and employed ReNet layer that has receptive field with size of 2×2, whereas the model in [7] concatenates feature maps from selected layers from VGG- 16. The detail of the architecture is shown in Fig. 2

4 Experiments

4.1 Image dataset

To train and test the model and later exploit the semantic segmentation of the medical image, we were considering the image of chest radiographs with nodule and non-nodule from Japan Society of Radiological Technology (JSRT) database. The chest radiographs are taken from the Japan Society of Radiological Technology database. This is a publicly available database with 247 PA chest radiographs collected from 13 institutions in Japan and one in the USA. The images were scanned from films to a

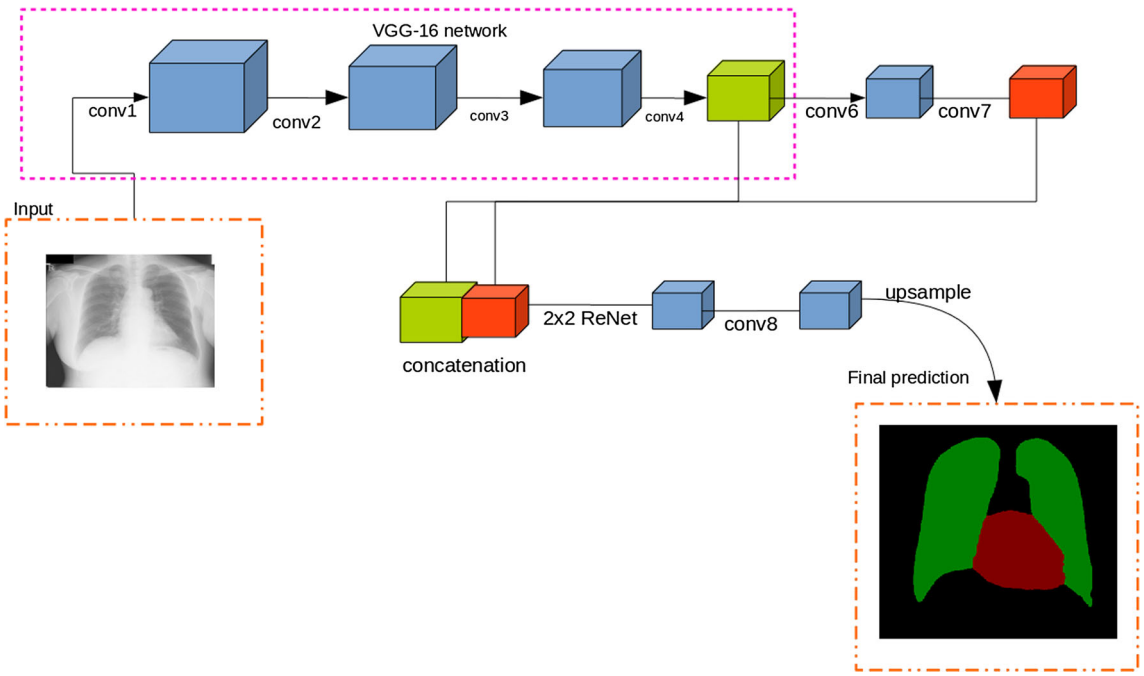

Fig. 2 ReNet with fully convolutional network structure

size of 2048×2048 pixels, a spatial resolution of 0.175 mm/pixel, and 12-bit gray levels. One hundred and fifty-four images contain exactly one pulmonary lung nodule each; the other 93 images contain no lung nodules. We adjust the size of the image to make comfortable for our work.

Table 1 Experimental parameters

Parameter	Value	Description
Network initialization		
$\mathcal{N}(\mu, \sigma^2)$	$\mathcal{N}(0, 0.01)$	Gaussian distribution
$\mathcal{U}(a, b)$	$\mathcal{U}(-0.04, 0.04)$	Uniform distribution
Train stage		
μ	0.9	Momentum
α	0.0001	Learning rate
Epochs	440	Termination criteria

4.2 Implementation details

In order to perform segmentation, we implemented the H-ReNet model [7] with our own modifications using the Caffe [29] deep learning framework. It is well known that one of the significant issues of the CNN is the network initialization [30] and the poorly initialized network cannot bet trained with momentum. Moreover, the huge amount of CNN is trained on the datasets which are composed of RGB images, but dataset used in this work is composed of gray images, and it sets some limitations on using pre-trained model for image segmentation because a number of networks parameters are different. As mentioned before, it is almost impossible to train the deep CNN from scratch on the small dataset.

We initialized all convolutional layers by sampling from a Gaussian distribution with $\mu = 0$ and $\sigma = 0.1$. The parameters of ReNet layers were randomly initialized by sampling

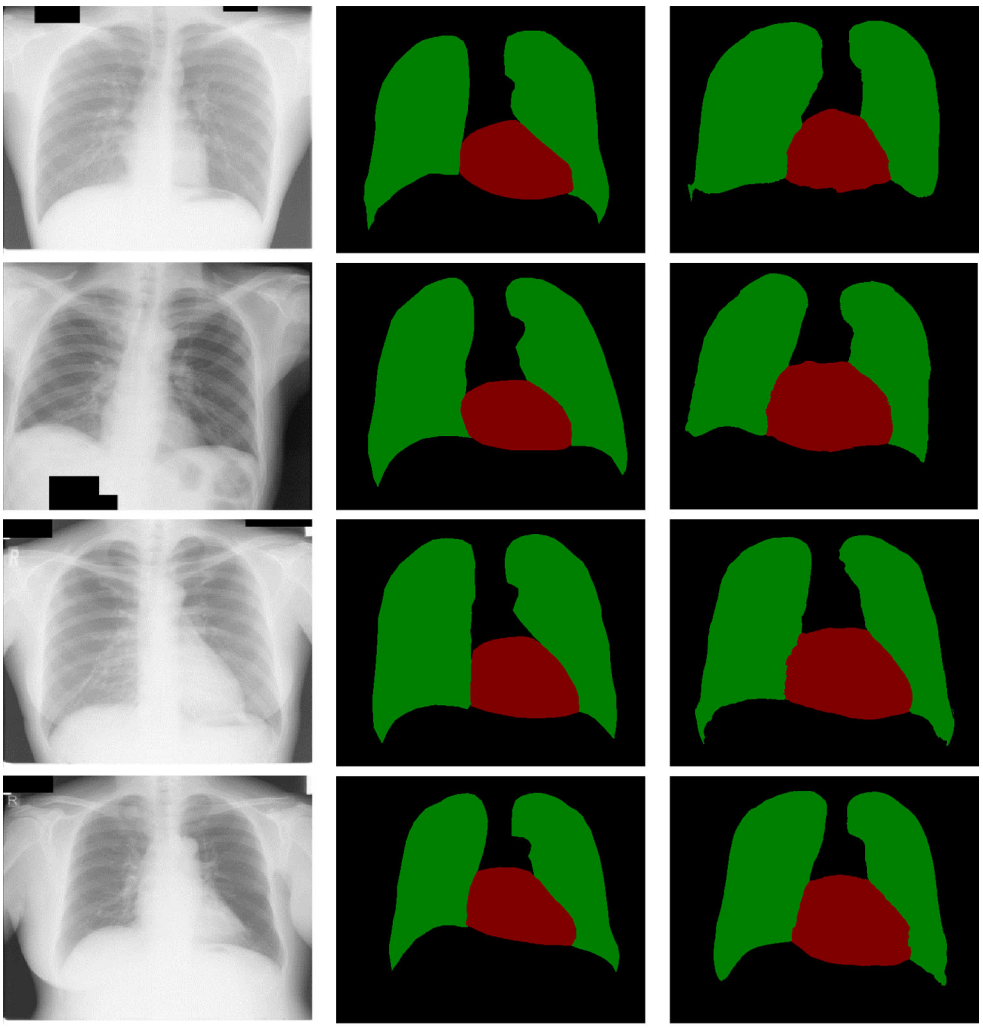

Fig. 3 An example of semantic segmentation on the JSRT dataset (*red* is heart; *green* is lungs). **a** Image, **b** ground truth, **c** prediction

from a uniform distribution over $[-0.4; 0.4]$. We also set the number of the filters in the upsampling layer to 3, since the dataset contains the three semantic regions: background, heart, and lungs. We train all layers of the network without employing "pre-train" and "fine-tune" concepts. The stochastic gradient descent (SGD) [30, 31] with momentum was used to train network. Updates rules are:

$$V_{t+1} = \mu V_t - \alpha \nabla L(W_t)$$

and $W_t = W_t + V_{t+1}$,where V is the update value and W is the updated weights. We followed the original paper [5], thus the learning rate α was set to 0.0001, and momentum μ was set to 0.9. Before each epoch, dataset was randomly shuffled. Training was done on a standard desktop with Ubuntu 16.04 and a NVIDIA GTX TITAN X with 12 GB memory. Training of the network takes around 12 h. The parameters of training stage and network initialization are summarized in Table 1.

The bottleneck of network architecture is ReNet layers, which are composed of LSTM units. Although the Caffe [29] deep learning framework was used for training and testing, where low-level programming is handled by C++, and convolutional operations are performed on GPU by employing CUDA, LSTM units still have poor performance. It takes almost 0.5 s on the desktop to perform prediction.

4.3 Experiments results

Before the network training, we removed overexposed images and resized the rest images to 400×500. Hence, our dataset is composed of 240 images and corresponding labels. The dataset was randomly split into train and test part, which contains 200 and 40 images correspondingly. As a preprocessing step, mean subtraction was applied to each pixel of the image.

As shown in Fig. 3, the trained network produces a decently labeled image. Moreover, it is also possible to train network only on the lungs extraction task, because some parts of heart and lungs are overlapped, and it affects the final result.

5 Conclusion

The existing method of image semantic segmentation was relying on various kinds of the images different from the medical image, and this paper exploits the medical image semantic segmentation by taking the advantage of the fully convolutional layer and spatially recurrent layer. Considering the Japan Society of Radiological Technology, which contains 247 images of the lung with 154 lung nodules and the rest 93 with no nodule, we perform the semantic segmentation of lungs and heart. Experiments results demonstrated that CNN can successfully tackle the medical image segmentation tasks. Moreover, it also can use to extract lungs or heart boundaries.

Acknowledgements This work is partially funded by the MOE—Microsoft Key Laboratory of Natural Language Processing and Speech, Harbin Institute of Technology, the Major State Basic Research Development Program of China (973 Program 2015CB351804), and the National Natural Science Foundation of China under Grant Nos. 61572155, 61672188, and 61572153. We would also like to acknowledge NVIDIA Corporation who kindly provided two sets of GPU.

Compliance with ethical standards

Conflict of interest The authors declared that they have no conflicts of interest to this work. We declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.

References

1. Shotton J, Winn J, Rother C Crimininsi A (2006) TextonBoost: joint appearance, shape and context modeling for multi-class object recognition and segmentation. In Proceedings of European conference on computer vision, vol 3951, Chapter 1, pp. 1–15
2. Jiang J, Trundle P, Ren J (2010) Medical image analysis with artificial neural networks. Comput Med Imaging Graph 34(8):617–631
3. Zheng S, Jayasumana S, Romera-Paredes B, Vineet V, Su Z, Du D, Huang C, Torr P (2015) Conditional random fields as recurrent neural networks. In: Proceedings of the ICCV, pp 1529–1537
4. Long J, Shelhamer E, Darrell T (2015) [Slices] fully convolutional networks for semantic segmentation. In: Cvpr 2015
5. Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet classification with deep convolutional neural networks. In: Proceedings of the NIPS, pp 1–9
6. Girshick R, Donahue J, Darrell T (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: 2014 IEEE conference on computer vision pattern recognition, pp 580–587
7. Yan Z, Zhang H, Jia Y, Breuel T, Yu Y (2016) Combining the best of convolutional layers and recurrent layers: a hybrid network for semantic segmentation. arXiv:1603.04871
8. Visin F, Ciccone M, Romero A, Kastner K, Kyunghyun C, Bengio Y, Matteucci M, Courville A (2016) ReSeg: a recurrent neural network-based model for semantic segmentation. In IEEE conference on computer vision pattern recognition workshops
9. Pinheiro PHO, Collobert R (2014) Recurrent convolutional neural networks for scene Labeling. In: Proceedings of the 31st international conference on Machine Learning, pp 82–90
10. Chen B-W, Wang J-C, Wang J-F (2009) A novel video summarization based on mining the story-structure and semantic relations among concept entities. IEEE Trans Multimedia 11(2):295–312
11. Chen B-W, Chen C-Y, Wang J-F (2013) Smart homecare surveillance system: behavior identification based on state transition support vector machines and sound directivity pattern analysis. IEEE Trans Syst Man Cybern Syst 43(6):1279–1289
12. Chen B-W, Tsai A-C, Wang J-F (2009) Structuralized context-aware content and scalable resolution support for wireless VoD services. IEEE Trans Consum Electron 55(2):713–720
13. Chen L-C, Barron JT Papandreou G Murphy K Yuille AL (2015) Semantic image segmentation with task-specific edge detection

using CNNs and a discriminatively trained domain transform. p 12
14. Gastal ESL, Oliveira MM (2011) Domain transform for edge-aware image and video processing. ACM Trans Graph 30(4):1
15. Chen L-C, Papandreou G, Kokkinos I, Murphy K, Yuille AL (2014) Semantic image segmentation with deep convolutional nets and fully connected CRFs. In: Iclr, pp 1–14
16. Ngo TA, Carneiro G (2015) Lung segmentation in chest radiographs using distance regularized level set and deep-structured learning and inference. In: IEEE international conference on image processing (ICIP), pp 2140–2143
17. Wolf I, Böttger T, Grunewald K, Schöbinger M, Fink C, Risse F, Kauczor HU, Meinzer HP (2007) Implementation and evaluation of a new workflow for registration and segmentation of pulmonary MRI data for regional lung perfusion assessment. Phys Med Biol 52(5):1261–1275
18. Candemir S, Jaeger S, Palaniappan K, Musco JP, Singh RK, Xue Z, Karargyris A, Antani S, Thoma G, McDonald CJ (2014) Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration. IEEE Trans Med Imaging 33(2):577–590
19. Chae S-H, Lee J, Won C, Pan SB (2014) Lung segmentation using prediction-based segmentation improvement for chest tomosynthesis. Int J Biosci Biotechnol 6(3):81–90
20. Li C, Xu C, Gui C, Fox MD (2010) Distance regularized level set evolution and its application to image segmentation. IEEE Trans Image Process 19(12):3243–3254
21. Zhang W, Zeng S, Wang D, Xue X (2015) Weakly supervised semantic segmentation for social images. In: Proceedings of the IEEE computer society conference on computer vision pattern recognition, vol 07, 12-June, pp. 2718–2726
22. Papandreou G, Chen L-C, Murphy KP, Yuille AL (2015) Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In Proceedings of the ICCV, pp 1742–1750
23. Vezhnevets A, Buhmann JM (2010) Towards weakly supervised semantic segmentation by means of multiple instance and multitask learning. In: Proceedings of the IEEE computer society conference on computer vision pattern recognition, pp 3249–3256
24. Xu J, Schwing AG, Urtasun R (2014) Tell me what you see and i will show you where it is. In: 2014 IEEE conference on computer vision pattern recognition (CVPR), pp 3190–3197
25. Rajchl M, Lee MCH, Oktay O, Kamnitsas K, Passerat-palmbach J, Bai W, Kainz B, Rueckert D (2017) DeepCut: object segmentation from bounding box annotations using convolutional neural networks. IEEE Trans Med Imaging 36(2):674–683
26. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. ImageNet Chall, pp 1–10. arXiv:1409.1556
27. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. Proceedings of the IEEE computer society conference computer vision pattern recognition vol 07, 12-June, pp 1–9
28. Zaremba W, Sutskever I, Vinylas O (2014) Recurrent neural network regularization. Arxiv preprint arXiv:1409.2329
29. Jia Y, Shelhamer E, Dohanue J et al. (2014) Caffe: convolutional architecture fpr fast feature embedding. arXiv:1408.5093
30. Sutskever I, Martens J, Dahl GE (2013) On the importance of initialization and momentum in deep learning. In Jwml W&Cp, vol 28, issue 2010, pp 1139–1147
31. Bottou L (2012) Stochastic gradient descent tricks. In: Montavon G, Orr GB, Müller KR (eds) Neural networks: tricks of the trade. Lecture notes in computer science, vol 7700. Springer, Berlin, Heidelberg