CHAPTER NINE

Automatic lesion detection with three-dimensional convolutional neural networks

Qi Dou[1], Hao Chen[1], Jing Qin[2] and Pheng-Ann Heng[1]
[1]Department of Computer Science and Engineering, The University of Hong Kong, Sha Tin, Hong Kong
[2]School of Nursing, The Hong Kong Polytechnic University, Hung Hom, Hong Kong

9.1 Introduction

Automatic lesion detection in medical images has been a fundamental and crucial topic in the area of medical image analysis. Accurate and efficient localization of the lesions are essential for many clinical procedures, such as disease diagnosis decision-making, primary cancer screening, management of early treatment, etc. For example, the cerebral microbleeds serve as important diagnostic biomarkers for brain vascular diseases, and can potentially cause neurologic dysfunction and cognitive impairment [1,2]. The pulmonary nodules are critical indicators of primary lung cancer, and timely surgical intervention of nodules help dramatically increase the survival rate of patients [3,4].

The automatic detection tasks are, however, very challenging. The lesions in medical images have very small size (i.e., at a scale of millimeter), especially when a patient is at an early-stage cancer or other diseases. Moreover, these small lesions are sparsely distributed throughout the anatomical area. The widespread and unpredictable lesion locations make complete and accurate detection even more difficult. In addition, the lesions themselves present large variations in characteristics and contextual information. There also exist many hard mimics, which are normal tissues but heavily resemble the appearance of lesions in scanned medical images. These complicated intra-/interclass variances set further obstacles for a computer-aided detection system to achieve a high sensitivity with a low false positive rate.

Typically, the automatic lesion detection system consists of two steps: (1) candidate screening, which would sensitively screen candidates but receive many false positives, and (2) false positive reduction, which aims to remove the false positives and produce the final detection results. In the last generation of computer-aided detection systems,

Biomedical Information Technology
ISBN 978-0-12-816034-3, https://doi.org/10.1016/B978-0-12-816034-3.00009-2

the first stage has relied on rule-based methods, such as curvature computation, voxel clustering, intensity thresholding and morphological operation. The second stage commonly has employed various classifiers, such as support vector machine (SVM), decision tree, random forest, etc., on the basis of handcrafted features, which are heuristically designed to describe the key characteristics of lesions, such as the intensity, size, sphericity, texture and contexts. The representation capability of those used low-level features have limited the accuracy of previous computer-aided detection systems.

Recently, with the remarkable success of deep convolutional neural networks (CNN) in image processing [5,6], the deep learning based representations have been broadly employed in medical image computing. With the unique nature of high dimensionality in medical images (e.g., computed tomography (CT) and magnetic resonance (MR) imaging), how to effectively unleash the power of CNN on 3-D volumetric medical data requires elaborated researches. One straight-forward way is to employ 2-D CNN based on each single slice and process the slices sequentially. Apparently, this solution disregards the spatial information along the third dimension. Alternatively, aggregations of adjacent slices or orthogonal planes (i.e., axial, coronal and sagittal) are useful to enhance complementary spatial information in the 3-D space. Nevertheless, these solutions are still suboptimal, as the employed 2-D kernels are independent from each other and the repeated patterns along the third dimension are insufficiently modeled.

In this chapter, we present 3-D convolutional neural network (3-D CNN), which aims to tailor highly representative and discriminative features for volumetric medical data. We further introduce a 3-D CNN based cascaded two-step framework, to efficiently and accurately perform the task of lesion detection from medical images. Two distinct case studies using the developed system are demonstrated with state-of-the-art performance achieved. Our early works related to this chapter were published in Refs. [7–9].

9.2 3-D convolutional neural network

Basically, a CNN classification model alternatively stacks convolutional (C) and subsampling—e.g., max-pooling (M)—layers. In a C layer, small feature extractors (kernels) sweep over the topology and transform the input into feature maps. In an M layer, activations within a neighborhood are abstracted to acquire invariance to local translations. After several C and M layers, feature maps are flattened into a feature vector, followed by fully connected (FC) layers. Finally, a softmax classification layer yields the prediction probability. This section describes the 3-D CNN for medical image analysis, which also follows that fundamental construction.

9.2.1 3-D convolutional kernel

In a typical C layer, a feature map is produced by convolving the input with convolution kernels, adding a bias term, and finally applying a nonlinear activation function. By denoting the *i-th* feature map of the *l-th* layer as $\boldsymbol{h}_i^l$ and the *k-th* feature map of the previous layer as $\boldsymbol{h}_k^{l-1}$, a C layer is formulated as:

$$\boldsymbol{h}_i^l = \boldsymbol{\sigma}\left(\sum_k \boldsymbol{h}_k^{l-1} * \boldsymbol{W}_{ki}^l + \boldsymbol{b}_i^l\right), \tag{9.1}$$

where $\boldsymbol{W}_{ki}^l$ and $\boldsymbol{b}_i^l$ are the filter and bias term connecting the feature maps of adjacent layers, the * denotes the convolution operation and the $\boldsymbol{\sigma}(\bullet)$ is the element-wise nonlinear activation function.

In 2-D natural image processing, the input of CNN usually consists of three color channels (i.e., RGB). Inspired by this, the most straightforward way to adapt 2-D CNN to support volumetric medical image processing is to replace the color channels with adjacent slices of the volume. As shown in Fig. 9.1A, given a volumetric image of size $X \times Y \times Z$, when we employ this scheme to generate a feature map, we first need to split the input volume along the third dimension into Z isolated slices, and then feed these Z isolated slices into the network. Correspondingly, Z 2-D kernels are formed, with each single slice swept over by a unique kernel (see the red line). However, this scheme cannot sufficiently leverage the spatial information, since the Z 2-D kernels are different from each other. In other words, due to the absence of kernel sharing across the third dimension, the encoded volumetric spatial information is inevitably deficient.

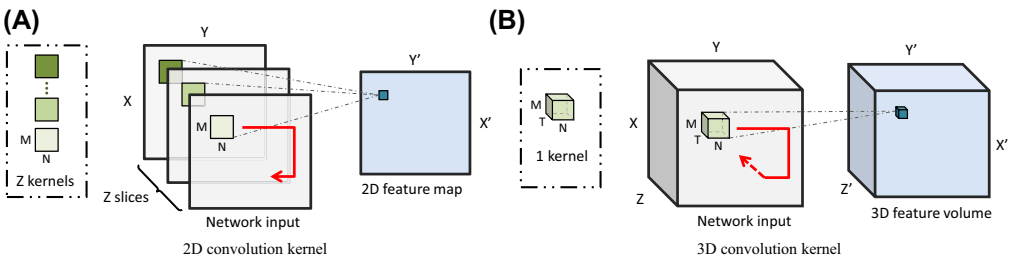

Figure 9.1 Comparison of using two- and three-dimensional convolution kernels given volumetric image with size of $X \times Y \times Z$ in terms of network input, kernel behavior and generated feature map. Red lines represent the moving direction of kernels—i.e., sweeping over the two- and three-dimensional topologies, respectively. (A) With the two-dimensional convolution (kernel size of $M \times N$), the volume is first split into Z isolated slices along the third direction and these slices are input to the network. Each generated feature map is a two-dimensional patch. (B) With the three-dimensional convolution (kernel size of $M \times N \times T$), the entire volume is input to the network. Each generated feature map is a three-dimensional volume. (Note that kernel sizes M, N and T need not to be equal. Best viewed in color.)

Learning feature representations from all three dimensions is vitally important for biomarker detection tasks from volumetric medical images. In this regard, we propose to set up the 3-D convolution kernel, in the pursuance of encoding richer spatial information of the volumetric data. In this case, the feature maps are 3-D blocks instead of 2-D patches (we call them *feature volumes* hereafter). As shown in Fig. 9.1B, given the same volumetric image of size $X \times Y \times Z$, when we employ a 3-D convolution kernel to generate a 3-D feature volume, the input to the network is the entire volumetric data. Consequently, a 3-D kernel is formed and it sweeps over the whole 3-D topology (see the red line). By leveraging the kernel sharing across all three dimensions, the network can take full advantage of the volumetric contextual information.

Generally, the following equation formulates the exploited 3-D convolution operation in an element-wise manner:

$$\boldsymbol{u}_{ki}^{l}(x,y,z) = \sum_{m,n,t} \boldsymbol{h}_{k}^{l-1}(x-m,y-n,z-t)\boldsymbol{W}_{ki}^{l}(m,n,t), \tag{9.2}$$

where $\boldsymbol{W}_{ki}^{l}$ denotes the 3-D kernel in the *l-th* layer which convolves over the 3-D feature volume $\boldsymbol{h}_{k}^{l-1}$, the $\boldsymbol{W}_{ki}^{l}(m,n,t)$ is an element-wise weight in the 3-D convolution kernel. Following Eq. (9.1) and Eq. (9.2), the 3-D feature volume $\boldsymbol{h}_{i}^{l}$ is obtained by summing over the 3-D convolution kernels:

$$\boldsymbol{h}_{i}^{l} = \boldsymbol{\sigma}\left(\sum_{k} \boldsymbol{u}_{ki}^{l} + \boldsymbol{b}_{i}^{l}\right). \tag{9.3}$$

9.2.2 3-D CNN hierarchical model

With the 3-D convolutional kernel and the layer, we can hierarchically construct a deep 3-D CNN model by stacking the C, M, and FC layers, as shown in Fig. 9.2. Specifically, in the C layer, a series of 3-D feature volumes are produced. In the M layer, the max-pooling operation, or any other down-sampling operation, is also conducted in the 3-D fashion—i.e., the feature volumes are subsampled based on a cubic neighborhood. In the following FC layer, the 3-D feature volumes are flattened into a feature vector as its input. The ultimate output layer employs the softmax activation to yield the prediction probabilities for the input image.

During 3-D CNN implementation, the nonlinear activation function (e.g., the ReLU or LeakyReLU) is used in C and FC layers. The 3-D convolution kernels are randomly initialized from the Gaussian distribution and trainable parameters in the network are updated using the standard backpropagation with stochastic gradient descent or other advanced optimizers. The loss function is derived according to the specific task aiming to solve, for example, the cross-entropy loss for classification tasks, the Dice loss for segmentation tasks, or the adversarial loss for generative model. The developed useful

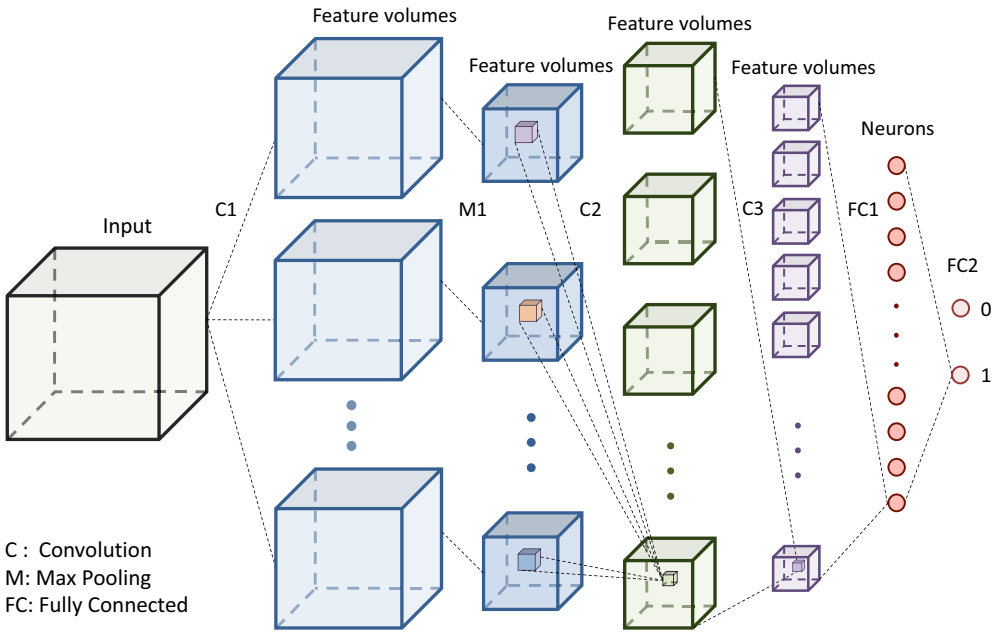

Figure 9.2 The hierarchical architecture of the 3-D CNN model.

strategies which can benefit the learning procedure, such as dropout, batch normalization, and residual connection, can be seamlessly embedded into the 3-D convolutional neural network.

9.3 Efficient fully convolutional architecture

One of the main concerns about exploiting CNN in medical imaging domain lies in the time performance, as many medical applications require prompt responses for further diagnosis and treatment. The situation is more rigorous when processing volumetric medical data. Directly applying 3-D CNNs to detect lesions using the traditional sliding window strategy is usually impracticable, especially when the input volumetric images are acquired with high resolutions, because thousands or even millions of 3-D block samples need to be analyzed. In most biomarker detection applications, the targets are usually sparsely distributed throughout the volume, such as the microbleed in the 3-D brain MR image. To this end, one promising solution for detection is to first obtain the candidates with a high sensitivity and then perform fine-grained discrimination only on these candidates, so that the computational cost can be greatly reduced. Previous work proposed to retrieve lesion candidates (also called regions-of-interest in some papers) by employing local statistical information, including size, intensity, shape and

other geometric features [10–12]. However, due to the large variations of lesions, only relying on these statistical values or handcrafted features are not effective enough.

We propose to use 3-D CNN to robustly screen candidates by leveraging high-level spatial representations of lesions learned from a large number of 3-D training samples. However, we still face the challenge of time performance when employing 3-D CNN to retrieve candidates with the traditional sliding window strategy. To this end, inspired by the 2-D fully convolutional networks (FCNs) [13], we propose to extend the strategy into a 3-D format for efficient retrieval of the candidates from volumetric medical images. The proposed 3-D FCN can take an arbitrary-sized volume as input and produce a 3-D score volume within a single forward propagation, and hence can greatly speed up the candidate retrieval procedure without damaging the sensitivity.

9.3.1 Fully convolutional transformation

In a 3-D CNN architecture, both the convolutional and down-sampling layers can process arbitrary-sized input, where convolution or max-pooling kernels sweep over the input and generate the corresponding-sized output. However, the traditional FC layers flatten the feature volumes into vectors thus dismissing the spatial relationships. These FC layers then utilize vector-matrix multiplications to generate the output, as shown in the following:

$$\boldsymbol{h}^l = \boldsymbol{\sigma}\left(\boldsymbol{W}^l \boldsymbol{h}^{l-1} + \boldsymbol{b}^l\right), \tag{9.4}$$

where $\boldsymbol{h}^{l-1} \in \mathbb{R}^P$ and $\boldsymbol{h}^l \in \mathbb{R}^Q$ are the feature vectors in the *(L-1)-th* and the *l-th* FC layers, respectively, $\boldsymbol{W}^l \in \mathbb{R}^{Q \times P}$ is the weight matrix and $\boldsymbol{b}^l$ denotes the bias term.

In traditional CNN, once trained, the weight $\boldsymbol{W}^l$ is with a fixed shape, and hence the FC layer has fixed input/output sizes. As a result, a network with traditional FC layers requires that the initial inputs have a fixed size. For example, when the network is trained based on 3-D samples of size $16 \times 16 \times 10$, errors will arise if we input a test sample of size $20 \times 16 \times 10$, due to the shape mismatch in the first dimension.

In this regard, we equivalently rewrite the FC layers into the following convolutional format:

$$\boldsymbol{h}_q^l = \boldsymbol{\sigma}\left(\sum_p \boldsymbol{h}_p^{l-1} * \boldsymbol{W}_{pq}^l + \boldsymbol{b}_q^l\right), \tag{9.5}$$

where each neuron in the FC layer is regarded as a $1 \times 1 \times 1$ feature volume, $\boldsymbol{W}_{pq}^l \in \mathbb{R}^{1\times1\times1}$ is the 3-D kernel and the * is the 3-D convolution operation described in Eq. (9.2). In this way, the vector-matrix multiplications are formulated as convolution operations with $1 \times 1 \times 1$ kernels. With the FC layers converted into convolutional layers, the network could therefore support arbitrary-sized input.

9.3.2 3-D score volume generation

During the training phase, a traditional 3-D CNN model is learned. Once training is done, to acquire the 3-D FCN model, the FC layers in the traditional 3-D CNN are transformed into the convolutional fashion. More specifically, the multiplication matrix $\boldsymbol{W}^l \in \mathbb{R}^{Q \times P}$ is reshaped into a 5D tensor $\boldsymbol{W}^l \in \mathbb{R}^{Q \times 1 \times P \times 1 \times 1}$ (the dimensions are ordered for the ease of implementation), and hence the weight matrix is converted into a series of convolution kernels. During the testing phase, the 3-D FCN model directly inputs a volume and outputs a 3-D score volume (with reduced resolution compared with the original input size). The value at each location of score volume indicates the probability of being a lesion.

There are some implementation issues needed to be handled when developing the 3-D FCN model. Specifically, when converting the traditional FC layers into the convolutional fashion by casting the 2-D multiplication matrix ($\mathbb{R}^{Q \times P}$) into the 5D tensor ($\mathbb{R}^{Q \times 1 \times P \times 1 \times 1}$), we should precisely maintain the spatial correlation. In addition, during whole volume testing, we need to ensure the dimension consistency in the logistic regression layer, where the feature volumes are first flattened into vectors, then applied to the softmax function and finally reshaped back to form the 3-D score volume. One alternative practice is to directly train the model in an FCN format, such that there are only convolutional and down-sampling layers in the network, without any FC layer.

9.3.3 Score volume index mapping

Due to successive layers of convolution and down-sampling operations, the size of the generated 3-D score volume is reduced compared with the original input. Actually, the 3-D score volume is a coarse version of the voxel-wise predictions which are produced by the sliding window strategy. Meanwhile, the locations on this coarse score volume can be traced back to the coordinates on the original input space.

Since all three dimensions follow the same index mapping mechanism, we demonstrate the mapping process with one dimension. In our formulation, indices are numbered from zero. Generally, for each C or M layer (supposing nonpadding convolution and nonoverlap pooling) in the model, the index mapping procedure with convolution or max-pooling operation can be calculated by:

$$x' = d \cdot x + \left\lfloor \frac{c-1}{2} \right\rfloor, \tag{9.6}$$

where x' and x denote the coordinates before and after the convolution or max-pooling operation; d and c represent the stride and kernel size, respectively; the $\lfloor \cdot \rfloor$ represents the floor function.

When mapping the location x_s in the coarse score volume back through the architecture toward the location x_o in the original input volume, we successively deduce

Table 9.1 The architecture of three-dimensional FCN screening model.

Layer	Kernel size	Stride	Output size	Feature volumes
Input	–	–	$16 \times 16 \times 10$	1
C1	$5 \times 5 \times 3$	1	$12 \times 12 \times 8$	64
M1	$2 \times 2 \times 2$	2	$6 \times 6 \times 4$	64
C2	$3 \times 3 \times 3$	1	$4 \times 4 \times 2$	64
C3	$3 \times 3 \times 1$	1	$2 \times 2 \times 2$	64
FC1	$2 \times 2 \times 2$	1	$1 \times 1 \times 1$	150
FC2	$1 \times 1 \times 1$	1	$1 \times 1 \times 1$	2

the index mapping procedures along all intermediate convolution and max-pooling layers until the initial input layer. For example, based on the screening network architecture shown in Table 9.1, for each position index x_s in the coarse score volume, we can obtain its corresponding index x_o in the original input as follows:

$$x_o = \left\lfloor \frac{c_1 - 1}{2} \right\rfloor + \left\lfloor \frac{c_2 - 1}{2} \right\rfloor + d_2 \bullet \left(x_s + \left\lfloor \frac{c_3 - 1}{2} \right\rfloor + \left\lfloor \frac{c_4 - 1}{2} \right\rfloor + \left\lfloor \frac{c_5 - 1}{2} \right\rfloor + \left\lfloor \frac{c_6 - 1}{2} \right\rfloor \right) = D \bullet x_s + C, \tag{9.7}$$

where, according to the network architecture, $c_1 = 5$, $c_2 = 2$, $d_2 = 2$, $c_3 = 3$, $c_4 = 3$, $c_5 = 2$, $c_6 = 1$, and we can calculate $D = 2$ and $C = 6$ for the $\boldsymbol{X}$ dimension.

As shown in Fig. 9.3, with this mechanism, each location in the 3-D score volume can be mapped back to the centroid of the corresponding receptive field of the neuron.

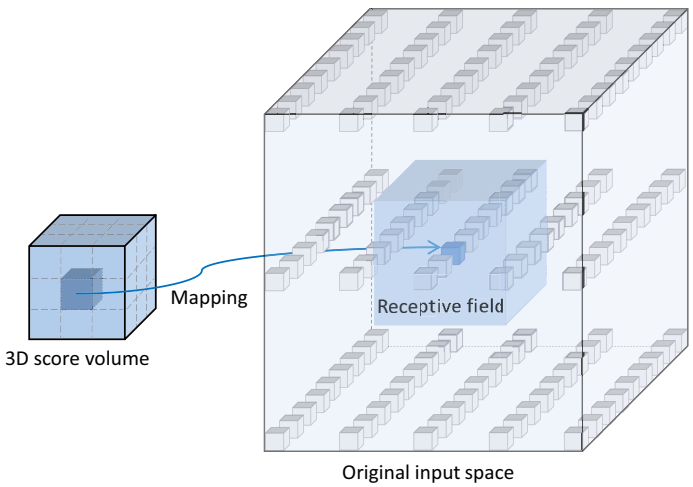

Figure 9.3 The mapping from the three-dimensional score volume onto the original input space.

Equivalently, if the cubic patch centered on the traced position is input to the traditional 3-D CNN, the prediction probability is indeed the value at the location on the coarse score volume. Consequently, the prediction scores are sparsely mapped back onto the input volume, and regions with high probabilities are retrieved as potential candidates.

9.4 Two-stage cascaded framework for detection

Based on 3-D CNN and the fully convolutional architecture, we build a detection network modeling. Specifically, we use a 3-D FCN model and a 3-D CNN model tailored for two different stages and integrate them into an efficient and robust detection framework. In this cascaded framework for lesion detection, each stage serves its own mission. The candidate screening stage with the 3-D FCN aims to accurately reject the background regions and rapidly retrieve a small number of potential candidates. The false positive reduction stage with the 3-D CNN focuses only on the screened set of candidates to further single out the true lesions from challenging mimics.

9.4.1 Candidate screening stage

The workflow of the candidate screening stage is presented in Fig. 9.4, including both training and testing phases. During the training phase, the positive samples are extracted from lesion regions and with augmentations to expand the training database. In practice, the network is trained with three substeps. We start from training an initial 3-D CNN with randomly selected nonlesion regions throughout the image as negative samples. Next, we add false positive samples acquired by applying the initial model on the training dataset. Finally, the initial model is fine-tuned with the enlarged training database which consists of positives, randomly selected negatives and supplemental false positives. In this way, the discrimination capability of the network is further enhanced. During the testing phase, the 3-D FCN model takes the whole volume as input and generates the corresponding coarse 3-D score volume.

Considering that the produced score volume could be noisy, we utilize the local nonmax suppression in a 3-D fashion as the postprocessing. Locations in the 3-D score volume are then sparsely traced back to coordinates in the original input space, according to the index mapping process. Finally, regions with high prediction probabilities are selected as the potential candidates.

9.4.2 False positive reduction stage

In this stage, 3-D small blocks are cropped centered on the screened candidate positions. The extracted 3-D candidate regions are classified by a newly constructed 3-D CNN model, to remove the remaining false positives. Note that the randomly selected nonlesion samples are not strongly representative, especially when we aim to distinguish true

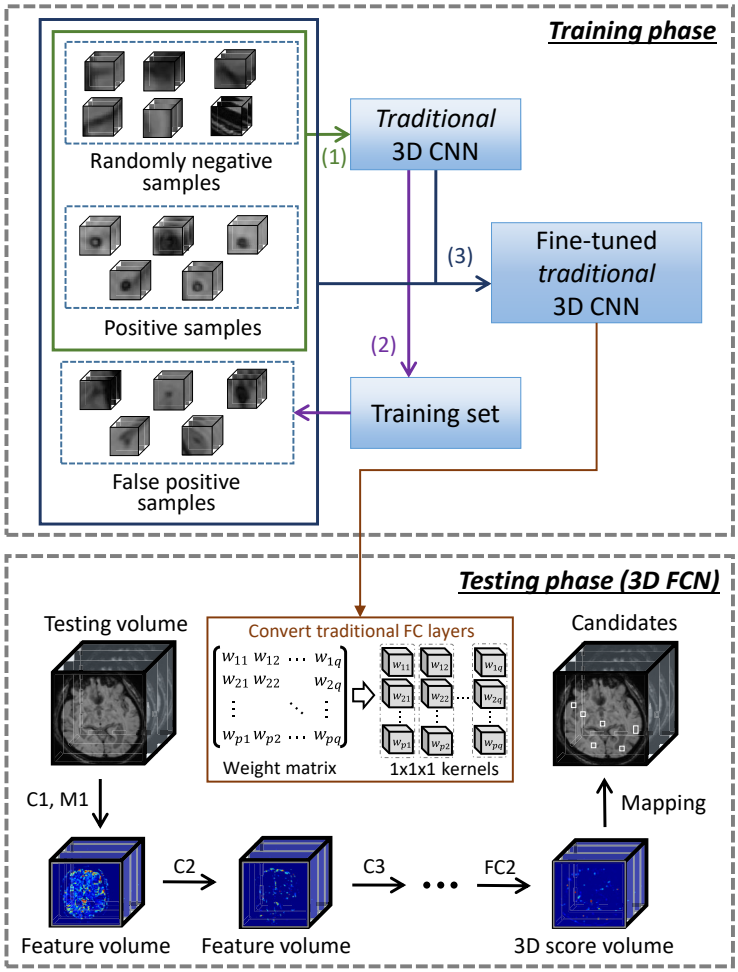

Figure 9.4 Illustration of the workflow of the screening stage. The training phase is conducted in three substeps: (1) train an initial *traditional* 3-D CNN with positive samples and randomly selected negative samples; (2) apply the initial model on training set and obtain false positive samples to enlarge the training database; (3) fine-tune the initial *traditional* 3-D CNN model with the enlarged database to strengthen its discrimination capability. Once training is done, the traditional FC layers are converted into the convolutional fashion (as shown in the brown box). During the testing phase, the 3-D FCN takes a whole volume as input, extracts representative feature volumes and finally produces a 3-D score volume to retrieve candidates.

lesions from their mimics. To generate representative samples and improve the discrimination capability of the 3-D CNN model, the obtained false positives (which take very similar appearance as lesions) on the training set in the screening stage are taken as negative samples when training the 3-D CNN in the second stage. The model ensemble can be employed in this stage to further improve the performance.

9.5 Case study I: cerebral microbleed detection in brain magnetic resonance imaging

9.5.1 Background of the application

Cerebral microbleeds (CMBs) refer to the small foci of chronic blood products, composed of the hemosiderin deposits that leak through pathological brain blood vessels [2]. This lesion is prevalent in patients with cerebrovascular and cognitive diseases (such as stroke and dementia), and also present in healthy aging populations. The existence of cerebral microbleeds and their distribution patterns have been recognized as important biomarker for diagnosis of the cerebrovascular diseases. For example, the CMB lobar distribution would suggest probable cerebral amyloid angiopathy, and the deep hemispheric or infratentorial microbleeds may imply probable hypertensive vasculopathy. The existence of CMBs would bring an increase in the risks of symptomatic intracerebral hemorrhage and recurrent ischemic stroke [14]. Furthermore, these lesions could structurally damage their nearby brain tissues, and further cause cognitive impairment and neurologic dysfunction [15]. In these regards, reliable detection of the CMB is crucial for cerebral diagnosis and may guide physicians in determining which drug to choose for necessary treatment, such as for stroke prevention.

Modern advances in MR imaging technologies make the paramagnetic blood products be more sensitive to screening [16], and hence facilitate the recognition of CMBs. As shown in Fig. 9.5, the cerebral microbleed is radiologically visualized as rounded hypointensities of small size within the susceptibility weighted imaging (SWI) MR data (refer to the yellow rectangle). In general, the clinical routine to detect the CMB is based on visual inspection and manual localizing, which is laborious and error-prone. Alternatively, computer-aided detection systems can assist to relieve the workload

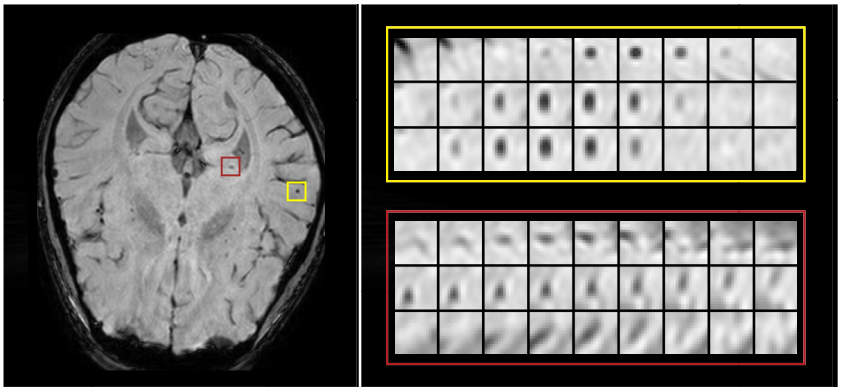

Figure 9.5 Illustration of a CMB and a CMB mimic denoted with the yellow and red rectangles, respectively. In each of the big rectangle, the rows demonstrate adjacent slices in axial, sagittal and coronal planes, from top to down. The importance of 3-D information can be observed.

on radiologists and also improve clinical efficiency. However, the automatic detection of CMBs can encounter several challenges: (1) the small size of lesions; (2) the widespread distributed locations of lesions; (3) the existence of hard mimics (e.g., flow voids, calcification and cavernous malformations) which would resemble the appearance of CMBs and heavily impede the detection process.

9.5.2 Dataset, preprocessing and evaluation metrics

Our employed dataset includes 320 SWI images with 1149 CMBs scanned from a 3.0T Philips Medical System with 3-D spoiled gradient-echo sequence using venous blood oxygen level dependent series with the following parameters: repetition time 17 ms, echo time 24 ms, volume size $512 \times 512 \times 150$, in-plane resolution 0.45×0.45 *mm*, slice thickness 2 *mm*, slice spacing 1 *mm* and a 230×230 mm^2 field of view. The subjects were from two separated groups: 126 cases with stroke (mean age $\pm$ standard deviation: 67.4 ± 11.3) and 194 cases of normal aging (mean age $\pm$ standard deviation: 71.2 ± 5.0).

The dataset was labeled by an experienced rater and was verified by a neurologist following the guidance of the Microbleed Anatomical Rating Scale [17]. We employed the Pearson correlation coefficient (PCC) to assess the interobserver agreement between the two raters [18]. Due to the large dataset and expensive manual annotation efforts, we tested the interobserver agreement with a subset of 20 subjects (including 10 cases with stroke and 10 cases of normal aging). The PCC turned out to be 0.91 ($P < .01$), which indicates a high degree of agreement between the two raters. Overall, a total of 1149 CMBs were annotated from the whole dataset and regarded as the ground truth in our experiments.

We randomly divided the dataset into three parts for training (230 cases with 924 CMBs), validation (40 cases with 108 CMBs) and testing (50 cases with 117 CMBs). In the preprocessing step, the volume intensities are normalized to the range of [0,1] with

$$I' = \frac{I - I_{min}}{I_{max} - I_{min}}, \tag{9.8}$$

where I and I' represent the original and normalized intensity value, respectively. The $I_{\max}$ is the maximum intensity value after trimming the top 1% grayscale intensities and the $I_{\min}$ is the minimum grayscale value of the volume.

We employed three metrics to quantitatively evaluate the performance on the task of CMB detection, including sensitivity (S), precision (P) and the average number of false positives per subject (FP_{avg}):

$$\mathrm{S} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \mathrm{FP}_{\mathrm{avg}} = \frac{\mathrm{FP}}{\mathrm{N}}, \tag{9.9}$$

where TP, FP and FN denote the total number of true-positive, false-positive and false-negative detection results, respectively. The N represents the number of subjects in the testing dataset.

9.5.3 Experimental results

For the first stage, we compared the candidate screening performance of the 3-D CNN based method with two state-of-the-art approaches, which utilize low-level statistical features [10,19]. We implemented these comparison approaches and employed them on our testing dataset. The results are listed in Table 9.2. The values of sensitivity mean the percentage of successfully retrieved CMBs while the values of FP_{avg} describe the number of remaining false positives per subject. The fewer false positives produced, the more powerful discrimination capability a screening method has. The proposed 3-D FCN model achieves the highest sensitivity with fewest average number of false positives, which highlights the efficacy of the proposed method.

Note that our method outperforms the other two methods by a large margin. We have also recorded the average time for screening each subject and the results are listed in Table 9.2. From the clinical perspective, the time performance of our method is satisfactory; processing a whole volume with a size of $512 \times 512 \times 150$ takes around 1 min. The method of [10] is slower than ours because it calculates local thresholds using a voxel-wise sliding window way. In contrast, the method of [19] merely exploits global thresholding on intensity and size, hence it has a much faster screening speed.

For the candidate screening stage, the retrieval accuracy is vitally important, because we cannot refind the CMBs that are missed by the screening stage in the following discrimination stage. Although [19] is faster, we achieved around 8% increase in sensitivity and reduced the number of FP_{avg} from 935.8 to 282.8, when compared with this method. These results provide a much more reliable basis for further fine discrimination. By employing the 3-D FCN, our method achieves a good balance between retrieval accuracy and speed. Typical candidate screening results by the proposed 3-D FCN are shown in Fig. 9.6. It is observed that high values on the score volume mostly correspond to CMB lesions. In addition, most of the backgrounds have been successfully suppressed as zeros. After thresholding, only a small number of candidates are obtained (see those white rectangles), which dramatically reduces the computational workload in the following stage.

Table 9.2 Comparison of different CMB lesion candidate screening methods.

Methods	Sensitivity	FP_{avg}	Time per subject (s)
Barnes et al. [10]	85.47%	2548.2	81.46
Chen et al. [19]	90.48%	935.8	**12.00**
3-D FCN model	**98.29%**	**282.8**	64.35

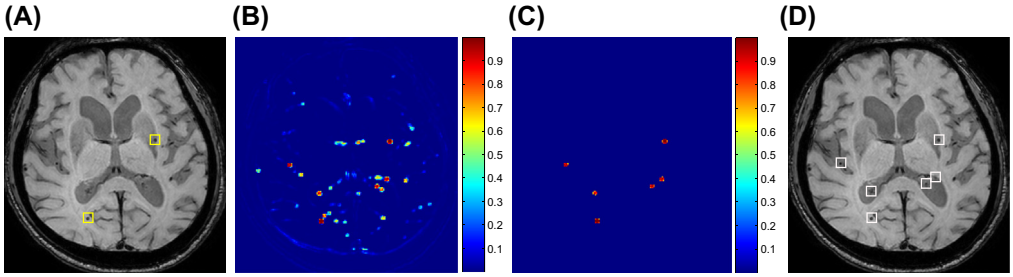

Figure 9.6 Typical results of the three-dimensional FCN screening model with score volume projection onto the axial plane. (A) Raw data with true CMBs (yellow rectangles). (B) two-dimensional projection of the score volume generated with FCN. (C) two-dimensional projection of the post-processed score volume. (D) Retrieved candidates (white rectangles). Best viewed in color.

Table 9.3 The architecture of three-dimensional CNN used for false positive reduction.

Layer	Kernel size	Stride	Output size	Feature volumes
Input	–	–	$20 \times 20 \times 16$	1
C1	$7 \times 7 \times 5$	1	$14 \times 14 \times 12$	32
M1	$2 \times 2 \times 2$	2	$7 \times 7 \times 6$	32
C2	$5 \times 5 \times 3$	1	$3 \times 3 \times 4$	64
FC1	–	–	$1 \times 1 \times 1$	500
FC2	–	–	$1 \times 1 \times 1$	100
FC3	–	–	$1 \times 1 \times 1$	2

In the second stage, we independently trained three models using the network architecture presented in Table 9.3. The differences of the three convolutional networks lie in the random weights initialization states and the training epochs. The deep network with a large number of parameters usually comes with a low bias and a large variance. By averaging multiple models with different weight initializations and different early stopping conditions, the model variance can be reduced, and thus the discrimination capability is further boosted [21].

We compared the performance of our lesion detection method with three other approaches. These methods were implemented on our dataset for direct comparison. The first one utilized handcrafted features based on shape and intensity information [10]. The second one constructed a random forest classifier with low-level features, which is widely used for 3-D object detection applications in medical imaging. The third one utilized a 2-D CNN and process the concatenated 2-D features with an SVM [19].

Table 9.4 shows the comparison results of different lesion detection methods and the FROC curves of these methods are presented in Fig. 9.7. It is clearly observed that our proposed methods outperform the other three comparison approaches by a significant margin with the highest detection sensitivity as well as the fewest false positive

Table 9.4 Evaluation of cerebral microbleed detection results.

Methods	Sensitivity	Precision	FP_{avg}
Barnes et al. [10]	64.96%	5.13%	28.10
Random forest [20]	85.47%	17.24%	9.60
2-D−CNN−SVM [19]	88.03%	22.69%	7.02
Ours (single)	92.31%	42.69%	2.90
Ours (average)	**93.16%**	**44.31%**	**2.74**

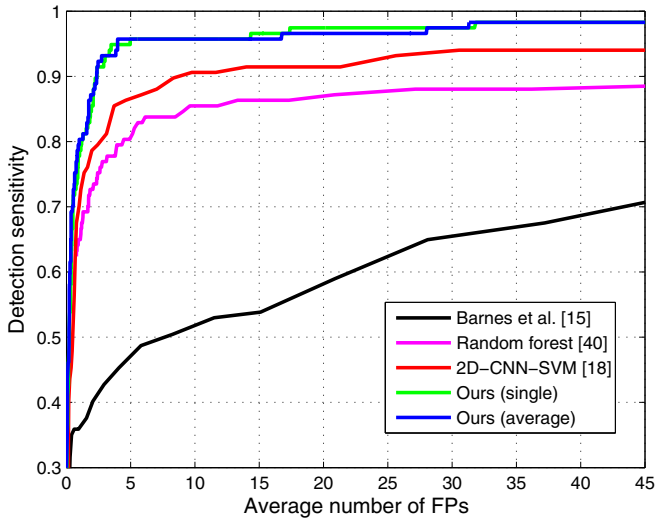

Figure 9.7 Comparison of FROC curves of different methods. The top two lines are results produced by our 3-D CNN based cascaded frameworks.

predictions. Although the 2-D−CNN−SVM method can not sufficiently leverage the 3-D spatial characteristics of the microbleed lesions, the high-level features even encoding limited spatial information obtained better detection performance than the other two methods employing traditional low-level features. The comparison results between our 3-D CNN based methods and the 2-D−CNN−SVM approach demonstrate that our framework benefits from the high-level features which can encode richer spatial information by taking advantage of the 3-D convolutional architectures. Utilizing the model average in the second discrimination stage can further improve the overall lesion detection performance. Fig. 9.8 present typical examples of successfully detected CMBs. In Fig. 9.8 left, there are a number of hard mimics (white rectangles) around the two true CMBs (green rectangles). Our method is able to precisely distinguish them. In Fig. 9.8 right, the two CMBs are sparsely distributed in the volume with one of them locating at almost the boundary of the volume. In this condition, our method can still accurately detect both of them.

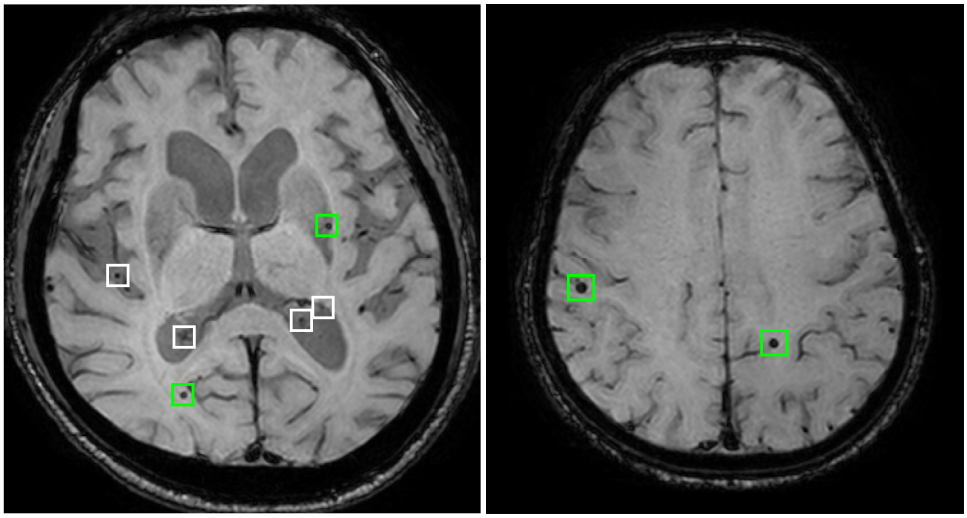

Figure 9.8 Examples of CMB detection results (viewed in axial planes). Green rectangles denote the correctly detected CMBs and white rectangles denote the removed false positive candidates by our method.

9.6 Case study II: lung nodule detection in chest computed tomography

9.6.1 Background of the application

Lung cancer has been among the leading cause of cancer deaths worldwide, and early detection of the lung nodules from low-dose CT scans is crucial for diagnosis of primary lung cancer and arrangement of necessary early treatment. In radiology scans, the pulmonary nodules are visible as small anatomical structures that are roughly spherical opacities within the pulmonary interstitium images [22]. Based on reliable detection of primary nodules, radiologists and surgeons can perform the size measurements and appearance characterizations for diagnosis of cancer malignancy [4] and, if necessary, conduct timely surgical intervention to increase the survival chances of the patients [3,23]. Annual lung cancer screening for those high-risk populations, such as smokers, has already been implemented in some countries, acquiring enormous CT data for clinical radiologists to analyze. It would be quite difficult, if not impossible, to manually screen the CT scans considering the huge requirement of manpower and the time costs.

Automated recognition of lung nodules in thoracic CT images is, however, among the most challenging problems in computer-aided detection [24]. First, the lung nodules come with large variations in sizes, shapes and locations [25], as presented in the green rectangle in Fig. 9.9. In addition, the contextual environments around the pulmonary

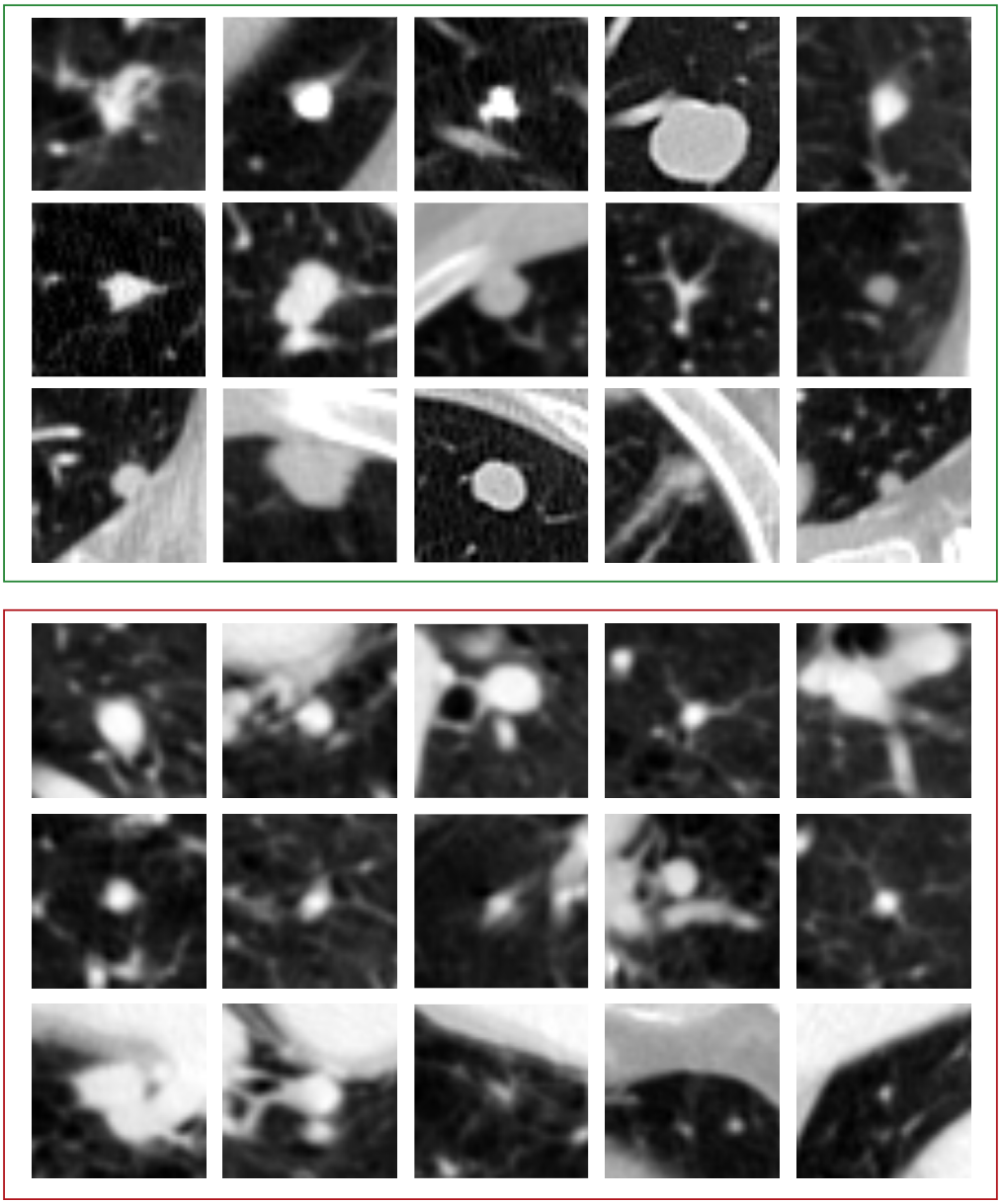

Figure 9.9 Examples of the pulmonary nodules with various sizes, shapes and locations (green rectangle), and the false positive candidates (red rectangle) which carry similar appearance and make the task challenging. Each example is a representative 2-D transverse plane of one nodule.

nodules are often diversified for different categories of the nodules, such as solitary nodules, ground-glass opacity nodules, cavity nodules and pleural nodules [26]. Second, some false positive candidates may carry very similar morphological appearance to the true lung nodules, as illustrated in the red rectangle in Fig. 9.9. The existence of these hard mimicking regions would heavily hinder the lesion detection process.

9.6.2 Improved learning strategy

Online Sample Filtering Strategy. For the candidate screening stage, we also employ the 3-D FCN, as it can not only encode rich volumetric spatial information to extract high-level representations for accurate candidate retrieval, but also rapidly generate the probability predictions in a volume-to-volume manner. More specifically, we construct a binary classification 3-D model which is consisted of five convolutional layers and one max-pooling layer. The screening network is learned with small 3-D patches of nodules and nonnodules and tested on the entire CT image in a fully convolutional manner (i.e., inputing the whole volumetric image and directly obtaining a 3-D score volume). In the following, we can retrieve lesion candidates based on the score volume with the suspicious probability of each location indicated, following the aforementioned mechanism of index mapping in Eq. (9.6).

However, given the severe imbalance between hard and easy samples, it is challenging to train the network and achieve a high-quality score volume. On one hand, an overwhelming amount of background samples are very easy to be recognized thus contribute little to model optimization. On the other hand, the number of hard samples (for example the mimics) is quite small, but they are challenging to be distinguished and therefore considered to be more informative for learning. Actually, the situation of sample imbalance is a very common problem in many biomedical detection tasks. Previous boosting methods would consecutively establish an ensemble of learners, where the misclassified hard samples from the former model were traced to train the next model, just as how we did for CMB detection. These methods repeatedly test the obtained model on all the training data, which would complicate the entire training process and incur additional computations (Fig. 9.10).

To tackle this problem, we further propose an online sample filtering scheme which can dynamically increase the proportion of hard training samples, borrowing the spirit of hard example mining originally employed for natural object detectors [27]. Superior to those previous boosting methods, our proposed scheme can select the hard examples on-the-fly during the stochastic gradient descent process, and it neither interrupts the normal learning process nor engages additional testing computations.

Our online sample scheme is constructed based on the observation that those hard samples usually produce higher classification loss compared with those easier ones. In this regard, we can dynamically obtain the hard samples based on the loss in every forward

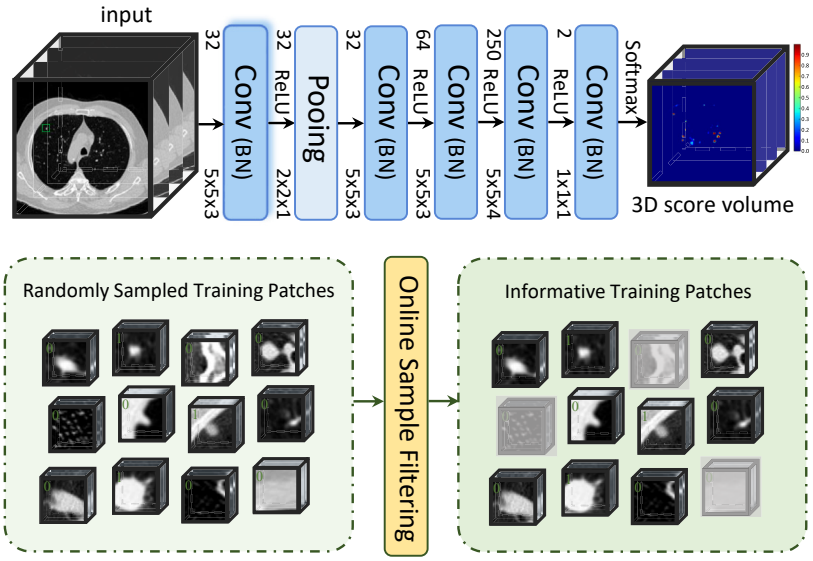

Figure 9.10 Candidate screening with the strategy of online sample filtering.

propagation during training. During the implementation, we randomly extract the initial training samples with a large batch size. After forward propagation of each batch, the samples are sorted by their loss, and we take the top 50% samples on which the current network performs worst as hard samples. Meanwhile, we randomly retain half of the remaining low-loss samples as easy samples. These are fairly reasonable choices to design the learning strategy, considering the balance for training patches. Finally, we exclude those less-informative examples from the current iteration of the optimization. By performing these online modifications to the stochastic gradient descent procedure, our proposed scheme is intuitive yet effective to train the model and speed up the convergence rate.

Multi-task Learning with Hybrid Loss. Aiming to accurately recognize the true pulmonary nodules from screened candidates, we exploit a 3-D CNN equipped with a 3-D variant of deep residual learning technique and a designed hybrid-loss objective function. We establish a modularized 3-D residual unit $\boldsymbol{x}_{\text{out}} = \boldsymbol{x}_{\text{in}} + F(\boldsymbol{x}_{\text{in}}, \{W_k\})$, where the $\boldsymbol{x}_{\text{in}}$ and $\boldsymbol{x}_{\text{out}}$ are its input and output; the $F(\bullet)$ represents 3-D residual transformation—i.e., a stack of convolutional, batch normalization and ReLU layers that are associated with the collection of parameters $\{W_k\}$. It has been evidenced that the utilized residual units can boost the information flow within the neural network and hence benefit the optimization.

Besides the small interclass variation between the true lung nodules and hard mimics, another challenge is that the proposal positions from 3-D FCN might deviate from the

ground truth pulmonary nodule centroids, due to the translation invariance inherited in the network. This would lead to shift of true lung nodule locations when we center at the FCN proposal positions to crop sample patches for processing in the second stage for false positive reduction. Under this setting, further leveraging localized annotations—i.e., where and how large a lung nodule is—can be beneficial to explicitly guide the learning to particularly focus on the targeted lesion regions. With these considerations, we design a novel hybrid-loss objective function, which jointly considers both classification errors and localized information. The constructed network simultaneously optimizes a classification branch and a regression branch, by sharing the network weights in early layers.

With a set of N training pairs of samples $\left\{\left(X^i, Y^i, G^i\right)\right\}_{i=1,\ldots,N}$, the shared early-layer parameters W_s and the classification branch parameters W_{cls} in the residual network, the classification loss is formulated as the negative log-likelihoods as follows:

$$\mathscr{L}_{\text{cls}} = -\frac{1}{N}\sum_i \log p\left(Y^i \middle| X^i, W_{\text{s}}, W_{\text{cls}}\right). \tag{9.10}$$

For the regression branch accounting nodule size and location, considering that we are targeting 3-D objects, our localization ground truth $G^i = \left(G_x^i, G_y^i, G_z^i, G_d^i\right)$ is represented by four parameters, with $\left(G_x^i, G_y^i, G_z^i\right)$ and G_d^i respectively denoting the centroid and diameter of the nodule. Denoting the position proposed via 3-D FCN by $P^i = \left(P_x^i, P_y^i, P_z^i\right)$, and the second stage cropped patch size by $S=(S_x, S_y, S_z)$, we define the continuous-valued regression target $T^i = \left(T_x^i, T_y^i, T_z^i, T_d^i\right)$ as follows:

$$\begin{aligned} T_x^i &= \frac{2\left(G_x^i - P_x^i\right)}{S_x}, T_y^i = \frac{2\left(G_y^i - P_y^i\right)}{S_y}, \\ T_z^i &= \frac{2\left(G_z^i - P_z^i\right)}{S_z}, T_d^i = \log\left(\frac{G_d^i}{\sqrt{S_x^2 + S_y^2 + S_z^2}}\right), \end{aligned} \tag{9.11}$$

where T^i specifies a scale-invariant translation and log-space size shift which is relative to the cropped patch size S. Considering the candidate proposal P^i is close to the ground truth lung nodule centroid, we divide their relative distance with half of the patch size for normalization purpose. Denoting the output of the regression branch by $\widehat{T}^i = f\left(X^i, W_s, W_{\text{reg}}\right)$, the loss from location information of each training sample i is:

$$\mathscr{L}_{\text{loc}}^i = \sum_{\in\{x,y,z,d\}} 1\left(Y^i = 1\right) \cdot \text{dist}\left(T_\gamma^i - \widehat{T}_\gamma^i\right), \tag{9.12}$$

where the function dist(a) $= 0.5a^2$ if $|a|<1$, otherwise $|a|-.5$, which is a robust L_1 loss and is validated to be less sensitive to outliers than the L_2 loss [28]. The $1(Y^i = 1)$ is the indicator function, with which we can only consider the localization loss for those positive samples, and ignore those nonnodule training samples without size notion. Overall, our hybrid-loss objective function is formulated as follows:

$$\mathscr{L} = \mathscr{L}_{\text{cls}} + \lambda \frac{1}{N_{\text{reg}}} \sum_i \mathscr{L}^i_{\text{loc}} + \beta\left(\left|\left|W_{\text{s}}\right|\right|_2^2 + \left|\left|W_{\text{cls}}\right|\right|_2^2 + \left|\left|W_{\text{reg}}\right|\right|_2^2\right). \tag{9.13}$$

The first term represents nodule classification loss. The second term denotes the localization loss, where the N_{reg} represents the number of positive samples considered in the regularization. The third term indicates the weight decay of the shared, classification and regression parameters. The λ and β are balancing weights of the terms.

9.6.3 Dataset, preprocessing and evaluation metrics

We evaluated our method on a large-scale benchmark dataset, which was released during the conference of ISBI 2016 for the LUNA16 Challenge. The dataset filtered out 888 CT scans from the publicly available Lung Image Database Consortium (LIDC) database [24]. The volumetric images were with a resolution in the transverse plane as 512×512, an element spacing of 0.74×0.74 mm^2, and variable slice thickness but not larger than 2.5 mm. The labels (including the location centroids and diameters) of pulmonary nodules were collected with a two-phase manual annotation process conducted by four experienced thoracic radiologists. During the labeling process, each radiologist marked the identified lesions as nonnodule, nodule $<$ 3 mm, and nodules $>=$ 3 mm. The challenge selected a total of 1186 lung nodules $>=$ 3 mm accepted by three or four radiologists as the ground truth. The annotations that were failed to be included in the reference standard (i.e., nonnodules, nodules $<$ 3 mm, and nodules annotated by merely one or two radiologists) were referred to as irrelevant findings.

For preprocessing the CT scans, we clipped the grayscale values into the interval of $(-1000, 400)$ Houndsfield units and normalized them into the range of $(0, 1)$. The mean intensity was subtracted before inputting the samples to the network. We conducted a series of augmentations for the positive nodule samples, including random translation within the radius region of the pulmonary nodule, flipping, random scaling between $[-0.9, +1.1]$, and random rotation of [90, 180, 270] degrees in the transverse plane.

When training the multi-tasking neural network, we set a relatively small training patch size $[30 \times 30 \times 10]$ in the candidate screening stage for fast processing, and the second stage utilized a larger size $[60 \times 60 \times 24]$ to include richer contextual information to accurately detect nodules. The 3-D fully convolutional model was randomly initialized from a Gaussian distribution $\mathscr{N}(0, 0.01)$, and we initialized the learning rate as 0.001. When training the hybrid-loss 3-D residual network, the first three convolutional

layers were initialized from the FCN model and the remaining parameters of deeper layers were randomly initialized as in Ref. [29]. The convolution layers in the residual units utilized padding to maintain dimension of the feature volumes. The trade-off parameters λ and β were set as 0.5 and 1e-4, respectively.

The detection performance was evaluated by measuring the sensitivity and average false positive rate per scan, as defined in the challenge. A predicted candidate location was counted as the true positive if it was positioned within the radius of a true lung nodule center. Detections of irrelevant findings were not considered (i.e., regarded as neither false positives nor true positives) in the evaluation. We conducted the free receiver operation characteristic (FROC) analysis by setting different thresholds on the raw prediction probabilities. The evaluation also computed the 95% confidence interval with the bootstrapping [30]. A competition performance metric (CPM) score [31], which was measured as the average sensitivity at seven predefined false positive rates: 1/8, 1/4, 1/2, 1, 2, 4 and 8 false positives per patient, was calculated.

9.6.4 Experimental results

To investigate the contribution of our proposed learning strategies, extensive ablation experiments are conducted for analysis. We first assess the capability of screening lung nodule candidates using the 3-D FCN trained to convergence with and w/o the online sample filtering (OSF) scheme. The results are presented in the first two columns of Table 9.5. We can observe that training with online sample filtering strategy significantly improves the candidate screening performance by increasing the sensitivity from 94.3% to 97.1% and reducing the FPs/scan rate from 286.2 to 219.1. The improvements present that selecting the high-loss samples (hard samples) with the online sample filtering strategy can greatly enhance the network's discrimination capability and improve the performance.

To evaluate the effectiveness of the residual learning technique and the hybrid-loss objective equipped in our model for false positive reduction, we implemented three different networks—i.e., plain deep network (DeepNet), residual network (ResNet), and our proposed novel hybrid-loss residual network (ResNet + HL)—according to the architecture illustrated in Fig. 9.11. Their results are presented in the last three columns of Table 9.5. With 1.0 FPs/scan, the three networks achieve detection sensitivities of 84.8%, 86.7%, and 90.5%, demonstrating that while the residual learning technique can

Table 9.5 Evaluation of the learning strategies in our detection framework.

Stages	Candidate screening		False positive reduction		
Methods	FCN	FCN + OSF	DeepNet	ResNet	ResNet + HL
Sensitivity	94.3%	97.1%	84.8%	86.7%	90.5%
FPs/scan	286.2	219.1	1.0	1.0	1.0

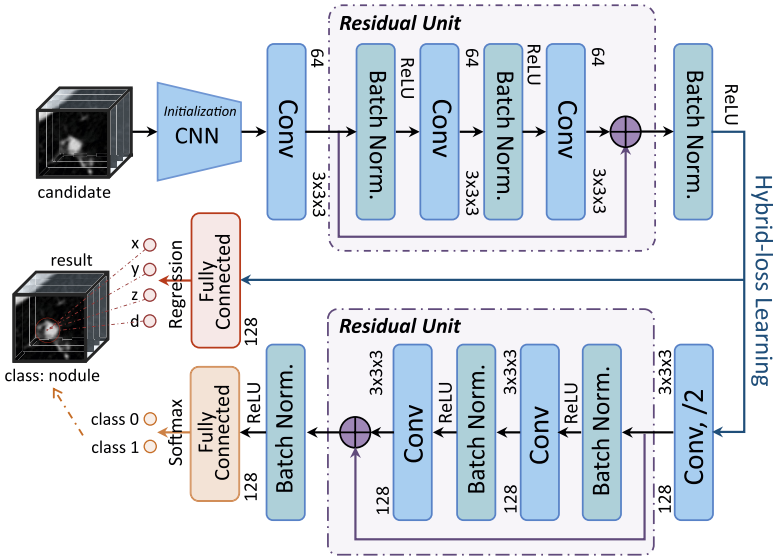

Figure 9.11 False positive reduction with multi-task and hybrid-loss learning.

improve the performance of traditional networks by facilitating gradients flow during optimization, the proposed hybrid-loss objective function can further boost the detection performance by additionally supervising the learning with location and size information. Fig. 9.12 presents the free-response receiver operating characteristic (FROC) curves of three networks, for more comprehensive comparison at a wider range of the false positive rates. It is observed that the proposed ResNet + HL continually obtains the best performance among these three configurations.

For overall lung nodule detection results, Table 9.6 reports the performance of our method and that of other approaches in the LUNA16 challenge. In fact, all participates employed deep learning based approaches, and we refer readers to Ref. [32], a comprehensive summary of LUNA16, to learn more details of other participating methods. It is observed from Table 9.6 that our proposed method achieves a CMP score of 0.839, which set state-of-the-art results. At the false positive rate of 0.5, 1, 2, four and eight per scan, our detection framework achieved the sensitivity of 81.9%, 86.5%, 90.6%, 93.3% and 94.6%, respectively, which are the highest among comparison methods. It is reported that, in real-world clinical practice, the FPs/scan scales between one and four are the mostly concerned [33]. Our proposed method achieves a sensitivity of 90.6% at two FPs/scan, highlighting its promising potential to be readily exploited in real clinical practice.

In Fig. 9.13, we depict typical examples of final detection results with the classification probability and regressed diameter indicated. We can observe that our model can recognize the various lung nodules with a high probability, as well as reliably predict the

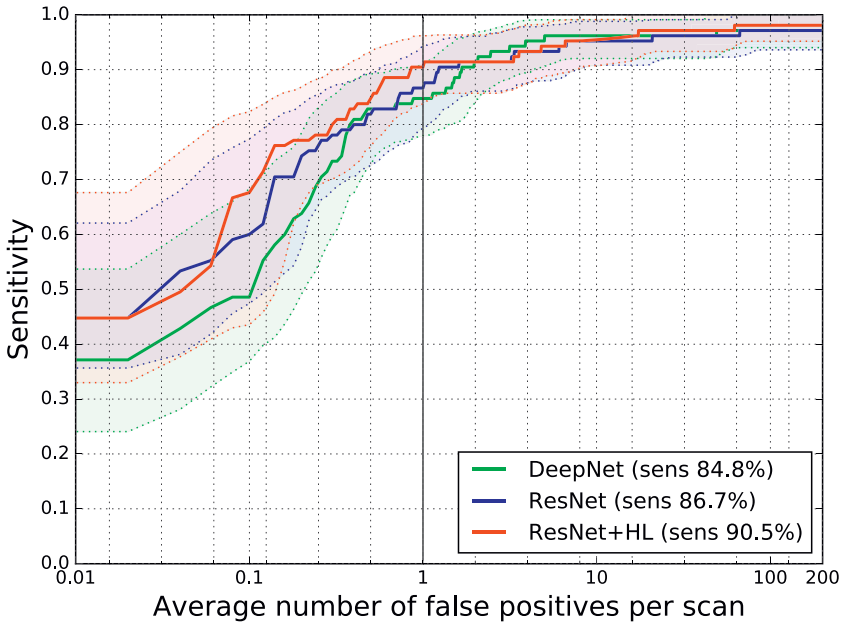

Figure 9.12 Comparison of FROC curves using different network configurations for lung nodule detection, with shaded areas presenting the 95% confidence interval.

Table 9.6 Comparison with other lung nodule detection methods in LUNA16 Challenge.

Teams	0.125	0.25	0.5	1	2	4	8	CPM score
DIAG_ConvNet	0.692	0.771	0.809	0.863	0.895	0.914	0.923	0.838
ZENT	0.661	0.724	0.779	0.831	0.872	0.892	0.915	0.811
Aidence	0.601	0.712	0.783	0.845	0.885	0.908	0.917	0.807
MOT_M5Lv1	0.597	0.670	0.718	0.759	0.788	0.816	0.843	0.742
VisiaCTLung	0.577	0.644	0.697	0.739	0.769	0.788	0.793	0.715
Etrocad	0.250	0.522	0.651	0.752	0.811	0.856	0.887	0.676
Our method	0.659	0.745	0.819	0.865	0.906	0.933	0.946	0.839

size of the detected nodules. Last but not least, our proposed lesion detection framework is quite efficient taking less than 1 minute for one case, which enables our method to be competent for performing large-scale data processing, such as the annual lung cancer screening program which is launched for high-risk populations.

9.7 Discussion

To illustrate the discrimination capability of intermediate features, the representations extracted by the 2-D CNN and 3-D CNN models on the CMB detection task

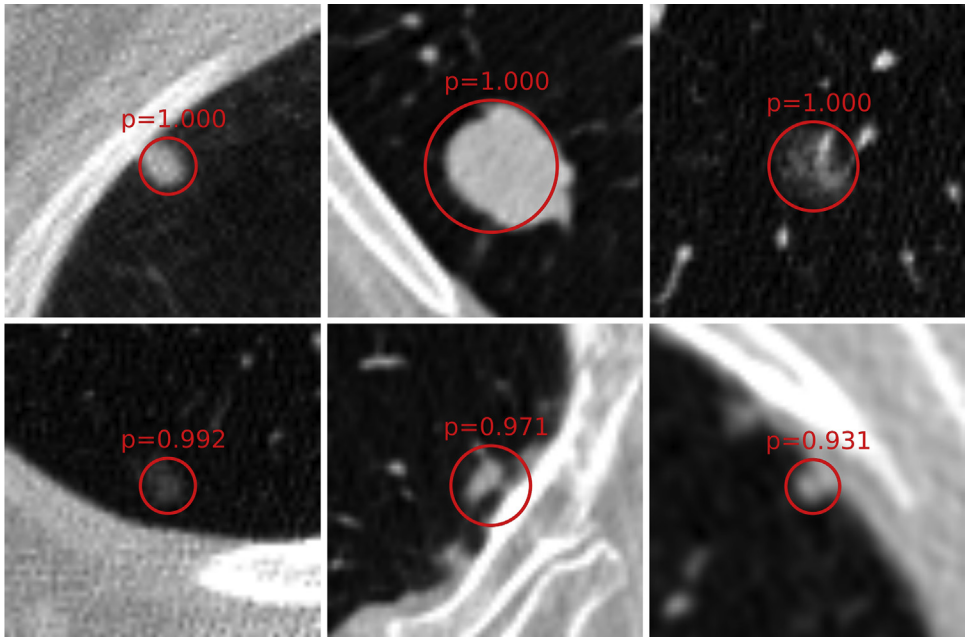

Figure 9.13 Examples of lung nodule detection results of our method with the prediction probability and diameter indicated in red.

are embedded into the 2-D plane using the t-SNE toolbox [34], as shown in Fig. 9.14. The CMB and non-CMB samples are distinctly separated based on the features extracted via our 3-D CNN. In contrast, embedding of the aggregated 2-D CNN representations do not present such a clear partition pattern, highlighting the discrimination capability of the 3-D CNN based features, which can encode richer spatial information within the volumetric medical data.

Meanwhile, we also visualize the 3-D convolution kernels of the first two convolutional layers in the 3-D FCN. Fig. 9.15A illustrates the C1 layer kernels (with size $5 \times 5 \times 3$), where each column represents a 3-D kernel which is demonstrated as three 5×5 maps. With a closer observation, we find that the learned kernels attend to the spherical shapes of the lesion as well as the intensity gradients between the microbleeds and surrounding background. More importantly, the observed slight changes of the three maps within each column prove that the 3-D kernels have effectively captured the spatial information across the third dimension of the volumetric data. Fig. 9.15B illustrates the C2 layer kernels (with size $3 \times 3 \times 3$), where each column represents a 3-D kernel which is visualized as three 3×3 maps. These kernels are relatively difficult for straightforward interpretation, because they try to construct some high-level concepts from the output activations of the bottom layer. Nevertheless, we can still observe that these kernels attain evidently organized patterns.

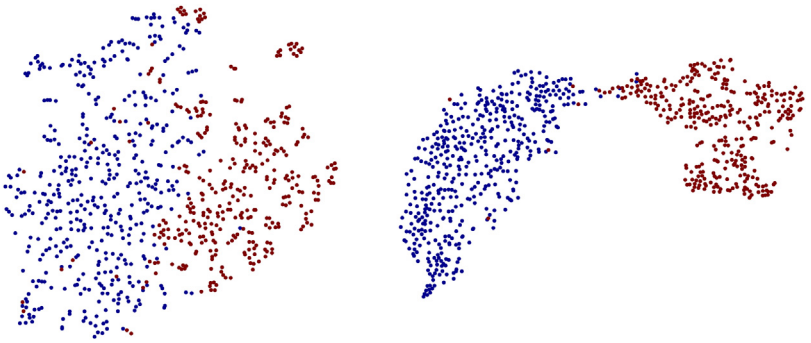

Figure 9.14 Feature embedding from the two-dimensional CNN (left) and three-dimensional CNN methods (right) with t-SNE toolbox. The red and blue colors correspond to the CMBs and non-CMBs, respectively. Best viewed in color.

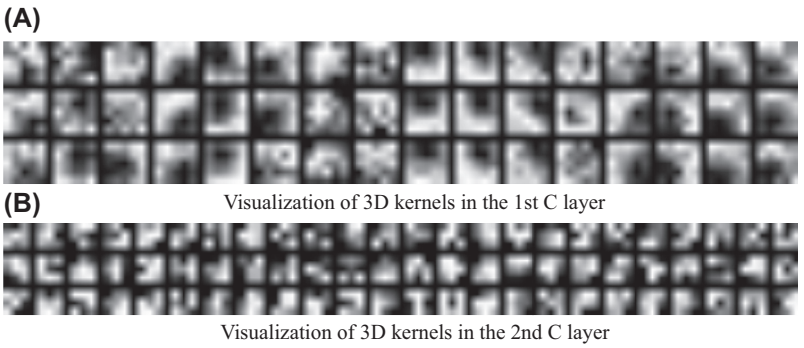

Figure 9.15 Visualization of typical learned filters in the screening three-dimensional CNN model: (A) visualization of the C1 layer kernels, where each column represents a three-dimensional kernel of size $5 \times 5 \times 3$, which is visualized as three 5×5 maps; (B) visualization of the C2 layer kernels, where each column represents a three-dimensional kernel of size $3 \times 3 \times 3$, which is visualized as three 3×3 maps.

With the design of the two-stage cascaded detection framework, we keep two aims in mind: efficiency and accuracy. For an automatic lesion detection system targeting real clinical practice, we believe that both of them are equally crucial. In the cascaded architecture, the first stage focuses on excluding massive background regions and screening potential candidates. In this stage, we develop the 3-D FCN to reduce computational cost, thus meet the requirement of efficiency. The second stage focuses on the small number of candidates and remove the difficult false positives which are with similar appearance to the lesions. In this stage, we employ a discrimination 3-D CNN to identify the true lesions with a high sensitivity and low false positive rate, thus meeting the requirement of accuracy. Quantitatively, with the first stage, we obtain hundreds of false positives per subject. After the second stage, only several false

positives remain. We can see that the second stage removes nearly 99% false positive candidates using the 3-D CNN discrimination model.

9.8 Conclusions

This chapter presents 3-D convolutional neural network based deep learning framework for automatic lesion detection in volumetric medical images. For efficiency, we further elaborate the 3-D FCN which inputs an arbitrary-sized volumetric image and directly outputs a 3-D prediction score volume within a single forward propagation. The two-stage cascaded framework has been extensively validated on two distinct challenging applications—i.e., cerebral microbleed detection in brain MR images and lung nodule detection in chest CT images—with outstanding performance demonstrated. There are appealing potentials to apply our efficient and accurate lesion detection system in real-world clinical practice.

Acknowledgments

We thank our colleagues Dr. Shi Lin, Dr. Vincent CT Mok, Dr. Defeng Wang, Dr. Lei Zhao, Mr. Lequan Yu, Ms. Yueming Jin and Mr. Huangjing Lin, for their early works which are valuable for the contents of this chapter. This work was supported by a grant from the Research Grants Council of HKSAR under General Research Fund (Project no. 14225616) and a grant from Hong Kong Innovation and Technology Commission under ITSP Tier two Platform Funding Scheme (Project no. ITS/426/17FP).

References

[1] S.M. Greenberg, M.W. Vernooij, C. Cordonnier, A. Viswanathan, R.A.-S. Salman, S. Warach, L.J. Launer, M.A. Van Buchem, M. Breteler, Cerebral microbleeds: a guide to detection and interpretation, The Lancet Neurology 8 (2) (2009) 165–174.
[2] A. Charidimou, A. Krishnan, D. JWerring, H. Rolf Jäger, Cerebral microbleeds: a guide to detection and clinical relevance in different disease settings, Neuroradiology 55 (6) (2013) 655–674.
[3] C.I. Henschke, et al., Early Lung Cancer Action Project: overall design and findings from baseline screening, The Lancet 354 (9173) (1999) 99–105.
[4] H. MacMahon, et al., Guidelines for management of small pulmonary nodules detected on CT scans: a statement from the Fleischner Society 1, Radiology 237 (2) (2005) 395–400.
[5] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, nature 521 (7553) (2015) 436.
[6] A. Krizhevsky, et al., Imagenet classification with deep convolutional neural networks, News in Physiological Sciences (2012) 1097–1105.
[7] D. Qi, H. Chen, L. Yu, L. Zhao, J. Qin, D. Wang, V.C.T. Mok, L. Shi, P.-A. Heng, Automatic detection of cerebral microbleeds from MR images via 3D convolutional neural networks, IEEE Transactions on Medical Imaging 35 (5) (2016) 1182–1195.
[8] D. Qi, H. Chen, Y. Jin, H. Lin, J. Qin, P.-A. Heng, Automated pulmonary nodule detection via 3d convnets with online sample filtering and hybrid-loss residual learning, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2017, pp. 630–638.
[9] H. Chen, D. Qi, L. Yu, J. Qin, L. Zhao, V.C.T. Mok, D. Wang, L. Shi, Pheng- Ann Heng, deep cascaded networks for sparsely distributed object detection from medical images, in: Deep Learning for Medical Image Analysis, Elsevier, 2017, pp. 133–154.
[10] S.R.S. Barnes, E.M. Haacke, M. Ayaz, A.S. Boikov, W. Kirsch, D. Kido, Semiautomated detection of cerebral microbleeds in magnetic resonance images, Magnetic Resonance Imaging 29 (6) (2011) 844–852.

[11] B. Ghafaryasl, Fedde van der Lijn, M. Poels, H. Vrooman, M. Arfan Ikram, W.J. Niessen, A. van der Lugt, M. Vernooij, M. de Bruijne, A computer aided detection system for cerebral microbleeds in brain MRI, in: Biomedical Imaging (ISBI), 2012 9th IEEE International Symposium on, IEEE, 2012, pp. 138–141.
[12] D. Qi, H. Chen, L. Yu, L. Shi, D. Wang, V.C.T. Mok, P. Ann Heng, Automatic cerebral microbleeds detection from MR images via independent subspace analysis based hierarchical features, in: 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2015, pp. 7933–7936.
[13] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.
[14] C. Cordonnier, R.A.-S. Salman, J. Wardlaw, Spontaneous brain microbleeds: systematic review, subgroup analyses and standards for study design and reporting, Brain 130 (8) (2007) 1988–2003.
[15] A. Charidimou, D.J. Werring, Cerebral microbleeds and cognition in cerebrovascular disease: an update, Journal of Neurological Sciences 322 (1) (2012) 50–55.
[16] J.D.C. Goos, W.M. van der Flier, D.L. Knol, P.J.W. Pouwels, P. Scheltens, F. Barkhof, M.P. Wattjes, Clinical relevance of improved microbleed detection by susceptibility weighted magnetic resonance imaging, Stroke 42 (7) (2011) 1894–1900.
[17] S.M. Gregoire, U.J. Chaudhary, M.M. Brown, T.A. Yousry, C. Kallis, H.R. Jäger, D.J. Werring, The microbleed anatomical rating scale (MARS) reliability of a tool to map brain microbleeds, Neurology 73 (21) (2009) 1759–1766.
[18] J. de Bresser, M. Brundel, M.M. Conijn, J.J. van Dillen, M.I. Geerlings, M.A. Viergever, P.R. Luijten, G.J. Biessels, Visual cerebral microbleed detection on 7T MR imaging: reliability and effects of image processing, American Journal of Neuroradiology 34 (6) (2013) E61–E64.
[19] H. Chen, L. Yu, D. Qi, L. Shi, V.C.T. Mok, P. Ann Heng, Automatic detection of cerebral microbleeds via deep learning based 3d feature representation, in: 2015 IEEE International Symposium on Biomedical Imaging (ISBI), IEEE, 2015, pp. 764–767.
[20] A. Liaw, M. Wiener, Classification and regression by randomForest, R News 2 (3) (2002) 18–22. URL, http://CRAN.R-project.org/doc/Rnews/.
[21] S. Geman, E. Bienenstock, R. Doursat, Neural networks and the bias/variance dilemma, Neural Computation 4 (1) (1992) 1–58.
[22] M. Tan, et al., A novel computer-aided lung nodule detection system for CT images, Medical Physics 38 (10) (2011) 5630–5645.
[23] T. Messay, et al., A new computationally efficient CAD system for pulmonary nodule detection in CT imagery, Medical Image Analysis 14 (3) (2010) 390–406.
[24] S.G. Armato III, et al., The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans, Medical Physics 38 (2) (2011) 915–931.
[25] D. Qi, H. Chen, L. Yu, J. Qin, Pheng-Ann Heng, Multilevel contextual 3-d cnns for false positive reduction in pulmonary nodule detection, IEEE Transactions on Biomedical Engineering 64 (7) (2017) 1558–1567.
[26] M. Firmino, et al., Computer-aided detection system for lung cancer in computed tomography scans: review and future prospects, BioMedical Engineering Online 13 (2014) 1–16.
[27] A. Shrivastava, A. Gupta, R. Girshick, Training region-based object detectors with online hard example mining, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 761–769.
[28] R. Girshick, Fast R-CNN, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1440–1448.
[29] K. He, X. Zhang, S. Ren, J. Sun, Identity mappings in deep residual networks, in: European Conference on Computer Vision, Springer, 2016, pp. 630–645.
[30] E. Bradley, R.J. Tibshirani, An Introduction to the Bootstrap, CRC press, 1994.
[31] M. Niemeijer, et al., On combining computer-aided detection systems, IEEE Transactions on Medical Imaging 30 (2) (2011) 215–223.

[32] Arnaud Arindra Adiyoso Setio, A. Traverso, B. van Ginneken, C. Jacobs, et al., Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge, Medical Image Analysis (2016) arXiv preprint arXiv:1612.08012.
[33] B. Van Ginneken, S.G. Armato, B. de Hoop, et al., Comparing and combining algorithms for computer-aided detection of pulmonary nodules in computed tomography scans: the ANODE09 study, Medical Image Analysis 14 (6) (2010) 707–722.
[34] L. Van der Maaten, Geoffrey Hinton, Visualizing data using t-SNE, Journal of Machine Learning Research 9 (2579–2605) (2008) 85.