Neural Networks 115 (2019) 1–10

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Multi-scale gradual integration CNN for false positive reduction in pulmonary nodule detection

Bum-Chae Kim[1], Jee Seok Yoon[1], Jun-Sik Choi, Heung-Il Suk *

Department of Brain and Cognitive Engineering, Korea University, Seoul, South Korea

HIGHLIGHTS

- Gradual feature extraction strategy that learns multi-scale feature representations.
- Multi-stream feature representations and abstract-level feature integration.
- Outperforming state-of-the-art methods on LUNA dataset by a large margin.

ARTICLE INFO

Article history:
Received 3 July 2018
Received in revised form 24 December 2018
Accepted 7 March 2019
Available online 18 March 2019

Keywords:
Multi-scale convolutional neural network
Multi-stream feature integration
False positive reduction
Pulmonary nodule detection
Lung cancer screening

ABSTRACT

Lung cancer is a global and dangerous disease, and its early detection is crucial for reducing the risks of mortality. In this regard, it has been of great interest in developing a computer-aided system for pulmonary nodules detection as early as possible on thoracic CT scans. In general, a nodule detection system involves two steps: (i) candidate nodule detection at a high sensitivity, which captures many false positives and (ii) false positive reduction from candidates. However, due to the high variation of nodule morphological characteristics and the possibility of mistaking them for neighboring organs, candidate nodule detection remains a challenge. In this study, we propose a novel Multi-scale Gradual Integration Convolutional Neural Network (MGI-CNN), designed with three main strategies: (1) to use multi-scale inputs with different levels of contextual information, (2) to use abstract information inherent in different input scales with gradual integration, and (3) to learn multi-stream feature integration in an end-to-end manner. To verify the efficacy of the proposed network, we conducted exhaustive experiments on the LUNA16 challenge datasets by comparing the performance of the proposed method with state-of-the-art methods in the literature. On two candidate subsets of the LUNA16 dataset, *i.e.*, V1 and V2, our method achieved an average CPM of 0.908 (V1) and 0.942 (V2), outperforming comparable methods by a large margin. Our MGI-CNN is implemented in Python using TensorFlow and the source code is available from https://github.com/ku-milab/MGICNN.

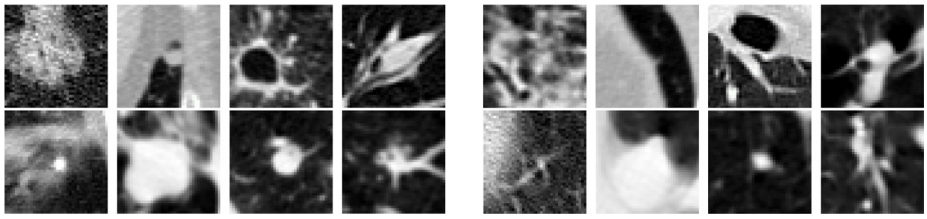
1. Introduction

Lung cancer is reported as the leading cause of death worldwide (Siegel, Miller, & Jemal, 2017). However, when detected at an early stage through thoracic screening with low-dose CT images and treated properly, the survival rate can be increased by 20% (National Lung Screening Trial Research Team et al., 2011). Clinically, pulmonary nodules are characterized as having round shape with a diameter of 3 mm $\sim$ 30 mm in thoracic CT scans (Gould et al., 2007). With this pathological knowledge, there have been efforts of applying machine-learning techniques for early and automatic detection of cancerous lesions, *i.e.*, nodules. To our knowledge, a computerized lung cancer screening system consists of two-steps: candidate nodule detection and False Positives (FPs) reduction. In the candidate nodule detection step, the system uses high sensitivity without concern for specificity to extract as many candidates as possible. Roughly, more than 99% of the candidates are non-nodules, *i.e.*, FPs (Setio et al., 2016), which should be identified and reduced in the second step correctly.

Pathologically, there are many types of nodules (*e.g.*, solids, non-solids, part-solids, calcified, *etc.* Ciompi et al., 2017) and their morphological characteristics such as size, shape, and strength are highly variable. In addition, there are many other structure in the thorax (*e.g.*, blood vessels, airways, lymph nodes) with morphological features similar to nodules (Gould et al., 2007; Roth et al., 2016). Fig. 1 shows an example of a nodule and a non-nodule. In these regards, it is very challenging to reduce FPs or to distinguish nodules from non-nodules, leading many researchers to devote their efforts on the step of false positive reduction (Cao et al., 2017; Dou, Chen, Yu, Qin, & Heng, 2017; Setio et al., 2016).

* Corresponding author.
E-mail address: hisuk@korea.ac.kr (H.-I. Suk).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.neunet.2019.03.003
0893-6080/



(a) Nodule (b) Non-nodule

Fig. 1. Examples of the pulmonary nodules and non-nodules. Both have a complex and similar morphological characteristics that must be distinguished between.

In earlier work, researchers had mostly focused on extracting discriminative morphological features with the help of pathological knowledge about nodule types and applied relatively simple linear classifiers such as logistic regression or support vector machine (Lee, Hara, Fujita, Itoh, & Ishigaki, 2001; Okumura, Miwa, Kako, Yamamoto, Matsumoto, Tateno, et al., 1998; Ye, Lin, Dehmeshki, Slabaugh, & Beddoe, 2009). Recently, with the surge of popularity and success in Deep Neural Networks (DNNs), which can learn hierarchical feature representations and class discrimination in a single framework, a myriad of DNNs has been proposed for medical image analysis (Dou et al., 2016; Esteva et al., 2017; Havaei et al., 2017; Hu et al., 2018; Shen, Wu et al., 2017; Suzuki, Armato, Li, Sone, & Doi, 2003). Of the various deep models, Convolutional Neural Networks (CNNs) have been applied most successfully for pulmonary nodule detection and classification in CT images (Jacobs et al., 2014; Liu, Hou, Qin, & Hao, 2018; Murphy et al., 2009; Roth et al., 2016; Setio et al., 2016; Setio, Jacobs, Gelderblom, & van Ginneken, 2015; Shen, Zhou et al., 2017). Moreover, in order to attain the network performance of computer vision applications, there were trials (Ciompi et al., 2015; Shin et al., 2016) to identify nodules with a deep model fine-tuned with pulmonary nodule data in the way of transfer learning (Razavian, Azizpour, Sullivan, & Carlsson, 2014; Yang & Pan, 2009).

From previous studies of nodule detection or classification in CT scans, we have two notable findings. The first is that it is helpful to exploit volume-level information, rather than 2D slice-level information (Ding, Li, Hu, & Wang, 2017; Roth et al., 2016; Setio et al., 2016). For example, Roth et al. (2016) proposed a 2.5D CNN by taking three orthogonal 2D patches as input for volume-level feature representation. Setio et al. (2016) proposed a multi-view CNN, which extracts hierarchical features from nine 2D slices with different angles of view, and groups the high-level features for classification. However, their method achieved limited performance in low-FP scans. Ding et al. (2017) proposed a 3D CNN with a 3D volumetric patch as input, and presented promising results in FP reduction.

The second is that performance can be enhanced by using multi-scale inputs with different levels of contextual information (Dou et al., 2017; Shen, Zhou, Yang, Yang, & Tian, 2015; Shen et al., 2017). Shen et al. (2015) proposed a multi-scale CNN and successfully applied nodule classification by combining contextual information at different image scales with the abstract-level feature representations. Dou et al. (2017) also designed a 3D CNN to encode multi-level contextual information to tackle the challenges of large variation in pulmonary nodules. The performance of pulmonary nodule classification using the 3D CNN is generally better than that of the 2D CNN (Ding et al., 2017). However, the 3D CNN is more difficult to train than the 2D CNN due to the large number of network parameters. Medical image data is relatively limited, so a 3D CNN may easily become over-fitted. It is also noteworthy that the multi-scale methods have proved their efficacy in computer vision tasks (Honari, Yosinski, Vincent, & Pal, 2016; Karpathy et al., 2014; Lin et al., 2016).

Inspired by the above-mentioned findings, in this study we propose a novel Multi-scale Gradual Integration CNN (MGI-CNN) for FP reduction in pulmonary nodule detection. In designing our network, we apply three main strategies. Strategy 1: We use 3D multi-scale inputs, each containing different levels of contextual information. Strategy 2: We design a network for Gradual Feature Extraction (GFE) from multi-scale inputs at different layers, instead of radical integration at the same layer (Dou et al., 2017; Karpathy et al., 2014; Shen et al., 2015, 2017). Strategy 3: For better use of complementary information, we consider Multi-Stream Feature Integration (MSFI) to integrate abstract-level feature representations. Our main contributions can be summarized as follows:

1. We propose a novel CNN architecture that learns feature representations of multi-scale inputs with a gradual feature extraction strategy.
2. With multi-stream feature representations and abstract-level feature integration, our network reduces many false positives.
3. Our method outperformed state-of-the-art methods in the literature by a large margin on the LUNA16 challenge datasets.

While the proposed network architecture extension is straightforward, to our best knowledge, this is the first work of designing a network architecture that integrates 3D contextual information of multi-scale patches in a gradual and multi-stream manner. Concretely, our work empirically proved the validity of integrating multi-scale contextual information in a gradual manner, which can be comparable to many existing work (Kamnitsas et al., 2017; Lin et al., 2016) that mostly considered radical integration of such information. Besides, our method also presents the effectiveness of learning feature representations from different orders of multi-scale 3D patches and combining the extracted features from different streams to further enhance the performance.

This paper is organized as follows. Section 2 introduces the existing methods in the literature. We then describe our proposed method in Section 3. The experimental settings and performance comparison with the state-of-the-art methods are presented in Section 4. In Section 5, we discuss key issues of the proposed method along with the experimental results. We conclude this paper by summarizing our work and suggesting the future direction for clinical practice in Section 6.

2. Related work

2.1. Computer-aided system for pulmonary nodule

There has been a continuous research effort to develop a Computer-Aided Detection (CADe) or Diagnosis (CADx) system for pulmonary nodules from CT scans, because of its great importance and fatality in providing the physicians a second

objective opinion, which allows to make less medical errors ultimately. Over the past three decades, a plethora of methods have been proposed for pulmonary nodule detection and diagnosis via advanced imaging processing and/or machine learning techniques (Lee et al., 2001; Okumura et al., 1998; Ye et al., 2009). Because of its importance and fatality in the clinic, there were many well-organized review articles, timely handling technical issues at the moment. Specifically, for the traditional techniques for pulmonary nodule detection, refer to Chan, Hadjiiski, Zhou, and Sahiner (2008) and Suzuki (2012).

Recently, with the surge of popularity and success in Deep Neural Networks (DNNs), which can learn hierarchical feature representations and class discrimination in a single framework, a myriad of DNNs has been proposed for medical image analysis (Dou et al., 2016; Esteva et al., 2017; Havaei et al., 2017; Hu et al., 2018; Shen et al., 2017; Suzuki et al., 2003). Of the various deep models, Convolutional Neural Networks (CNNs) have been applied most successfully for pulmonary nodule detection and classification in CT images (Jacobs et al., 2014; Liu et al., 2018; Murphy et al., 2009; Roth et al., 2016; Setio et al., 2016, 2015; Shen et al., 2017). For a review in DNNs for CADe and CADx, refer to Cheng et al. (2016).

2.2. Volumetric contextual information

Automatic lung cancer screening systems classify nodules using specific algorithms to extract nodule morphological characteristics. Okumura et al. (1998) distinguished solid nodules by using a Quoit filter that could detect only isolated nodules. In the case of isolated nodules, the graph of the pixel values becomes 'sharp', and the nodule is detected when the annular filter passes through the graph. However, filters that use only one characteristic of nodules have difficulty in distinguishing diverse nodule types. Li, Sone, and Doi (2003) proposed point, line, and surface shape filters for finding nodule, blood vessel, and airway in a thoracic CT. This is a detection method that considers various types of nodules, effectively reducing the FP response of the automatic lung cancer screening system. However, handcrafted features still do not detect complex types of nodules (*e.g.*, part-solid or calcified nodules). Hence, to detect the more elusive types of nodule, researchers attempted to use volumetric information about the nodule and its surrounding area. Jacobs et al. (2014) extracted volumetric information from various types of bounding boxes that defined the region around a nodule to classify part-solid nodules. That volumetric information includes 107 phenotype features and 21 context features of the nodule and various nodule area with diverse sizes of a bounding box. For the classification, the GentleBoost classifier (Friedman, Hastie, Tibshirani, & Stanford, 1998) learned a total of 128 features and obtained 80% of sensitivity at 1.0 FP/scan. However, the method was inefficient in distinguishing the various types of nodules because it must be reconfigured to filter each nodule type.

Recently, DNNs have been successfully used to substitute the conventional pattern-recognition approaches that first extract features and then train a classifier separately, thanks to their ability of discovering data-driven feature representations and training a classifier in a unified framework. Among various DNNs, CNN-based methods reported promising performance in classifying nodules correctly. Roth et al. (2016) proposed 2.5D CNN that used three anatomical planes (sagittal, coronal, and axial) to extract 3D volumetric information. Their 2.5D CNN also classified organs similar to nodules, such as lymph nodes. This study inspired some researchers in the field of pulmonary nodule detection. Setio et al. (2016) proposed a multi-view CNN that extracted volumetric information with an increased number of input patches. Furthermore, to better consider contextual information, they used groupings of high-level features from each 2D CNN in a 9-view (three times more than 2.5D CNN's anatomical plane) by achieving promising performance, compared with the methods using hand-crafted features. However, this effort could not fully utilize all the 3D volumetric information that could be useful to further enhance the performance. Ding et al. (2017) tried to build a unified framework by applying a deep CNN for both candidate nodule detection and nodule identification. Specifically, they designed a deconvolutional CNN structure for candidate detection on axial slices and a three-dimensional deep CNN for the subsequent FP reduction. In the FP reduction step, they used a dropout method by achieving a sensitivity of 0.913 in average FP/scan on the LUNA16 dataset. Although they claimed to use 3D volumetric information, they did not consider the information between the small patches that were extracted in a large patch.

2.3. Multi-scale contextual information

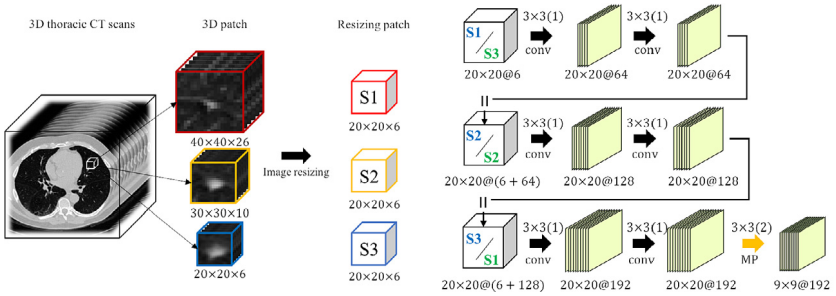
From an information quantity perspective, it may be reasonable to use morphological and structural features in different scales and thus effectively integrating multi-scale contextual information. Shen et al. (2015) proposed Multi-scale CNN (MCNN) as a method for extraction of high-level features from a single network by converting images of various scales to the same size. The high-level features are jointly used to train a classifier, such as support vector machine or random forest, for nodule classification. Dou et al. (2017) used three different architectures of 3D CNN, each one of which was trained with the respective receptive field of an input patch empirically optimized for the LUNA16 challenge dataset. To make a final decision, they integrated label prediction values from patches of three different scales by a weighted sum at the top layers. However, the weights for each scale were determined manually, rather than learning from training samples.

Shen et al. (2017) proposed a Multi-Crop CNN to automatically extract nodule salient information by employing a novel multi-crop pooling strategy. In particular, they cropped different regions from convolutional feature maps and then applied a max-pooling operation different times. To give more attention on the center of the patches, they cropped out the neighboring or surrounding information during multi-crop pooling, which could be more informative to differentiate nodules from non-nodules, *e.g.*, other organs.

In this paper, unlike the methods of Roth et al. (2016) and Setio et al. (2016), we exploit 3D patches to best utilize the volumetric information and thus enhancing the performance in FP reduction. Further, to utilize contextual information from different scales, we exploit a multi-scale approach similar to Dou et al. (2017) and Shen et al. (2017). However, instead of radical integration of multi-scale contextual information at a certain layer (Shen et al., 2015, 2017), we propose to gradually integrate such information in a hierarchical fashion. It is also noteworthy that we still consider the surrounding regions of a candidate nodule to differentiate from other organs, which can be comparable to Shen et al. (2017).

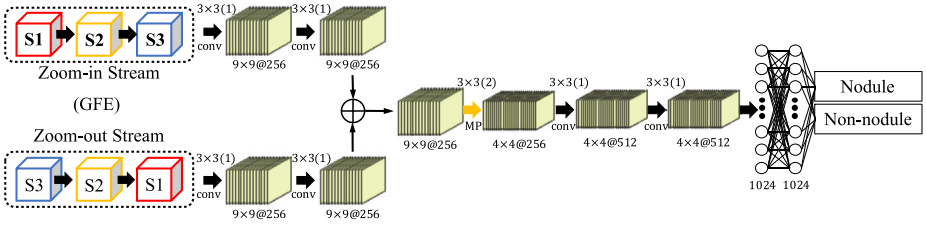
3. Multi-scale gradual integration convolutional neural network (MGI-CNN)

In this section, we describe our novel method of Multi-scale Gradual Integration Convolutional Neural Network (MGI-CNN) in Fig. 2 for pulmonary nodule detection, which consists of two main components: Gradual Feature Extraction (GFE) and Multi-Stream Feature Integration (MSFI). For each candidate nodule, we extract 3D patches at three different scales $40 \times 40 \times 26$, $30 \times 30 \times 10$, and $20 \times 20 \times 6$ by following Dou et al.'s work (Dou et al., 2017). We then resize three patches to 20×20



(a) 3D Patch Extraction: Given the coordinates of a candidate nodule, we extract three patches in different scales and then resize them to the same size, *i.e.*, $S1$, $S2$, and $S3$.

(b) Gradual Feature Extraction: The multi-scale patches with different levels of contextual information ($S1$-$S2$-$S3$ or $S3$-$S2$-$S1$) are integrated in a gradual manner.



(c) The architecture of the proposed multi-scale gradual integration CNN.

Fig. 2. Overview of the propose framework for FP reduction in pulmonary nodule detection. The notations of $\|$ and $\oplus$ denote, respectively, concatenation and element-wise summation of feature maps. The numbers above the thick black or yellow arrows present a kernel size, *e.g.*, 3×3 and a stride, *e.g.*, (1) and (2). (conv: convolution, MP: max-pooling). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

$\times$ 6, denoted as $S1$, $S2$, and $S3$, respectively, as input to the proposed network (Fig. 2a). Note that patches of $S1$, $S2$, and $S3$ have the same center coordinates but pixels in the patches are different in resolution.

3.1. Gradual feature extraction

Inspired by the human visual system, which retrieves meaningful contextual information from a scene by changing the field of view, *i.e.*, by considering contextual information at multiple scales (Zhang, Atkinson, & Goodchild, 2014), we first propose a scale-ordered GFE network presented in Fig. 2b. In integrating morphological or structural features from patches at different scales, the existing methods (Shen et al., 2015, 2017) combined features from multiple patches all at once. Unlike their methods, in this paper, we extract features by gradually integrating contextual information from different scales in a hierarchical manner. For gradual feature representation from multi-scale patches, *i.e.*, $S1$, $S2$, and $S3$, there are two possible scenarios, *i.e.*, $S1 - S2 - S3$ ('*zoom-in*') or $S3 - S2 - S1$ ('*zoom-out*').

For the zoom-in scenario of $S1-S2-S3$, a patch at one scale $S1$ is first filtered by the corresponding local convolutional kernels and the resulting feature maps $F1$ are concatenated ($\|$) with the patch at the next scale $S2$, *i.e.*, $F1 \ \| \ S2$. In our convolution layer, $F1$ is the result of two repeated computations of a spatial convolution and a non-linear transformation by a Rectifier Linear Unit (ReLU) (Nair & Hinton, 2010). Our convolution kernel uses zero padding to keep the size of the output feature maps equal to the size of an input patch and thus valid to concatenate the resulting feature maps and another input patch $S2$ in different scale. The $F1 \ \| \ S2$ tensor is then convolved with kernels of the following convolution layers, producing feature maps $F12$, which now represent the integrated contextual information from $S1$ and $S2$. The feature maps $F12$ are then concatenated with the patch at the next scale $S3$ and the tensor of $F12 \ \| \ S3$ is processed by the related kernels, resulting in feature maps $F123$. The feature maps $F123$ represent the final integration of various types of contextual information in patches $S1$, $S2$, and $S3$. The number of feature maps in our network increases as additional inputs are connected so that the additional information can be extracted from the information of the preceding inputs and the contextual information of the sequential inputs. For the zoom-out scenario of $S3-S2-S1$, the same operations are performed but with input patches in the opposite order.

In the zoom-in scenario, the network is provided with patches at an increasing scale. So, the field of view in a zoom-in network is gradually reduced, meaning that the network gradually focuses on a nodule region. Meanwhile, the zoom-out network has a gradually enlarging field of view, and thus the network finds morphological features combined with the neighboring contextual information by gradually focusing on the surrounding region. In our network architecture, the feature maps extracted from the previous scale are concatenated to the patch of the next scale with zero padding, and then fed into the following convolution layer. By means of our GFE method, our network sequentially integrates contextual features according to the order of the scales. It is noteworthy that the abstract feature representations from two different scenarios, *i.e.*, zoom-in and zoom-out, carry different forms of information.

3.2. Multi-stream feature integration (MSFI)

Rather than considering a single stream of information flow, either $S1 - S2 - S3$ or $S3 - S2 - S1$, it will be useful to consider

Table 1
Statistics of the two datasets, *i.e.*, V1 and V2, for FP reduction in the LUNA16 challenge. The numbers in parentheses denote the number of nodule-labeled candidates in each dataset that match with the radiologists' decisions. The numbers outside parentheses denote the number of all nodules and non-nodules.

Dataset	Candidates	
	Nodule	Non-nodule
V1	1351 (1120)	549,714
V2	1557 (1166)	753,418

multiple streams jointly and to learn features accordingly. With two possible scenarios of zoom-in and zoom-out, we define the information flow of $S1 - S2 - S3$ as '*zoom-in stream*' and the information flow of $S3 - S2 - S1$ as '*zoom-out stream*'.

As the zoom-in and zoom-out streams focus on different scales of morphological and contextual information around the candidate nodule in a different order, the learned feature representations from different streams can be complementary to each other for FP reduction. Hence, it is desirable to combine such complementary features in a single network and to optimize the feature integration from the two streams in an end-to-end manner. To this end, we design our network to integrate contextual information from the two streams as presented in Fig. 2c and call it as MSFI. The proposed MSFI is then followed by additional convolutional layers and fully-connected layers to fully define our MGI-CNN, as shown in Fig. 2c.

To summarize, our proposed network consists of 'zoom-in' and 'zoom-out' GFE networks, followed by an MSFI network. In each of the 'zoom-in' and 'zoom-out' GFE networks, two convolution layers were inserted between successive two-scale patch inputs, and the number of feature maps increased by a multiple of 64, applying the same kernel size of 3×3. Those GFE networks then go through a max pooling layer with 3×3 kernels and a stride of 2, and two additional convolutional layers with 256 filters applying kernels of 3×3 in size. Finally, the MSFI network integrates the outputs of each GFE network by element-wise summation followed by a max pooling layer with 3×3 kernels with a stride of 2, two convolutional layers with 512 features applying kernels of 3×3 in size, and two fully connected layers with 1024 units for each. The output of the proposed network is a single sigmoidal neuron for binary classification, *i.e.*, nodule vs. non-nodule.

4. Experimental settings and results

4.1. Experimental settings

We performed the experiments on the LUng Nodule Analysis 2016 (LUNA16) challenge (Setio et al., 2017) datasets[2] by excluding patients whose slice thickness exceeded 2.5 mm. LUNA16 includes samples from 888 patients in the LIDC–IDRI open database (Armato et al., 2011), which contains annotations of the Ground Truth (GT) collected from the two-step annotation process by four experienced radiologists. After each radiologist annotated all the candidates on the CT scans, each candidate nodule with the agreement of at least three radiologists was approved as GT. There are 1186 GT nodules in total. For the FP reduction challenge, LUNA16 provides the center coordinates of candidate nodules, the respective patient's ID, and the label information, obtained by commercially available systems. Specifically, there are two versions (V1 and V2) of datasets: The V1 dataset provides 551,065 candidate nodules obtained with Jacobs et al. (2014), Murphy et al. (2009) and Tan, Deklerck, Jansen, Bister, and Cornelis (2011), of which 1351 and 549,714 candidates are, respectively, labeled as nodules and non-nodules. In comparison with the four radiologists' decisions, 1120 nodules out of the 1351 are matched with GTs; The V2 dataset includes 754,975 candidate nodules detected with five different nodule detection systems (Jacobs et al., 2014; Murphy et al., 2009; Setio et al., 2015; Tan et al., 2011; Traverso, Torres, Fantacci, & Cerello, 2017). Among the 1557 nodule-labeled candidates, 1166 are matched with the GTs, *i.e.*, four radiologists' decisions. Table 1 summarizes the statistics of the candidate nodules of two datasets for FP reduction in LUNA16.

By using the 3D center coordinates of the candidates provided in the dataset, we extracted a set of 3D patches from thoracic CT scans at scales of $40 \times 40 \times 26$, $30 \times 30 \times 10$, and $20 \times 20 \times 6$, which covered, respectively, 99%, 85%, and 58% of the nodules in the dataset, by following Dou et al.'s work (Dou et al., 2017). The extracted 3D patches were then resized to $20 \times 20 \times 6$ by nearest-neighbor interpolation. For faster convergence, we applied a min–max normalization to patches in the range of $[-1000, 400]$ Hounsfield units (HU)[3] (Hounsfield, 1980).

Regarding the network training, we initialized network parameters with Xavier's method (Glorot & Bengio, 2010). We also used a learning rate of 0.003 by decreasing with a weight decay of 2.5% in every epoch and the number of epochs of 40. For non-linear transformation in convolution and fully-connected layers, we used a ReLU function. To make our network robust, we also applied a dropout technique to fully connected layers with a rate of 0.5. For optimization, we used a stochastic gradient descent with a mini-batch size of 128 and a momentum rate of 0.9.

For performance evaluation, we used a Competitive Performance Metric (CPM) (Niemeijer et al., 2011) score, a criterion used in the FP reduction track of the LUNA16 challenge for ranking competitors. Concretely, a CPM is calculated with 95% confidence by using bootstrapping (Efron & Tibshirani, 1994) and averaging sensitivity at seven predefined FP/scan indices, *i.e.*, 0.125, 0.25, 0.5, 1, 2, 4, and 8. For fair comparison with other methods, the performance of our methods reported in this paper were obtained by submitting the probabilities of being nodule for candidate nodules to the website of the LUNA16 challenge. To better justify the validity of the proposed method, we also counted the number nodules and non-nodules correctly classified and thus to present the effect of reducing FPs.

We evaluated the proposed method with 5-fold cross-validation by following the LUNA16's instruction for splitting each fold into train and test set. In regards to the use of a separate validation set for model selection, due to high computational time in training, we have not applied a nested cross-validation for V2 dataset. Instead, we conducted a preliminary experiment with V1 dataset, and found out that a reasonably high performance was obtained around 40 epochs. Based on this observation, we have predefined the number of epochs to 40 for V2 dataset. To avoid a potential bias problem due to the high imbalance in the number of samples between nodules and non-nodules, we augmented the nodule samples by 90°, 180°, and 270° rotation on a transverse plane and 1-pixel shifting along the x, y, and z axes. Consequently, the ratio between the number of nodules to non-nodules was approximately 1 : 6. The detailed numbers of training, validation, and test samples are presented in Table 2.

4.2. Performance comparison

To verify the validity of the proposed method, *i.e.*, MGI-CNN, we compared with the existing methods (Ding et al., 2017; Dou et al., 2017; Sakamoto et al., 2017; Setio et al., 2016) in the literature that achieved state-of-the-art performance on V1 and/or

[2] Available at 'https://luna16.grand-challenge.org'.

[3] A quantitative scale for describing radio density.

Table 2
Statistics of the training, validation, and test samples used in each run with a 5-fold cross-validation scheme. Numbers in parentheses denote the number of validation and test samples, respectively. Due to high computational time in training, we have not applied a nested cross-validation for V2 dataset. Instead, we conducted a preliminary experiment with V1 dataset and found out that a reasonably high performance was obtained around 40 epochs. Hence, for V2 dataset, we set the number of epochs 40 with no validation step involved.

Dataset		Run #1	Run #2	Run #3	Run #4	Run #5
V1	Scans	533 (177/178)	532 (178/178)	533 (177/178)	536 (175/177)	530 (181/177)
	Nodule	665 (334/352)	677 (379/295)	745 (309/297)	762 (249/340)	709 (369/273)
	Augmentation	53,865	54,837	60,345	61,722	57,429
	Non-nodule	330,549 (64,787/154,378)	330,639 (73,481/145,594)	330,771 (66,576/152,367)	330,453 (69,618/149,643)	330,683 (67,595/151,436)
V2	Scans	710 (−/178)	710 (−/178)	710 (−/178)	711 (−/177)	711 (−/177)
	Nodule	1205 (−/352)	1262 (−/295)	1260 (−/297)	1217 (−/340)	1284 (−/273)
	Augmentation	97,605	102,222	102,060	98,577	104,004
	Non-nodule	599,040 (−/154,378)	607,824 (−/145,594)	601,051 (−/152,367)	603,775 (−/149,643)	601,982 (−/151,436)

Table 3
The CPM scores of the competing methods for the FP reduction task on the dataset V2 and V1 in LUNA16 challenge.

		0.125	0.25	0.5	1	2	4	8	Average
V2	Proposed MGI-CNN	**0.904**	**0.931**	**0.943**	0.947	0.952	0.956	0.962	**0.942**
	Ding et al. (2017)	0.797	0.857	0.895	0.938	0.954	0.970	0.981	0.913
	Dou et al. (2017)	0.677	0.834	0.927	**0.972**	**0.981**	**0.983**	**0.983**	0.908
	Setio et al. (2016)	0.669	0.760	0.831	0.892	0.923	0.944	0.960	0.854
V1	Proposed MGI-CNN	**0.880**	**0.894**	**0.907**	**0.912**	**0.914**	**0.919**	**0.927**	**0.908**
	Dou et al. (2017)	0.678	0.738	0.816	0.848	0.879	0.907	0.922	0.827
	Sakamoto, Nakano, Zhao, and Sekiyama (2017)	0.760	0.794	0.833	0.860	0.876	0.893	0.906	0.846
	Setio et al. (2016)	0.692	0.771	0.809	0.863	0.895	0.914	0.923	0.838

V2 datasets of the LUNA16 challenge. Concisely, Setio et al.'s method (Setio et al., 2016) uses 9-view 2D patches, Ding et al.'s method (Ding et al., 2017) takes 3D patches as input, and Dou et al.'s method (Dou et al., 2017) uses multi-level 3D patches. Sakamoto et al.'s 2D CNN (Sakamoto et al., 2017) eliminates the predicted nonconformity in the training data by raising the threshold in every training iteration. Table 3 summarizes the CPM scores over seven different FP/scan values on the V2 and V1 datasets, respectively.

First, on the large-sized V2 dataset, the proposed MGI-CNN was superior to all other competing methods by a large margin in the average CPM. Notably, when comparing with Dou et al.'s method (Dou et al., 2017), which also uses a 3D CNN with the same multi-scale patches as ours, our method increased the average CPM by 0.034 (∼3% improvement). It is also noteworthy that while the sensitivity of our method at 1, 2, 4, and 8 FP/scan was lower than Dou et al. (2017) and Ding et al. (2017), our method still achieved the best performance at the 0.125, 0.25, and 0.5 FP/scan. That is, for a low FP rate, which is the main goal of the challenge, our method outperformed those methods.

Over the V1 dataset, our method obtained the highest CPMs under all conditions of the FP/scan as presented in Table 3. Again, when compared with Dou et al.'s and Setio et al.'s work (Setio et al., 2016), our method made promising achievements by increasing the average CPM by 0.081 (∼10% improvement) and by 0.070 (∼8.35% improvement). In comparison with Sakamoto et al.'s method (Sakamoto et al., 2017) that reported the highest CPM among the competing methods, our MGI-CNN increased by 0.062 (∼7.3% improvement).

4.3. Effects of the proposed strategies

To show the effects of our strategies in constructing a multi-scale CNN, *i.e.*, GFE in Fig. 2b and MSFI in Fig. 2c, we also conducted experiments with the following Multi-scale CNNs (MCNNs):

- MCNN with Radical integration of Input patches (MCNN-RI): taking multi-scale 3D patches concatenated at the input-level, *i.e.*, $S1\|S2\|S3$, as presented in Fig. 3.
- MCNN with radical integration of Low-level feature Representations (MCNN-LR): integrating multi-scale information with feature maps of the first convolution layer as presented in Fig. 4.
- MCNN with zoom-in gradual feature integration (MCNN-ZI): integrating multi-scale patches gradually in the order of $S1-S2-S3$, *i.e.*, the upper network pathway of the proposed network in Fig. 2c.
- MCNN with zoom-out gradual feature integration (MCNN-ZO): integrating multi-scale patches gradually in the order of $S3-S2-S1$, *i.e.*, the lower network pathway of the proposed network in Fig. 2c.

To make these networks have similar capacity, we designed network architectures to have a similar number of tunable parameters: MCNN-RI (9,463,320), MCNN-LR (9,466,880), MCNN-ZI (9,464,320), MCNN-ZO (9,464,320), MGI-CNN (9,472,000), where the number of tunable parameters are in parentheses. We conducted this experiment on the V2 dataset only, because the V1 dataset is a subset of the V2 dataset and reported the results in Table 4.

First, regarding the strategy of gradual feature extraction, the methods of MCNN-ZI and MCNN-ZO obtained 0.937 and 0.939 of the average CPM, respectively. While the methods with radical integration of contextual information either in the input layer (MCNN-RI) or in the first convolution layer (MCNN-LR) achieved 0.939 and 0.929 of the average CPM. Thus, MCNN-ZI and MCNN-ZO showed slightly higher average CPM scores than MCNN-RI and MCNN-LR. However, in terms of FPs reduction, the power of the gradual feature extraction became notable. That is, while MCNN-RI and MCNN-LR misidentified 383 and 309 non-nodules as nodules, MCNN-ZI and MCNN-ZO failed to remove 279 and 267 non-nodule candidates.

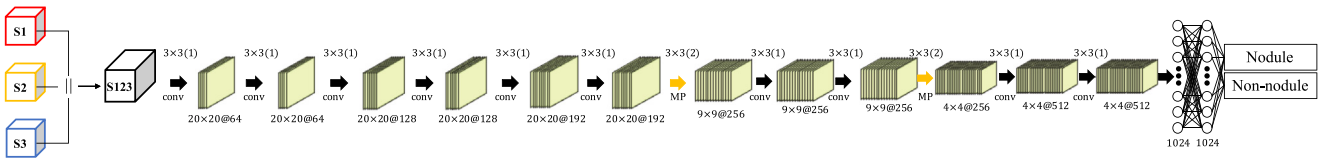


Fig. 3. Architecture of a multi-scale convolutional neural network with radical integration. '||' denotes concatenation of feature maps. The numbers above the thick black or yellow arrows present a kernel size, *e.g.*, 3 × 3 and a stride, *e.g.*, (1) and (2). (conv: convolution, MP: max-pooling). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

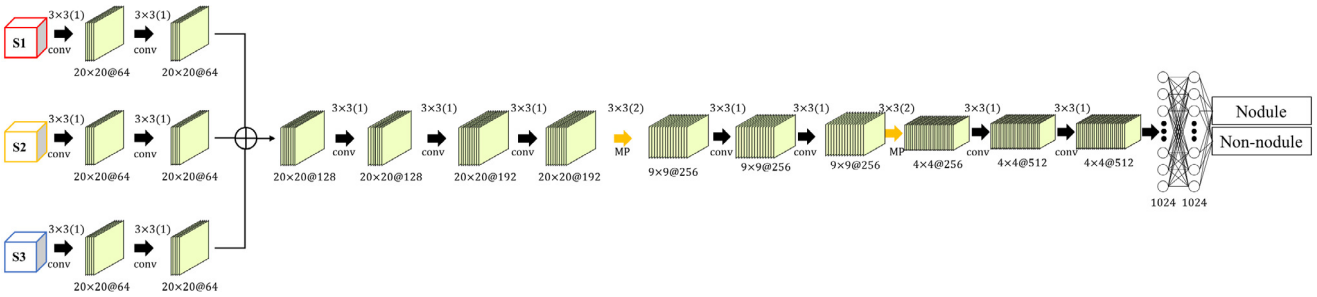


Fig. 4. Architecture of a multi-scale convolutional neural network with radical integration of low-level feature representations. '⊕' denotes element-wise summation of feature maps. The numbers above the thick black or yellow arrows present a kernel size, *e.g.*, 3 × 3 and a stride, *e.g.*, (1) and (2). (conv: convolution, MP: max-pooling). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 4
The CPM scores and the number of True Positives (TPs) and False Positives (FPs) of Multi-scale CNNs (MCNN) with different ways of integrating contextual information from input patches. (MCNN-RI: Radical integration of Input patches; MCNN-LR: MCNN with radical integration of Low-level feature representations; MCNN-ZI: MCNN with zoom-in gradual feature integration; MCNN-ZO: MCNN with zoom-out gradual feature integration, for details refer to the main contexts) The *p*-value denotes a statistical significance test (vs. Proposed MGI-CNN).

	CPM								TP in GT	FP	*p*-value
	0.125	0.25	0.5	1	2	4	8	Average			
MCNN-RI	0.887	0.921	0.939	0.943	0.947	**0.958**	0.962	0.936	1159	383	0.032
MCNN-LR	0.879	0.907	0.926	0.935	0.945	0.954	0.962	0.929	1156	309	0.008
MCNN-ZI	0.893	0.920	0.937	0.945	0.951	0.956	0.960	0.937	1160	279	0.018
MCNN-ZO	0.899	0.920	0.939	0.945	0.951	0.957	**0.965**	0.939	**1161**	267	0.078
Proposed MGI-CNN	**0.904**	**0.931**	**0.943**	**0.947**	0.952	0.956	0.962	**0.942**	**1161**	**232**	–

Second, as for the effect of multi-stream feature integration, the proposed MGI-CNN overwhelmed all the competing methods by achieving the average CPM of 0.942. Further, in FP reduction, MGI-CNN reported only 232 mistakes in filtering out non-nodule candidates. In comparison with MCNN-ZI and MCNN-ZO, the proposed MGI-CNN made 47 and 35 less mistakes, respectively, and thus achieving the best performance in FPs reduction.

It is also worth mentioning that the networks of MCNN-RI, MCNN-LR, MCNN-ZI, MCNN-ZO achieved better performance than the competing methods of Ding et al. (2017), Dou et al. (2017) and Setio et al. (2016) in average CPM. From this comparison, it is believed that the network architectures with the number of tunable parameters of approximately 9.4M had better power of learning feature representations than those of Ding et al. (2017), Dou et al. (2017) and Setio et al. (2016) for FP reduction in pulmonary nodule detection.

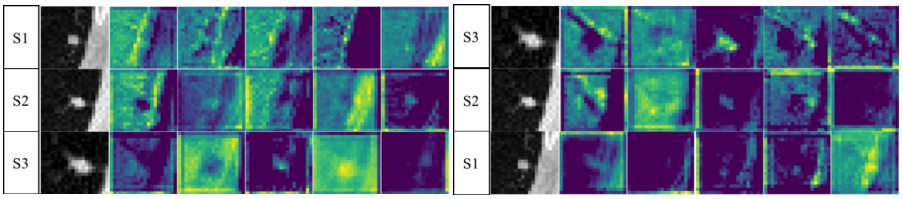
Furthermore, the complementary features from the two different streams of GFE should be integrated properly without lowering the performance of FP reduction. To fully utilize the morphological and contextual information while reducing the chance of information loss, we integrate such information with the abstract-level feature representations through MSFI. With an effective integration method, it is possible to compensate for the loss of information that may occur through the feed-forward propagation of the network, especially the max-pooling layer. To combine the feature maps of two streams, we consider three different methods: concatenation, element-wise summation, and 1 × 1 convolution (Table 5). In our experiments, there was no significant difference in the statistical significance in the CPM score between element-wise summation and 1 × 1 convolution (p-value = 0.146), but the element-wise summation method achieved the lowest number of FPs, which is the ultimate goal of our work.

Table 5
Performance changes of average (Avg.) CPM according to different stream-integration methods. 'TP in GT' denotes the number of true positives that are also included in GT. FP and FN stand for false positive and false negative, respectively. The *p*-value denotes a statistical significance test (vs. Element-wise sum).

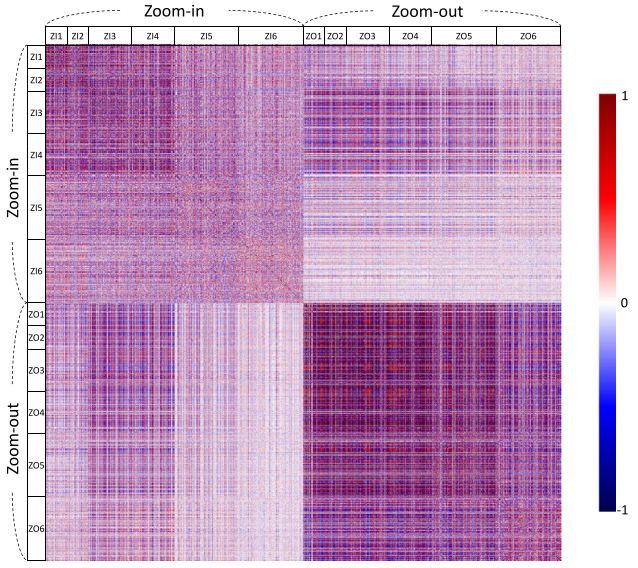
	Avg. CPM	TP in GT	FP	FN	*p*-value
Concatenation	0.939	1160	263	105	0.046
1 × 1 conv	**0.942**	1160	253	**93**	0.146
Element-wise sum	**0.942**	1161	**232**	98	–

5. Discussions

The major advantages of the proposed method can be summarized by two points. First, as shown in Fig. 5, our MGI-CNN could successfully discover morphological and contextual features at different input scales. In Fig. 5a, we observe that the feature maps in the zoom-in network (*i.e.*, each column in the figure) gradually integrate contextual information in the nodule region.



(a) Samples of feature maps from the zoom-in stream (b) Samples of feature maps from the zoom-out stream



(c) Representational similarity analysis between all layers of zoom-in stream (ZI#) and zoom-out stream (ZO#)

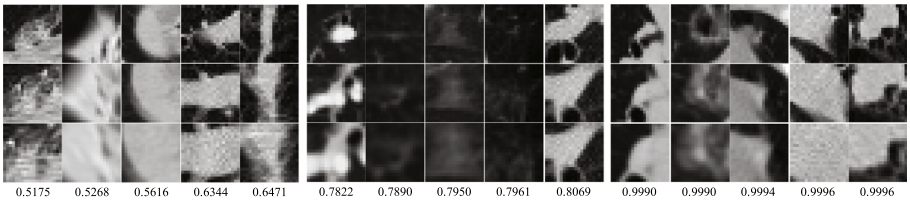
Fig. 5. Examples of the feature maps extracted before concatenation with other scale inputs in the zoom-in/zoom-out stream of the proposed MGI-CNN. The feature maps show gradually extracted contextual information in nodule regions. (a) The zoom-in stream feature maps in the first row show the features of the small-scale patch, and the last row shows the features of the largest scale patch. (b) The zoom-out stream feature maps, on the contrary, show the features of the largest scale patch in the first row and the feature of the smallest scale patch in the last row. (c) Pearson's correlation r is shown for each feature of each layer. There is a high correlation between adjacent layers and very low correlation between each streams, suggesting that each layer contributes to the performance of the model.

Each sample feature map was extracted from the middle of the sagittal plane in the 3D feature map before concatenation with the next scale input. A similar but reversed pattern in integrating the contextual information can be observed in the zoom-out network (each column in Fig. 5b). The illustrated feature maps in each row (Figs. 5a and b) were chosen randomly from the first convolution layer attached to the input patches in different scales from 'zoom-in' and 'zoom-out' GFE networks, respectively. In regards to the concern of feature redundancy between two streams, we have conducted representational similarity analysis by calculating Pearson's correlation coefficients between all pairs of feature maps from the same or different layers in 'zoom-in' and 'zoom-out' stream. The results is presented in Fig. 5c. In the figure, we observe relatively high intra-stream relations (ZI–ZI and ZO–ZO) but low inter-stream (ZI–ZO and ZO–ZI) relations. These different ways of integrating contextual information and extracting features from multi-scale patches could provide complementary information, and thus could enhance performance in the end. Second, our proposed abstract feature integration is useful in terms of information utilization. It is possible to maximize the FP reduction by integrating features at the abstract-level.

With regard to complementary features integration at the abstract-level, we considered three different strategies, *i.e.*, concatenation, element-wise summation, 1×1 convolution (Lin, Chen, & Yan, 2013), commonly used in the literature. The resulting performances are presented in Table 5. Although there is no significant difference among the four methods in average CPM, from a FP reduction perspective, the element-wise summation reported 232 number of FPs, reducing by 31 (vs. concatenation), 103 (vs. skip-connection), and 31 (vs. 1×1 convolution). In this regard, we used element-wise summation in our MGI-CNN.

The 3D patches fed into our network were resized to fit the input receptive field size, *i.e.*, $20 \times 20 \times 6$. Such image resizing may cause information loss or corruption in the original-sized patches. However, as we can see in Fig. 5, the 3D patches of size $20 \times 20 \times 6$, in which the nodule still occupies most of the patch, was not affected by the resizing operation. This means that even if the surrounding region information is lost by resizing, the information of the nodule itself could be preserved.

We visually inspected the misclassified candidate nodules. In particular, we first clustered the 232 FPs by our MGI-CNN into three groups based on the their probabilities as nodule: Low Confidence (LC; $0.5 \leq p < 0.7$), Moderate Confidence (MC; $0.7 \leq p < 0.9$), and High Confidence (HC; $p > 0.9$). The number of FP patches for each group was 33 (LC), 47 (MC), and 152 (HC), respectively. Fig. 6 presents the representative FP 3D patches for three groups. One noticeable thing from the LC and HC groups is that the extracted 3D patches mostly seem to be a subpart of a large tissue or organ, and thus our network failed to find configural patterns necessary to differentiate from non-nodules. For the MC group, patches show relatively low contrasts, which is possibly due to our normalization during preprocessing



(a) Low confidence $(0.5 \leq p < 0.7)$ (b) Moderate confidence $(0.7 \leq p < 0.9)$ (c) High confidence $(0.9 \leq p < 1)$

Fig. 6. Examples of the candidate nodules misclassified to nodule by our MGI-CNN. Based on the output probabilities as nodule, samples are clustered into three groups. The first, second, and third rows correspond to $S1$, $S2$, and $S3$ scale, respectively. The number at the bottom of each column is the output probability as nodule.

(Section 4.1). These observations motivate us to extend our network to accommodate an increased number of patches with larger scales and patches normalized in different ways. This would be an interesting direction to further improve the performance.

Regarding a clinical use, there are some limitations of our method as well as the existing ones in the literature. First, there exists a high variation across images in most large medical image datasets. This is due to arbitrations in the dataflow pattern in the imaging devices across centers and even across sessions. These arbitrations will cause an interference with proper readout of the CT scans and will affect the quality of pulmonary nodule detection. LUNA16 organizers went through robust registration and normalization for all CT scans to make these interference as little as possible. Second, since our method is trained in a supervised manner, the confidence of a radiologist's ground truth of nodules is of great importance. There are about 850 nodule candidates per scan, and only 0.002% of them are nodules. The sheer number of candidates and the imbalance between number of nodules and non-nodules make it very hard for radiologists to make a perfect detection. Due to these reasons, LUNA16 organizers hired 4 experienced radiologists to make a more confident detection for all scans. For its general use and better clinical supportive system, it is of importance for an interactive use with a clinician. To be specific, a system needs incrementally update its model parameters to take account for any false positives or false negatives identified by clinicians during practice.

From a system's perspective, instead of developing a full pulmonary nodule detection system, which usually consists of a candidate detection part and a FP reduction part, this study mainly focused on improving the FP reduction component. As the proposed approach is independent of candidate screening methods, our network can be combined with any candidate detector. If the proposed network is combined with more high-performance candidate detection methods, we presume to obtain better results.

6. Conclusion

In this paper, we proposed a novel multi-scale gradual integration CNN for FP reduction in pulmonary nodule detection on thoracic CT scans. In our network architecture, we exploited three major strategies: (1) use of multi-scale inputs with different levels of contextual information, (2) gradual integration of the information inherent in different input scales, and (3) multi-stream feature integration by learning in an end-to-end manner. With the first two strategies, we successfully extracted morphological features by gradually integrating contextual information in multi-scale patches. Owing to the third strategy, we could further reduce the number of FPs. In our experiments on the LUNA16 challenge datasets, our network achieved the highest performance with an average CPM of 0.908 on the V1 dataset and an average CPM of 0.942 on the V2 dataset, outperforming state-of-the-art methods by a large margin. In particular, our method obtained promising performances in low FP/scan conditions.

Our current work mostly focused on FP reduction given coordinates of many candidate nodules. We believe that our network can be converted to accomplish positive nodule detection on the low-dose CT scans directly with minor modifications, such as replacing the fully-connected layers with 1×1 convolution layers. For clinical practice, it is also important to classify nodules into various subtypes of solid, non-solid, part-solid, perifissural, calcified, spiculated (Ciompi et al., 2017), for which different treatments can be used. Thus, it will be our forthcoming research direction.

Acknowledgments

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2017-0-01779, A machine learning and statistical inference framework for explainable artificial intelligence).

References

Armato, S. G., McLennan, G., Bidaut, L., McNitt-Gray, M. F., Meyer, C. R., Reeves, A. P., et al. (2011). 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), 915–931.

Cao, P., Liu, X., Yang, J., Zhao, D., Li, W., Huang, M., et al. (2017). A multi-kernel based framework for heterogeneous feature selection and over-sampling for computer-aided detection of pulmonary nodules. *Pattern Recognition*, *64*, 327–346.

Chan, H.-P., Hadjiiski, L., Zhou, C., & Sahiner, B. (2008). Computer-aided diagnosis of lung cancer and pulmonary embolism in computed tomography - A review. *Academic Radiology*, *15*(5), 535–555. http://dx.doi.org/10.1016/j.acra.2008.01.014, URL http://www.sciencedirect.com/science/article/pii/S1076633208000470.

Cheng, J.-Z., Ni, D., Chou, Y.-H., Qin, J., Tiu, C.-M., Chang, Y.-C., et al. (2016). Computer-aided diagnosis with deep learning architecture: Applications to breast lesions in US images and pulmonary nodules in CT scans. *Scientific Reports*, *6*, 24454. http://dx.doi.org/10.1038/srep24454.

Ciompi, F., Chung, K., van Riel, S. J., Setio, A. A. A., Gerke, P. K., Jacobs, C., et al. (2017). Towards automatic pulmonary nodule management in lung cancer screening with deep learning. *Scientific Reports*, *7*(46479).

Ciompi, F., de Hoop, B., van Riel, S. J., Chung, K., Scholten, E. T., Oudkerk, M., et al. (2015). Automatic classification of pulmonary peri-fissural nodules in computed tomography using an ensemble of 2D views and a convolutional neural network out-of-the-box. *Medical Image Analysis*, *26*(1), 195–202.

Ding, J., Li, A., Hu, Z., & Wang, L. (2017). Accurate Pulmonary Nodule Detection in Computed Tomography Images Using Deep Convolutional Neural Networks, arXiv preprint arXiv:1706.04303 (pp. 1–9).

Dou, Q., Chen, H., Yu, L., Qin, J., & Heng, P. A. (2017). Multi-level contextual 3D CNNs for false positive reduction in pulmonary nodule detection. *IEEE Transactions on Biomedical Engineering*, *64*(7), 1558–1567.

Dou, Q., Chen, H., Yu, L., Zhao, L., Qin, J., Wang, D., et al. (2016). Automatic detection of cerebral microbleeds from MR images via 3D convolutional neural networks. *IEEE Transactions on Medical Imaging*, *35*(5), 1182–1195.

Efron, B., & Tibshirani, R. (1994). *An introduction to the bootstrap* (p. 436). Chapman & Hall.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., et al. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639), 115–118.

Friedman, J., Hastie, T., Tibshirani, R., & Stanford, Y. (1998). Additive logistic regression: a statistical view of boosting. *The Annals of Statistics*, *28*(2), 337–407.

Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the international conference on artificial intelligence and statistics* (pp. 249–256).

Gould, M. K., Fletcher, J., Iannettoni, M. D., Lynch, W. R., Midthun, D. E., Naidich, D. P., et al. (2007). Evaluation of patients with pulmonary nodules: When is it lung cancer? ACCP evidence-based clinical practice guidelines (2nd edition). *Chest*, *132*(3), 108S–130S.

Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., et al. (2017). Brain tumor segmentation with deep neural networks. *Medical Image Analysis*, *35*, 18–31.

Honari, S., Yosinski, J., Vincent, P., & Pal, C. (2016). Recombinator networks: Learning coarse-to-fine feature aggregation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1–11).

Hounsfield, G. (1980). Computed medical imaging. *Science*, *210*(4465).

Hu, Z., Tang, J., Wang, Z., Zhang, K., Zhang, L., & Sun, Q. (2018). Deep learning for image-based cancer detection and diagnosis - A survey. *Pattern Recognition*, *83*, 134–149.

Jacobs, C., van Rikxoort, E. M., Twellmann, T., Scholten, E. T., de Jong, P. A., Kuhnigk, J.-M., et al. (2014). Automatic detection of subsolid pulmonary nodules in thoracic computed tomography images. *Medical Image Analysis*, *18*(2), 374–384.

Kamnitsas, K., Ledig, C., Newcombe, V. F., Simpson, J. P., Kane, A. D., Menon, D. K., et al. (2017). Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Medical Image Analysis*, *36*, 61–78.

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1725–1732).

Lee, Y., Hara, T., Fujita, H., Itoh, S., & Ishigaki, T. (2001). Automated detection of pulmonary nodules in helical CT images based on an improved template-matching technique. *IEEE Transactions on Medical Imaging*, *20*(7), 595–604.

Li, Q., Sone, S., & Doi, K. (2003). Selective enhancement filters for nodules, vessels, and airway walls in two- and three-dimensional CT scans. *Medical Physics*, *30*(8), 2040–2051.

Lin, M., Chen, Q., & Yan, S. (2013). Network in Network, arXiv preprint arXiv:1312.4400.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2016). Feature Pyramid Networks for Object Detection, arXiv preprint arXiv:1612.03144.

Liu, X., Hou, F., Qin, H., & Hao, A. (2018). Multi-view multi-scale CNNs for lung nodule type classification from CT images. *Pattern Recognition*, *77*, 262–275.

Murphy, K., van Ginneken, B., Schilham, A. M., De Hoop, B., Gietema, H., & Prokop, M. (2009). A large-scale evaluation of automatic pulmonary nodule detection in chest CT using local image fatures and k-nearest-neighbour classification. *Medical Image Analysis*, *13*(5), 757–770.

Nair, V., & Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In *Proceedings of international conference on machine learning* (pp. 807–814).

National Lung Screening Trial Research Team, Aberle, D. R., Adams, A. M., Berg, C. D., Black, W. C., Clapp, J. D., Fagerstrom, R. M., et al. (2011). Reduced lung-cancer mortality with low-dose computed tomographic screening. *New England Journal of Medicine*, *365*(5), 395–409.

Niemeijer, M., Loog, M., Abramoff, M. D., Viergever, M. A., Prokop, M., & van Ginneken, B. (2011). On combining computer-aided detection systems. *IEEE Transactions on Medical Imaging*, *30*(2), 215–223.

Okumura, T., Miwa, T., Kako, J.-I., Yamamoto, S., Matsumoto, R., Tateno, Y., et al. (1998). Automatic detection of lung cancers in chest CT images by variable N-Quoit filter. In *Proceedings of international conference on pattern recognition, Vol. 2* (pp. 1671–1673).

Razavian, A. S., Azizpour, H., Sullivan, J., & Carlsson, S. (2014). CNN Features off-the-shelf: An astounding baseline for recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 512–519).

Roth, H. R., Lu, L., Liu, J., Yao, J., Seff, A., Cherry, K., et al. (2016). Improving computer-aided detection using convolutional neural networks and random view aggregation. *IEEE Transactions on Medical Imaging*, *35*(5), 1170–1181.

Sakamoto, M., Nakano, H., Zhao, K., & Sekiyama, T. (2017). Multi-stage neural networks with single-sided classifiers for false positive reduction and its evaluation using lung X-Ray (CT) images. In *Proceedings of International Conference Image Analysis and Processing* (pp. 370–379).

Setio, A. A. A., Ciompi, F., Litjens, G., Gerke, P., Jacobs, C., Van Riel, S. J., et al. (2016). Pulmonary nodule detection in CT images: False positive reduction using multi-view convolutional networks. *IEEE Transactions on Medical Imaging*, *35*(5), 1160–1169.

Setio, A. A. A., Jacobs, C., Gelderblom, J., & van Ginneken, B. (2015). Automatic detection of large pulmonary solid nodules in thoracic CT images. *Medical Physics*, *42*(10), 5642–5653.

Setio, A. A. A., Traverso, A., De Bel, T., Berens, M. S., van den Bogaard, C., Cerello, P., et al. (2017). Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge. *Medical Image Analysis*, *42*, 1–13.

Shen, D., Wu, G., & Suk, H.-I. (2017). Deep learning in medical image analysis. *Annual Review of Biomedical Engineering*, *19*, 221–248.

Shen, W., Zhou, M., Yang, F., Yang, C., & Tian, J. (2015). Multi-scale convolutional neural networks for lung nodule classification. In *Proceedings of international conference on information processing in medical imaging* (pp. 588–599). Springer.

Shen, W., Zhou, M., Yang, F., Yu, D., Dong, D., Yang, C., et al. (2017). Multi-crop convolutional neural networks for lung nodule malignancy suspiciousness classification. *Pattern Recognition*, *61*, 663–673.

Shin, H. C., Roth, H. R., Gao, M., Lu, L., Xu, Z., Nogues, I., et al. (2016). Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Transactions on Medical Imaging*, *35*(5), 1285–1298.

Siegel, R. L., Miller, K. D., & Jemal, A. (2017). Cancer statistics, 2017. *CA: A Cancer Journal for Clinicians*, *67*(1), 7–30.

Suzuki, K. (2012). A review of computer-aided diagnosis in thoracic and colonic imaging. *Quantitative Imaging in Medicine and Surgery*, *2*(3).

Suzuki, K., Armato, S. G., Li, F., Sone, S., & Doi, K. (2003). Massive training artificial neural network (MTANN) for reduction of false positives in computerized detection of lung nodules in low-dose computed tomography. *Medical Physics*, *30*(7), 1602–1617.

Tan, M., Deklerck, R., Jansen, B., Bister, M., & Cornelis, J. (2011). A novel computer-aided lung nodule detection system for CT images. *Medical Physics*, *38*(10), 5630–5645.

Traverso, A., Torres, E. L., Fantacci, M. E., & Cerello, P. (2017). Computer-aided detection systems to improve lung cancer early diagnosis: state-of-the-art and challenges. *Proceedings of Journal of Physics: Conference Series*, *841*, 012013.

Yang, Q., & Pan, S. J. (2009). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, *22*, 1345–1359.

Ye, X., Lin, X., Dehmeshki, J., Slabaugh, G., & Beddoe, G. (2009). Shape-based computer-aided detection of lung nodules in thoracic CT images. *IEEE Transactions on Biomedical Engineering*, *56*(7), 1810–1820.

Zhang, J., Atkinson, P. M., & Goodchild, M. F. (2014). *Scale in spatial information and analysis* (p. 345). CRC Press, Taylor and Francis.