International Journal of Computer Assisted Radiology and Surgery (2018) 13:585–595
https://doi.org/10.1007/s11548-017-1696-0

ORIGINAL ARTICLE

Agile convolutional neural network for pulmonary nodule classification using CT images

Xinzhuo Zhao[1] · Liyao Liu[1] · Shouliang Qi[1] · Yueyang Teng[1] · Jianhua Li[1] · Wei Qian[1,2]

Received: 2 June 2017 / Accepted: 20 December 2017 / Published online: 23 February 2018

Abstract

Objective To distinguish benign from malignant pulmonary nodules using CT images is critical for their precise diagnosis and treatment. A new Agile convolutional neural network (CNN) framework is proposed to conquer the challenges of a small-scale medical image database and the small size of the nodules, and it improves the performance of pulmonary nodule classification using CT images.

Methods A hybrid CNN of LeNet and AlexNet is constructed through combining the layer settings of LeNet and the parameter settings of AlexNet. A dataset with 743 CT image nodule samples is built up based on the 1018 CT scans of LIDC to train and evaluate the Agile CNN model. Through adjusting the parameters of the kernel size, learning rate, and other factors, the effect of these parameters on the performance of the CNN model is investigated, and an optimized setting of the CNN is obtained finally.

Results After finely optimizing the settings of the CNN, the estimation accuracy and the area under the curve can reach 0.822 and 0.877, respectively. The accuracy of the CNN is significantly dependent on the kernel size, learning rate, training batch size, dropout, and weight initializations. The best performance is achieved when the kernel size is set to 7 × 7, the learning rate is 0.005, the batch size is 32, and dropout and Gaussian initialization are used.

Conclusions This competitive performance demonstrates that our proposed CNN framework and the optimization strategy of the CNN parameters are suitable for pulmonary nodule classification characterized by small medical datasets and small targets. The classification model might help diagnose and treat pulmonary nodules effectively.

Keywords Lung cancer · Nodule classification · Deep learning · Convolutional neural network

Introduction

Lung cancer is the leading cause of cancer deaths in the world [1]. It has become the first killer among cancers in China, partially due to the asymptomatic growth of this cancer [2,3]. In the majority of cases, it is too late for successful therapy once the patient develops the first symptoms. However, there is a survival rate of 47% if the lung cancer is detected early according to the American Cancer Society. Therefore, early determination of whether a pulmonary nodule is benign or malignant is important.

Computed tomography (CT) scanners can provide continuous high-resolution, near-isotropic thin sections throughout the lungs in a single-breath hold. These CT images delineate the location, size and shape of the suspicious pulmonary nod-

Liyao Liu: Joint first author.

✉ Shouliang Qi
qisl@bmie.neu.edu.cn

Xinzhuo Zhao
xzhzhao@mail.neu.edu.cn

Liyao Liu
6021200664@qq.com

Yueyang Teng
tengyy@bmie.neu.edu.cn

Jianhua Li
lijh@bmie.neu.edu.cn

Wei Qian
wqian@bmie.neu.edu.cn

[1] Sino-Dutch Biomedical and Information Engineering School, Northeastern University, Life Science Building, 500 Zhihui Street, Hun’nan District, Shenyang 110169, China

[2] College of Engineering, University of Texas at El Paso, 500 West University Avenue, El Paso, Texas 79968, USA

ules [4]. Via imaging processing techniques, some computer aided diagnosis (CAD) systems have been implemented to estimate the malignance of the detected pulmonary nodules [5]. In these systems, after the lung nodules are segmented, various types of image features (e.g., intensity) are extracted [6]. Then, machine-learning classifiers are used to predict the malignance [7]. However, several challenges have faced these handcrafted feature-based CAD systems. First, the handcrafted features depend on the segmentation of the lung nodule. However, this step is challenging and contentious because whether there is ground truth is open to debate and the reproducibility of the segmentation is contingent [8]. Second, the handcrafted features are based on prior knowledge, which is dependent on the ability of the designers of the CAD system. These challenges make the handcrafted feature-based CAD systems difficult for clinical applications.

Deep learning, especially the convolutional neural network (CNN), might have the potential to address the aforementioned challenges, considering its significant success in object recognition and localization in nature images [9]. One of the advantages of the CNN is that it can be fed raw images without previous image preprocessing, which is highly amenable to image analysis. Deep learning consists of increased numbers of layers, which permits higher levels of abstraction and improved predictions from data [10].

Many deep learning networks with more layers and flexible structures have been proposed since LeNet-5 [11]. For example, AlexNet [12] contains eight learned layers, and VGG-VD [13] has 16-layer and 19-layer CNN structures. GoogLeNet [14], a 22-layer deep network that contains inception architectures, is proposed to manage the contradiction between increasing the training parameters and overfitting. ResNet [15] is approximately 20 times deeper than AlexNet and 8 times deeper than VGGNet. By increasing the depth, the network can better approximate the target function with increased nonlinearity and achieve better feature representations.

The applications of the CNN to medical images are quite different from those for nature images in several respects. CNN requires a large number of labeled training data acting as ImageNet. However, large datasets are not always available because of the extremely expensive expert annotations and scarcity of the disease images [9]. Moreover, instead of containing RGB channels as in natural images, medical images are grayscale images.

These important differences between medical and nature images have prompted investigators to study whether CNNs can be used effectively for lung nodule classification. Hua et al. [16] first introduced the CNN to nodule classification in CT images and found that it outperforms the conventional handcrafted feature-based CAD frameworks. Sun et al. [17] found that the deep belief networks (DBN) performed best followed by CNNs and stacked denoising autoencoder (SDAE). Cheng et al. [18] developed an SDAE CAD system with an accuracy of 94.40%. Kumar et al. [19] utilized the autoencoder to extract image features and used the decision tree to realize classification. Wei et al. [20] utilized multiscale CNNs to capture features from raw nodule patches and classified the nodules with SVM, achieving an accuracy of 86.84%. They further proposed a multi-crop CNN to increase the accuracy to 87.14% [21].

To further increase the accuracy of classification of lung nodules, the new CNN network architecture and optimization strategy for the learning parameters are required. In this paper, the Agile CNN architecture, which is suitable for small datasets of lung nodule CT images, is proposed and implemented. In the Agile CNN, which has only two convolutional layers, a small number of kernels (20 kernels in C1 and 1000 kernels in C2) are adopted. Compared with those famous deep or deeper structures, such as the GoogleNet, ResNet, and VGGNet, the Agile structure has relatively fewer layers (only 2 convolutional layers), and thus, it is called the "Agile" CNN. Additionally, the number of parameters to calculate for the training is determined by the number of the layers, the number of the kernels, and the sizes of the kernels. In addition, more parameters for training require more input data, which is not always feasible in medical applications. Thus, the strategy for the optimization of learning parameters of CNNs is clarified to increase the accuracy of the classification and to avoid overfitting at the same time. Finally, a classification model for classifying the malignant pulmonary nodules from the others based on CT scan images is obtained.

Materials and methods

Dataset of lung nodule CT images

The images in the current study are generated from the Lung Image Database Consortium image collection (LIDC-IDRI) [22–24]. So far, it contains 1018 cases. Each subject includes images from a clinical thoracic CT scan and an associated XML file that records the results of a two-phase image annotation process performed by four radiologists. Each radiologist independently reviewed each CT scan and marked lesions that belonged to one of three categories ("nodule $>$ or $=$ 3 mm", "nodule $<$ 3 mm" and "non-nodule $>$ or $=$ 3 mm"). Then, the nodules are marked with 5 malignancy levels, from 1 to 5.

To generate the training dataset, several steps are implied. First, we select nodules that are larger than 3 mm for this study. Since each nodule is labeled by four radiologists, those that are recognized by fewer than three of the radiologists are eliminated. At the same time, we label the nodules according to their malignancy levels, i.e., the average rating of the four radiologists. Levels 1 to 2.5 are considered to be benign, and

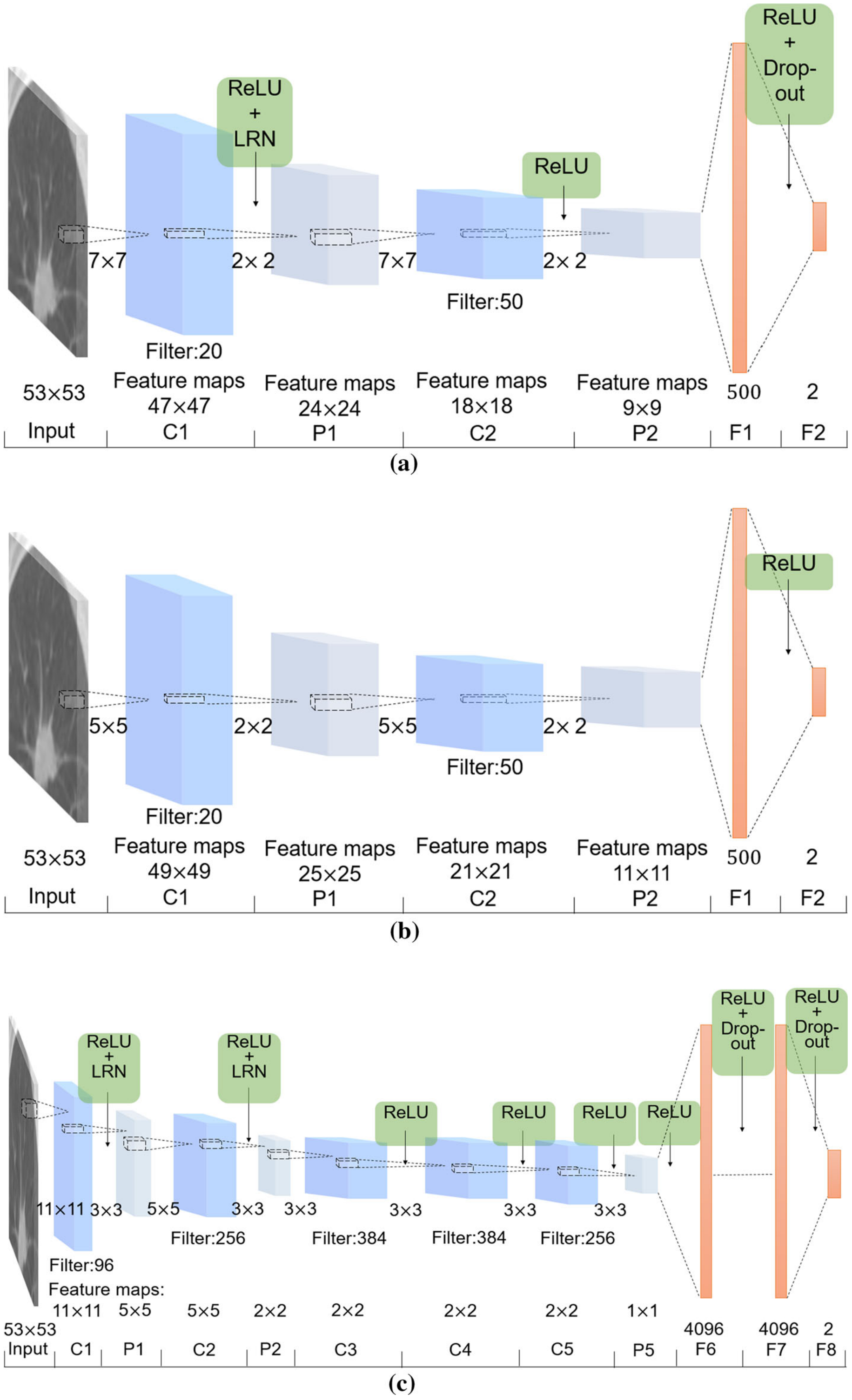

Fig. 1 Schematic structure of the proposed Agile CNN, LeNet and AlexNet. **a** Presents the schematic structure of the proposed agile CNN. It is a hybrid structure of LeNet and AlexNet, combining the layer settings of LeNet and the parameter settings of AlexNet. **b** Shows the schematic structure of LeNet. LeNet has two convolutional layers, two pooling layers, and two fully connected layers, and it is designed for the image size of 28 × 28. **c** AlexNet is designed for the image size of 256 × 256. It has more convolutional layers, with the layers of ReLU, LRN, and dropout in this framework

levels 3.5 to 5 are denoted as malignant, and there is elimination of all of the intermediate cases (level 3). In total, there are 743 nodules left, with 375 malignant nodules and 368 benign nodules. Then, because of the varying image resolutions, the nodules are resampled using spline interpolation with a fixed resolution with 0.5 mm/voxel along two axes [25]. Third, the nodule areas are annotated based on the union of the radiologists' truth files, obtaining the minimum bounding rectangle of each slice. The size of the cropped patch of slices is fixed at 53 by 53. Instead of centralizing the nodules, they are located at random positions of the patch. To utilize the background information, the surrounding pixels are preserved. Since the

Table 1 Computer environment

Computer environment	Detail
Computer	HP Z840
GPU	NVIDIA Quadro M2000 4 GB
CPU	Intel Xeon E5-2640v3 2.60 GHz
Memory	RAM 64G
Operating system	Linux ubuntu 16.04 64-bit
CUDA	CUDA 8.0
cuDNN	cuDNN 5.1

CNN used in this paper is a 2D structure, each slice of a nodule is cropped as a patch.

CNN experiment

CNN architecture

The CNN architecture consists of a number of convolutional and pooling layers optionally followed by fully connected layers. The convolutional layer is composed of several small matrices or "kernels" that are convolved throughout the whole input image, which work as filters. The output of this convolution is called a "feature map". These feature maps are the input for the pooling layer, which aggregates contiguous values to one scalar with functions such as mean or max [26]. In the following parts, the convolutional layers are labeled Cx, the pooling layers Px, and the fully connected layers Fx, where x is the layer index.

The Agile CNN framework is proposed in the current study, as shown in Fig. 1. It is a hybrid structure of LeNet and AlexNet, which combines the layer settings of LeNet and the parameter settings of AlexNet. In other words, we start from the LeNet framework, add the layers of ReLU, LRN, and dropout into this framework, and construct the Agile CNN. Inspired by LeNet, the proposed CNN has two convolutional layers, two pooling layers, and two fully connected layers. Layer C1 has 20 feature maps. Every unit in each feature map is connected to a 7 × 7 neighborhood in the input. The size of the input patch is 53 × 53, and the size of the feature maps in C1 is 47 × 47, which prevents a connection from the input from falling out of the boundary. In P1, every unit in each feature map is connected to a 2 × 2 neighborhood in the corresponding feature map in C1. Then, layer C2 has 50 feature maps. The other settings are the same as the previous layers. Finally, F1 and F2 follow after layer P2. The number of neuron units in F1 and F2 is 500 and 2, respectively.

The experiment environment is listed in Table 1. The platform that we work on is Caffe 1.0 (Convolutional Architecture for Fast Feature Embedding).

Inspired by AlexNet, two convolutional layers in our framework are followed by a rectified linear unit (ReLU) layer and a local response normalization (LRN) layer [27]. A deep CNN with ReLU trains several times faster than those with sigmoid and other logistic functions. ReLU is a non-saturating neuron, which avoids the problem of gradient vanishing [12]. The non-saturating nonlinearity of ReLU can be shown as

$$\varphi(x) = \max(0, x) \tag{1}$$

In addition to the ReLU layer, the LRN scheme aids the generalization of the network. The performance of LRN appears to be a type of "lateral inhibition." At the LRN layer, each input $a_{x,y}^{i}$ is divided by an expression:

$$b_{x,y}^{i} = a_{x,y}^{i} / \left(1 + \left(\frac{\alpha}{n}\right) \sum_{i} x_{i}^{2}\right)^{\beta} \tag{2}$$

Here, $a_{x,y}^{i}$ is the input neuron at position (x, y) applied by kernel i, where the sum runs over the adjacent kernel maps at the same spatial position. There are two modes of LRN [28]: one is the in-channel mode and the other is the cross-channel mode. Here, the cross-channel mode is selected with $n = 5$, $\alpha = 0.0001$ and $\beta = 0.75$ as in AlexNet's set.

Motivated by Srivastava [29], the dropout is applied in F1, while setting the output hidden neuron to zero with a probability of 0.5. The role is to reduce the complex coadaptations of the neurons in F1.

Parameters for optimization

After determining the architecture of the CNN, another important task is to optimize the parameters to improve the performance of the proposed CNN for lung nodule classification. There are four main parts: (1) kernel size; (2) learning rate; (3) batch size; and (4) weight initialization.

A. Kernel size

The kernel size and the kernel number are two significant parameters that affect the learning efficiency of the system. The number of parameters to be learned is proportional to the kernel size, the previous kernel number and the current kernel number. In our architecture, the number of learned parameters can be calculated as 7 × 7 × 20 × 50, which is 49,000. The number of learned parameters must adapt to the number of training images, which not only guarantees the richness of image features but also avoids overfitting.

B. Learning rate

Beyond choosing a single global learning rate, it is clear that picking a different learning rate η can improve the convergence. Whenever the loss function stops to decay, the learning rate is multiplied by a factor γ. During the whole training experiment, the learning rate has decayed for several times. In each instance of decay, the learning rate η must be multiplied by γ. In Caffe, the decay time of the learning rate is defined as a "step."

C. Batch size

The ability to perform generalization by a network also relates to the batch size [30]. CNNs with large batch sizes tend to make the training and testing functions converge to sharp minimizers, which leads to the neural network to having poor generalizability. In contrast, CNNs with small batch sizes consistently converge to flat minimizers.

D. Weight initialization

The initial values of the weights have a significant effect on the training process. If the randomly chosen weights are all very large, then the ReLU will saturate, which results in small gradients that make learning slow. If the randomly chosen weights are very small, then the gradients will also be very small. Intermediate weights with a Gaussian distribution with a mean of 0 and a standard deviation of 0.01 have two advantages: (1) the gradients are sufficiently large that learning can proceed, and (2) the network will learn the linear part of the mapping before the more difficult nonlinear part.

Training, testing, and parameter optimization

We train and evaluate CNNs using tenfold cross-validation. The 743 nodules are split into training, validation, and testing datasets. In each fold of the cross-validation, 10% patients are used to test the architecture. To augment the training and validating datasets, each slice is cropped four times randomly and rotated three times with the angles of 90, 180, and 270 degrees. Each of them is flipped horizontally and vertically. Data augmentation is not applied to the testing dataset. To evaluate the effect of each parameter on the performance of the CNN, several groups of control experiments were designed. During the experiments, the same conditions were maintained except for in one particular factor, and then, the effect of this varied factor was evaluated.

To study the effect of the kernel size on the performance of the proposed CNN, it was varied from 3 × 3 to 9 × 9 while freezing the other parameters. The variation of the kernel size is described in detail in Table 2. This group of experiments is used to observe the effect of different convolution kernel sizes on the performance of our architecture.

The effect of the learning rate and the number of steps is also investigated. The variation in the learning rate and the decay times (steps) of the learning rate during the train iteration process are exhibited in Table 2. The decay factor γ is set to 0.1. Moreover, the effect of the batch size is also studied while altering it from 32 to 64. Along with the batch size, the influence of dropout is also checked.

Table 2 Summary of the optimized parameters

Parameter	Variation	
Kernel size	C1	C2
	3 × 3	3 × 3
	5 × 5	5 × 5
	7 × 7	**7 × 7**
	9 × 9	9 × 9
	9 × 9	7 × 7
	7 × 7	5 × 5
	5 × 5	3 × 3
Learning rate	0.01	
	0.005	
	0.001	
	0.0005	
	0.0001	
Decay times (learning rate)	4	
	4.5	
	5	
	5.5	
	6	
Batch size	**32**	
	64	
Weight initialization	Xavier	
	Gaussian	**0.001**
		0.005
Bias initialization	Constant 1	
	Constant 0	
Dropout	**Yes**	
	No	

The bold indicates the optimized parameters which are adopted in the current study

For the weights of C1, C2, and F1, an experiment is performed to compare the Xavier and Gaussian initialization methods. The standard deviation of the Gaussian is also shifted from 0.001 to 0.005 for further experiments. In addition, a constant (0 or 1) is used to initialize the bias of C1, C2, F1 and F2.

Results

CNN classification performance

After the comparison of the layers and parameter settings, the Agile CNN structure was optimized. It contains two convolutional layers and two fully connected layers, all of which are followed by ReLU and LRN. Finally, we set two convolutional layer kernel sizes to be 7 × 7. Dropout is used in the fully connected layer. The base learning rate is 0.0005, with a 5.5 times reduction to diminish the learning speed.

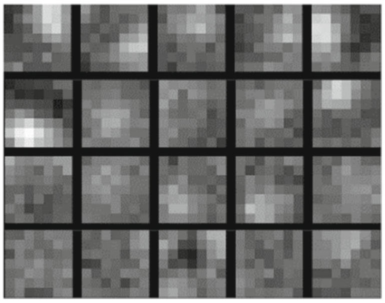

Fig. 2 Visualization of the 20 kernels in the C1. There are 20 kernels with the size of 7 × 7 in the first convolutional layer. The kernels are presented in 4 rows and 5 columns without any particular order

The batch size is 32. The weights of the C1, C2 and F2 are initialized by Gaussian, with a standard deviation of 0.01. For the initialization of the weights of the F1, the standard deviation of the Gaussian is 0.001. The bias of the C2 and F1 is initialized as the constant of 0.1, and the bias of the C1 and F2 is initialized as the constant of 0.0.

The CNN algorithm helps the computer learn its own features, instead of using handcrafted features. The visualization of the weights is one of the important methods of evaluating the features. Figure 2 shows the visualization of the weights in C1. These kernels not only present a noise pattern but also exhibit a high correlation. Additionally, they also lack structural patterns. These characteristics implicate under-training. The features in C2, as shown in Fig. 3, obviously perform better, displaying a smoother pattern, containing more shape information, and demonstrating that C2 is well trained.

After 50,000 iterations, 10% of the nodules are utilized to evaluate the performance of our framework. We obtain the final classification model with a test accuracy of 0.822 and an AUC of 0.877. Figure 4 shows the test accuracy of the Agile CNN structure, LeNet, and AlexNet. The Agile CNN has an accuracy of 0.822, which is higher than that of LeNet, which is 0.648, and that of AlexNet, which is 0.782. The T-test and Wilcoxon signed ranks test are used to compare the classification accuracy of the Agile CNN structure with that of LeNet and AlexNet. The significance test results are shown in Table 3. It was found that the test accuracy of the Agile CNN structure is significantly higher than that of LeNet (T-test, $p < 0.05$) and that of AlexNet (Wilcoxon signed ranks test, $p < 0.08$). Figure 5 presents the Receiver Operating Characteristic (ROC) curve of the test result of the Agile CNN structure for nodule classification.

Table 4 summarizes the methods and accuracy reported in recently published papers using deep learning algorithms based on the LIDC dataset. The accuracy of our method is higher than that of Kumar et al. [19] and Sun et al. [17], but lower than that of Wei et al. [20,21].

The effect of different parameter settings

Figure 6 shows the accuracy of validation with iterations for seven structures with different kernel sizes. The kernel size of 7 × 7 for two convolutional layers is the best, as indicated by the red crossline. The relationship between the accuracy and kernel size is not monotonic: the accuracy increases with the kernel size while it is smaller than 9 × 9, it reaches the peak at the size of 7 × 7, and it decreases with increasing size.

Table 5 shows the validation accuracy of the proposed CNN with different learning rates and steps. On the left side of Table 5, the learning rate (Lr) remains unchanged (constant learning speed) when the learning “step” equals zero during one whole iteration. It is found that the validation accuracy of the last row ($Lr = 0.0001$) is higher compared with the former rows. However, the loss of the last row is also larger, which indicates that the validation process starts overfitting. The right side of Table 5 shows the performance of CNN with different steps and a constant learning rate of 0.0005.

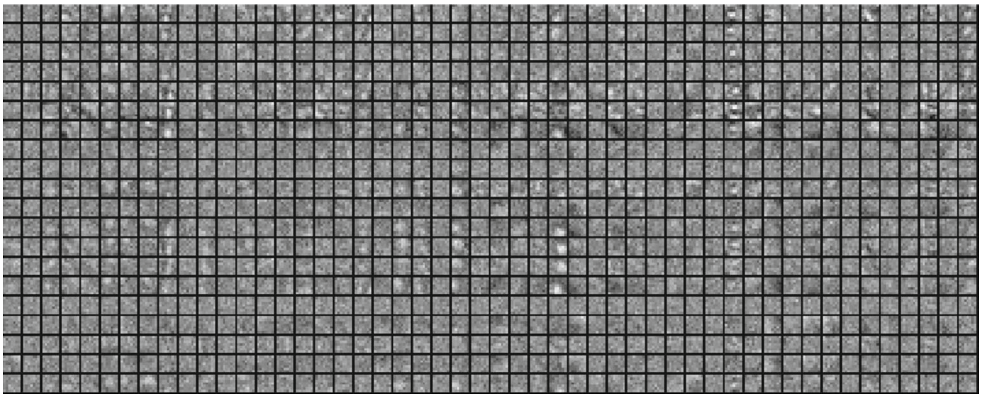

Fig. 3 Visualization of the 1000 kernels in the C2. There are 1000 kernels with the size of 7 × 7 in the second convolutional layer. The kernels are presented in 20 rows and 50 columns without any particular order

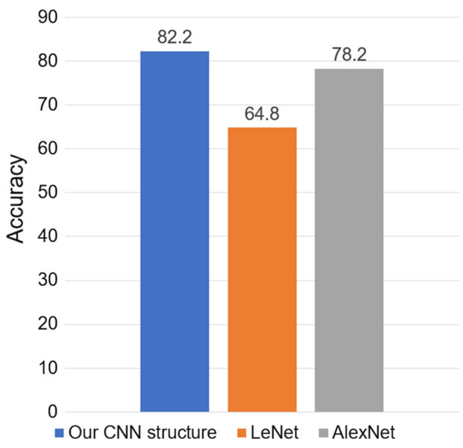

Fig. 4 Test accuracy of the Agile CNN structure, LeNet and AlexNet. The Agile CNN has an accuracy of 0.822, which is higher than that of LeNet, which is 0.648, and AlexNet, which is 0.782

Table 3 Significance tests of the classification accuracy of the Agile CNN structure, LeNet, and AlexNet

Model	T test	Wilcoxon signed ranks test
Our CNN structure versus LeNet	$p = 5.52 \times e - 05$	$p = 2.45 \times e - 04$
Our CNN structure versus AlexNet	$p = 0.135$	$p = 0.076$

Table 4 Summary of the methods and accuracy of the recently published papers using LIDC as a dataset

Experiments	Year	Method	Accuracy (%)
Wei et al. [21]	2017	Multi-crop CNN	87.14
		Multi-scale CNN	86.53
		CNN	86.32
Sun et al. [17]	2016	CNN	79.76
		CNN	81.19
		SDAE	79.29
Wei et al. [20]	2015	Multi-scale CNN to extract feature with the SVM classifier	86.84
Kumar et al. [19]	2015	Autoencoder to extract feature with the decision tree classifier	75.01
Our method		CNN	82.23

With each reduction step, the learning rate is multiplied by 0.1. Finally, the base learning rate of 0.0005 and step of 5.5 achieves the best result, which not only obtains the highest accuracy but also reduces the loss value.

Figure 7 shows the validation accuracy of the experimented CNNs with different validation batch sizes and with or without dropout. There are mainly four observations. First, it is shown that dropout can reduce the loss value and prevent overfitting. Second, the CNN with a batch size of 32 performs better than that with a batch size of 64. Third, Xavier performs worse than the Gaussian, and thus, a Gaussian with a constant bias should be used to initialize the weights. Fourth, for the initialization of the weight of F1, a standard deviation of the Gaussian of 0.001 is better than 0.005, as set in AlexNet. For the initialization of the bias of the C2 and F1, the constant of 0 is better than 1, as set in ImageNet.

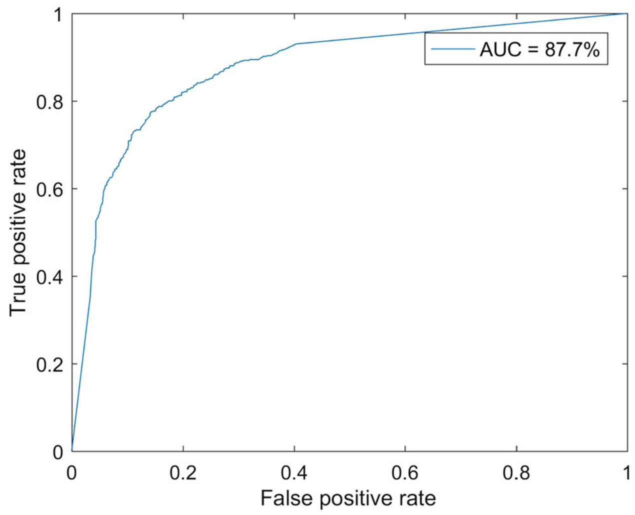

Fig. 5 Receiver operating characteristic (ROC) curve of the test accuracy for the proposed nodule classification model

Discussion and future work

In this paper, we proposed the Agile CNN for pulmonary nodule malignancy classification using CT images. This Agile CNN is designed to be a feature extractor and classifier. The parameters of the CNN are optimized through a series of experiments to obtain a high accuracy and low loss. The results demonstrate that the Agile CNN presents competitive performance. From our viewpoint, three main contributions are included in the current work. First, a method is proposed through combining LeNet and AlexNet into the Agile CNN. Second, a strategy for optimizing the parameters in the CNN is developed. Third, a model for classification is obtained for discriminating the benign and malignant lung nodules using CT images.

CNN framework and performance

For the CNN layer setting, the hybrid structure is based on the improved LeNet [11] and AlexNet [12]. The LeNet CNN model is applied to check the handwriting and is suitable

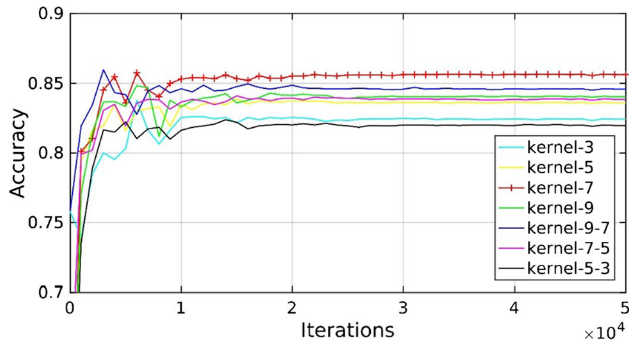

Fig. 6 Validation accuracy of various kernel sizes in the proposed nodule classification model. The red cross line achieves the highest accuracy, which means that the kernel size of 7 × 7 is the best. Based on all of the lines, there is a tendency that the accuracy increases with the size of the kernel, and it reaches the highest accuracy. Furthermore, by observing the red cross line, it is found that the accuracy increases quickly at the beginning. After a small fluctuation, the accuracy tends to be stable. With more iterations, the accuracy does not decrease. This finding illustrates that the model does not overfit

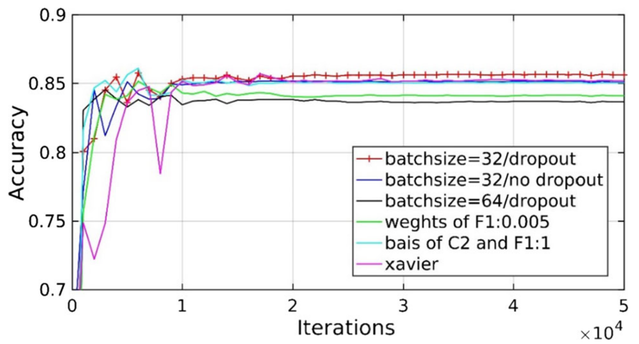

Fig. 7 Validation accuracy of various batch sizes, dropout and weights and bias initialization in the proposed nodule classification model. Comparing the red cross line with the dark blue line, it is found that dropout helps to increase the validation accuracy. The batch size of 64 (black line) underperforms the batch size of 32 (dark blue). The pink line presents the Xavier initialization, which underperforms that initialized by the Gaussian method (the red cross line). The azure line (the standard deviation of the Gaussian is 0.001) and the green line (the standard deviation of the Gaussian is 0.0.005) are used to show the influence of the standard deviation of the Gaussian. It is proven that the standard deviation of the Gaussian being 0.001 is better than 0.005

Table 5 Validation accuracy of the proposed CNN with different learning rates and steps

Learning rate (Step = 0)	Accuracy	Step (Learning rate = 0.0005)	Accuracy
0.01	0.7576	4	0.8424
0.005	0.7576	4.5	0.8462
0.001	0.8245	5	0.8404
0.0005	**0.8408**	**5.5**	**0.8564**
0.0001	0.8449	6	0.8503

The bold indicates the optimized parameters which are adopted in the current study

for a small dataset and images with a small size, such as our experimental dataset. However, the AlexNet model utilizes a deeper CNN that is suitable for a large dataset with large input images. For the CNN structure used in medical image classification tasks, Sun et al. [17], Setio et al. [31], and He et al. [15] have all built architectures with three convolutional layers. Based on our experiment with a two-layer CNN, three layers are more likely to overfit. Additionally, a structure with two convolutional layers followed by two fully connected layers can obtain a higher accuracy. This proposed structure performs well for a small-scale and small-size medical dataset.

The deep learning framework can have an important impact on the accuracy of the lung nodule classification. Our results have shown that the proposed CNN achieves better performance than the Autoencoder to extract features with the Decision Tree classifier [19] and DBNs, as well as SDAE [17]. Even using the CNNs, the framework of our model is more suitable for lung nodule classification than that of Sun et al. [17], which has three convolutional levels. However, it is noted that our methods cannot reach the accuracy of Wei et al. [20,21]. There are three possible reasons: (1) The 3D CNN and multi-scale strategy is used [20,21]; (2) The more complicated network with a multi-crop strategy is adopted; (3) more features, including histogram of oriented gradient (HOG) and local binary patterns (LBP), are extracted and combined with the features extracted using CNN.

CNN parameter settings

To further improve the performance, we optimized the parameter settings of the CNN. For the kernel size, Sun et al. [17] utilized kernel sizes of 12, 8, and 6. Setio et al. [31] designed their structure with 5, 3, and 3. Compared with the three-layer CNN structure, our two-layer structure uses larger kernels, which are 7 and 7. The large size of the kernels can create a wider receptive field. It is also different from Shin et al. [32], which set the batch size at 50, Sun et al. [17] at 100, and Hinton et al. [33] and Setio et al. [31] at 128, and our smaller batch size of 32, as demonstrated above, performs better. For the initialization of the weights, Xavier does not have an effect on the network convergence compared with the Gaussian initialization, based on our observations.

Evaluation of the misclassified samples

Figure 8 gives some patches of true negative (TN), false positive (FP), true positive (TP), and false negative (FN) results. The first row is TN, and the third row is TP. One general feature can be found: the small-size nodules with regular surroundings belong to the negative (or benign) class; the large size nodules with irregular surroundings have a higher probability of belonging to the positive (or malignant) class.

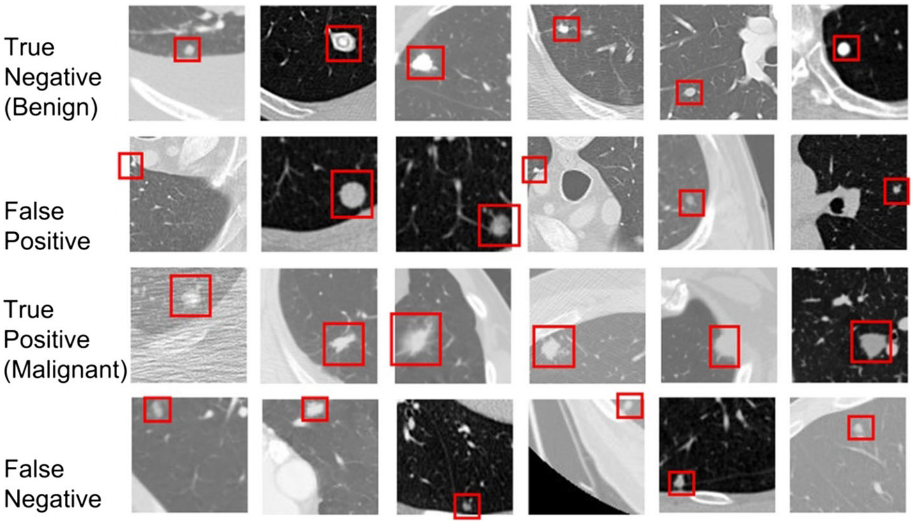

Fig. 8 Examples for the lung nodule classification results. The first and the third rows are examples of correctly classified and the second and fourth rows are misclassified

The nodules in the second row, which are labeled benign, are classified as malignant by mistake. Except for the nodules in the second and third columns, the other nodules in this row are very small. The last row is for the FN samples. The common feature of these six images is that their nodule sizes are all small, which is easily misidentified as benign nodules.

As seen in Fig. 8, the size of the nodule appears to be the prime reason for misclassification. To validate this observation, the average sizes of the bounding rectangle of the four classified categories are measured. The average size of the bounding rectangle of TN, FP, TP, and FN is 10.9, 11.3, 22.3, and 14.8. The difference in the average size of the bounding rectangle between TN and FP is not very obvious, yet there is a trend in that the larger size benign nodules have a greater probability of being classified as malignant and the small-size malignant nodules are more likely to be classified as benign.

Limitations and future work

One of the limitations of this work is that the three channels of input are homogeneous. Both LeNet and AlexNet are designed for color images, while our medical images are gray scale images, which results in an inability to make full use of all channels.

Another limitation is that the current CNN classifier utilizes the independent 2D patch as the input.

The misclassified patches shown in Fig. 8 are also difficult for a radiologist to diagnose, because the candidate nodules are diagnosed based on the information in the front and back slices. As a result, 2.5D or 3D input will be used in the future.

Moreover, the original sample patients in the LIDC dataset total to only 1018. Compared with the natural images in ImageNet, it has too small a number to be calculated by deep learning. Further research, such as transfer learning and fine-tuning, is needed.

Finally, performing diagnosis based on only medical images has its own limitations. Usually, doctors reach a conclusion based on many types of medical information. It is an impossible mission to affirm whether the nodule is benign or malignant solely based on medical images.

In addition to the methods mentioned above to conquer the drawbacks, more innovative studies will be performed for further research. First, we prefer to utilize the multi-modality strategy, which combines the general CT scan with the contrast-enhanced CT scan to determine the malignance of a nodule. The contrast-enhanced images usually contain more information about the vessel distribution and can distinguish tissues from lung effusion. These ample input data contribute to a more accurate result. Second, instead of using the CNN, many other supervised deep learning methods can be utilized to classify a nodule or tumor, such as deep reinforcement learning, generative adversarial nets. These deep learning models can be used to extract the features of medical images, which are significant for image analysis. Third, features that are gathered from different training models can be fused as input data to SVMs or other classifiers. Finally, if data acquisition is available, then the medical imaging information will combine genomic knowledge to contribute to better diagnosis.

Conclusions

In this paper, we constructed one new Agile CNN for pulmonary nodule classification using CT images, on which we investigated the effects of kernel size, learning rate, training batch size, dropout, and weight initialization on the accuracy and loss of the proposed CNN model. This Agile structure achieves a relatively high accuracy of 0.822 and an AUC of 0.877, and it is less prone to overfitting, which is associated with other redundant CNN structures. The results have shown that the proposed CNN framework and the optimization strategy for the CNN parameters might be suitable for pulmonary nodule classification when characterized by small medical datasets and small targets. The classification model might help to diagnose and treat pulmonary nodules effectively, and the strategy of optimizing the CNN parameters should be referential for other medical applications that could use CNNs. Further research will explore the three-dimensional input data, transfer learning and fine-tuning.

Acknowledgements The authors acknowledge the National Cancer Institute and the Foundation for the National Institutes of Health and their critical role in the creation of the free publicly available LIDC-IDRI Database used in this study.

Funding This study was funded by the National Natural Science Foundation of China under Grant (Nos. 81671773, 61672146).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval All procedures performed in these studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. For this type of study, formal consent is not required.

References

1. Siegel R, Naishadham D, Jemal A (2013) Cancer statistics. CA-Cancer J Clin 63(1):11–30. https://doi.org/10.3322/caac.21166
2. Chen D, Zheng R, Peter D, Baade PD, Zhang S, Zeng H, Bray F, Jemal A, Yu X, He J (2015) Cancer statistics in China. CA-Cancer J Clin 66(2):115–132. https://doi.org/10.3322/caac.21338
3. Valente IR, Cortez PC, Neto EC, Soares JM, De Albuquerque VH, Tavares JM (2016) Automatic 3D pulmonary nodule detection in CT images: a survey. Comput Methods Prog Biomed 124(C):91–107. https://doi.org/10.1016/j.cmpb.2015.10.006
4. Gridelli C, Rossi A, Carbone DP, Guarize J, Karachaliou N, Mok T, Petrella F, Spaggiari L, Rosell R (2015) Non-small-cell lung cancer. Nat Rev Dis Primers 2(T3):N1. https://doi.org/10.1038/nrdp.2015.9
5. Elbaz A, Beache GM, Gimelfarb G, Suzuki K, Okada K, Elnakib A, Soliman A, Abdollahi B (2013) Computer-aided diagnosis systems for lung cancer: challenges and methodologies. Int J Biomed Imaging 1:942353–942353. https://doi.org/10.1155/2013/942353
6. Doi K (2007) Computer-aided diagnosis in medical imaging: historical review, current status and future potential. Comput Med Imag Gr 31(4–5):198–211. https://doi.org/10.1016/j.compmedimag.2007.02.002
7. Parmar C, Grossmann P, Bussink J, Lambin P, AertsH J (2015) Machine learning methods for quantitative radiomic biomarkers. Sci Rep 5:13087. https://doi.org/10.1038/srep13087
8. Gillies RJ, Kinahan PE, Hricak H (2015) Radiomics: images are more than pictures, they are data. Radiology 278(2):563. https://doi.org/10.1148/radiol.2015151169
9. Greenspan H, van Ginneken B, Summers RM (2016) Guest editorial deep learning in medical imaging: overview and future promise of an exciting new technique. IEEE Trans Med Imaging 35(5):1153–1159. https://doi.org/10.1109/TMI.2016.2553401
10. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444. https://doi.org/10.1038/nature14539
11. LeCun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proc IEEE 86(11):2278–2324. https://doi.org/10.1109/5.726791
12. Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet classification with deep convolutional neural networks. In: NIPS 2012, pp 1097–1105
13. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. Comput Sci. arXiv:1409.1556
14. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. In: CVPR 2015, pp 1–9. https://doi.org/10.1109/CVPR.2015.7298594
15. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: CVPR, pp 770–778
16. Hua KL, Hsu CH, Hidayati SC, Cheng W, Chen Y (2015) Computer-aided classification of lung nodules on computed tomography images via deep learning technique. Onco Targets Ther 8:2015–2022. https://doi.org/10.2147/OTT.S80733
17. Sun W, Zheng B, Qian W (2016) Computer aided lung cancer diagnosis with deep learning algorithms. In: SPIE Medical Imaging 9785 2016:97850Z-97850Z-8. https://doi.org/10.1117/12.2216307
18. Cheng JZ, Ni D, Chou YH, Qin J, Tiu CM, Chang YC, Huang CS, Chen CM (2016) Computer-aided diagnosis with deep learning architecture: applications to breast lesions in us images and pulmonary nodules in CT scans. Sci Rep 6:24454. https://doi.org/10.1038/srep24454
19. Kumar D, Wong A, Clausi DA (2015) Lung nodule classification using deep features in CT images. Comput Robot Vis 2015:133–138. https://doi.org/10.1109/CRV.2015.25
20. Shen W, Zhou M, Yang F, Yang C, Tian J (2015) Multi-scale convolutional neural networks for lung nodule classification. IPIM 2015:588–599. https://doi.org/10.1007/978-3-319-19992-4_46
21. Shen W, Zhou M, Yang F, Yu D, Dong D, Yang C, Zang Y, Tian J (2017) Multi-crop convolutional neural networks for lung nodule malignancy suspiciousness classification. Pattern Recognit 61:663–673. https://doi.org/10.1016/j.patcog.2016.05.029
22. Armato SG, McLennan G, Bidaut L, McNitt-Gray MF, Meyer CR, Reeves AP (2015) Data from LIDC-IDRI. The Cancer Imaging Archive. https://doi.org/10.7937/K9/TCIA.2015.LO9QL9SX
23. Armato SG, McLennan G, Bidaut L, McNitt-Gray MF, Meyer CR, Reeves AP, Zhao B, Aberle DR, Henschke CI, Hoffman EA, Kazerooni EA, MacMahon H, Van Beeke EJ, Yankelevitz D, Biancardi AM, Bland PH, Brown MS, Engelmann RM, Laderach GE, Max D, Pais RC, Qing DP, Roberts RY, Smith AR, Starkey A, Batrah P, Caligiuri P, Farooqi A, Gladish GW, Jude CM, Munden RF, Petkovska I, Quint LE, Schwartz LH, Sundaram B, Dodd LE, Fenimore C, Gur D, Petrick N, Freymann J, Kirby J, Hughes B, Casteele AV, Gupte S, Sallamm M, Heath MD, Kuhn MH, Dharaiya E, Burns R, Fryd DS, Salganicoff M, Anand V, Shreter U, Vastagh S, Croft BY (2011) The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. Med Phys 38(2):915–931. https://doi.org/10.1118/1.3528204
24. Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, Moore SM, Phillips S, Maffitt DR, Tarbox L, Prior F (2013) The cancer

imaging archive (TCIA): maintaining and operating a public information repository. J Digit Imaging 26(6):1045–1057. https://doi.org/10.1007/s10278-013-9622-7

25. Sun W, Tseng TLB, Zhang J, Qian W (2016) Enhancing deep convolutional neural network scheme for breast cancer diagnosis with unlabeled data. Comput Med Imaging Gr 57:4–9. https://doi.org/10.1016/j.compmedimag.2016.07.004
26. Arevalo J, González FA, Ramos-Pollán R, Oliveira JL, Lopez MAG (2016) Representation learning for mammography mass lesion classification with convolutional neural networks. Comput Methods Programs Biomed 127:248–257. https://doi.org/10.1016/j.cmpb.2015.12.014
27. Nair V, Hinton GE (2010) Rectified linear units improve restricted boltzmann machines (ICML-10), pp 807–814
28. Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, Guadarrama S, Darrell T (2014) Caffe: convolutional architecture for fast feature embedding. ACMMM 2014:675–678. https://doi.org/10.1145/2647868.2654889
29. Srivastava N (2013) Improving neural networks with dropout. University of Toronto. http://www.cs.toronto.edu/~nitish/msc_thesis.pdf. Accessed 18 Feb 2013
30. Keskar NS, Mudigere D, Nocedal J, Smelyanskiy M, Tang PTP (2016) On large-batch training for deep learning: generalization gap and sharp minima. arXiv:1609.04836
31. Setio AAA, Ciompi F, Litjens G, Gerke PK, Jacobs C, van Riel S, Wille MMW, Naqibullah M, Sanchez CI, van Ginneken B (2016) Pulmonary nodule detection in CT images: false positive reduction using multi-view convolutional networks. IEEE Trans Med Imaging 35(5):1160–1169. https://doi.org/10.1109/TMI.2016.2536809
32. Shin HC, Roth HR, Gao M, Lu L, Xu Z, Nogues I, Yao J, Mollura D, Summers RM (2016) Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans Med Imaging 35(5):1285–1298. https://doi.org/10.1109/TMI.2016.2528162
33. Hinton GE, Srivastava N, Krizhevsky A, Sutskever I, Salakhutdinov R (2012) Improving neural networks by preventing co-adaptation of feature detectors. arXiv:1207.0580