Original Article

An efficient deep convolutional neural network model for visual localization and automatic diagnosis of thyroid nodules on ultrasound images

Jialin Zhu[1], Sheng Zhang[1], Ruiguo Yu[2], Zhiqiang Liu[2], Hongyan Gao[3], Bing Yue[1], Xun Liu[4], Xiangqian Zheng[5], Ming Gao[5], Xi Wei[1]

[1]Department of Diagnostic and Therapeutic Ultrasonography, Tianjin Medical University Cancer Institute and Hospital, National Clinical Research Center of Cancer, Key Laboratory of Cancer Prevention and Therapy, Tianjin's Clinical Research Center for Cancer, Tianjin, China; [2]College of Intelligence and Computing, Tianjin University, Tianjin Key Laboratory of Cognitive Computing and Application, Tianjin Key Laboratory of Advanced Networking, Tianjin, China; [3]Tianjin Xiqing District Women and Children's Health and Family Planning Service Center, Tianjin, China; [4]Department of Ultrasonography, the Fifth Central Hospital of Tianjin, Tianjin, China; [5]Department of Thyroid and Neck Tumor, Tianjin Medical University Cancer Institute and Hospital, National Clinical Research Center of Cancer, Key Laboratory of Cancer Prevention and Therapy, Tianjin's Clinical Research Center for Cancer, Tianjin, China

Correspondence to: Xi Wei, MD, PhD, Professor. Department of Diagnostic and Therapeutic Ultrasonography, Tianjin Medical University Cancer Institute and Hospital, National Clinical Research Center of Cancer, Key Laboratory of Cancer Prevention and Therapy, Tianjin's Clinical Research Center for Cancer, Huan Hu West Road, Tianjin 300060, China. Email: weixi@tmu.edu.cn; Ming Gao, MD, PhD, Professor. Department of Thyroid and Neck Tumor, Tianjin Medical University Cancer Institute and Hospital, National Clinical Research Center of Cancer, Key Laboratory of Cancer Prevention and Therapy, Tianjin's Clinical Research Center for Cancer, Huan Hu West Road, Tianjin 300060, China. Email: gaoming68@aliyun.com; gaomingheadneck@163.com.

Background: The aim of this study was to construct a deep convolutional neural network (CNN) model for localization and diagnosis of thyroid nodules on ultrasound and evaluate its diagnostic performance.

Methods: We developed and trained a deep CNN model called the Brief Efficient Thyroid Network (BETNET) using 16,401 ultrasound images. According to the parameters of the model, we developed a computer-aided diagnosis (CAD) system to localize and differentiate thyroid nodules. The validation dataset (1,000 images) was used to compare the diagnostic performance of the model using three state-of-the-art algorithms. We used an internal test set (300 images) to evaluate the BETNET model by comparing it with diagnoses from five radiologists with varying degrees of experience in thyroid nodule diagnosis. Lastly, we demonstrated the general applicability of our artificial intelligence (AI) system for diagnosing thyroid cancer in an external test set (1,032 images).

Results: The BETNET model accurately detected thyroid nodules in visualization experiments. The model demonstrated higher values for area under the receiver operating characteristic (AUC-ROC) curve [0.983, 95% confidence interval (CI): 0.973–0.990], sensitivity (99.19%), accuracy (98.30%), and Youden index (0.9663) than the three state-of-the-art algorithms ($P<0.05$). In the internal test dataset, the diagnostic accuracy of the BETNET model was 91.33%, which was markedly higher than the accuracy of one experienced (85.67%) and two less experienced radiologists (77.67% and 69.33%). The area under the ROC curve of the BETNET model (0.951) was similar to that of the two highly skilled radiologists (0.940 and 0.953) and significantly higher than that of one experienced and two less experienced radiologists ($P<0.01$). The kappa coefficient of the BETNET model and the pathology results showed good agreement (0.769). In addition, the BETNET model achieved an excellent diagnostic performance (AUC =0.970, 95% CI: 0.958–0.980) when applied to ultrasound images from another independent hospital.

Conclusions: We developed a deep learning model which could accurately locate and automatically diagnose thyroid nodules on ultrasound images. The BETNET model exhibited better diagnostic performance than three state-of-the-art algorithms, which in turn performed similarly in diagnosis as the

experienced radiologists. The BETNET model has the potential to be applied to ultrasound images from other hospitals.

Keywords: Thyroid nodule; artificial intelligence (AI); deep convolutional neural network (deep CNN); localization; ultrasound diagnosis

Submitted Apr 06, 2020. Accepted for publication Oct 20, 2020.
doi: 10.21037/qims-20-538
View this article at: http://dx.doi.org/10.21037/qims-20-538

Introduction

According to the American Cancer Society and Cancer Statistics Center, more than 500,000 new cases are diagnosed with thyroid cancer every year (567,233 in 2018), accounting for the highest incidence of all endocrine tumors (1). The global incidence of thyroid cancer is increasing yearly and has grown faster than that of other malignant tumors in recent years (2).

In clinical practice, ultrasound is the main examination used for both the screening and diagnosis of thyroid diseases. Ultrasound diagnosis of thyroid carcinoma is mainly based on the thyroid imaging report and data system (TI-RADS) (3-6). Previous research has shown that the accuracy of TI-RADS ranges from 29.0% to 84.0% (7,8). In addition, diagnostic accuracy varies greatly according to the level of experience of the ultrasound technician. Indeed, the main limitation of ultrasound is that it relies on the technician's expertise, and less experienced radiologists are more likely to misdiagnose a cancer, which increases the need for a greater number of fine needle aspiration biopsies (FNAB) (9,10). Even experienced radiologists have difficulty eliminating subjective opinion from their diagnoses. Therefore, there is an urgent need for an accurate, objective, efficient, and stable diagnostic method for classifying thyroid nodules.

More recently, there has been a growing interest in the automatic classification of thyroid nodules and considerable research progress has been made (11-13). In the traditional machine learning method, characteristic extraction is used for the automatic classification of the thyroid nodule ultrasound images. Using this method, researchers must manually extract the image features, including the shape, margin, and composition of the nodule, and then input these into the model for classification (14,15). However, manual extraction not only requires professional expertise but is also costly, and, in the case of large data sets, it significantly increases the burden of work for physicians.

Recently, the convolutional neural network (CNN) algorithm has achieved remarkable success in analyzing radiological, pathological, or clinical imaging classification tasks, including grading of diabetic retinopathy, assessment of skin lesions, and surveillance for acute neurologic events (16-18). However, most of the current CNNs are limited to classification, and do not visually localize lesions (16-20). In this study, we constructed a deep CNN model, named the Brief Efficient Thyroid Network (BETNET), for localization and classification of thyroid nodules. We aimed to verify whether this model could automatically locate and classify thyroid nodules, and whether it could achieve the same high level of diagnostic accuracy as that of experienced radiologists.

Methods

Patients and datasets

All ultrasound images included in the training, validation, and internal test sets were obtained from the thyroid imaging database at Tianjin Medical University Cancer Institute and Hospital, China. Images included in the external test set were obtained from Peking University BinHai Hospital, China. Consecutive patients in these two medical centers who underwent diagnostic thyroid ultrasound examination and subsequent surgery in the same hospital were included in the study. Inclusion criteria were as follows: (I) the presence of thyroid nodules without any previous local therapy; (II) an ultrasound image scan performed within 1 month before surgery; and (III) the type of thyroid nodule confirmed by histologic examination. Exclusion criteria were as follows: (I) images from anatomical sites that were judged non-tumorous according to postoperative pathology; (II) nodules with incomplete (one or both orthogonal plane images missing) or unclear ultrasound images; (III) cases with incomplete clinicopathological information; (IV) nodules that had received previous local therapy before image acquisition.

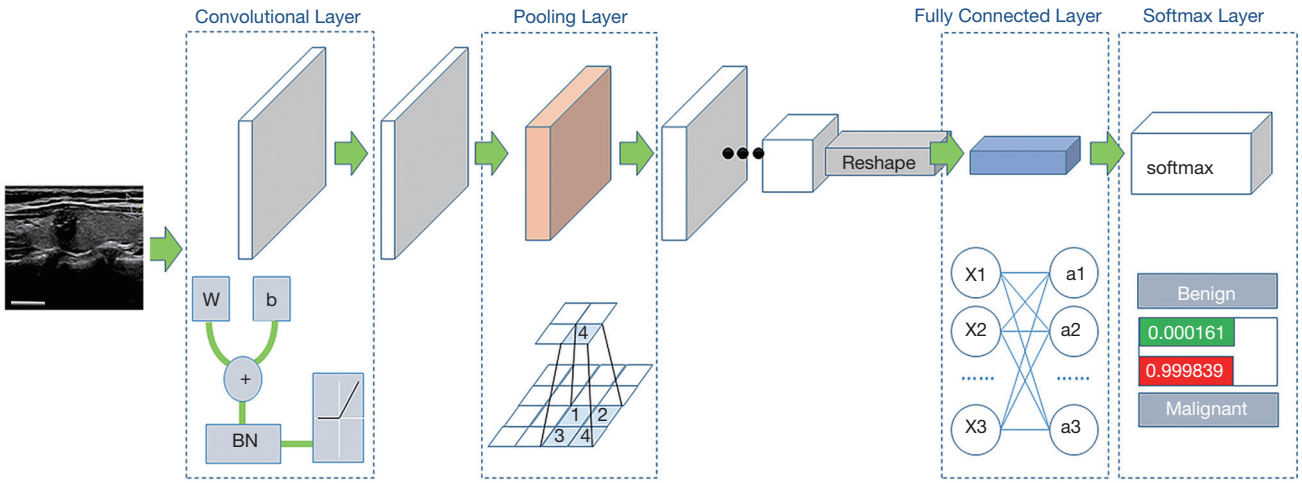

Figure 1 Framework of the BETNET model. In the convolution layer: W = weight, b = bias, and BN = batch normalization. The maximum pool operation was used in the pool layer. Reshape is a process that changes the feature map with the size from the previous layer and is then input into the fully connected layer. Softmax mapped the output of the fully connected layer to the results of the classification. The white scale bars represent 1 cm on the ultrasound image.

From January 2015 to June 2017, 16,401 ultrasound images from 5,895 patients were obtained for the training set. From August 2017 to December 2017, 1,000 ultrasound images from 414 patients were obtained for the validation set. The internal test set was composed of 300 images from 117 patients consecutively examined in January 2018 at Tianjin Medical University Cancer Institute and Hospital. The external test set included 1,032 images from 261 patients who underwent imaging and surgery between January 2015 and April 2017 at Peking University BinHai Hospital. All patients received surgical treatment, and all images were pathologically confirmed postoperatively. All ultrasound images and pathological examination reports were deidentified before being transferred to the investigators.

This study was evaluated and approved by the Ethics Committee of the Tianjin Medical University Cancer Institute and Hospital and conformed to the provisions of the Declaration of Helsinki. As this retrospective study was deemed to carry minimal risk, the requirement for informed patient consent was waived.

Ultrasound image acquisition

All ultrasound examinations were performed with the Phillips EPIQ 5, IU 22, HD11, (Philips Healthcare, Eindhoven, The Netherlands), GE Logiq 9 (GE Healthcare, Milwaukee, WI, USA), and Aplio 500 (Toshiba Medical Systems, Tokyo, Japan) devices equipped with either a 5–12 MHz or a 4.8–11 MHz linear array probe. Image quality control was performed for the four datasets by authors HYG and XQZ. The criteria for the ultrasound image selection were as follows: (I) a maximum section of the thyroid nodule containing the whole nodule; (II) the vertical section of the maximum section of the lesion; (III) the maximum section of the nodule with measurement marks including the maximum diameter and vertical diameter lines. Images were removed if the anatomical sites were non-cancerous according to the pathological report. Images not containing the thyroid nodules were deleted, such as images showing only blood vessels, adipose tissue, muscle tissue, or lymph nodes. Ultrasound images acquired with color Doppler flow imaging were deleted.

Deep CNN

In this study, our BETNET model was designed as a classification model of thyroid nodules based on the Visual Geometry Group-19 (VGG-19) model (21,22). VGG-19 was designed to process the multi-classification of natural images, so we considered it an appropriate model for fine tuning the two-category classification of ultrasound images in this paper. The framework and the structure of the BETNET model are shown in *Figure 1*, *Table 1*, and Tables S1,S2.

The surrounding black areas were removed from the ultrasound images (23), and the images were resized to

Table 1 Structure of the BETNET model

Layer name	Output size	Activation	Layer design
conv_block1_1	112×112×64	ReLU	3×3 conv, 64
conv_block1_2	112×112×64	ReLU	3×3 conv, 64
Pool_1	56×56×64		
conv_block2_1	56×56×128	ReLU	3×3 conv, 128
conv_block2_2	56×56×128	ReLU	3×3 conv, 128
Pool_2	28×28×128		
conv_block3_1	28×28×256	ReLU	3×3 conv, 256
conv_block3_2	28×28×256	ReLU	3×3 conv, 256
conv_block3_3	28×28×256	ReLU	3×3 conv, 256
conv_block3_4	28×28×256	ReLU	3×3 conv, 256
Pool_3	14×14×256		
conv_block4_1	14×14×512	ReLU	3×3 conv, 512
conv_block4_2	14×14×512	ReLU	3×3 conv, 512
conv_block4_3	14×14×512	ReLU	3×3 conv, 512
conv_block4_4	14×14×512	ReLU	3×3 conv, 512
Pool_4	7×7×512		
conv_block5_1	7×7×512	ReLU	3×3 conv, 512
conv_block5_2	7×7×512	ReLU	3×3 conv, 512
conv_block5_3	7×7×512	ReLU	3×3 conv, 512
conv_block5_4	7×7×512	ReLU	3×3 conv, 512
Reshape	25,088		
FC_1	4,096	ReLU	250,888×4,096
FC_2	4,096	ReLU	4,096×4,096
FC_3	2	ReLU	4,096×2
Softmax	2		

conv_block, convolution block; ReLU, rectified linear Unit; FC, fully connected.

112×112 pixels by the bilinear interpolation method before being put into the BETNET model. The red, green, and blue (RGB) images, 112×112 pixels in size, were input into the BETNET model which was composed of 16 convolutional layers, 3 fully connected layers, and the softmax layer. The working process of the model was divided into five stages. The first two stages comprised two convolutional layers and a pooling layer, the third and fourth stages comprised four convolutional layers and a pooling layer, and the last stage comprised four convolutional layers (*Table 1*, Tables S1,S2).

As shown in *Figure 1*, in the convolution layer the following abbreviations were used: W = weight, b = bias, BN = batch normalization. Each convolution operator was followed by BN, and then activated by a rectified linear unit (ReLU). "Convolution-BN-ReLU" constitutes a generalized convolution unit. The maximum pool operation was used in the pooling layer. The diagram below in the pooling layer shows the maximum pool. Reshape refers to the process that changed the feature map with the size of Width × Height × Channel into a one-dimensional vector of 60*1 from the previous layer and was then input into

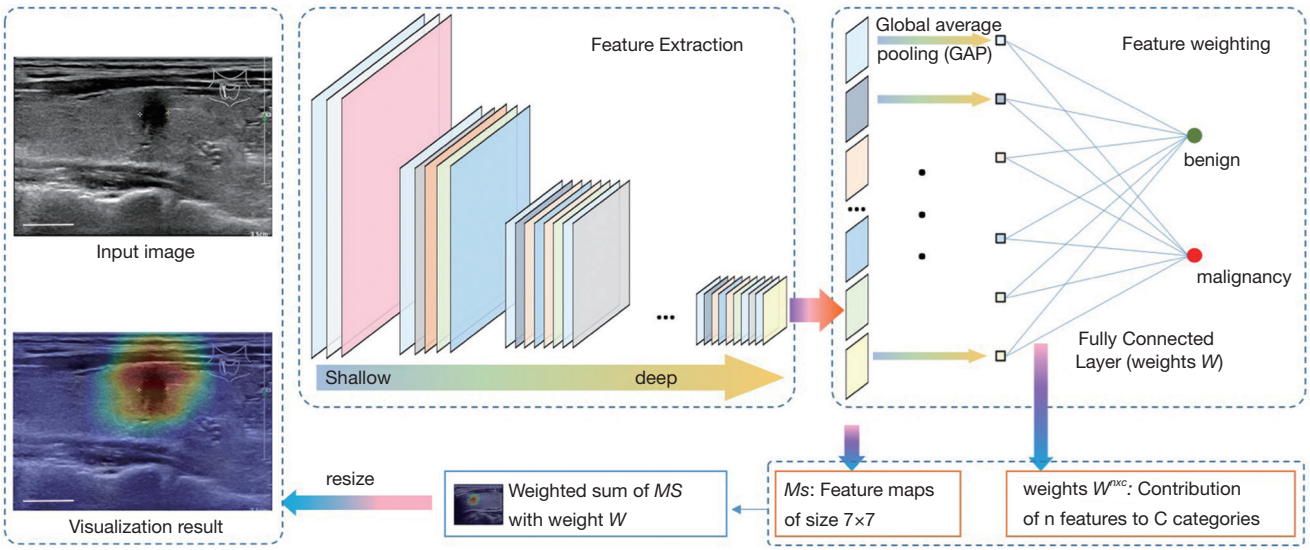

Figure 2 Schematic diagram of the visualization method. Features of the ultrasound images were extracted by multiple operations such as convolution and pooling, and then the deepest abstract features were combined to obtain classification results. Next, the classification result was obtained by nonlinearly transforming the weighted sum of the pixel values in the last feature map. The contribution of each pixel formed a matrix, and the visualized image (heat map) was generated by performing color mapping on the matrix. The white scale bars represent 1 cm on the ultrasound images.

the fully connected layer. Each node in the fully connected layer was connected to all the nodes in the previous layer, which was used to synthesize the features extracted from the previous layer. Softmax mapped the output of the full connected layer to the results of the classification. The probability value of the image belonged to each category.

Training of the model

We trained the BETNET model with 1,000 thyroid nodule images and used the validation set to evaluate its diagnostic performance. The BETNET model was then trained by increasing the number of training set images, and the model's diagnostic performance was constantly tested by the validation set.

Visualization experiments

The convolutional network used for classification followed this general process: first, image features (feature maps) were extracted by multiple operations such as convolution and pooling, and then the deepest abstract features were combined to obtain the classification result. More specifically, the classification result was obtained by nonlinearly transforming the weighted sum of the pixel values in the last feature map. Therefore, for an image entered into the model, each pixel's contribution to the result could be calculated by inverse reasoning according to the model's parameters. The contribution of each pixel formed a matrix, and the visualized image (heat map) was generated by performing color mapping on the matrix. The warm tone region of the visualization results was the most concerned part that the neural network could recognize. On the contrary, the cold tone region was the least important feature when the neural network classifying benign and malignant (*Figure 2*).

The diagnostic performance evaluation of the model

First, we compared the diagnostic performance of the BETNET model with three state-of-the-art machine learning algorithms [SE_Net (24), SE_inception_v4 (25), and Xception (26)], which were more advanced than the currently and widely used deep learning models such as ResNet (27) and DenseNet (28). Next, we used the test set to estimate the BETNET model, and five radiologists with differing experience levels in thyroid nodule diagnosis independently distinguished the benign and malignant images in the test set. Among them, three radiologists (Doctors A, B, and C: authors XW, SZ, and JLZ, respectively) had more than 7 years' experience in the

ultrasound diagnosis of thyroid nodules, and two doctors (Doctors D and E: authors BY and XL, respectively) had fewer than 2 years' diagnostic experience and worked in other hospitals. All doctors were blind to information regarding the patients' clinical history, previous classification results, or previous biopsy results. We compared the diagnostic performance differences between the radiologists and the BETNET model.

General applicability test

In this section, we aimed to investigate the general applicability of our artificial intelligence (AI) system for diagnosing thyroid cancer. We demonstrated this by testing the BETNET model on a data set of ultrasound images (n=1,032) from Peking University BinHai Hospital, which included 502 benign nodule images and 530 malignant nodule images.

Data and statistical analysis

The data are presented as means and standard deviations for continuous variables, and as the number of patients and images for categorical variables. The diagnostic performance of the computer-aided design (CAD) system and the radiologists was evaluated by analyzing sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and accuracy. The differences in diagnostic performance between the BETNET model and the radiologists were compared using the Z test and McNemar's test. We used the kappa (κ) coefficient to measure the agreement between the BETNET model prediction, the five radiologists' diagnoses, and the pathological results. We categorized the κ coefficient as follows: poor (0–0.20), fair (0.20–0.40), moderate (0.40–0.60), good (0.60–0.80), and excellent (0.80–1.00). The area under the receiver operating characteristic (ROC) curve with a 95% confidence interval (CI) was calculated to compare the differences in diagnosing thyroid cancer between the BETNET model and the radiologists.

All statistical analyses were performed using SPSS version 23.0 for Windows (SPSS, Chicago, IL, USA) and MedCalc version 15.0 for Windows (MedCalc Software, Ostend, Belgium). A significant difference was defined as a P value <0.05.

Results

Demographic features of the patients in four data sets

A total of 18,733 ultrasound images from 6,687 patients were used in this research. Of these, 16,401 images from 5,895 patients were obtained for the training set, which contained 6,760 (41.22%) benign nodule images and 9,641 (58.78%) malignant nodule images. The benign nodules included nodular goiter, adenomatous goiter, thyroid granuloma, and follicular adenoma. The malignant nodules contained papillary thyroid carcinoma, medullary thyroid carcinoma, and follicular thyroid carcinoma according to the pathological results. For the validation set, 1,000 ultrasound images from 414 patients were obtained, 500 of which were benign. The internal test set comprised 300 images from 117 patients and consisted of 73 (24.33%) benign nodule images and 227 (75.67%) malignant nodule images. The mean age of patients in the training, validation, and internal test sets was 44.22±12.49, 46.30±12.08, and 44.49±9.99 years, respectively.

The external test set comprised 1,032 images from 261 patients and consisted of 502 (48.64%) benign nodule images and 530 (51.36%) malignant nodule images. The benign nodules included nodular goiter, adenomatous goiter, and follicular adenoma, and the malignant nodules contained papillary thyroid carcinoma, lymphoma, and medullary thyroid carcinoma according to the pathological results of Peking University BinHai Hospital. The mean age of patients in the external test sets was 52.01±11.83 years. Demographic data for the four data sets are shown in *Table 2*.

Visual localization and diagnostic performance of the BETNET model

The model successfully focused on the thyroid nodule areas in the ultrasound images. As seen in *Figure 3*, the orange box area denotes the thyroid nodule location, and the heat map represents the visual location result. We also investigated the effect of the number of images in the training set on the algorithm performance using the validation set. By increasing the number of images in the training set, the area under the ROC curve (AUC) of the BETNET model was improved from 0.765 (95% CI: 0.737–0.771) to 0.978 (95% CI: 0.967–0.986). Accuracy also improved from 76.5% to 97.8%, and sensitivity and specificity increased in waves. Although there were several up and downs of accuracy, sensitivity, specificity, and AUC in training, the general trend was upward with increasing image numbers. The details are shown in *Table 3*.

The comparison of the BETNET model with three advanced deep learning algorithms

The diagnostic performance of all four deep learning

Table 2 Demographic data for four data sets

Parameter	Training set	Validation set	Internal test set	External test set
Patients	5,895	414	117	261
Male, n (%)	1,509 (25.60%)	113 (27.29%)	22 (18.80%)	64 (24.52%)
Female, n (%)	4,386 (74.40%)	301 (72.71%)	95 (81.20%)	197 (75.48%)
Age (years)	44.22±12.49	46.30±12.08	44.49±9.99	52.01±11.83
Total images	16,401	1,000	300	1,032
Benign images, n (%)	6,760 (41.22%)	500 (50%)	73 (24.33%)	502 (48.64%)
Malignant images, n (%)	9,641(58.78%)	500 (50%)	227 (75.67%)	530 (51.36%)

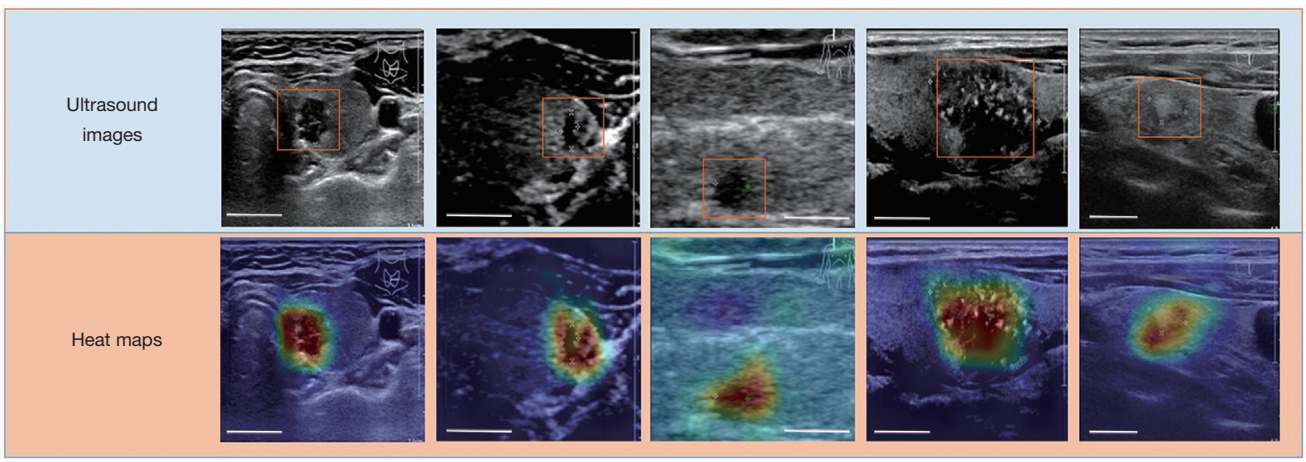

Figure 3 Localization of thyroid nodules. The orange box area is the location of the thyroid nodules, and the heat map represents the visualization results. Each column shows the same ultrasound image. The warm tone region of the visualization image is the most important part that the neural network could recognize. By contrast, the cold tone region is the least important feature. The white scale bars represent 1 cm on the ultrasound images.

algorithms is shown in *Table 4*. In the validation dataset, the BETNET model demonstrated the highest value for the AUC (0.983, 95% CI: 0.973–0.990), which was significantly higher than the other three models (P<0.05). Also, the BETNET model demonstrated the highest values for sensitivity (99.19%), accuracy (98.3%), and the Youden index (0.9663). However, the BETNET model demonstrated a lower value in specificity (97.45%) than the SE_NET (98.40%) and the SE_inception_v4 (98.00%).

The diagnostic performance of the BETNET model compared with the radiologists

In the internal test set, the diagnostic performance of the BETNET model and five radiologists with differing levels of experience is shown in *Table 5* and *Figure 4*. The BETNET model demonstrated a high level of accuracy in identifying thyroid cancer in the test set compared with the experienced radiologists. The accuracy of the BETNET model was 91.33%, which was markedly higher than one of the experienced radiologists (85.67%) and two of the less experienced radiologists (69.33–77.67%). The AUC of the BETNET model was 0.951 (95% CI: 0.920–0.972), which was similar to the more experienced doctors A (AUC =0.940, 95% CI: 0.906–0.964) and B (AUC =0.953, 95% CI: 0.923–0.974), and significantly higher than the more experienced doctor C (AUC =0.896, 95% CI: 0.856–0.928) and the two less experienced doctors D and E (AUC =0.833, 95% CI: 0.782–0.870; AUC =0.788, 95% CI: 0.737–0.833). When measuring the agreement between the BETNET model prediction and

Table 3 Number of images in the training set and diagnostic performance of the BETNET model

Images	Accuracy	Sensitivity	Specificity	AUC	95% CI	
2,000	0.765	0.878	0.652	0.765	0.737	0.771
4,000	0.827	0.828	0.826	0.827	0.809	0.842
6,000	0.911	0.890	0.932	0.911	0.892	0.928
8,000	0.955	0.952	0.958	0.955	0.940	0.967
10,000	0.949	0.944	0.954	0.949	0.933	0.962
12,000	0.971	0.972	0.970	0.971	0.959	0.980
14,000	0.965	0.968	0.962	0.965	0.952	0.976
16,401	0.978	0.980	0.976	0.978	0.967	0.986

AUC, area under the ROC curve; CI, confidence interval.

Table 4 Comparison of the diagnostic performance of the BETNET model with three machine-learning algorithms in the validation dataset

Parameters	BETNET	SE_Net	SE_inception_v4	Xception
AUC, 95% CI	0.983 (0.973–0.990)	0.963 (0.949–0.974)	0.971 (0.959–0.980)	0.964 (0.951–0.975)
Sensitivity (%)	99.19	94.20	96.20	97.80
Specificity (%)	97.45	98.40	98.00	95.00
Accuracy (%)	98.3	96.3	97.1	96.4
Youden index	0.9663	0.9276	0.9420	0.9287
P	–	0.0004*	0.0337*	0.0027*

AUCs of the BETNET model and the other three models were calculated by DeLong *et al.*'s. method. P: The difference of AUCs between the BETNET model and other three models was compared by Z-test; *, P<0.05. AUC, area under the ROC curve; CI, confidence interval.

Table 5 Diagnostic performance of the BETNET model and the doctors with different experience levels in the internal test set

Parameters	BETNET model	Experienced doctor A	Experienced doctor B	Experienced doctor C	Doctor D	Doctor E
Accuracy	91.33%	93.67%	94.33%	85.67%	77.67%	69.33%
Sensitivity	93.39%	93.39%	93.39%	81.94%	72.69%	60.35%
Specificity	84.93%	94.52%	97.26%	97.26%	93.15%	97.26%
AUC	0.951	0.940	0.953	0.896	0.833	0.788
95% CI	0.920–0.972	0.906–0.964	0.923–0.974	0.856–0.928	0.782–0.870	0.737–0.833
P1	–	0.5494	0.8676	0.0059*	0.000*	0.000*
P2	–	0.296	0.188	0.000*	0.000*	0.000*
κ value	0.769	0.836	0.855	0.670	0.521	0.409

AUCs of the BETNET model and doctors were calculated by DeLong *et al.'s* method. P1: the difference of AUCs between the BETNET model prediction and doctor was compared by Z-test; *, P<0.05. P2: Measures the agreement between the BETNET model prediction and doctors. McNemar's test was used for the statistical analysis; *, P<0.05. κ value: Measures the agreement between the BETNET model prediction, five doctors, and the pathological result. AUC, area under the ROC curve; CI, confidence interval.

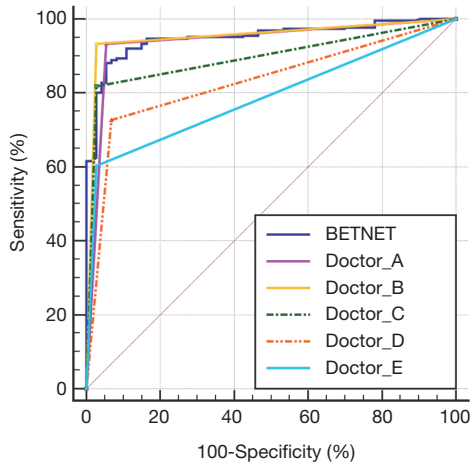

Figure 4 AUCs of the BETNET model and five doctors in the internal test set. The AUC of the BETNET model was 0.951 (95% CI: 0.920–0.972), which was similar to that of skilled doctors A (AUC =0.940, 95% CI: 0.906–0.964) and B (AUC =0.953, 95% CI: 0.923–0.974), and significantly higher than that of skilled doctor C (AUC =0.896, 95% CI: 0.856–0.928) and the two less experienced doctors (AUC =0.833, 95% CI: 0.782–0.870; AUC =0.788, 95% CI: 0.737–0.833).

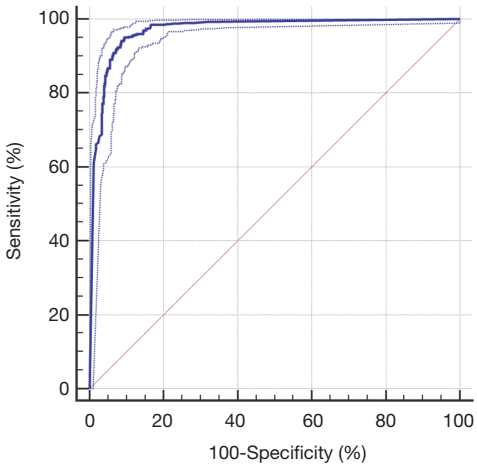

Figure 5 The performance of the BETNET model in the external test set. The BETNET model achieved an accuracy of 93.80%, with a sensitivity of 95.09%, a specificity of 90.445% and an AUC of 0.970 (95% CI: 0.958–0.980).

the five radiologists, performance of the BETNET model was consistent with experienced doctors A and B (P>0.05) and was superior to doctors C, D, and E (P≤0.001). The κ coefficient of the BETNET model and the pathology result was rated as good (κ=0.769) compared with a coefficient of 0.670–0.855 for the experienced radiologists and a coefficient of 0.409–0.521 for the less experienced doctors.

Generalizability of the BETNET model

To investigate the generalizability of our AI system to the diagnosis of thyroid cancer, we tested the ultrasound images from Peking University BinHai Hospital, which were not contained in the training set. In this test, the BETNET model achieved an accuracy of 93.80%, with a sensitivity of 95.09% and a specificity of 90.44% for differentiating between benign and malignant thyroid nodules. The ROC curve is shown in *Figure 5*. The AUC of the BETNET model for diagnosing thyroid cancer was 0.970 (95% CI: 0.958–0.980).

Discussion

In this study, we developed the BETNET model based on deep learning to accurately locate and automatically classify thyroid nodules on ultrasound. This model simultaneously predicts the label and constructs the attention heat map, thus highlighting the most important part of the image, which has several advantages: (I) It can locate thyroid nodules on ultrasound images, which is a further step towards neural network visualization and exploration of the "black box" phenomenon (29). (II) It can help identify the location of thyroid nodules and contribute to a fully automatic computer diagnosis. (III) The results of visual localization prove that our model classifies ultrasound images according to the characteristics of the nodules, rather than being affected by other non-nodular areas. The diagnostic accuracy is more credible compared with other networks with only a classification function. (IV) The accuracy of our model is higher than that of the three state-of-the-art algorithms. (V) The number of images required in training the model was relatively small, which is convenient for popularization and application.

Training the BETNET model was completed automatically by computer. There was no need to manually extract the characteristics of the image or the marked region of interest (ROI) from the ultrasound image used for training. Some researchers have applied CNN to the automatic classification of the thyroid nodules and achieved a relatively satisfying accuracy (30,31). However, most of the current deep learning methods are semi-automatic and require their data sets to be marked manually, insofar as the

radiologist needs to draw the margin of the nodule or ROI on the ultrasound image of the thyroid nodule (11,32). As such, the BETNET model has the potential to reduce the workload of busy professional physicians.

The BETNET model, despite its simplicity in structure, exhibited excellent diagnostic abilities in identifying thyroid cancer. It demonstrated the highest value for AUC, sensitivity, accuracy, and the Youden index in the validation dataset, compared with the three state-of-the-art deep learning models (SE_Net, SE_inception_v4, and Xception). These in turn are more advanced than deep learning models such as ResNet and DenseNet which are more commonly used in current practice. Furthermore, in the test set, the BETNET model's high level of accuracy and high AUC were extremely similar to those of experienced radiologists.

In previous studies, CNN models have been used to diagnose thyroid carcinoma, but most of these studies had small sample sizes or relatively low accuracy (33,34). Recently, Li *et al.* (19) structured a model for the diagnosis of thyroid cancer based on ResNet50 and Darknet19. However, the algorithm of this model was complicated and diagnostic accuracy was reported as 85.7–88.9%. In comparison, our algorithm was trained on a relatively small dataset but achieved better accuracy, which suggests a greater application prospect. Since deep learning performance was strongly correlated with the amount of available training data, it is feasible that a higher performing algorithm could be developed with a larger training dataset (29). As our results demonstrated, with an increase in the number of images added in the training set, the accuracy, sensitivity, specificity, and the AUC of the BETNET model were all improved. Therefore, we speculate that as the number of images in the training set continues to increase, the diagnostic performance of the BETNET model may outstrip that of skilled radiologists.

In this study, we also compared the BETNET model's diagnostic ability with that of two less experienced doctors working at other hospitals. We found that both doctors showed an inferior diagnostic ability compared with the BETNET model, with lower accuracy, sensitivity, specificity, and AUC. In ultrasound imaging, human evaluation is subjective and dependent on the individual doctor's experience (35). In contrast, the BETNET model provides consistent predictions for the same input, which could potentially eliminate the problem of interobserver variability. Furthermore, this model could provide radiological decision support to medical centers where an experienced radiologist is unavailable and thereby reduce the need for unnecessary FNAB and surgeries.

Our network represents a generalized platform that can be universally applied to ultrasound images from different medical centers. When we applied the BETNET model to ultrasound images from another hospital using different types of ultrasound equipment to ours, the BETNET model also achieved excellent accuracy, sensitivity, and specificity. This resulting high accuracy suggests that the BETNET model has the potential to effectively learn from different types of ultrasound images with a high degree of generalization. This could be of benefit for screening programs and could be conducive to wide dissemination across all medical fields, particularly in low-resource or remote areas, resulting in a wide range of beneficial clinical and public health outcomes.

There were some limitations to our study. The training set did not contain ultrasound images from other hospitals in China or other countries, which limited the access to images from a greater diversity of patients and different ultrasound equipment. Also, the sample size of the training set was not large enough in this study, especially of the benign nodules, which might have led to a sample bias. Further, the BETNET model was developed on specialized computer hardware suitable for retrospective analysis and therefore could not provide real-time diagnostic results. We are currently in the process of building internet applications and a website to provide free and real-time access to the developed CNN model. In further investigations, we will be conducting a multicenter, large-data, and prospectively designed study.

Conclusions

In this study, we developed a deep learning model for visual localization and automatic diagnosis of thyroid nodules on ultrasound images. Our model simultaneously predicts the label and constructs the attention heat map, which enables identification of the important parts of the image when the model predicts a thyroid nodule. The BETNET model exhibited a better diagnostic performance than three state-of-the-art algorithms, which in turn was similar to that of experienced radiologists. Furthermore, the BETNET model represents a generalized platform that has the potential to assist radiologists working across different medical centers.

Acknowledgments

Funding: This work was partially funded by the National

Natural Science Foundation of China grant (81771852) and the Tianjin Major Science and Technology Project of Artificial Intelligence (18ZXZNSY00300).

Footnote

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at http://dx.doi.org/10.21037/qims-20-538). Dr. XW reports grants from the National Natural Science Foundation of China, and grants from the Tianjin Municipal Science and Technology Bureau during the course of the study; Dr. RY reports grants from the Tianjin Municipal Science and Technology Bureau during the course of the study. The other authors have no conflicts of interest to declare.

Ethical Statement: This study was evaluated and approved by the Ethics Committee of the Tianjin Medical University Cancer Institute and Hospital (no. Ek2019030) and it conformed to the provisions of the Declaration of Helsinki. This retrospective study was deemed to carry minimal risk and therefore the requirement for informed patient consent was waived.

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2018;68:394-424.
2. Global Burden of Disease Cancer C, Fitzmaurice C, Akinyemiju TF, Al Lami FH, Alam T, Alizadeh-Navaei R, Allen C, Alsharif U, Alvis-Guzman N, Amini E, Anderson BO, Aremu O, Artaman A, Asgedom SW, Assadi R, Atey TM, Avila-Burgos L, Awasthi A, Ba Saleem HO, Barac A, Bennett JR, Bensenor IM, Bhakta N, Brenner H, Cahuana-Hurtado L, Castaneda-Orjuela CA, Catala-Lopez F, Choi JJ, Christopher DJ, Chung SC, Curado MP, Dandona L, Dandona R, das Neves J, Dey S, Dharmaratne SD, Doku DT, Driscoll TR, Dubey M, Ebrahimi H, Edessa D, El-Khatib Z, Endries AY, Fischer F, Force LM, Foreman KJ, Gebrehiwot SW, Gopalani SV, Grosso G, Gupta R, Gyawali B, Hamadeh RR, Hamidi S, Harvey J, Hassen HY, Hay RJ, Hay SI, Heibati B, Hiluf MK, Horita N, Hosgood HD, Ilesanmi OS, Innos K, Islami F, Jakovljevic MB, Johnson SC, Jonas JB, Kasaeian A, Kassa TD, Khader YS, Khan EA, Khan G, Khang YH, Khosravi MH, Khubchandani J, Kopec JA, Kumar GA, Kutz M, Lad DP, Lafranconi A, Lan Q, Legesse Y, Leigh J, Linn S, Lunevicius R, Majeed A, Malekzadeh R, Malta DC, Mantovani LG, McMahon BJ, Meier T, Melaku YA, Melku M, Memiah P, Mendoza W, Meretoja TJ, Mezgebe HB, Miller TR, Mohammed S, Mokdad AH, Moosazadeh M, Moraga P, Mousavi SM, Nangia V, Nguyen CT, Nong VM, Ogbo FA, Olagunju AT, Pa M, Park EK, Patel T, Pereira DM, Pishgar F, Postma MJ, Pourmalek F, Qorbani M, Rafay A, Rawaf S, Rawaf DL, Roshandel G, Safiri S, Salimzadeh H, Sanabria JR, Santric Milicevic MM, Sartorius B, Satpathy M, Sepanlou SG, Shackelford KA, Shaikh MA, Sharif-Alhoseini M, She J, Shin MJ, Shiue I, Shrime MG, Sinke AH, Sisay M, Sligar A, Sufiyan MB, Sykes BL, Tabares-Seisdedos R, Tessema GA, Topor-Madry R, Tran TT, Tran BX, Ukwaja KN, Vlassov VV, Vollset SE, Weiderpass E, Williams HC, Yimer NB, Yonemoto N, Younis MZ, Murray CJL, Naghavi M. Global, Regional, and National Cancer Incidence, Mortality, Years of Life Lost, Years Lived With Disability, and Disability-Adjusted Life-Years for 29 Cancer Groups, 1990 to 2016: A Systematic Analysis for the Global Burden of Disease Study. JAMA Oncol 2018;4:1553-68.
3. Horvath E, Majlis S, Rossi R, Franco C, Niedmann JP, Castro A, Dominguez M. An ultrasonogram reporting system for thyroid nodules stratifying cancer risk for clinical management. J Clin Endocrinol Metab 2009;94:1748-51.
4. Park JY, Lee HJ, Jang HW, Kim HK, Yi JH, Lee W, Kim SH. A proposal for a thyroid imaging reporting and data system for ultrasound features of thyroid carcinoma. Thyroid 2009;19:1257-64.
5. Haugen BR, Alexander EK, Bible KC, Doherty GM, Mandel SJ, Nikiforov YE, Pacini F, Randolph GW, Sawka AM, Schlumberger M, Schuff KG, Sherman SI, Sosa JA, Steward DL, Tuttle RM, Wartofsky L. 2015 American Thyroid Association Management Guidelines for Adult Patients with Thyroid Nodules and Differentiated Thyroid Cancer: The American Thyroid Association Guidelines

Task Force on Thyroid Nodules and Differentiated Thyroid Cancer. Thyroid 2016;26:1-133.
6. Tessler FN, Middleton WD, Grant EG, Hoang JK, Berland LL, Teefey SA, Cronan JJ, Beland MD, Desser TS, Frates MC, Hammers LW, Hamper UM, Langer JE, Reading CC, Scoutt LM, Stavros AT. ACR Thyroid Imaging, Reporting and Data System (TI-RADS): White Paper of the ACR TI-RADS Committee. J Am Coll Radiol 2017;14:587-95.
7. Yoon JH, Han K, Kim EK, Moon HJ, Kwak JY. Diagnosis and Management of Small Thyroid Nodules: A Comparative Study with Six Guidelines for Thyroid Nodules. Radiology 2017;283:560-9.
8. Hoang JK, Middleton WD, Farjat AE, Langer JE, Reading CC, Teefey SA, Abinanti N, Boschini FJ, Bronner AJ, Dahiya N, Hertzberg BS, Newman JR, Scanga D, Vogler RC, Tessler FN. Reduction in Thyroid Nodule Biopsies and Improved Accuracy with American College of Radiology Thyroid Imaging Reporting and Data System. Radiology 2018;287:185-93.
9. Choi SH, Kim EK, Kwak JY, Kim MJ, Son EJ. Interobserver and intraobserver variations in ultrasound assessment of thyroid nodules. Thyroid 2010;20:167-72.
10. Park CS, Kim SH, Jung SL, Kang BJ, Kim JY, Choi JJ, Sung MS, Yim HW, Jeong SH. Observer variability in the sonographic evaluation of thyroid nodules. J Clin Ultrasound 2010;38:287-93.
11. Choi YJ, Baek JH, Park HS, Shim WH, Kim TY, Shong YK, Lee JH. A Computer-Aided Diagnosis System Using Artificial Intelligence for the Diagnosis and Characterization of Thyroid Nodules on Ultrasound: Initial Clinical Assessment. Thyroid 2017;27:546-52.
12. Yoo YJ, Ha EJ, Cho YJ, Kim HL, Han M, Kang SY. Computer-Aided Diagnosis of Thyroid Nodules via Ultrasonography: Initial Clinical Experience. Korean J Radiol 2018;19:665-72.
13. Reverter JL, Vázquez F, Puig-Domingo M. Diagnostic Performance Evaluation of a Computer-Assisted Imaging Analysis System for Ultrasound Risk Stratification of Thyroid Nodules. AJR Am J Roentgenol 2019;213:1-6.
14. Yu Q, Jiang T, Zhou A, Zhang L, Zhang C, Xu P. Computer-aided diagnosis of malignant or benign thyroid nodes based on ultrasound images. Eur Arch Otorhinolaryngol 2017;274:2891-7.
15. Xia J, Chen H, Li Q, Zhou M, Chen L, Cai Z, Fang Y, Zhou H. Ultrasound-based differentiation of malignant and benign thyroid Nodules: An extreme learning machine approach. Comput Methods Programs Biomed 2017;147:37-49.
16. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, Venugopalan S, Widner K, Madams T, Cuadros J, Kim R, Raman R, Nelson PC, Mega JL, Webster DR. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. JAMA 2016;316:2402-10.
17. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542:115-8.
18. Titano JJ, Badgeley M, Schefflein J, Pain M, Su A, Cai M, Swinburne N, Zech J, Kim J, Bederson J, Mocco J, Drayer B, Lehar J, Cho S, Costa A, Oermann EK. Automated deep-neural-network surveillance of cranial images for acute neurologic events. Nat Med 2018;24:1337-41.
19. Li X, Zhang S, Zhang Q, Wei X, Pan Y, Zhao J, Xin XJ, Qin CX, Wang XQ, Li JX, Yang F, Zhao YH, Yang M, Wang QH, Zheng ZM, Zheng XQ, Yang XM, Whitlow CT, Gurcan MN, Zhang L, Wang XD, Pasche B, Gao M, Zhang W, Chen KX. Diagnosis of thyroid cancer using deep convolutional neural network models applied to sonographic images: a retrospective, multicohort, diagnostic study. Lancet Oncol 2019;20:193-201.
20. Liu C, Chen S, Yang Y, Shao D, Peng W, Wang Y, Chen Y, Wang Y. The value of the computer-aided diagnosis system for thyroid lesions based on computed tomography images. Quant Imaging Med Surg 2019;9:642-53.
21. Simonyan K and Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. 2014; CoRR arXiv:1409.1556.
22. Global Burden of Disease Cancer Collaboration, Kim HG, Lee KM, Kim EJ, Lee JS. Improvement diagnostic accuracy of sinusitis recognition in paranasal sinus X-ray using multiple deep learning models. Quant Imaging Med Surg 2019;9:942-51.
23. Ying X, Yu ZH, Yu RG, Li XW, Yu M, Zhao MK, Liu K. Thyroid Nodule Segmentation in Ultrasound Images Based on Cascaded Convolutional Neural Network. International Conference on Neural Information Processing. ICONIP 2018:373-84.
24. Hu J, Shen L, Albanie S, Sun G, Wu E. Squeeze-and-Excitation Networks. IEEE Trans Pattern Anal Mach Intell 2020;42:2011-23.
25. Szegedy C, Loffe S, Vanhoucke V, Alemi AA. Inception-v4, inception-resnet and the impact of residual connections on learning. Available online: https://arxiv.org/abs/1602.07261
26. Chollet F. Xception: Deep Learning with Depthwise

Separable Convolutions [Internet]. arXiv [cs.CV]. 2016. Available online: http://arxiv.org/abs/1610.02357. Accessed 9 Feb 2018.
27. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016:770-8.
28. Huang G, Liu ZL, Maaten VD, Weinberger KQ. Densely connected convolutional networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017:4700-8.
29. Walsh SLF, Calandriello L, Silva M, Sverzellati N. Deep learning for classifying fibrotic lung disease on high-resolution computed tomography: a case-cohort study. Lancet Respir Med 2018;6:837-45.
30. Ma J, Wu F, Jiang T, Zhu J, Kong D. Cascade convolutional neural networks for automatic detection of thyroid nodules in ultrasound images. Med Phys 2017;44:1678-91.
31. Ma J, Wu F, Zhu J, Xu D, Kong D. A pre-trained convolutional neural network based method for thyroid nodule diagnosis. Ultrasonics 2017;73:221-30.
32. Seo JK, Kim YJ, Kim KG, Shin I, Shin JH, Kwak JY. Differentiation of the Follicular Neoplasm on the Gray-Scale US by Image Selection Subsampling along with the Marginal Outline Using Convolutional Neural Network. Biomed Res Int 2017;2017:3098293.
33. Gao L, Liu R, Jiang Y, Song W, Wang Y, Liu J, Wang J, Wu D, Li S, Hao A, Zhang B. Computer-aided system for diagnosing thyroid nodules on ultrasound: A comparison with radiologist-based clinical assessments. Head Neck 2018;40:778-83.
34. Wang L, Yang S, Yang S, Zhao C, Tian G, Gao Y, Chen Y, Lu Y. Automatic thyroid nodule recognition and diagnosis in ultrasound imaging with the YOLOv2 neural network. World J Surg Oncol 2019;17:12.
35. Lee HJ, Yoon DY, Seo YL, Kim JH, Baek S, Lim KJ, Cho YK, Yun EJ. Intraobserver and Interobserver Variability in Ultrasound Measurements of Thyroid Nodules. J Ultrasound Med 2018;37:173-8.

Cite this article as: Zhu J, Zhang S, Yu R, Liu Z, Gao H, Yue B, Liu X, Zheng X, Gao M, Wei X. An efficient deep convolutional neural network model for visual localization and automatic diagnosis of thyroid nodules on ultrasound images. Quant Imaging Med Surg 2021;11(4):1368-1380. doi: 10.21037/qims-20-538

