future internet

MDPI

Article

Tooth-Marked Tongue Recognition Using Gradient-Weighted Class Activation Maps

Yue Sun [1], Songmin Dai [1], Jide Li [1], Yin Zhang [1] and Xiaoqiang Li [1,2,*]

[1] School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China; flyyue@shu.edu.cn (Y.S.); laodar@shu.edu.cn (S.D.); iavtvai@shu.edu.cn (J.L.); zhangyin6998@t.shu.edu.cn (Y.Z.)

[2] Shanghai Institute for Advanced Communication and Data Science, Shanghai University, Shanghai 200444, China

* Correspondence: xqli@shu.edu.cn

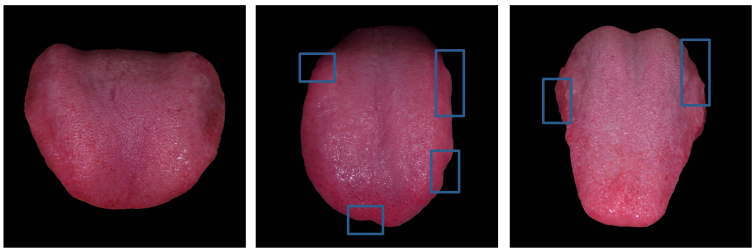
Received: 11 January 2019; Accepted: 13 February 2019; Published: 15 February 2019

Abstract: The tooth-marked tongue is an important indicator in traditional Chinese medicinal diagnosis. However, the clinical competence of tongue diagnosis is determined by the experience and knowledge of the practitioners. Due to the characteristics of different tongues, having many variations such as different colors and shapes, tooth-marked tongue recognition is challenging. Most existing methods focus on partial concave features and use specific threshold values to classify the tooth-marked tongue. They lose the overall tongue information and lack the ability to be generalized and interpretable. In this paper, we try to solve these problems by proposing a visual explanation method which takes the entire tongue image as an input and uses a convolutional neural network to extract features (instead of setting a fixed threshold artificially) then classifies the tongue and produces a coarse localization map highlighting tooth-marked regions using Gradient-weighted Class Activation Mapping. Experimental results demonstrate the effectiveness of the proposed method.

Keywords: tooth-marked tongue; convolutional neural network; gradient-weighted class activation maps

1. Introdution

Inspection of the tongue is one of the most important diagnostic methods in traditional Chinese medicine (TCM). According to [1], medical experts diagnose diseases by observing patient tongue color, tongue shape, and other characteristics of the tongue. Different features of the tongue reflect the internal state of the body and the health of the organs. Thus, tongue diagnosis has been widely applied to clinical analysis for thousands of years [2]. Tooth-marked tongue, a kind of abnormal tongue, is one appearance of the tongue when there are teeth marks along the lateral borders [3]. Medical experts believe that the tooth-marked tongue is caused by spleen deficiency, which provides guidance for clinical syndrome differentiation [4]. The appearances of the tooth-marked tongue are shown in Figure 1; (a) is a normal tongue image for reference. (b.1) and (b.2) are tooth-marked tongue images with teeth-marked regions shown in blue boxes. According to previous surveys, the incidence of tooth-marked tongue in the crowd is about 56%, of which the severe ones accounts for 11% [5]. However, the recognition of tooth-marked tongue is a challenging task for TCM practitioners. The appearance of tooth-marked tongues has a great number of variations, such as different colors, different shapes, and different types of teeth marks [6]. Therefore, clinical effectiveness of the diagnosis heavily depends on the TCM practitioner's experience. For this reason, more and more computer researchers have begun to combine image processing with pattern recognition technology to establish an objective and quantitative TCM recognition system [7,8].



(a) Healthy tongue (b.1) Tooth-marked tongue (b.2) Tooth-marked tongue

Figure 1. Examples of a heathy tongue and tooth-marked tongues. The tooth-marked regions in (b.1) are obvious, while the tooth-marked regions on (b.2) are difficult to identify.

The recognition of tooth-marked tongues can be viewed as a fine-grained classification problem, but it is more challenging than distinguishing between subcategories due to some specific difficulties in the field of tongue diagnosis. Firstly, the number of tongue images is limited because of personal privacy and image acquisition limitations. Secondly, a tongue image is labeled as a tooth-marked tongue or a nontooth-marked tongue, and the locations of the tooth-marked regions are not available. Moreover, existing approaches have a lack of decomposability into intuitive and understandable components, making tongue diagnoses hard to interpret. These questions lead us to seek help from Gradient-weight Class Activation Mapping (Grad-CAM). Grad-CAM was proposed by Selvaraju et al. [9] to provide visual explanations of the Convolutional Neural Network (CNN). It uses the gradients of any target concept, flowing into the final convolutional layer to produce a coarse localization map highlighting the important regions in the image for predicting the concept. It was shown that even if there is no location information when training a classification network, convolutional neural networks still have remarkable abilities for localizing objects. In this work, we adopted the Grad-CAM technique to help us analyze the tooth-marked tongue. We present a method that accurately classifies tooth-marked tongue and localizes the important regions in the image for predicting the pathology without bounding boxes. Through the visual interpretation of the tooth-mark problem, we also explore the effect of different receptive field sizes on the classification results. The experimental result shows that our method provides excellent interpretability while improving tongue recognition accuracy.

The remainder of this paper is organized as follows. Section 2 reviews the related work briefly. Section 3 describes the proposed method for tooth-marked tongue recognition in detail. Section 4 presents the detailed process of the method and results of experiments. Finally, this study is concluded in Section 5.

2. Related Work

2.1. Tongue Diagnosis

In the past few decades, some researchers have been contributing to the field of computerized tongue diagnosis, including tongue examination system establishment and tongue analysis. Chiu et al. [10] built a computerized tongue examination system for the purpose of quantizing the tongue properties in traditional Chinese medical diagnoses. Zhang et al. [11] established the relationship between tongue appearances and diseases using Bayesian network classifiers based on quantitative features. Many works have also proposed techniques for tongue segmentation [12], tongue image color analysis [13,14], and tongue shape analysis [15].

In the study of the tooth-marked tongue, the threshold of tongue concavity is an important indicator for classifying the tooth-marked tongue. Zhang [16] pointed out that the tooth-marked tongue is very common in tongue images. It is fatter than the normal tongue, the texture is more tender, and the color is paler. Li [17] proposed a method, based on specific thresholds, to extract features of tooth-marked tongues. Firstly, in order to find suspicious tooth-marked regions, he set a threshold for

the curvature change of the tongue edge. Secondly, he scanned the edge of the tongue image with a diamond-shaped box. Finally, the R-value of the box, which represents the color of the tongue image, was defined as a feature to classify the tooth-marked tongue. Wang et al. [18] calculated the slope and length of the tongue image, and used the threshold of this information to identify tooth-marked tongues. Shao et al. [19] defined features of tongues which focused on the change of curvature and brightness. They classified tooth-marked tongues by thresholding these feature values. Recently, some researchers have used CNN features to extract tooth-marked features. In [6], a method for extracting features using CNN, using a multi-instance classifier for final classification, was proposed.

2.2. Visual Explanation

CNN have significantly improved the performance of many computer vision tasks, such as image classification [20] and object detection [21]. There have been many recent studies exploring CNN visualizations. Zeiler et al. [22] used deconvolutional networks to visualize what patterns activate each unit and discovered the performance contribution from different model layers. Springenberg et al. [23] used guided backpropagation to make modifications to 'raw' gradients that resulted in qualitative improvements. Zhou et al. [24] indicated that a CNN learns object detectors while being trained to identify scenes. They proved that in a single forward-pass, the same network can perform both object classification and object localization. Mahendran et al. [25] reversed the characteristics of different convolutional layers and analyzed the visual coding of CNN. They showed that certain layers in the CNN retain accurate information, such as varying degrees of geometric features. Class Activation Mapping (CAM) was proposed by Zhou et al. [24]. This approach highlighted the class-specific discriminative regions by modifying the image classification CNN architecture. It replaced fully-connected layers with convolutional layers and global average pooling, and generated the CAM by mapping back the predicted category score to the previous convolutional layer. These methods are not only suitable for common datasets, but also for a variety of medical imaging tasks, such as cancer classification [26] and pneumonia detection [27].

3. Method

3.1. Problem Formulation

The tooth-marked tongue recognition task is a binary classification problem, where the input is a picture X taken in the standard image acquisition environment [14], and the output is a binary label $y \in \{0, 1\}$ indicating the absence or presence of tooth-marked tongue, respectively. For each example in the training set, we optimize the weighted binary cross entropy loss

$$L(X, y) = -w_+ \cdot y \log p(Y = 1|X) - w_- \cdot (1 - y) log p(Y = 0|X), \quad (1)$$

where $p(Y = i|X)$ is the probability that the network assigns to the label i, $w_+ = |N|/(|P| + |N|)$, and $w_- = |P|/(|P| + |N|)$ with $|P|$ and $|N|$ the number of positive samples and negative samples of tooth-marked tongue in the training set, respectively.

3.2. Model Architecture

As stated in Section 1, robust features, which can combine color, shape, and texture information of tongues, are needed to describe the tooth-marked symptom. In this paper, we use the CNN to extract a fixed-length feature vector of the tongue image. As shown in Figure 2, the proposed method takes a tongue image as input and outputs the probability of tooth-marked along with a heatmap localizing the most indicative tooth-marked regions in the image.

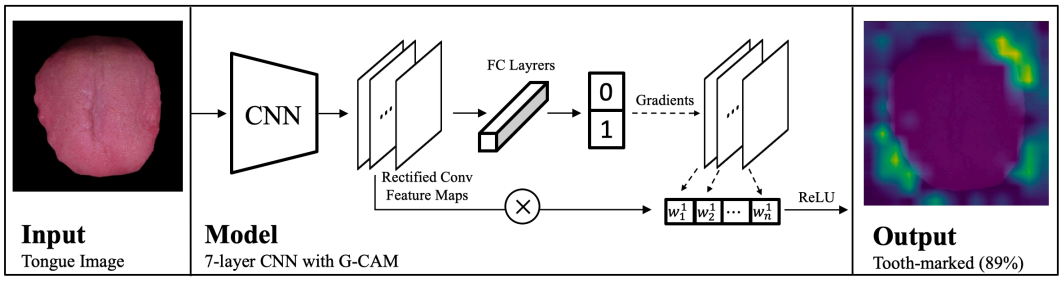


Figure 2. Our method is designed to output the probability of a tooth-marked tongue and localize regions in the image most indicative of the pathology. In this example, given a tongue image as input, we forward-propagate the image through the Convolutional Neural Network (CNN) and then compute a raw score (89%) for the class of tooth-marked tongue. Then, we reset the output to one for tooth-marked tongue predictions while zero for nontooth-marked ones. This signal is further backpropagated to the rectified convolutional feature maps of interest to acquire their corresponding gradients, which we combine to compute the coarse Gradient-weight Class Activation Mapping (Grad-CAM) localization (heatmap) which represents where the model has to look to make the particular decision.

The proposed network has seven weight layers, five of which are convolutional layers and the rest of which are fully connected layers (FC layers). Input images are downscaled to 256 × 256 and randomly cropped to 224 × 224. The convolution kernel has a size of 3 × 3 and a stride of 1, and the kernel channel for each convolutional layer is 128, except that the first layer is 64. We apply max-pooling with a size of 2 × 2 with a stride of 2 on each feature map to reduce the filter responses to a lower dimension. Instead of traditional sigmoid or tanh neurons, we use Rectified Linear Units (ReLUs) in each convolution layer and the full connection layer [20], which enables the network to converge several times faster while achieving almost identical performance. We use 0.7 dropout, followed by the fifth pooling layer, to reduce overfitting in the model training procedures. The last FC layer is a 2-way fully connected layer, that represents whether the image is a tooth-mark tongue or not. We use softmax to output the probability of each category as a classification function.

Many previous works have shown that the fully-connected layers lose spatial information about the image, but the convolutional layers naturally preserve this information, while deeper features can capture higher levels of the visual construct. Therefore, in [28], it is conjectured that the last convolutional layers have the best expression between abstract semantics and specific spatial information, and the neurons in these convolutional layers can look up the semantic information of a particular class. Grad-CAM uses the gradient values of different convolutional layers to analyze the importance of each neuron for classification [9]. In order to generate the class-discriminative localization map, $L_{Grad-CAM} \in \mathbb{R}^{u \times v}$ of width u and height v for tooth-marked tongue class, the score of the gradient for tooth-marked tongue class y is calculated, and the feature maps A^k to a convolutional layer is obtained (i.e., $\frac{\partial y}{\partial A^k}$). These gradients are fed back into global average pooling to obtain the weights of the neurons for the tooth-marked tongue α_k:

$$\alpha_k = \overbrace{\frac{1}{Z}\sum_i\sum_j}^{\text{global average pooling}} \underbrace{\frac{\partial y}{\partial A_{ij}^k}}_{\text{gradients via backprop}} . \tag{2}$$

This weight α_k represents the weight of feature map k to the tooth-marked tongue class, and ∂A_{ij}^k represents the pixel value at the (i, j) position in the feature map k. Global Average Pooling (GAP) [24] outputs the spatial average of each unit in the feature map. After obtaining the weights of the tooth-marked tongue class for all feature maps, the weights can be summed to obtain the heat map.

We calculate the weighted combination of forwarding activation maps, and further process the results by a ReLU function,

$$L_{Grad-CAM} = ReLU(\sum_k \alpha_k A^k). \quad (3)$$

We apply the ReLU function to the linear combination of maps because we only focus on the features that have a positive impact on the tooth-marked tongue. In the heat map, the highlighted areas represent pixels that contribute a large amount to the tooth-marked tongue classification.

4. Experiment and Discussion

In this section, we present four different experiments results of the proposed method. The first is the result on five-fold cross-validation, which is used to evaluate the performance of the proposed method. The second is the comparison with other works, such as Shao et al. [19] and Li et al. [6]. The third is the comparison of different receptive field sizes of CNN models. The last is the visual explanations of the most indicative regions of tooth-marked tongue using Grad-CAM. The experiments results are evaluated by the following five metrics: (1) Accuracy; (2) Precision; (3) Recall; (4) F1 Score; and (5) F2 Score. TP, FP, TN, and FN represent true positive, false positive, true negative, and false negative, respectively.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \quad (4)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad (5)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad (6)$$

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \quad (7)$$

$$\text{F2 Score} = 5 \times \frac{\text{Precision} \times \text{Recall}}{4 \times \text{Precision} + \text{Recall}}. \quad (8)$$

4.1. Dataset

As described in [14], the tongue image data should be collected in a uniform environment and contain as many high-quality images as possible. The dataset we used was provided by Shanghai Daosh Medical Technology Company, Ltd, Shanghai, China. It contained images taken at three different times for a total of 645 tongue images. These images were labeled by Chinese medicine experts. Among them, 346 nontooth-marked tongue images were marked as negative examples, and 299 tooth-marked tongue images were marked as positive examples [6].

4.2. Training

We used the above dataset to train our CNN model, described in Section 3.2. Before inputting the images into the network, we adapted some images, preprocessing to separate the tongue body from the background, and downscaled the images to 256 × 256 and cropped them randomly to 224 × 224. Since each person's tongue color was slightly different, and the color of the tongue has little effect on the recognition of the tooth-marked tongue, we also augmented the training data with random horizontal flipping and brightness adjustments.

The network was trained end-to-end using Adam with standard parameters (β_1 = 0.9 and β_2 = 0.999) [29]. We trained the model using minibatches of size 16. We used an initial learning rate of 0.001, which was decayed by a factor of 0.8 following every 2000 epochs, and we stopped our training after 12,000 epochs, since the accuracy was basically stable beyond this point.

4.3. Test

Five-fold cross-validation was used to test the proposed method. Since our dataset included a total of 645 tongue images obtained at three different times, the tongue images were randomly divided into five groups, each group containing 129 tongue images. Each time, four groups were used for training and the other one was used for testing. Table 1 shows the results of each cross-validation experiment. The proposed method was relatively stable and the average classification accuracy reached 78.6%.

Table 1. Five-fold cross-validation results.

	Accuracy	Precision	Recall	F1 Score	F2 Score
Fold_1	76.0%	74.2%	73.8%	0.74	0.74
Fold_2	77.5%	74.4%.	75.6%	0.75	0.76
Fold_3	79.8%	79.0%	79.0%	0.79	0.79
Fold_4	82.2%	84.1%.	80.0%	0.82	0.81
Fold_5	77.5%	70.7%.	82.1%	0.76	0.80
Average	78.6%	76.5%	78.1%	0.77	0.78

4.4. Comparison

We conducted experiments on our dataset using another three methods, mentioned in Section 2.1, which were proposed by Wang [18], Shao [19], and Li [6]. The cross-validation settings used in these experiments were the same as in Section 4.3. The average accuracy and recall of these four methods are recorded in Figure 3.

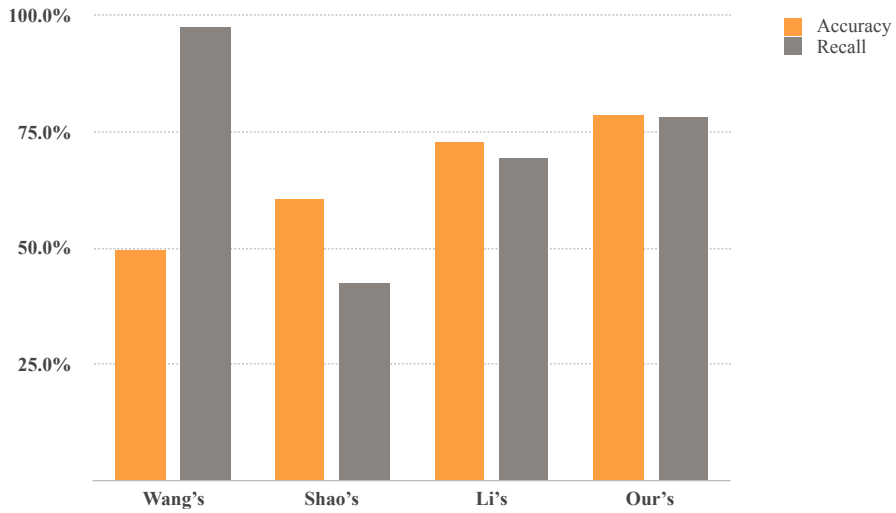


Figure 3. Comparison with other tooth-marked classification methods. Wang and Shao set thresholds based on concavity information, while Li's and our methods extracted features using CNN. Orange bars represent the accuracy and gray bars represent the recall of these four methods.

Most of traditional methods are designed based on the experience and ideas of the researchers. They are highly interpretable but not very accurate (or robust). As we can see from the results, though Wang's method had a high recall, it only used concavity information; which could easily misjudge the concave regions on the healthy tongue. Thus, the overall accuracy was low. Shao's method effectively improved the accuracy, but the recall is not guaranteed. These two methods both extract features manually and set thresholds that match the specific dataset. They don't have good generalization ability and fail to achieve a good balance between accuracy and recall. Li extracts tooth-marked tongue features using a VGG16 model, while we extract features using the model described in Section 3.2. For a more detailed comparison of these two methods, we provide the receiver operating characteristic

(ROC) curves in Figure 4. We also calculated the Area Under Curve (AUC) values of these two methods. Li's method had an AUC of 0.81 and our method had an AUC of 0.85. These two methods both had stable performance, while our model used a shallower network and achieved better results.

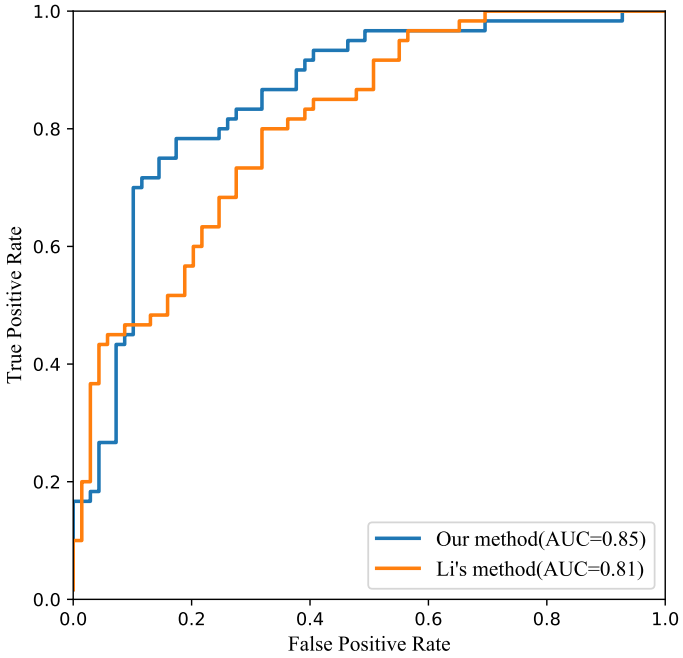


Figure 4. ROC curves for Li's method and our method. Li's method has an AUC of 0.81 and our method has an AUC of 0.85.

We provide two examples, showing the explanations from Li's method and ours, in Figure 5. The left column is the original tongue image and the tooth-marked regions are provided by TCM practitioners which are marked by blue boxes. The middle column is the visualizations provided by Li's method and the right column is the visualizations provided by our method. Although both of the methods classify the tongue image correctly, by generating Grad-CAM visualizations, we find that these two models are different in their attention. In the first row, the region on the left of the tongue has deceived Li's method, but in our method, the true region of the teeth mark is successfully locked down. In the second row, our method can correctly find the regions of the teeth mark and ignore other irrelevant regions. Li's method not only highlights the tooth-marked region but also highlights the non-toothed regions, such as the upper part of the tongue. Non-toothed regions are given a high weight, meaning that the model does not focus on the regions that have the greatest impact on the classification, which is why our method's classification accuracy is higher.

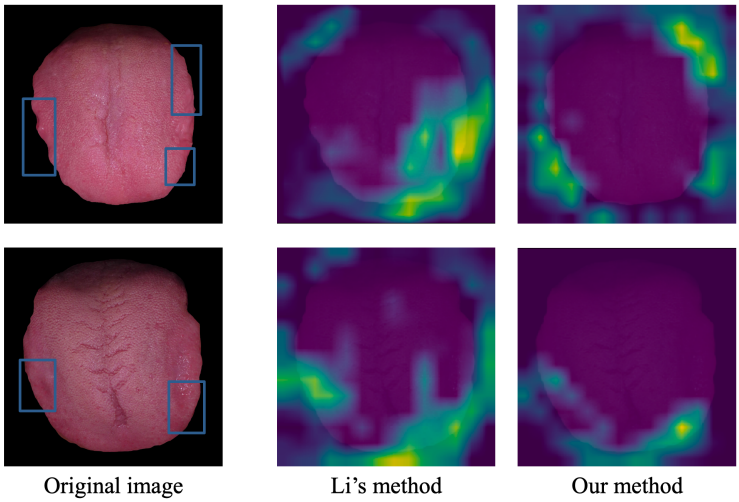


Figure 5. Grad-CAM explanations for Li's method and our method. We can see that, even though both methods made the right decision, these two models are different in their attention.

4.5. Effects of Parameters

Several parameters are involved in our CNN model design. In this section, we examine how these parameters affect the network performance. Figure 6 shows how the performance varies with respect to the number of convolution (Conv) kernels. The network architecture starts from a 256-kernel in 2-Conv layer to 5-Conv layer and halving parameters. We find that, although many mainstream networks use 256 or more kernels every layer, our model tends to prefer 128 kernels. This may be because the background of the tongue image is single, the position of the tongue is clear, and too many parameters will lead to under-fitting of the model.

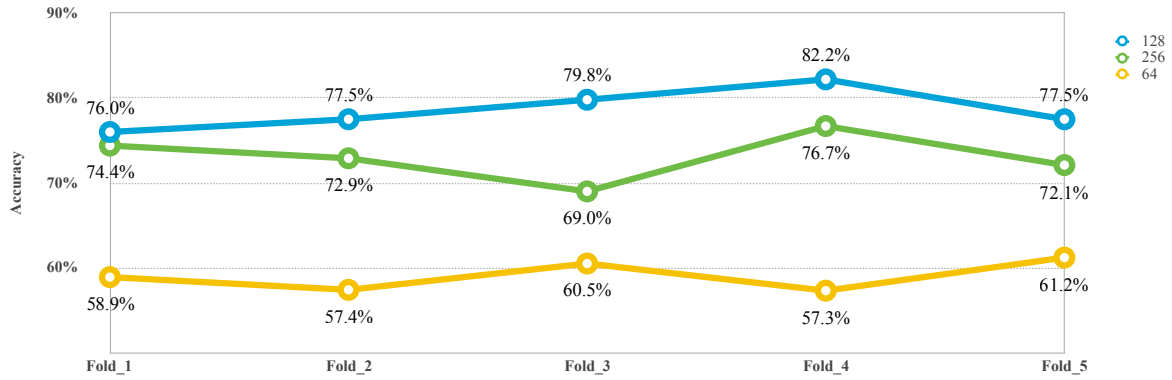


Figure 6. Accuracy with respect to varying the different number of convolutional kernels. The blue, green, and yellow lines represent 128, 256, and 64 kernels, respectively.

Through observation, we found that the size of the tooth mark region accounts for about 1/8 of the whole picture. Therefore, in theory, the neuron receptive field for detecting tooth marks doesn't need to be too large. For this reason, we explored the influence of different sizes of receptive fields on the classification results. The receptive field was defined as the region in the input space that a particular CNN's feature is looking at [30]. Our choices for kernel size include (3×3, 5×5, and 7×7), and for convolutional layers include (3, 4, 5, and 6). Based on [30], we compute the size of receptive fields in different network models. The experiment results are shown in Tables 2 and 3. We find that the model with a 3×3 kernel size is the most effective. As stated in [31], 3×3 kernel layers have more non-linear rectifications, which makes the decision function more discriminative. A 5-Conv layer whose receptive field is 94 performs better than others. In [9], it was found that the

best-looking visualizations are often obtained after the deepest convolutional layer in the network, and localizations get progressively worse at shallower layers. This is because the later convolutional layers capture high-level semantic information and retain spatial information, while the shallower layers have smaller receptive fields and only concentrate on local features. However, in our research, we find that a 5-Conv layer network performs better than a 6-Conv layer network. So, the accuracy of image classification does not depend entirely on the number of network layers, but also on the specific classification tasks. It is necessary to understand the problem deeply and build a suitable network model to solve the specific problem.

Table 2. Comparison between different kernel sizes with the same convolutional layers, and the 3×3 kernel size is the most effective. (Conv Layer: 5-Conv layer).

Kernel Size	Receptive Field	Accuracy	Precision	Recall	F1 Score	F2 Score
3 × 3	**94**	**78.6%**	**75.9%**	**78.1%**	**0.77**	**0.78**
5 × 5	156	76.0%	74.1%	74.1%	0.74	0.74
7 × 7	218	73.6%	71.1%	72.9%	0.72	0.73

Table 3. Comparison between different convolutional layers with the same kernel size, and the 5-Conv layer is the most effective. (Kernel Size: 3×3).

Conv Layer	Receptive Field	Accuracy	Precision	Recall	F1 Score	F2 Score
3-Conv	22	71.3%	70.4%.	62.1%	0.66	0.64
4-Conv	46	73.6%	77.3%.	56.1%	0.65	0.59
5-Conv	**94**	**78.6%**	**75.9%**	**78.1%**	**0.77**	**0.78**
6-Conv	190	72.1%	72.5%.	67.7%	0.70	0.69

4.6. Model Interpretation

To interpret the model predictions, we analyze how the localizations change qualitatively as we perform Grad-CAM with respect to different features maps in our model. As we can see from Figure 7, in the first few layers, CNN pays more attention to the edge and color information which are important for the next layers. Then, later convolutional layers start to detect the texture associated with the tooth marks. The Grad-CAM highlights the discriminative regions, which are usually some indentations along the lateral borders, and the color and brightness of them differ from the normal regions. It is interesting to see how our recognition method can serve as a tool to understand the network better by providing a localized high-resolution visualization of the tooth-marked regions. It shows precise localization to support the model's prediction. We think it helps to judge which model is more effective, as stated in Section 4.4, and also helps doctors to analyze the region of the tooth-marked tongue, rather than simply giving the classification results.

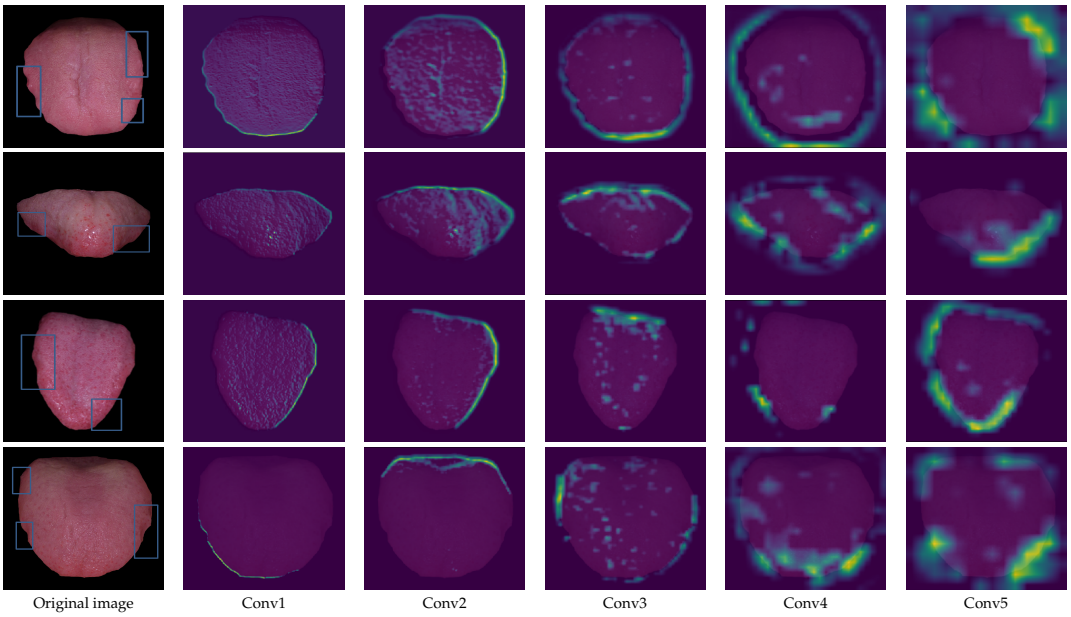


Figure 7. Grad-CAM localizations for the "tooth-marked tongue" category on different convolutional layer feature maps in our model. The first column is the original tongue image and teeth-marked regions are contained in blue boxes. The remaining columns of each row corresponds to Conv1–Conv5.

5. Conclusions

Automated recognition of tooth-marked tongue would not only have benefit in clinical settings, but it would also be invaluable in the delivery of health care to populations with inadequate access to diagnostic imaging specialists. In this paper, we have presented a tooth-marked tongue recognition method based on deep features and localized discriminative regions using Grad-CAM. The experiments showed that the proposed method has greatly improved accuracy, compared to previous methods. We mainly improve the problem in the following three aspects: (1) Analysis of the entire tongue picture—no need to cut small patches. (2) Analyzing the influence of different receptive field sizes on the classification results and finding the receptive field size suitable for the tooth-marked tongue problem. (3) Enhancing the interpretability of the CNN algorithm. Future work includes two aspects: (1) More new tongue samples will be acquired. Since we use a deep CNN as a feature extractor, the proposed model will benefit a lot from a larger dataset. (2) We will improve the architecture of the CNN model to further improve the accuracy and reduce the computation cost, and use patient history (or other relevant characteristics) in the future.

Author Contributions: Methodology, Experimental analysis and Paper Writing, Y.S.; Writing-review and Data analysis, S.D.; Data and Writing Correction, Y.Z. and J.L.; The work was done under the supervision and guidance of X.L.; All the authors revised the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Shen, Z.Y. Basic theory of traditional Chinese medicine. *Chin. J. Integr. Tradit. West. Med.* **1997**, *17*, 643.
2. Pang, B.; Zhang, D.; Wang, K. Tongue image analysis for appendicitis diagnosis. *Inf. Sci.* **2005**, *175*, 160–176. [CrossRef]
3. McLean, N. Color atlas of oral diseases. *Br. J. Plast. Surg.* **2004**, *100*, 1299–1300. [CrossRef]
4. Li, W.; Luo, J.; Hu, S.; Xu, J.; Zhang, Z. Towards the Objectification of Tongue Diagnosis: the Degree of Tooth-marked. In Proceedings of the IEEE International Symposium on It in Medicine and Education, Xiamen, China, 12–14 December 2008; pp. 592–595.

5. Ren, Y.; Rong, L.; Ying, Z. Study on the correlation between dental scar tongue and constitution of traditional Chinese medicine in physical examination population. *World Sci. Technol.-Mod. Tradit. Chin. Med.* **2012**, *14*, 2283–2289.
6. Li, X.; Yin, Z.; Cui, Q.; Yi, X.; Yi, Z. Tooth-Marked Tongue Recognition Using Multiple Instance Learning and CNN Features. *IEEE Trans. Cybern.* **2019**, *49*, 380–387. [CrossRef] [PubMed]
7. Zhang, D.; Pang, B.; Li, N.; Wang, K.; Zhang, H. Computerized diagnosis from tongue appearance using quantitative feature classification. *Am. J. Chin. Med.* **2005**, *33*, 859–866. [CrossRef] [PubMed]
8. Wang, Y.G.; Yang, J.; Zhou, Y.; Wang, Y.Z. Region partition and feature matching based color recognition of tongue image. *Pattern Recognit. Lett.* **2007**, *28*, 11–19. [CrossRef]
9. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In Proceedings of the International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 618–626.
10. Chiu, C.C. A novel approach based on computerized image analysis for traditional Chinese medical diagnosis of the tongue. *Comput. Methods Progr. Biomed.* **2000**, *61*, 77–89. [CrossRef]
11. Pang, B.; Zhang, D.; Li, N.; Wang, K. Computerized tongue diagnosis based on Bayesian networks. *IEEE Trans. Biomed. Eng.* **2004**, *51*, 1803–1810. [CrossRef] [PubMed]
12. Zuo, W.; Wang, K.; Zhang, D.; Zhang, H. Combination of polar edge detection and active contour model for automated tongue segmentation. In Proceedings of the International Conference on Image and Graphics, Hong Kong, China, 18–20 December 2004; pp. 270–273.
13. Yu, S.; Yang, J.; Wang, Y.; Zhang, Y. Color Active Contour Models Based Tongue Segmentation in Traditional Chinese Medicine. In Proceedings of the International Conference on Bioinformatics and Biomedical Engineering, Wuhan, China, 6–8 July 2007; pp. 1065–1068.
14. Wang, X.; Zhang, B.; Yang, Z.; Wang, H.; Zhang, D. Statistical analysis of tongue images for feature extraction and diagnostics. *IEEE Trans. Image Process.* **2013**, *22*, 5336–5347. [CrossRef] [PubMed]
15. Huang, B.; Wu, J.; Zhang, D.; Li, N. Tongue shape classification by geometric features. *Inf. Sci.* **2010**, *180*, 312–324. [CrossRef]
16. Zhang, Y. Research on Analysis Method of Tongue and Teeth-Marked Tongue. Ph.D. Thesis, Beijing University of Chinese Medicine, Beijing, China, 2005.
17. Li, J.F.; Li, N.M.; Wang, K.Q.; Zhang, H.Z. Extracting feature of teeth-marked tongue image. In Proceedings of the Diagnosis Section of China Society of Integrated Traditional Chinese and Western Medicine, Fuzhou, China, 1 July 2009; pp. 100–105.
18. Wang, H.; Zhang, X.; Cai, Y. Research on Teeth Marks Recognition in Tongue Image. In Proceedings of the Academic Conference on National Diagnosis of Chinese Society of Integrated Traditional Chinese and Western Medicine, Shenzhen, China, 30 May–1 June 2014; pp. 80–84.
19. Shao, Q.; Li, X.; Fu, Z. Recognition of teeth-marked tongue based on gradient of concave region. In Proceedings of the International Conference on Audio, Language and Image Processing, Shanghai, China, 7–9 July 2015; pp. 968–972.
20. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. In Proceedings of the International Conference on Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–6 December 2012; pp. 1097–1105.
21. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
22. Zeiler, M.D.; Fergus, R. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2014; pp. 818–833.
23. Springenberg, J.T.; Dosovitskiy, A.; Brox, T.; Riedmiller, M. Striving for Simplicity: The All Convolutional Net. *arXiv* **2014**, arXiv:1412.6806.
24. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning Deep Features for Discriminative Localization. In Proceedings of the Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27 June–30 June 2016; pp. 2921–2929.
25. Mahendran, A.; Vedaldi, A. Understanding deep image representations by inverting them. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 5188–5196.

26. Esteva, A.; Kuprel, B.; Novoa, R.A.; Ko, J.; Swetter, S.M.; Blau, H.M.; Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **2017**, *542*, 115–118. [CrossRef] [PubMed]
27. Rajpurkar, P.; Irvin, J.; Zhu, K.; Yang, B.; Mehta, H.; Duan, T.; Ding, D.; Bagul, A.; Langlotz, C.P.; Shpanskaya, K.; et al. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. *arXiv* **2017**, arXiv:1711.05225.
28. Mahendran, A.; Vedaldi, A. Visualizing Deep Convolutional Neural Networks Using Natural Pre-images. *Int. J. Comput. Vis.* **2016**, *120*, 233–255. [CrossRef]
29. Kingma, D.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2014**, arXiv:1412.6980.
30. Girshick, R.B.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
31. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2015**, arXiv:1409.1556.