Clustered-CAM: Visual Explanations for Deep Convolutional Networks for Thyroid Nodule Ultrasound Image Classification

Ali Eskandari[*1] 2007269@buckingham.ac.uk
Hongbo Du[1] hongbo.du@buckingham.ac.uk
Alaa AlZoubi[†1] alaa.alzoubi@buckingham.ac.uk
[1] *School of Computing, University of Buckingham, Buckingham, UK*

Editors: Under Review for MIDL 2022

Abstract

Explaining the CNN classification decision is crucial for the system acceptance in critical applications such as tumour recognition in 2D Ultrasound images. Generating saliency maps that highlight the image regions contributing to the final CNN decision is one of the most common techniques. In this paper, we propose a clustering-based approach to group similar feature maps before assigning importance scores to produce a more accurate and less sensitive visual explanation for CNN models for thyroid nodule classification in US images. Our study with a dataset of 864 ultrasound images shows that the Clustered-CAM achieved a lower average drop and higher percent increase in confidence comparing to the-state-of-the-art techniques. We demonstrate that Clustered-CAM is an effective and promising approach for visualising the CNN model decisions for thyroid nodule recognition.

Keywords: Deep Learning Visualization, Thyroid Cancer Recognition, Ultrasonography.

1. Introduction

Automated thyroid nodule classification is critical for early detection of thyroid cancer and reduction in mortality rates. Some existing CNN-based solutions for classifying thyroid nodules in Ultrasound (US) images have matched or outperformed radiologists, but the "black-box" nature of such models leads to poor interpretability of model decisions. Explaining the decisions of such models is essential to accept the system in the clinical practice; therefore, few attempts (Eskandari et al., 2021) have been made towards CNN decision comprehension. Recently, techniques such as Grad, Group, and Ablation-CAM have been developed to visually explain decisions of CNN models for natural image classification. The output of Grad-CAM reflects subtle changes in the prediction, which might not be large enough to alter the decision of the CNN. Group-CAM (Zhang et al., 2021) attempted to solve this issue by splitting the activation maps into groups (without considering their similarities), summing up the sub-activations, and compute the saliency map. Ablation-CAM freezes each feature map from the final convolutional layer and then assesses whether or not the prediction class is unchanged. EGrad-CAM (Eskandari et al., 2021) shows that several feature maps of the final convolutional layer have either no contributions or similar characteristics. This led to the question over whether *a group of highly similar feature maps*

[*] Contributed equally
[†] Contributed equally

be represented in one saliency map that captures accurate model decision. Inspired by the Ablation and EGrad-CAM, we developed a new method, Clustered-CAM, for visualizing CNN model decisions. We aim at investigating the behaviour of ablating a group of similar feature maps in the visualization of the CNN final decision and increase the accuracy of the saliency map outputs.

2. Data and Method

Data Collection and Pre-processing: Two datasets of US images of thyroid nodules were collected from two hospitals in China. Dataset A of 421 benign and 298 malignant images was used to train the CNN model TNet. Dataset B of 65 benign and 80 malignant was used as a test set. Each image in both datasets contains one nodule with a set of points on the boundary provided by expert radiologist. Fine Needle Aspiration confirmed the nodule status. Region of Interest (RoI) was then derived from the point coordinates on the boundary. We added a small margin (8% of the nodule width and height) around the nodule to include some surrounding tissues. The RoI images were resized to 224×224×3 using bicubic interpolation and then used to train the CNN model.

Visual Explanations of CNN Model: We used CNN architecture and training hypermeters in (Eskandari et al., 2021) to build a CNN model TNet. Our Clustered-CAM method efficiently generates saliency maps by clustering the feature maps. Figure 1 shows our Cluster-CAM approach. For an input image, similar feature maps in the last convolutional layer of TNet are clustered into n groups $(G_1, ..., G_i, ..., G_n)$. Then, we ablate all feature maps (F_{G_i}) in group G_i, calculate the changes in the activation score (pre-softmax) of class c, and assign the computed weight to all feature maps within the cluster G_i. In particular, we first flatten each feature map in the last convolutional layer into a vector v. Then, we group the flattened feature maps using the k-means algorithm and cosine similarity. Second, we freeze all feature maps in each cluster and estimate the changes in the activation score of class c. Equation 1 shows the significance value (score-weighted) of each feature map within the same cluster $(w^c_{G_i})$ and the heatmap visualisation (M_c).

$$w^c_{G_i} = (S_c - S^{G_i}_c)/S_c;\ M_c = ReLU(\sum_{i=1}^{n} w^c_{G_i} F_{G_i}) \tag{1}$$

where S_c is the activation score of class c, and $S^{G_i}_c$ is the score for class c when feature map(s) in cluster G_i is ablated. The ReLU used to remove the effect of negative weights in class c, i.e., visualising those feature map(s) that their absence decreases the class score S_c.

3. Experimental Results

To determine the classification accuracy of the TNet model, a stratified 10-folds cross-validation was applied on Dataset A. TNet models achieved an average accuracy of 86.5% (i.e. 83.9% TPR, and 88.6% TNR). To analyse the models' decision visualisation output, we selected the best model (accuracy of 86.3%, with 87.7% TNR and 84.4% TPR) that has high overall accuracy and balanced TPR and TNR. The same model was tested on Dataset

Table 1: Comparison between our Clustered-CAM (K=64) with state-of-the-art Grad, EGrad and Ablation-CAM (ADC: lower is better; PIC: higher is better)

Metrics	**Grad-CAM**	**EGrad-CAM**	**Ablation-CAM**	**Clustered-CAM**
ADC %	40.35	40.35	36.70	**34.09**
PIC %	25.52	25.52	24.14	**28.28**

B, achieving 84% accuracy (89% TPR and 77% TNR). Using the selected TNet model, we compare the performance of our Clustered-CAM against the state-of-the-art methods over Dataset B. To evaluate the trustworthiness of each method, for each image in Dataset B, we generated a heatmap by capturing the importance of each image-region in the final TNet decision. We use two metrics to evaluate the performance of explanation maps: Average Drop in Confidence (ADC) and Percent Increase in Confidence (PIC), as given below:

$$ADC\% = \frac{1}{N}\sum_{i=1}^{N}\frac{max(0, Y_i^c - O_i^c)}{Y_i^c} * 100; \; PIC\% = \sum_{i=1}^{N}\left(\frac{1_{Y_i^c} < O_i^c}{N}\right) * 100 \qquad (2)$$

The confidence score of Y_i^c is calculated when the original image is used as input. O_i^c is computed when the explanation map is used as input. N is the number of images in dataset B. In PIC Equation, $1_{Y_i^c}$ returns 1 if the argument is true. Table 1 shows the performance of Clustered-CAM (64 clusters). We also evaluated the method with 16 and 32 clusters where ADC of 36.47 and 38.22, and PIC of 26.90 and 24.83, were achieved respectively. Besides, the average computation time of analysing Dataset B using our Clustered-CAM is 3.75 seconds comparing to 14.19 seconds of Ablation-CAM. Figure 2 compares the visualisation output of different techniques.

4. Conclusion

This study presents a new method (Clustered-CAM) to investigate the efficacy of applying ablation on a group of similar feature maps for accurate saliency maps generation. Intrigued by the results, we plan to extend our analysis into other cancer types (e.g. breast). We also want to investigate the performance of our Clustered-CAM using different CNN models.

Acknowledgments: This research is funded by TenD AI Medical Technologies Ltd, Shanghai, China.

References

Ali Eskandari, Hongbo Du, and Alaa AlZoubi. Towards linking cnn decisions with cancer signs for breast lesion classification from ultrasound images. In *Annual Conference on Medical Image Understanding and Analysis*, pages 423–437. Springer, 2021.

Qinglong Zhang, Lu Rao, and Yubin Yang. Group-cam: Group score-weighted visual explanations for deep convolutional networks. *arXiv preprint arXiv:2103.13859*, 2021.

Appendix A

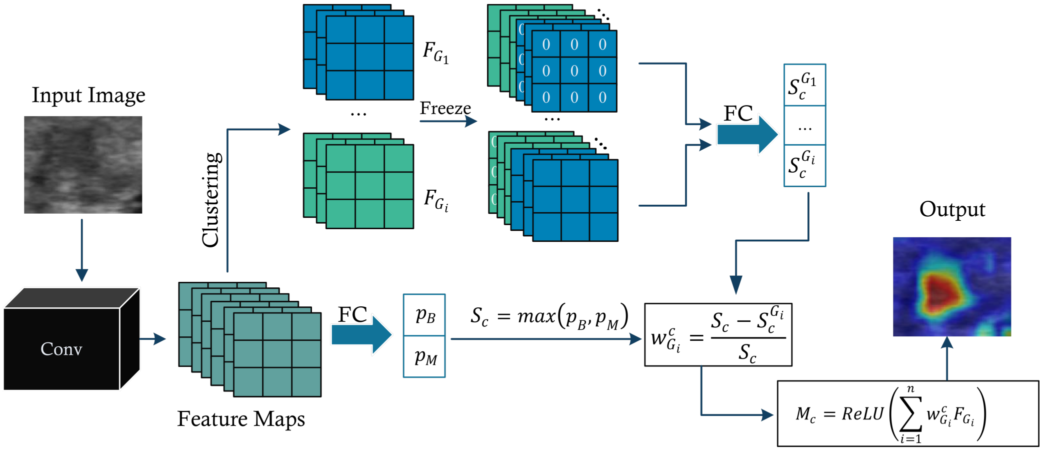

Figure 1: Overview of the Proposed Clustered-CAM.

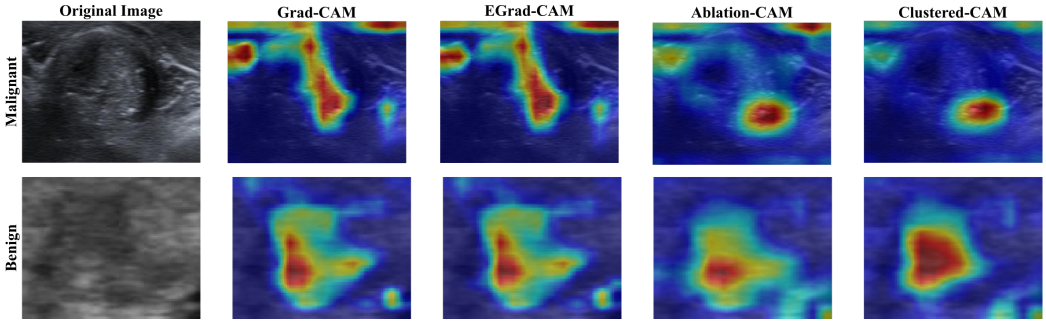

Figure 2: Visualisation Output of Different Techniques for correctly classified benign and malignant thyroid nodules by TNet. Clustered-CAM has lower ADC (malignant case) and higher PIC (benign case).