THROAT INFECTION DETECTION USING DEEP LEARNING

Yelamanchili Sai Vamsi, Student, Department of Electronics and Computer Engineering, Vignan's Institute of Information Technology, India.
Nemani Prathusha, Student, Department of Electronics and Computer Engineering, Vignan's Institute of Information Technology, India.
Koidala Lohith, Student, Department of Electronics and Computer Engineering, Vignan's Institute of Information Technology, India.
Yelamanchili Venkata Sai Sinith Dora, Student, Department of Electronics and Computer Engineering, Vignan's Institute of Information Technology, India.
Vadaboyina Appalaraju, Student, Department of Electronics and Computer Engineering, Vignan's Institute of Information Technology, India.

Abstract

The recent global pandemic of corona virus disease (COVID-19) encourages the use of automated healthcare systems for patients with respiratory symptoms. Throat infection, also called as pharyngitis leads to damage of the tissues of the at throat area. The walls or structures of the throat bulges and this could be very painful. Therefore, there is a need for an effective automated technique to detect the throat infection. In this work, the authors present a study on automated detection of throat infection using modern deep learning algorithms. The authors have applied a Convolution Neural Network (CNN) based architecture ResNet50, and CoAtNet model that is based on both Convolution Neural Network (CNN) and Vision Transformer (ViT) architecture for detection of throat infection, and found which one of these models gave the best accuracy for this particular problem. The results are provided as classification accuracy. The authors have found that the CoAtNet model gives the highest accuracy of 96.6%.

Keywords: Deep Learning, Transfer Learning, Convolution Neural Network (CNN), Vision Transformer (ViT), Image classification, Throat infection.

Introduction

Detection of throat infections involves in the medical partitioners manually examining the patients with their mouth opened. Without proper precautions this could be very dangerous especially during this recent COVID19 pandemic where the virus spread through air. Thus, well-developed, trustworthy and efficient automatic classification processes can reduce the risk & efforts for the medical partitioners and allow them to treat more patients.

The challenge here is the availability of the labeled dataset for this particular use case. [1] Previous studies had employed Generative Adversarial Neural Networks (GANs) [2] for data augmentation to address this issue. However, training the GANs is computationally very expensive. And they cannot be used for most cases where synthetic data fail to capture new information. In the present study, the authors proposed a method to detect throat infection using deep learning without having to use GANs for data augmentation. The proposed method helps medical partitioners to during the examination of the patients. [3] Convolution Neural Networks (CNN) had been the go-to networks when it comes to automatic medical image analysis tasks. However, the [4] Transformer based architectures, Vision Transformer (ViT) which is now prominent in the field of Natural Language Processing (NLP), has begun to catch up with CNNs, particularly in image classification applications.

A comparative study with existing SOTA deep learning models one based on CNN architecture called ResNet50, a 50-layer Residual network based on the concept of creating deeper layers with improved classification accuracy for complicated tasks [6] and the other based on both CNN and ViT architecture called [7] CoAtNet is performed. It shows that CoAtNet (trained with only basic data augmentation) out performs ResNet50 (trained with data augmented using GAN) with 96.6%.

Statement of the Problem

In order to improve the automation involved in the diagnosis which in turn leads is better reach of medical facilities, especially in the areas that involves the mouth we need to come up with an effective algorithm to process the throat images.

Objectives of the study

- To identify the best algorithm for automatic throat infection detection.
- To analyze the behavior of the predictions based on various parameters such as of accuracy, sensitivity and specificity.

Dataset used

This collection contains 208 photos of normal throats and 131 images of throat infections. Figure (1) shows a portion of the dataset that was used.

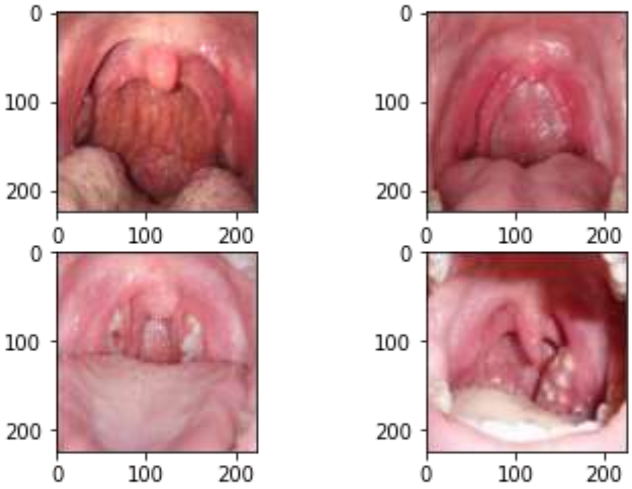

Figure 1. The top row two are normal throat images and the bottom row images are the infected throat images.

Review of Literature

In [1], authors had compared the performance between Inception-v3, ResNet50 & MobileNet-v2, to detect pharyngitis using dataset containing two categories: 131 photos of pharyngitis and 208 images of normal throat (the same dataset we used in our study). All of the DL models are based on CNNs, and they used CycleGan [9] for data augmentation, as well as basic image augmentation like left and right flipping, random rotation from 10° to 10°, width and height translation from 5% to +5%, zooming in and out from 0% to 20%, and randomized brightness change from 10% to 10%. And showed ResNet50 trained on CycleGan synthesized images gives the highest accuracy.

In research [10], authors had proposed a vision system that is based on Histogram of Gradients (HOG) integrated with a deep neural network (DNN). The model was trained on a dataset with only 40 images in total with 25 normal throat images and rest 15 images of infected throats.

Research [11] had illustrated image segmentation can be used as feature extraction method to get the region of interest (ROI), solely based on pixel intensity values across the three channels (Red, Blue & Green) and with popular machine learning algorithm called K-Nearest-Neighborhood (KNN) an accuracy up to 93.75% can be achieved. For this study the authors had taken a dataset with total 56 images with 26 of them being normal throat images and the rest 26 being the infected throat images.

In both the above-mentioned studies the models fail to capture the spatial information of the images.

Research Methodology

In this study, we make use of the existing SOTA DL algorithms to understand which architecture is best suited for our use case. The experimentation was performed according to the following workflow in fig. 2

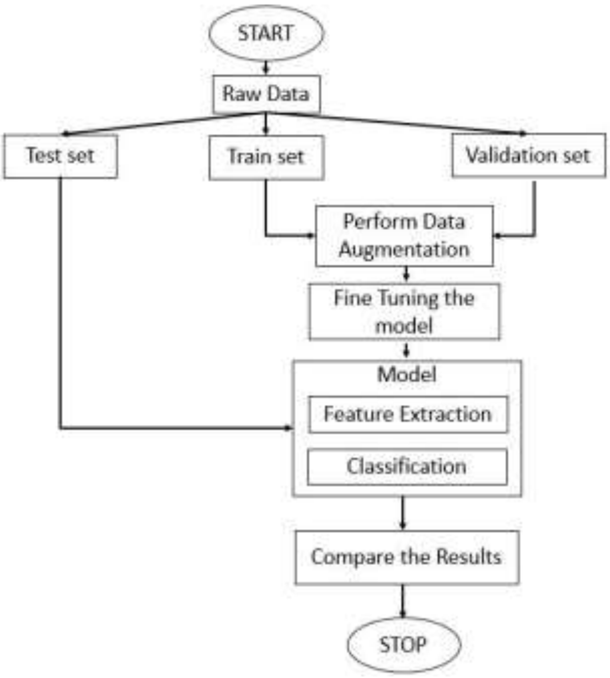

Figure 2. Overall Experiment Workflow

Basic data augmentation techniques are employed, including right and left flip, vertical flip, translation in height and width from 6% to +6%, randomized rotations from 25° to 25°, zooming in and out from 4% to 26%, shear from 10° to 10°, and random change in brightness from 40% to 40%.

The training set was made up of 1500 normal throat photos and 800 infected throat images, while the validation and test sets were made up of 600 normal throat images and 400 pharyngitis images.

Then we used the CoAtNet model pre-trained on the ImageNet dataset [12] for transfer learning and fine-tuned the model.

Finally, the both ResNet50 and CoAtNet models are compared based on the defined performance metrics including accuracy, sensitivity & specificity.

A. Convolution Neural Network (CNN)

[2] CNNs has an ability to compress the image data while preserving the vital information in the image. The architecture is divided into two sections: a convolution layer and a pooling layer [13].

The convolution layer is used to get vital information from the image called feature maps. It achieves this by performing a mathematical operation called convolution with certain filters.

The pooling layers are for reducing the resolution of the convolved feature maps to cut the computational costs. An outline of the CNN architecture is given in Figure (3).

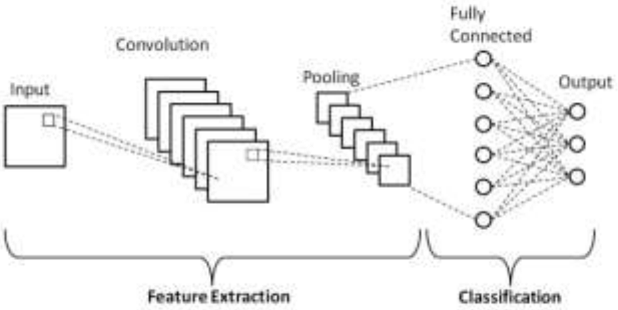

Figure 3. Convolution Neural Network Architecture

The convolution layer relies on a static filter to abstract the information.

$$y_i = \sum_{j \in \mathcal{L}(i)} w_{i-j} \odot x_j \quad \text{(Depth-wise convolution)}, \quad (1)$$

Here, L(i) denotes a local neighborhood of i and x_i, $y_i \in R^D$ are the input and output at position i.

B. Vision Transormer (ViT)

Ever since they first got introduced by Dosovitskiy et al [4], ViT models had given CNNs a very tough competition. The reason is that CNNs are poor when it comes to capturing information at the pixel level because the static nature of the window size.

An outline of the ViT architecture design is given in Figure (4). Principally, how a ViT works is the input images were initially gets fragmented into smaller segments of same size called patches. The Transformer encoder is fed with a sequence of one-dimensional patch embeddings, where self-attention modules are utilized in determining the relation-based weighted sum of each hidden layer's outputs. Being able to learn the global dependencies in the input images is the reason why Transformers are able to perform so well.

Self-attention [14] is based on the idea that at any given point there are only certain specific regions in an image where there is a need of focusing rather than considering the entire image as it is. Just link the way human vision works. It relates every single position with every other position of the same sequence. Pair wise similarity between the pair (x_i, x_j) is calculated as shown in the below equation:

$$y_i = \sum_{j \in \mathcal{G}} \underbrace{\frac{\exp\left(x_i^\top x_j\right)}{\sum_{k \in \mathcal{G}} \exp\left(x_i^\top x_k\right)}}_{A_{i,j}} x_j \quad \text{(self-attention)}, \quad (2)$$

where G indicates the global spatial space.

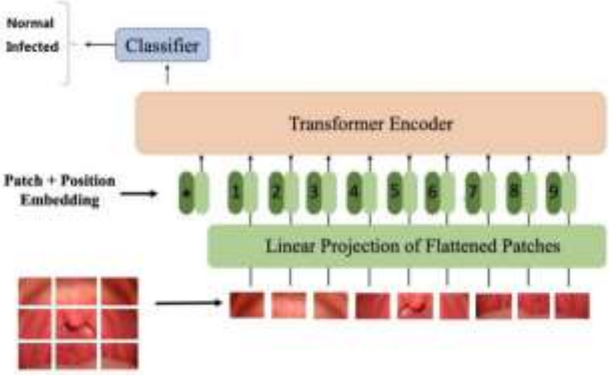
Normal
Infected
Classifier
Transformer Encoder
Patch + Position Embedding
*
1
2
3
4
5
6
7
8
9
Linear Projection of Flattened Patches

Figure 4. Vision Transformer Architecture

C. ResNet50

[5] ResNet50 is by far the best CNN based architecture. It solves the issue of the vanishing gradient problem by introducing skip connections. It was created by Kaiming for residual learning, which can be interpreted as the derivation of input properties from a specific layer. ResNet can accomplish this by using shortcut connections with each of the thirty-three filters, directly attaching the input of the nth layer to the (n + m)th layer. Weights will boost the neighboring layer throughout the model’s training, and will adjust the weights to preserve the prior layer.

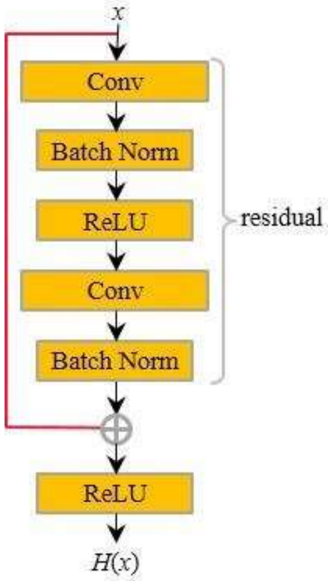

Figure 5. Residual block structure in ResNet50

A non-linear activation function, rectified linear unit (Relu) [15] is the activation function which is used in convolutional layers.

f(y) = max(y, 0)

D. CoAtNet

Though transformers have a higher model capacity than convolutional networks, their generalization might be inferior due to a lack of the appropriate inductive bias. In order to successfully integrate the advantages of both designs, the authors of [6] introduced CoAtNet.

By stacking the attention layers & convolution layers in a certain order CoAtNet achieves better generalization and capacity.

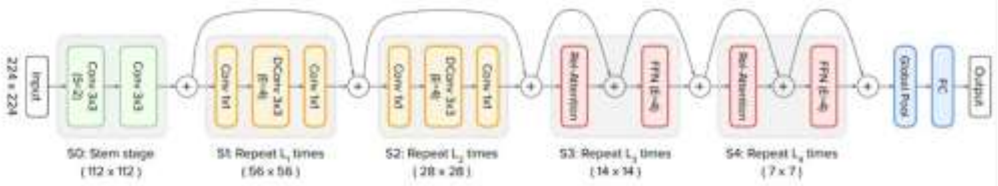

Figure 6. The CoAtNet architecture

By summing the attention matrix with a fixed convolution filer, before or after the Softmax normalization, i.e.,

$$y_i^{\text{post}} = \sum_{j \in \mathcal{G}} \left(\frac{\exp\left(x_i^\top x_j\right)}{\sum_{k \in \mathcal{G}} \exp\left(x_i^\top x_k\right)} + w_{i-j} \right) x_j$$

(or)

$$y_i^{\text{pre}} = \sum_{j \in \mathcal{G}} \frac{\exp\left(x_i^\top x_j + w_{i-j}\right)}{\sum_{k \in \mathcal{G}} \exp\left(x_i^\top x_k + w_{i-k}\right)} x_j. \quad (3)$$

E. Transfer learning & Fine tuning

Transfer learning is the process at which the knowledge learnt to perform a task often referred as base task is transferred and used in performing another different desired task called target task. For training on the target task along with the features from the base task, it's related dataset will be used. This will work only when these tasks share similar feature sets [16].

Back propagation is how neural networks learn. During the training period all the weights in the network gets adjusted to their optimal values that yields the highest performance. All the knowledge of the network is nothing but these weights. So, if we don't make these weights persistent all the knowledge transferred from the base task will be lost due to back propagation. To make the weights persistent we freeze the respective layers. Hence while training for the target task, the top layers of pre-trained models aren't frozen, which means they learn through back propagation, although the rest of the model is frozen. Fine-tuning [17] is the process of updating the weights during back-propagation. Because the mean and variance of those layers might not be similar to those of target dataset, fine-tuning of the top layers is necessary. Finally, fine-tuning the upper layers will compensate for the target dataset's mean or variance.

For comparison, all trials are run on the same fixed training and testing dataset. The final evaluation is done using 4-fold cross-validation (CV). The output layer is changed in both TL trials using a classifier with the [18] Softmax activation function. The Adam [19] optimizer is employed, and the models are trained for 250 epochs (with early stopping). Weighted cross-entropy loss was choosen as the loss function. To accommodate our unbalanced training dataset, fine-tuning is conducted, optimised by stochastic gradient descent (SGD) [20] with 0.8 for momentum for the CNN & ViT based model, i.e., CoAtNet. A batch size of 20 is chosen for fine-tuning, with a cosine decay learning rate of 0.0001 and 10 linear warmup steps, and with early stopping the entire training dataset had gone through back propagation for 250 times.

Results and Discussion

We tested and compared the CNN based model ResNet50 with CNN & ViT based model CoAtNet for transferring pretrained models to classify throat images into two categories: normal and infected.

The ResNet50 model was trained with CycleGan generated synthetic images along with basic augmentation. While the CoAtNet model was trained with base dataset with basic data augmentation.

The performance of the CNN & ViT model (CoAtNet) and the CNN model (ResNet50) acquired from the TL demonstrates that the CNN & ViT model beats the SOTA CNN networks.

The tiny amount of the dataset used in training the CoAtNet model compared to the larger dataset used in training the ResNet50 model is an intriguing aspect of our findings, demonstrating the potential strength of attention-based models in medical throat image processing. The relationship between spatial information, or more especially the large-scale interdependence between distinct patches, is considerably more evident in the throat pictures, which might explain the efficacy of the CNN & ViT based model on such a short dataset.

We used the true positive (A), true negative (B), false positive (C), and false negative (D) data to calculate accuracy, sensitivity, and specificity as performance metrics:

$$\text{Sensitivity} = ((A) / (A + D))$$

$$\text{Specificity} = ((C) / (C + D))$$

$$\text{Accuracy} = ((A + C) / (A + C + B + D))$$

In terms of the number of pictures, A, B, C, and D were tallied. A represents the number of photographs that were accurately identified as infected when they were infected, and B represents the number of images that were wrongly identified as infected when they were normal. On the other hand, C represents the number of photos that were properly identified as normal provided that they are normal, and D represents the number of images that were wrongly identified as normal given that they are infected. The average accuracy of the 4-fold cross-validation was obtained by taking the sum of all the accuracy values from all rounds of the cross-validation and divide it with the number of accuracy values from all the rounds.

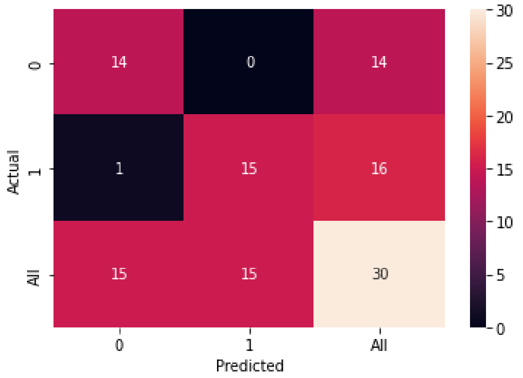

Figure 7. Confusion matrix of the CoAtNet model.

The above Figure (7) shows the confusion matrix for the CoAtNet model. With normal encoded as zero and infected encoded as one.

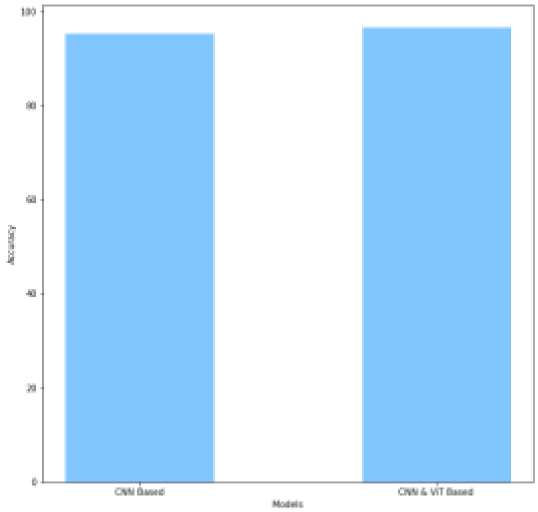

Figure 8. shows the bar graph representation for the models with their respective accuracies.

The final results are illustrated in the table below.

Table 1. Classification performance for throat infection detection.

Evaluation	Accuracy (%)	Sensitivity (%)	Specificity (%)
ResNet50 (using CycleGan for data augmentation)	95.3	92.9	96.8
CoAtNet	**96.6**	93.7	93.3

Conclusion

We used a publicly available dataset from the Medeley Data repository for this investigation. The CoAtNet model, which is built on both CNN and ViT, surpassed the ResNet50 model with 96.6 percent accuracy without even being trained on GAN produced synthetic pictures, according to the results. When trained using GAN produced syntactic pictures, the suggested model's accuracy can be improved even further. To improve the model's accuracy, image segmentation can be used as a feature extraction procedure in the future. A smartphone application might be created that allows users to diagnose their throat infection without having to go to the doctor.

References

1. Yoo, Tae Keun, Joon Yul Choi, Yeon II Jang, Ein Oh and Ik Hee Ryu. “Towards automated severe pharyngitis detection with smartphone using deep learning networks.” Computers in Biology and Medicine 125 (2020): 103980 -103980.
2. Creswell, Antonia, et al. "Generative adversarial networks: An overview." IEEE Signal Processing Magazine 35.1 (2018): 53-65.
3. G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. “Densely connected convolutional networks”. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages2261–2269, 2017.
4. Kalyan, Katikapalli Subramanyam, Ajit Rajasekharan, and Sivanesan Sangeetha. "Ammus: A survey of transformer-based pretrained models in natural language processing." arXiv preprint arXiv:2108.05542 (2021).
5. Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929 (2020).
6. N. Sharma, V. Jain, and A. Mishra, “An analysis of convolutional neural networks for image classification,” Procedia Computer Science, vol. 132, pp. 377–384, 2018.
7. Dai, Zihang, et al. "Coatnet: Marrying convolution and attention for all data sizes." Advances in Neural Information Processing Systems 34 (2021).
8. Yoo, TaeKeun (2020), “Toward automated severe pharyngitis detection with smartphone camera using deep learning networks”, Mendeley Data, V2, doi: 10.17632/8ynyhnj2kz.2.
9. Zhu, Jun-Yan, et al. "Unpaired image-to-image translation using cycle-consistent adversarial networks." Proceedings of the IEEE international conference on computer vision. 2017.
10. Tobias, Rogelio Ruzcko & De Jesus, Luigi Carlo & Mital, Matt Ervin & Lauguico, Sandy & Bandala, Argel & Vicerra, Ryan & Dadios, Elmer. (2019). “Throat Detection and Health Classification Using Neural Network”. 38-43. 10.1109/IC3I46837.2019.9055535.
11. Askarian, Behnam & Yoo, Seung-Chul & Chong, Jo. (2019). “Novel Image Processing Method for Detecting Strep Throat (Streptococcal Pharyngitis) Using Smartphone”. Sensors. 19. 3306. 10.3390/s19153307.
12. Deng, Jia, et al. "Imagenet: A large-scale hierarchical image database." 2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009.
13. Gholamalinezhad, Hossein, and Hossein Khosravi. "Pooling methods in deep neural networks, a review." arXiv preprint arXiv:2009.07485 (2020).
14. Shaw, Peter, Jakob Uszkoreit, and Ashish Vaswani. "Self-attention with relative position representations." arXiv preprint arXiv:1803.02155 (2018).
15. Agarap, Abien Fred. "Deep learning using rectified linear units (relu)." arXiv preprint arXiv:1803.08375 (2018).
16. Yosinski J, Clune J, Bengio Y, and Lipson H. How transferable are features in deep neural networks? In Advances in Neural Information Processing Systems 27 (NIPS ’14), NIPS Foundation, 2014.
17. Reyes AK, Caicedo JC, Camargo JE. Fine-tuning Deep Convolutional Networks for Plant Recognition. CLEF (Working Notes). 2015 Sep 8;1391.
18. Wang, Meiqi, et al. "A high-speed and low-complexity architecture for softmax function in deep learning." 2018 IEEE Asia Pacific Conference on Circuits and Systems (APCCAS). IEEE, 2018.
19. Zhang, Zijun. "Improved adam optimizer for deep neural networks." 2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS). IEEE, 2018.
20. Ketkar, Nikhil. "Stochastic gradient descent." Deep learning with Python. Apress, Berkeley, CA, 2017. 113-132.