Fine-Tuning Deep Learning Architectures for Early Detection of Oral Cancer

Roshan Alex Welikala[1], Paolo Remagnino[1], Jian Han Lim[2], Chee Seng Chan[2], Senthilmani Rajendran[3], Thomas George Kallarakkal[2], Rosnah Binti Zain[2,4], Ruwan Duminda Jayasinghe[5], Jyotsna Rimal[6], Alexander Ross Kerr[7], Rahmi Amtha[8], Karthikeya Patil[9], Wanninayake Mudiyanselage Tilakaratne[2,5], John Gibson[10], Sok Ching Cheong[2,3] and Sarah Ann Barman[1]

[1] Kingston University, Surrey, KT1 2EE, United Kingdom
[2] University of Malaya, 50603 Kuala Lumpur, Malaysia
[3] Cancer Research Malaysia, 47500 Subang Jaya, Malaysia
[4] MAHSA University, Bandar Saujana Putra, 42610 Jenjarom, Malaysia
[5] University of Peradeniya, Peradeniya, 20400, Sri Lanka
[6] BP Koirala Institute of Health Sciences, Dharan, 56700, Nepal
[7] New York University, New York, NY 10010, USA
[8] Trisakti University, Kota Jakarta Barat, Jakarta 11440, Indonesia
[9] Jagadguru Sri Shivarathreeshwara University, Mysuru, 570 015 Karnataka, India
[10] University of Aberdeen, Aberdeen, AB25 2ZD, United Kingdom
r.welikala@kingston.ac.uk

Abstract. Oral cancer is most prevalent in low- and middle-income countries where it is associated with late diagnosis. A significant factor for this is the limited access to specialist diagnosis. The use of artificial intelligence for decision making on oral cavity images has the potential to improve cancer management and survival rates. This study forms part of the MeMoSA® (**M**obil**e** **Mo**uth **S**creening **A**nywhere) project. In this paper, we extended on our previous deep learning work and focused on the binary image classification of 'referral' vs. 'non-referral'. Transfer learning was applied, with several common pre-trained deep convolutional neural network architectures compared for the task of fine-tuning to a small oral image dataset. Improvements to our previous work were made, with an accuracy of 80.88% achieved and a corresponding sensitivity of 85.71% and specificity of 76.42%.

Keywords: Deep Learning, Oral Cancer, Oral Potentially Malignant Disorders.

1 Introduction

Oral cancer is one of the most common cancers worldwide, with an estimated 354,864 new cases and 177,384 deaths in 2018 [1]. The disease disproportionately affects low- and middle-income countries (LMICs). Oral cancer is typically associated with late diagnosis, particularly in LMICs, and as a result survival rates are low [2]. Significant factors associated with late diagnosis are poor awareness and the limited access to specialist diagnosis.

A major advantage is that oral cancer is often preceded by visible oral lesions termed as oral potentially malignant disorders (OPMDs) which can be detected from a clinical oral examination performed by a trained healthcare practitioner. Screening programs, if in place, offer early diagnosis and can lead to a reduction in mortality rates and morbidity. Telemedicine using images captured via mobile phones [3] would allow for remote consultation by specialists and may improve the referral accuracy of screening programs.

Artificial intelligence (AI) has the potential to classify images according to specific disease types or even provide descriptive summaries. However, achieving a high-level of performance for the binary classification of ‘referral’ vs. ‘non-referral’ would be the first step towards translation into clinical practice (following robust clinical evaluation). With a telemedicine approach, this would assist primary healthcare providers who may not be trained in identifying high-risk oral lesions in sending through only relevant cases to the specialists.

Recent methods related to the automated early detection of oral cancer made use of the convolutional neural network (CNN) which is a deep learning based AI technique designed for inputs in the form of images. Deep learning enables features to be automatically learnt at multiple levels of abstraction which allow complex patterns to be derived. Uthoff [4] used a CNN to classify pairs of autofluorescence and white light images as suspicious and not suspicious. Aubreville [5] used a CNN to classify laser endomicroscopy images as clinically normal and carcinogenic. Whilst custom CNN architectures can be built for a specific task, there are several popular architectures well known for achieving state-of-the-art performance on the ImageNet dataset [6] at their time of release. Among these are VGG [7], InceptionV3 [8], ResNet [9] and Xception [10].

Transfer learning is a technique where a model trained on one task is repurposed on a second related task. The biggest benefit of transfer learning shows when the target dataset is small, this is due to very large datasets being required to train deep learning models. It is common to use CNN architectures pre-trained on the ImageNet dataset which contains 1.2 million images with 1000 classes (e.g. tiger, pizza, speedboat). If a dataset is very small (e.g. < 1000 images) then best practice is to use a pre-trained CNN as a fixed feature extractor, if not as small (e.g. > 1000 images) then fine-tuning the CNN can produce superior results. Due to overfitting concerns with small datasets, it is advisable to keep the initial layers frozen (which capture universal low-level features such as edges, curves and blobs) and only fine-tune the latter part of the network.

Our previous work [11] focused on using ResNet to tackle early detection of oral cancer. ResNet was used to explore image classification and object detection, along with classifying according to different levels of disease categorization. In this paper, we provide a short extension to this work, focused on the binary image classification of ‘referral’ vs. ‘non-referral’. We compared the performance of some common pre-trained CNN architectures (VGG, InceptionV3 and ResNet) when applied to our oral image dataset, whilst exploring issues of fine-tuning with respect to a small dataset.

2 Materials

This study forms part of the MeMoSA$^{®}$ (**M**obil**e** **Mo**uth **S**creening **A**nywhere) project [3], in which images are currently in the process of being gathered and annotated from clinical experts from across the world. At this initial phase of the project, the number of annotated oral cavity images stands at 2155.

From this dataset, 1180 images were of class 'non-referral' and 975 images were of class 'referral'. The 'non-referral' class comprised of a mixture of images without lesions and images with lesions but not requiring referral. The 'referral' class comprised of images with lesions that required referral for low risk OPMD, high risk OPMD, cancer and other reasons. The images were of varying size, the largest was 5472 x 3648 pixels and the smallest was 119 x 142 pixels. The dataset was split into training, validation and test sets as detailed in Table 1. Further details on the dataset can be found in [11].

Table 1. Image numbers according to the class label and dataset type.

Class	Training	Validation	Test	Total
Non-referral	949	125	106	1180
Referral	795	82	98	975

3 Method

Five different CNN architectures were trained on our dataset for the binary image classification of 'referral' vs. 'non-referral'. The softmax layer with a 1000 outputs was changed to two outputs to represent the two classes (equivalent to sigmoid function). The training involved freezing the initial part of the networks and fine-tuning the latter part of the networks, which included the convolutional layers responsible for high-level features. These architectures are detailed in Table 2; the stated top1/top5 accuracies were reported by Keras [12] for performance of the ImageNet dataset.

Table 2. CNN architectures.

Architecture	Description	Top-1	Top-5
VGG-16	13 convolutional and 3 fully-connected (FC) layers (including the softmax layer). Its novelty was to go deeper.	71.3%	90.1%
VGG-19	A deeper variant to VGG-16.	71.3%	90.0%
Inception-V3	48 layers with no FC layers except for the softmax layer. Its novelty was the concatenation of feature maps generated by filters of multiple sizes. Among the first to use batch normalization.	77.9%	93.7%
ResNet-50	50 layers with no FC layers except for the softmax layer. Its novelty was to popularize skip connections with residual blocks to combat training issues associated with very deep networks. Among the first to use batch normalization.	74.9%	92.1%
ResNet-101	A deeper variant to ResNet-50.	76.4%	92.8%

3.1 Technical Details

Backpropagation and stochastic gradient descent (SGD) with momentum of 0.9 was used for training. Images were rescaled to 224 x 224 pixels, except for InceptionV3 which used 299 x 299 pixels. The training data was augmented with horizontal/vertical flipping, scaling, translation and rotation.

SGD mini-batch size was 128 images. A weighted loss function was used to correct for the slight class imbalance in the training data. The models were initialized with pre-trained ImageNet weights and fine-tuned from the second last convolutional layer for VGG, from second last Inception block for InceptionV3 and from conv4_1 for ResNet. The training strategy varied based on the architecture, e.g. VGG-19 was trained for 100 epochs at a learning rate of 0.001. Varying levels of weight decay were used for regularization. The models were built on the training set and hyperparameters were derived from performance on the validation set.

A Nvidia GeForce RTX 2080 Ti graphics card with 11GB memory was used for training. This implementation used Keras and TensorFlow.

3.2 Batch Normalization for Transfer Learning

Batch Normalization (BN) targets the vanishing gradient problem by standardizing the output of the previous layer, it speeds up the training process and it enables deeper networks to be trained. During training BN uses the mean and variance of the current mini-batch to normalize, and during inference BN uses fixed batch statistics derived from the moving mean and variance that was estimated during training.

BN works well when fine-tuning the entire network. But when part of the network is frozen (due to limited data) the behavior of BN can cause discrepancies between training and inference. Consider the frozen part of the network; for training BN uses the current mini-batch statistics and for inference BN uses fixed batch statistics derived from the original dataset. This works well if the data is from the same/similar domain as ImageNet, but leads to poor results if the domain is different (i.e. oral cancer). This was rectified when BN in the frozen part of the network was set to use moving mean and variance that was estimated during training for both training and inference.

An additional issue, when the dataset is small and the domain is different, is to achieve representative fixed batch statistics (used for inference) for the data. Training for long enough resolves this, but this is problematic with limited data. We find a BN momentum value of 0.9 helped towards achieving better statistics.

These issues affected the IncpetionV3 and ResNet models which used BN throughout their architecture.

4 Results

Evaluation was performed on the test set. As the classes were approximately balanced in the test set, we used accuracy as a single performance metric to compare the architectures (as detailed in Table 3). For each architecture, a confidence score threshold that produced the best operating point defined by the accuracy was selected. The best performing architecture was VGG19 with an accuracy of 80.88%. This corresponds to a sensitivity of 85.71% and a specificity of 76.42%, with further metrics detailed in Table 4. Examples of outputs from VGG-19 are provided in Fig. 1.

Table 3. Image classification results for several CNN architectures.

Architecture	Accuracy (%)	Architecture	Accuracy (%)
VGG-16	80.39	ResNet-50	74.51
VGG-19	80.88	ResNet-101	76.96
Inception-V3	76.47		

Table 4. Further metrics on the image classification results for VGG-19.

Performance metric	(%)	Performance metric	(%)
Sensitivity	85.71	False positive rate	23.58
Specificity	76.42	Precision	77.06
Positive predictive value	77.06	Recall	85.71
Negative predictive value	85.26	F_1 score	81.16
False negative rate	14.29	Accuracy	80.88

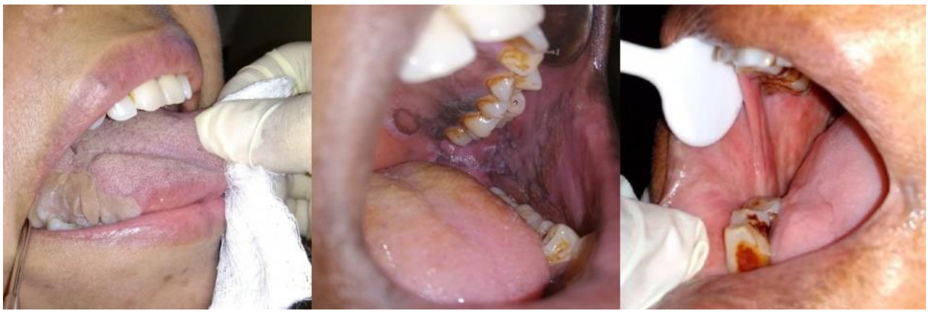

Fig. 1. Examples of results. Left: correctly classified as ‘non-referral’ with a class probability of 0.82. Middle: correctly classified as ‘referral’ with a class probability of 0.85. Right: incorrectly classified as ‘non-referral’ with a class probability of 0.83.

5 Discussion and Conclusion

In this paper, we demonstrate the performance of deep learning based systems for the image classification of ‘referral’ vs. ‘non-referral’ with respect to oral cancer. The best performing model achieved a sensitivity of 85.71% and a specificity of 76.42% for the identification of images that required referral. An accuracy of 80.88% and F_1 score of 81.16 %; it surpassed the F_1 score of 78.30% reported in previous work [11].

After exploring several CNN architectures, we demonstrate that the VGG architectures produced superior results for our dataset, with the VGG-19 coming out on top. Despite InceptionV3 and ResNet being more complex and deeper networks, they were surpassed by the easier to fine-tune VGG models. Aside from overfitting issues with deeper networks (although VGG does contain FC layers which can also cause overfitting); InceptionV3 and ResNet present batch normalization issues when being used for fine-tuning on small datasets of a different domain. Therefore, the VGG models currently provide a more stable and reliable approach, better showing the potential of AI. However, the other architectures offer more potential to learn complex patterns and will be used to produce superior results when our dataset is larger

Our future scope is to pre-train our models on datasets from a similar domain as our data (e.g. skin cancer), this is likely to improve results. But the most important target is to build a large dataset as this is key to deep learning. This will enable fine-tuning of the entire network, or even training the architectures from scratch and training custom made architectures. We plan to focus on the interpretability of our models to support clinical confidence in AI decision making, briefly covered in the appendix.

In conclusion, we have shown potential for AI to be incorporated into a mobile phone based telemedicine approach for the early detection of oral cancer. These promising early results are set to improve as the MeMoSA® project continues and the dataset grows.

Acknowledgments. We would like to thank the Medical Research Council for providing funding (MR/S013865/1).

Appendix

Fig. 2-3 provide gradient-weighted class activation mappings (Grad-CAM) [13] to demonstrate where the VGG-19 model was looking. The model appears to approximately focus on the lesion when making the decision of ‘referral’. We feel our models could benefit from using a trainable attention mechanism [14].

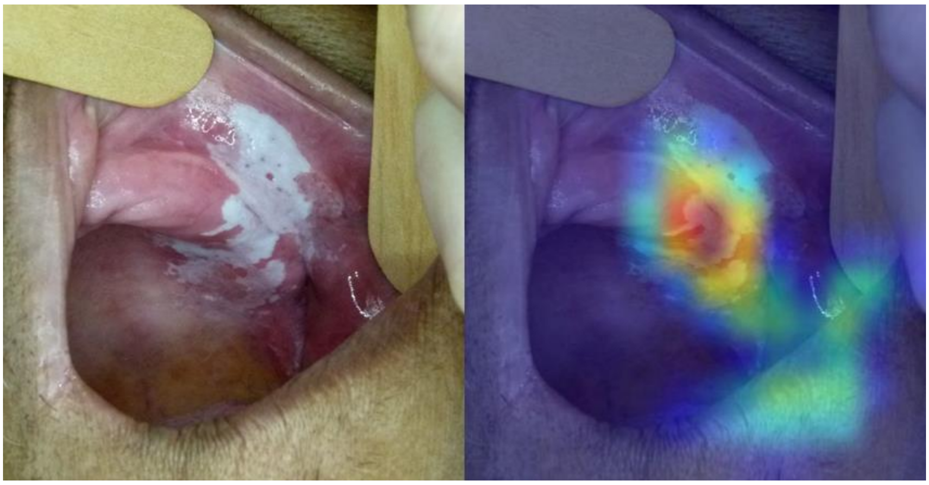

Fig. 2. Correctly classified as ‘referral’ with a class probability of 0.787. Left: original image. Right: Grad-CAM for class ‘referral’.

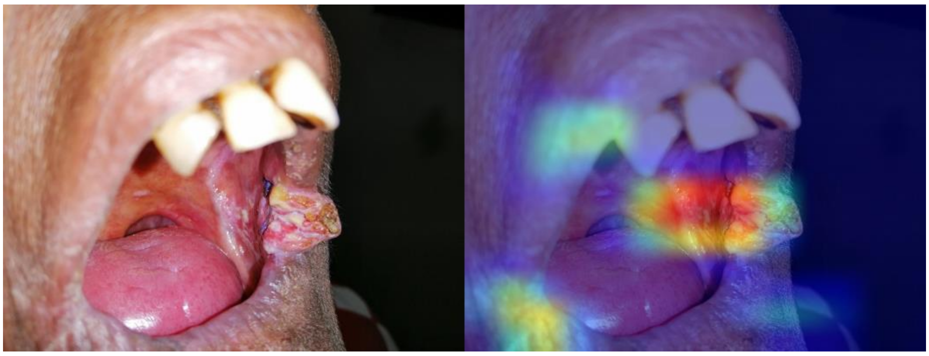

Fig. 3. Correctly classified as ‘referral’ with a class probability of 0.791. Left: original image. Right: Grad-CAM for class ‘referral’.

References

1. Bray, F., et al.: Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians 68(6), 394–424 (2018).
2. Doss, J.G., et al.: Validity of the FACT-H&N (v 4.0) among Malaysian oral cancer patients. Oral oncology 47(7), 648–652 (2011).
3. Haron, N., et al.: m-Health for Early Detection of Oral Cancer in Low-and Middle-Income Countries. Telemedicine and e-Health 26(3), 278-285 (2020).
4. Uthoff, R.D., et al.: Point-of-care, smartphone-based, dual-modality, dual-view, oral cancer screening device with neural network classification for low-resource communities. PLoS One 13(12), e0207493 (2018).
5. Aubreville, M., et al.: Automatic classification of cancerous tissue in laserendomicroscopy images of the oral cavity using deep learning. Scientific reports 7(1), 1-10 (2017).
6. Deng, J., et al.: Imagenet: A large-scale hierarchical image database. IEEE conference on computer vision and pattern recognition, 248–255 (2009).
7. Simonyan, K., et al.: Very deep convolutional networks for large-scale image recognition. arXiv Prepr., arXiv1409.1556 (2014).
8. Szegedy, C., et al.: Rethinking the inception architecture for computer vision. IEEE conference on Computer Vision and Pattern Recognition, 2818-2826 (2016).
9. He, K., et al.: Deep residual learning for image recognition. IEEE conference on Computer Vision and Pattern Recognition, 770–778 (2016).
10. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. IEEE conference on Computer Vision and Pattern Recognition, 1251-1258 (2017).
11. Welikala, R., et al.: Automated detection and classification of oral lesions using deep learning for early detection of oral cancer. IEEE Access 8, 132677-132693 (2020).
12. Keras, https://keras.io/api/applications/, last accessed 2020/06/22.
13. Selvaraju, R.R., et al.: Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. International Journal of Computer Vision 128(2), 336-359 (2020).
14. Jetley, S., et al.: Learn to pay attention. arXiv preprint., arXiv:1804.02391 (2018).