Fundamentals of Deep Learning

(CSE477s)

Project

Team 4

Abdelrahman Ezz Eldin Ismail	2101000
Abdulrahman Ahmed Saeed	2100811
Omar Ashraf Abdelsattar	2100354

1. What is the problem statement of the paper?

The paper addresses the **challenge of early and cost-effective detection of oral cancer**, particularly in **low- and middle-income countries (LMICs)** where access to specialists and healthcare resources is limited. It focuses on developing **automated systems for detecting and classifying oral lesions**—specifically oral potentially malignant disorders (OPMDs) using deep learning techniques applied to **mobile phone images**, enabling scalable screening solutions.

2. What are the objectives of the paper, and do you think the authors managed to achieve these goals? Explain

Objectives:

- To develop and assess **deep learning-based systems** (image classification and object detection) for **automated detection and classification of oral lesions**.
- To create a large dataset with **composite annotations** from multiple clinicians.
- To demonstrate the **feasibility of using mobile phone images** and AI for early diagnosis in resource-limited settings.

Did they achieve these?
Yes, **partially**:

- They successfully developed two deep learning models (ResNet-101 and Faster R-CNN) and evaluated them on an initial dataset of 2,155 images.
- They introduced a **novel composite annotation strategy**, which enhanced the data quality by resolving clinician disagreements.
- While the results showed promising F1 scores **(87.07% for lesion classification and 78.30% for referral classification)**, the performance of object detection was lower **(F1 = 41.18%)**, indicating that further refinement and larger datasets are needed for robust deployment.
- Overall, the study demonstrated **proof-of-concept** rather than a fully deployable solution.

3. What is the DL method used in this paper?

The paper used two deep learning methods:

- **Image Classification**: Implemented using **ResNet-101**, a deep convolutional neural network with 101 layers they made use of this architecture due to its widespread use and high reported performances, Transfer learning which is used is a machine learning technique where a model trained on one task is re-purposed on a second related task. Transfer learning is popular in deep learning given the enormous resources required to train deep learning models, and the model used is pre-trained on ImageNet and fine-tuned on the oral lesion dataset.
- **Object Detection**: Implemented using **Faster R-CNN**, The Faster R-CNN was a two-stage approach. The RST stage was the region proposal network (RPN) which generated a sparse set of object/region proposals each with an objectness score. The second stage is known as the detection network

which classified the region proposals into object classes and background. Both networks shared a common set of convolution layers. These common layers form the backbone/base of the framework which was a CNN (can be referred to as the base CNN), whose output from an intermediate convolutional layer provided rich hierarchical features for the input image, with ResNet-101 as the backbone and Feature Pyramid Networks (FPN) for enhanced feature extraction, also pre-trained and fine-tuned.

4. What are the other state-of-the-art methods that can be applied to the same problem?

Several state-of-the-art methods have been applied or could be applied to the problem of detecting and classifying oral lesions, including:

- **Mask R-CNN**: Used by Anantharaman et al. for instance segmentation of oral lesions like cold sores and canker sores.
- **YOLO (You Only Look Once)** and **SSD (Single Shot MultiBox Detector)**: One-stage object detectors that offer faster inference at the cost of some accuracy, and are viable for mobile and real-time applications.
- **VGG-based CNNs**: Used by Uthoff et al. for classifying autofluorescence and white-light oral images as "suspicious" or "not suspicious" with high sensitivity and specificity.
- **Attention-based CNNs**: Recent developments in attention mechanisms (e.g., Residual Attention Network, Show-Attend-and-Tell models) can potentially help focus on relevant parts of the image for better lesion detection and classification.

These methods offer alternatives or improvements in accuracy, speed, or interpretability, depending on the task.

5. Would you apply any of the other methods other than the DL method used in this paper? Explain your answer.

Yes, due to limited access to high-end computational resources, we explored and implemented alternative methods. Instead of ResNet-101 and Faster R-CNN, we used:

- A **custom CNN model** for image classification of oral lesions, which was lightweight and suitable for our hardware.
- **YOLO (You Only Look Once)** for object detection, which is more efficient and offers real-time performance even on modest GPUs.

These alternatives enabled us to implement the key ideas from the paper while staying within resource limits. YOLO offered a good trade-off between speed and accuracy, making it a viable choice for mobile health applications.

6. What datasets have been used in this paper? Do you think the result is generalizable for any datasets?
The authors used a custom dataset consisting of **2,155** mobile phone images of the oral cavity, collected from 396 participants in Tamil Nadu, India. The images were annotated using a composite label generated from multiple clinicians' inputs to address inter-observer variability.

In our implementation, we used a publicly available Kaggle dataset titled *Oral Cancer Images for Classification*, which originally contained **1,238 images**. We removed **40 corrupted images**, resulting in a working dataset of **1,198 images**.

While the results are promising, generalizability is limited because:

The original paper's dataset and the Kaggle dataset may not fully represent global diversity in lesion appearance due to

- imaging conditions like Lighting, camera quality, and image acquisition protocols may vary across different populations and devices.
- region, ethnicity, cultural, genetic, and environmental factors might influence lesion appearance.

Our model showed good performance on the Kaggle dataset, but further validation on broader, multi-center datasets would be necessary to confirm generalizability.

7. Discuss the results presented in the paper. Compare the results with other state-of-the-art methods used to solve this problem.
Results:

- **ResNet-101 classification model** achieved **F1 scores of 87.07%** (lesion classification) and **78.30%** (referral classification).
- **Faster R-CNN object detection** yielded a lower F1 score of **41.18%**, suggesting challenges in precise lesion localization.
- Composite annotation helped improve classification performance by addressing inter-observer variability.

Comparison:

- Uthoff et al. (2019) used a VGG-based CNN for binary classification of autofluorescence images, achieving high sensitivity and specificity.
- Anantharaman et al. used **Mask R-CNN** for instance segmentation of oral lesions with better localization.
- YOLO and SSD models, although not evaluated in this paper, are known to offer faster and reasonably accurate detection in mobile and embedded environments.

Overall, the classification performance in this study is competitive, but the detection model underperforms compared to state-of-the-art segmentation methods like Mask R-CNN.

8. What would you like to criticize about the paper? Could you suggest any improvements?
Criticisms:

- The dataset is limited in size and diversity, affecting the robustness and generalizability of the model.
- The object detection model's performance (F1 = 41.18%) is relatively low, suggesting inadequate localization capability.
- There is no ablation study showing the effect of different backbone networks or preprocessing techniques.
- The paper lacks discussion on model explainability, which is crucial in medical AI applications.

Suggestions for improvement:

- Include larger and more diverse datasets, possibly via collaboration with other hospitals or institutions.
- Experiment with alternative architectures (e.g., EfficientDet, YOLOv5, or transformer-based vision models like DETR).
- Incorporate explainability tools (e.g., Grad-CAM, LIME) to provide visual evidence supporting the model's predictions.
- Conduct prospective validation or a pilot clinical study to test the model in real-world scenarios.

9. Have you implemented the paper using your own code? Do your results agree with the authors? What are the differences and why?
Yes, we attempted to implement the paper using our own code, but we encountered **significant computational limitations** that influenced our approach.

What we tried:

- We initially attempted to use **ResNet-101**, as done in the original paper, with pre-trained weights.
- When ResNet-101 proved too resource-intensive for our hardware, we attempted a lighter alternative using **ResNet-50**, also with pre-trained ImageNet weights.
- Unfortunately, even ResNet-50 was too computationally demanding for effective training and evaluation on our systems.

What we did instead:

- We designed and trained a **custom CNN model** for image classification of oral lesions.
- For object detection, we did use **Faster R-CNN**, just like in the paper. However, we trained it on the Kaggle dataset.
- Annotation for object detection was done manually using **VIA (VGG Image Annotator)**.
- The dataset of the paper was also private, so we used the **Kaggle Oral Cancer Images for Classification** dataset. Out of 1,238 images, 40 corrupted ones were removed, and the remaining 1,198 were used for training and testing.

Results:

The paper’s classification F1 score **(87.07%)** is slightly higher than ours **(83%)**, which is expected since they used ResNet-101, a much deeper fine-tuned model with a larger dataset.

Our implementation of **Faster R-CNN** significantly outperformed the paper's object detection results. We achieved an F1 score of **71.79%**, compared to the paper's **41.18%**. This likely stems from:

- Our cleaner and more consistent dataset (Kaggle images).
- Better quality annotations using **VIA (VGG Image Annotator)**.
- Possibly improved hyperparameter tuning or fewer ambiguous lesion samples.
- Our **Mean IoU of 76.37%** indicates good localization accuracy, reinforcing the model's utility for real-world lesion detection.

Conclusion:

While we could not use high-capacity models like ResNet-101 due to hardware limitations, our custom CNN and Faster R-CNN implementation still yielded competitive—and in the case of object detection, superior—results. This suggests that:

- The feasibility and effectiveness of deep learning for oral lesion screening is reinforced.
- Lighter, more optimized approaches can be just as valuable in real-world applications, especially in resource-constrained environments like rural clinics.