TRANSFER LEARNING FOR ENDOSCOPY DISEASE DETECTION AND SEGMENTATION WITH MASK-RCNN BENCHMARK ARCHITECTURE

Shahadate Rezvy[1,4]*, Tahmina Zebin*[2]*, Barbara Braden*[3]*, Wei Pang*[4]*, Stephen Taylor*[5]*, Xiaohong W Gao*[1]

[1]School of Science and Technology, Middlesex University London, UK
[2]School of Computing Sciences, University of East Anglia, UK
[3]Translational Gastroenterology Unit, John Radcliffe Hospital, University of Oxford, UK
[4]School of Mathematical & Computer Sciences, Heriot-Watt University, UK
[5]MRC Weatherall Institute of Molecular Medicine, University of Oxford, UK

ABSTRACT

We proposed and implemented a disease detection and semantic segmentation pipeline using a modified mask-RCNN infrastructure model on the EDD2020 dataset[1]. On the images provided for the phase-I test dataset, for 'BE', we achieved an average precision of 51.14%, for 'HGD' and 'polyp' it is 50%. However, the detection score for 'suspicious' and 'cancer' were low. For phase-I, we achieved a dice coefficient of 0.4562 and an F2 score of 0.4508. We noticed the missed and mis-classification was due to the imbalance between classes. Hence, we applied a selective and balanced augmentation stage in our architecture to provide more accurate detection and segmentation. We observed an increase in detection score to 0.29 on phase -II images after balancing the dataset from our phase-I detection score of 0.24. We achieved an improved semantic segmentation score of 0.62 from our phase-I score of 0.52.

1. INTRODUCTION

Endoscopy is an extensively used clinical procedure for the early detection of cancers in various organs such as esophagus, stomach, colon, and bladder [1]. In recent years, deep learning methods were used in various endoscopic imaging tasks including esophago-gastro-duodenoscopy (EGD), colonoscopy, and capsule endoscopy (CE) [2]. Most of these were inspired by artificial neural network-based solutions for accurate and consistent localization and segmentation of diseased region-of-interests enable precise quantification and mapping of lesions from clinical endoscopy videos. This enables critical and useful detection techniques for monitoring and surgical planning.

For oesophageal cancer detection, Mendel *et al.* [3] proposed an automatic approach for early detection of adenocarcinoma in the esophagus by using high-definition endoscopic images (50 cancer, 50 Barrett). They adapted and fed the data set to a deep Convolutional Neural Network (CNN) using a transfer learning approach. The model was evaluated to leave one patient out cross-validation. With sensitivity and specificity of 0.94 and 0.88, respectively. Horie *et al.* [4] reported AI diagnoses of esophageal cancer including squamous cell carcinoma (ESCC) and adenocarcinoma (EAC) using CNNs. The CNN correctly detected esophageal cancer cases with a sensitivity of 98%. CNN could detect all small cancer lesions less than 10 mm in size. It has reportedly distinguished superficial esophageal cancer from advanced cancer with an accuracy of 98%. Very recently, Gao *et al.* [5] investigated the feasibility of mask-RCNN (Region-based convolutional neural network) and YOLOv3 architectures to detect various stages of squamous cell carcinoma (SCC) cancer in real-time to detect subtle appearance changes. For the detection of SCC, the reported average accuracy for classification and detection was 85% and 74% respectively.

For colonoscopy, deep neural networks based solutions were implemented to detect and classify colorectal polyps in research presented by the authors in reference [6, 7, 8]. For gastric cancer, Wu *et al.* [9] identified EGC from non-malignancy with an accuracy of 92.5%, a sensitivity of 94.0%, a specificity of 91.0%, a positive predictive value of 91.3%, and a negative predictive value of 93.8%, outperforming all levels of endoscopists. In real-time unprocessed EGD videos, the DCNN achieved automated performance for detecting EGC and monitoring blind spots. Mori *et al.* [10] and Min *et al.* [2] provided a comprehensive review of some recent literature in this field.

For Endoscopy Disease Detection and Segmentation Grand Challenge, we proposed and implemented a disease detection and semantic segmentation pipeline using a modified mask-RCNN architecture. The rest of the paper is organized as follows. Section 2 introduces the dataset for the task. Section 3 presents our proposed architecture with various settings and procedural stages, with results presented and discussed in Section 4. Finally, conclusions are drawn in Section 5.

[1]https://edd2020.grand-challenge.org

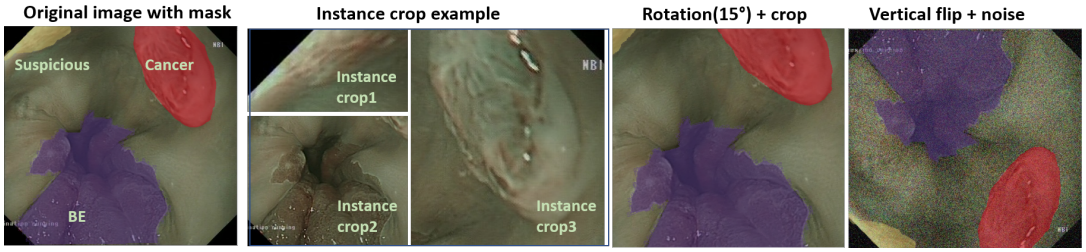

Fig. 1. Augmentation methods applied on the images including transformation such as rotation, flip and instance cropping.

Table 1. Class-wise object distribution [1]

Disease Category (Class name)	Objects
Non-dysplastic Barrett's oesophagus (BE)	160
Subtle pre-cancerous lesion (Suspicious)	88
Suspected Dysplasia (HGD)	74
Adenocarcinoma (Cancer)	53
Polyp	127

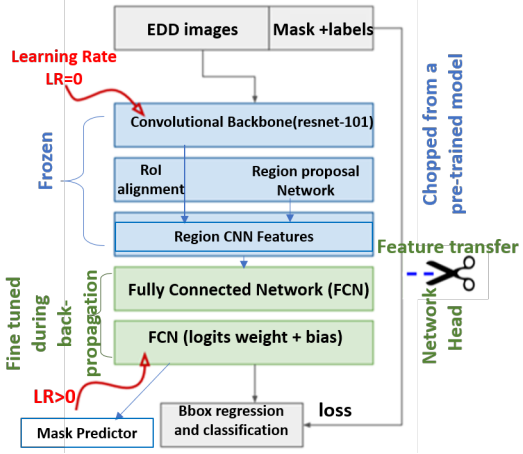

Fig. 2. Illustration of the mask-RCNN architecture adapted for transfer learning on the EDD dataset

2. DATASET DESCRIPTION AND IMAGE AUGMENTATION

The annotated dataset provided for the competition contained 388 frames from 5 different international centers and 3 organs (colon, esophagus, and stomach) targeting multiple populations and varied endoscopy video modalities associated with pre-malignant and diseased regions. The dataset is labeled by medical experts and experienced post-doctoral researchers. It came with object-wise binary masks and bounding box annotation. The class-wise object distribution in the dataset is shown in Table 1. A detailed description of the dataset can be found at [1].

We separated a small subset from the original training set with various class labels as our external validation set. This subset had 25 images, and was programmatically chosen to have similar size and resolution as the images in phase-I test dataset of 24 images. This set with ground truth labels served as a checkpoint for us to the trained model's performance.

We applied image augmentation techniques [11] on the rest of the images with their associated masks. Our observation of the dataset revealed a co-location of 'BE' regions with 'suspicious, cancer and HGD' area. We also noticed an imbalance between classes and images coming from various organs. Hence, we opted for an instance cropping stage in our pipeline that produced multiple images from these co-located images, each with one target object and other objects are removed by a selective cropping mechanism (example shown on Figure 1). We kept 10% padding around the ground truth bounding box provided for the instance. This isolated the instances of 'cancer', 'suspicious' and 'HGD' regions from co-localized 'BE' regions. We applied transformations such as rotation, flip and crop on the individual classes and instances to increase our training data. We then used the 'WeightedRandomSampler' from the PyTorch data loader to form the final balanced training set of almost equal class representation. This set included 1670 instances in total. Figure 1 illustrates some of the augmentation methods we applied in our pipeline.

3. METHODS

We implemented the Endoscopic disease detection and semantic segmentation pipeline for the EDD2020 challenge using a modified mask-RCNN [12] architecture trained in the feature-representation transfer learning mode. Mask-RCNN was proposed as an extension of Faster R-CNN and the architecture has reportedly outperformed all the previous state-of-the-art models used for the instance segmentation task on various image datasets. We used PyTorch, torchvision, imgaug, pycoco-creator, maskrcnn-benchmark [13], apex, and OpenCV libraries in python for generating various functions of the pipeline.

3.1. Pre-trained model backbone and network head removal

We removed the network head or the final layers of the pre-trained model with a Resnet-101 backbone [12] that was initially trained on the COCO dataset. This stage is crucial as the pre-trained model was trained for a different classification task. The removal of network head removed weights and bias associated to class score, bounding box predictor and mask

predictor layers. It is then replaced with new untrained layers with desired number of classes for the new data. We adjusted a six-class network head for the EDD2020 dataset (five assigned classes+ Background). We fed the augmented dataset and and the associated masks into the mask-RCNN model architecture as illustrated in figure 2.

3.2. Ttransfer learning stages

At the initial stage, we froze the weights of the earlier layers of the pre-trained Resnet-101 backbone to help us extract the generic low-level descriptors or patterns from the endoscopy image data. Later layers of the CNN become progressively more specific to the details of the output classes of the new data-set. Then a newly added network head is trained for adapting the weights according to the patterns and distribution of the new dataset. The network head is updated and fine tuned during model training. The training of the model has been done offline on an Ubuntu machine with Intel(R) Core i9-9900X CPU @ 3.50GHz, 62GB memory and a GeForce RTX 2060 GPU. The final model was fine- tuned with an Adam optimizer with a learning rate of 0.0001 and a categorical cross-entropy for 50000 epochs. To be noted, the dataset after augmentation is still quite small, so we employed a five-fold cross-validation during training to avoid the over-fitting of the model.

4. RESULTS AND EVALUATION SCORE

Equations (1) to (3) in this section summarises the detection and segmentation matrices we are using to evaluate the performance of a model trained on this dataset [1]. The metric, mean average precision (mAP) measures the ability of an object detector to accurately retrieve all instances of the ground truth bounding boxes. The higher the mAP the better the performance. In Equation (1), N = 5 and AP_i indicates Average precision of individual disease class i for this dataset.

$$mAP = \frac{1}{N} \sum_i AP_i \quad (1)$$

$$score_d = 0.6 \times mAP_d + 0.4 \times IoU_d \quad (2)$$

$$score_s = 0.25 * \sum_i precision + recall + F_1 + F_2 \quad (3)$$

For the detection task, the competition uses a a final mean score ($score_d$), which is a weighted score of mAP and IoU and formula is presented in Equation (2). Here, IoU - intersection over union measures the overlap between the ground truth and predicted bounding boxes. For scoring of the semantic segmentation task, an average measure ($score_s$) is calculated as per Equation (3), which is the average score of F_1-score (Dice Coefficient), F_2-score, precision and recall. A detail description of these matrices can be found in [1].

Table 2. Validation set bounding-box detection and segmentation score before and after fine-tuning

Fine-tuning	Task	mAP	$AP(50), AP(75)$	$AP(m), AP(l)$
No	bbox	0.291	0.361; 0.319	0.450; 0.328
No	segment	0.254	0.347; 0.252	0.250; 0.292
Yes	bbox	0.479	0.689; 0.600	0.675; 0.493
Yes	segment	0.513	0.683; 0.549	0.563; 0.566

Table 3. Out-of-sample detection and segmentation score

Training Dataset	(Test data)	$score_d$	$score_s$
Original+ Flip, rotate, crop	(Phase-I)	0.2460	0.5243
Original+Instance-crop+class-balance	(Phase-II)	0.2906	0.6264

4.1. Results on validation dataset

Table 2 summarises average precision performances on the isolated validation dataset (25 images with ground-truth masks) to get an estimate of the test set performance. Class-wise precision values were presented for two IoU thresholds. For $AP(50)$, only candidates over 50% region comparing ground truth were counted and we achieved about 36.1% average precision for bounding box detection and 34.7% average precision for pixel-to-pixel segmentation. For $AP(75)$, only the candidates over 75% IoU value are counted. Average precision values were counted for large ($AP(l)$) and medium-sized ($AP(m)$) objects in the images and the accuracy ranged from 32.27% to 45% respectively. To be noted, we omitted $AP(s)$ for small object (area $< 32pixel^2$) due to the absence of such small objects in the test dataset. However, such low values are indicative of the model being overfit and we applied parameter-tuning to the fully connected network layers along with realistic and balanced augmentation. This significantly improved the mAP for for both bounding box and segmentation mask to 47.9% and 51.3% respectively (shown in row 3 and 4 on Table 2).

4.2. Results on the test dataset: Phase-I and Phase-II

For phase-I, we received 24 images and Figure 3 shows detection and segmentation output from some of the images from this test set. From the scores available on the leaderboard, for 'BE', we achieved average precision value of 51.14%, for 'HGD' and 'polyp' it is 50%. However, the score for 'suspicious' and 'cancer' areas were very low. We attained a dice coefficient of 0.4562 and an F2 score of 0.4508. We noticed the missed and mis-classification was due to the imbalance between classes. Hence, before phase-II submission, we retrained the model after applying a 'WeightedRandomSampler' for selective and balanced sampling of the augmented dataset. During phase-II, we received 43 images and we

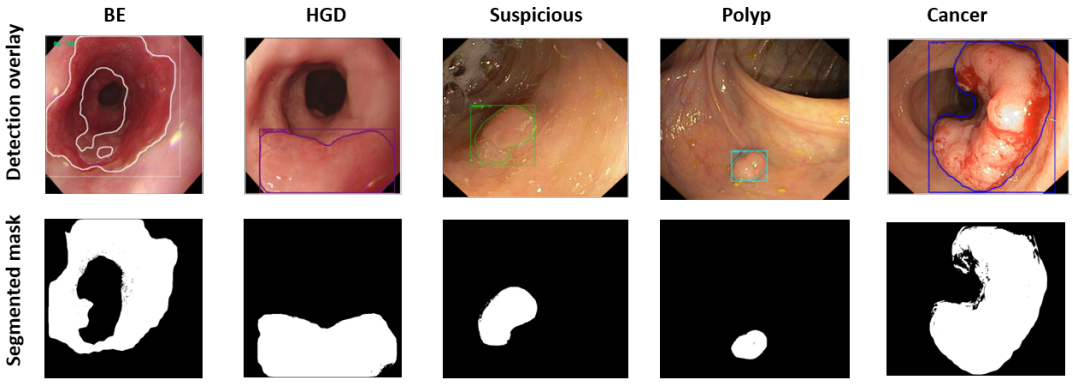

Fig. 3. Semantic segmentation results on some of the images from the test dataset

retrained the model with a balanced augmentation dataset. From the leader-board scores available at this stage, the final detection score $score_d$ and semantic segmentation score $score_s$ is listed in Table 3. In the table, we observed an increase in detection score to 0.29 when a class balancing and instance cropping is applied on the training dataset. We had a score of 0.24 on phase-I which we obtained with generic augmentation techniques applied on the data. We achieved an improved semantic segmentation score of 0.62 as well from our phase-I score 0f 0.52. The final model had an standard deviation of 0.082 in the mAP_d value and deviation was 0.33 in the semantic score.

5. DISCUSSION & CONCLUSION

As balanced augmentation has improved both detection and segmentation score in this task, application of generative adversarial network-based augmentation techniques in future can contribute to a more generalised and robust model. Additionally, we assumed that the detected object was spread uniformly across a detected region as the patch was classified as a specific disease type (cancer, polyp) depending on the patch-specific feature. However, the idea of one uniform region of cancer or polyp or BE is not always the case in practice. Very often, multifocal patches of cancer, low-grade and high-grade dysplasia are scattered across the surface of the lesion. Further improvements are required to deal with bubble, saturation, instrument and other visible artefacts in the dataset [14]. This will improve the model's performance by avoiding false detection in these regions and will provide more accurate and realistic solution for endoscopic disease detection cases.

6. REFERENCES

[1] Sharib Ali, Noha Ghatwary, Barbara Braden, Lamarque Dominique, Adam Bailey, Stefano Realdon, Cannizzaro Renato, Jens Rittscher, Christian Daul, and James East. Endoscopy disease detection challenge 2020. 2020.

[2] Jun Ki Min, Min Seob Kwak, and Jae Myung Cha. Overview of deep learning in gastrointestinal endoscopy. *Gut and liver*, 13(4):388, 2019.

[3] Robert Mendel, Alanna Ebigbo, Andreas Probst, et al. Barrett's esophagus analysis using convolutional neural networks. In *Image Processing for Medicine 2017*, pages 80–85. Springer, 2017.

[4] Yoshimasa Horie, Toshiyuki Yoshio, Kazuharu Aoyama, Yoshimizu, et al. Diagnostic outcomes of esophageal cancer by artificial intelligence using convolutional neural networks. *Gastrointestinal endoscopy*, 89(1):25–32, 2019.

[5] Xiaohong W Gao, Barbara Braden, Stephen Taylor, and Wei Pang. Towards real-time detection of squamous pre-cancers from oesophageal endoscopic videos. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pages 1606–1612, Dec 2019.

[6] Yoriaki Komeda, Hisashi Handa, et al. Computer-aided diagnosis based on convolutional neural network system for colorectal polyp classification: preliminary experience. *Oncology*, 93:30–34, 2017.

[7] Teng Zhou, Guoqiang Han, Bing Nan Li, et al. Quantitative analysis of patients with celiac disease by video capsule endoscopy: A deep learning method. *Computers in biology and medicine*, 85:1–6, 2017.

[8] Lequan Yu, Hao Chen, Qi Dou, Jing Qin, and Pheng Ann Heng. Integrating online and offline three-dimensional deep learning for automated polyp detection in colonoscopy videos. *IEEE journal of biomedical and health informatics*, 21(1):65–75, 2016.

[9] Lianlian Wu, Wei Zhou, Xinyue Wan, Jun Zhang, et al. A deep neural network improves endoscopic detection of early gastric cancer without blind spots. *Endoscopy*, 51(06):522–531, 2019.

[10] Yuichi Mori, Tyler M Berzin, and Shin-ei Kudo. Artificial intelligence for early gastric cancer: early promise and the path ahead. *Gastrointestinal endoscopy*, 89(4):816–817, 2019.

[11] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):60, 2019.

[12] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[13] Francisco Massa and Ross Girshick. maskrcnn-benchmark: Fast, modular reference implementation of Instance Segmentation and Object Detection algorithms in PyTorch. https://github.com/facebookresearch/maskrcnn-benchmark, 2018.

[14] Sharib Ali, Felix Zhou, Christian Daul, Barbara Braden, Adam Bailey, Stefano Realdon, James East, Georges Wagnières, Victor Loschenov, Enrico Grisan, Walter Blondel, and Jens Rittscher. Endoscopy artifact detection (EAD 2019) challenge dataset. *CoRR*, abs/1905.03209, 2019.