ENDOSCOPIC ARTEFACT DETECTION WITH ENSEMBLE OF DEEP NEURAL NETWORKS AND FALSE POSITIVE ELIMINATION

Gorkem Polat, Deniz Sen, Alperen Inci, Alptekin Temizel

Graduate School of Informatics
Middle East Technical University, Ankara, Turkey
{gorkem.polat, deniz.sen_01, alperen.inci, atemizel}@metu.edu.tr

ABSTRACT

Video frames obtained through endoscopic examination can be corrupted by many artefacts. These artefacts adversely affect the diagnosis process and make the examination of the underlying tissue difficult for the professionals. In addition, detection of these artefacts is essential for further automated analysis of the images and high-quality frame restoration.
In this study, we propose an endoscopic artefact detection framework based on an ensemble of deep neural networks, class-agnostic non-maximum suppression, and false-positive elimination. We have used different ensemble techniques and combined both one-stage and two-stage networks to have a heterogeneous solution exploiting the distinctive properties of different approaches. Faster R-CNN, Cascade R-CNN, which are two-stage detector, and RetinaNet, which is single-stage detector, have been used as base models. The best results have been obtained using the consensus of their predictions, which were passed through class-agnostic non-maximum suppression, and false-positive elimination.

Index Terms— Endoscopic artefact detection, Faster R-CNN, Feature pyramid networks, RetinaNet

1. INTRODUCTION

Endoscopic imaging is a widely used clinical procedure to inspect hollow organs and collect tissue samples for further examination. However, video frames captured during endoscopic examination are corrupted by many artefacts due to several factors such as lighting and shape of the organ. In order to perform a detailed endoscopic procedure, it is required to detect and localize these artefacts. This is also an essential process for high-quality frame restoration and developing computer-assisted endoscopy tools.

There are many challenges in artefact detection in endoscopic images. Analysis of the dataset provided by EAD2020 Challenge [1], reveals two major problems. Firstly, there is a class imbalance problem. While artefacts such as specularity account for nearly 34% of all detections, instrument class accounts for only 1.7%. Three classes (specularity, artifact, and bubbles), in total, account for 82% of all bounding boxes. Secondly, there is a scale imbalance problem. There are various bounding boxes that cover almost the entire frame and various bounding boxes only have very few pixels. Hence, the parameters of the object detection algorithms should be chosen carefully in the light of these observations to detect both small and large objects. We adopted an approach based on an ensemble of object detectors. Despite being slower, we mainly focused on two-stage networks due to their ability to detect small and very close objects and used Faster R-CNN [2] and Cascade R-CNN [3]. In addition, we used a single-stage detector, RetinaNet [4], as a complementary model in the ensemble.

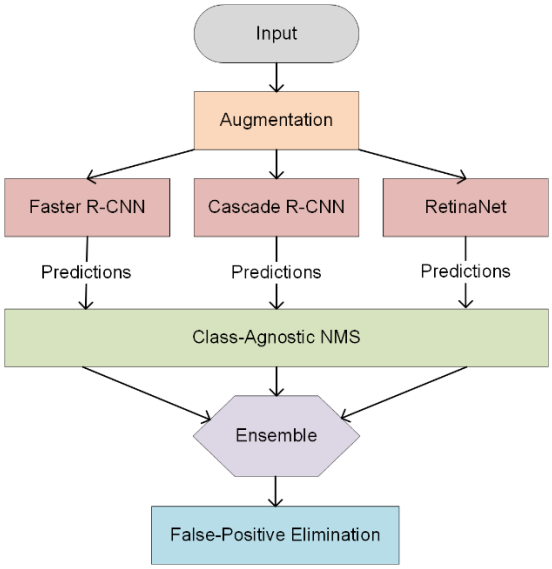

Figure 1: Flowchart of the proposed method.

2. PROPOSED METHOD

The flowchart of the proposed approach is given in Figure 1. We use three base models. The outputs of these base models are then fed into a class agnostic non-maximum suppression algorithm independently before combining the results through an ensemble model. Then a false-positive elimination is applied to the output of the ensemble model. In the remainder of this section, we describe these steps in more detail.

2.1. Base Models

We use two two-stage models: Faster R-CNN [2], Cascade R-CNN [3] and one single-stage model: RetinaNet [4] as base models. Examination of the previous studies in this

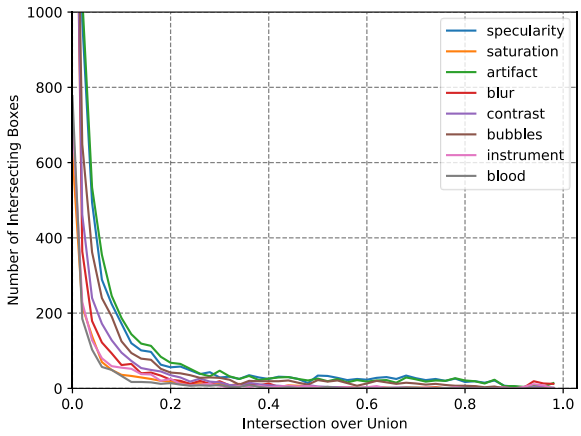

Figure 2: IoU histogram of each class with the other seven classes Vertical axis is clipped to provide better visualization.

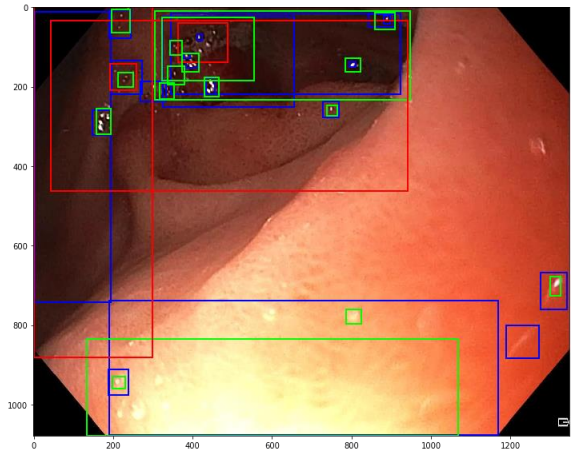

Figure 3: Blue: Ground-truth bounding boxes. Red: Bounding boxes eliminated after the FP reduction step. Green: Remaining predicted boxes after elimination.

domain reveals that feature pyramid network (FPN) [5] and ResNet [6] architectures achieve promising results [7]. Therefore, these networks have been selected as the basis for our models.

The first model is based on Faster R-CNN [2] and uses FPN as a backbone. Although FPNs are compute and memory intensive, they are good at extracting features at different scales. Since the dataset consists of objects in a wide variety of sizes, FPNs are an important element of the proposed network. We used a ResNet50 model with FPN as a backbone of this model. Standard convolutional and fully connected heads have been used for box predictions.

The second model is Cascade R-CNN [3]. While it is a similar model to Faster R-CNN, it is claimed to alleviate the problem of overfitting at training. Cascade R-CNN consists of consecutive detectors which are trained sequentially with increasing intersection-over-union (IoU) thresholds. This architecture is reported to be more selective against close false positives. Again, we used a ResNet50 model with FPN as a backbone.

In addition to these two-stage object detectors, we trained and used a RetinaNet [4] as our third model. RetinaNet is a single-stage method and, as such, does not use a region proposal network. It has one backbone network that extracts features and two subnetworks for object classification and bounding box regression. An important difference of this network from other single-stage networks (e.g. YOLO, SSD) is the use of focal loss. Focal loss is an extension to cross-entropy loss that puts a focus on sparse hard examples. It changes the weight of loss according to the performance of the model on different examples.

2.2. Class-Agnostic Non-Maximum Suppression (NMS)

In the original Faster R-CNN architecture, NMS operation is performed on each class independently. Yet, these architectures are generally designed considering non-medical datasets such as COCO [8] or PASCAL VOC [9], which have high overlap ratios among the bounding boxes of different classes.

However, it is not expected to have frequent overlaps between different objects in the endoscopic images. To validate this assertion, we calculated the IoU values for each class with the other classes. Figure 2 shows the IoU histogram of each class with the other seven classes. As seen in this figure, EAD Challenge dataset does not exhibit high number of overlaps between class bounding boxes. On the other hand, the original object detector predictions result in a high IoU between classes. Therefore, we propose a class-agnostic procedure where the model predictions are passed through the NMS process together for all classes. As a consequence of this process, if an artefact is detected by multiple models with high IoU, the ones having the lower confidence scores are eliminated. A threshold of 0.4 IoU has been used to perform this class-agnostic NMS step.

2.3. Ensemble of Models

Two different ensemble methods, affirmative and consensus, have been used [10]. In the affirmative method, the outputs of different models are merged, and NMS operation is applied on the result. It can be regarded as the union of all bounding boxes. In the consensus method, only the bounding boxes for which the majority of the models agree are kept. This method is similar to the ensemble of models in classification problems.

2.4. False-Positive (FP) Elimination

Although class-agnostic NMS discards the bounding boxes that have high IoU with other bounding boxes in the detector network, the IoU threshold (0.4) might be still too high for the same class types. For example, if the intersection of two bubble bounding boxes has very low probability but model predicts bounding boxes that have high IoU, it implies that there is redundancy and one of them should be removed. Therefore, we have examined the IoU histogram of each class

individually and determined a class-specific threshold. When there are bounding boxes with higher IoU values than the threshold, the ones having lower confidence scores are removed. Thresholds are determined according to the 1.5 interquartile range (IQR) above the 3rd quartile. Thresholds for elimination are given in Table 1. This process is applied after the ensemble operation. An example image demonstrating the effect of this step is shown in Figure 3.

Table 1: IoU thresholds for false-positive elimination.

Class	Threshold	Class	Threshold
Specularity	0.13	Contrast	0.19
Saturation	0.21	Bubbles	0.12
Artifact	0.17	Instrument	0.24
Blur	0.4	Blood	0.11

3. EXPERIMENTAL DESIGN

We have evaluated the performance of the individual models and their combination through affirmative and consensus ensemble models. In addition, we have evaluated the effect of adding a false-positive elimination step on the outputs of these models. We have used the EAD Challenge dataset [1] throughout the experiments. The dataset contains 2200 images and 1555 of them, corresponding to 70%, have a dimension of 512x512. Therefore, we rescaled all the images to that size in order to fix the input size. In order to prevent overfitting, 10% of the overall dataset (~250 images) has been set aside for validation. The rest of the images in the dataset have been used for training. The training dataset has been expanded by image augmentation techniques. Each image has been transformed by horizontal flipping and 90°, 180°, 270° rotations. As a result, there was an eight-fold increase in the training dataset size. We have observed that augmenting the dataset results in better generalization.

For the best performance, anchor box sizes should match the object bounding boxes. For this purpose, we calculated the statistics of the ground-truth object bounding boxes. Figure 4 shows the histogram of the bounding boxes. According to this figure, most of the bounding boxes are located in the region where bounding boxes are smaller than the median area (256x512); therefore, 12, 25, and 80-pixel sizes for both width and height have been used for smaller boxes. For the mid-size and larger bounding boxes, 256 and 384 pixels have been chosen respectively. Each anchor box

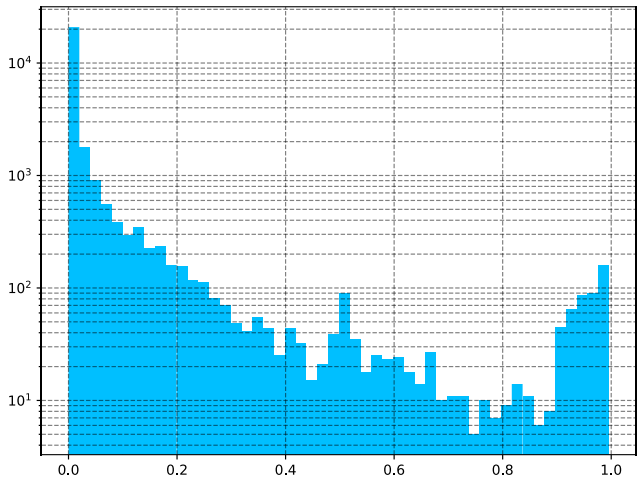

Figure 4: Histogram of normalized bounding box sizes, where 1.0 corresponds to an area of 512x512**.**

size [12, 25, 80, 256, 384] was mapped to the corresponding feature map layer in [p_2, p_3, p_4, p_5, p_6] respectively where p_n is the n^{th} feature map layer. Three different aspect ratios (width/height): 0.5, 1, and 2, were used for each anchor box.

The total number of iterations was 200000 for Faster R-CNN and Cascade R-CNN and 90000 for RetinaNet. Learning rate scheduling by a factor of 10 was used for all three models. Scheduling has been done at iterations 130000 and 180000 for Faster R-CNN, at iterations 150000 and 190000 for Cascade R-CNN and at iterations 60000 and 80000 for RetinaNet.

We used PyTorch [11] and Detectron2 API [12] to train the models on a workstation with two NVIDIA RTX2080 GPUs. Faster R-CNN and Cascade R-CNN models took 15 hours to train, and RetinaNet model took 11 hours to train using a single GPU. We used the other GPU to train different models in parallel. For all three models, weights of pretrained models on COCO dataset have been used.

The results are given in Table 2. In addition to the results using different network types, ensemble models and their version with class-agnostic NMS and FP elimination steps are also provided. Ensemble methods utilize all three networks.

Table 2: Experimental results.

Method	Without Class-Agnostic NMS		With Class-Agnostic NMS	
	mAP	mIoU	mAP	mIoU
Faster R-CNN with FPN	45.66	40.78	44.20	42.82
Cascade R-CNN with FPN	45.98	32.23	44.07	35.03
RetinaNet	45.09	36.44	43.91	41.22
Ensemble (affirmative)	**47.91**	26.03	47.12	30.28
Ensemble (consensus)	47.29	42.89	45.96	45.19
Ensemble (affirmative) with FP elimination	46.92	32.21	46.54	34.25
Ensemble (consensus) with FP elimination	46.86	44.65	45.71	**45.91**

4. EXPERIMENTAL RESULTS AND DISCUSSION

According to the results in Table 2, while the individual networks have very similar mAP values, the Faster R-CNN model has a higher mIoU. The affirmative ensemble gives the highest mAP score, which is expected as some true positives missed by a model can be detected by the other models. On the other hand, a higher number of false positives are generated, which adversely affects its mIoU score. The consensus ensemble has the highest mIoU value among the methods not utilizing FP elimination. Although class-agnostic NMS and FP reduction steps decrease the mAP values marginally, they eliminate many false-positives and give higher mIoU scores, resulting in a more balanced mAP and mIoU scores. For example, when FP elimination is applied to the ensemble (affirmative) result, in return to a 0.99 points decrease in mAP, there is a 6.18 points increase in mIoU. Increasing mIoU by such an elimination mechanism adversely affect mAP. Because, in some cases, object detectors do not perform well; models may detect artefacts incorrectly and boxes which have true classes but low confidence scores are suppressed by wrongly detected high confidence boxes. It is observed that FP elimination works better if there is a lower mIoU. Since there is a trade-off between mAP and mIoU, these steps can be utilized to have more robust object detectors. Different score metrics are used for different object detection tasks. In this work, we have used post-processing techniques to have a balanced mAP and mIoU scores.

The highest scores are obtained using the consensus ensemble of the classifiers, which were passed through a class-agnostic NMS, and FP reduction as the final step.

Object detectors are generic and they are not developed considering the domain-specific challenges. In addition, these networks have many internal parameters and these parameters need to be tuned for the particular application. Hence, it is not sufficient to use more advanced models and a comprehensive understanding of the characteristics of the data is of essence.

To integrate the domain knowledge into detection architecture, we have qualitatively observed that some classes such as specularity and saturation have bounding boxes overlapping with each other. While removal of the one that has less confidence seems to be a solution, this is not ideal since, in a number of cases, the one that has less confidence is the true class. Therefore, specific algorithms should be included in the detection framework to tackle this problem.

5. CONCLUSIONS

In this study, we have trained three different object detectors for endoscopic artefact detection. We have used ensemble techniques to utilize all three individual networks. Applying a class-agnostic NMS to each of them independently resulted in a better trade-off between mAP and mIoU scores. As a final step, FP elimination is applied, which resulted in more robust results.

In this work, we have focused on using lighter networks and taken ensemble of weak classifiers approach. Use of lighter networks made the hyperparameter tuning possible in feasible time periods and allowed us to experiment with various network parameters. In the future, more sophisticated networks, such as ResNeXt or ResNet152, which require more time to train and tune parameters could also be investigated.

6. ACKNOWLEDGEMENTS

We would like to thank MİTAŞ TI Tower AG for donation of the workstation and GPUs used in this work.

7. REFERENCES

[1] S. Ali *et al.*, "Endoscopy artifact detection (EAD 2019) challenge dataset," *arXiv 1905.03209*, 2019. [Online]. Available: http://dx.doi.org/10.17632/C7FJBXCGJ9.1.

[2] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 91--99.

[3] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into High Quality Object Detection," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6154–6162.

[4] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal Loss for Dense Object Detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2980–2988.

[5] T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2117–2125.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[7] S. Ali *et al.*, "An objective comparison of detection and segmentation algorithms for artefacts in clinical endoscopy," *Sci. Rep.*, vol. 10, no. 1, pp. 1–15, 2020.

[8] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *European Conference on Computer Vision*, 2014, pp. 740–755.

[9] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes (VOC) Challenge," *International Journal of Computer Vision*, vol. 88, no. 2. pp. 303–338, 2010.

[10] J. Heras and A. Casado-Garcia, "Ensemble Methods for Object Detection," in *24th European Conference on Artificial Intelligence (ECAI2020)*, 2020.

[11] A. Paszke *et al.*, "PyTorch: An Imperative Style, High-Performance Deep Learning Library." 2019.

[12] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2," 2019. [Online]. Available: https://github.com/facebookresearch/detectron2.