Evaluating object detector ensembles for improving the robustness of artifact detection in endoscopic video streams

Pedro Esteban CHAVARRIAS-SOLANO[1], Carlos Axel GARCIA-VEGA[1], Francisco Javier LOPEZ-TIRO[1], Gilberto OCHOA-RUIZ[1], Thomas BAZIN[2], Dominique LAMARQUE[2], Christian DAUL[3]

[1]Escuela de Ingenieria y Ciencias, Tecnologico de Monterrey
Av. Eugenio Garza Sada 2501 Sur, Tecnológico, 64849 Monterrey, N.L., Mexico

[2] Hôpital Ambroise Paré
9 Avenue Charles de Gaulle, 92100 Boulogne-Billancourt, France

[3] CRAN (UMR 7039), Université de Lorraine and CNRS,
2 avenue de la Forêt de Haye, 54518 Vandœuvre-lès-Nancy cedex, France

A00344305@tec.mx, A01754346@tec.mx, A01799045@tec.mx, gilberto.ochoa@tec.mx
thomasbazin@icloud.com, dominique.lamarque@aphp.fr, christian.daul@univ-lorraine.fr

Résumé – Dans cette contribution nous utilisons une méthode ensembliste d'apprentissage profond pour combiner la prédiction de deux détecteurs individuels à un étage (c'est-à-dire YOLOv4 et Yolact) dans le but de détecter les artefacts vus dans des images endoscopiques. Cette ≪ stratégie ensembliste ≫ a permis d'améliorer la robustesse des modèles individuels sans nuire à leurs capacités de calcul en temps réel. L'efficacité de notre approche a été démontrée en entraînant et testant les deux modèles individuels et diverses configurations ensemblistes sur le jeu de données ≪ Endoscopic Artifact Detection Challenge ≫. Des expériences poussées montrent la supériorité, en termes de précision moyenne, de l'approche ensembliste par rapport aux modèles individuels et aux travaux de l'état de l'art.

Abstract – In this contribution we use an ensemble deep-learning method for combining the prediction of two individual one-stage detectors (i.e., YOLOv4 and Yolact) with the aim to detect artefacts in endoscopic images. This ensemble strategy enabled us to improve the robustness of the individual models without harming their real-time computation capabilities. We demonstrated the effectiveness of our approach by training and testing the two individual models and various ensemble configurations on the "Endoscopic Artifact Detection Challenge" dataset. Extensive experiments show the superiority, in terms of mean average precision, of the ensemble approach over the individual models and previous works in the state of the art.

1 Introduction

Endoscopy is a technique that has been widely used over two centuries by physicians to screen the interior of otherwise inaccessible sites in the human body [11]. Nowadays, it is the primary diagnostic and therapeutic tool for managing gastrointestinal (GI) malignancies [9] and the primary instrument in minimally invasive surgery (MIS) procedures [11]. As the endoscope provides a high quality video signal, it has fostered the development of a great deal of tasks related to image analysis and computer vision [10]. Some of the most promising applications in this area are related to the detection of diseases or pre-cancerous lesions such as polyps, ulcer, bleeding, Celiac and Chron's disease. Recently, the use of computer vision (CV) approaches in endoscopy has caught the attention of the artificial intelligence (AI) and medical research communities due to the advent of deep learning approaches. Typically used of computer vision in endoscopy fall into two main categories : Computer-aided detection (CADe) and computer-aided diagnosis (CADx) systems [8].

Even though promising applications have been demonstrated, there are several challenges that must be addressed before AI can be successfully deployed in endoscopic interventions in real clinical settings : for instance, images can be corrupted by various types of artefacts, making object detection and instance segmentation methods less robust and unable to generalize , among other pressing issues [1]. In order to tackle these problems, various public datasets have been released to develop tools capable of handling these complex settings and various competitions are organized every year in major conferences.

For instance, the EndoCV is a crowd-sourced challenge that aims to address these issues by developing reliable and robust CADe/x endoscopy systems [1]. In recent editions of this challenge, the teams that obtained the two highest detection scores implemented an ensemble of one two-stage and one single-stage detectors [1]. The main advantage of employing two-stage detectors is the overall improvement in terms of robustness. However, these methods have an important limitation in the clinical context, which is related to the very high inference time, as they cannot be used for CADe or CADx applications.

More recently, several proposals in the challenge have been proposed to overcome these limitations. In [4], an ensemble of RetinaNet, Cascade and Faster R-CNN was proposed with a class-agnostic NMS stage after each model. An ensemble of Faster R-CNN RetinaNet and Faster R-CNN was implemented in [5]. On the other hand, a YOLACT implementation with Non-Maximum Suppression (NMS) was proposed by [12].

Even though multiple proposals have started to tackle the robustness issues discussed above, we consider that there is still a lack of robust yet real-time capable object detector that would enable various applications for the CV community in endoscopy. A way of dealing with this problem of robustness is to adopt a model ensambling strategy, but the models need to be preferably lightweight (to avoid becoming too power hungry and memory intensive) and run in real-time [6].

To investigate how this can be achieved, in this paper we present a comparison between two single-stage object detectors, YOLOv4 and YOLACT, and we propose an ensemble using both of them. The main contribution consists on developing an ensemble mechanism using only single-stage detectors aiming to improve the robustness of the detection task while maintaining a low memory footprint and inference time.

This paper is organized as follows : In Section II, a brief description of EndoCV challenge, YOLOv4 and YOLACT architectures is given, followed by the ensemble method that was implemented for this dataset. Section III covers the results that were obtained during this comparison and section IV concludes the article with some future avenues of research.

2 Data and methods

This study compares the use of two single-stage detectors, YOLOv4 and YOLACT against an ensemble of both of them, which we dub CEM. These models were trained and tested with the Endoscopy Artifact Detection challenge 2020 [1].

2.1 EAD Dataset

The EndoCV challenge consists of Endoscopy Disease Detection (EDD) and Endoscopy Artifact Detection (EAD) tasks. The EAD sub-challenge contains diverse endoscopy video frames that were collected from seven institutions. The dataset covers eight different artefact classes that were identified by clinical experts as specularity, saturation, artefact, blur, contrast, bubbles, instrument and blood. The dataset is composed by 2,532 images, classified into two types of data : single and sequence frames [1] for testing different models.

At the top of Fig. 1, five images from the single frame category are shown. While, at the bottom of the same figure, five images belonging to the sequence frames category are shown.

For the detection task, 31,069 bounding boxes were given, being specularity, the class with more occurrences, followed by artefact and bubbles. Whereas, the class with less occurrences is blood, followed by instrument, blur, and saturation. A chart describing the data distribution is given in Fig. 2.

2.2 Object Detectors

The first single-stage object detector that was trained and evaluated with this dataset was YOLOv4 [2]. The architecture of this model was set to work with CSPDarknet53 as its backbone, an additional SPP module, PANet path-aggregation neck and a YOLOv3 head.

The other single-stage object detector that was selected for this study is YOLACT's lightweight instance segmentation detector [7], whose architecture closely follows the one of RetinaNet, without the use of focal loss. We set our backbone detector to be ResNet50 with a modified FPN, applying smooth-L1 loss to train box regressors and encode box regression coordinates, as SSD.

2.3 Ensemble

An ensemble method combines the predictions done by multiple object detectors into a final output [3]. This method can be understood as a voting system in which every model in the ensemble submits its own predictions, aiming that the final decisions accuracy overcomes the accuracy of every learning method alone. We implemented two voting strategies :

— **Consensus :** This voting strategy needs the majority of the models to detect the same object in order to consider that prediction a final output. Since we are structuring our ensemble method with just two single-stage object detectors, this approach works exactly as the unanimous voting strategy, in which both YOLOv4 and YOLACT models need to agree with the same prediction.

— **Affirmative :** Unlike the consensus and unanimous strategies, this approach requires just a single model, either YOLOv4 or YOLACT, to do a prediction in order to take it into account for the final output. In other words, all predictions done by each detector will be considered for the final result.

2.4 Training

The training procedure of both object detectors, YOLOv4 and YOLACT, was performed on an NVIDIA DGX-1 system with eight Tesla V100 GPUs. Before the training procedure was done, input images were resized to 416 x 416 pixels and augmented using different data augmentation techniques : flips, blur and hue, gamma and equalized histogram and a combination of all of them.

The YOLOv4 model was trained using a learning rate of 0.001, a batch size of 64, subdivision equals to 16, momentum value of 0.949 and weight decay of $5e^{-4}$ on a single GPU. The traditional YOLACT model was also trained using a single GPU with a learning rate of 0.001, batch size of 16 images, weight decay of $5e^{-4}$, a momentum value of 0.9. The ensemble method was implemented by using the generated object detector models from the training stage.

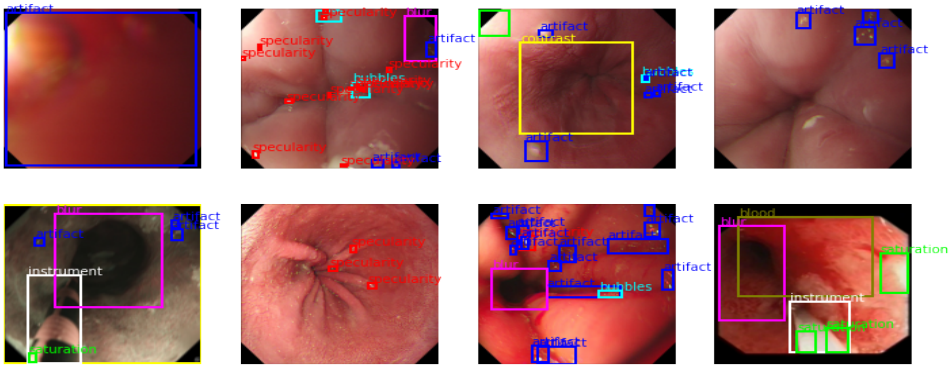

FIGURE 1 – Samples from EAD2020 dataset images [1]

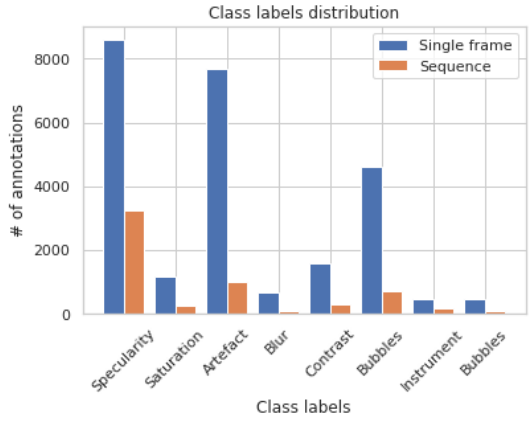

FIGURE 2 – Class distribution from EAD2020 dataset [1]

2.5 Metrics

The performance of the methods was evaluated using an standard evaluation metric : mean average precision (mAP). This metric measures the ability of a model to accurately capture all instances of the ground truth annotations.

This metric was evaluated at three different intersection over union values (IoU) : 0.25, 0.50, and 0.75. The IoU value measures the overlap of two different bounding boxes and is used to determine whether or not a prediction is correct with respect to the ground truth.

3 Results

This section compares qualitatively and quantitatively the obtained results after completing an evaluation of the selected models : YOLOv4, YOLACT and the ensemble method.

3.1 Quantitative Results

The mAP metric was used to evaluate all three methods at three different IoU : 25, 50, and 75 percent. The graphs in Fig. 3 show the mAP values obtained by each model with different data augmentation strategies : original data only, geometric data augmentation, distortion (blur and hue), photo-metric (gamma and equalized histogram) and flips. The red crosses indicate the YOLOv4 mAP values, while the blue circles denote the YOLACT mAP values and finally, the green stars highlight the mAP values of the consensus ensemble method (CEM).

From the figure, we can observe that the ensemble model outperforms the other methods for at 75 IoU, while preserving a low inference time given by the two one-stage detectors.

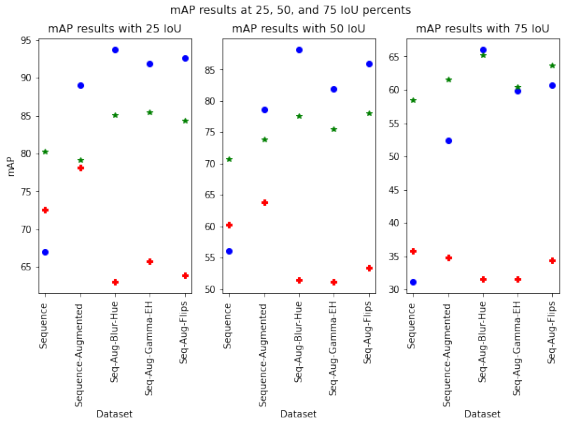

FIGURE 3 – Mean average precision values at 25, 50, and 75 IoU values. Red crosses : YOLOv4 mAP values, Blue Circles : YOLACT mAP values and Green stars : CEM

One of the main objectives of the challenge is to address generalization issues, therefore, the test data that was released to the participants was collected from different sources not presented in the training data. In this study, the public training data was randomly split to construct the train and test datasets since official test data is not publicly available. Even though the models evaluated during the challenge and our models were not evaluated over the same data, table 1 presents the results obtained by the three best performing teams, the given baseline and the methods described in this study, i.e. YOLOv4, YOLACT and CEM, with the data augmentation that achieves the best mAP for the CEM method. From the table we can observe that all our methods outperform previous competitors challenge, as well as the baselines by a large margin. YOLACT has the hi-