

Deep Learning Model Generalization with Ensemble in Endoscopic Images

Ayoung Hong^{a,b}, Giwan Lee^b, Hyunseok Lee^c, Jihyun Seo^{d,e} and Doyeob Yeo^e

^aRobotics Engineering Convergence, Chonnam National University, Gwangju, South Korea

^bDepartment of AI Convergence, Chonnam National University, Gwangju, South Korea

^cDaegu-Gyeongbuk Medical Innovation Foundation, Daegu, South Korea

^dDepartment of Computer Convergence Software, Korea University, Sejong, South Korea

^eElectronics and Telecommunications Research Institute, Daejeon, South Korea

Abstract

Owing to the rapid development of deep learning technologies in recent years, autonomous diagnostic systems are widely used to detect abnormal lesions such as polyps in endoscopic images. However, the image characteristics, such as the contrast and illuminance, vary significantly depending on the center from which the data was acquired; this affects the generalization performance of the diagnostic method. In this paper, we propose an ensemble learning method based on k -fold cross-validation to improve the generalization performance of polyp detection and polyp segmentation in endoscopic images. Weighted box fusion methods were used to ensemble the bounding boxes obtained from each detection model trained for data from each center. The segmentation results of the data center-specific model were averaged to generate the final ensemble mask. We used a Mask R-CNN-based model for both the detection and segmentation tasks. The proposed method achieved a score of 0.7269 on the detection task and 0.7423 ± 0.2839 on the segmentation task in Round 1 of the EndoCV2021 challenge.

Keywords

EndoCV2021, polyp detection, polyp segmentation, ensemble, k -fold cross-validation

1. Introduction

Colorectal polyps are abnormal tissue growths that create flat bumps or tiny mushroom-like stalks on the lining of the colon or rectum. Although these abnormal cells often cause rectal bleeding and abdominal pain due to partial bowel obstruction, most colorectal polyps are asymptomatic [1]. However, it is very important to examine the growth of polyps since it is highly related to colorectal cancer.

Colonoscopy is a common procedure used to examine the inside wall of the colon by using a camera attached at its tip, and if necessary, polyps are removed by inserting the instrument through its channel during the procedure. Doctors localize polyps and examine their size and shape using the captured colonoscopic images; however, the polyp detection rate varies


3rd International Workshop and Challenge on Computer Vision in Endoscopy (EndoCV2021) in conjunction with the 18th IEEE International Symposium on Biomedical Imaging ISBI2021, April 13th, 2021, Nice, France

✉ ahong@jnu.ac.kr (A. Hong); 216297@jnu.ac.kr (G. Lee); hs.lee@dgmif.re.kr (H. Lee); goyangi100@korea.ac.kr (J. Seo); yeody@etri.re.kr (D. Yeo)

ORCID 0000-0002-3839-4194 (A. Hong); 0000-0002-8343-2356 (J. Seo); 0000-0002-6510-8763 (D. Yeo)



© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

depending on the abilities of individual clinicians [2]. This has led to the development of automatic polyp detection systems using methods based on computer vision [3]. The recent advancements in deep learning technologies have contributed to new polyp detection methods using artificial intelligence.

The Endoscopy Computer Vision Challenge (EndoCV) was initiated in 2019 as Endoscopic Artefact Detection (EAD) [4] to realize reliable computer-assisted endoscopy by detecting multiple artifacts such as pixel saturation, motion blur, defocus, bubbles, and debris. This year, the EndoCV challenge [5] aims to develop a polyp detection and segmentation method that works for endoscopic images obtained from multiple centers. In this paper, we propose an ensemble learning method based on k -fold cross-validation to improve the generalization performance.

2. Related work

2.1. Detectron2

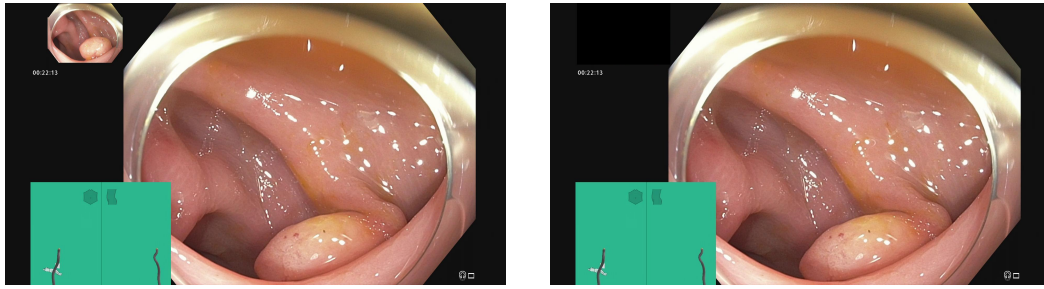
Detectron2 is a PyTorch [6]-based software system to provide the implementations of object detection and segmentation networks, developed by Facebook AI Research [7]. It provides object detection algorithms, including Faster R-CNN [8] and RetinaNet [9], and includes implementations of instance segmentation with Mask R-CNN [10] and panoptic segmentation with Panoptic FPN [11]. The implemented networks are provided with pre-trained weights using the COCO dataset that facilitates transfer learning.

2.2. Mask R-CNN

Mask R-CNN is a model proposed to perform image detection and segmentation simultaneously [10], by extending Faster R-CNN with a Feature Pyramid Network (FPN), mask branch, and Region of Interest Align (RoIAlign) [8]. Faster R-CNN creates RoI from only one feature map and performs classification and bounding box regression. However, it is difficult to collect detailed feature data and required to create an anchor box with many scales and ratios to detect objects of various sizes in a feature map that induces an inefficient learning process. To overcome this limitation, FPN was introduced [12]. FPN enables the detection of a large object in a small map and a small object in a large map by creating an anchor box of the same scale in each map. To perform segmentation, Faster R-CNN adopts RoI Pooling, whereas Mask R-CNN fixes RoI through RoIAlign [10]. In this method, pixel values are generated using bilinear interpolation [13], which leads to a more accurate segmentation model by preventing position distortion. In the EndoCV2021 challenge, polyp detection and segmentation must be performed together; therefore, Mask R-CNN is used in this study.

2.3. Robust Model Selection Methods

To train a deep learning classification model with high accuracy, it is important to avoid overfitting of the model on the training dataset and to have robustness for data that the model has never seen before. This implies that we need to choose a model with optimal hyperparameters



(a) An original training endoscopic image from center 4 with a sub-endoscope image part. (b) The endoscopic image after removing the sub-endoscope image part.

Figure 1: Removing the sub-endoscope image part in the training dataset.

that have small bias and variations for general data. The Holdout evaluation is a simple model selection method that divides the data into training and test sets. Typically, 4/5 of the available data are used as a training set, and the remaining data are used as a test set. However, the performance of this method relies heavily on the method of selecting the test set from the entire dataset. This could cause the model to overfit the test set.

One of the most popular model selection methods is k -fold cross-validation [14]. In this method, we randomly separate the entire data into k exclusive subsets. Then, the deep learning classifier trains using $k - 1$ subsets and tests using the other subset by repeating the process k times. The performance of the model is the average score obtained over the k times training. This prevents the model from overfitting on the test data and helps in achieving better prediction performance over data that it has never seen before.

3. Proposed Method

In this section, we describe how polyp detection and segmentation were performed in the EndoCV2021 competition. We used ensemble inference based on k -fold cross-validation to increase the generalization performance. In addition, we used several data augmentation techniques provided by Detectron2 and preprocessed the input images for training each mask R-CNN model to improve both the segmentation and detection performances of each mask R-CNN model.

3.1. Preprocessing

Since the characteristics of the image change depending on the center from where the endoscopic image is acquired, we used data augmentation to prevent overfitting during training and to improve the generalization performance. Data augmentation uses the same technique for the detection and segmentation of each mask R-CNN model. We applied the augmentation techniques provided in Detectron2, namely RandomBrightness, RandomContrast, RandomSaturation, and RandomLighting.

As shown in Figure 1, several endoscopic images in the training set included a secondary

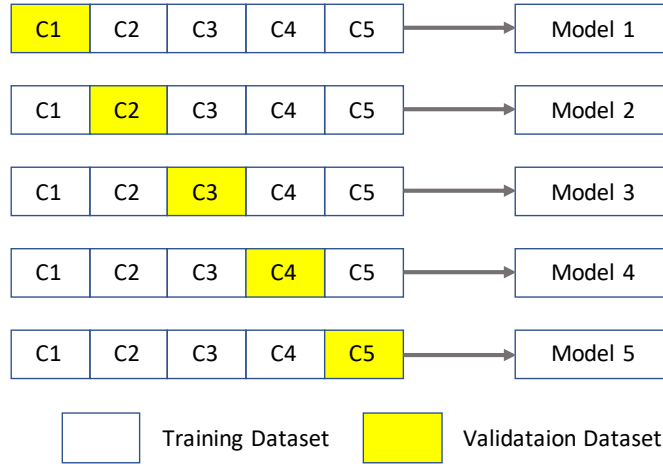


Figure 2: Ensemble training based on k -fold cross-validation for data from five centers. The white boxes and yellow boxes denote the training and validation sets for each mask R-CNN model, respectively.

endoscopic image in addition to the main endoscopic image. Because the ground truth labels for the provided detection/segmentation tasks do not consider the sub-endoscope image part, we removed the sub-endoscope image part in the training dataset to improve the detection/segmentation performance.

3.2. Ensemble

3.2.1. Training

The training dataset provided in the EncoCV 2021 competition was given as data from five different centers. As shown in Figure 2, we trained five mask R-CNN models for ensemble inference based on k -fold cross-validation. While training a single mask R-CNN model, the data acquired from all five data centers were not used; instead, only the data acquired from the remaining four data centers were used. In this study, the learned mask R-CNN model excluding center i is called “Model i ” ($i = 1, 2, 3, 4, 5$).

3.2.2. Inference

Ensemble inference was performed by combining the detection/segmentation inference results of Models 1–5, as shown in Figure 3. In detection, bounding boxes ensemble the results of five models using the weighted box fusion technique [15]. In segmentation, the mask created from each model has a value of 0 for the background and 1 for the polyp. We averaged the segmentation masks from Models 1–5. When the inference result was above the given threshold, it was determined to be the final polyp mask; otherwise, it was considered as background. In this study, the threshold was set to 0.6. In addition, only segmentation/detection results with confidence values greater than 0.5 in each model were used in the ensemble scores.

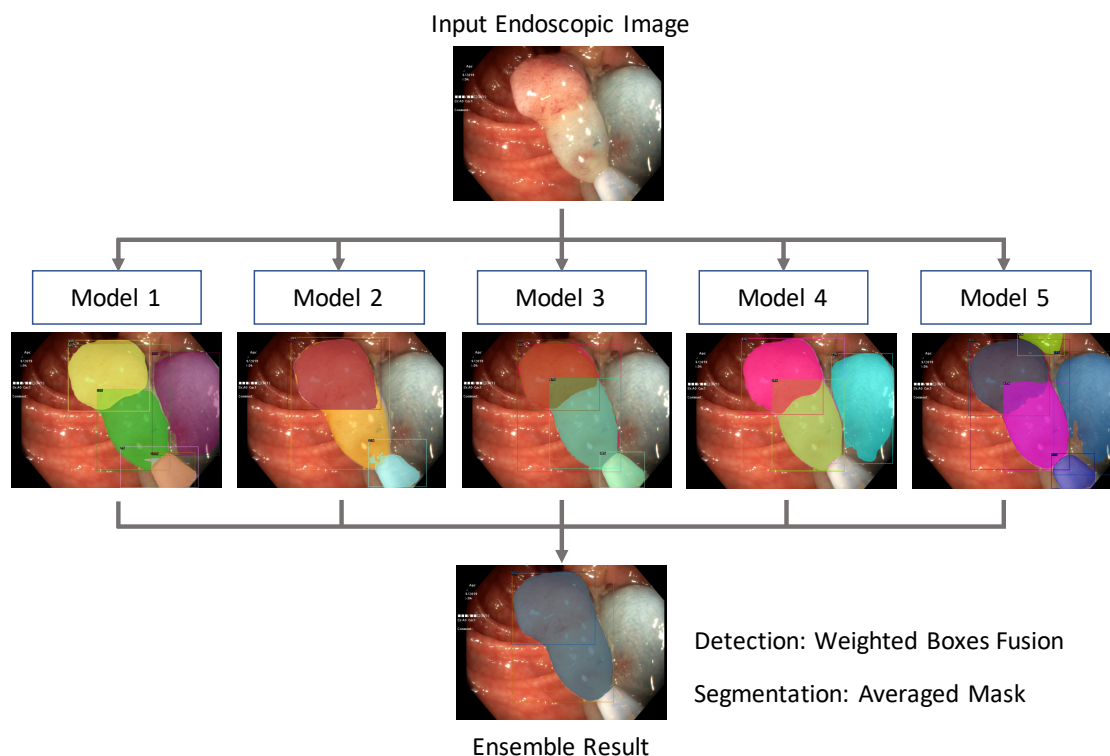


Figure 3: Ensemble inference based on k -fold cross-validation for data from five centers. To ensemble the results from each mask R-CNN model, we use a weighted box fusion technique for the detection task and use an averaged mask for the segmentation task.

4. Experimental Results

In this section, we compare the performance of the proposed ensemble method with that of a single model trained with all the datasets from five centers.

4.1. Settings of experiments

At the time of writing this paper, not all scores were provided for the competition test set; therefore, validation was performed using the given dataset. In addition to the data provided by the five centers, the competition also provided a sequential image dataset. The sequential dataset was kept separately and was used only to evaluate the model's performance. The ensemble model and the single model were trained using the default trainer provided by Detectron2. A stochastic gradient descent optimizer with LR= 0.001 and a warm-up scheduler was used in the experiments. Image augmentations such as random crop and flip were used to avoid overfitting. We conducted the experiments with training steps of 50,000 and stored the checkpoints for every 1,000 steps. Among the saved checkpoints, the one with the best performance on the validation set was selected as the final model weight. For a more sophisticated evaluation, various evaluation metrics were used: Jaccard, Dice, F_2 , Precision, Recall, and Accuracy, for

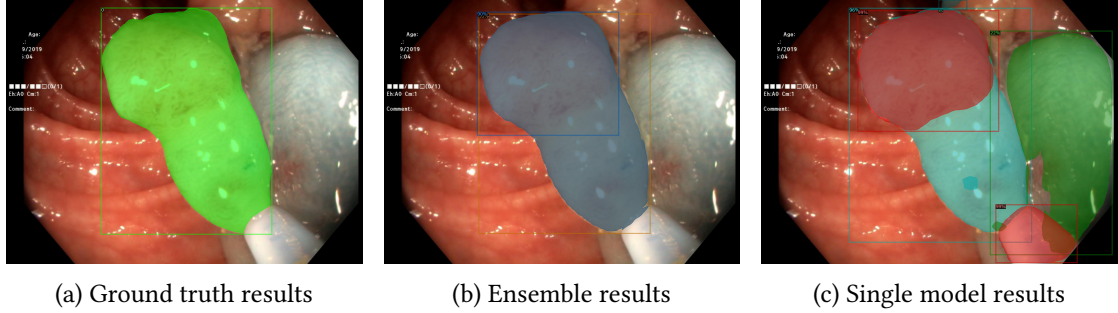


Figure 4: Comparison of the ensemble result (b) and the result of the single model trained using data from five centers at once (c). In (c), there are two objects (green and pink) inferred as polyps for non-polyp objects. In contrast, in (b), only the polyp region is found.

segmentation task; and mAP (Mean Averaged Precision at IoU=0.50:.05:.95), APs, APm and API (AP for small, medium and large), for the detection task. The organisers also provided leaderboard scores that incorporated out-of-sample generalisation metric [4].

4.2. Evaluation results

As shown in Figures 3 and 4, the inference results of models 1–5 and the single model show many false positive regions. However, the ensemble technique based on k -fold cross-validation significantly reduced the number of false positive regions.

The segmentation task result is shown in Table 1. The performance of each individual model (Models 1 to 5) was significantly lower than that of the single model. However, the ensemble model showed better performance than the single model, except for recall. In particular, the precision increased significantly because the number of false positives was reduced in the ensemble model.

The experimental results for the detection task show the similar aspects, as shown in Table 2. The performance of an individual model was either better or worse than that of the single model. However, the ensemble model showed better performance than the single model, except for APm. It should be noted that API improved significantly.

Table 3 shows the average computing time taken for the detection and segmentation inferences using the single model and the proposed ensemble model. The computing time was measured and averaged for 1793 images in the test dataset. Although the ensemble was performed using five single models, the inference time of the ensemble model takes less than five times compared to that of the single model. In addition, the ensemble process of a single image took on average 0.008 seconds and 0.014 seconds for detection and segmentation tasks, respectively.

5. Conclusions

In this paper, we proposed an ensemble learning method based on k -fold cross-validation to improve the generalization performance of polyp detection and segmentation in endoscopic images. Five mask R-CNN models were trained using only the training data collected from four of

Table 1

Segmentation performances for the test dataset. Single indicates the model that was trained using the data of five centers at once.

	Jaccard	Dice	F_2	Precision	Recall	Accuracy
Model 1	0.4700	0.5372	0.5780	0.6173	0.7693	0.9113
Model 2	0.5208	0.5717	0.5914	0.7557	0.6907	0.9431
Model 3	0.5345	0.5774	0.5709	0.8712	0.6057	0.9584
Model 4	0.5494	0.5854	0.5903	0.8484	0.6369	0.9542
Model 5	0.5503	0.5991	0.6151	0.7553	0.7204	0.9425
Single	0.5518	0.6044	0.6229	0.7647	0.7235	0.9434
Ensemble	0.5707	0.6192	0.6290	0.8122	0.7053	0.9554

Table 2

Detection performances for the test dataset. Single indicates the model that was trained using the data of five centers at once.

	mAP	APs	APm	API
Model 1	0.3223	0.0223	0.2753	0.3791
Model 2	0.2820	0.0147	0.0755	0.3711
Model 3	0.2327	0.0707	0.2527	0.2508
Model 4	0.3177	0.1233	0.3185	0.3463
Model 5	0.3555	0.1199	0.3279	0.3989
Single	0.3063	0.1015	0.3232	0.3329
Ensemble	0.3976	0.1264	0.3098	0.4585

Table 3

An average inference time consumption (secs) for a single image.

	Detection	Segmentation
Single	0.175	0.183
Model 1–5	0.463	0.506
Model 1–5 + Ensemble	0.471	0.520

the five centers, and the inference results of the five models were ensemble. In the ensemble, the weighted boxes fusion technique was used for detection, and the ensemble segmentation mask was created by averaging the inference results of five segmentation masks. In the experiment, the detection and segmentation performances were measured using the sequential image dataset provided in the EndoCV2021 competition as a test dataset. The experimental results showed that in both detection and segmentation tasks, the performance of the mask R-CNN using the ensemble technique was better than that of the mask R-CNN model trained using data from all five centers at once. The proposed method achieved a score of 0.7269 on the detection task and 0.7423 ± 0.2839 on the segmentation task in Round 1 of the EndoCV2021 challenge.

Acknowledgement

This work was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No.NRF-2020R1F1A1072201) and in part by the National Research Council of Science & Technology (NST) grant from the Korean government (MSIP) (No. CRC-15-05-ETRI).

References

- [1] J. H. Bond, Polyp guideline: diagnosis, treatment, and surveillance for patients with colorectal polyps, *Am J Gastroenterol* 95 (2000) 3053.
- [2] G. Urban, P. Tripathi, T. Alkayali, M. Mittal, F. Jalali, W. Karnes, P. Baldi, Deep learning localizes and identifies polyps in real time with 96% accuracy in screening colonoscopy, *Gastroenterology* 155 (2018) 1069–1078.
- [3] J. Bernal, J. Sánchez, F. Vilarino, Towards automatic polyp detection with a polyp appearance model, *Pattern Recognit* 45 (2012) 3166–3182.
- [4] S. Ali, F. Zhou, B. Braden, A. Bailey, S. Yang, G. Cheng, P. Zhang, X. Li, M. Kayser, R. D. Soberanis-Mukul, S. Albarqouni, X. Wang, C. Wang, S. Watanabe, I. Oksuz, Q. Ning, S. Yang, M. A. Khan, X. W. Gao, S. Realdon, M. Loshchenov, J. A. Schnabel, J. E. East, G. Wagnieres, V. B. Loschenov, E. Grisan, C. Daul, W. Blondel, J. Rittscher, An objective comparison of detection and segmentation algorithms for artefacts in clinical endoscopy, *Scientific Reports* 10 (2020) 2748. doi:10.1038/s41598-020-59413-5.
- [5] S. Ali, D. Jha, N. Ghatwary, S. Realdon, R. Cannizzaro, M. A. Riegler, P. Halvorsen, C. Daul, J. Rittscher, O. E. Salem, D. Lamarque, T. de Lange, J. E. East, Polypgen: A multi-center polyp detection and segmentation dataset for generalisability assessment, *arXiv* (2021).
- [6] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, Pytorch: An imperative style, high-performance deep learning library, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, R. Garnett (Eds.), *Advances in Neural Information Processing Systems* 32, Curran Associates, Inc., 2019, pp. 8024–8035.
- [7] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, R. Girshick, Detectron2, <https://github.com/facebookresearch/detectron2>, 2019.
- [8] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, *arXiv preprint arXiv:1506.01497* (2015).
- [9] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: *Proc IEEE Int Conf Comput Vis*, 2017, pp. 2980–2988.
- [10] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask r-cnn, in: *Proc IEEE Int Conf Comput Vis*, 2017, pp. 2961–2969.
- [11] A. Kirillov, R. Girshick, K. He, P. Dollár, Panoptic feature pyramid networks, in: *Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit*, 2019, pp. 6399–6408.
- [12] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks

- for object detection, in: Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit, 2017, pp. 2117–2125.
- [13] E. J. Kirkland, Bilinear interpolation, in: *Advanced Computing in Electron Microscopy*, Springer, 2010, pp. 261–263.
- [14] M. Anthony, S. B. Holden, Cross-validation for binary classification by real-valued functions: theoretical analysis, in: *Proceedings of the eleventh annual conference on Computational learning theory*, 1998, pp. 218–229.
- [15] R. Solovyev, W. Wang, T. Gabruseva, Weighted boxes fusion: Ensembling boxes from different object detection models, *Image and Vision Computing* (2021) 1–6.