Medico Multimedia Task at MediaEval 2021: Transparency in Medical Image Segmentation

Steven A. Hicks[1,2], Debesh Jha[1,3], Vajira Thambawita[1,2], Hugo L. Hammer[2], Thomas de Lange[4,5], Sravanthi Parasa[6], Michael A. Riegler[1], and Pål Halvorsen[1,2]

[1]SimulaMet, Norway
[2]Oslo Metropolitan University, Norway
[3]UiT The Arctic University of Norway, Norway
[4]Sahlgrenska University Hospital Mölndal, Sweden
[5]University of Gothenburg, Sweden
[6]Swedish Medical Center, USA

ABSTRACT

The *Medico Multimedia Task* focuses on providing multimedia researchers with the opportunity to contribute to different areas of medicine using multimedia data to solve several subtasks. This year, the task focuses on transparency within machine learning-based medical segmentation systems, where the use case is gastrointestinal endoscopy. In this paper, we motivate the organization of this task, describe the development and test dataset, and present the evaluation process used to assess the participants' submissions.

1 INTRODUCTION

Finding and removing colon polyps is an essential step in preventing colorectal cancer. Current procedures have a high miss rate, whereas computer-aided diagnosis systems can reduce the probability that diagnosticians overlook a polyp during a colonoscopy. As machine learning becomes more common in high-risk fields like medicine, the need for transparent systems becomes more critical. In this case, transparency is defined as giving as much detail as possible on the different parts that make up a machine learning pipeline, including everything from data collection to final prediction. This task focuses on high-performing, efficient, and transparent algorithms for polyp segmentation.

The Medico Multimedia Task is held for the fifth time at the MediaEval benchmark. We continue the tradition of using medical data to develop machine learning models that solve real-world issues in medicine [2, 4–6]. Like last year, we use the gastrointestinal tract as the medical use case, where automatic polyp segmentation is the primary focus. However, this year, we have more training data and add an additional task that focuses on transparency in the submitted solutions. The task is of interest to the researchers working with multimedia segmentation, deep learning (semantic segmentation), computer vision and trustable and transparent AI systems.

MediaEval'21, December 6-8 2021, Bergen, Norway and Online

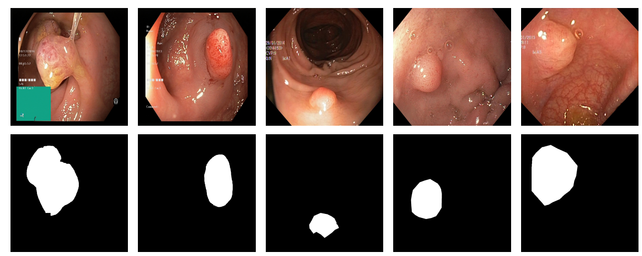

Figure 1: Examples taken from the development part of the polyp segmentation dataset HyperKvasir. Note that the images have been resized from their original image dimensions.

2 DATASET DETAILS

The provided dataset is based on HyperKvasir [1][1], which is currently the largest public gastrointestinal dataset. We combined the segmentation part of HyperKvasir with additional images and masks to create the development and testing datasets for this task. The development dataset contains 1, 360 images of polyps and corresponding image masks, while the test dataset consists of 200 image pairs collected from the same distribution as the development dataset. The additional images added to the development dataset were collected from the testing datasets that were used in two previous tasks, namely EndoTect [3] and last year's Medico [4]. The ground truth for the provided dataset was created by an experienced computer scientist, which was then verified by an expert gastroenterologist with over ten years of experience. Example images and corresponding segmentation masks can be seen in Figure 1.

3 TASK DESCRIPTIONS AND EVALUATION

The 2021 edition of the Medico Multimedia Task provides three different subtasks, namely *the polyp segmentation task*, *the efficient segmentation task*, and *the transparent machine learning systems task*. The polyp segmentation subtask is the only subtask that is required, the other two are optional. Each task allows for a total of five submissions each.

[1]https://datasets.simula.no/hyper-kvasir

3.1 Subtask 1: Polyp Segmentation

The *polyp segmentation task* targets high-performing polyp segmentation systems. Using the provided development dataset, participants are asked to develop models that automatically segment the presence of colon polyps in a given image. Submission to this task should be a zip file containing a predicted segmentation mask using the *.png* file format for each image in the testing dataset. Each predicted mask should use the same resolution as the input image and have the same filename.

Submissions will be evaluated based on the correctness of the predicted masks using various segmentation metrics like pixel accuracy, precision, recall, Sørensen–Dice coefficient (Dice), and Intersection over Union (IoU). The primary metric used to rank the submissions will be IoU. The participants will receive a *.csv* file containing the evaluation metrics for each run.

3.2 Subtask 2: Efficient Segmentation

The *efficient segmentation task* aims for efficient segmentation systems while still obtaining a satisfactory prediction accuracy. Model efficiency is measured in the number of frames that a model can process per second. The motivation behind this is the need for real-time detection systems used during live endoscopy procedures. For the system to be considered *real-time*, it should be able to process at least 30 frames per second. To participate in this subtask, participants must use the development dataset to train a polyp segmentation model. Furthermore, this task also requires the participants to submit a Docker image of their implementation to be evaluated on the organizers' hardware. The Docker submission should generate a *.csv* submission file that contains the name of the segmented image and the time (in seconds) used to perform the segmentation. A detailed description of the preparation and submission requirements of the Docker image is available on the official GitHub repository[2].

Models will be evaluated based on the performance metrics used to evaluate the polyp segmentation task and the number of frames that can be segmented per second. Submission will be ranked based on a balanced metric between predictive performance and speed. All submissions are evaluated on what can be considered consumer-grade hardware, that is, a computer running Arch Linux with an Intel Core i9-10900K processor, an Nvidia GeForce RTX 3090 graphics processing unit (GPU), and 32 gigabytes of RAM.

3.3 Subtask 3: Transparent Machine Learning Systems

The goal of the *transparent machine learning system task* is to promote more transparency in medical applications of machine learning. The motivation behind this task is rooted in a general lack of transparency in medical machine learning research [7]. A lot of work is often published using private data, closed-source implementations, and lackluster evaluations, making the systems not very reproducible or transparent. We leave it to the participants to determine what makes a machine learning system transparent. Still, some ideas include failure analysis, ablation studies, model explanations, open and commented source code, and detailed implementation descriptions.

Submissions to this task will be evaluated by a committee comprised of at least three computer scientists and expert gastroenterologists that are familiar with AI. The committee will evaluate the submissions from different perspectives. For example, the medical doctors will look at the system from a clinic point of view, assessing transparency based on how it can be used in the clinic. The computer scientists will look at the technical transparency of the submissions, like source code descriptions and the clarity of the implementation. Each team that submits to this task will receive a report on the level of transparency determined by the evaluation committee.

4 DISCUSSION AND OUTLOOK

Automatic segmentation of polyps in the gastrointestinal tract is a problem that is highly requested by medical doctors working in the field. Being more transparent about the work that goes into developing these methods would not only help doctors make more informed decisions on what systems should be used, but can also aid in further development by future researchers. We hope that this task will encourage the multimedia community to aid in the development of computer-assisted finding segmentation, and further motivate the use of transparent and open implementations of machine learning systems in medicine.

REFERENCES

[1] Hanna Borgli, Vajira Thambawita, Pia H Smedsrud, Steven Hicks, Debesh Jha, Sigrun L Eskeland, Kristin Ranheim Randel, Konstantin Pogorelov, Mathias Lux, Duc Tien Dang Nguyen, Dag Johansen, Carsten Griwodz, Håkon K Stensland, Enrique Garcia-Ceja, Peter T Schmidt, Hugo L Hammer, Michael A Riegler, Pål Halvorsen, and Thomas de Lange. 2020. HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Scientific Data* 7, 1 (2020), 283. https://doi.org/10.1038/s41597-020-00622-y

[2] Steven Hicks, Pål Halvorsen, Trine B. Haugen, Jorunn M. Andersen, Oliwia Witczak, Konstantin Pogorelov, Hugo L. Hammer, Duc Tien Dang Nguyen, Mathias Lux, and Michael Riegler. 2019. Medico multimedia task at MediaEval 2019. In *Proc. of MediaEval 2019 CEUR Workshop.*

[3] Steven A. Hicks, Debesh Jha, Vajira Thambawita, Pål Halvorsen, Hugo L. Hammer, and Michael A. Riegler. 2021. The EndoTect 2020 Challenge: Evaluation and Comparison of Classification, Segmentation and Inference Time for Endoscopy. In *Pattern Recognition. ICPR International Workshops and Challenges*, Alberto Del Bimbo, Rita Cucchiara, Stan Sclaroff, Giovanni Maria Farinella, Tao Mei, Marco Bertini, Hugo Jair Escalante, and Roberto Vezzani (Eds.). Springer International Publishing, Cham, 263–274.

[4] Debesh Jha, Steven Hicks, Krister Emanuelsen, Håvard Johansen, Dag Johansen, Thomas de Lange, Michael Riegler, and Pål Halvorsen. 2020. Medico Multimedia Task at MediaEval 2020: Automatic Polyp Segmentation. In *Proc. of MediaEval 2020 CEUR Workshop.*

[5] Konstantin Pogorelov, Michael Riegler, Pål Halvorsen, Steven Hicks, Kristin Ranheim Randel, Duc Tien Dang Nguyen, Mathias Lux, Olga Ostroukhova, and Thomas de Lange. 2018. Medico multimedia task at MediaEval 2018. In *Proc. of MediaEval 2018 CEUR Workshop.*

[6] Michael Riegler, Konstantin Pogorelov, Pål Halvorsen, Carsten Griwodz, Thomas Lange, Kristin Randel, Sigrun Eskeland, Duc Tien Dang Nguyen, Mathias Lux, and Concetto Spampinato. 2017. Multimedia for medicine: the medico task at Mediaeval 2017. In *Proc. CEUR Worksh. Multim. Bench. Worksh.*

[7] Hannah Stower. 2020. Transparency in medical AI. *Nature Medicine* (2020). https://doi.org/10.1038/s41591-020-01147-y

[2] https://github.com/multimediaeval/2021-Medico-Multimedia