ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

Leveraging endoscopic data with Contrastive Learning for Crohn's disease detection

Robin Ghyselinck, Jérôme Fink, Bruno Dumas, Benoît Frenay *

University of Namur - NaDI
Rue Grangagnage, 21, 5000 Namur - Belgium

Abstract. This study contributes to the automatic detection of Crohn's Disease (CD), a gastrointestinal inflammatory condition. In particular, our approach deals with the challenge of data scarcity for CD by pre-training Vision Transformers (ViT) on Hyper-Kvasir and LDPolyp, two large colonscopic datasets that represent over one million images from a similar domain, using a Contrastive Loss (CL) mechanism. This approach significantly outperforms models pre-trained on ImageNet as well as models pre-trained with a Cross-Entropy Loss on the Crohn-IPI dataset.

1 Introduction

Since its inception in the mid-1990s, Wireless Capsule Endoscopy (WCE) has dramatically transformed the landscape of gastrointestinal diagnostics by facilitating early treatment[1], in particular for Crohn's Disease (CD), an inflammatory bowel disease [2]. For a typical WCE procedure, between 3 to 4 meters of the small bowel are recorded, such that gastroenterologists have to analyse 50,000 images per patient, requiring between 30 and 60 minutes [3].

The creation of automated systems to assist the diagnosis would save time and reduce the risk of errors. The creation of the Crohn-IPI dataset [4] marked a significant milestone in CD research, offering 3,498 images from 63 patients annotated for 7 types of CD lesions. Deep learning has proven successful on this dataset in the past for detecting the presence of CD lesions on an image.

Our research introduces a novel approach to CD diagnosis through the development of models pre-trained on 1 million images from two public colonoscopic datasets, the LDPolyp [5] and hyper-kvasir[6] datasets, with a contrastive loss (CL) [7] and vision transformer (ViT) [8]. This methodology harnesses the power of large-scale image datasets to overcome the data scarcity challenge, and fine-tunes a final model for a binary classification task on the Crohn-IPI dataset (i.e., anomalous or normal image). This approach significantly outperforms models with identical architecture that are pre-trained on ImageNet or that are pre-trained on the same datasets using a cross-entropy loss as well as replicated results from Vallee [4] using a ResNet-34. Our work shows that small datasets that perform a specific task can benefit from pre-training using larger datasets from different, yet related domain. Section 2 introduces previous works that deals with the use of deep learning for CD detection and explains contrastive

*The present research benefited from computational resources made available on Lucia, the Tier-1 supercomputer of the Walloon Region, infrastructure funded by the Walloon Region under the grant agreement n°1910247.

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

learning, Section 3 describes the methodology that is followed, Section 4 describes the experimental setup, Section 5 presents and discusses the results that are obtained. Finally, Section 6 concludes on the study and opens the path to further works.

2 Related works

Vallée et al. [4] introduced the first version of the Crohn-IPI dataset, containing 3,218 images from 39 patients with annotations from two gastroenterologists. Each image is marked either as pathological (with one out of six possible classes) or normal. In their initial work, Vallée et al. merged all six anomalies as one class and obtained a 89.3% accuracy [9]. This work was further continued with a modified dataset [4] that contains 3,498 images with seven anomalous classes and a normal one. Those classes were converted into normal or abnormal, and four models are assessed on the binary classification tasks where the authors reach a 94.6% accuracy with a 5-fold cross-validated training and a ResNet34 [10]. However, the average results of those models are presented with no confidence interval nor any indication of the variance between individual folds. Moreover, no hyper-parameter nor any experimental information are shared. Also, the authors claimed that the dataset is balanced, which is subject to discussion because 61% of the data is normal. Consequently, it makes it difficult to interpret the robustness of the results from a statistical standpoint. Also, with no indication of the training setup, results are hard to replicate.

Xing et al. [11] used Wireless Capsule Endoscopy (WCE) data from two private videos (35,053 unlabeled images) for self-supervised contrastive pre-training, and fine-tuned their models on CD classification with the Crohn-IPI dataset. They used several decoders, such as MoCo v2, BYOL, and Barlow Twins with reported accuracy between 91% and 93%. However, their work used pre-training with same domain data (i.e. WCE) in an unsupervised fashion, which means that it did not leverage the information that is contained in the annotations. Moreover, only the accuracy is presented as a performance metric, with no indication of the variance between the 5 folds. The private nature of their WCE data makes it impossible to replicate their results.

Contrastive Learning (CL) has become popular for pre-training. It is an effective method for learning latent space representations of data by maximizing the agreement between differently augmented views of the same data, while minimizing the similarity between representations of distinct data points at the same time [12]. This approach has shown superior capability in learning latent space representations compared to traditional cross-entropy loss, primarily due to its emphasis on understanding the nuances and variances within the data itself. Typically, one can use CL in the pre-training phase of a model to offer a significant advantage, in particular for fine-tuning on a given downstream task where annotated data is scarce.

This work showcases the use of CL for pre-training Vision Transformers (ViT) on data from another medical domain where they are more abundant (i.e., en-

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

doscopic images). These models are subsequently fine-tuned on the Crohn-IPI dataset for a binary classification task. Contrary to previous work, results are clearly discussed and take into account the unbalanced nature. We show the interest of knowledge transfer from another domain with the use of CL.

3 Methodology

This section explains the methodology used to exploit the Crohn-IPI dataset [9] for binary classification. Data are converted from 8 classes into two: normal or abnormal. Three different experiments are conducted to assess the benefit of using i) a larger dataset to pre-train a ViT, ii) a larger dataset from a domain close to Crohn-IPI and, iii) contrastive learning with a larger dataset.

3.1 Detection using Backbones pre-trained on Imagenet

To constitute a baseline, one can use a Vision Transformer (ViT) [8] pre-trained on ImageNet [13], and fine-tuned for the specific binary classification task. To do so, its classification head is removed, and replaced with a linear layer with input size 768, and output size 1. In addition, a ResNet-34 pre-trained on ImageNet is fine-tuned on the same task with the intent to replicating the same model as Vallée [4]. For the binary classification task, we replaced the fully-connected layer with a linear layer of input size 512, output size 256, a ReLU, and a last fully-connected layer of input size 256, output size 1.

3.2 Detection using Supervised Cross-Entropy Loss Backbone

ImageNet contains 1 million images of 1,000 different classes that are not close to the domain being researched, i.e., WCE data. To assess the benefit from pre-training a backbone on data from a closer domain, one can use colonoscopy data, and fine-tune the model for binary-classification in a similar fashion as described above. We pre-train a ViT with a Cross-Entropy (CE) loss by leveraging two public colonoscopy datasets, LDPolyp [5] and Hyper-Kvasir [6], which have been merged by Tian et al. [14]. In their combined dataset, Tian et al. provide binary annotations (normal or anomaly) for more than one million colonoscopic images that were initially classified in more than two classes. As in the previous experiment, this backbone is then fine-tuned on the Crohn-IPI dataset.

3.3 Detection using Supervised Contrastive Loss Backbone

The supervised Contrastive Loss (CL) [7] is chosen to leverage the labels available in the two large datasets. This loss can be mathematically expressed as

$$L_i = \frac{-1}{|\{P(i)\}|} \sum_{j \in \{P(i)\}} \log \frac{\exp(\mathbf{z_i} \cdot \mathbf{z_p}/\tau)}{\sum_{a \in \{A(i)\}} \exp(\mathbf{z_i} \cdot \mathbf{z_a}/\tau)} \tag{1}$$

where $\{P(i)\}$ represents the set of indices of all positive examples in the batch for the anchor i, excluding i itself. $\mathbf{z_l}$ contains the data embeddings. $\{A(i)\}$ denotes

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

the set of indices for all other examples in the batch, excluding i, encompassing both positive and negative examples. The $\cdot$ symbol denotes the dot product and exp the exponent function. Finally, τ is a temperature scaling parameter that moderates the separation margin. Using the same ViT backbone, the contrastive model is pre-trained on the two colonoscopic datasets. After training, this model is fine-tuned for the binary classification on Crohn-IPI dataset as stated in the previous sections.

4 Experimental setup

This section explains pre-training and fine-tuning of the experiments as well as selected performance metrics.

4.1 Pre-training

The models trained with CL and CE on the colonoscopic dataset were both trained using 4 Tesla A100 in parallel for 50 epochs with PyTorch 2.2.1, with a one-cycle cosine learning rate warm up [15] that goes up to 1×10^{-3} is used. A batch size of 512, a drop out rate of 0.3, a temperature parameter τ of 0.1, the AdamW [16] optimizer and without any color channel regularization. Several values were tested for those parameters using grid-search and the best performing models were kept. To validate the best models, the one leading to the smallest loss are retained. Default weights for ViT pre-trained on ImageNet are collected from the PyTorch library.

4.2 Model fine-tuning

The fine-tuned models are trained with a 5-fold cross-validation strategy using 5 Tesla A100, a batch size of 256, an initial learning rate of 1×10^{-2}, a one-cycle cosine learning rate warm up scheduler is used. No color channel regularization and no dropout are applied. The following data augmentation techniques are used: multiple of 90° random rotations, random flip and mirror, resizing to 236×236 pixels, random cropping to 224×224 pixels. Several values were experimented with, and the ones leading to the best results were kept. To manage the data imbalancing, a multinomial sampler is used to ensure that each training epoch draws 50% of images from each class.

4.3 Performance metrics

To account for the data imbalancing when reporting performance metrics, the balanced accuracy, precision and recall are assessed. Moreover, the Jaccard score (Intersection over Union (IoU)) gives an indication of the quality of the predictions while the Area Under the Curve (AUC) Receiver Operating Characteristics (ROC) is a common choice for assessing the quality of models dealing with medical data.

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

5 Experimental results

Table 1 shows the average performance metrics with the 5-fold cross-validation for each of the three models together with their 95% Confidence Interval (CI). The model that uses the backbone pre-trained with a contrastive loss shows superior performance across all metrics. The results obtained using the CB models significantly outperforms the other methods across all five metrics and validate the interest of using a contrastive loss in pre-training a backbone model on data that are close to the domain. Quite surprisingly, a pre-training on endoscopic data with a cross-entropy loss does not result in better results, compared to a model pre-trained on ImageNet, whereas one could have expected that data from a closer domain would be beneficial. One possible explanation is that the CL helps capture domain-specific characteristics more effectively than a Cross-Entropy loss.

Table 1: The Contrastive Backbone (CB) gives better results than other methods. Average results on 5-fold for CB, the Cross-Entropy Backbone (C-E B), the ViT ImageNet Backbone (VIB), and the ResNet-34 ImageNet Backbone (RIB). Data in bold show the best results with a 95% confidence interval.

Meth.	**Bal. Acc.**	**Prec.**	**Recall**	**IOU**	**ROC AUC**
CB	**92.1**(±1.7)	**88.8**(±2.3)	**92.4**(±3.5)	**82.7**(±3.0)	**96.9**(±0.1)
C-E B	86.3(±2.2)	82.5(±4.6)	85.5(±4.7)	72.3(±3.5)	93.4(±1.4)
VIB	89.2(±1.6)	85.3(±1.9)	89.4(±2.6)	77.4(±2.7)	95.1(±0.1)
RIB	85.6(±2.3)	81.5(±3.1)	83.7(±4.0)	70.3(±4.0)	92.0(±1.7)

6 Conclusion

Our study showcases a compelling methodology for improving Crohn's Disease (CD) diagnosis with the use of deep learning techniques. By adopting an approach that pre-trains models on vast amounts of colonoscopic data using Contrastive Loss and vision transformers, the data scarcity issue in the field of CD detection is mitigated. Indeed, a contrastive loss backbone pre-trained on the LDPolyp and Hyper-Kvasir datasets significantly outperforms models pre-trained on ImageNet or trained from scratch with a cross-entropy loss. This emphasises the importance of using data from a close domain as well as the interest of contrastive learning in medical image analysis. Moreover, the comparative analysis of different pre-training strategies depicts a nuanced relationship between domain specificity and model effectiveness, suggesting that closer domain data does not always guarantee superior results. In the future, the proposed backbone could be used on other endoscopic datasets, such as bronchoscopies, laparoscopies or rhinoscopies. Also, one could assess the impact of additional data on pre-training or evaluate the performance on multi-class classification.

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

References

[1] Rami Eliakim. The impact of panenteric capsule endoscopy on the management of crohn's disease. *Therapeutic Advances in Gastroenterology*, 10(9):737–744, July 2017.

[2] Joana Torres, Saurabh Mehandru, Jean-Frédéric Colombel, and Laurent Peyrin-Biroulet. Crohn's disease. *The Lancet*, 389(10080):1741–1755, April 2017.

[3] Mark E. McAlindon, Hey-Long Ching, Diana Yung, Reena Sidhu, and Anastasios Koulaouzidis. Capsule endoscopy of the small bowel. *Annals of Translational Medicine*, 4(19):369–369, October 2016.

[4] Rémi Vallée, Astrid De Maissin, Antoine Coutrot, Harold Mouchère, Arnaud Bourreille, and Nicolas Normand. Crohnipi: An endoscopic image database for the evaluation of automatic crohn's disease lesions recognition algorithms. In Barjor S. Gimi and Andrzej Krol, editors, *Medical Imaging 2020: Biomedical Applications in Molecular, Structural, and Functional Imaging*. SPIE, February 2020.

[5] Yiting Ma, Xuejin Chen, Kai Cheng, Yang Li, and Bin Sun. *LDPolypVideo Benchmark: A Large-Scale Colonoscopy Video Dataset of Diverse Polyps*, page 387–396. Springer International Publishing, 2021.

[6] 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. Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Scientific Data*, 7(1), August 2020.

[7] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2020.

[9] Remi Vallee, Astrid de Maissin, Antoine Coutrot, Nicolas Normand, Arnaud Bourreille, and Harold Mouchere. Accurate small bowel lesions detection in wireless capsule endoscopy images using deep recurrent attention neural network. In *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, September 2019.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2016.

[11] Jing Xing and Harold Mouchere. Contrastive self-supervised learning on crohn's disease detection. In *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, December 2022.

[12] Ting Chen, Calvin Luo, and Lala Li. Intriguing properties of contrastive losses. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 11834–11845. Curran Associates, Inc., 2021.

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, June 2009.

[14] Yu Tian, Guansong Pang, Fengbei Liu, Yuyuan Liu, Chong Wang, Yuanhong Chen, Johan Verjans, and Gustavo Carneiro. *Contrastive Transformer-Based Multiple Instance Learning for Weakly Supervised Polyp Frame Detection*, page 88–98. Springer Nature Switzerland, 2022.

[15] Leslie N. Smith and Nicholay Topin. Super-convergence: Very fast training of residual networks using large learning rates, 2018.

[16] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.