Polyp segmentation generalisability of pretrained backbones

Author
Edward Sanderson, Bogdan J. Matuszewski – Computer Vision and Machine Learning (CVML) Group, School of Engineering, University of Central Lancashire, Preston, UK Institute

Citation
Sanderson, E., Matuszewski, B.J. Polyp segmentation generalisability of pretrained backbones.

Introduction

Due to the low availability of annotated data for training polyp segmentation models, e.g. Sanderson and Matuszewski (2022), which typically take the form of an autoencoder with UNet-style skip connections (Ronneberger et al., 2015), it is common practice to pretrain the encoder, also known as the backbone. This has almost exclusively been done in a supervised manner with ImageNet-1k (Deng et al., 2009). However, we recently demonstrated that pretraining backbones in a self-supervised manner generally provides better fine-tuned performance, and that models with ViT-B (Dosovitskiy et al., 2020) backbones typically perform better than models with ResNet50 (He et al., 2016) backbones (Sanderson and Matuszewski, 2024).

In this paper, we extend this work to consider generalisability. I.e., we assess performance on a different dataset to that used for fine-tuning, accounting for variation in network architecture and pretraining pipeline (algorithm

and dataset). This reveals how well models generalise to a somewhat different distribution to the training data, which arise in deployment as a result of different cameras, demographics of patients, and other factors. Our results provide further insights into the strengths and weaknesses of existing architectures and pretraining pipelines that should inform the future development of polyp segmentation models.

Analysis

We consider 12 polyp segmentation models pretrained and fine-tuned in a previous study (Sanderson and Matuszewski, 2024), specifically those fine-tuned on Kvasir-SEG (Jha et al., 2020). Each model is either a ResNet50 encoder with a DeepLabV3+ (Chen et al., 2018) decoder, or a ViT-B encoder with a DPT (Ranftl et al., 2021) decoder. Additionally, each model was pretrained on either Hyperkvasir-unlabelled (Borgli et al., 2020) or ImageNet-1k in a self-supervised manner using either MoCo v3 (Chen et al., 2021), Barlow Twins (Zbontar et al., 2021) (ResNet50 only), or MAE (He et al., 2022) (ViT-B only); or pretrained in a supervised manner (ImageNet-1k only); or not pretrained at all.

We evaluate performance on the full CVC-ClinicDB dataset (Bernal et al., 2015) with mDice, mIoU, mPrecision, and mRecall. The results are reported in Table 1, where we also specify each model's rank on each metric, as well as any change in rank relative to the model's evaluation on the Kvasir-SEG test set (Sanderson and Matuszewski, 2024). The results show that self-supervised pretraining on ImageNet-1k generally provides the best generalisation, that supervised pretraining on ImageNet-1k is generally better than self-supervised pretraining on Hyperkvasir-unlabelled, and that any considered pretraining is better than no pretraining. These findings are consistent with the evaluation on the Kvasir-SEG test set.

However, the model pretrained with MAE on ImageNet-1k, which performs best on the Kvasir-SEG test set, reduces its rank on every metric, notably dropping from rank 1 to 4 on mDice. In contrast, models with a ResNet50 backbone generally improve their ranking, implying greater generalisability

TABLE 1: Performance of models fine-tuned on the Kvasir-SEG training set and tested on CVC-ClinicDB. In addition to reporting the value of each metric, we also indicate the rank of each model, as well as any change in this rank relative to the model's evaluation on the Kvasir-SEG test set. For conciseness, we abbreviate Hyperkvasir-unlabelled to HK, ImageNet-1k to IN, and Barlow Twins to BT

Backbone	Pretraining		mDice		mIoU		mPrecision		mRecall	
arch.	Data	Algo.	Value	Rank	Value	Rank	Value	Rank	Value	Rank
ResNet50	HK	MoCo v3	0.789	9 (↑1)	0.686	10	0.785	10	0.856	3 (↑7)
		BT	0.801	8 (↑1)	0.709	8 (↑1)	0.831	8 (↑1)	0.848	7 (↓2)
	IN	MoCo v3	0.843	1 (↑3)	0.760	1 (↑3)	0.867	3 (↑5)	0.874	1
		BT	0.826	5	0.735	6 (↑1)	0.858	6	0.854	4 (↑2)
		Supervised	0.822	6	0.735	5	0.899	1 (↑4)	0.811	10 (↓2)
	None	None	0.520	12	0.394	12	0.496	12	0.724	12
ViT-B	HK	MoCo v3	0.789	10 (↓2)	0.696	9 (↓1)	0.812	9 (↓2)	0.848	8 (↓1)
		MAE	0.828	3	0.743	3	0.852	7 (↓4)	0.858	2 (↑1)
	IN	MoCo v3	0.830	2	0.742	4 (↓2)	0.861	4 (↓2)	0.849	5 (↓3)
		MAE	0.827	4 (↓3)	0.746	2 (↓1)	0.868	2 (↓1)	0.848	6 (↓2)
		Supervised	0.809	7	0.717	7 (↓1)	0.860	5 (↓1)	0.832	9
	None	None	0.637	11	0.519	11	0.670	11	0.759	11

than models with a ViT-B backbone, which generally experience a drop in ranking, and the best generalisation is achieved by the model with a ResNet50 backbone that was pretrained on ImageNet-1k using MoCo v3, notably improving from rank 4 to 1 on mDice. To better understand this, we compare the distribution of instance-wise Dice scores from each model's evaluation on the Kvasir-SEG test set against the distribution from its evaluation on CVC-ClinicDB in Fig. 1. This indicates that all models experience a drop in overall performance that primarily arises from a higher variance. However, the portion of each distribution for the highest Dice scores shows that most models with ResNet50 backbones achieve better performance on some instances of CVC-ClinicDB than any in the Kvasir-SEG test set, while models with ViT-B backbones fail to exceed their maximum Dice score across the Kvasir-SEG test set when evaluated on CVC-ClinicDB. We verify that all models experience a drop in performance, and quantify the relative drop, in Fig. 2, which reveals that most models with ResNet50

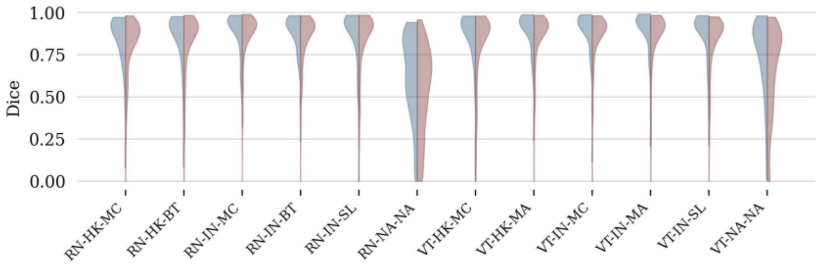

FIGURE 1
Comparison of the distribution of instance-wise Dice score from each model's evaluation on the Kvasir-SEG test set (blue) against the distribution from its evaluation on CVC-ClinicDB (red). For conciseness, we denote ResNet50s with RN, ViT-Bs with VT, Hyperkvasir-unlabelled with HK, ImageNet-1k with IN, MoCo v3 with MC, Barlow Twins with BT, MAE with MA, supervised pretraining with SL, and no pretraining with NA-NA.

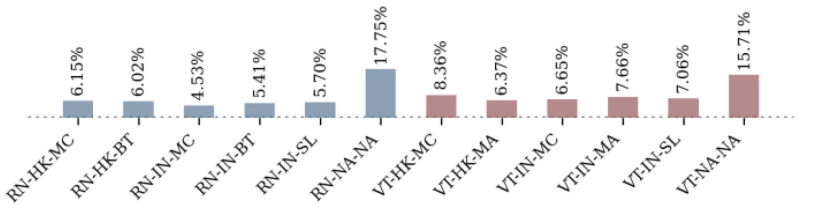

FIGURE 2
Relative drop in mDice from each model's evaluation on the Kvasir-SEG test set to its evaluation on CVC-ClinicDB. For conciseness, we denote ResNet50s with RN, ViT-Bs with VT, Hyperkvasir- unlabelled with HK, ImageNet-1k with IN, MoCo v3 with MC, Barlow Twins with BT, MAE with MA, supervised pretraining with SL, and no pretraining with NA-NA. For clarity, the results for ResNet50 models are coloured blue and the results for ViT-B models are coloured red.

backbones do indeed experience less of a drop, potentially as a result of their improvement in maximum Dice score, explaining the improvement in ranking.

Conclusion

In this paper, we showed that previous findings, regarding pretraining pipelines for polyp segmentation, hold true when considering generalisability. However, our results imply that models with ResNet50 backbones typically generalise better, despite being outperformed by models with ViT-B backbones in evaluation on the test set from the same dataset used for fine-tuning. We expect that this is a result of the larger complexity of the models with ViT-B backbones allowing for overfitting on the distribution underlying the training data. However, this challenges the assumption that the considered pretraining pipelines should help prevent this, and more work is required to better understand the relationships between architecture, pretraining pipeline, and performance on different distributions of data, as well as the amount of training data.

References

Bernal, J., Sánchez, F. J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., and Vilariño, F. (2015). Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. Computerized medical imaging and graphics 43, 99–111

Borgli, H., Thambawita, V., Smedsrud, P. H., Hicks, S., Jha, D., Eskeland, S. L., et al. (2020). Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. Scientific data 7, 283

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European conference on computer vision (ECCV). 801–818

Chen, X., Xie, S., and He, K. (2021). An empirical study of training self-supervised vision transformers. CoRR abs/2104.02057

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition (Ieee), 248 255

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. (2022). Masked autoencoders are scalable vision learners. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 16000–16009

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition. 770–778

Jha, D., Smedsrud, P. H., Riegler, M. A., Halvorsen, P., de Lange, T., Johansen, D., et al. (2020). Kvasir-seg: A segmented polyp dataset. In MultiMedia Modeling: 26th International Conference, MMM 2020 Daejeon, South Korea, January 5–8, 2020, Proceedings, Part II 26 (Springer), 451–462

Ranftl, R., Bochkovskiy, A., and Koltun, V. (2021). Vision transformers for dense prediction. In Proceedings of the IEEE/CVF international conference on computer vision. 12179–12188

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18 (Springer), 234–241

Sanderson, E. and Matuszewski, B. J. (2022). Fcn-transformer feature fusion for polyp segmentation. In Annual conference on medical image understanding and analysis (Springer), 892–907

Sanderson, E. and Matuszewski, B. J. (2024). A study on self-supervised pretraining for vision problems in gastrointestinal endoscopy. IEEE Access 12, 46181–46201. doi:10.1109/ACCESS.2024.3381517

Zbontar, J., Jing, L., Misra, I., LeCun, Y., and Deny, S. (2021). Barlow twins: Self-supervised learning via redundancy reduction. In International Conference on Machine Learning (PMLR), 12310–12320