Multi-task SwinV2 transformer for polyp classification and segmentation

Author
Kerr Fitzgerald – Computer Vision and Machine Learning (CVML) Group, University of Central Lancashire, Preston, United Kingdom
Jorge Bernal – Computer Vision Center and Computer Science Department, Universitat Autònoma de Barcelona, Barcelona, Spain
Yael Tudela – Computer Vision Center and Computer Science Department, Universitat Autònoma de Barcelona, Barcelona, Spain
Bogdan J. Matuszewski – Computer Vision and Machine Learning (CVML) Group, University of Central Lancashire, Preston, United Kingdom

Citation
Fitzgerald, K., Bernal, J., Tudela, Y., Matuszewski, B.J. Multi-task SwinV2 transformer for polyp classification and segmentation.

Abstract
Motivated by the aim of improving polyp classification performance on the CVC-HDClassif dataset, joint classification-segmentation multi-task learning using a SwinV2 Transformer UNet based architecture has been explored.

Introduction
The gold standard of polyp screening and removal procedures is widely considered to be colonoscopy, which allows clinicians to navigate through the colon and visually inspect for abnormalities in real time. However, colonoscopy does have limitations as not all polyps are consistently identified

(A.M. Leufkens et al., 2012). Therefore research has focused on developing computer aided systems to support clinicians in improving the detection rate and characterization of polyps, with deep learning systems achieving current state-of-the-art performance across a variety of polyp imaging tasks.

Whilst polyp segmentation models are showing signs of reaching maturity (K. Fitzgerald et al., 2024) (RG. Dumitru, D. Peteleaza & C. Craciun, 2023), polyp classification has been identified as a critical area for further research. One reason for this is the lack of openly available datasets which contain polyp classification labels. The novel CVC-HDClassif dataset (Y. Tudela et al., 2023) contains 788 training, 113 validation, and 225 testing images, with corresponding ground truth segmentation maps and polyp histology labels (adenomatous vs non adenomatous).

System Design and Methodology

Previous models for medical imaging multi-task learning employ UNet (O. Ronneberger et al., 2015) style architectures with the addition of a classification head at a selected stage of the network (B. Oliveira et al., 2023) (C. Li, J. Liu & J. Tang, 2024). Such multitask learning models lead to improved classification performance, which is hypothesized to occur due to the mixing of detailed spatial information needed for segmentation and global contextual information needed for classification. Motivated by the improved classification performance of these models and the excellent performance of SwinV2 systems when used as encoders in segmentation systems, a SwinV2 UNet system (Z. Liu et al., 2022) with a classification head added the to the final encoder layer was developed for joint polyp classification and segmentation. A description of the SwinV2 UNet style architecture can be found in (K. Fitzgerald et al., 2024). The classification head flattens the tensor from the final encoder stage and then passes this sequentially through two Fully Connected (FC) layers using a dropout rate of 50%. A final FC layer and activation function is used to generate a final classification prediction. The architecture of the SwinV2 UNet segmentation-classification model is shown in Figure 1.

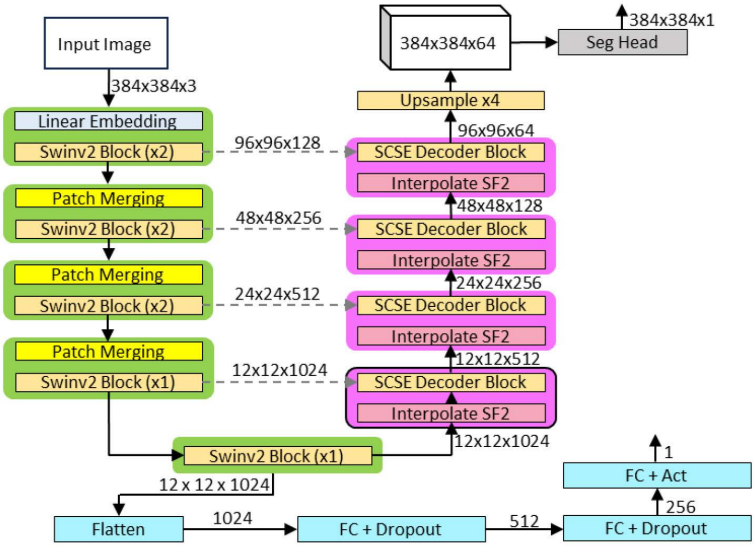

FIGURE 1
SwinV2 UNet system with a classification head added to the final block of the encoder. The decoder block makes use of 'Spatial and Channel Squeeze and Excitation' (SCSE) modules.

The SwinV2 UNet model was implemented using PyTorch and the encoder was initialized using ImageNet-22K (J. Deng et al., 2009) weights available from the PyTorch Image Models Library (R. Wightman, 2019). Since the CVC-HDClassif test split has not been released by the dataset authors, only the training and validation splits were utilized in this study. Due to the relatively small number of images available for training and validation, standard on-the-fly data augmentations (e.g. color variations and geometrical transforms) were applied to the training set using the Albumentations library (A. Buslaev et al., 2018). Static data augmentations were applied to the validation set to stabilize accuracy scores. The AdamW algorithm (I. Loshchilov & F. Hutter, 2019) was used for model optimization alongside a cosine learning rate schedular. To train the SwinV2 UNet model on the CVCHDClassif dataset, a combined segmentationclassification loss function (EQ1) was used.

$$L_{total} = (\gamma_{class} \cdot L_{class}) + (\gamma_{seg} \cdot L_{seg}) \quad \text{EQ1}$$

Where γ_{class} represents a classification weighting factor, L_{class} is the Binary Cross Entropy (BCE) classification loss (E. Bekele & W. Lawson, 2019), γ_{seg} represents a segmentation weighting factor, and L_{seg} represents the segmentation loss which is a combination of the pixel based Binary Cross Entropy (BCE) loss and dice loss (S. Jadon, 2020). The mean training and validation classification losses and accuracies were recorded to examine model performance. Ablation studies showed that setting the classification weight to a very small value (γ_{class} = 1E^{-6}) for the first 15 training epochs allowed the model to achieve strong segmentation performance, before then changing the classification weight to the value of 1 (γ_{class} = 1. The performance of the SwinV2 UNet multitask learning model was also compared to a standard SwinV2 classification model and the fully convolutional EfficienetV2M model (M. Tan & Q. Le, 2021). These models used the same training methodology (excluding the task of segmentation) and required fewer training epochs before signs of overfitting occurred.

Preliminary Results and Discussion

The SwinV2-UNet model shows excellent segmentation performance on the validation set, achieving 90.88 mDice and 85.13 mIoU scores. The training and validation classification losses and accuracies are shown for each model in Figure 2.

For the classification task, the SwinV2-UNet model reached a maximum accuracy of 84.82%, which represents a substantial improvement over the maximum accuracy of 78.86% achieved by the SwinV2 classification model. The SwinV2 classification model is likely overfitting to the training data due to the limited dataset size and network complexity. This highlights the potential for multi-task learning approaches to enhance generalizability performance on classification tasks by leveraging spatial information supplied by segmentation data. The EfficientNetV2 model achieved the highest maximum validation accuracy of 85.42%. The EfficientNetV2 model is likely to offer benefits over Transformer based architectures for small dataset sizes

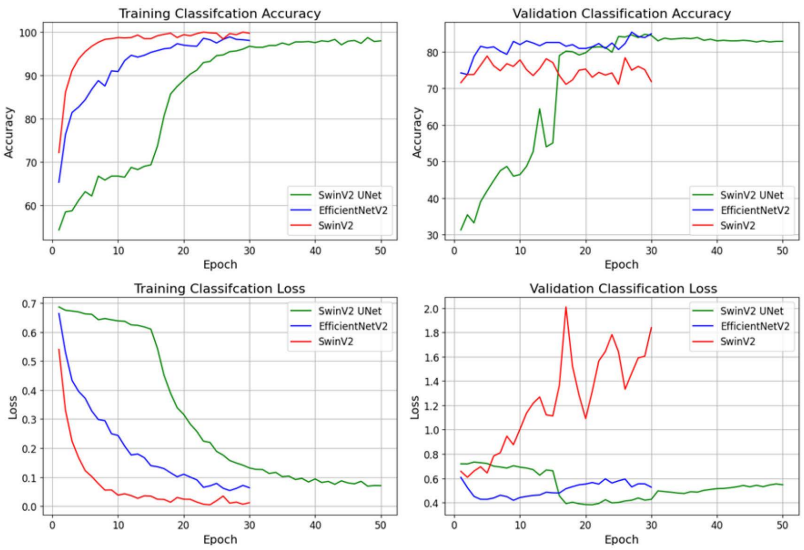

FIGURE 2
Training and validation performance of the models. The increase in classification accuracy at epoch 15 for the SwinV2 UNet model is due to the classification weight change.

due to the inherent inductive biases contained within fully convolutional architectures (A. Dosovitskiy et al., 2021). Further model refinements and larger multitask polyp segmentation-classification datasets will be beneficial to fully investigate and leverage the advantages of multi-task learning frameworks.

References

A. Buslaev et al., 2018. *Albumentations: Fast and Flexible Image Augmentations.* [Online] Available at: ttps://albumentations.ai/[Accessed 16 Janurary 2024].

A. Dosovitskiy et al., 2021. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.* Virtual, s.n.

A.G. Roy, N. Navab & C. Wachinger, 2018. *Concurrent Spatial and Channel 'Squeeze & Excitation' in Fully Convolutional Networks.* s.l., s.n., p. 421–429.

A.M. Leufkens, M.G.H. van Oijen, F.P. Vleggaar & P.D. Siersema, 2012. Factors influencing the miss rate of polyps in a back-to-back colonoscopy study. *Endoscopy,* 44(5), pp. 470-475.

B. Oliveira et al., 2023. A multi-task convolutional neural network for classification and segmentation of chronic venous disorders. *Nature Scientific Reports.*

D.A. Corley, 2014. Adenoma Detection Rate and Risk of Colorectal Cancer and Death. *New England Journal of Medicine,* Volume 370, pp. 1298-1306.

E. Sanderson & B.J. Matuszewski, 2022. *FCN-Transformer Feature Fusion for Polyp Segmentation.* s.l., s.n., pp. 892-907.

I. Loshchilov & F. Hutter, 2019. *Decoupled Weight Decay Regularization.* New Orleans, s.n.

J. Deng et al., 2009. *ImageNet: A large-scale hierarchical image database.* s.l., s.n.

J. Lee, S.W. Park, Y.S. Kim, K.J. Lee, H.S, P.H. Song, W.J. Yoon & J.S. Moon, 2017. Risk factors of missed colorectal lesions after colonoscopy. *Medicine,* 96(27).

K. Fitzgerald et al., 2024. Polyp Segmentation With the FCB-SwinV2 Transformer. *IEEE Access,* Volume vol. 12, pp., pp. 38927-38943.

M. Abe, 2022. *Swin V2 Unet/Upernet.* [Online]
Available at: https://www.kaggle.com/code/abebe9849/swin-v2-unet-upernet [Accessed January 2023].

M. Tan & Q. Le, 2021. *EfficientNetV2: Smaller Models and Faster Training.* s.l., s.n.

M.J. Whitson, C.A. Bodian, J. Aisenberg & L.B. Cohen, 2012. Is production pressure jeopardizing the quality of colonoscopy? A survey of U.S. endoscopists' practices and perceptions. *Gastrointestinal Endoscopy,* 75(3), pp. 641-648.

N.H. Kim, Y.S. Jung, W.S. Jeong, H.J. Yang, S.K. Park, K. Choi & D.I. Park, 2017. Miss rate of colorectal neoplastic polyps and risk factors for missed polyps in consecutive colonoscopies. *Intestinal Research,* 15(3), pp. 411-418.

O. Ronneberger et al., 2015. *U-Net: Convolutional Networks for Biomedical Image Segmentation.* s.l., s.n., pp. 234-241.

P. Iakubovskii, 2019. Segmentation Models Pytorch. *GitHub Repository.*

R. Wightman, 2019. *PyTorch Image Models.* s.l., s.n.

RG. Dumitru, D. Peteleaza & C. Craciun, 2023. Using DUCK-Net for polyp image segmentation. *Nature Scientific Reports,* Volume 13.

World Health Organization, 2023. *Colorectal cancer.* [Online] Available at: https://www.who.int/news-room/fact-sheets/detail/colorectal-cancer [Accessed December 2023].

Y. Tudela et al., 2023. *Towards Fine-Grained Polyp Segmentation and Classification.* s.l., Cham: Springer Nature, pp. 32-42.

Z. Liu et al., 2022. *Swin Transformer V2: Scaling Up Capacity and Resolution.* s.l., s.n., pp. 12009-12019.

Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin & B. Guo, 2021. *Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows.* s.l., s.n., pp. 10012-10022.