Mask2Tasks: Leveraging Segmentation to Enhance Classification Performance in Histopathological Colorectal Images

Hieu Le Xuan, Minh Hoang Le, Viet V Truong, Huy Phan Quang & Hoang Vu Huy
{lexuanhieu131297}@gmail.com

Abstract

In this study, we explore the enhancement of colorectal image classification accuracy with the aid of a segmentation task. We introduce Mask2Tasks, a deep neural network for joint colorectal image classification and segmentation which is trained using a novel two-stage training approach. Numerical results have demonstrated its effectiveness in both classification and multi-task learning scenarios.

1 Introduction

Recent advances in deep learning have proven its potential in the medical imaging domain; one such application is the diagnosis of histopathological colorectal images. However, the accuracy and robustness of deep learning algorithms can be severely undermined by the lack of training data. In such data sparsity regimes, multi-task learning (MTL) emerges as a viable approach that can improve the generalization, efficiency, and interpretability of deep learning models thanks to shared features that can be useful for multiple related tasks (Lee & Son (2022)). On the other hand, among the most pressing issues in multi-task learning is the conflicting gradient problem (Liu et al. (2023)) where certain tasks dominate the others in different training stages (Lee et al. (2021)) leading to training suboptimality or even failure to converge.

In this work, we investigate the utilization of multi-task learning methods in improving the performance of histopathological colorectal image classification. First, we propose a deep learning architecture, named Mask2Tasks, along with a novel two-stage training strategy for joint colorectal image classification and segmentation. Second, we prove by numerical results that our training scheme effectively enhances the performance of the diagnosis of colorectal histopathological, as well as consistently outperforms several popular multi-task learning methods in the considered settings. Finally, post-hoc analyses of the training dynamics and the learned deep features reveal interesting insights for the improved performance of the proposed method, thus suggesting its suitability for multi-task learning in colorectal histopathological imaging.

2 Methodology

Our proposed Mask2Tasks architecture (see Figure 1) is derived from Mask2Former, a powerful image segmentation architecture (Cheng et al. (2022)). To facilitate multi-task functionality, a classifier head comprising one global average pooling layer (GAP) and three subsequent fully connected layers with ReLU activation functions is integrated into the original architecture. This classifier head employs low-level features given by the encoder and is trained with a standard cross-entropy loss.

Our Mask2Tasks is trained using a novel two-stage training strategy. In the initial phase, the model is trained for histopathological image segmentation. In the second phase, the encoder is frozen and the classification head is fine-tuned for the classification task. By first focusing on the segmentation task, it is expected that the encoder can learn more robust and diverse features in a faster, more optimized manner. The inspiration for this design is discussed in the next section.

Experiments: Our method is evaluated on the EBHI-Seg dataset (Shi et al. (2023)) which comprises 2298 histopathology section images representing diverse tumor differentiation stages, each accompanied by a binary segmentation mask and an associated stage label. Reported metrics include accuracy for the classification task and the Jaccard Index for the segmentation task. Since this is a small-sized dataset, 5-fold cross-validation is conducted for reliability. For a fair comparison, similar architectures have been used in all experiments. Table 1 shows the performance of our approach in the classification task. Results for the segmentation task are provided in the Appendix.

3 Results, Discussion, and Conclusion

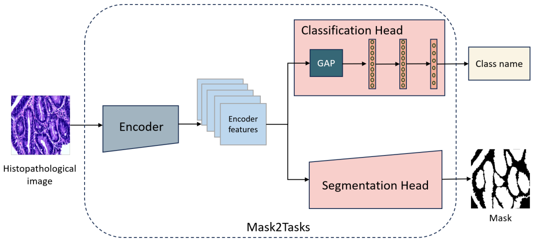

Figure 1: Mask2Tasks architecture.

Model	Accuracy ↑
SwinTransformer	0.874 ±0.015
Resnet50	0.895 ±0.013
EfficientNet	0.856 ±0.010
MTL (uniform weight)	0.916 ±0.012
MTL + Linear Scalarization	0.920 ±0.008
MTL + PCGrad	0.926 ±0.011
MTL + FAMO	0.140 ±0.072
Our Approach	**0.936** ±0.004

Table 1: Performance comparison in the classification task.

First, in the colorectal classification task (Table 1), it is observed that multi-task methods consistently outperform single-task ones. To understand the underlying difference in decision-making, gradient-based saliency maps (Simonyan et al. (2014)) (Figure 2 in the Appendix) have been used. As evident in the saliency maps (Figure 2), compared with single-task training approaches, classifiers trained by multi-task learning tend to rely on more tissue regions. We argue that these spatially diverse discriminative features are attained from image segmentation. This remark aligns with the common sense that to perform well in image segmentation, detailed visual traits of the whole image must be effectively encoded. In contrast, overfitted classifiers tends to focus only on restricted or irrelevant regions of the image. For instance, in Figure 2, the saliency maps sometimes largely emphasize non-tissue, uninformative areas. More evidence can be found in Figure 4 in the Appendix.

Second, our proposed method outperforms other multi-task approaches in the classification task (Table 1), as well as being slightly better in average gain per task (Table 5). A possible explanation is that our proposed method handles the problem of conflicting gradients better. To understand how conflicting gradients happen as well as their impacts, Figure 3 in the Appendix demonstrates the learning curve of the uniformly weighted objective function where each task loss curve is presented in its original scale. It can be seen that the joint loss tends to be dominated by the segmentation component throughout the entire training process. Overall this matches the observation by several other works (Lee et al. (2021), Kendall et al. (2017)) that uniformly weighted multi-task training can lead to sub-optimal performance. Furthermore, in our case, this imbalance is hard to correct by fixed scaling optimization methods such as Linear Scalarization, since each task has different training dynamics. Specifically, while the classification term converges fast, the segmentation component takes significantly longer to settle. Hence, it is likely that when being jointly optimized, the classification term is prone to rely on premature features in the early training stage, and barely improve after reaching convergence. Once overfitted, this component might not be able to fully exploit more robust segmentation-guided features that are acquired later as originally intended. On the other hand, sophisticated gradient manipulation methods such as PCGrad (Yu et al. (2020)) and FAMO (Liu et al. (2023)) can dynamically balance the task losses; however, this might not always be guaranteed in practical settings (Liu et al. (2023)). Thus they can only attain certain levels of mitigation rather than a complete riddance of conflicting gradients. On the contrary, our two-stage training strategy optimizes only one task at a time, hence cleverly circumventing the loss scale imbalance problem.

In conclusion, one would argue that the two-stage training is prone to training suboptimality. While this might happen in certain cases, we have proven that the proposed training regime is useful for this particular problem of colorectal histopathological imaging, and recommend it as an efficient and simple multi-task learning option for the joint colorectal classification-segmentation tasks.

URM Statement

The authors acknowledge that at least one key author of this work meets the URM criteria of ICLR 2024 Tiny Papers Track.

References

URL https://huggingface.co/.

Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. 2022.

Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7482–7491, 2017. URL https://api.semanticscholar.org/CorpusID:4800342.

Jae-Han Lee, Chul Lee, and Chang-Su Kim. Learning multiple pixelwise tasks based on loss scale balancing. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5087–5096, 2021. doi: 10.1109/ICCV48922.2021.00506.

Sungjae Lee and Youngdoo Son. Multitask learning with single gradient step update for task balancing. *Neurocomputing*, 467:442–453, 2022. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2021.10.025. URL https://www.sciencedirect.com/science/article/pii/S0925231221015034.

Bo Liu, Yihao Feng, Peter Stone, and qiang liu. FAMO: Fast adaptive multitask optimization. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=zMeemcUeXL.

Liyu Shi, Xiaoyan Li, Weiming Hu, Haoyuan Chen, Jing Chen, Zizhen Fan, Minghe Gao, Yujie Jing, Guotao Lu, Deguo Ma, Zhiyu Ma, Qingtao Meng, Dechao Tang, Hongzan Sun, Marcin Grzegorzek, Shouliang Qi, Yueyang Teng, and Chen Li. Ebhi-seg: A novel enteroscope biopsy histopathological hematoxylin and eosin image dataset for image segmentation tasks. *Frontiers in Medicine*, 10, 2023. ISSN 2296-858X. doi: 10.3389/fmed.2023.1114673. URL https://www.frontiersin.org/articles/10.3389/fmed.2023.1114673.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Workshop at International Conference on Learning Representations*, 2014.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 5824–5836. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/3fe78a8acf5fda99de95303940a2420c-Paper.pdf.

A Appendix

A.1 Experiment Setup

All experiments in this study are conducted using the EBHI-Seg dataset Shi et al. (2023). Specifically, we excluded 2 images from the dataset with missing masks and partitioned the entire dataset into 5 mutually exclusive folds. For all experiments, we assessed them utilizing 5-fold cross-validation, with reported metrics representing mean performance across 5 folds.

The Mask2Former architecture is adopted from HuggingFace (Hug). Based on this architecture, we developed our multi-task architecture named Mask2Tasks. This architecture is utilized in both the uniformly weighted multi-task learning approach and our approach. We trained all models using the same set of hyperparameters, with each model undergoing 100 epochs. With our approach, the

Table 4: Performance of our approach compared with MTL on segmentation task
(Jaccard's Index - higher is better).

	Image Classes					
	Polyp	Normal	Adeno	Serrated	Low grade	High grade
Mask2Tasks	**0.933 ±0.006**	**0.926 ±0.007**	**0.830 ±0.051**	**0.931 ±0.028**	**0.912 ±0.006**	**0.846 ±0.047**
Unet	0.308	0.263	0.808	0.886	0.849	0.816
SegNet	0.886	0.684	0.646	0.832	0.864	0.812
MedT	0.643	0.562	0.595	0.509	0.808	0.707

training process includes 2 stages: The first stage involves 40 epochs, followed by the second stage lasting 60 epochs. Specific hyperparameter details are provided in Table 2. For reproducibility, the download link to the training code is publicly available in the supplementary material file.

A.2 Additional Results

Table 3 details the performance comparison between our proposed approach and the single-stage multi-task learning with uniform weighting. A notable observation is that our approach demonstrated identical performance to the MTL method on the segmentation task. This observation hinted that, given our approach that involves segmentation training in only 40 epochs, it exhibits a faster convergence rate compared to the vanilla MTL method.

Additionally, we presented in Table 4 a performance evaluation of Mask2Tasks against other segmentation networks on six classes of images from the original paper Shi et al. (2023). It can be seen that our approach surpasses these baseline approaches by a substantial margin.

Table 2: Hyperparameters utilized for training

Hyperparameter	Value
Total training epochs	100
Learning rate	0.0001
Batch size	8
Optimizer	ADAM

Table 3: Performance of our approach compared with MTL on segmentation task

Method	Jaccard Index ↑
Our Approach	**0.880 ±0.008**
MTL (uniform weight)	0.884 ±0.004
MTL + Linear Scalarization	0.883 ±0.005
MTL + PCGrad	0.887 ±0.006
MTL + FAMO	0.003 ±0.001

Table 5: Comparison of single-stage multi-task learning approaches and our method based on the average per-task performance gain over single-task learning.

Method	△ m% ↑
MTL (uniform weight)	2.630
MTL + Linear Scalarization	2.802
MTL + PCGrad	3.372
MTL + FAMO	-91.820
Our Approach	3.547

Table 6: Comparison of the impact of varying numbers of training epochs in stage 1 of our method on both segmentation and classification performance.

Number of epoch	IoU	Accuracy
40 (Current Method)	0.883	0.936
50	0.889	0.931
60	0.890	0.931

Figure 2 illustrates the qualitative results, specifically the saliency maps, generated by our method and other approaches using a sample from the EBHI-Seg dataset. Additionally, Figure 3 presents the training curves. Table 5 compares our method and other approaches in MTL. Similar to (Liu et al. (2023)), we utilize the average per-task performance gain of the multi-task learning model over the single-task learning model: $\triangle m\% = \frac{1}{K}\sum_{k=1}^{K}\left(M_{m,k} - M_{b,k}\right)/M_{b,k}$, with $M_{b,k}$, $M_{m,k}$, and K respectively being the performance metric of the single task learner b, the multi-task learner m in the k-th task and the total number of task, as the primary metric for MTL performance. Figure 4 displays additional qualitative saliency maps as well as predictions generated by our approach, a conventional classifier, the uniform weight multi-task model, and other MTL approaches on a few sample images.

Table 6 illustrates a comparison of varying numbers of training epochs in the first stage of our method on performance of both tasks. After reviewing the table, we decided to select the number of epoch to be 40, considering that increasing this number did not significantly improve accuracy or segmentation performance. In addition, choosing a lower number also saves computational resources.

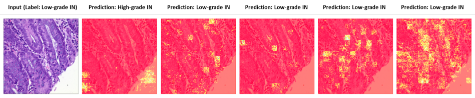

Figure 2: Sample input (left), saliency map produced by a SwinTransformer classifier with wrong prediction, single-stage uniformly-weighted multi-task learning, the single-stage MTL with Linear Scalarization method, the single-stage MTL with PCGrad method, and our approach with correct predictions. Focused regions are highlighted in yellow.

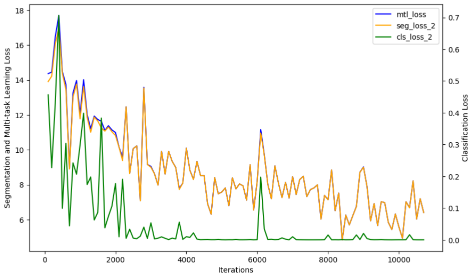

Figure 3: Learning curves of the single-stage uniformly-weighted objective and its components.

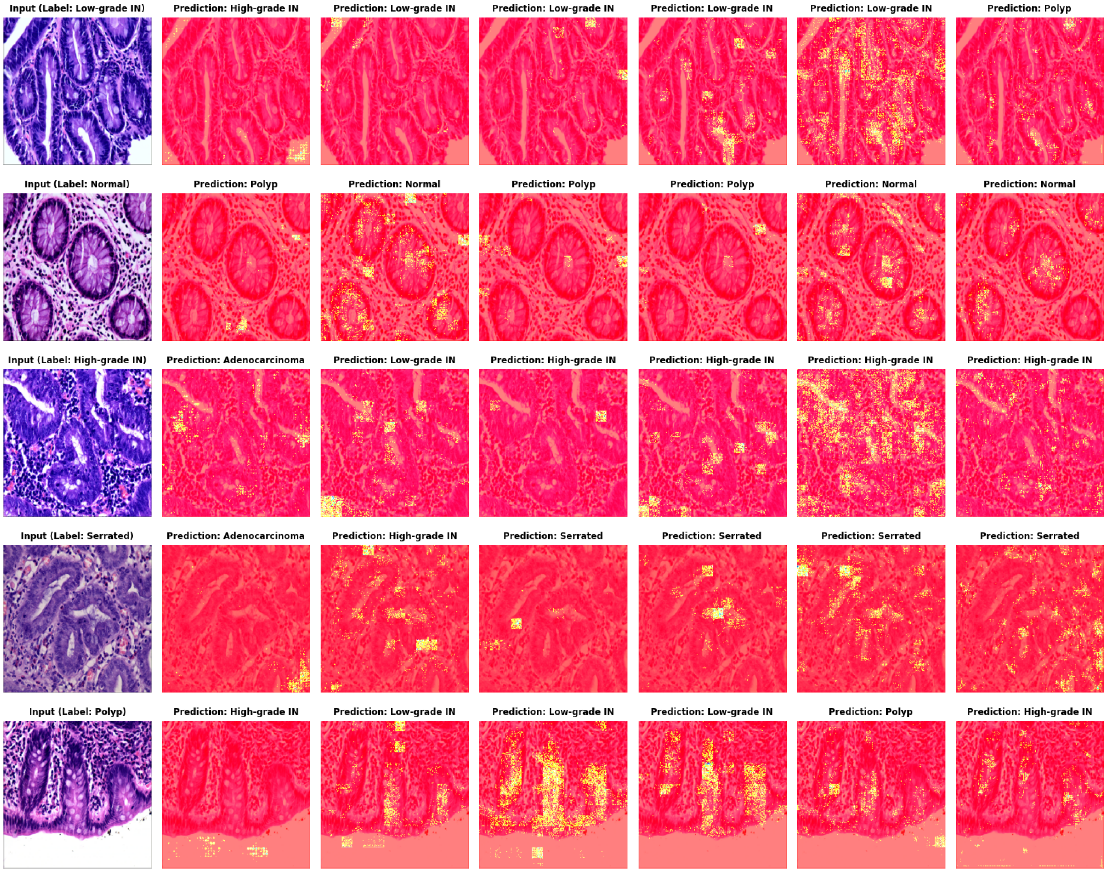

Figure 4: Qualitative results on several samples from the EBHI-Seg dataset. From left to right: Input images (with labels), saliency maps (and predictions) produced by a SwinTransformer classifier, the single-stage uniform weight MTL method, the single-stage MTL with Linear Scalarization method, the single-stage MTL with PCGrad method, our approach and a single task CNN. Focused regions are highlighted in yellow.