Pathology-tailored Data Augmentation for Colorectal Cancer Grading

João D. Nunes [1, 2]
joao.d.fernandes@inesctec.pt
Tânia Pereira[1]
tania.pereira@inesctec.pt
Jaime S. Cardoso [1, 2]
jaime.cardoso@inesctec.pt

[1] INESC TEC - Institute for Systems and Computer Engineering, Technology and Science
Porto, Portugal

[2] Faculty of Engineering - University of Porto
Porto, Portugal

Abstract

Deep Learning (DL) is the go-to solution in many Computer Vision (CV) tasks, including Haematoxylin and Eosin (H&E)-stained Whole Slide Image (WSI) analysis. However, DL models usually generalize poorly to out-of-distribution (o.o.d.) samples. In computational pathology (CPath), this is a known limitation, in spite of the several strategies that have been proposed to attenuate this challenge and improve generalization. In this work, we propose data augmentation to learn more robust representations of H&E-stained WSIs for colorectal dysplasia grading and demonstrate that data augmentations tailored to CPath could be useful to generalize DL models to novel environments.

1 Introduction

Deep Learning (DL) algorithms are the dominant paradigm in computational Pathology (CPath), however, a known limitation of DL methods is that they are prone to learn spurious correlations, which translates to DL models with high performance on independent and identically distributed (i.i.d.) data, but that often fail to generalize to out-of-distribution (o.o.d.) samples [5]. Learning stable representations is thus a central objective in DL. In this work, we consider pathology-tailored data augmentations, primarily techniques focused on increasing robustness to known factors of variation in Whole Slide Images (WSIs) and evaluate if augmenting the data helps DL models generalize.

2 Related Work

Due to the gigapixel size of WSIS, these are usually split into smaller patches for processing. But as the target label of the WSI is determined from specific tissue Regions of Interest (ROIs), this means not all tiles share the same label as the WSI, ultimately making WSI analysis a Multiple Instance Learning (MIL) paradigm. The most common approaches for learning instance representations in CPath comprise: 1) pretraining on the ImageNet dataset; 2) Self-supervised representation learning; and 3) Fine-tuning pretrained models on a small amount of annotated tiles. For instance, Ciga et al. [1] resort to SimCLR and demonstrate that self-supervised domain-specific pretraining on large-scale datasets allows learning effective representations of WSI data. Other methods consider attention mechanisms and demonstrate models pretrained on ImageNet are good feature extractors, meaning only the bag feature aggregation would need to be optimized [4, 8]. Neto et al. [5], on the contrary, demonstrate supervised fine-tuning on a small set of annotated tiles benefits downstream MIL. However, despite the outstanding success in i.i.d. data, the proposed methods were shown to not generalize to o.o.d. samples. In this work, we consider the supervised pretraining strategy of [5] but adopt a data augmentation policy tailored to pathology to increase robustness.

3 Methods

We consider a ResNet-34, and Quadratic Weighted Kappa (QWK) Loss [5] for grading Haematoxylin and Eosin (H&E)-stained WSIs of Colorectal Cancer (CRC) into Non-Neoplastic, Low-grade Dysplasia (LG), and High-grade Dysplasia (HG). We train our models on samples from the *CRC* dataset [5]. More specifically, we resort to the set of 967 WSIs annotated at the tile (instance) level. Previous work has adopted a successful weakly supervised approach [5], but they have also demonstrated the benefits of supervised pretraining of the MIL paradigm. We thus focus on the pretraining stage and analyse if we can increase generalization to novel environments. We restrict the experiments in this paper to the annotated data. We split the 967 annotated WSIs into training (567), validation (200), and testing (200) splits. We further extract a training set with 100 WSIs and another with 250 from the 567 training split. Besides, we evaluate the models in slide-level classification on the public test set of the *CRC* dataset [5]. To evaluate model performance in o.o.d. data, we also included the PAIP [6] and TCGA [2, 3] CRC datasets.

Data augmentation is a central part of this work. To prevent corrupting the semantics of the data, while simulating distribution shifts, we adopt a data augmentation policy previously validated in CPath [10], as described in Table 1. Aside from geometric transformations, we also consider colour augmentations to account for H&E-stain variability. For instance, we consider Haematoxylin-Eosin-DAB (HED) colour augmentation, which is based on a colour deconvolution technique, using a predetermined stain matrix [7]. As the staining protocol demands the application of each stain independently, randomly and independently perturbing the individual stain components increases model robustness to variability in stain concentration.

Transform	Hyperparameters
Hue Shift	shift limits = $[-0.125, 0.125]$
Contrast Shift	shift limits = $[-0.2, 0.2]$
Saturation Shift	shift limits = $[-0.125, 0.125]$
HED Jitter	$\alpha = 0.1 \beta = 0.0075$
Flips (Horizontal/Vertical)	-
Rotation	limits = $[-90°, 90°]$
Gaussian Blur	max kernel size = 15, σ limits = $[0.1, 2]$
Gaussian Noise	$\mu = 0, \sigma = [0, 0.1]$

Table 1: Data Augmentation tailored to CPath, as proposed in [10].

3.1 Implementation Details

The experiments are implemented in *python* 3.10.8 with *PyTorch* 1.13.0 on a Tesla V100 32GB GPU. For data augmentation, we use *Albumentations* 1.3.0, and *scikit-image* 0.19.3. We consider Stochastic Gradient Descent (SGD) with momentum($\mu = 0.9$), weight decay $\lambda = 3e-4$, and initial learning rate $\gamma = 1e-4$. We run each training loop for 30 epochs with a batch size of 32. Besides, we reduce γ by a factor of 10^{-1} when $\mathcal{L}_{QWK}$ did not decrease at least by 10^{-4} in 10 steps of 500 iterations. We select for testing the model with the best QWK in the validation set.

4 Results and Discussion

4.1 Instance-level Classification

Considering training and validation results (Table 2), i.e., tile-level supervision, we observe that the models with data augmentation achieve, in general, similar QWK (validation set) when compared with ResNet-34 without data augmentation. When evaluated on the held-out test split (Table 2), the baseline ResNet-34 without data augmentation outperforms the ResNet-34 with data augmentation for larger training set sizes.

4.2 Whole Slide Image Classification

In these experiments, we consider evaluating results in i.i.d. data and assess performance on H&E-stained WSI colorectal dysplasia grading. We resort to a heuristic to predict the slide target, where we define the bag label as the most frequent label of the top 7 instances with the highest expected severity, $\mathbb{E}(\hat{C}_{s,n})$ [5]:

# WSIs (train)	data aug	qwk loss (train)	qwk (val)	qwk (test)	acc (test)
100	✓	0.1493	0.8235	0.8430	0.8400
	✗	0.1017	0.8077	0.8279	0.8070
250	✓	0.1433	0.8431	0.8284	0.8243
	✗	0.1478	0.8422	0.8431	0.8319
567	✓	0.1516	0.8385	0.8337	0.8208
	✗	0.1314	0.8431	0.8470	0.8422

Table 2: Results of the ResNet-34 (pre-trained on ImageNet) with and without data augmentation.

# WSIs	data aug	acc	qwk	acc (bin)
100	✓	0.5485	0.462	0.8484
	✗	0.5563	0.4221	0.8462
250	✓	0.6187	0.5776	0.8618
	✗	0.6210	0.5711	0.8640
567	✓	0.6042	0.5404	0.8584
	✗	0.6198	0.5912	0.8729

Table 3: Effects of dataset size. The WSI dysplasia grade is given by majority voting the predictions from the top seven tiles with the highest expected severity, $\mathbb{E}(\hat{C}_{s,n})$.

$$\mathbb{E}(\hat{C}_{s,n}) = \sum_{i=1}^{K} i \times p\left(\hat{C}_{s,n} = C^{(i)}\right) \tag{1}$$

where $\hat{C}_{s,n}$ is a random variable on the set of possible class labels $\{C^{(1)}, ..., C^{(K)}\}$, and $p\left(\hat{C}_{s,n} = C^{(i)}\right)$ are the K model predictions.

In Table 3 we assess the effects of dataset size on model performance when evaluating the WSI classification performance. To compute the binary accuracy we aggregate the LG and HG classes. In most experiments, the method with data augmentation achieves the worst performance.

4.3 Out of Distribution Evaluation

As a final study of robustness, we evaluate the whole-slide classification performance on the PAIP CRC [6] (Table 4) and TCGA CRC [2, 3] (Table 5) test sets. We observe that in PAIP CRC [6] all methods perform particularly well, especially those trained on the 250 and 567 H&E-stained WSIs sets. However, in TCGA CRC [2, 3] the performance of the ResNet-34 method with no data augmentation is unstable with increasing training set size. While it achieves an accuracy of only $40,09\%$ with the 100 WSIs split, the accuracy increases to $93,10\%$ in the model trained on the 250 whole-slides and drops to $71,12\%$ in the other training set. As expected, our findings support other empirical evidence [9, 10] that pathology-tailored data augmentations increase model robustness.

#WSIs (train)	data aug	acc	acc (bin)	sens
100	✗	0.8700	0.8700	0.8700
	✓	1.0000	1.0000	1.0000
250	✗	0.9600	1.0000	1.0000
	✓	1.0000	1.0000	1.0000
567	✗	0.9600	1.0000	1.0000
	✓	0.9900	1.0000	1.0000

Table 4: Out of distribution (o.o.d) evaluation on the PAIP CRC [6] dataset.

5 Conclusion

In this document, we focus on increasing the robustness of DL models for WSI analysis. We assess the effects of pathology-tailored data augmentation in representation learning and compare model robustness in o.o.d. samples. We observe that while the effects of data augmentation are not evident when generalizing in the i.i.d. setting, the technique leads to more robust representations as models generalize to o.o.d. data.

6 Aknowledgments

This work is co-financed by Component 5 - Capitalization and Business Innovation, integrated in the Resilience Dimension of the Recovery and Resilience Plan within the scope of the Recovery and Resilience Mechanism (MRR) of the European Union (EU), framed in the Next Generation EU, for the period 2021 - 2026, within project HfPT, with reference 41 and by National Funds through the Portuguese funding agency, FCT-Foundation for Science and Technology Portugal, a PhD Grant Number 2022.12385.BD

#WSIs (train)	data aug	acc	acc (bin)	sens
100	✗	0.4009	0.6379	0.6379
	✓	0.9655	1.0000	1.0000
250	✗	0.9310	0.9397	0.9397
	✓	1.0000	1.0000	1.0000
567	✗	0.7112	0.8276	0.8276
	✓	1.0000	1.0000	1.0000

Table 5: Out of distribution (o.o.d) evaluation on the TCGA CRC [2, 3] dataset.

References

[1] Ozan Ciga, Tony Xu, and Anne Louise Martel. Self supervised contrastive learning for digital histopathology. *Machine Learning with Applications*, 7:100198, 2022. ISSN 2666-8270. doi: https://doi.org/10.1016/j.mlwa.2021.100198. URL https://www.sciencedirect.com/science/article/pii/S2666827021000992.

[2] Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, et al. The cancer imaging archive (tcia): maintaining and operating a public information repository. *Journal of digital imaging*, 26:1045–1057, 2013.

[3] S Kirk, Y Lee, CA Sadow, S Levine, C Roche, E Bonaccio, and J Filiippini. Radiology data from the cancer genome atlas colon adenocarcinoma [tcga-coad] collection. the cancer imaging archive, 2016.

[4] Ming Y. Lu, Drew F. K. Williamson, Tiffany Y. Chen, Richard J. Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5(6):555–570, Jun 2021. ISSN 2157-846X. doi: 10.1038/s41551-020-00682-w. URL https://doi.org/10.1038/s41551-020-00682-w.

[5] Pedro C. Neto, Sara P. Oliveira, Diana Montezuma, João Fraga, Ana Monteiro, Liliana Ribeiro, Sofia Gonçalves, Isabel M. Pinto, and Jaime S. Cardoso. imil4path: A semi-supervised interpretable approach for colorectal whole-slide images. *Cancers*, 14(10), 2022. ISSN 2072-6694. doi: 10.3390/cancers14102489. URL https://www.mdpi.com/2072-6694/14/10/2489.

[6] Pathology AI Platform, 2020. URL http://wisepaip.org/paip.

[7] Arnout C Ruifrok, Dennis A Johnston, et al. Quantification of histochemical staining by color deconvolution. *Analytical and quantitative cytology and histology*, 23(4):291–299, 2001.

[8] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in neural information processing systems*, 34:2136–2147, 2021.

[9] Chetan L. Srinidhi, Seung Wook Kim, Fu-Der Chen, and Anne L. Martel. Self-supervised driven consistency training for annotation efficient histopathology image analysis. *Medical Image Analysis*, 75:102256, 2022. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2021.102256. URL https://www.sciencedirect.com/science/article/pii/S1361841521003017.

[10] David Tellez, Geert Litjens, Péter Bándi, Wouter Bulten, John-Melle Bokhorst, Francesco Ciompi, and Jeroen van der Laak. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Medical Image Analysis*, 58:101544, 2019. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2019.101544. URL https://www.sciencedirect.com/science/article/pii/S1361841519300799.