Masked Attention as a Mechanism for Improving Interpretability of Vision Transformers

Clément Grisi CLEMENT.GRISI@RADBOUDUMC.NL
Jeroen van der Laak
Geert Litjens
Computational Pathology Group, Radboudumc, Netherlands

Abstract

Vision Transformers are at the heart of the current surge of interest in foundation models for histopathology. They process images by breaking them into smaller patches following a regular grid, regardless of their content. Yet, not all parts of an image are equally relevant for its understanding. This is particularly true in computational pathology where background is completely non-informative and may introduce artefacts that could mislead predictions. To address this issue, we propose a novel method that explicitly masks background in Vision Transformers' attention mechanism. This ensures tokens corresponding to background patches do not contribute to the final image representation, thereby improving model robustness and interpretability. We validate our approach using prostate cancer grading from whole-slide images as a case study. Our results demonstrate that it achieves comparable performance with plain self-attention while providing more accurate and clinically meaningful attention heatmaps.

Keywords: Vision Transformers, attention, computational pathology, prostate cancer

1. Introduction

Adoption of whole-slide imaging technology in pathology has contributed to the growing availability of large digitized datasets. This shift towards digital pathology has fostered computer vision research to support and augment pathologists with deep learning algorithms. However, conventional deep learning methods are ill-equipped to handle the enormous sizes of whole-slide images (WSIs), which usually exceed the memory capacity of graphics processing units. A popular approach to overcome this memory bottleneck involves partitioning WSIs into smaller, more manageable patches (Campanella et al., 2019). This technique aligns well with the operational mechanics of Vision Transformers (ViTs) (Dosovitskiy et al., 2021), sparking increased interest in their application within computational pathology (Shao et al., 2021; Chen et al., 2022, 2024).

While ViTs process images by breaking them into smaller patches following a regular grid, they do not account for the fact that all parts of an image are not equally relevant or informative. Uniform background areas are often less informative than more cluttered, dense areas. In computational pathology, background is not only low-informative but fundamentally devoid of any diagnostic value. Including tokens that correspond to background areas may introduce artefacts that could mislead predictions and compromise model interpretability, possibly resulting in clinically irrelevant hotspots in attention maps. To address this issue, we propose a simple method that explicitly masks background in the attention

© 2024 CC-BY 4.0, C. Grisi, J.v.d. Laak & G. Litjens.

mechanism of ViTs. By doing so, we ensure tokens corresponding to background do not contribute to the final image representation. Omitting visually present but diagnostically irrelevant information should not only sharpen the signal-to-noise ratio, but also result in attention heatmaps that are both more visually coherent and easier to interpret.

2. Proposed Method

Hierarchical Vision Transformer. The inherent hierarchical structure within whole-slide images spans across various scales, from tiny cell-centric regions containing fine-grained information, up to the entire slide which exhibits the overall intra-tumoral heterogeneity of the tissue microenvironment. Drawing inspiration from this layered structure, our model consists of a Hierarchical Vision Transformer that processes whole-slide images at three nested scales (Grisi et al., 2023). Slides are unrolled into non-overlapping 2048×2048 regions, capturing macro-scale interactions between clusters of cells. These are further unrolled into non-overlapping 256×256 patches, depicting cell-to-cell interactions. A pre-trained ViT-S/16 is used to embed these patches into feature vectors. Then, a second Transformer aggregates the representations of 256×256 patches within larger 2048×2048 regions. Finally, a third Transformer pools region-level tokens into a slide-level representation that is projected to class logits for loss computation (Appendix A, Figure 2).

Masked Attention. When extracting regions from whole-slide images, only those containing tissue are retained as fully background regions contain no informative content. However, when these regions are further unrolled into non-overlapping 256×256 patches, some patches may still contain no tissue (Appendix C, Figure 4). To ensure that region-level representations are exclusively derived from patches containing tissue, we propose a novel *masked attention* method. By leveraging fine-grained tissue segmentation masks, our approach explicitly nullifies the contribution of entirely background patches during self-attention, thereby enhancing the quality of extracted features. We provide a pseudo code implementation in Appendix B.

3. Experimental Results

Dataset. To assess the robustness of the proposed method, we use the PANDA dataset (Bulten et al., 2022). It is the largest publicly available dataset of H&E stained prostate WSIs to date, with $11,554$ prostate biopsies curated from two different sites. All slides are provided at a pixel spacing close to 0.50 μm, together with their ISUP score (Appendix D).

Data Preprocessing & Evaluation Metric. We used an internally developed model to automatically segment tissue in each slide, from which we extract non-overlapping 2048×2048 regions at the resolution closest to 0.50 μm (Appendix C). We split PANDA development set into 5 cross-validation folds, stratifying on the ISUP score. To evaluate the model's classification performance, we report averaged quadratic weighted kappa scores on the tuning set, as well as on the combined public and private test sets.

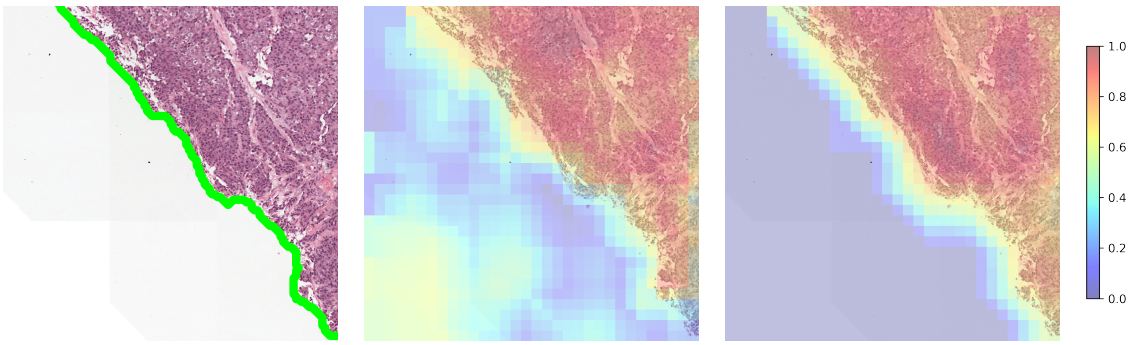
Prostate Cancer Grading. For each fold, we pretrain the first Transformer on the training set via the student-teacher knowledge distillation framework DINO (Caron et al., 2021). This Transformer is used as a feature extractor to embed each slide into a $(M_i^{2048}, 64, 384)$

feature vector, where M_i stands for the number of 2048×2048 regions extracted in the i-th slide. The last two Transformers are then jointly trained to map these sequences to ISUP scores. We formulate the classification problem as a regression task and use the Mean Squared Error loss to leverage the ordinal nature of the ISUP scores. Classification results are summarized in Table 1. Masked self-attention achieves comparable performance with plain self-attention.

Attention Mechanism	Tune Score	Combined Test Score
Plain self-attention	0.945 ± 0.003	0.899 ± 0.008
Masked self-attention	0.946 ± 0.003	0.899 ± 0.009

Table 1: ISUP score classification results. We report quadratic weighted kappa, averaged over the 5 cross-validation folds.

Model Interpretability. Attention heatmaps offer a streamlined form of model interpretability by revealing the specific image features that the model has learned to associate with particular classes. Figure 1 shows attention heatmaps for the region-level Transformer. While some background patches display high attention values in plain self-attention heatmaps (Figure 1(b)), all background patches are given no attention in masked self-attention heatmaps (Figure 1(c)). Additional visualizations at the slide level are provided in Appendix E.



(a) tissue segmentation (b) plain self-attention (c) masked self-attention

Figure 1: Region-level attention maps.

4. Conclusion

In conclusion, our proposed masked attention strategy improves model interpretability by explicitly excluding irrelevant patches from contributing to self-attention in Vision Transformers. This approach is particularly beneficial in computational pathology where the inclusion of non-informative background content can introduce artefacts that can compromise model reliability. Our results demonstrate that masked attention achieves comparable performance with plain self-attention while providing more accurate and clinically meaningful heatmaps. This method has the potential to enhance the accuracy, robustness, and

interpretability of ViT-based models in digital pathology, ultimately contributing to improved diagnostic accuracy.

References

Wouter Bulten, Kimmo Kartasalo, Po-Hsuan Cameron Chen, Peter Ström, Hans Pinckaers, Kunal Nagpal, Yuannan Cai, David F. Steiner, Hester van Boven, Robert Vink, Christina Hulsbergen-van de Kaa, Jeroen van der Laak, Mahul B. Amin, Andrew J. Evans, Theodorus van der Kwast, Robert Allan, Peter A. Humphrey, Henrik Grönberg, Hemamali Samaratunga, Brett Delahunt, Toyonori Tsuzuki, Tomi Häkkinen, Lars Egevad, Maggie Demkin, Sohier Dane, Fraser Tan, Masi Valkonen, Greg S. Corrado, Lily Peng, Craig H. Mermel, Pekka Ruusuvuori, Geert Litjens, Martin Eklund, Américo Brilhante, Aslı Çakır, Xavier Farré, Katerina Geronatsiou, Vincent Molinié, Guilherme Pereira, Paromita Roy, Günter Saile, Paulo G. O. Salles, Ewout Schaafsma, Joëlle Tschui, Jorge Billoch-Lima, Emíio M. Pereira, Ming Zhou, Shujun He, Sejun Song, Qing Sun, Hiroshi Yoshihara, Taiki Yamaguchi, Kosaku Ono, Tao Shen, Jianyi Ji, Arnaud Roussel, Kairong Zhou, Tianrui Chai, Nina Weng, Dmitry Grechka, Maxim V. Shugaev, Raphael Kiminya, Vassili Kovalev, Dmitry Voynov, Valery Malyshev, Elizabeth Lapo, Manuel Campos, Noriaki Ota, Shinsuke Yamaoka, Yusuke Fujimoto, Kentaro Yoshioka, Joni Juvonen, Mikko Tukiainen, Antti Karlsson, Rui Guo, Chia-Lun Hsieh, Igor Zubarev, Habib S. T. Bukhar, Wenyuan Li, Jiayun Li, William Speier, Corey Arnold, Kyungdoc Kim, Byeonguk Bae, Yeong Won Kim, Hong-Seok Lee, Jeonghyuk Park, and the PANDA challenge consortium. Artificial intelligence for diagnosis and gleason grading of prostate cancer: the panda challenge. *Nature Medicine*, 28(1):154–163, Jan 2022. ISSN 1546-170X. doi: 10.1038/s41591-021-01620-2. URL https://doi.org/10.1038/s41591-021-01620-2.

Gabriele Campanella, Matthew G. Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J. Busam, Edi Brogi, Victor E. Reuter, David S. Klimstra, and Thomas J. Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine*, 25(8):1301–1309, Aug 2019. ISSN 1546-170X. doi: 10.1038/s41591-019-0508-1. URL https://doi.org/10.1038/s41591-019-0508-1.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.

Richard J. Chen, Chengkuan Chen, Yicong Li, Tiffany Y. Chen, Andrew D. Trister, Rahul G. Krishnan, and Faisal Mahmood. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16144–16155, June 2022.

Richard J. Chen, Tong Ding, Ming Y. Lu, Drew F. K. Williamson, Guillaume Jaume, Andrew H. Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, Mane Williams, Lukas Oldenburg, Luca L. Weishaupt, Judy J. Wang, Anurag Vaidya, Long Phi Le, Georg Gerber, Sharifa Sahai, Walt Williams, and Faisal Mahmood. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862,

Mar 2024. ISSN 1546-170X. doi: 10.1038/s41591-024-02857-3. URL https://doi.org/10.1038/s41591-024-02857-3.

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, 2021.

Clément Grisi, Geert Litjens, and Jeroen van der Laak. Hierarchical vision transformers for context-aware prostate cancer grading in whole slide images. *arXiv preprint arXiv:2312.12619*, 2023.

Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, and Yongbing Zhang. Transmil: Transformer based correlated multiple instance learning for whole slide image classification, 2021.

Appendix A. Architecture Overview

Figure 2 shows the multi-stage Hierarchical Vision Transformer architecture we use in this work. It features three Vision Transformers, followed by a simple linear classifier that projects the slide-level embedding onto the desired number of classes.

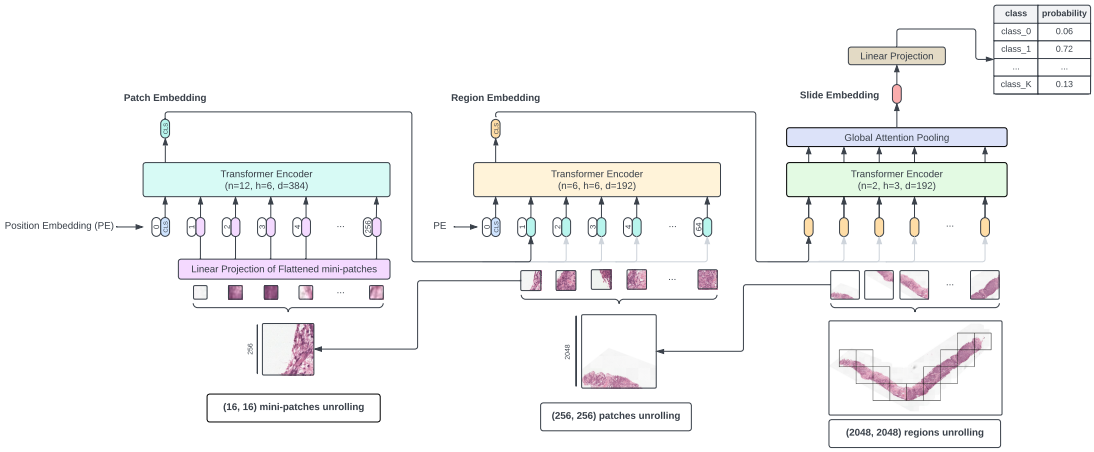


Figure 2: Overview of our Hierarchical Vision Transformer for whole-slide image analysis. This figure illustrates the multi-scale processing of whole-slide images.

Appendix B. Masked Attention Pseudo Code

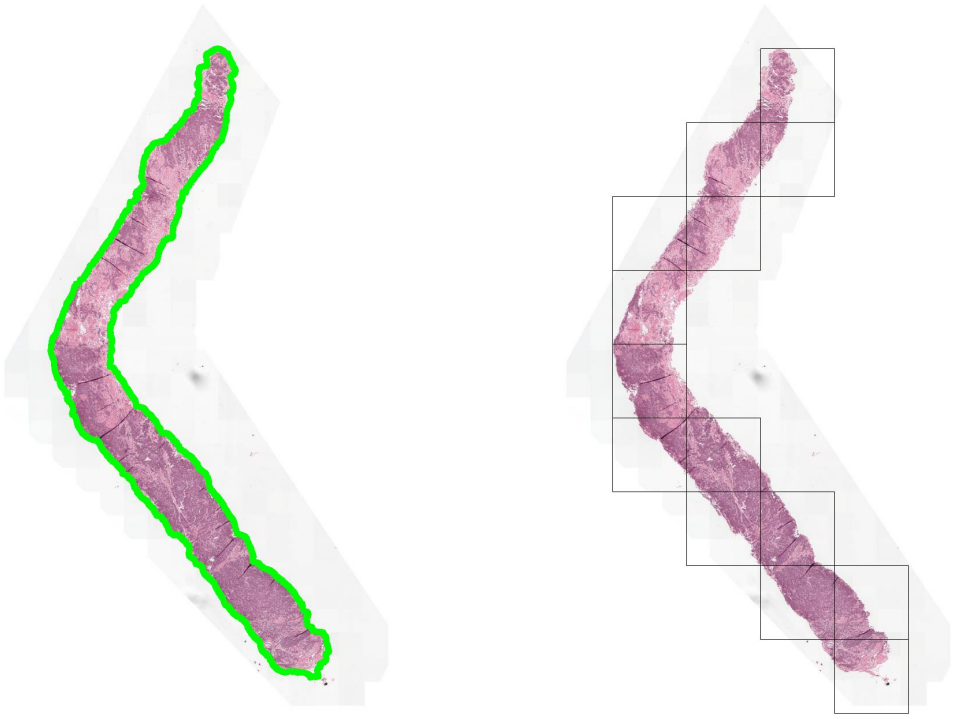
The *masked attention* module expects the input sequence x – of shape $(M_i^{2048}, 64, 384)$ – as well as a *pct* tensor of shape $(M_i^{2048}, 1, 64)$ containing the tissue percentage for each 256×256 patch within each 2048×2048 regions in a slide.

Algorithm 1: Nullifying the contribution of background patches

Input: x of shape (M, 64, 384), *pct* of shape (M, 1, 64)
Output: x_{attended}, the attended tensor
q, k, v ← self.qkv(x)
raw_attn ← (q @ k.T) * scale
pct ← *pct*.unsqueeze(1).expand(-1, self.num_heads, -1, -1)
masked_attn ← raw_attn.masked_fill(*pct* == 0, float("-inf"))
attn ← masked_attn.softmax(dim=-1)
x_{attended} ← (attn @ v).T

Appendix C. Data Preprocessing

Figure 3 shows an example result of our tissue segmentation and region extraction algorithm. Due to potential tissue segmentation irregularities, regions containing fewer than 10% tissue were discarded.



(*a*) tissue segmentation (*b*) 2048×2048 regions at 0.50 μm

Figure 3: Example result of data preprocessing pipeline

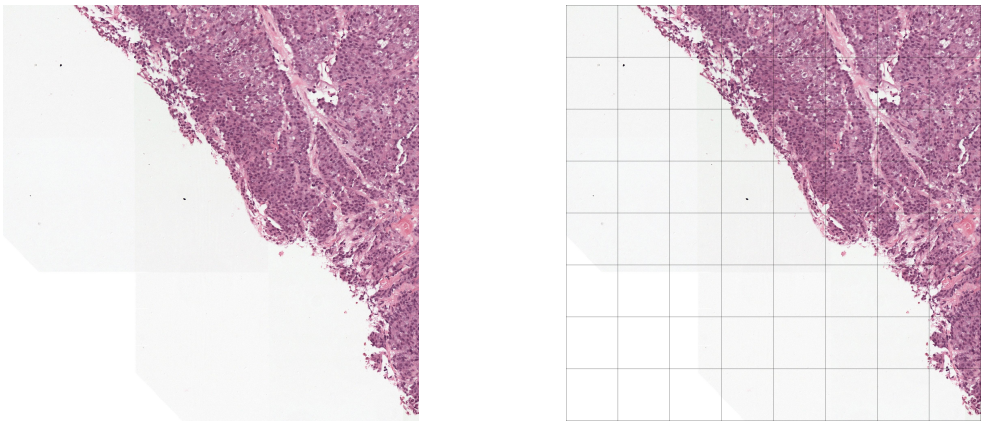


Figure 4: Unrolling a 2048×2048 region into non-overlapping 256×256 patches

Appendix D. PANDA Dataset Details

In Table 2, we provide a summary of the main characteristics of the PANDA dataset.

Table 2: PANDA dataset summary

Center	Scanner	Spacing (μm)	# dev	# public test	# private test
Radboud	3DHistech	0.48	5160	195	333
Karolinska	Leica	0.50	2193	97	150
Karolinska	Hamamatsu	0.45	3263	101	62

Pathologists classify tumors into different growth patterns by analyzing the histological architecture of the tumor tissue. Tissue specimens are then categorized into one of five groups based on the distribution of these patterns in the tumor. Figure 5 shows the grade group distribution for the development set, the public test set and the private test set.

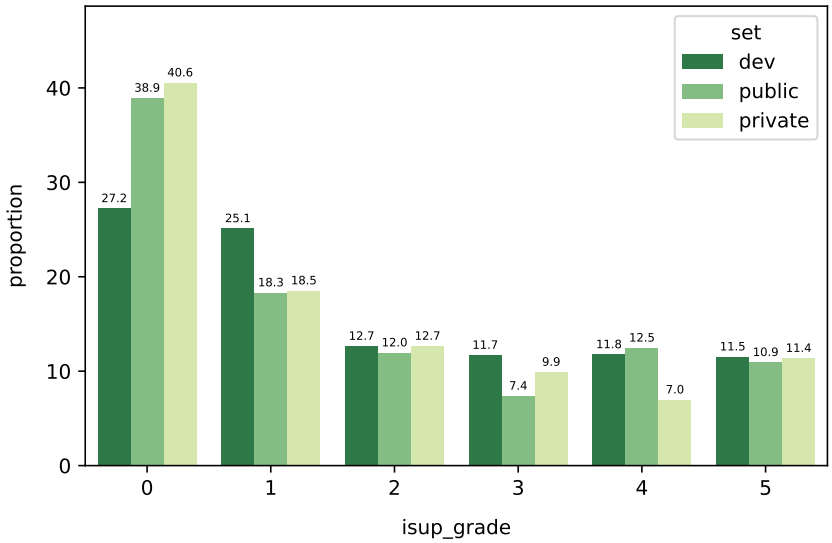
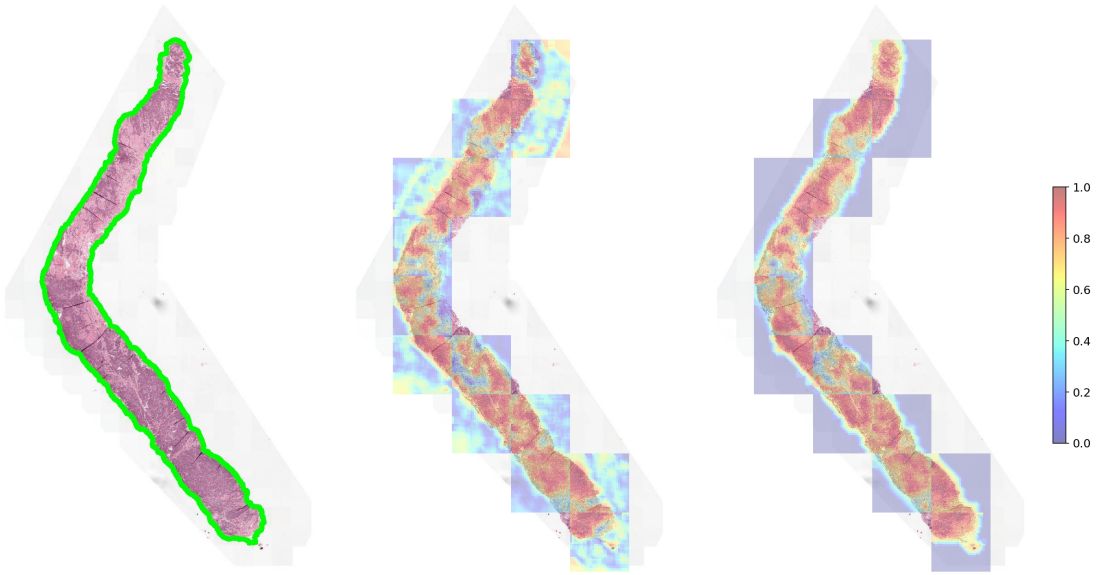


Figure 5: PANDA label distribution

Appendix E. Stitched Attention Heatmaps

Stitched attention heatmaps provide a comprehensive visualisation of the model attention, offering a more intuitive understanding of which parts of the slide contribute most significantly to the model's decision-making process.



(a) tissue segmentation (b) plain self-attention (c) masked self-attention

Figure 6: Stitched region-level attention heatmaps