Caption generation from histopathology whole-slide images using pre-trained transformers

Author name(s) withheld EMAIL(S) WITHHELD
Address withheld

Editors: Under Review for MIDL 2023

Abstract

The recent advent of *foundation models* and *large language models* has enabled scientists to leverage large-scale knowledge of pretrained (vision) transformers and efficiently tailor it to downstream tasks. This technology can potentially automate multiple aspects of cancer diagnosis in digital pathology, from whole-slide image classification to generating pathology reports while training with pairs of images and text from the diagnostic conclusion. In this work, we orchestrate a set of weakly-supervised transformer-based models with a first aim to address both whole-slide image classification and captioning, addressing the automatic generation of the conclusion of pathology reports in the form of image *captions*. We report our first results on a multicentric multilingual dataset of colon polyps and biopsies. We achieve high diagnostic accuracy with no supervision and cheap computational adaptation.

Keywords: Whole slide images, histopathology, multi-modal training, caption generation

1. Introduction

Recent advances in the field of deep learning are showing increasing capability of bridging the gap between *language* understanding and *vision*. Such technology is particularly suited for the field of medical imaging, where *multimodal* data with pairs of images and text from electronic health records are clinically available. With the adoption of digital pathology workflows, an increasing amount of gigapixel whole-slide images is produced clinically, containing a wealth of information for deep learning development. However, the promise of computer algorithms as a support for pathology diagnosis and potentially aiding the generation of pathology reports often relies on supervised learning, presenting a challenge due to the substantial amount of labeled data required, together with the time-consuming interpretation of histopathology whole slide images (WSIs). The authors in (Gamper and Rajpoot, 2021) provide evidence that models pre-trained on dense digital pathology images learn highly informative representations for caption generation. Nevertheless, their proposed method involves an exhausting effort to extract and process captions in figures from text books. Similarly, previous studies (Zhang et al., 2020; Tsuneki and Kanavati, 2022) show that caption generation is a viable task in digital pathology but the authors do not apply self-supervised and pre-trained models. Motivated by this, we demonstrate the benefit from fine-tuning pre-trained weakly supervised transformers on the downstream task of pathology caption generation. We orchestrate a two-stage pipeline where we first learn highly informative image and text representations using the CLIP training regime (Radford et al., 2021). In the second stage, we utilize extracted WSI representations from the first stage to condition a pre-trained bio-gpt-2 (Luo et al., 2022) language model to generate

© 2023 CC-BY 4.0, Anonymous.

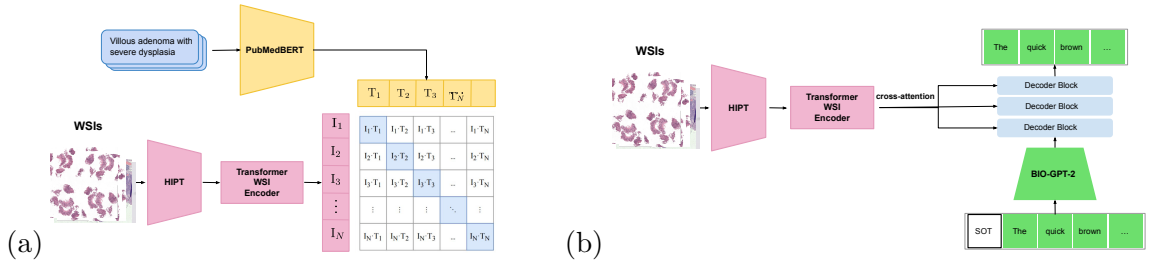

Figure 1: In the first stage (a) we perform Contrastive WSI-caption pre-training, while in (b), the decoder blocks are conditioned on the WSI embeddings trained in (a).

Original Caption	**GPT-3.5 Cleaned**	**Generated Caption**
biopsies distal colon: chronic inflammation, in partially active and slightly histiocytary. no specific characteristics. the microscopic preparations from elsewhere have been requested for revision.	chronic inflammation, no specific characteristics.	no abnormalities, no dysplasia or malignancy. cyclic inflammation.
biopt colon transversum: adenocarcinoma.	adenocarcinoma.	Metastasis of adenocarcinoma best suited to primary process.
1) fragments of tubular adenoma with high degree dysplasia.	tubular adenoma with high degree dysplasia.	adenocarcinoma on villous adenoma. no lymphovascular invasion is identified. enced enced enced ED ED ED ED

Table 1: The original caption, processed caption and the generated caption given the WSI. The last example shows a failed caption generation.

captions. Moreover, pathology captions may include irrelevant or noisy information, such as running text unrelated to any observed lesion in the WSI. To address this, we explore the use of GPT-3.5-turbo (Ouyang et al., 2022) to pre-process the captions and remove extraneous information. As a result, we introduce a novel approach towards generating text from pathology images. Our experiments show promising diagnostic accuracy from the generated captions suggesting the viability of our approach in digital pathology.

2. Method

Data We collected 5729 gigapixel-size whole slide images of colon polyps and biopsies scanned at 0.25 micron per pixel spacing originating from two labs from two countries. Each whole slide image from Lab 1 was associated with one diagnostic report. For Lab 2, a report referred to several slides (from the same paraffin block), because of this the slides were packed together to form one image and linked to a single caption. Each WSI-caption pair was labelled with one of five diagnostic labels: normal, hyperplasia, low-grade dysplasia, high-grade dysplasia, or adenocarcinoma. These labels were not used during training of the pipeline. A subset of 569 patient-split WSI-caption pairs was reserved for testing, which served as the basis for evaluating our results. The captions were machine translated (Tiedemann and Thottingal, 2020) from two languages to English. The captions were subsequently pre-processed using GPT-3.5-turbo, which was prompted with ten examples of how to restructure the captions. These processed captions are then used for training in the two-stage pipeline. Three examples are shown in Table 1.

Unpretrained caption model	Pre-trained caption model	GPT-3.5 cleaned pre-trained caption model	WSI supervised classifier
0.65 (±0.20)	0.70 (±0.21)	**0.73** (±0.15)	0.76 (±0.16)

Table 2: Mean F1-scores over the five diagnostic classes for each model.

Architecture In the first stage of our pipeline, we train a CLIP model, which consists of a HIPT model (Chen et al., 2022) to encode WSIs and PubmedBERT (Gu et al., 2020) to encode medical text. The HIPT model is a hierarchical transformer model trained using DINO (Caron et al., 2021) on TCGA (Liu et al., 2018). HIPT encodes a 4096x4096 WSI region to a vector of size 192. In this manner, we extract a sequence of embeddings that represents one (packed) WSI. Subsequently, we train a transformer encoder on this sequence to pool the features and map them to the same dimensionality as the caption embeddings. We kept both the image and language pre-trained models frozen and only extract the WSI and caption embeddings. In the second stage, we extract the learnt WSI embeddings from the previous stage and condition three transformer decoder layers on top of a pre-trained bio-gpt-2 model. This generates a diagnostic caption that is conditioned on a WSI.

Evaluation We evaluated the generated captions by assessing their diagnostic ability by manually classifying the captions to one of the five diagnostic labels. We could then compare the mean F1-score of four distinct models: (1) A supervised WSI classifier that we treat as a baseline, (2) a caption model without a pre-trained bio-gpt-2 decoder, (3) a pre-trained bio-gpt-2 caption model and (3) a pre-trained bio-gpt-2 caption model trained on the processed gpt-3.5-turbo caption data. We further note that the baseline (1) was a transformer WSI encoder (Figure 1) with a classifier head.

3. Results and Discussion

The findings presented in Table 2 and in the third column of Table 1 suggest that the pre-trained model embeddings contain highly informative features. The application of GPT-3.5-turbo to clean the captions results in a significant improvement over the baseline models in terms of diagnostic accuracy and the quality of the generated captions. Our captioning model has a diagnostic accuracy close to a supervised classifier while having the weakly-supervised advantage. We also find that the model is susceptible to the manner in which the captions are presented. The templates in which the captions are written by the pathologists differ between the two labs and by prompt-style caption pre-processing we are able to normalize them. But a more structured validation strategy is needed to evaluate the fidelity of the generated captions through pathologist review. Furthermore, in Table 2 we see typical language modelling failure mode where some words are repeated or hallucinated, determining the primary cause of this is left to be explored. Another area for improvement is expanding the pre-training of HIPT to encompass a broader range of pathological conditions to enhance its effectiveness in under-represented disease manifestations. Despite the use of large state-of-the-art transformer models, we fine-tuned our pipeline on a single A100 (40GB) GPU in less than 20 minutes. Our work highlights the need for further development in large scale pre-trained models in the field of digital pathology.

Acknowledgments

Acknowledgments withheld.

References

Mathilde Caron, Hugo Touvron, Ishan Misra, Herv'e J'egou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9630–9640, 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, 2022.

Jevgenij Gamper and Nasir Rajpoot. Multiple instance captioning: Learning representations from histopathology textbooks and articles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16549–16559, June 2021.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing, 2020.

Jianfang Liu, Tara M. Lichtenberg, Katherine A. Hoadley, Laila M. Poisson, Alexander J. Lazar, Andrew D. Cherniack, Albert J. Kovatich, Christopher C. Benz, Douglas A. Levine, Adrian V. Lee, Larsson Omberg, Denise M. Wolf, Craig D. Shriver, Vésteinn Thorsson, and Hai Hu. An integrated tcga pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell*, 173:400 – 416.e11, 2018.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. Biogpt: Generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6), November 2022. URL https://www.microsoft.com/en-us/research/publication/biogpt-generative-pre-trained-transformer-for-biomedical-text-generation-and-mining/.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe. Training language models to follow instructions with human feedback. *ArXiv*, abs/2203.02155, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021. URL https://arxiv.org/abs/2103.00020.

Jörg Tiedemann and Santhosh Thottingal. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal, 2020.

Masayuki Tsuneki and Fahdi Kanavati. Inference of captions from histopathological patches. In *International Conference on Medical Imaging with Deep Learning*, 2022.

Renyu Zhang, Christopher Weber, Robert Grossman, and Aly A Khan. Evaluating and interpreting caption prediction for histopathology images. In *Machine Learning for Healthcare Conference*, pages 418–435. PMLR, 2020.