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

Bryan Cardenas Guevara, Niccolò Marini, Stefano Marchesin, Witali Aswolinskiy, Robert-Jan Schlimbach, Damian Podareanu, Francesco Ciompi

ExaMode Consortium

Overview

- Typically in machine learning workflows for digital histopathology we use expensive supervised signals
- We orchestrate a set of weakly-supervised transformer-based models with a first aim to address both whole-slide image classification and captioning.
- Our proposed pipeline shows competitive results and emphasizes the need for pre-trained foundation models in digital histopathology.

Models

CLIP A multi-modal model to learn representations between images and text. CLIP uses a contrastive loss. Both the text and the image representations should be one vector. [1].

HIPT Is a hierarchical transformer model that computes a latent of size 192 from a 4096×4096 WSI region. This model was trained with the DINO. [2].

Bio-GPT is a domain-specific generative Transformer language model pre-trained on large scale biomedical literature. [3]

Previous Work

Approaches:

- Rely on supervised signals
- use captions from textbooks [4]
- Use recurrent networks trained from scratch [5]

Problems:

- Labelling is expensive in histopathology
- No extensive use of weakly supervised pre-trained models

Conclusion

Using GPT-3.5-turbo to clean the captions improves 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.

Despite using large transformer models, we fine-tuned our pipeline on a single GPU in 20 minutes. Our work highlights the need for large scale pre-trained models in the field of digital pathology.

References

[1] A. R. et al;. Learning transferable visual models from natural language supervision. *CoRR*, 2021.

[2] R. J. C. et al;. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning.

[3] L. R. et al;. Biogpt: Generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*.

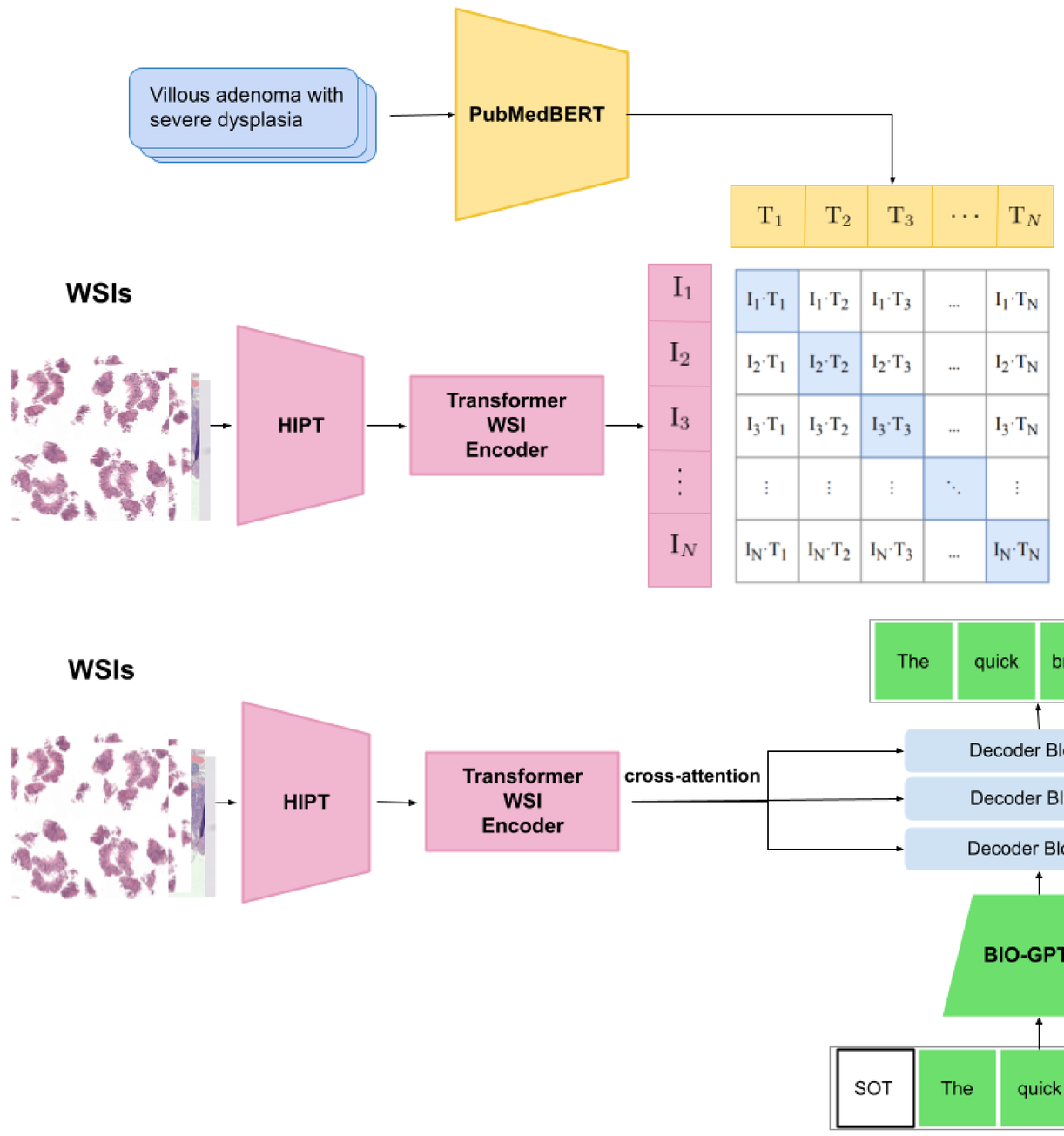
[4] Gamper et al;. Multiple instance captioning: Learning representations from histopathology textbooks and articles. In *CVPR*.

[5] Z. R. et al;. Evaluating and interpreting caption prediction for histopathology images. In *Machine Learning for Healthcare Conference*.

Method

Solution: We address the automatic generation of the conclusion of pathology reports in the form of image captions. Our pathology reports come from two labs and are written originally in two languages: Dutch and Italian.

We use an NMT model to translate the captions. To further get rid of extraneous information and to normalize the format of the reports we prompt GPT-3.5-turbo.



Proposed Pipeline:

1. Extract embeddings with pre-trained models to represent WSIs and text.
2. Train CLIP using these embeddings.
3. Train an additional decoder layer on top of a pre-trained Bio-GPT model conditioned on the CLIP-trained WSI embeddings.

Results

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)

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