

# HistoClipCap : Utilizing ClipCap for histopathology image caption generation

Ammar Rais, Yongjin Kang, and Dr. Dong-Jin Kim  
Department of Data Science, Hanyang University

## Motivations

Deriving a **text description** from an image is a crucial task for disease diagnosis in **histopathology**. However, caption generation models in this domain are typically **heavy**. Therefore, we aimed to explore a **lighter method** for this domain.

We assess the *ClipCap* method to explore the feasibility of training a lighter caption generation model for the histopathology domain.

## Data Preparation

We began on **Quilt-1M** dataset, which contains one million images-text pairs. Because the dataset was sourced from screenshot of online videos, some images required preprocessing procedures.

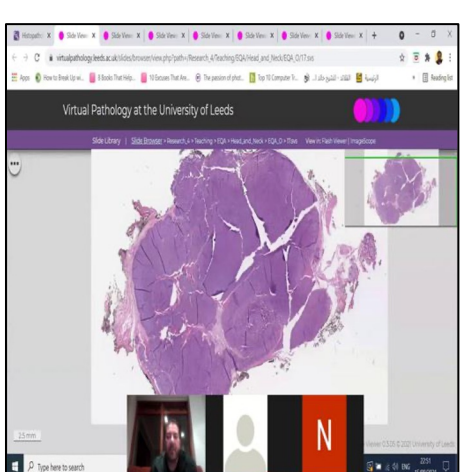


Figure 1: Masking human face

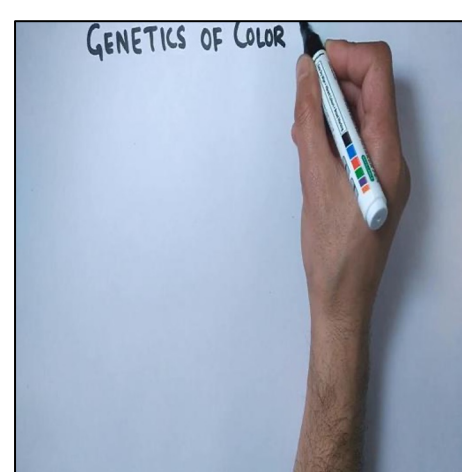
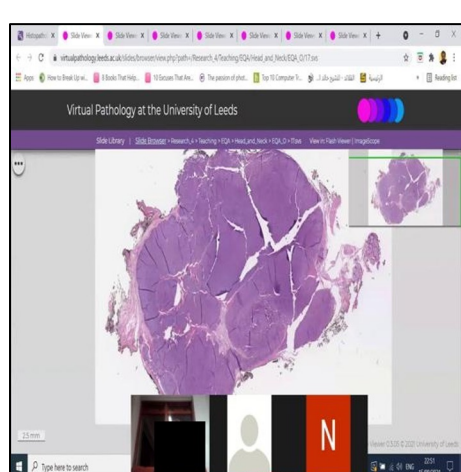


Figure 2: Non-included images



We adjusted or removed such abnormal images to get our final dataset, which contains 80k images.

## Model development

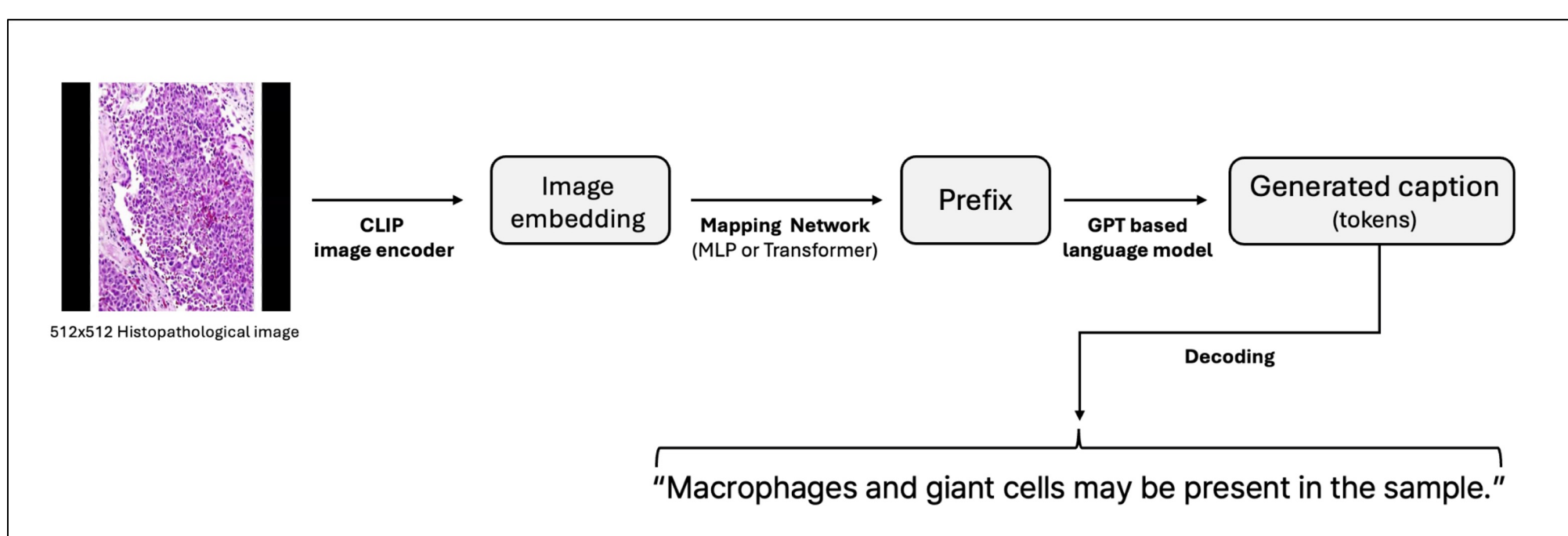


Figure 3: ClipCap Architecture

- ▶ *ClipCap* is a method to build caption generation model using pretrained **CLIP** and **language model**.
- ▶ It introduces '**mapping network**' to close modality gap between 1. **CLIP** and 2. **language model**. We implemented 1. **QuiltNet** and 2. **BioGPT**, which are pretrained for histopathology domain.
- ▶ We testified **MLP**, **Transformer** mapping network.

## Conclusion and Future Works

We found that the ClipCap architecture shows potential as a lightweight solution for building a caption generation model for histopathology. Nevertheless, several limitations and future works remain.

- Model should achieve higher evaluation metric scores to be considered for meaningful uses.
- Techniques related to prefix such as prefix-tuning might help generate better captions.

## Result

Model	BLUE-1	METEOR	ROUGE_L	CIDEr	#Params (M)	Epoch
MLP+BioGPT tuning	0.118	<b>0.093</b>	0.126	0.000	401	20
Transformer+BioGPT frozen	<b>0.142</b>	0.078	<b>0.154</b>	<b>0.013</b>	72	<b>10</b>
Ablation study						
Transformer+BioGPT tuning	0.124	0.102	0.133	0.000	419	30
MLP+BioGPT frozen	0.139	0.075	0.143	0.005	<b>55</b>	50

Table 1: Quantitative result

- ▶ **Transformer+BioGPT frozen** resulted best overall scores and fastest convergence.
- ▶ Transformer worked better as a mapping network.
- ▶ Language model fine-tuning did not seem to always improve the performance.

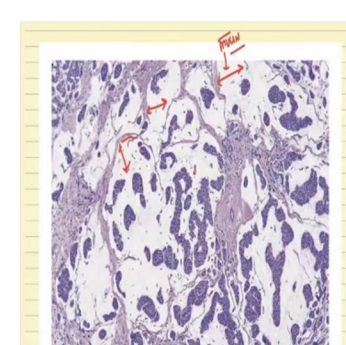
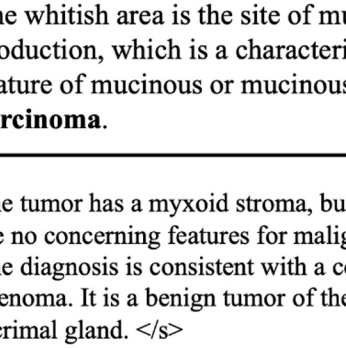
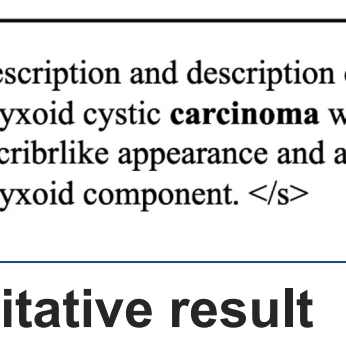

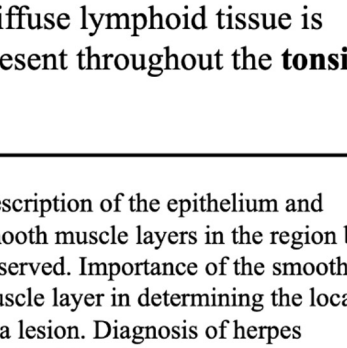
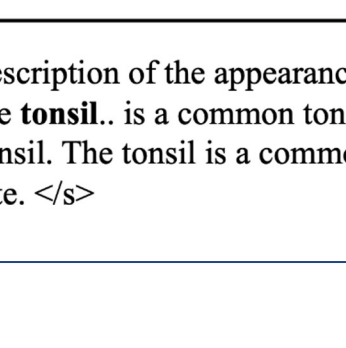
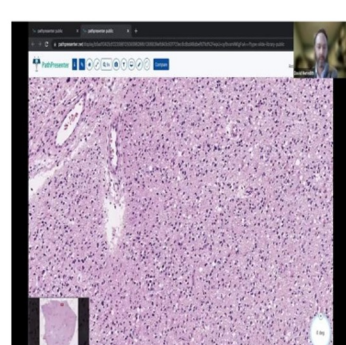
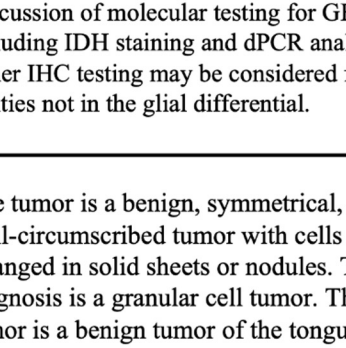
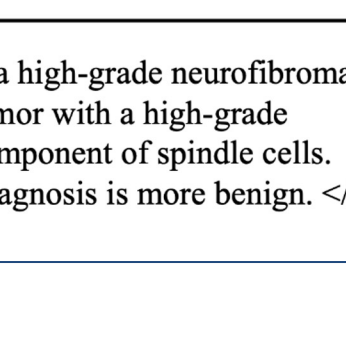
Ground Truth	MLP + BioGPT tuning	Transformer + BioGPT frozen
 The whitish area is the site of mucin production, which is a characteristic feature of mucinous or mucinous carcinoma.	 The tumor has a myxoid stroma, but there are no concerning features for malignancy. The diagnosis is consistent with a colloid adenoma. It is a benign tumor of the lacrimal gland. </s>	 description and description of a myxoid cystic carcinoma with a cribriform appearance and a myxoid component. </s>
 Diffuse lymphoid tissue is present throughout the tonsil.	 Description of the epithelium and smooth muscle layers in the region being observed. Importance of the smooth muscle layer in determining the location of a lesion. Diagnosis of herpes esophagitis. </s>	 description of the appearance of the tonsil. is a common tonsil site. </s>
 Discussion of molecular testing for GBMs, including IDH staining and dPCR analysis. Other IHC testing may be considered for entities not in the glial differential.	 The tumor is a benign, symmetrical, well-circumscribed tumor with cells arranged in solid sheets or nodules. The diagnosis is a granular cell tumor. The tumor is a benign tumor of the tongue. </s>	 is a high-grade neurofibroma tumor with a high-grade component of spindle cells. Diagnosis is more benign. </s>

Figure 4: Qualitative result

- ▶ **Transformer+BioGPT frozen** better captured keywords. (eg. 'carcinoma', 'tonsil')
- ▶ **MLP+BioGPT tuning** was less effective at predicting EOS token at appropriate point, often resulting in captions that hit maximum sequence length.