

Advanced AI for Histopathological Whole Slide Image Classification and Captioning

by

S M Taslim Uddin Raju

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2025

© S M Taslim Uddin Raju 2025

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see *Statement of Contributions* included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

The research presented in this thesis has yielded the following publications:

1. **S M Taslim Uddin Raju**, Md. Milon Islam, Md Rezwanul Haque, Hamdi Altaheri, and Fakhri Karray, “GNN-ViTCap: GNN-Enhanced Multi-Instance Learning with Vision Transformer for Classification and LLM-based Captioning”, International Joint Conference on Neural Networks (IJCNN). [In Press].
2. Md Rezwanul Haque, Md. Milon Islam, **S M Taslim Uddin Raju**, Hamdi Altaheri, Lobna Nassar, and Fakhri Karray, “Multimodal Depression Detection through Mutual Transformer”, IEEE SMC 2025. [Under Review].
3. **S M Taslim Uddin Raju**, Abdul Raqeeb Mohammad, Md. Milon Islam, and Fakhri Karray, “TransUAAE-CapGen: Caption Generation from Histopathological Patches through Transformer and UNet-Based Adversarial Autoencoder”, IEEE International Conference on Systems, Man, and Cybernetics (IEEE SMC), IEEE, Borneo Convention Centre Kuching, Sarawak, Malaysia, 6 - 10 October 2024

In papers 1 and 3, I took the lead on the conceptualization, methodology development, investigative processes, validation, and the drafting of the original manuscript. Md. Milon Islam provided significant contributions to the development and validation of the methodology, and the writing of the original draft.

For paper 2, I contributed to the development of methodology, related works, data presentation, discussion, and drafting, with Md. Rezwanul Haque taking responsibility for the conceptualization, methodology, literature review, discussion, and original draft.

Across all papers, Prof. Fakhri Karray provided thorough reviews, detailed revision, invaluable guidance, and suggestions. Meanwhile, Dr. Hamdi Altaheri further enhanced the work with his review, revision, guidance, and further recommendations.

Chapters 1 through 5 of this dissertation incorporate content from papers 1 to 3.

Abstract

Microscopic assessment of histopathology images is vital for accurate cancer diagnosis and treatment. Classification and captioning Whole Slide Images (WSIs) for pathological analysis is an essential but not extensively explored aspect of computer-aided pathological diagnosis. Challenges arise from insufficient datasets and the effectiveness of model training. Generating automatic caption reports for various gastric adenocarcinoma images is another challenge. Moreover, microscopic WSIs face challenges such as redundant patches and unknown patch positions due to subjective pathologist captures.

The thesis is divided into two folds. At first, a hybrid method referred to as TransUAAE-CapGen is introduced to generate histopathological captions from WSI patches. The TransUAAE-CapGen architecture consists of a hybrid UNet-based Adversarial Autoencoder (AAE) for feature extraction and a transformer for caption generation. The hybrid UNet-based AAE extracted complex tissue properties from histopathological patches, transforming them into low-dimensional embeddings. The embeddings are then fed into the transformer to generate concise captions.

The second fold focuses on how to reduce redundant patches and handle the unknown patch positions due to subjective pathologist captures. To address these challenges, a novel GNN-ViTcap framework is introduced for classification and caption generation from histopathological microscopic images. A visual feature extractor is used to extract feature embeddings. The redundant patches are removed by using deep embedded clustering to dynamically cluster the images and extracting representative images through a scalar dot attention mechanism. The graph is formed by constructing edges from similarity matrix, ensuring that each node is connected to its closest neighbors. Therefore, graph neural network is utilized to extract and represent contextual information from both local and global areas. The aggregated image embeddings are then projected into the language model's input space using a linear layer and combined with input caption tokens to fine-tune the large language models for caption generation.

Our proposed methods are validated using the BreakHis and PatchGastric microscopic datasets. Several ablation studies have been performed to validate our methods. Experimental analysis demonstrates that the proposed TransUAAE-CapGen and GNN-ViTcap architectures outperform state-of-the-art approaches. Our approaches can be integrated into clinical workflows to facilitate early diagnosis and treatment planning, ultimately enhancing patient care using whole slide images.

Acknowledgements

I would like to express my heartfelt gratitude to my supervisor, Prof. Fakhri Karray, for his unwavering support, insightful guidance, and the incredible research opportunities he provided throughout my Master's journey.

In addition, I am also thankful to my colleague, Md. Milon Islam. Our countless brainstorming sessions at the CPAMI lab sparked many innovative ideas, and his assistance in reviewing, revising, and refining our manuscripts, along with the invaluable input from Dr. Hamdi Altaheri, has been essential to our work.

I would like to extend my sincere appreciation to the members of my thesis committee, Dr. Seyed Majid Zahedi and Prof. Ali Elkamel, for their thoughtful evaluations and constructive feedback, which have significantly enriched my research.

Above all, I owe my deepest thanks to my wife, Farhana Sultana Taposi. Her endless love, selfless sacrifice, and unwavering support have been my anchor through every challenge. Without her, none of this would have been possible.

Dedication

This thesis is dedicated to my wife, parents, colleagues, and those who helped me push through until the end.

Table of Contents

Author’s Declaration	ii
Statement of Contributions	iii
Abstract	iv
Acknowledgements	v
Dedication	vi
List of Figures	x
List of Tables	xii
List of Abbreviations	xiii
1 Introduction	1
1.1 Problem Definition	1
1.2 Motivation	3
1.3 Scope of the Work	4
1.4 Contributions	4
1.5 Thesis Organization	5

2	Literature Review	6
2.1	Multiple Instance Learning in WSI	6
2.1.1	Graph-based Aggregation Strategy	7
2.1.2	Attention-based Aggregation Strategy	7
2.2	Vision Language Pretraining in WSI	8
2.3	Biomedical Language Models for WSI Captioning	9
2.4	Summary	11
3	Proposed Architecture: WSI Classification and Captioning	12
3.1	TransUAAE-CapGen Architecture	12
3.1.1	Overview	12
3.1.2	Feature Extraction	14
3.1.3	Caption Generator through Transformer	18
3.2	GNN-ViTCap Architecture	22
3.2.1	MIL Formulation	25
3.2.2	Large Language Models	25
3.2.3	Vision Encodings	27
3.2.4	Attention-based Deep Embedded Clustering	27
3.2.5	Graph-Based Aggregation	30
3.2.6	Visual Embedding Projection	31
3.2.7	Traninging Procedure	31
3.3	Summary	32
4	Experimental Results and Discussions	34
4.1	Datasets	34
4.1.1	BreakHis Dataset	34
4.1.2	PatchGastric Dataset	35
4.2	Performance Evaluation	36

4.2.1	Classification Metrics	36
4.2.2	Image Captioning Metrics	37
4.3	Experimental Analysis: TransUAAE-CapGen Architecture	39
4.3.1	Experimental Setup	40
4.3.2	Training Parameters	40
4.3.3	Caption Generation Performance	40
4.3.4	Comparison with State-of-the-Art	43
4.4	Experimental Analysis: GNN-ViTCap Architecture	45
4.4.1	Implementation Details and Training Phase	45
4.5	Results and Discussions: GNN-ViTCap Architecture	48
4.5.1	Image Classification Results	48
4.5.2	Image Captioning Results	49
4.5.3	Discussions	51
4.5.4	Ablation Studies	55
4.6	Summary	57
5	Conclusion and Future Works	58
5.1	Conclusion	58
5.2	Limitations and Future Works	59
	References	60

List of Figures

1.1	(a) Scanner WSI where the absolute position of each patch is known. (b) Microscopic WSI lacks patch position and contains redundant patches due to multiple captures from a pathologist’s subjective perspective.	2
3.1	Architecture of our proposed TransUAAE-CapGen model for histopathological caption generation.	13
3.2	Architecture of our proposed hybrid UNet-based AAE model for feature extraction.	15
3.3	The loss curve of the training set varies with the number of epochs of our proposed adversarial-UNet encoder.	18
3.4	Architecture of the transformer model for caption generation.	19
3.5	Overview of the GNN-ViTCap framework for microscopic whole slide images classification and captioning. (i) extracting the patches from whole slide image, (ii) extracting image embeddings using a visual feature extractor, (iii) removing redundancy through deep embedded clustering, (iv) extracting representative images with scalar dot attention mechanism, (v) constructing a graph neural network (GNN) using the similarity of representative patches to capture contextual information within clusters (local) and between different clusters (regional), (vi) applying global mean pooling, which aggregates all node representations, (vii) classifying microscopic WSI using aggregated image embeddings and, (viii) projecting the aggregated image embeddings into the language model’s input space using a linear layer, and combining these projections with input caption tokens fine-tunes the LLMs for caption generation.	23
4.1	Sample images of benign and malignant classes from BreakHis dataset with various magnifying factors [1].	35

4.2	Example of biopsy samples with associated captions and 300×300 pixel patches extracted at $20\times$ magnification [2].	36
4.3	Loss curve representation on the PatchGastric dataset using proposed TransUAAE-CapGen model along with four baseline architectures. (a) Training loss (b) Validation loss.	42
4.4	Loss curves on the BreakHis dataset using the proposed GNN-ViTCap with various feature extractors. (a) Training loss, (b) Validation loss.	46
4.5	Loss curves on the PatchGastric dataset using the proposed GNN-ViTCap with various LLMs. (a) Training loss, (b) Validation loss.	47
4.6	t-SNE feature visualizations for the GNN-ViTCap for the BreakHis test dataset. (a) ResNet-34+DEC+GNN-MIL, (b) ViT+DEC+GNN-MIL.	50
4.7	Performance of the proposed GNN-ViTCap architecture on the PatchGastric test dataset using various LLMs	53
4.8	Performance metrics with various numbers of representative image clusters using attention-based deep embedded clustering method. (a) BreakHis, (b) PatchGastric datasets.	56

List of Tables

3.1	Conceptual building block of proposed AdversarialUNet for feature extraction. The final extracted feature dimension is $P \times 18 \times 18 \times 512$	16
4.1	Performance metric (%) of our proposed methods for caption generation on test set of PatchGastric dataset.	41
4.2	Illustration of qualitative results using TransUAAE-CapGen model on test dataset.	44
4.3	Comparison of our proposed method with existing methods using the PatchGastric dataset.	45
4.4	Performance of GNN-ViTCap against SOTA methods on the BreakHis test dataset for classification.	49
4.5	Performance metrics of the proposed GNN-ViTCap against SOTA methods for caption generation on PatchGastric test dataset.	52
4.6	Illustration of qualitative results using the GNN-ViTCap architecture on the PatchGastric test dataset.	54
4.7	Performance evaluation <i>with</i> ($w/$) and <i>without</i> (w/o) representative images on the BreakHis and PatchGastric datasets.	55
4.8	Comparison of AUC (BreakHis) and BLEU@4 (PatchGastric) performances across various graph layers in graph neural networks. Minimal layers enhance performance by focusing on nearby images and leveraging <i>similarity</i>	57

List of Abbreviations

AAE Adversarial Autoencoders [4](#), [5](#), [12–14](#), [17–19](#), [32](#), [40](#), [43](#), [44](#), [58](#)

ABMIL Attention-based Aggregation Operator [2](#), [7](#), [48](#), [51](#)

BLEU Bilingual Evaluation Understudy [38](#)

CIDEr Consensus-based Image Description Evaluation [39](#)

CNN Convolutional Neural Networks [44](#)

DCNN Deep Convolutional Neural Network [8](#)

DEC Deep Embedded Clustering [27](#), [28](#)

DSMIL Dual-Stream MIL [7](#), [8](#), [48](#), [51](#)

DTFD-MIL Double-Tier Feature Distillation MIL [8](#), [48](#), [51](#)

FFN Feed Forward Network [22](#)

GANs Generative Adversarial Networks [17](#)

GAT Graph Attention Networks [7](#), [30](#), [31](#), [45](#)

GCN Graph Convolutional Networks [7](#)

GCNN Graph Convolutional Neural Network [7](#)

GNN Graph Neural Networks [7](#), [12](#)

GNN-MIL Graph Neural Network based Multiple Instance Learning [5](#), [33](#), [51](#), [53](#)

GPT-3 Generative Pretrained Transformer 10

H&E Hematoxylin and Eosin 1, 35

KL Kullback-Leibler 28

LLMs Large Language Models 3, 9, 26, 33, 51, 54, 58, 59

LSTM Long Short-Term Memory 2, 4, 5, 10, 43, 49, 54

METEOR Metric for Evaluation of Translation with Explicit Ordering 38, 39

MIL Multiple Instance Learning 2, 5–11, 25, 48, 49, 51, 53

MLP Multilayer Perceptron 32

MSE Mean Squared Error 17

PE Positional Encoding 20

PFI Platinum-Free Intervals 7

RNNs Recurrent Neural Networks 2, 4, 5, 9

ROI Region of Interest 2

ROUGE Recall-Oriented Understudy for Gisting Evaluation 39

RPNs Region Proposal Networks 2, 3

seq2seq Sequence-to-Sequence 19, 20

SGMT Subtype-Guided Masked Transformer 10, 43, 44, 49, 53

SOTA State-of-the-Art 5, 11, 36, 43, 48, 49, 51, 53

TCGA The Cancer Genome Atlas 8

TF-IDF Inverse Document Frequency 9

TransMIL Transformer-based MIL 2, 8, 48, 51

ViTs Vision Transformers 3, 59

WSI Whole Slide Image 1–3, 5–14, 18, 20–22, 25, 27, 30, 31, 33–37, 40, 43, 44, 51, 58

Chapter 1

Introduction

This chapter introduces histopathological [Whole Slide Image \(WSI\)](#) analysis for classification and caption generation, highlighting the role of deep learning in computational pathology. It discusses key challenges in [WSIs](#) processing, including large image sizes, spatial dependencies, and redundant information, and explores the significance of vision-language models in improving interpretability and diagnostic accuracy. The scope of this research is defined through the development of the proposed TransUAAE-CapGen and GNN-ViTCap architectures. Finally, the chapter highlights the unique contributions of this dissertation in advancing automated pathological captioning and classification.

1.1 Problem Definition

Histopathology is the microscopic examination of tissue, serving as the benchmark for cancer diagnosis, treatment decisions, and prognosis [2, 3]. The traditional process involves pathologists using a microscope to evaluate tissue samples for morphological irregularities to provide essential information for clinical decision-making. However, the advancement of deep learning technologies has significantly propelled computational histopathology, particularly through training models with gigapixels [WSIs](#) obtained from [Hematoxylin and Eosin \(H&E\)](#)-stained specimens [4]. Although cancer detection and classification are crucial, pathologists typically prepare diagnostic reports based on their observations of [H&E](#)-stained slides. These detailed captions provide invaluable insights that enhance the diagnostic process. Therefore, automated pathological report generation can make model predictions more understandable and provide pathologists with richer contextual information to guide their decision-making. Thus, collecting high-quality

WSI-text pairs is crucial for advancing visual-language models and promoting innovation in computational pathology [5].

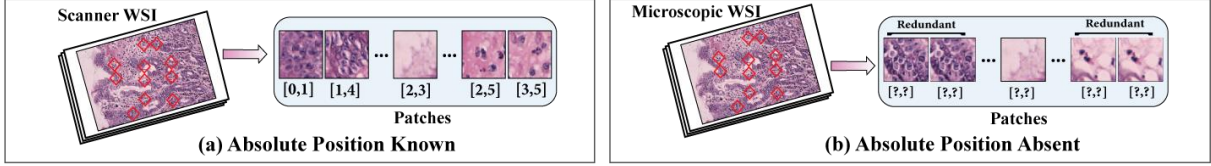


Figure 1.1: (a) Scanner WSI where the absolute position of each patch is known. (b) Microscopic WSI lacks patch position and contains redundant patches due to multiple captures from a pathologist’s subjective perspective.

However, histopathology WSIs have limitations due to their large size, complicating computational analysis. **Multiple Instance Learning (MIL)** emerged as a popular approach for WSI classification by segmenting large images into smaller patches (i.e., instances) to form a slide-level (bag-level) representation. However, MIL assumes patches from the same WSI are independent, neglecting vital tissue context and spatial interactions [6]. In contrast, pathologists consider spatial organization and relationships between patches for comprehensive tissue analysis. Recently, microscopic imaging offers cost and memory advantages as an alternative to scanned WSIs. However, it introduces complexities such as the absence of absolute positional data and redundant patches from multiple subjective captures. Figure 1.1(a) demonstrates the scanner WSI where the absolute position of the patches is known. In contrast, Figure 1.1(b) presents the microscopic WSI lacking the absolute position of patches and the presence of redundant patches due to multiple captures from a pathologist’s subjective perspective. Existing approaches such as **Attention-based Aggregation Operator (ABMIL)** [6], and **Transformer-based MIL (TransMIL)** [7] struggle with global modeling and capturing long-distance dependencies due to limited consideration of inter-patches relationships. Graph-based MIL approaches [8] also encounter challenges in microscopic image analysis within patients, as the absence of absolute patches prevents the construction of effective graph edges. Furthermore, the redundancy in microscopic images leads to excessively dense and repetitive graph connections, which reduces the models’ ability to effectively capture global context and long-distance dependencies.

Moreover, recurrent architectures such as **Recurrent Neural Networks (RNNs)** and **Long Short-Term Memory (LSTM)** have become popular in the field of pathological image captioning due to their capacity to model sequential data [2, 9]. However, these methods struggle with vanishing gradients and cannot resolve long-range dependencies. However, these models suffer from vanishing gradients and have trouble resolving long-range dependencies. Conventional approaches to natural image captioning depend on either dense grid features [10] or sparse **Region of Interest (ROI)** features extracted from **Region Proposal Networks (RPNs)** [11]. However, these ap-

proaches are not capable of directly translating to the context of pathological images. RPNs may find it challenging to diagnose these lesions because the attention regions within a WSI lesion are complex and dispersed. Furthermore, dense grid features can lead to lengthy caption sequences, which can increase memory usage and cause computational inefficiencies. These problems could potentially make the model more sensitive to noise or prone to overfitting.

1.2 Motivation

Automated pathological captioning plays a crucial role in generating text descriptions of WSIs, aiding in the analysis and interpretation of pathological images [2]. Such automated captioning produces concise and targeted descriptions of features within WSIs, allowing pathologists to efficiently focus on specific areas of clinical interest [12]. The integration of computational approaches is vital to enhance the accuracy and consistency of pathological investigations, contributing to improved computer-aided diagnosis. Nevertheless, automated pathological captioning remains an underexplored research area primarily due to the scarcity of high-quality medical datasets [13]. Moreover, WSIs can be categorized into two main types: scanner-based and microscope-based. Scanner-based WSIs are systematically divided into multiple patches with clearly defined absolute positions. In contrast, microscope-based analysis offers a cost-effective solution, saving both cost and memory. However, it relies heavily on the subjective selection of patches by pathologists, causing absolute patch positions to remain unknown and leading to redundancy due to subjective interpretation. This variability introduces further complexity in developing reliable automated captioning systems. Deep learning methods have significantly advanced digital pathology, enabling precise and automated analysis of WSIs to generate pathological reports [14]. However, these models must adapt to the large image sizes, varied data quality, and complex nature of pathology. While they can generate reports, the expertise of pathologists remains indispensable [15]. Recently, Large Language Models (LLMs) have drawn significant attention for their advanced capability to process and understand complex text [16]. Biomedical language models demonstrate outstanding ability at caption generation by effectively translating complex medical images into clinically relevant descriptions [17, 18, 19]. In addition, Vision Transformers (ViTs) leverage transformer-based architectures to transform visual data into high-dimensional representations, enhancing image analysis [20]. Integrating ViTs with biomedical language models facilitates multimodal integration of visual and textual data, thereby enhancing caption generation and classification tasks for more accurate medical image interpretation [21, 22, 23].

1.3 Scope of the Work

This research introduces two novel architectures, TransUAAE-CapGen and GNN-ViTCap, for histopathological whole slide image classification and captioning. TransUAAE-CapGen utilizes a UNet-based [Adversarial Autoencoders \(AAE\)](#) for feature extraction and a transformer for caption generation. GNN-ViTCap leverages attention-based deep-embedded clustering and graph-based aggregation to improve classification and captioning from microscopic images. These methods integrate vision transformers and biomedical language models to enhance interpretability and accuracy. The study aims to address challenges such as redundant image patches, spatial relationships, and the lack of high-quality WSI-text datasets.

1.4 Contributions

The main contributions of this thesis are twofold. Firstly, we developed a hybrid architecture, TransUAAE-CapGen, to generate captions for histopathological images. The method consists of a UNet-based [AAE](#) for feature extraction and a transformer for caption creation. The AAE, with its UNet autoencoder, extracts complex tissue properties and converts them into low-dimensional embeddings. These embeddings are then used by the transformer to produce succinct captions. Transformers exhibit exceptional context-association capabilities, offering a promising alternative to [RNNs](#) or [LSTM](#)-based methods. The contributions can be summarized as follows:

- A hybrid UNet-based [AAE](#) model is proposed for feature extraction. This model effectively uses skip connections to maintain spatial information throughout both the encoder and decoder paths. The UNet architecture captures local and global features by merging, contracting and expanding paths with skip connections.
- The adversarial UNet overcomes the latent space regularization limitation inherent in traditional autoencoders. By integrating the adversarial component, the model offers better control over the latent space, producing efficient lower-dimensional representations of large images.
- The TransUAAE-CapGen model employs the transformer architecture, which aids in mapping large sequences of latent representations of images to their respective captions, leading to accurate caption generation.

Secondly, we proposed a novel GNN-ViTCap method for classification and caption generation from histopathological microscopic images. The method comprises a visual feature extractor, attention-based deep embedded clustering, graph-based multiple instance learning, and the

high capabilities of biomedical language models. An attention-based deep embedded clustering method selects the most representative images to remove the redundant images, and graph-based aggregation [Graph Neural Network based Multiple Instance Learning \(GNN-MIL\)](#) leverages the spatial relationships between image patches. The large language model combined with visual feature embeddings exhibits exceptional context-association capabilities, offering a promising alternative to [RNNs](#) or [LSTM](#)-based methods. The contributions can be summarized as follows:

- Applying a visual feature extractor to extract feature embeddings from histopathological image patches.
- Developing an attention-based deep embedded clustering method and selecting representative images using a scaled dot attention mechanism to remove redundancy in microscopic images.
- Constructing the graph neural network based on the similarity of representative images to capture spatial and contextual information within and between clusters.
- Projecting aggregated image embeddings into the language model’s input space using a linear layer, then combining them with caption tokens to fine-tune the LLMs for caption generation.

1.5 Thesis Organization

The rest of this thesis is organized as follows. Chapter 2 reviews the [State-of-the-Art \(SOTA\)](#) techniques on [WSI](#) classification and captioning, covering recent advancements in computational pathology and deep learning-based approaches. The chapter also explores the most recent literature on [MIL](#) methods, including graph-based and attention-based approaches, alongside vision-language pretraining models to enhance [WSI](#) interpretability and diagnostic insights. Chapter 3 first demonstrates the details of the proposed TransUAAE-CapGen architecture, including hybrid UNet-based [AAE](#) feature extraction and transformer models. Therefore, the chapter also presents the GNN-ViTcap architecture, including feature extraction, self-attention-based deep-embedded clustering, GNN-MIL, and large language models. Chapter 4 describes the microscopic [WSI](#) datasets, the experimental setup used to evaluate the proposed methods, research questions, comprehensive evaluations, experimental results, and ablation studies. Lastly, Chapter 5 summarizes the findings, discusses the limitations, and mentions potential directions for future research.

Chapter 2

Literature Review

In this chapter, we explore the advancements in deep learning methodologies for histopathological image analysis, focusing on novel architectures designed to handle the challenges of **WSIs**. We first divide the related works into two subsections, focusing on **MIL** strategies and the roles of graph-based and attention-based methods in **WSI** analysis. Therefore, we examine the current literature on advancements in vision-language pretraining and biomedical language models, which align visual features with textual data to enhance the interpretability of **WSIs**. We discuss the strengths and weaknesses of all reviewed works.

2.1 Multiple Instance Learning in WSI

Multiple instance learning is now a significant method for analyzing **WSIs** with pyramid structure and giga-pixel dimensions in digital pathology, highlighting the applications of cancer subtyping, staging, grading, and tissue segmentation [6, 24, 25]. The main challenge associated with **MIL** is to combine the large number of instance features to generate a comprehensive bag feature for the specific task. Traditional methods for aggregating patch features in a handcrafted manner typically utilize non-parametric operators including mean pooling and max pooling. **MIL** methods can be divided into two categories: structure-specific or graph-based, and structure-free or attention-based.

2.1.1 Graph-based Aggregation Strategy

Recent research has shown that [Graph Convolutional Networks \(GCN\)](#), [Graph Neural Networks \(GNN\)](#) and other graph-oriented techniques are capable of capturing the global structural information within knowledge graphs. Li et al. [26] developed a graph convolutional neural network [Graph Convolutional Neural Network \(GCNN\)](#) named DeepGraphSurv to enhance survival analysis using whole slide images from cancer patients. The approach combined local patch features with global topological information through spectral graph convolution. This integration captured both intra-modal and inter-modal features within the WSI, fostering robust survival predictions. The experimental study revealed that the proposed DeepGraphSurv method achieved higher C-index scores on TCGA [27] and NLST [28] datasets, outperforming traditional models. However, the performance of the method may decline with poor image quality due to its dependence on fine-tuned patch features. In another study, Bontempo et al. [29] proposed a novel graph-based multi-instance learning approach, DAS-MIL, for classifying histological whole slide images. This method leveraged a pyramidal [GNN](#) integrated with self-knowledge distillation to improve information flow across multiple resolutions. The experimental study revealed that the proposed DAS-MIL technique achieved accuracy scores of 94.5% from the CAMELYON-16 [30] dataset and 92.5% from the TCGA lung cancer [27] dataset. Bontempo et al. [31] introduced the GDS-MIL framework, integrating a [Graph Attention Networks \(GAT\)](#) with [MIL](#) to enhance the prediction of [Platinum-Free Intervals \(PFI\)](#) in [WSIs](#). The method was validated with the "Ome18" dataset [32], comprising 176 [WSIs](#) from 77 patients diagnosed with high-grade serous ovarian cancer. The findings showed that the GDS-MIL approach achieved an accuracy of 0.704 and an AUC of 0.661, outperforming existing models.

2.1.2 Attention-based Aggregation Strategy

The structure-free method typically employs attention-based [MIL](#), utilizing attention scores to weight instance embeddings, thereby learning the slide-level representation. Ilse et al. [6] introduced an [ABMIL](#) utilizing a learnable neural network to enhance the contribution of each instance via trainable attention weights. [ABMIL](#) method addressed the multiple instance learning problem for histopathological images, where a bag of instances is assigned a single class label. The [MIL](#) problem was formulated by representing the bag label as a Bernoulli distribution [33], where neural networks parameterized the probability of the bag label. The experimental study was carried out with two histopathology datasets, breast cancer [34] and colon cancer [35], to evaluate the proposed architecture. The attention-based model outperformed max/mean pooling on the breast and colon cancer datasets, achieving accuracy of 75.5% and 90.4%, respectively. Li et al. [36] proposed a [Dual-Stream MIL \(DSMIL\)](#) framework to classify [WSIs](#) using multi-scale features.

The features are extracted from patches through a self-supervised contrastive learning approach. These features from various magnification levels are then fused using a pyramidal strategy, where higher magnification features are concatenated with those from lower magnifications, enhancing the informational depth. The findings revealed that the [DSMIL](#) approach appraised an accuracy of 89.92% from the CAMELYON-16 [\[30\]](#) dataset and 94.76% from [The Cancer Genome Atlas \(TCGA\)](#) lung [\[27\]](#) dataset. Although the model leverages self-supervised contrastive learning for robust patch embeddings, the method’s extensive patch fusion significantly increases computational demands.

In another study, Shao et al. [\[7\]](#) presented a novel [TransMIL](#) to analyze both morphological and spatial aspects in [WSI](#) classification. [TransMIL](#) method leveraged the self-attention technique, utilizing the output data of a transformer network to encode the mutual correlations among instances. The developed system was found to obtain an AUC scores of 93.09% for CAMELYON-16 [\[30\]](#) and 96.03% and 98.82% for TCGA-NSCLC [\[27\]](#) and TCGA-RCC [\[27\]](#), respectively. The [TransMIL](#) approach is optimized for 20× magnification, limiting its use at higher magnifications due to increased computational and memory requirements. Zhang et al. [\[37\]](#) designed the [Double-Tier Feature Distillation MIL \(DTFD-MIL\)](#) method to increase the total number of bags through the use of pseudo-bags, resulting in a double-tier [MIL](#) system to utilize inherent features effectively. This network determined the probability of the instance within the structure of attention-based [MIL](#) and employed this derivation to assist in generating and analyzing the image features. The proposed system is evaluated with two benchmark datasets: CAMELYON-16 [\[30\]](#) and TCGA [\[27\]](#), and achieved an accuracy of 90.8% and 89.4% for the CAMELYON-16 [\[30\]](#) and TCGA [\[27\]](#) datasets, respectively. Pal et al. [\[38\]](#) presented a deep [MIL](#) to detect the abnormal cell using cervical histopathology image samples. [Deep Convolutional Neural Network \(DCNN\)](#) extracted the high-level features from cervical cell images. The proposed sparse attention-based [MIL](#) obtained an average accuracy of 84.55% on the test set.

2.2 Vision Language Pretraining in WSI

Recent studies have focused on pre-training multimodal models that combine visual and textual data to analyze histopathology images [\[39\]](#).

A few works, including CLIP [\[21\]](#) and ALIGN [\[22\]](#), and Florence [\[40\]](#) have shown that training on large and various web-sourced datasets of paired image samples and captions enables networks to build excellent zero-shot transfer abilities. These models attain this by employing prompts that deploy cross-modal alignment between visuals and texts learned during the pre-training phase. CLIP-based methodologies have demonstrated robust performance and adaptability by coherently mapping numerous image-text pairings sourced from the internet into a

unified embedding space. MI-Zero [41] is introduced to leverage the zero-shot transfer abilities of contrastively correlated visual and text architectures for histopathological WSI samples. This system enabled pretrained encoders to execute various downstream diagnostic tasks without requiring extra labeling. Moreover, the MI-Zero architecture redefines zero-shot transfer within the MIL framework, resolving the computational difficulties of performing inference on very large images. In PathAlign [5], the use of BLIP-2 [42] improved image-text alignment for giga-pixel WSIs. This framework obtained vision-language alignment by linking WSIs with their associated medical texts from pathology records. The pre-trained LLMs and the WSI encoder are aligned in this research to create efficient and significant diagnostic reports from WSIs. CPath-Omni [43] was introduced for patch and WSI level image analysis, such as classification and captioning that contained 15 billion LLMs parameters. CPath-CLIP, a CLIP-based visual processor, is designed in this architecture to combine various vision models and include an LLM as a text encoder to develop a more robust CLIP framework.

2.3 Biomedical Language Models for WSI Captioning

Biomedical language models have become crucial methods for generating informative captions from histopathological images, improving the interpretability and usability of visual data in healthcare environments [44, 45, 46]. These architectures utilize big pre-trained language models that have been fine-tuned using biomedical literature and annotated image datasets to generate relevant explanations of WSIs.

Tsuneki et al. [2] utilized high-resolution WSIs to create an automated captioning system. The authors pretrained EfficientNetB3 and DenseNet121 models on ImageNet to extract features from patches extracted from the WSIs. These features were then processed using global and 3×3 average pooling to reduce their sizes, and fed into an RNNs-based decoder to generate captions. This approach demonstrated potential in automating caption generation for medical images. Alada et al. [47] presented a comparative analysis of various embedding models for analyzing pathology reports, including BioBERT [45], BioMed-RoBERTa [48], Clinical BioBERT [44], and Inverse Document Frequency (TF-IDF) [49]. The authors found that merging TF-IDF with pre-existing model word embeddings yielded superior results compared to using contextualized word embeddings or TF-IDF alone. The best performance was observed for the model that integrated Clinical BioBERT’s embeddings with TF-IDF vectors, achieving an accuracy of 93.77%. The paper highlights how machine learning can extract valuable insights from pathology reports, which is important for interpreting digital pathology reports. Selivanov et al. [50] developed an architecture that automatically generates captions for clinical images. The authors combined radiological scans with patient information from medical records using two language models,

Show-Attend-Tell [51] and Generative Pretrained Transformer (GPT-3) [52]. The models were tested on two medical datasets, Open-I [53] and MIMIC-CXR [54], as well as a general-purpose dataset named MS-COCO [55]. The results demonstrated its efficient applicability to chest X-ray image captioning. The paper emphasizes the use of attention mechanisms and pretrained transformers to generate clear and informative results, which can aid in clinical decision-making. Additionally, the authors suggest exploring interactive training methods like active learning and dialogue-based ChatGPT further to improve the performance of medical image captioning models. In other work, Zhang et al. [56] introduced a two-step method to analyze whole slide images. First, the authors used a triplet loss-based autoencoder to encode patches within the WSI and categorized them into clusters. Then, patches were sampled from each cluster, and a pretrained ResNet-18 was used to extract features for the patches. Attention pooling was applied to reduce feature vector dimensions, which were then used as input for a LSTM decoder to generate captions. This method selected distinct areas with similar features, avoiding random patch selection from the high-resolution image. However, a limitation of this method lies in its access to only a certain number of patches during the caption generation process rather than the entire image.

In another research, Lu et al. [41] presented MI-Zero, an innovative framework for leveraging the zero-shot transfer abilities of contrastively coordinated image and text models on gigapixel histopathology WSIs. The authors created the largest known dataset of web-sourced image caption pairs for pathology and pre-trained their text encoder on over 550k pathology reports. Based on multiple instance learning, the MI-Zero framework enables multiple downstream diagnostic tasks without requiring additional labels. The optimal model achieved a median zero-shot precision of 70.2% over three distinct real-world tasks related to cancer subtyping. Qin et al. [57] developed the Subtype-Guided Masked Transformer (SGMT) for pathological captioning of whole slide images. The proposed method utilized a Transformer architecture to process sequences of image patches and generate corresponding captions. A subtype-guided token leveraged subtype information to enhance model learning, while an Asymmetric Masked Mechanism addressed the large size and redundancy of WSIs. The method was evaluated using the PatchGastricADC22 [2] dataset and achieved a BLEU@4 score of 55.11% and a CIDEr score of 4.83. The limitation of the SGMT method is the lack of patch position information in the PatchGastricADC22 dataset, which may hinder the model’s ability to capture spatial relationships between patches. Zhou et al. [23] introduced a multimodal multi-task MIL system called PathM3 for WSI classification and caption generation. PathM3 leveraged a query-based transformer to align WSIs with diagnostic texts, overcoming limited text data during training. The method utilized a correlation module to capture instance relationships and integrated multi-task learning for improved performance. The experimental study revealed that the proposed PathM3 technique achieved a classification accuracy of 86.40% and a BLEU@4 score of 0.520 for caption generation on the PatchGastricADC22 [2] dataset. However, the method’s reliance on high-quality, labeled data remains a limitation for

broader deployment.

2.4 Summary

In this chapter, we reviewed the [SOTA](#) techniques and advancements in histopathological image analysis. We started by analyzing the use of [MIL](#) to handle the large-scale image in [WSIs](#), focusing on its applications in cancer subtyping and tissue segmentation. We discussed various [MIL](#) methods, such as graph-based and attention-based approaches, showing how they improve feature aggregation and prediction accuracy by capturing spatial and contextual details in [WSIs](#). Therefore, we studied vision-language pretraining models that help align visual features with text, making [WSIs](#) more interpretable and improving diagnostic insights. Furthermore, we explored the use of biomedical language models to generate informative captions from histopathological images.

Chapter 3

Proposed Architecture: WSI Classification and Captioning

In this chapter, we introduce two novel deep learning architectures designed specifically for histopathological image analysis and caption generation. The first, **TransUAAE-CapGen**, leverages a hybrid UNet-based **AAE** for hierarchical feature extraction from **WSI** patches and a transformer module for sequential caption generation. The architecture extracts local and global features while ensuring reconstruction quality and regularizing the latent space through adversarial training. The second architecture, **GNN-ViTCap**, integrates a vision encoder with attention-driven deep embedded clustering to select representative patch features, followed by a **GNN** module to aggregate local and regional information. Therefore, a visual embedding projection aligns the aggregated features with a language model’s input space, allowing for **WSI** classification and captioning.

3.1 TransUAAE-CapGen Architecture

3.1.1 Overview

In this study, we first propose a TransUAAE-CapGen architecture to generate captions from **WSI** patches. Figure 3.1 demonstrates our proposed TransUAAE-CapGen architecture for generating histopathological captions from **WSI** patches. This architecture incorporates the UNet with **AAE** for feature extraction and a transformer for caption generation. The TransUAAE-CapGen model consists of a hybrid UNet-based **AAE** for feature extraction and a transformer for caption generation. In the hybrid feature extractor, UNet is integrated into **AAE** as its generator. UNet-based

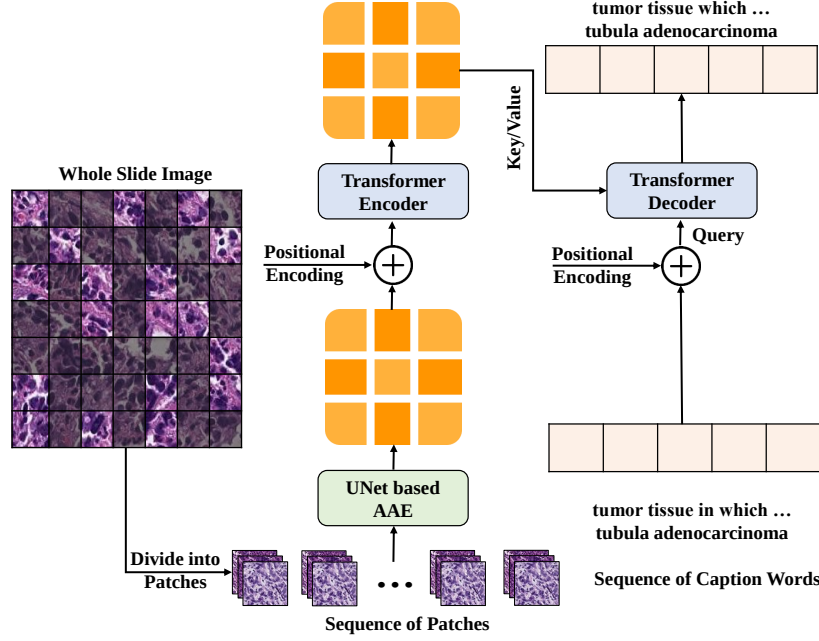


Figure 3.1: Architecture of our proposed TransUAAE-CapGen model for histopathological caption generation.

AAE extracts crucial features and minimizes information loss. AAE reconstructs the input as well as generate data that closely resembles the input. The adversarial interaction between the generator and discriminator in the AAE refines the latent space, promoting high-quality data generation. Moreover, UNet integrated into AAE facilitates the learning of hierarchical features, capturing local features at shallower layers and global features at deeper layers, respectively. After training our hybrid feature extractor model on the WSI, the trained model is used to extract the features and feed them to the transformer with their associated captions. The transformer architecture is selected for its excellent handling of long-range dependencies, faster training due to parallelization, and subtle context understanding through self-attention. The process involves partitioning a WSI x , into a set of patches. Each x consists of P number of patches with $w_p \times w_p$ uniform dimensions. The caption corresponding to WSI is denoted as C . Next, we will present details of our proposed TransUAAE-CapGen model architecture including the feature extraction and the loss functions. The overall system of our proposed TransUAAE-CapGen model is described algorithmically in Algorithm 1.

Algorithm 1: TransUAAE-CapGen for Histopathological Caption Generation

Input: Histopathological WSIs X with related captions C
Output: Generated captions $\hat{C}$ for given WSIs
Step 1: Patch Extraction: Extract patches P from each WSI $x \in X$;
Step 2: Preprocessing: Preprocess patches and captions C ;
Step 3: Feature Extraction: Train hybrid UNet-based AAE model for feature extraction ;
for each WSI $x \in X$ do
 Step 4: Feature Representation: Extract features F_{patch} using trained UNet-based AAE;
 Step 5: Feature Concatenation: Concatenate features F_{patch} based on patient IDs;
end
Step 6: Model Training: Train transformer model on concatenated features and captions C
 using categorical cross-entropy loss in (3.13);
for each WSI $x \in X$ do
 Step 7: Caption Generation: Generate captions $\hat{C}$ using trained transformer model;
end
Output: Generated captions $\hat{C}$ for all WSIs

3.1.2 Feature Extraction

Figure 3.2 presents our proposed hybrid architecture for feature extraction from histopathological patches. This architecture incorporates an AAE along with UNet for feature extraction.

UNet Architecture

In this study, the UNet [58] architecture is exploited for semantic segmentation at the pixel level to extract features from WSIs. The encoder and decoder modules, along with skip connections, extract the important features efficiently from the input images at various levels. The encoder module consists of four encoder blocks. The encoder block contains 3×3 convolutions along with batch normalization, ReLU activation, and dropout layers, respectively. Each encoder block is followed by 2×2 max-pooling layers. Max pooling layers downsample the feature map, enhancing the understanding of context and reducing computational costs. On the other hand, the decoder module consists of four decoder blocks. There is an upsampling layer before each block. Each decoder module employs the same sequence of convolutional layers as in the encoder model, refining the segmentation results. The feature maps of the upsampling layer are concatenated with skip connections from the corresponding encoder features map in the encoder block. The skip connection helps to enrich the network with detailed spatial information. Finally, the output layer of the decoder block reconstructs the image to produce the output. This reconstruction is

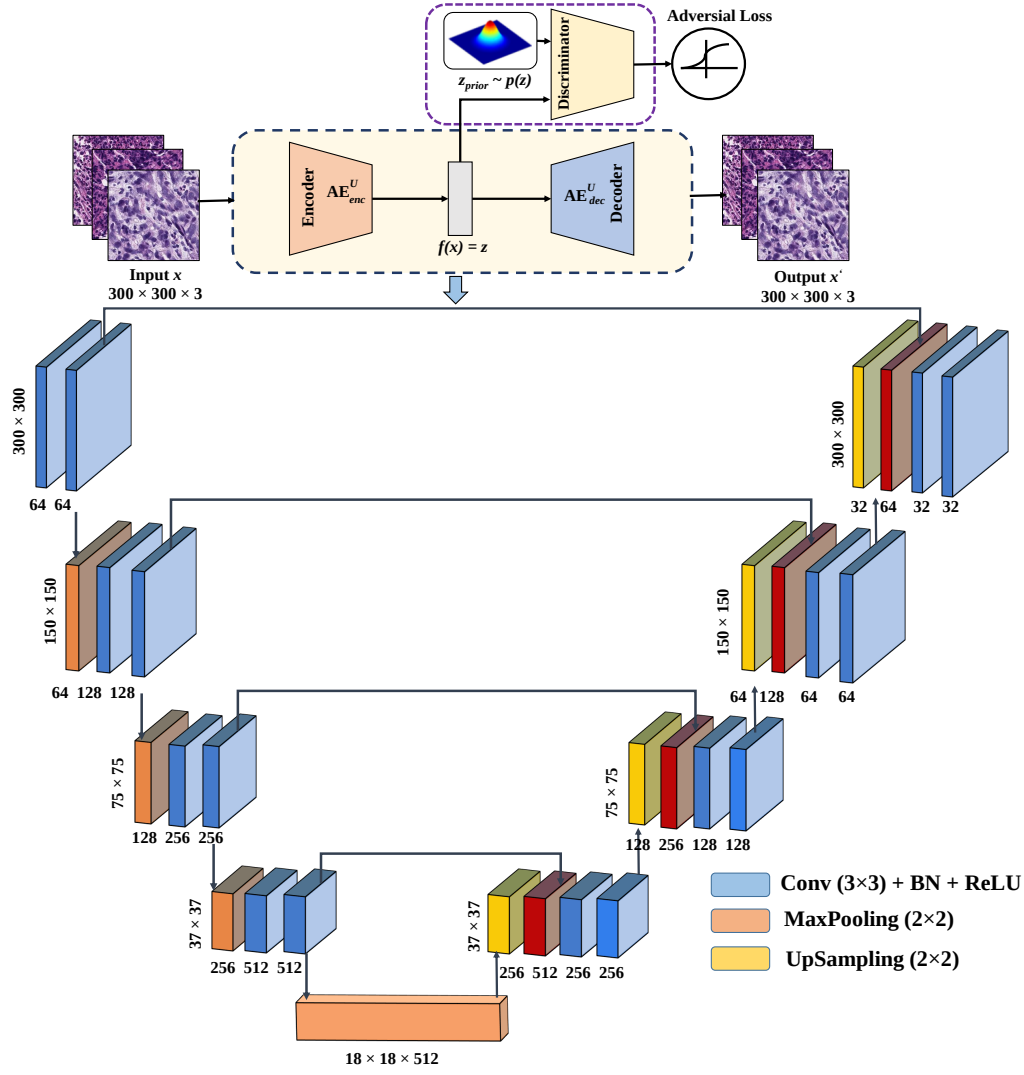


Figure 3.2: Architecture of our proposed hybrid UNet-based AAE model for feature extraction.

obtained through a 3×3 convolutional layer without activation functions, enabling the network to generate the reconstructed image with precise spatial details.

Table 3.1: Conceptual building block of proposed AdversarialUNet for feature extraction. The final extracted feature dimension is $P \times 18 \times 18 \times 512$

Layer	#Filters	Kernel size	Output shape	#Parameters	Layer	#Filters	Kernel size	Output shape	#Parameters
Encoder					Decoder				
Conv2d	64	3 × 3	300 × 300 × 64	1,792	ConvTranspose2d	256	2 × 2	36 × 36 × 256	524,544
BatchNorm2d	64	—	300 × 300 × 64	128	ReLU	256	—	36 × 36 × 256	0
ReLU	64	—	300 × 300 × 64	0	Conv2d	256	3 × 3	37 × 37 × 256	1,769,728
Dropout	64	—	300 × 300 × 64	0	BatchNorm2d	256	—	37 × 37 × 256	512
Conv2d	64	3 × 3	300 × 300 × 64	36,928	ReLU	256	—	37 × 37 × 256	0
BatchNorm2d	64	—	300 × 300 × 64	128	Dropout	256	—	37 × 37 × 256	0
ReLU	64	—	300 × 300 × 64	0	ConvTranspose2d	128	2 × 2	74 × 74 × 256	131,200
Dropout	64	—	300 × 300 × 64	0	ReLU	128	—	74 × 74 × 128	0
MaxPool2d	64	—	150 × 150 × 64	0	Conv2d	128	3 × 3	75 × 75 × 128	442,496
Conv2d	128	3 × 3	150 × 150 × 128	73,856	BatchNorm2d	128	—	75 × 75 × 128	256
BatchNorm2d	128	—	150 × 150 × 128	256	ReLU	128	—	75 × 75 × 128	0
ReLU	128	—	150 × 150 × 128	0	Dropout	128	—	75 × 75 × 128	0
Dropout	128	—	150 × 150 × 128	0	ConvTranspose2d	64	2 × 2	150 × 150 × 64	32,832
Conv2d	128	3 × 3	150 × 150 × 128	147,584	ReLU	64	—	150 × 150 × 64	0
BatchNorm2d	128	—	150 × 150 × 128	256	Conv2d	64	3 × 3	150 × 150 × 64	110,656
ReLU	128	—	150 × 150 × 128	0	BatchNorm2d	64	—	150 × 150 × 64	128
Dropout	128	—	150 × 150 × 128	0	ReLU	64	—	150 × 150 × 64	0
MaxPool2d	128	—	75 × 75 × 128	0	Dropout	64	—	150 × 150 × 64	0
Conv2d	256	3 × 3	75 × 75 × 256	295,168	ConvTranspose2d	32	2 × 2	300 × 300 × 32	8,224
BatchNorm2d	256	—	75 × 75 × 256	512	ReLU	32	—	300 × 300 × 32	0
ReLU	256	—	75 × 75 × 256	0	Conv2d	32	3 × 3	300 × 300 × 32	27,680
Dropout	256	—	75 × 75 × 256	0	BatchNorm2d	32	—	300 × 300 × 32	64
Conv2d	256	3 × 3	75 × 75 × 256	590,080	ReLU	32	—	300 × 300 × 32	0
BatchNorm2d	256	—	75 × 75 × 256	512	Dropout	32	—	300 × 300 × 32	0
ReLU	256	—	75 × 75 × 256	0	Conv2d	32	3 × 3	300 × 300 × 3	867
Dropout	256	—	75 × 75 × 256	0	Decoder	32	3 × 3	300 × 300 × 3	0
MaxPool2d	256	—	37 × 37 × 256	0	Discriminator				
Conv2d	512	3 × 3	37 × 37 × 512	1,180,160	Linear	—	—	256	42,467,584
BatchNorm2d	512	—	37 × 37 × 512	1,024	BatchNorm1d	—	—	256	512
ReLU	512	—	37 × 37 × 512	0	Linear	—	—	128	32,896
Dropout	512	—	37 × 37 × 512	0	BatchNorm1d	—	—	128	256
Conv2d	512	3 × 3	37 × 37 × 512	2,359,808	Linear	—	—	1	129
BatchNorm2d	512	—	37 × 37 × 512	1,024	Discriminator	—	—	1	0
ReLU	512	—	37 × 37 × 512	0					
Dropout	512	—	37 × 37 × 512	0					
MaxPool2d	512	—	18 × 18 × 512	0					

UNet-Based AAE Architecture

AAE [59] is a probabilistic autoencoder that incorporates adversarial training to align the aggregated posterior of its hidden code vector with a specified prior distribution. AAE utilizes Generative Adversarial Networks (GANs) for variational inference [60]. The integration of UNet into AAE aids in acquiring hierarchical features, where local features are captured at shallower layers and global features are captured at deeper layers. This hybrid architecture leverages the strengths of both the UNet and the AAE training mechanism to regularize the latent space and capable of generalizing well to unseen data. The conceptual building block of our proposed hybrid feature extraction architecture is shown in Table 3.1. The reconstruction loss $\mathcal{L}_{recon}$ for the UNet model is computed as the Mean Squared Error (MSE) between the input image x and its reconstruction x' :

$$\mathcal{L}_{recon} = MSE(x, x') = \frac{1}{n} \|x_i - x'_i\|_2^2 \quad (3.1)$$

The adversarial network ensures that the generated feature vectors are different from those of real images. The adversarial loss of the UNet-based AAE architecture is defined as $\mathcal{L}_{adv}$:

$$\mathcal{L}_{adv} = \mathbb{E}_x[\log(D_{AAE}^U(x))] + \mathbb{E}_z[\log(1 - D_{AAE}^U(G^U(z)))] \quad (3.2)$$

Here, D_{AAE}^U is the discriminator of hybrid architecture and G^U denotes the encoder space of UNet. The generator aims to minimize the adversarial loss function, while the discriminator aims to maximize it [60]. The discriminator loss is calculated as follows $\mathcal{L}_D$:

$$\begin{aligned} \mathcal{L}_D = & -\frac{1}{n} \sum_{i=1}^n [y_i \cdot \log(D_{AAE}^U(z_i)) \\ & + (1 - y_i) \cdot \log(1 - D_{AAE}^U(z'_i))] \end{aligned} \quad (3.3)$$

where y_i is the true label (1 for real, 0 for fake), z_i is latent space generated by the encoder, and $z'_i \sim p(z)$ is a sample from true prior. Therefore, the generator loss can be calculated as $\mathcal{L}_G$:

$$\mathcal{L}_G = \frac{1}{n} \sum_{i=1}^n \log(1 - D_{AAE}^U(G^U(z))) \quad (3.4)$$

Therefore, the total loss $\mathcal{L}_{total}$ for the hybrid UNet-based AAE model is a weighted sum of the reconstruction loss $\mathcal{L}_{recon}$ and the adversarial loss $\mathcal{L}_{adv}$, controlled by a hyperparameter λ :

$$\mathcal{L}_{total} = \lambda \cdot \mathcal{L}_{recon} + (1 - \lambda) \cdot \mathcal{L}_{adv} \quad (3.5)$$

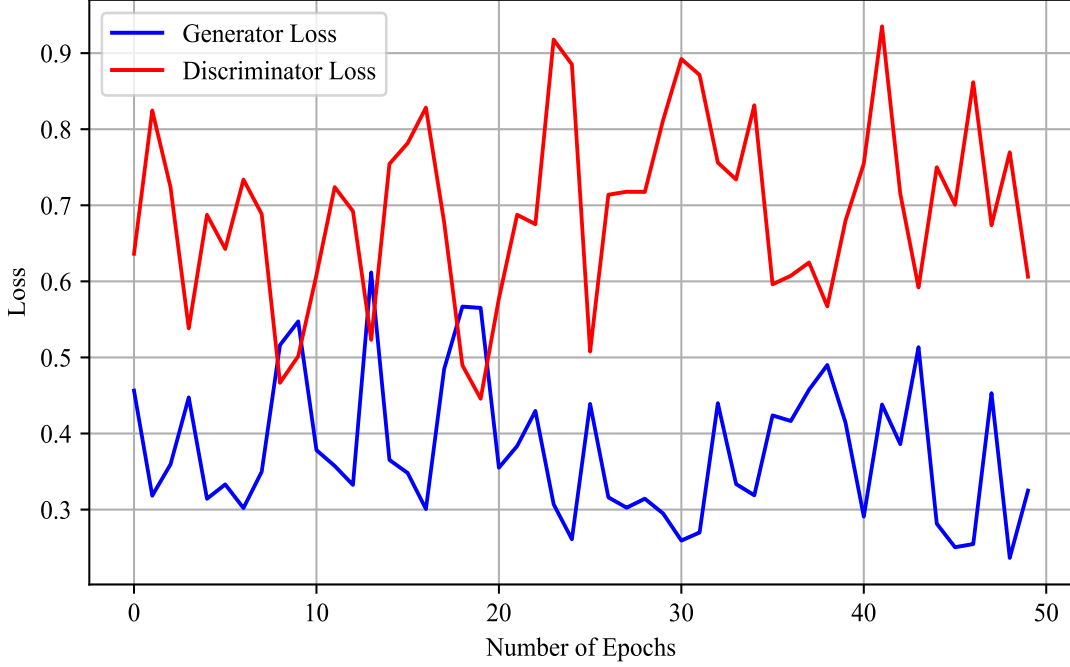


Figure 3.3: The loss curve of the training set varies with the number of epochs of our proposed adversarial-UNet encoder.

Training Phase

The hybrid UNet-based [AAE](#) feature extraction model is trained on histopathological patches. The training is performed over 50 epochs using 30,000 histopathological image samples. Figure 3.3 shows the loss curve of our proposed UNet-based [AAE](#) model. Our UNet-based [AAE](#) architecture is designed with a latent space of dimensions $(18 \times 18 \times 512)$. Features are extracted from the histopathological images using the trained UNet-based [AAE](#) encoder. Then, we concatenate the list of features for each image based on their respective patient IDs. Furthermore, we concatenate these feature sequences along with their corresponding captions. These concatenated datasets are then prepared to be fed into the transformer model for caption generation. The final dimension of our extracted features for each [WSI](#) with a P number of patches is $P \times 18 \times 18 \times 512$.

3.1.3 Caption Generator through Transformer

The sequence of the extracted features and corresponding textual captions are then fed into a transformer model. The dimension of the extracted features is represented as $(P \times X)$, where

P denotes the number of patches for each patient histopathological image and X represents the number of features extracted from each patch. This procedure effectively set up a [Sequence-to-Sequence \(seq2seq\)](#) model for the image captioning of histopathological patches. Figure 3.4 illustrates the architecture of the Transformer model for caption generation.

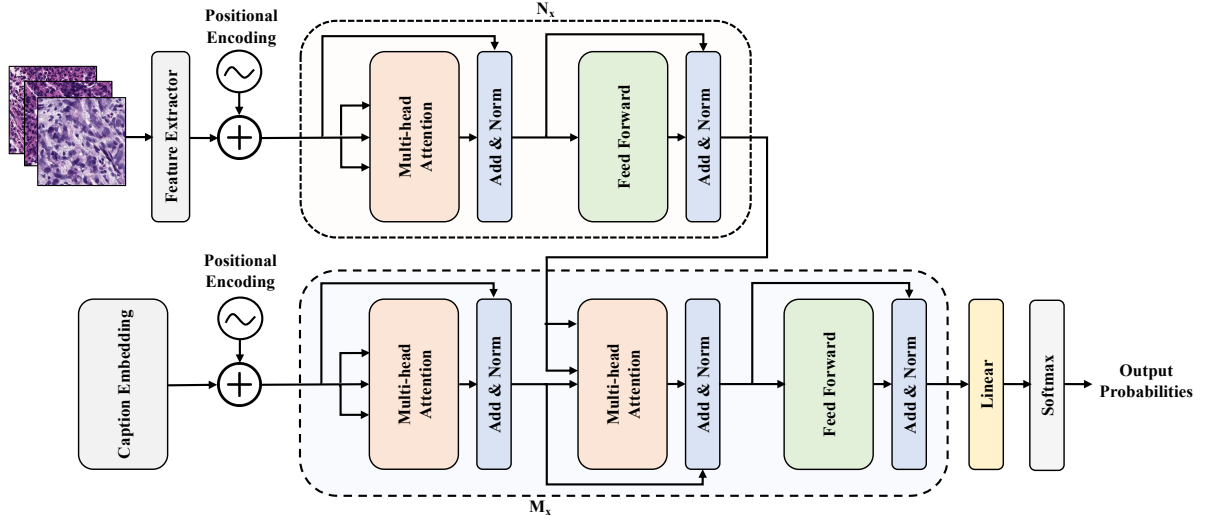


Figure 3.4: Architecture of the transformer model for caption generation.

Encoder

The encoder block consists of $N_x = 3$ identical layers. Each layer has two sublayers comprising a multi-head attention mechanism and a position-wise fully connected feed-forward neural network. A residual connection [61] is applied to each of these sublayers, followed by layer normalization. The output of each sublayer is $LayerNorm(x + sublayer(x))$, where the sublayer implemented the $sublayer(x)$ function itself. The sub-layers and embedding layers of the model produce outputs with a dimensionality of d_m to make it easier for residual connections. Similarly, the outputs in the decoder layers are also set to dimension d_m to ensure compatibility with the encoder output, given the presence of residual connections between the encoder and the decoder. However, in this paper, the dimension of the multi-head attention output is denoted as X , representing the number of features extracted by the UNet-based AAE model. To ensure the residual connection between the encoder and the decoder, the feed-forward sublayer maps the multi-head attention output dimension from X to d_m , aligning it with the output dimension of the decoder sub-layers.

Decoder

The decoder block consists of $M_x = 3$ identical layers. In addition to the two sub-layers present in each encoder layer, an auxiliary sub-layer is integrated into the structure of the decoder. This component enables the decoder to focus on the output stemming from the encoder. Like the encoder architecture, residual connections are applied around each sub-layer, followed by layer normalization to maintain stability. Furthermore, adjustments are implemented in the self-attention sub-layer of the decoder to prevent anticipatory attention, ensuring that predictions are solely influenced by preceding information. The output dimension of sub-layers is maintained at d_m to ensure consistency throughout the model architecture.

Input Embeddings

A transformer model operates as a [seq2seq](#) architecture. The input to the encoder is a sequence with dimensions of $(P \times X)$. Padding is uniformly applied to each patient ID to ensure uniformity in the number of patches across all sequences. Input embedding is applied to the captions of each patient, serving as input to the decoder. The embedded input for the decoder assumes a two-dimensional shape of $(\text{vocab_size} \times \text{number of unique words})$, facilitating the representation of vocabulary size.

Positional Encoding

In the absence of recurrence or convolution, the transformer model relies on positional encoding to understand the sequential order of input sequences. Therefore, prior to entering the transformer encoder, the input embedding (such as extracted [WSI](#) feature map) is added with [Positional Encoding \(PE\)](#) to preserve sequential information. The [PE](#) is computed as follows:

$$\begin{aligned} PE_{(pos, 2i)} &= \sin(pos/10000^{2i/d_m}) \\ PE_{(pos, 2i+1)} &= \cos(pos/10000^{2i/d_m}) \end{aligned} \tag{3.6}$$

where pos denotes the position of an extracted patch feature in the sequence during encoding or a word within the sentence during decoding, and i represents the respective dimension of the embedding. [PE](#) assigns distinct frequencies to different dimensions of the embedded positional vector denoted by sinusoidal or cosinusoidal functions. Thus, various positions within the sequence are associated with unique phases within these periodic functions.

Multi-Head Attention

In transformer architecture, a specialized attention mechanism, referred to as scaled dot-product attention, is utilized. This mechanism diverges from conventional attention mechanisms by calculating attention weights through dot-product operations. In multi-head attention, scaled dot-product attention is concurrently executed across various heads. First, the positional encoded patch feature map $F_{patch} \in \mathbb{R}^{L_{patch} \times d_m}$ is mapped into three matrices: query $Q \in \mathbb{R}^{L_{patch} \times d_q}$, key $K \in \mathbb{R}^{L_{patch} \times d_k}$, and values $V \in \mathbb{R}^{L_{patch} \times d_v}$. Here, L_{patch} denotes the length of the patches for the **WSI** of the histopathological image. Therefore, the dot-product of the query and key are scaled by a factor of $1/\sqrt{d_k}$, which guarantees numerical stability. Furthermore, a softmax function is applied to obtain weights for the corresponding values. The dot-product attention is computed as follows:

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (3.7)$$

The multi-head attention mechanism extends the scaled dot-product attention by incorporating h distinct heads for the query, key, and value. Each head functions autonomously and concurrently. For the i^{th} head, the computation of the attention output is as follows:

$$Z_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \quad (3.8)$$

The multi-head attention concatenates individual attention heads (Z_1 to Z_h) as (3.9).

$$MultiHead(Q, K, V) = Concat(Z_1, \dots, Z_h)W^O \quad (3.9)$$

where, $W_i^Q \in \mathbb{R}^{d_m \times d_q}$, $W_i^K \in \mathbb{R}^{d_m \times d_k}$, and $W_i^V \in \mathbb{R}^{d_m \times d_v}$ denotes the linearly projected matrices that maps the matrices Q , K and V subspaces, respectively. W^O denotes an output projection matrix that aggregates the self-attention outputs from all attention heads.

The output from the multi-head attention, denoted as Z_{MHA} is processed through an addition and normalization step to produce the final output Z_{AN} . This output is then passed to the subsequent feed-forward layer.

$$Z_{AN} = LayerNorm(Z_{MHA} + F_{patch}) \quad (3.10)$$

In this paper, we employ $h = 8$ parallel attention layers, or heads. The encoder utilizes a $d_m = 512$, while the decoder operates with a $d_m = 344$. For each of these, we set $d_k = d_v = d_m/h = 192$ for the encoder and $d_k = d_v = d_m/h = 43$ for the decoder

Feed Forward Network

In a transformer model, the output of Z_{AN} is fed to another sublayer named the **Feed Forward Network (FFN)**. It consists of two linear transformations with ReLU activation. **FFN** is used in each of the encoder and decoder blocks to process the output (Z_{AN}) of the self-attention mechanism. The output of **FFN** is computed as follows:

$$Z_{FFN} = \text{ReLU}(Z_{AN}W_1 + b_1)W_2 + b_2 \quad (3.11)$$

where, $W_1 \in \mathbb{R}^{d_{model} \times d_h}$, $W_2 \in \mathbb{R}^{d_h \times d_{model}}$ represent the weights and $b_1 \in \mathbb{R}^{d_h}$, $b_2 \in \mathbb{R}^{d_{model}}$ denote the biases. d_h represents the size of the hidden state. Finally, Z_{FFN} is fed to *Add and Norm* to generate the transformer encoder's final output (Z_{TE_n}).

$$Z_{TE_n} = \text{LayerNorm}(Z_{FFN} + Z_{AN}) \quad (3.12)$$

We used the same process to generate the transformer decoder output Z_{TD_e} . In this paper, our proposed TransUAAE-CapGen model, the input dimensionality of the **FFN** in the encoder is set to encoder $d_m = 512$, with an output dimensionality of 344. In the decoder, the dimensionality of the **FFN** input and output is set to decoder $d_m = 344$. The total number of trainable parameters of the transformer architecture is 4.1M.

Loss Function

The loss for the transformer model is the categorical cross-entropy loss between the generated caption and the actual caption:

$$\mathcal{L}_{\text{TransUAAE-CapGen}} = -\frac{1}{T} \sum_{t=1}^T \log(p(C_t | C_{<t}, x)) \quad (3.13)$$

where T is the length of the caption, $C_{<t}$ is the sequence of tokens before time step t , and $p(C_t | C_{<t}, x)$ is the probability of the true token C_t given by the TransUAAE-CapGen model for the input **WSI** x .

3.2 GNN-ViTCap Architecture

Figure 3.5 demonstrates the proposed GNN-ViTCap architecture for classification and captioning from microscopic whole slide images.

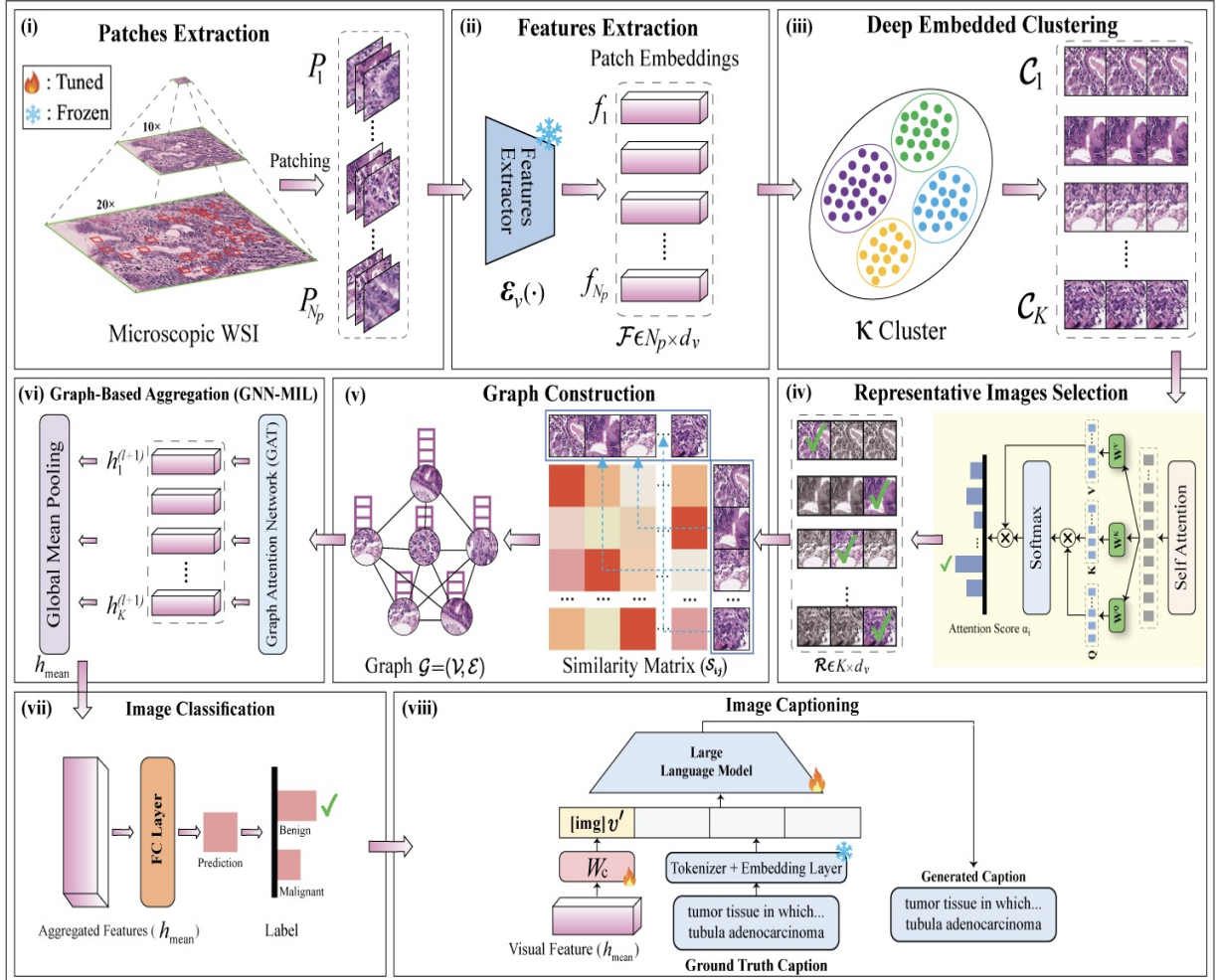


Figure 3.5: Overview of the GNN-ViTCap framework for microscopic whole slide images classification and captioning. (i) extracting the patches from whole slide image, (ii) extracting image embeddings using a visual feature extractor, (iii) removing redundancy through deep embedded clustering, (iv) extracting representative images with scalar dot attention mechanism, (v) constructing a graph neural network (GNN) using the similarity of representative patches to capture contextual information within clusters (local) and between different clusters (regional), (vi) applying global mean pooling, which aggregates all node representations, (vii) classifying microscopic WSI using aggregated image embeddings and, (viii) projecting the aggregated image embeddings into the language model’s input space using a linear layer, and combining these projections with input caption tokens fine-tunes the LLMs for caption generation.

Algorithm 2: GNN-ViTCap architecture for classification and captioning

Input: Dataset: $D = \{(X^{(s)}, Y^{(s)}, C^{(s)}) \mid s = 1, \dots, N\}$, Each $X^{(s)}$ comprises N_p patches $\{p_k^{(s)}\}_{k=1}^{N_p}$, $\mathcal{K} \leftarrow$ number of clusters, $N \leftarrow$ number of patients s , $\epsilon \leftarrow$ convergence threshold, $L \leftarrow$ number of layers.

Output: Predicted class $\hat{y}_{(s)}$ and generated captions $\hat{C}_{(s)}$

```

for  $s \leftarrow 1$  to  $N$  do
  /* Feature Extraction Module */
  Extract features:  $f_k^{(s)} = \mathcal{E}_v(p_k^{(s)}) \ \forall k = 1, \dots, N_p$ ;
  Concatenate features:  $\mathcal{F}^{(s)} = [f_1^{(s)}; \dots; f_{N_p}^{(s)}]$ ;
  /* Clustering Module */
  Initialize cluster centroids,  $\mu_k^{(s)}$ ;
  while  $|\mathcal{L}_{Clu}^{(Curr)} - \mathcal{L}_{Clu}^{(Prev)}| < \epsilon$  do
    Assign clusters using Student's  $t$ -distribution:  $q_{ik}^{(s)} = \frac{(1 + \|f_i^{(s)} - \mu_k^{(s)}\|^2)^{-1}}{\sum_{j=1}^K (1 + \|f_i^{(s)} - \mu_j^{(s)}\|^2)^{-1}}$ ;
    Compute target distribution  $t_{ik}^{(s)}$ ;
    Minimize Kullback-Leibler (KL) divergence  $\mathcal{L}_{Clu}$ ;
    Update cluster centroids  $\mu_k^{(s)}$ ;
  end
  /* Representative images selection */
  for each cluster  $k \leftarrow 1$  to  $\mathcal{K}$  do
    Extract cluster features:  $\mathcal{Z}_k = \{f_i^{(s)} \mid i \in \mathcal{C}_k\}$ ;
    Compute attention scores  $e_i$ , and weights  $\alpha_i$ , using scalar dot attention mechanism;
    Select representative embeddings  $r_k^{(s)}$ ;
  end
  Aggregate representative images:  $\mathcal{R}^{(s)} = [r_1^{(s)}, \dots, r_K^{(s)}]$ ;
  /* Graph-Based Aggregation Module */
  Construct similarity matrix using cosine similarity:  $\mathcal{S}_{ij}^{(s)} = \langle \frac{r_i^{(s)}}{\|r_i^{(s)}\|_2}, \frac{r_j^{(s)}}{\|r_j^{(s)}\|_2} \rangle, \ \forall i, j \in \{1, \dots, K\}$ ;
  Compute edge matrix using Gumbel Softmax:  $\mathcal{E}_{i,j}^{(s)} = 1$ , if  $\mathcal{S}_{i,j}^{(s)} = \max_{k \in \mathcal{N}^{(i)}} \sigma_{\text{gsf}}(\mathcal{S}_{k,j}^{(s)})$ , otherwise  $\mathcal{E}_{i,j}^{(s)} = 0$ ;
  Create graph  $\mathcal{G}_{(s)} = (\mathcal{V}_{(s)}, \mathcal{E}_{(s)})$ ,  $\mathcal{V}_{(s)} = \mathcal{R}_{(s)}$ ,  $\mathcal{E}^{(s)} = \mathcal{E}_{i,j}^{(s)}$ ;
  Apply Graph Attention Networks (GAT) for node aggregation;
  for each GAT layer  $l \leftarrow 0$  to  $L - 1$  do
    for each node  $v \in \mathcal{V}_{(s)}$  do
      Aggregate node features,  $h_v^{(l+1)(s)}$ ;
    end
  end
  Compute the global mean pooling,  $h_{\text{mean}}^{(s)}$  from  $h_v^{(l+1)(s)}$ ;
  /* WSI classification */
  Compute the predicted label,  $\hat{y}_{(s)} = \text{MLP}(h_{\text{mean}}^{(s)})$ ;
  /* WSI caption generation */
  Compute visual embedding projection,  $v'_{(s)} = h_{\text{mean}}^{(s)} \cdot \mathcal{W}_c$ ;
  Generate caption  $\hat{C}_{(s)}$  using language models with visual prefix  $v'_{(s)}$  and start-of-sequence token embeddings;
end
return Class label  $\hat{y}_{(s)}$  and generated captions  $\hat{C}_{(s)}$ 

```

3.2.1 MIL Formulation

Multiple Instance Learning organizes data into *bags* containing multiple *instances*, with labels provided only at the bag level [62]. MIL can be divided into two approaches: instance-based models and bag embedding-based models [63]. Bag embedding-based models are preferred for their richer representations, enhancing WSI classification effectiveness [7]. WSI classification is formulated as MIL problem, where each slide is considered a *bag* and its patches are *instances*. Consider a binary classification problem with a dataset $\mathcal{D} = \{X_i, Y_i\}_{i=1}^B$ where B is the total number of bags. Each bag X_i is represented as a set of instances $X_i = \{x_{i,1}, x_{i,2}, \dots, x_{i,n_i}\}$, where n_i is the number of instances in the i^{th} bag. The label $Y_i \in \{0, 1\}$ is assigned to the bag as a whole rather than to individual instances. According to the standard MIL assumption, the bag label Y_i is described as:

$$Y_i = \begin{cases} 0 & \text{iff } \sum_{j=1}^{n_i} y_{i,j} = 0 \\ 1, & \text{otherwise} \end{cases} \quad (3.14)$$

which can be modeled using max-pooling [64]. In another approach, the bag label $\hat{Y}_i$ can be determined using an aggregation function followed by a classifier as:

$$\hat{Y}_i = g \left(\sigma_{\text{AvgPool}} \left(f(x_{i,1}), f(x_{i,2}), \dots, f(x_{i,n_i}) \right) \right) \quad (3.15)$$

where $f(\cdot)$ is an instance-level feature extractor, $\sigma_{\text{AvgPool}}(\cdot)$ is a permutation-invariant function such as MIL pooling operation, and $g(\cdot)$ is the classifier to predict the bag label [6, 65].

Multiple instance learning can be formulated in two main ways, based on the functions $f(\cdot)$ and $g(\cdot)$. In the instance-based approach, $f(\cdot)$ is an instance classifier that produces a score for each instance, and $g(\cdot)$ is a pooling function that combines these scores to yield a bag-level output. In the embedding-based approach, $f(\cdot)$ is used to transform each instance into an embedding, and $g(\cdot)$ aggregates all instance embeddings into a single bag representation that is then used to produce the final bag score. The embedding-based approach learns a bag-level representation under direct supervision from the bag label and generally achieves higher accuracy than instance-based approach [66]. However, it is often more challenging to identify the specific instances that drive the classifier’s decision in the embedding-based framework [67].

3.2.2 Large Language Models

The integral component of the proposed method involves utilizing both encoder-decoder and decoder-only language model architectures, specifically designed for autoregressive text generation. The image caption $C = C_1, \dots, C_T$ is tokenized into multiple tokens and passed through

an embedded layer to generate a sequence of embeddings. These embeddings are then fed to the decoder, which produces output embeddings. A linear classifier predicts the next token based on these outputs. The LLMs are trained by minimizing the cross-entropy loss between the predicted tokens and the ground truth, thereby fine-tuning the model for accurate caption generation. To address the caption generation task, the proposed architecture utilizes a combination of four LLMs: ClinicalT5-Base [17], employed as an encoder-decoder model, and BioGPT [18], LLamaV2-Chat [16], and BiomedGPT [19], which are utilized as decoder-only models. These models effectively capture linguistic structures and generate coherent text.

ClinicalT5-Base

ClinicalT5-Base [17] is based on the T5 [68] encoder-decoder architecture and is pre-trained on a large corpus of clinical text data. It has 220M parameters distributed across 12 layers, providing a well-balanced structure for handling domain-specific language. During its pre-training phase, the model learns to understand and generate specialized medical terminology, which is crucial for clinical content tasks. In this study, we adapt the ClinicalT5-Base for histopathological caption generation by feeding visual features from whole slide images into the encoder. The decoder then takes caption tokens as input to generate textual descriptions. This combination of visual and textual information allows the model to produce accurate, contextually relevant captions.

BioGPT

BioGPT [18] is a transformer-based language model with a decoder-only structure using the GPT architecture [69]. It has around 347M parameters, depending on the configuration. The model is pretrained on 15M PubMed abstracts focused on biomedical topics. After pretraining, it is fine-tuned on specialized biomedical datasets for tasks like relation extraction, question answering, and document classification.

LLamaV2-Chat

LLama2-Chat [16] is a transformer-based model that's been fine-tuned for conversational tasks. It uses a decoder-only architecture and incorporates grouped query attention to improve scalability. The model comes in various configurations, including 7B, 13B, and 70B parameters, and its training involves a wide range of publicly available datasets. This diverse pretraining helps LLama2-Chat generate contextually relevant responses across a broad topic. In this study, we fine-tuned LLama2-Chat for histopathological caption generation by feeding visual features from whole slide images, known as LLamaV2-Chat.

BiomedGPT

BiomedGPT [19] is a vision-language model designed for biomedical tasks. It combines a BERT-style [70] encoder to process text with a GPT-style [69] autoregressive decoder for generating outputs. Using a multi-head attention mechanism, it can focus on different representation subspaces simultaneously, improving its ability to understand complex data. In addition, it employs a decoupled positional encoding method for both text and image data, incorporating a 1D relative position bias for text and a 2D relative position bias for images to enhance performance. The model has three configurations: BiomedGPT-S (33M parameters), BiomedGPT-M (93M parameters), and BiomedGPT-B (182M parameters). BiomedGPT utilizes a unified vocabulary for text and image tokens for multimodal tasks, which is essential for generating effective multimodal outputs.

3.2.3 Vision Encodings

The GNN-ViTCap integrates a vision encoder with language models to simultaneously process visual and text inputs. For each patient s , given a microscopic whole slide image $X^{(s)}$, associated with the label $Y^{(s)}$ and the corresponding caption $C^{(s)}$. The WSI $X^{(s)}$ comprises a collection of N_p patches or images, $\mathcal{P}^{(s)} = \{p_k^{(s)}\}_{k=1}^{N_p}$, where $p_k^{(s)}$ denotes the k -th patch of patient s . The number of patches N_p varies depending on the specific WSI and the individual characteristics of patient s . The pretrained vision encoder $\mathcal{E}_v$ processes each patch $p_k^{(s)}$ to produce a fixed-length embedding feature vector, a.k.a., a patch token:

$$f_k^{(s)} = \mathcal{E}_v(p_k^{(s)}) \in \mathbb{R}^{1 \times d_v} \quad (3.16)$$

where, d_v is the dimension of the visual embeddings. The patch-level embeddings are concatenated to represent the entire WSI of each patient s :

$$\mathcal{F}_{(s)} = \left[f_1^{(s)}; f_2^{(s)}; \dots; f_{N_p}^{(s)} \right] \in \mathbb{R}^{N_p \times d_v} \quad (3.17)$$

3.2.4 Attention-based Deep Embedded Clustering

In the proposed architecture, an attention-based Deep Embedded Clustering (DEC) [71] is utilized to remove the redundant images from each patient.

Deep Embedded Clustering

Let $F_{(s)} = \{f_1^{(s)}, f_2^{(s)}, \dots, f_{N_p}^{(s)}\}$ denotes the set of features extracted from the images of patient s . The initial cluster centroids $\mu^{(s)} = \{\mu_k^{(s)}\}_{k=1}^K$ are used to assign these features into K image clusters, where K denotes the total number of clusters. The probability of assigning each feature embedding $f_i^{(s)}$ to cluster k is determined using the Student's t -distribution as follows:

$$q_{ik}^{(s)} = \frac{\left(1 + \|f_i^{(s)} - \mu_k^{(s)}\|^2 / \alpha\right)^{-\frac{\alpha+1}{2}}}{\sum_{j=1}^K \left(1 + \|f_i^{(s)} - \mu_j^{(s)}\|^2 / \alpha\right)^{-\frac{\alpha+1}{2}}} \quad (3.18)$$

where $\alpha = 1$ indicates the degree of freedom of Student's t -distribution, and $q_{ik}^{(s)}$ denotes the soft assignment probability of the i -th image from patient s to the k -th cluster. The higher value of $q_{ik}^{(s)}$ indicates a stronger likelihood that the feature embedding $f_i^{(s)}$ belongs to cluster k . Therefore, an auxiliary target distribution $\mathcal{T}_{(s)} = \{t_{ik}^{(s)}\}$ is introduced based on the soft assignments $\mathcal{Q}_{(s)} = \{q_{ik}^{(s)}\}$ to refine the clustering process. The target distribution emphasizes high-confidence assignments and is computed as:

$$t_{ik}^{(s)} = \frac{\left(q_{ik}^{(s)}\right)^2 / \sum_{i=1}^{N_p} q_{ik}^{(s)}}{\sum_{j=1}^K \left(q_{ij}^{(s)} / \sum_{i=1}^{N_p} q_{ij}^{(s)}\right)^2} \quad (3.19)$$

where, $\sum_{i=1}^{N_p} q_{ik}^{(s)}$ prevents bias toward larger clusters by balancing the influence of clusters with more assignments. The [Kullback-Leibler \(KL\)](#) divergence between the target distribution $\mathcal{T}_{(s)}$ and the soft assignment distribution $\mathcal{Q}_{(s)}$ is used as the clustering loss.

$$\mathcal{L}_{\text{clu}} = \mathcal{KL}(\mathcal{T}_{(s)} || \mathcal{Q}_{(s)}) = \sum_{i=1}^{N_p} \sum_{k=1}^K t_{ik}^{(s)} \log \left(\frac{t_{ik}^{(s)}}{q_{ik}^{(s)}} \right) \quad (3.20)$$

The objective of clustering is to minimize the distance between feature embeddings and their cluster centroids. By iteratively minimizing $\mathcal{L}_{\text{clu}}$, the clustering process refines both the embeddings and cluster centroids, resulting in compact and well-separated clusters.

Attention-Based Representative Images

In our approach, [DEC](#) is utilized to cluster images, thereby capturing the underlying relationships among images within same region. To address redundancy and select the most representative

image in each cluster, an attention-based selection mechanism is utilized [72]. This approach reduces intra-cluster redundancy while enhancing inter-cluster communication and information sharing. Let C_k denotes the set of image indices corresponding to cluster k . Therefore, each cluster k contains $N_k = |C_k|$ images, and the feature embeddings of these images are extracted as follows:

$$\mathcal{Z}_k = \{f_i^{(s)} \mid i \in C_k\} \in \mathbb{R}^{N_k \times d_v} \quad (3.21)$$

Then, the embeddings $\mathcal{Z}_k$ are mapped into query, key, and value representations using learnable linear projections.

$$Q^{\text{attn}} = \mathcal{Z}_k \mathcal{W}_{\text{attn}}^Q, \quad K^{\text{attn}} = \mathcal{Z}_k \mathcal{W}_{\text{attn}}^K, \quad V^{\text{attn}} = \mathcal{Z}_k \mathcal{W}_{\text{attn}}^V \quad (3.22)$$

where $\mathcal{W}_{\text{attn}}^Q, \mathcal{W}_{\text{attn}}^K, \mathcal{W}_{\text{attn}}^V \in \mathbb{R}^{d_v \times d_v}$ are parameter matrices. For each patch $i \in C_k$, an attention score is defined based on the element-wise product of its query $Q_i^{\text{attn}} \in \mathbb{R}^{d_v}$ and key $K_i^{\text{attn}} \in \mathbb{R}^{d_v}$:

$$e_i = \frac{\sum (Q_i^{\text{attn}} \odot K_i^{\text{attn}})}{\sqrt{d_v}}, \quad i = 1, \dots, N_k. \quad (3.23)$$

where, $\sqrt{d_v}$ is a scaling factor to ensure balanced weight distributions. Therefore, the softmax function is applied across all patches in the cluster to obtain normalized attention weights:

$$\alpha_i = \frac{\exp(e_i)}{\sum_{j=1}^{N_k} \exp(e_j)}, \quad i = 1, \dots, N_k \quad (3.24)$$

These weights $\alpha_i \in \mathbb{R}$ indicate the relative significance of each patch i within the same cluster. Therefore, a scalar importance score is computed for each patch using its corresponding value embedding $V_i^{\text{attn}} \in \mathbb{R}^{d_v}$:

$$\text{score}_i = \sum_{m=1}^{d_v} \alpha_i \cdot V_{i,m}^{\text{attn}} \quad (3.25)$$

Furthermore, the image with the highest score is selected as the representative image within cluster k , and its corresponding embedding is defined as:

$$r_k^{(s)} = f_{i_k^*}^{(s)} \in \mathbb{R}^{d_v} \quad \text{where} \quad i_k^* = \arg \max_{i \in C_k} \text{score}_i \quad (3.26)$$

Therefore, the final representative features for patient s is computed as follows:

$$\mathcal{R}_{(s)} = [r_1^{(s)}, r_1^{(s)}; \dots; r_K^{(s)}]^T \in \mathbb{R}^{K \times d_v} \quad (3.27)$$

where K is the total number of clusters.

3.2.5 Graph-Based Aggregation

Scanner-based WSI provides defined absolute positions and coordinates, which facilitate the identification of neighboring patches. In contrast, microscopy-based WSI lacks absolute coordinates, making it more challenging to determine neighboring patches. To address this, a graph-based aggregation approach is utilized to identify neighboring representative images based on their relative positions, thereby maximizing their interactions.

Constructing Graph

For each patient s , the graph $\mathcal{G}^{(s)}$ is constructed using a similarity matrix $S^{(s)} \in \mathbb{R}^{K \times K}$, where each node denotes a representative image. The similarity between nodes is calculated using cosine similarity as:

$$S_{ij}^{(s)} = \left\langle \frac{r_i^{(s)}}{\|r_i^{(s)}\|_2}, \frac{r_j^{(s)}}{\|r_j^{(s)}\|_2} \right\rangle, \quad \forall i, j \in \{1, \dots, K\} \quad (3.28)$$

where each element $S_{i,j}^{(s)}$ represents the similarity between image features $r_i^{(s)}$ and $r_j^{(s)}$. Therefore, an edge matrix $\mathcal{E}_{i,j}^{(p)}$ is created by applying the Gumbel Softmax function σ_{gsf} [73] to the entire similarity matrix $S_{ij}^{(s)}$. This process selects the most similar neighbors for each node, resulting in:

$$\mathcal{E}_{i,j}^{(s)} = \begin{cases} 1, & \text{if } S_{i,j}^{(s)} = \max_{k \in \mathcal{N}(i)} S_{k,j}^{(s)} \\ 0, & \text{otherwise} \end{cases} \quad (3.29)$$

where $\mathcal{E}_{i,j}^{(s)} \in \{0, 1\}$ indicates the presence of edge between nodes i and j and $\mathcal{N}(i)$ denotes the set of neighbors for node i . Therefore, the graph $\mathcal{G}_{(s)}$ is defined by its set of nodes $\mathcal{V}_{(s)}$ and edges $\mathcal{E}_{(s)}$:

$$\mathcal{G}_{(s)} = (\mathcal{V}_{(s)}, \mathcal{E}_{(s)}) \quad (3.30)$$

where $\mathcal{V}_{(s)} = \mathcal{R}_{(s)}$ denotes the set of nodes corresponding to the representative images for patient s , and $\mathcal{E}_{(s)} \subseteq \mathcal{V}_{(s)} \times \mathcal{V}_{(s)}$ are edges determined by $\mathcal{E}_{i,j}^{(s)}$.

Graph Neural Network

In the graph $\mathcal{G}^{(s)}$, each node corresponds to a representative image, and the edges denote the relationships between these images. GAT [74] improve attention mechanisms to assign adaptive

weights to neighboring nodes during aggregation. For each layer $l = 0, 1, \dots, L - 1$, the features of node v can be updated as follows:

$$h_v^{(l+1)(s)} = \rho \left(\sum_{u \in \mathcal{N}(v)} \beta_{vu}^{(l)(s)} \mathcal{W}^{(l)} h_u^{(l)(s)} \right) \quad (3.31)$$

where, $\mathcal{W}^{(l)} \in d_v \times d_v$ is the weight matrix at layer l and ρ is an activation function, such as LeakyReLU [75]. The attention coefficient $\beta_{vu}^{(l)}$ between nodes u and v at layer l is computed as:

$$\beta_{vu}^{(l)(s)} = \frac{\exp \left(\rho \left(a^{(l)T} \left[\mathcal{W}^{(l)} h_v^{(l)(s)} \parallel \mathcal{W}^{(l)} h_u^{(l)(s)} \right] \right) \right)}{\sum_{w \in \mathcal{N}(v)} \exp \left(\rho \left(a^{(l)T} \left[\mathcal{W}^{(l)} h_v^{(l)(s)} \parallel \mathcal{W}^{(l)} h_w^{(l)(s)} \right] \right) \right)} \quad (3.32)$$

where $a^{(l)}$ denotes a learnable weight vector at layer l . After L GAT layers, the final node representations $h_v^{(L)}$ are obtained, where each $h_v^{(L)} \in \mathbb{R}^{d_{\text{out}}}$. Consequently, the WSI representation $h_{\text{mean}}^{(s)} \in \mathbb{R}^{d_{\text{out}}}$ is generated by applying global mean pooling, which aggregates all node representations:

$$h_{\text{mean}}^{(s)} = \frac{1}{K} \sum_{v=1}^K h_v^{(L)(s)} \quad (3.33)$$

3.2.6 Visual Embedding Projection

The visual embeddings generated by the vision encoder are not directly compatible with the language model. This incompatibility arises because the vision embeddings have different dimensions and feature distributions compared to the language model's input space. To address this, the aggregated image embeddings $h_{\text{mean}}^{(s)} \in \mathbb{R}^{d_{\text{out}}}$, are transformed using a linear projection matrix $W_c \in \mathbb{R}^{d_{\text{out}} \times d_{\text{model}}}$. This projection maps the visual embeddings into a d_{model} -dimensional space compatible with the language model's input embeddings. The visual prefix can be computed as:

$$v'_{(s)} = h_{\text{mean}}^{(s)} \cdot W_c \in \mathbb{R}^{d_{\text{model}}} \quad (3.34)$$

Therefore, the visual prefix is integrated with the input caption token embeddings to fine-tune the language model for caption generation.

3.2.7 Training Procedure

The GNN-ViTCap architecture is trained for two tasks: image classification and image captioning.

Image Classification

For the image classification, the aggregated image embeddings $h_{\text{mean}}^{(s)}$, are fed into **Multilayer Perceptron (MLP)** to predict the target variable:

$$\hat{y}_{(s)} = \text{MLP}(h_{\text{mean}}^{(s)}) \quad (3.35)$$

where, $\hat{y}_{(s)}$ represents the predicted probability for patient s . The binary cross-entropy loss function is then used for optimization and is defined as:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N} \sum_{i=1}^N [y_i \cdot \log \hat{y}_i + (1 - y_i) \cdot \log(1 - \hat{y}_i)] \quad (3.36)$$

where, N is the number of patients, y_i is the ground-truth label, and $\hat{y}_i$ is the predicted probability for each patient. Therefore, the total loss for the whole slide image classification can be characterized as:

$$\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{BCL}} + \mathcal{L}_{\text{Clu}} \quad (3.37)$$

Caption Generation

In the caption generation task, the visual prefix v' , combined with the start-of-sequence token embeddings of the caption, is fed into the language model. The language model then autoregressively generates caption tokens C_t based on v' and the previously generated tokens C_1 to C_{t-1} . The loss for caption generation is calculated using the negative log-likelihood of the ground-truth captions:

$$\mathcal{L}_{\text{Cap}} = -\frac{1}{N} \sum_{i=1}^N \sum_{t=1}^T \log p_{\theta}(C_{i,t} | v'_i, C_{i,1}, \dots, C_{i,t-1}) \quad (3.38)$$

where, T is the maximum caption length, $C_{i,t}$ is the t -th token of the ground-truth caption for patient i , and p_{θ} represents the probability predicted by the language model with parameters θ . Therefore, the total loss for caption generation for the whole slide image can be characterized as:

$$\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{Cap}} + \mathcal{L}_{\text{Clu}} \quad (3.39)$$

3.3 Summary

In this chapter, we presented two advanced architectures tailored for histopathological image analysis. The **TransUAAE-CapGen** model combines a UNet-based **AAE** for efficient feature

extraction with a transformer network that leverages self-attention and positional encoding for high-quality caption generation. In contrast, the **GNN-ViTCap** framework utilizes vision encodings, attention-based deep embedded clustering, and graph-based aggregation to extract representative patch features, enabling effective [WSI](#) classification via [GNN-MIL](#) and detailed caption generation through integration with [LLMs](#).

Chapter 4

Experimental Results and Discussions

In this chapter, we present a comprehensive analysis of the quantitative and qualitative results from our benchmarking experiments on histopathological whole slide image classification and caption generation. We evaluate the effectiveness of the proposed TransUAAE-CapGen and GNN-ViTCap architectures using a range of performance metrics, including model complexity (number of parameters), classification Precision, Recall, F_1 -score, and Area Under the Curve (AUC). For caption generation, we assess the models using BLEU, METEOR, ROUGE, and CIDEr scores. To further refine our approach, we explore different architectural variations, examine key trade-offs, and provide insights into selecting the most suitable model and quantization strategy for various real-world applications.

4.1 Datasets

To evaluate our proposed architectures for histopathological [WSI](#) classification and captioning, we use two datasets: BreakHis and PatchGastic. BreakHis offers high-resolution images of breast tumors, which allow for a detailed analysis of tissue patterns and cancer subtypes. In contrast, PatchGastic includes image patches from stomach adenocarcinoma endoscopic biopsies that are used for caption generation.

4.1.1 BreakHis Dataset

The BreakHis dataset [76] comprises 7,909 microscopic histopathology biopsy images from 82 patients. Each image is classified into benign and malignant tumor categories. In this study, labels

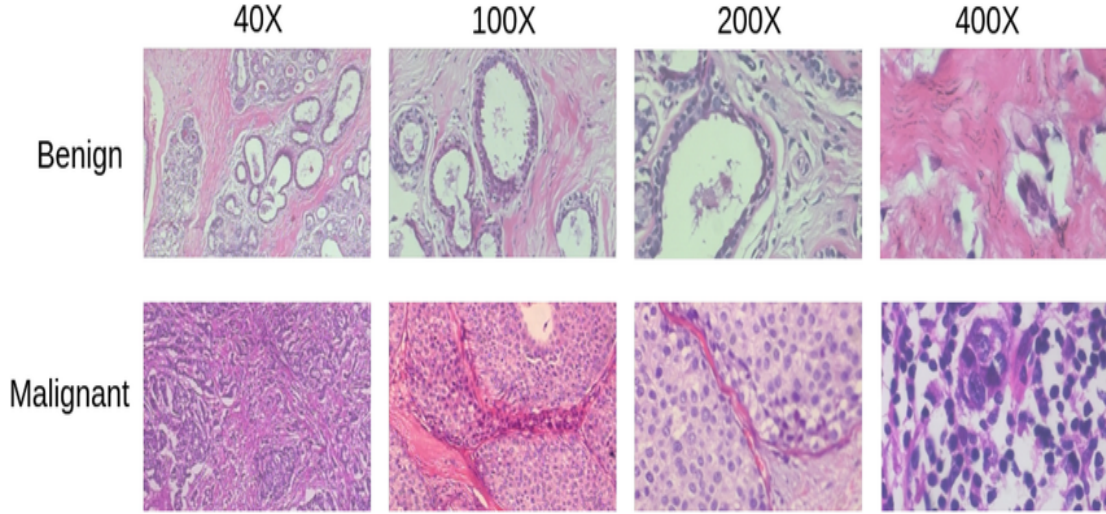


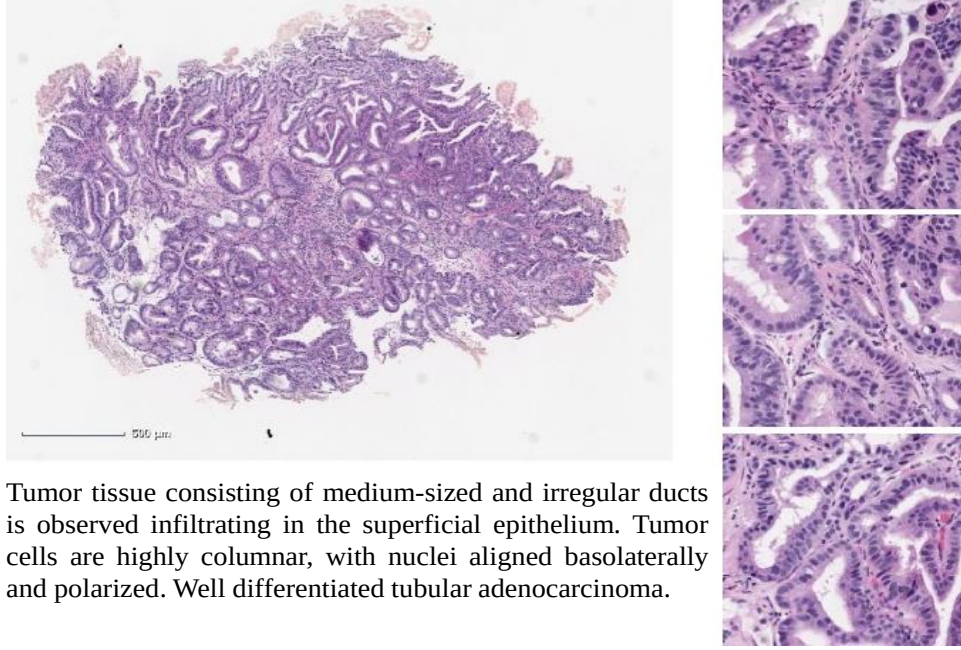
Figure 4.1: Sample images of benign and malignant classes from BreakHis dataset with various magnifying factors [1].

are assigned only at the patient level within the dataset, without annotations for each image patch. The sample images are obtained from breast tissue biopsy slides stained with [H&E](#), a standard staining technique used in histopathology for highlighting cellular structures. The images are captured in RGB TrueColor with a 24-bit color depth (8 bits per channel) using magnifying factors of 40 $\times$, 100 $\times$, 200 $\times$, and 400 $\times$. Each image in the dataset is stored in an uncompressed graphics format to preserve high-quality histopathological details, with a fixed resolution of 700 $\times$ 460 pixels. Figure 4.1 demonstrates sample examples of BreakHis dataset with various magnifying factors. The dataset presents a class imbalance, with a greater number of malignant samples than benign ones, posing a challenge for machine learning models.

4.1.2 PatchGastric Dataset

The PatchGastric dataset [2] comprises pairs consisting of patches from a stomach adenocarcinoma endoscopic biopsy sample and a corresponding histopathological caption. The patches are extracted from whole slide images and aligned with the captions extracted from the respective diagnostic reports. The dataset consists of a total of 262,777 patches that are derived from 991 [H&E](#)-stained slides. Each slide, at a magnification of 20 times, originates from a unique patient. In the patches, each pixel is the dimension of 300 $\times$ 300. The vocabulary of the captions comprises 344 unique words, with the length of each sentence length at a maximum of 47 words. Figure 4.2 provides a visual representation of the instances of patches retrieved from [WSIs](#) and

their corresponding text captions.



Tumor tissue consisting of medium-sized and irregular ducts is observed infiltrating in the superficial epithelium. Tumor cells are highly columnar, with nuclei aligned basolaterally and polarized. Well differentiated tubular adenocarcinoma.

Figure 4.2: Example of biopsy samples with associated captions and 300×300 pixel patches extracted at 20 $\times$ magnification [2].

4.2 Performance Evaluation

To demonstrate the effectiveness and feasibility of [SOTA](#) approaches alongside the proposed TransUAAE-CapGen and GNN-ViTcap architectures, we utilize two distinct sets of evaluation metrics: WSI image classification metrics and [WSI](#) captioning metrics. These metrics provide a comprehensive evaluation of the overall performance capabilities of the TransUAAE-CapGen and GNN-ViTcap architectures. Here, TransUAAE-CapGen is utilized for caption generation from [WSIs](#), while GNN-ViTcap is employed for both classification and caption generation tasks.

4.2.1 Classification Metrics

For [WSI](#) classification, the evaluation metrics include Area Under the Curve (AUC), Precision (Pr), Recall (Rc), and F_1 -Score (F_1). Here, TP , FP , TN , and FN represent True Positive,

False Positive, True Negative, and False Negative, respectively. The four evaluation metrics are defined as follows [77]:

- **Precision:** Precision quantifies the accuracy of positive predictions. It is defined as the ratio of true positive outcomes to the total predicted positive outcomes.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (4.1)$$

- **Recall:** Recall, also known as Sensitivity, measures the proportion of actual positives that are correctly identified.

$$\text{Recall} = \frac{TP}{TP + FN} \quad (4.2)$$

- **F_1 -Score:** F_1 -Score is the harmonic mean of Precision and Recall. This score is crucial for imbalanced datasets as it balances the contributions of precision and recall, providing a more robust evaluation of model performance.

$$F_1\text{-Score} = 2 \cdot \frac{Pr \times Re}{Pr + Re} \quad (4.3)$$

- **AUC:** AUC quantifies how well the model differentiates between positive and negative cases, where higher values reflect stronger discriminative performance. This curve plots the true positive rate (TPR) against the false positive rate (1 - Specificity) at various threshold settings.

$$\text{AUC} = \int_0^1 \text{TPR}(t) dt \quad (4.4)$$

AUC and F_1 -Score are the primary evaluation metrics used to assess the performance of the proposed GNN-ViTCap architecture, considering the significant imbalance in microscopic WSI datasets.

4.2.2 Image Captioning Metrics

For WSI captioning, our proposed methods are evaluated using BLEU [78], METEOR [79], ROUGE [80], and CIDEr [81]. The score ranges for BLEU, METEOR, and ROUGE-L span from 0 to 1 (1 – 100%), while CIDEr ranges from 0 to 10. In all four metrics, higher values indicate better quality in the generated captions. The details of each evaluation metric are defined as follows:

- **BLEU: Bilingual Evaluation Understudy (BLEU)** [78] is a widely used metric for assessing text quality, primarily in machine translation. It compares captions based on modified n -gram precision to generate BLEU scores. The precision for each n -gram size (unigrams, bigrams, trigrams, etc.) $P(n)$ is computed as:

$$P(n) = \frac{Matched(n)}{H(n)} \quad (4.5)$$

where $Matched(n)$ represents the count of n -gram tuples in the hypothesis as well as present in the actual caption, while $H(n)$ denotes the count of n -gram tuples in the hypothesis. The brevity penalty ρ is computed as follows:

$$\rho = \exp \left(\min \left(0, 1 - \frac{n}{L_{cap}} \right) \right) \quad (4.6)$$

where n represents the length of the hypothesis and L_{cap} denotes the length of the actual caption. The BLEU score is then calculated as follows:

$$BLEU = \rho \left(\prod_{i=1}^N P(i) \right)^{\frac{1}{N}} \quad (4.7)$$

where N is the maximum order of n -grams. In this case, the value of n is assigned as 1, 2, 3, and 4, corresponding to BLEU-1, BLEU-2, BLEU-3, and BLEU@4, respectively.

- **METEOR: Metric for Evaluation of Translation with Explicit Ordering (METEOR)** [79] relies on unigram alignment between the generated and actual captions. It uses the harmonic mean of unigram precision and recall, with a higher weighting assigned to recall over precision. The calculation of unigram precision P_e and recall R_e in METEOR is expressed as follows:

$$P_e = \frac{w_m}{w_t}, R_e = \frac{w_m}{w_r} \quad (4.8)$$

where w_m is the number of unigrams in the candidate translation as well as present in the actual caption, w_t and w_r are the number of unigrams in the candidate and the actual translation, respectively. The harmonic mean of precision and recall F_μ is computed as:

$$F_\mu = \frac{10P_eR_e}{R_e + 9P_e} \quad (4.9)$$

The penalty ρ is calculated using:

$$\rho = 0.5 \left(\frac{C_k}{u_m} \right)^3 \quad (4.10)$$

where, C_k represents the number of chunks, and u_m signifies the count of mapped unigrams. Finally, the **METEOR** score M for a segment is calculated as:

$$METEOR = F_\mu(1 - \rho) \quad (4.11)$$

- **ROUGE**: **Recall-Oriented Understudy for Gisting Evaluation (ROUGE)** [80] measures the frequency of n -gram matches in the total of manually annotated descriptions, while other methods account for the frequency in the aggregation of all generated descriptions. ROUGE-N and ROUGE-L are often used in tasks like document summarization and video captioning. The ROUGE-N score can be expressed as follows:

$$ROUGE - N = \frac{\sum_{S \in AS} \sum_{g_n \in S} MatchedCount(g_n)}{\sum_{S \in AS} \sum_{g_n \in S} TotalCount(g_n)} \quad (4.12)$$

where $MatchedCount(g_n)$ represents the maximum number of n -grams that appear both in a candidate summary and a collection of actual summaries. On the other hand, $TotalCount(g_n)$ is the total number of n -grams present in the actual summaries.

- **CIDEr**: **Consensus-based Image Description Evaluation (CIDEr)** [81] metric for n -grams of length n is calculated using the mean cosine similarity between the candidate and actual captions. It can be computed as:

$$CIDEr_n(c_j, S_j) = \frac{1}{n} \sum_k \frac{h_n(c_j) \cdot h_n(s_{jk})}{||h_n(c_j)|| \cdot ||h_n(s_{jk})||} \quad (4.13)$$

where $h_n(c_j)$ is a vector of all n -grams of length n in the candidate caption, and $||h_n(c_j)||$ is its magnitude. The same applies to $h_n(s_{jk})$ for the actual caption. The overall **CIDEr** score combines the scores from n -grams of different lengths:

$$CIDEr(c_j, S_j) = \sum_{n=1}^N w_n CIDEr_n(c_j, S_j) \quad (4.14)$$

where uniform weights $w_n = 1/N$ have been found to work best empirically. c_j is the candidate captions, S_j is the set of actual captions, and s_{jk} is the k -th actual caption in S_j .

4.3 Experimental Analysis: TransUAAE-CapGen Architecture

In this section, we present the experimental analysis of the TransUAAE-CapGen architecture using the PatchGastric dataset.

4.3.1 Experimental Setup

The experiments were conducted on an Apple MacBook Pro equipped with the Apple M3 Pro chip. The system specifications are: 18-core GPU, 18 GB of RAM, and a 500 GB SSD, running macOS. The models were implemented using the PyTorch [82] framework.

4.3.2 Training Parameters

The dataset is partitioned in a ratio of 7(688) : 1(100) : 2(203) for model training, validation, and testing. For the feature extraction, initially, we used two pretrained models on ImageNet (EfficientNetB3 [83] and DenseNet121 [84]) to extract features from the penultimate layer. Therefore, we implemented a hybrid UNet-based AAE model for feature extraction. Finally, our proposed TransUAAE-CapGen model is trained with 50 epochs with a learning rate = 0.001 and a weight decay = epoch/length(train). For evaluation, the model weights from the final epoch are employed with a batch size configured to 16. Figure 4.3 shows the loss curve of our TransUAAE-CapGen architecture along with pre-trained model on the training set of PatchGastric dataset.

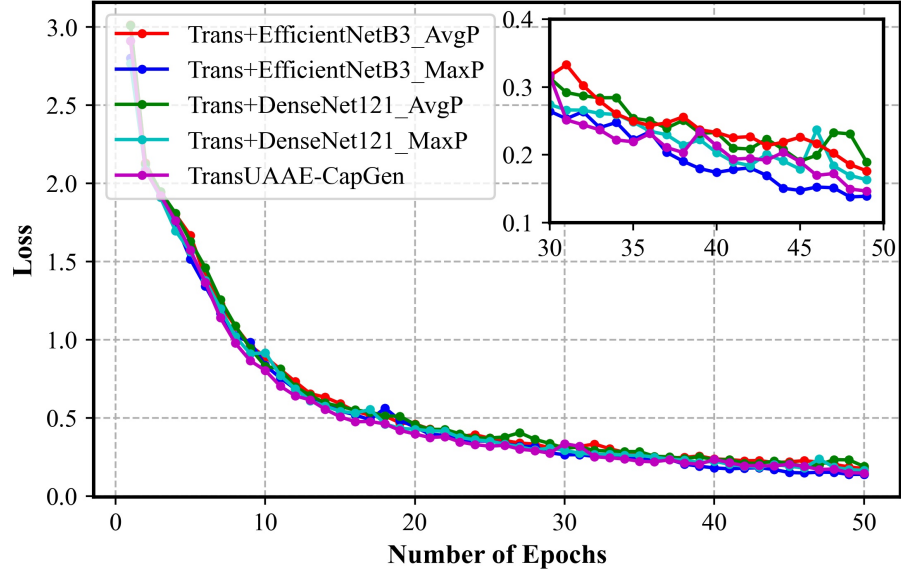
4.3.3 Caption Generation Performance

In our study, we introduced a TransUAAE-CapGen architecture comprise of transformer and UNet-based AAE to generate captions from a set of WSI patches instead of depending solely on a single patch. In the initial stage, we applied a 3×3 pooling (either average or max) and global average pooling to the EfficientNetB3 and DenseNet121 models. Therefore, the extracted features are fed into the transformer to generate the histopathological caption. For the EfficientNetB3 model, the extracted features output for a single WSI is $P \times 9 \times 9 \times 1536$ for max pooling, where P number of patches from a specific WSI. Therefore, the subsequent global average pooling results is an output of $P \times 1 \times 1 \times 1536$. In the case of average pooling, the extracted features for a single WSI is $P \times 4 \times 4 \times 1536$, and subsequently, the global average pooling generates an output of $P \times 1 \times 1 \times 1536$.

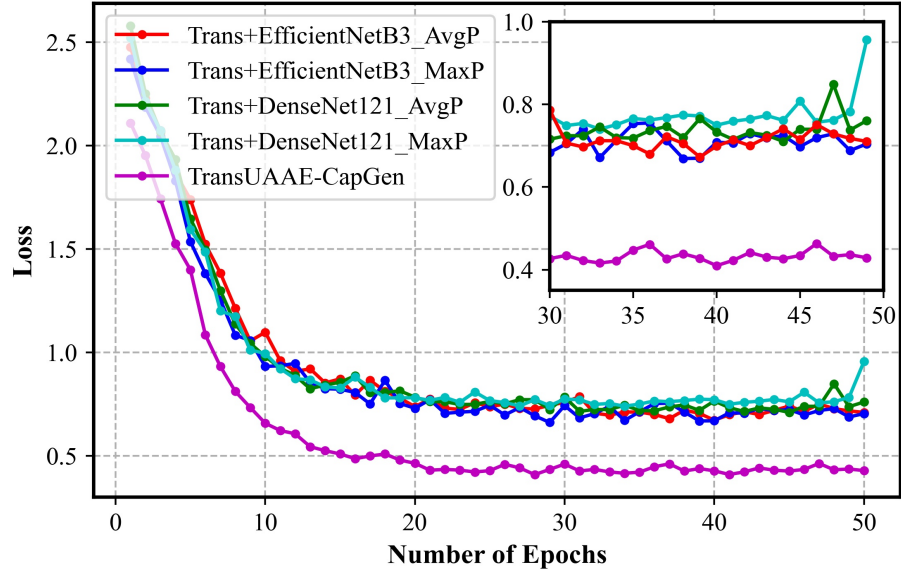
On the other hand, with the DenseNet121 model, the extracted feature output for a single WSI is $P \times 9 \times 9 \times 1024$ for max pooling and $P \times 4 \times 4 \times 1024$ for average pooling, respectively. Therefore, the global average pooling result is an output of $P \times 1 \times 1 \times 1024$. As the number of patches of WSI varies from patient to patient, we applied padding to keep all the WSI the same number of patches. The final features dimension of the EfficientNetB3 model for the transformer encoder is (1500×1536) , and DenseNet121 is (1500×1024) , respectively. Finally, we deployed our hybrid UNet-based AAE feature extractor to extract more robust and informative features. In

Table 4.1: Performance metric (%) of our proposed methods for caption generation on test set of PatchGastric dataset.

Model	Feature Dimension	BLEU-1	BLEU-2	BLEU-3	BLEU@4	METEOR	ROUGE	CIDEr
Trans+EfficientNetB3_AvgP	$P \times 9 \times 9 \times 1536$	84.5	79.7	76.4	73.9	51.8	81.7	6.50
Trans+EfficientNetB3_MaxP	$P \times 4 \times 4 \times 1536$	88.3	85.1	82.9	81.2	57.2	87.4	7.34
Trans+DenseNet121_AvgP	$P \times 9 \times 9 \times 1024$	86.8	82.6	79.6	77.2	54.2	85.7	6.81
Trans+DenseNet121_MaxP	$P \times 4 \times 4 \times 1024$	88.4	84.8	82.5	80.7	56.2	87.5	7.45
Our(TransUAAE-CapGen)	$P \times 18 \times 18 \times 512$	92.1	90.3	88.3	86.8	59.6	89.3	7.72



(a)



(b)

Figure 4.3: Loss curve representation on the PatchGastric dataset using proposed TransUAAE-CapGen model along with four baseline architectures. (a) Training loss (b) Validation loss.

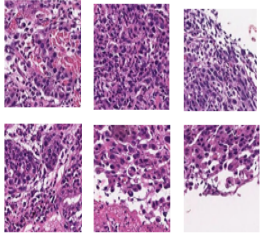
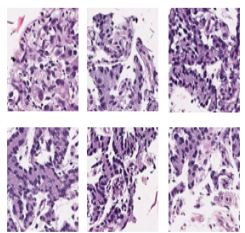
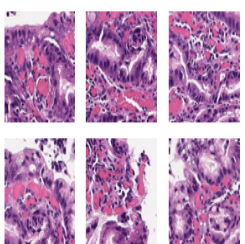
the UNet-based AAE model, when a WSI is divided into P patches, the feature extraction step produces output dimensions of $P \times 18 \times 18 \times 512$ after 2×2 max pooling and global average pooling then reduces these dimensions to $P \times 1 \times 1 \times 512$. The final feature dimension of the UNet-based AAE model for the transformer encoder is (1500×512) . Therefore, we feed the extracted features to the transformer to generate the caption.

Our experiments are conducted on the PatchGastric dataset to evaluate the performance of our proposed TransUAAE-CapGen architecture and compare it with other existing methods. Table 4.1 demonstrates the result of our proposed model on the test dataset for caption generation. As shown in Table 4.1, the results indicate that our proposed TransUAAE-CapGen model significantly outperforms the other models in all metrics. The estimated accuracies of the TransUAAE-CapGen model are BLEU@4 = 86.8%, METEOR = 59.6%, ROUGE = 89.3% and CIDEr = 7.72%, respectively. The Trans+EfficientNetB3_AvgP, Trans+EfficientNetB3_MaxP, Trans+DenseNet121_AvgP, and Trans+DenseNet121_MaxP models are obtained the BLEU@4 scores of 73.9%, 81.2%, 77.2%, and 80.7%, respectively. These outcomes substantiate the efficiency of our proposed TransUAAE-CapGen method for image captioning on the PatchGastric dataset. Table 4.1 shows qualitative examples of the PatchGastricADC22 dataset. We compared generated captions using our proposed TransUAAE-CapGen model and reference captions.

4.3.4 Comparison with State-of-the-Art

In this section, we present a comparison of our proposed TransUAAE-CapGen method with existing SOTA methods using the PatchGastric dataset. Looking at individual-related studies in Table 4.3, in [2], the authors used two pretrained ImageNet models to extract the features and applied the LSTM model to generate the caption, obtaining the highest BLUE@4 score of 32.3%. In [23], the author introduced a multi-modal, multi-task and multiple instance learning (PathM3) model to classify and generate the caption from WSI and obtained a BLUE@4 score of 52.0%. In another study, Qin et al. [57] developed a SGMT model using the transformer and achieved a BLUE@4 score of 55.11%. However, our method employs a UNet-based AAE for extracting more meaningful spatial features from images, which are then fed into a transformer for caption generation. This approach has yielded significant improvements in performance metrics, as shown in Table 4.3. TransUAAE-CapGen obtained the BLEU@4 score of 86.8%, METEOR score of 59.6%, ROUGE score of 89.3%, and CIDEr score of 7.72%. These results are significantly higher than LSTM-based models [2], PathM3 [23] and SGMT model [57], demonstrating the effectiveness of our approach. The superiority of our model can be attributed to the use of transformers over LSTM [2] and the hybrid UNet-based AAE feature extraction model. Transformers are shown to be more effective than LSTM for tasks that require understanding the context and dependencies in the data. On the other hand, the SGMT [57] model used a maximum patch size of 128 for

Table 4.2: Illustration of qualitative results using TransUAAE-CapGen model on test dataset.

WSI	TransUAAE-CapGen	Reference
	atypical epithelium cells with and proliferating atypical forming solid tubular observed is observed moderately differentiated adenocarcinoma solid	atypical epithelial cells infiltrating and proliferating while forming a glandular cavity is observed moderately differentiated tubular adenocarcinoma
	highly columnar with adenocarcinoma disordered papillary localization proliferates well a densely adenocarcinoma papillary differentiated adenocarcinoma	highly columnar epithelium with disordered nuclear localization proliferates showing fusion ductal construction differentiated tubular adenocarcinoma
	large differentiated tubular adenocarcinoma well differentiated adenocarcinoma solid type with ductal tissue the adenocarcinoma chromatin adenocarcinoma	poorly differentiated adenocarcinoma solid type with tumor tissue diffusely infiltrated with atypical epithelial cells

WSI, whereas our TransUAAE-CapGen used a maximum patch size of 1400, which reduces the chance of information loss. Furthermore, in SGMT [57], the authors used pretrained Convolutional Neural Networks (CNN) models to extract the features and applied the transformer model for caption generation. However, our proposed architecture explored the UNet-based AAE architecture, which is significant in extracting more meaningful spatial features from images. The UNet-based AAE for feature extraction combined with a transformer provides a robust and effective solution for caption generation.

Table 4.3: Comparison of our proposed method with existing methods using the PatchGastric dataset.

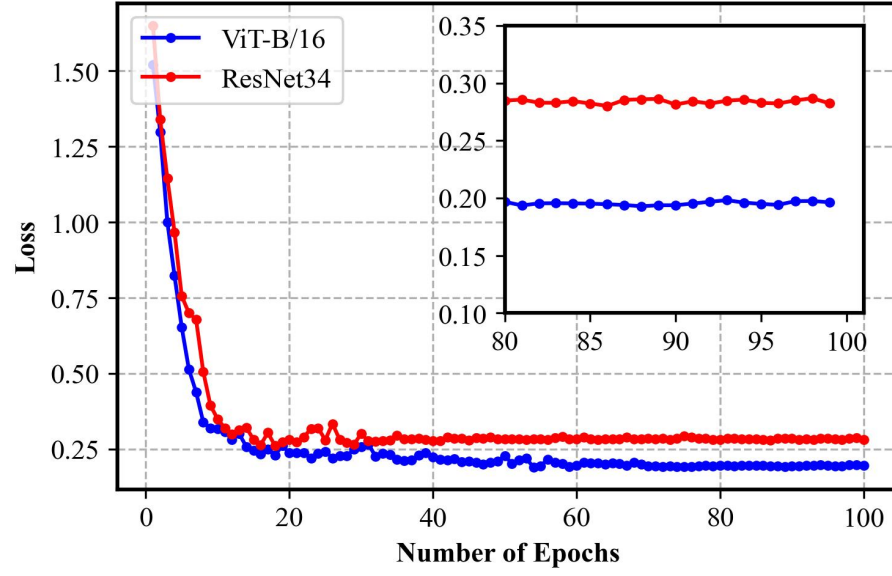
Model	B	M	R	C
LSTM+EfficientNetB3_AvgP [2]	28.3	30.54	49.73	2.31
LSTM+EfficientNetB3_MaxP [2]	32.4	27.75	46.49	1.62
LSTM+DenseNet121_AvgP [2]	27.2	28.21	46.95	1.52
LSTM+DenseNet121_MaxP [2]	32.3	26.85	45.47	1.48
PathM3 [23]	52.0	39.4	—	—
SGMT [57]	55.11	43.17	69.68	4.83
Our (TransUAAE-CapGen)	86.8	59.6	89.3	7.72

* B = BLEU@4(%), M = METEOR(%), R = ROUGE(%), C = CIDEr

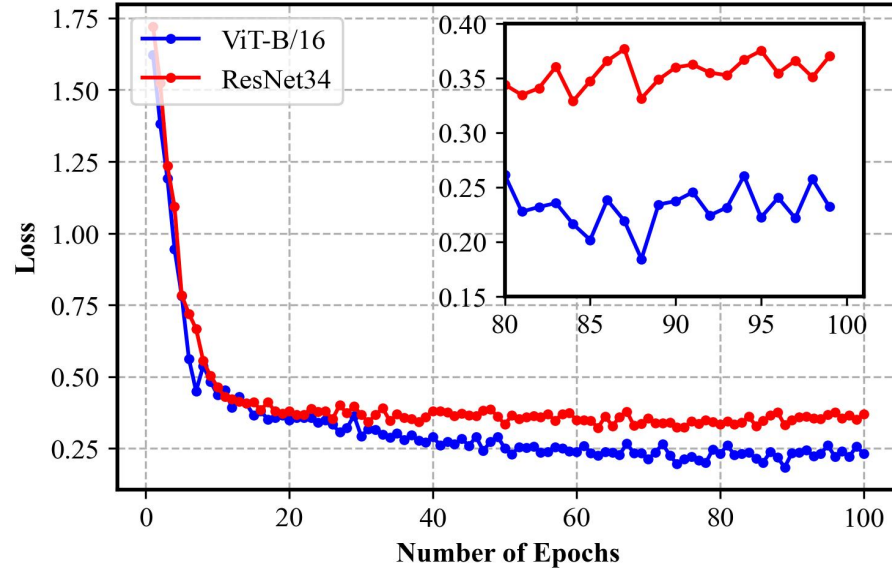
4.4 Experimental Analysis: GNN-ViTCap Architecture

4.4.1 Implementation Details and Training Phase

For feature extraction, an ImageNet-21k pre-trained ViT-B/16 [20] is used as the visual backbone, which has 12-layers of Vision Transformer that encodes images into embeddings of dimension $d_v = 768$. Images are processed at a resolution of 224×224 pixels with 16×16 patch size. Additionally, a ResNet-34 pre-trained on ImageNet [85] with 34 convolutional layers produces embeddings of dimension $d_v = 512$. For attention-based deep embedded clustering, the number of clusters is set to $\mathcal{K} = 8$ for the BreakHis dataset and $\mathcal{K} = 50$ for the PatchGastric dataset, with a convergence threshold of $\epsilon = 10^{-4}$. The hidden layer dimension of the graph neural network is $d_{\text{out}} = 512$ with GAT layer $L = 3$. A linear projection layer with a dimension of $d_{\text{model}} = 768$ maps image features into the large language model input space. The tokenizer functions of our chosen LLMs facilitated the conversion of text into a format suitable for model processing. All methods are trained using cross-entropy loss with a learning rate of 10^{-3} and a dropout rate of 0.3. The training is set to run for 100 epochs using the Adam optimizer [86], a weight decay of 10^{-2} , and a batch size of 16 for both the training and evaluation phases. The proposed architecture is implemented using PyTorch [82] and the Deep Graph Library (DGL) [87]. Experiments are performed on an NVIDIA RTX A6000 graphics card GPU with 48 GB of memory. Figure 4.4 and Figure 4.5 demonstrate the loss curves of the GNN-ViTCap architecture on the BreakHis and PatchGastric training sets, respectively

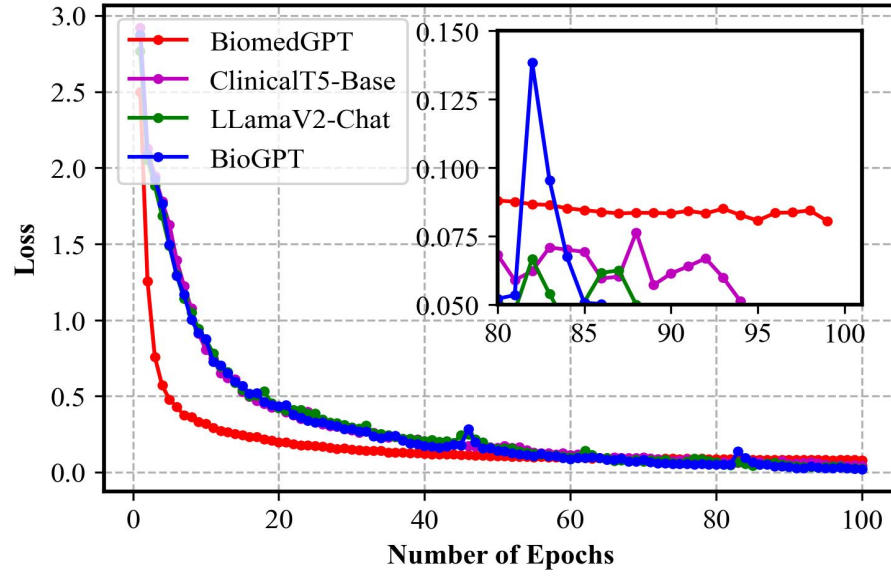


(a)

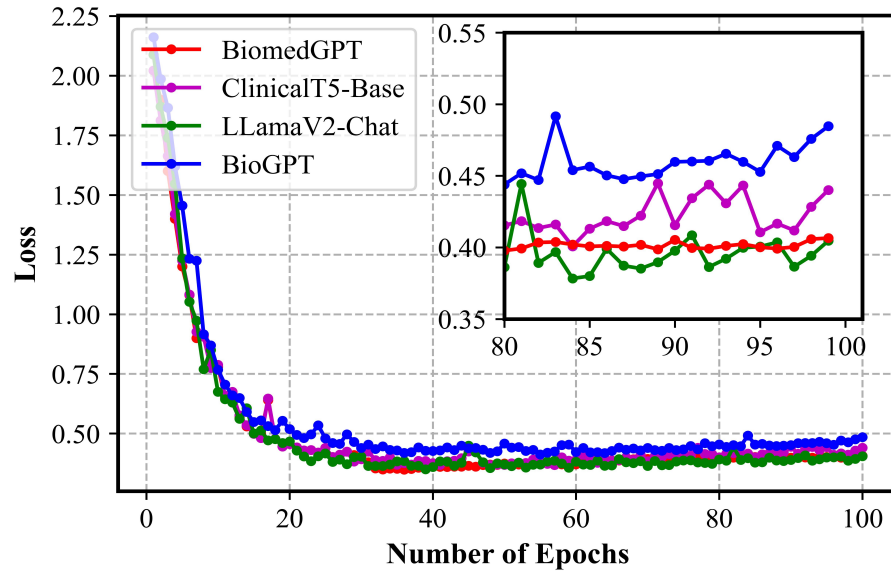


(b)

Figure 4.4: Loss curves on the BreakHis dataset using the proposed GNN-ViTCap with various feature extractors. (a) Training loss, (b) Validation loss.



(a)



(b)

Figure 4.5: Loss curves on the PatchGastric dataset using the proposed GNN-ViTCap with various LLMs. (a) Training loss, (b) Validation loss.

4.5 Results and Discussions: GNN-ViTCap Architecture

The main objective of this experiment is to explore the following questions:

- Q_1 : *Does the proposed GNN-MIL perform better than SOTA MIL methods for microscopic WSI classification?*
- Q_2 : *Does the spatial positional information of patches impact the performance of model for caption generation?*
- Q_3 : *Do LLMs perform better than LSTM or traditional transformer models for image captioning of WSI?*
- Q_4 : *Do in-domain LLMs perform better than generalized LLMs for generating captions in histopathological image analysis?*

4.5.1 Image Classification Results

Table 4.4 demonstrates the performance of the proposed GNN-ViTCap and other SOTA methods for benign and malignant tumor classification on the BreakHis test dataset. In this experiment, the visual features of microscopic images are extracted using either a pretrained ResNet-34 or ViT-B/16 model. The SOTA MIL-based methods are conducted with the same configuration of the BreakHis test dataset to ensure a fair comparison. Overall, the proposed GNN-ViTCap ($ViT + DEC + GNN - MIL$) architecture achieved the best classification performance across all metrics: F_1 -Score of 0.934, Precision of 0.926, Recall of 0.942, and AUC of 0.963. The findings demonstrate an improvement of 2.3% in the F_1 -Score and 5.7% in AUC over other SOTA MIL-based methods for classification, as shown in Table 4.4. The SOTA ABMIL [6] method achieves an F_1 -Score of 0.900 and an AUC of 0.871, while TransMIL [7] records an F_1 -Score of 0.886 and an AUC of 0.862. Similar performance gaps are observed when comparing with DSMIL [36] (with an F_1 -Score of 0.856 and an AUC of 0.869) and DTFD-MIL [37] (with an F_1 -Score of 0.911 and an AUC of 0.887).

The SOTA MIL-based methods, including ABMIL, DSMIL, TransMIL, and DTFD-MIL, were designed for scanner-based whole slide images where the absolute positions of the patches are unknown. Therefore, the existing MIL-based methods show lower performance compared to the proposed GNN-ViTCap ($ViT + DEC + GNN - MIL$) when using the microscopic images dataset. The reason is that the proposed GNN-ViTCap removes redundant images or regions using a deeply embedded clustering method and selects the most representative image patches based

on attention score. Moreover, the GNN-based MIL identifies neighboring representative images based on their relative positions and aggregates the features for classification.

Figure 4.6 depicts the feature distribution from the GNN-ViTCap method on the BreakHis test dataset using t-SNE visualization. Some Benign samples overlap with the Malignant category, leading to misclassifications. These misclassifications occur because the GNN-ViTCap method struggles to distinguish these categories but can still detect clusters.

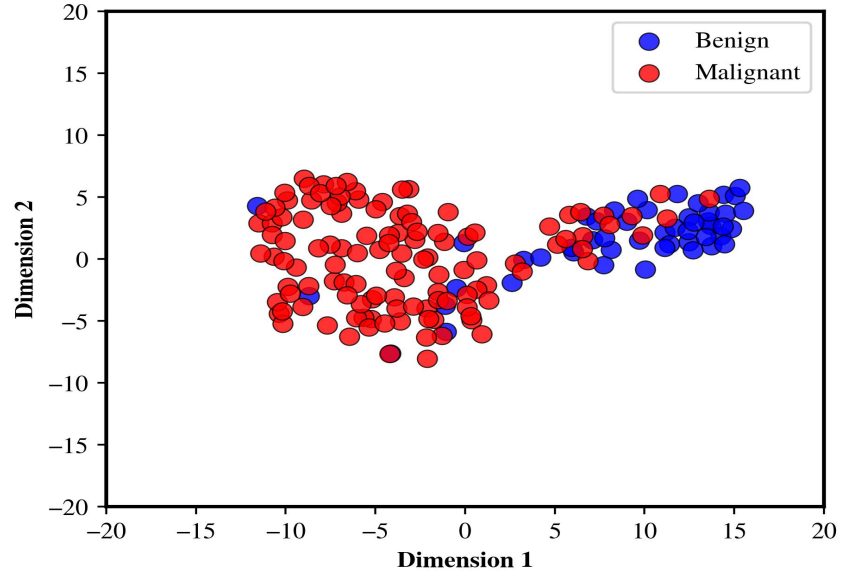
Table 4.4: Performance of GNN-ViTCap against SOTA methods on the BreakHis test dataset for classification.

Methods	Precision	Recall	F_1 -Score	AUC
ABMIL [6]	0.835	0.922	0.900	0.871
DSMIL [36]	0.872	0.842	0.856	0.869
TransMIL [7]	0.865	0.908	0.886	0.862
DTFD-MIL [37]	0.854	0.925	0.911	0.887
GNN-ViTCap (ResNet-34)	0.917	0.925	0.921	0.906
GNN-ViTCap (ViT-B/16)	0.926	0.942	0.934	0.963

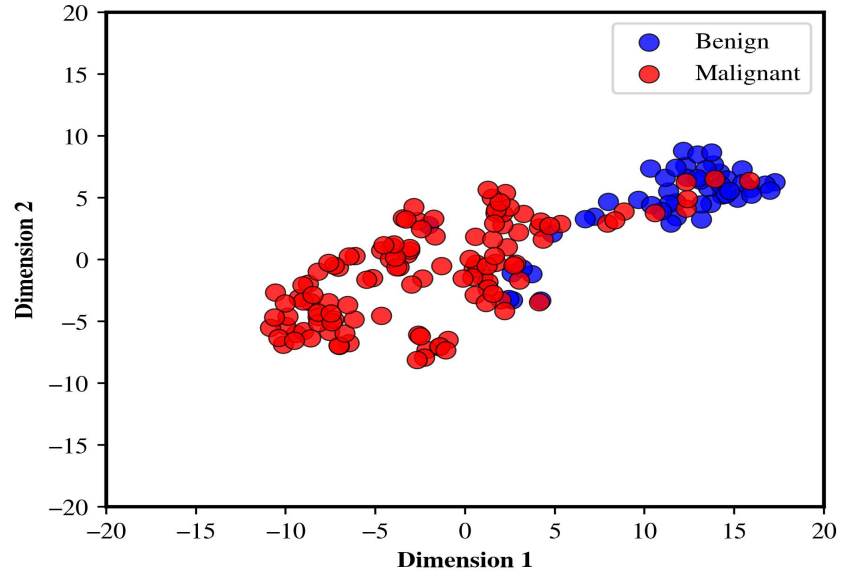
4.5.2 Image Captioning Results

Table 4.5 demonstrates the comparative analysis of the proposed GNN-ViTCap and other SOTA methods for image captioning on the PatchGastric test dataset. In the experiments, four large language models (ClinicalT5-Base, BioGPT, LLamaV2-Chat, and BiomedGPT) are fine-tuned using visual features along with corresponding captions. The proposed GNN-ViTCap ($ViT + DEC + GNN - MIL + BiomedGPT$) architecture achieved the highest scores across all metrics: a BLEU@4 score of 0.811, METEOR of 0.567, ROUGE of 0.865, and CIDEr of 7.42. The results indicate that our proposed GNN-ViTCap ($ViT + DEC + GNN - MIL + BiomedGPT$) method significantly outperforms the other SOTA methods in all metrics. The findings also demonstrate an improvement of 26% on BLEU@4 and 13.5% on METEOR over existing caption generation methods, as evident in Table 4.5. On the same dataset, the second-best BLEU@4 score of 0.796 and METEOR of 0.557 are obtained from the GNN-ViTCap ($ViT + DEC + GNN - MIL + LLamaV2 - Chat$) architecture.

The SOTA approach, PatchCap [2] obtained the highest BLUE@4 score of 0.324 using EfficientNetB3 and LSTM models. The other SOTA approaches, PathM3 [23] and SGMT [57]



(a)



(b)

Figure 4.6: t-SNE feature visualizations for the GNN-ViTCap for the BreakHis test dataset. (a) ResNet-34+DEC+GNN-MIL, (b) ViT+DEC+GNN-MIL.

achieved BLEU@4 of 0.520 and 0.551, respectively, using transformer-based language models. The superior performance of the GNN-ViTCap architecture is attributed to its effective feature extraction using a visual encoder and the selection of the most significant visual features through deep embedded clustering. Moreover, our proposed architecture explored the [GNN-MIL](#), which aggregates features with patch positional encoding, and the integration with in-domain large language models provides a robust and effective solution for caption generation.

Furthermore, in Figure 4.7, a bar chart is utilized to illustrate the BLEU@4, METEOR, and ROUGE scores achieved by each [LLMs](#) for caption generation. This visual representation allows a better comparison of performance across the three metrics, showing how each model handles both n-gram overlaps and semantic aspects in the generated text. BiomedGPT shows the highest scores on all three metrics, underscoring its robust generation capabilities and consistency. Meanwhile, ClinicalT5-Base, BioGPT, and LLamaV2-Chat, demonstrate competitive performance, though slightly lower than BiomedGPT overall. In addition, Table 4.6 presents qualitative examples of the PatchGastric dataset by comparing generated captions from the proposed GNN-ViTCap method with reference captions.

4.5.3 Discussions

In this work, a novel GNN-ViTCap architecture is proposed for cancer tumor classification and caption generation from microscopic whole slide images. GNN-ViTCap architecture comprises a visual feature extractor, attention-based deep embedded cluster, GNN-based MIL, and large language models. Experimental results on both BreakHis and PatchGastric datasets demonstrate the effectiveness of the proposed GNN-ViTCap.

Q₁: Does the proposed GNN-MIL perform better than SOTA MIL methods for microscopic WSI classification?

The [SOTA MIL](#)-based methods, including [ABMIL](#) [6], [DSMIL](#) [36], [TransMIL](#) [7], and [DTFD-MIL](#) [37], were designed exclusively for scanner-based WSIs, whereas spatial information is absent in microscopic [WSIs](#). However, the spatial information of patches in whole slide images is crucial for cancer diagnosis. Therefore, the proposed GNN-ViTCap (*ViT/ResNet* – 34 + *DEC* + *GNN* – *MIL*) learns spatial information from graph data. In GNN-ViTCap, each graph node represents a [WSI](#) patch, and edges are determined by the embedded features of these patches, capturing the spatial relationships between different regions. As a result, GNN-ViTCap efficiently captures both spatial information and patch correlations, leading to superior feature representations. Moreover, the GNN-ViTCap architecture removes redundant images or regions using a deep embedded clustering method and selects the most representative image patches based

Table 4.5: Performance metrics of the proposed GNN-ViTCap against SOTA methods for caption generation on PatchGastric test dataset.

Methods	Visual Encoder	Language Model	BLEU-1	BLEU-2	BLEU-3	BLEU@4	METEOR	ROUGE	CIDEr
PatchCap [2]	EfficientNetB3_AvgP	LSTM	-	-	-	0.283	0.305	0.497	2.31
	EfficientNetB3_MaxP		-	-	-	0.324	0.277	0.465	1.62
	DenseNet121_AvgP		-	-	-	0.272	0.282	0.469	1.52
	DenseNet121_MaxP		-	-	-	0.323	0.268	0.454	1.48
PathM3 [23]	ViT-g/14	Flan-T5	-	-	-	0.520	0.394	-	-
SGMT [57]	CNN	Transformer	-	-	-	0.551	0.432	0.697	4.83
		BioGPT	0.802	0.748	0.713	0.686	0.485	0.766	5.72
GNN-ViTCap	ViT-B/16	ClinicalT5-Base	0.851	0.804	0.774	0.753	0.526	0.826	6.72
		LLamaV2-Chat	0.877	0.838	0.813	0.796	0.557	0.856	7.25
		BiomedGPT	0.886	0.851	0.828	0.811	0.567	0.865	7.42

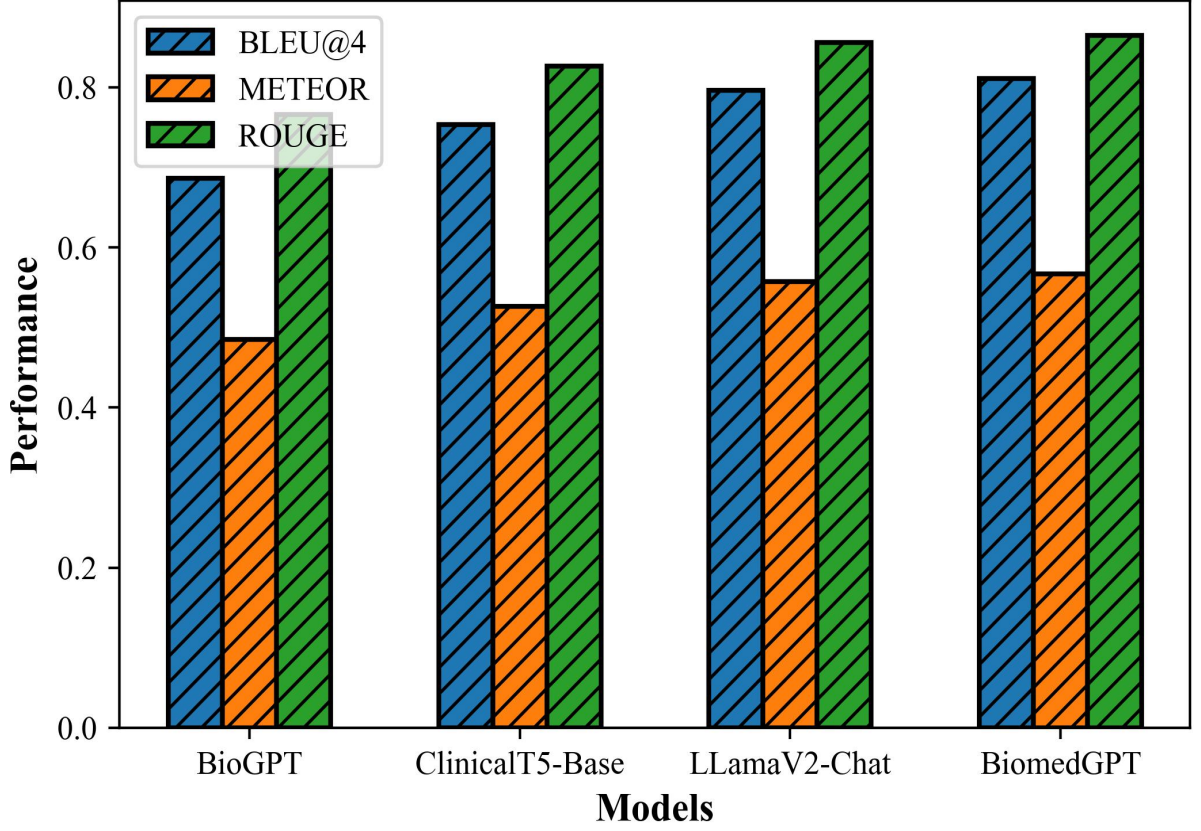


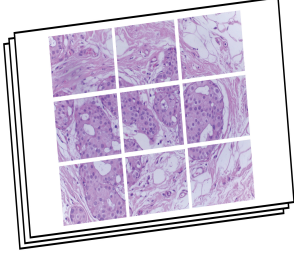
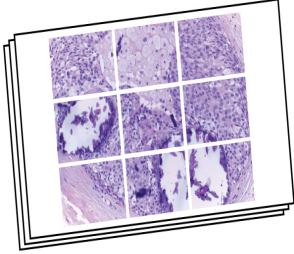
Figure 4.7: Performance of the proposed GNN-ViTCap architecture on the PatchGastric test dataset using various LLMs

on attention scores. Therefore, the proposed GNN-ViTCap method outperforms the [SOTA MIL](#)-based methods for classification.

Q₂: Does the spatial positional information of patches impact the performance of the model for caption generation?

The proposed GNN-ViTCap (*ViT + DEC + GNN – MIL + LLMs*) method achieved better results than the [SGMT](#) [57], method for caption generation using the PatchGastric dataset. Both GNN-ViTCap and [SGMT](#) utilize transformer-based language models, but [SGMT](#) achieved a lower BLEU@4 score due to lack of consideration for patch positional encoding. The graph-based aggregation ([GNN-MIL](#) within GNN-ViTCap leverages the spatial relationships between image patches, enabling the model to generate more accurate and relevant captions.

Table 4.6: Illustration of qualitative results using the GNN-ViTCap architecture on the PatchGas-tric test dataset.

Microscopic WSI	Ground Truth	GNN-ViTCap
	tumor tissue, in which medium to small irregular ducts infiltrate and proliferate in the submucosa can be seen in the epithelium well differentiated tubular adenocarcinoma	tumor tissue in which medium sized small irregular ducts with or proliferate in the sub-mucosa can be irregular in the epithelium well differentiated tubular adenocarcinoma
	in the superficial epithe-lium tumor tissue that invades by forming medium sized to small irregular ducts is observed moderately differentiated adeno-carcinoma	on the superficial epithe-lium tumor tissue that infil-trates by forming medium sized to small irregular ducts is observed moder-ately differentiated tubular

Q₃: Do LLMs perform better than LSTM or traditional transformer models for image captioning of WSI?

Furthermore, the proposed GNN-ViTCap (*ViT + DEC + GNN – MIL + LLMs*) outperforms *LSTM* [2] and traditional transformer models [57] in image captioning of whole slide images. Large language models have superior contextual understanding and efficient integration of multimodal data, which enable them to generate more accurate and coherent captions. The integration of advanced *LLMs* within the GNN-ViTCap framework facilitates more coherent and contextually accurate captioning.

Q₄: Do in-domain LLMs perform better than generalized LLMs for generating captions in histopathological image analysis?

Moreover, PathM3 [23] introduced a multi-modal, multi-task, and multiple instance learning model for caption generation using ViT feature extraction and distilled versions of *LLMs* models (Flan-T5). However, PathM3 obtained a lower BLEU@4 score compared to the proposed GNN-ViTCap (*ViT + DEC + GNN – MIL + BimedGPT*) method, which utilizes in-domain *LLMs*.

Fine-tuning language models on domain-specific data enables them to produce more relevant and precise descriptions, tailored to the nuances of medical imaging.

4.5.4 Ablation Studies

This section presents the results of ablation studies conducted to evaluate the effectiveness of each module in the GNN-ViTCap architecture.

With vs. Without Representative Images

Table 4.7 indicates that excluding representative images leads to decreased performance compared to including them, as demonstrated by both the BreakHis and PatchGastric datasets. This performance difference arises because the model cannot effectively capture contextual relationships across diverse areas. By selecting a distinctive image for every region, the model is better equipped to capture detailed contextual information, thereby enhancing interactions across different image areas and improving overall performance. The GNN-ViTCap method achieves an F_1 -Score of 0.934 and an AUC of 0.963 with representative images, compared to an F_1 -Score of 0.902 and an AUC of 0.916 without representative images, for classification using the BreakHis microscopic dataset. Therefore, the GNN-ViTCap ($ViT + DEC + GNN - MIL$) method demonstrates an increase of 3.2% in F_1 -Score and 4.7% in AUC with representative images on the BreakHis dataset. In contrast, the GNN-ViTCap ($ViT + DEC + GNN - MIL + BiomedGPT$) method illustrates an increase of 4.8% in BLEU@4 score and 3.1% in METEOR with representative images for caption generation using PatchGastric microscopic dataset.

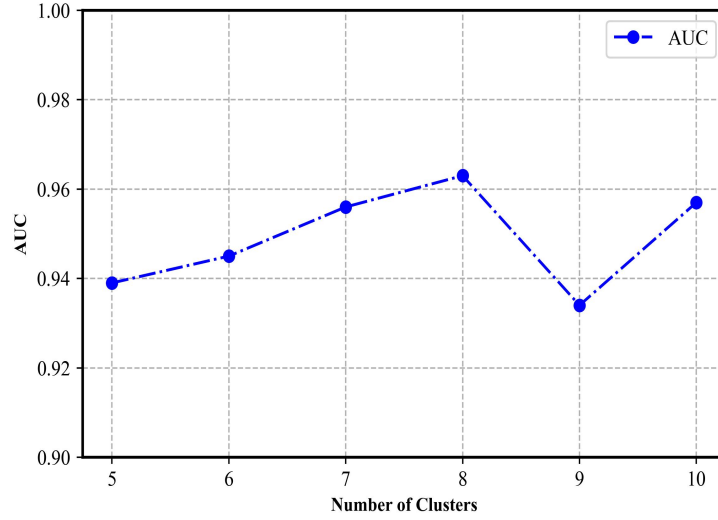
Table 4.7: Performance evaluation *with* ($w/$) and *without* (w/o) representative images on the BreakHis and PatchGastric datasets.

	BreakHis		PatchGastric	
Cluster	F_1 -Score	AUC	BLEU@4	METEOR
w/o Rep. Images	0.902	0.916	0.763	0.536
w/ Rep. Images	0.934	0.963	0.811	0.567

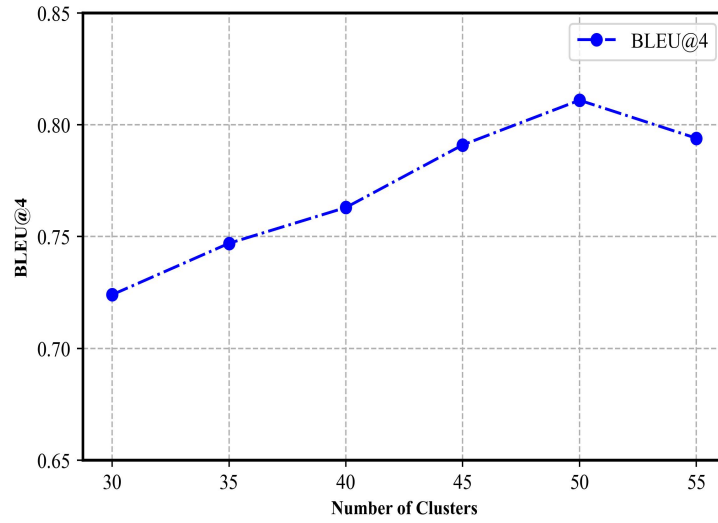
Representative Image Clusters

Figure 4.8 illustrates the impact of varying the number of representative image clusters on both datasets. For the BreakHis dataset, the number of clusters is varied between $\mathcal{K} = 5$ and 10, with

the highest AUC of 0.963 achieved when $\mathcal{K} = 8$. In contrast, for the PatchGastric dataset, the number of clusters is set to $\mathcal{K} = 30$ to 55, with the highest BLEU@4 score of 0.811 observed when $\mathcal{K} = 50$.



(a)



(b)

Figure 4.8: Performance metrics with various numbers of representative image clusters using attention-based deep embedded clustering method. (a) BreakHis, (b) PatchGastric datasets.

Graph Layers Effect

Table 4.8 demonstrates the impact of graph edge generation using the *similarity* method with varying numbers of graph layers. For the BreakHis dataset, the AUC increases from 0.943 at 1 layer to a maximum of 0.963 at 3 layers. In contrast, for the PatchGastric dataset, the BLEU@4 score starts at 0.753 at 1 layer, increases to a maximum of 0.811 at 3 layers, and slightly decreases to 0.792 at 4 layers. These results suggest that while performance improves with the addition of layers, the effect diminishes beyond a certain point.

Table 4.8: Comparison of AUC (BreakHis) and BLEU@4 (PatchGastric) performances across various graph layers in graph neural networks. Minimal layers enhance performance by focusing on nearby images and leveraging *similarity*.

	No. of Graph Layers			
Dataset	1	2	3	4
BreakHis	0.943	0.945	0.963	0.952
PatchGastric	0.753	0.784	0.811	0.792

4.6 Summary

In this chapter, we first introduced the datasets utilized to evaluate our proposed architectures in histopathological image classification and caption generation. We outlined the performance metrics used to assess our architectures, including classification metrics such as Precision, Recall, $F1$ -score, and AUC, as well as captioning metrics like BLEU, METEOR, ROUGE, and CIDEr. We reported the benchmarking performance of TransUAAE-CapGen and GNN-ViTCap, comparing them against state-of-the-art methods. Our study also explored various architectural configurations and conducted ablation studies to investigate the impact of different design choices. The findings highlight the effectiveness of our proposed methods in capturing meaningful histopathological features and generating accurate textual descriptions. The results demonstrate the feasibility of our architectures for efficient and accurate histopathological image analysis in practical applications.

Chapter 5

Conclusion and Future Works

This chapter presents the concluding remarks, discusses the limitations of the study, and highlights potential directions for future research.

5.1 Conclusion

Histopathological [WSI](#) analysis is an advancing field with the potential to improve diagnostic accuracy and medical decision-making. In this thesis, at first, we proposed a TransUAAE-CapGen architecture to generate the caption from [WSI](#) patches. The TransUAAE-CapGen model is based on a hybrid UNet-based [AAE](#) for feature extraction and a transformer for caption generation. The hybrid UNet-based [AAE](#) model efficiently extracts key features and reduces information loss by learning hierarchical features. The transformer architecture is used for caption generation due to its superior handling of long-range dependencies and understanding of context. Our proposed method is validated using the PatchGastric histopathological dataset. Later, we propose a novel GNN-ViTCap architecture for classification and caption generation from microscopic whole slide image images. The GNN-ViTCap method is based on a visual feature extractor, attention-based deep embedded clustering, GNN-MIL aggregation, and [LLMs](#). The deep embedded clustering method dynamically clusters images to reduce redundancy, while self-attention extracts the most representative images. Graph-based aggregation (GNN-MIL) leverages the spatial relationships between image patches and captures the contextual information. Therefore, [LLMs](#) are used for caption generation due to their exceptional context-association capabilities. Our proposed method is validated using the BreakHis and PatchGastric datasets.

Experimental results demonstrate the method’s effectiveness in microscopic whole slide image classification and captioning, providing healthcare professionals with a valuable tool for better interpreting medical images.

5.2 Limitations and Future Works

Despite the promising results achieved by the proposed methods, certain limitations remain. The TransUAAE-CapGen method is sensitive to hyperparameter selection, which can lead to instability in training and potential model collapse. Similarly, the GNN-ViTCap model relies on a fixed clustering mechanism, where selecting the optimal K value is crucial—an improper choice can lead to information loss and reduced performance. Moreover, fine-tuning LLMs is computationally expensive, making real-world deployment resource intensive. The lack of diverse, high-quality WSI-text paired datasets also remains a challenge, affecting the generalizability of the models.

In the future, more advanced and efficient vision-language architectures can be explored to enhance model robustness and efficiency in pathological image analysis. The integration of ViTs into TransUAAE-CapGen can improve hierarchical feature extraction and enhance contextual understanding. For GNN-ViTCap, adaptive clustering techniques could be explored to dynamically determine optimal cluster numbers, minimizing information loss. Additionally, parameter-efficient fine-tuning strategies for LLMs, such as low-rank adaptation or adapter-based methods, can be employed to reduce computational costs while maintaining performance. Expanding high-quality datasets and validating models in real-world clinical environments could further strengthen their applicability. Lastly, improving model interpretability and explainability can be essential for fostering trust and adoption in medical practice.

References

- [1] Niful Islam, Khan Md Hasib, MF Mridha, Sultan Alfarhood, Mejdl Safran, and MK Bhuyan. Fusing global context with multiscale context for enhanced breast cancer classification. *Scientific Reports*, 14(1):27358, 2024.
- [2] Masayuki Tsuneki and Fahdi Kanavati. Inference of captions from histopathological patches. In *International Conference on Medical Imaging with Deep Learning*, pages 1235–1250. PMLR, 2022.
- [3] Zizhao Zhang, Pingjun Chen, Mason McGough, Fuyong Xing, Chunbao Wang, Marilyn Bui, Yuanpu Xie, Manish Sapkota, Lei Cui, Jasreman Dhillon, et al. Pathologist-level interpretable whole-slide cancer diagnosis with deep learning. *Nature Machine Intelligence*, 1(5):236–245, 2019.
- [4] Andrew H Song, Guillaume Jaume, Drew FK Williamson, Ming Y Lu, Anurag Vaidya, Tiffany R Miller, and Faisal Mahmood. Artificial intelligence for digital and computational pathology. *Nature Reviews Bioengineering*, 1(12):930–949, 2023.
- [5] Faruk Ahmed, Andrew Sellergren, Lin Yang, Shawn Xu, Boris Babenko, Abbi Ward, Niels Olson, Arash Mohtashamian, Yossi Matias, Greg S Corrado, et al. Pathalign: A vision-language model for whole slide images in histopathology. *arXiv:2406.19578*, 2024.
- [6] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018.
- [7] 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.
- [8] David Ahmedt-Aristizabal, Mohammad Ali Armin, Simon Denman, Clinton Fookes, and Lars Petersson. A survey on graph-based deep learning for computational histopathology. *Computerized Medical Imaging and Graphics*, 95:102027, 2022.

- [9] Samar Elbedwehy, Tamer Medhat, Taher Hamza, and Mohammed F Alrahmawy. Enhanced descriptive captioning model for histopathological patches. *Multimedia Tools and Applications*, 83(12):36645–36664, 2024.
- [10] Lingxiang Wu, Min Xu, Jinqiao Wang, and Stuart Perry. Recall what you see continually using gridlstm in image captioning. *IEEE Transactions on Multimedia*, 22(3):808–818, 2019.
- [11] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2016.
- [12] Djamila-Romaissa Beddiar, Mourad Ouassalah, and Tapio Seppänen. Automatic captioning for medical imaging (mic): a rapid review of literature. *Artificial Intelligence Review*, 56(5):4019–4076, 2023.
- [13] Yuxuan Sun, Hao Wu, Chenglu Zhu, Sunyi Zheng, Qizi Chen, Kai Zhang, Yunlong Zhang, Xiaoxiao Lan, Mengyue Zheng, Jingxiong Li, et al. Pathmmu: A massive multimodal expert-level benchmark for understanding and reasoning in pathology. *arXiv:2401.16355*, 2024.
- [14] Neofytos Dimitriou, Ognjen Arandjelović, and Peter D Caie. Deep learning for whole slide image analysis: an overview. *Frontiers in medicine*, 6:264, 2019.
- [15] Jocelyn Barker, Assaf Hoogi, Adrien Depeursinge, and Daniel L Rubin. Automated classification of brain tumor type in whole-slide digital pathology images using local representative tiles. *Medical Image Analysis*, 30:60–71, 2016.
- [16] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv:2302.13971*, 2023.
- [17] Qiuhaio Lu, Dejing Dou, and Thien Nguyen. Clinicalt5: A generative language model for clinical text. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5436–5443, 2022.
- [18] 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):bbac409, 2022.

- [19] Kai Zhang, Rong Zhou, Eashan Adhikarla, Zhiling Yan, Yixin Liu, Jun Yu, Zhengliang Liu, Xun Chen, Brian D Davison, Hui Ren, et al. A generalist vision–language foundation model for diverse biomedical tasks. *Nature Medicine*, pages 1–13, 2024.
- [20] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115:211–252, 2015.
- [21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
- [22] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.
- [23] Qifeng Zhou, Wenliang Zhong, Yuzhi Guo, Michael Xiao, Hehuan Ma, and Junzhou Huang. Pathm3: A multimodal multi-task multiple instance learning framework for whole slide image classification and captioning. *arXiv:2403.08967*, 2024.
- [24] Ming Y Lu, Drew FK 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, 2021.
- [25] Dalila Barbosa, Marcos Ferreira, Geraldo Braz Junior, Marta Salgado, and António Cunha. Multiple instance learning in medical images: A systematic review. *IEEE Access*, 2024.
- [26] Ruoyu Li, Jiawen Yao, Xinliang Zhu, Yeqing Li, and Junzhou Huang. Graph cnn for survival analysis on whole slide pathological images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 174–182. Springer, 2018.
- [27] Carolyn Hutter and Jean Claude Zenklusen. The cancer genome atlas: creating lasting value beyond its data. *Cell*, 173(2):283–285, 2018.
- [28] CR Taylor and Richard M Levenson. Quantification of immunohistochemistry—issues concerning methods, utility and semiquantitative assessment ii. *Histopathology*, 49(4):411–424, 2006.

- [29] Gianpaolo Bontempo, Angelo Porrello, Federico Bolelli, Simone Calderara, and Elisa Ficarra. Das-mil: distilling across scales for mil classification of histological wsis. In *International conference on medical image computing and computer-assisted intervention*, pages 248–258. Springer, 2023.
- [30] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes Van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM Van Der Laak, Meyke Hermsen, Quirine F Manson, Maschenka Balkenhol, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, 318(22):2199–2210, 2017.
- [31] Gianpaolo Bontempo, Nicola Bartolini, Marta Lovino, Federico Bolelli, Anni Virtanen, and Elisa Ficarra. Enhancing pfi prediction with gds-mil: A graph-based dual stream mil approach. In *International Conference on Image Analysis and Processing*, pages 550–562. Springer, 2023.
- [32] Adamantia Nikolaidi, Elena Fountzilias, Florentia Fostira, Amanda Psyrris, Helen Gogas, and Christos Papadimitriou. Neoadjuvant treatment in ovarian cancer: new perspectives, new challenges. *Frontiers in Oncology*, 12:820128, 2022.
- [33] Eric W Weisstein. Bernoulli distribution. <https://mathworld.wolfram.com/>, 2002.
- [34] Elisa Drelie Gelasca, Jiyun Byun, Boguslaw Obara, and BS Manjunath. Evaluation and benchmark for biological image segmentation. In *2008 15th IEEE international conference on image processing*, pages 1816–1819. IEEE, 2008.
- [35] Korsuk Sirinukunwattana, Shan E Ahmed Raza, Yee-Wah Tsang, David RJ Snead, Ian A Cree, and Nasir M Rajpoot. Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. *IEEE Transactions on Medical Imaging*, 35(5):1196–1206, 2016.
- [36] Bin Li, Yin Li, and Kevin W Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14318–14328, 2021.
- [37] Hongrun Zhang, Yanda Meng, Yitian Zhao, Yihong Qiao, Xiaoyun Yang, Sarah E Coupland, and Yalin Zheng. Dtf-d-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18802–18812, 2022.

- [38] Anabik Pal, Zhiyun Xue, Kanan Desai, Adekunbiola Aina F Banjo, Clement Akinfolarin Adepiti, L Rodney Long, Mark Schiffman, and Sameer Antani. Deep multiple-instance learning for abnormal cell detection in cervical histopathology images. *Computers in Biology and Medicine*, 138:104890, 2021.
- [39] Dibalo Chanda, Milan Aryal, Nasim Yahya Soltani, and Masoud Ganji. A new era in computational pathology: A survey on foundation and vision-language models. *arXiv:2408.14496*, 2024.
- [40] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv:2111.11432*, 2021.
- [41] Ming Y Lu, Bowen Chen, Andrew Zhang, Drew FK Williamson, Richard J Chen, Tong Ding, Long Phi Le, Yung-Sung Chuang, and Faisal Mahmood. Visual language pre-trained multiple instance zero-shot transfer for histopathology images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19764–19775, 2023.
- [42] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.
- [43] Yuxuan Sun, Yixuan Si, Chenglu Zhu, Xuan Gong, Kai Zhang, Pingyi Chen, Ye Zhang, Zhongyi Shui, Tao Lin, and Lin Yang. Cpath-omni: A unified multimodal foundation model for patch and whole slide image analysis in computational pathology. *arXiv:2412.12077*, 2024.
- [44] Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. *arXiv:1904.03323*, 2019.
- [45] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.
- [46] Zabir Al Nazi and Wei Peng. Large language models in healthcare and medical domain: A review. In *Informatics*, volume 11, page 57. MDPI, 2024.
- [47] Aishwarya Krishna Allada, Yuanxin Wang, Veni Jindal, Morteza Babee, Hamid R Tizhoosh, and Mark Crowley. Analysis of language embeddings for classification of unstructured

- pathology reports. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 2378–2381. IEEE, 2021.
- [48] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv:1907.11692*, 2019.
 - [49] Ari Aulia Hakim, Alva Erwin, Kho I Eng, Maulahikmah Galinium, and Wahyu Muliady. Automated document classification for news article in bahasa indonesia based on term frequency inverse document frequency (tf-idf) approach. In *2014 6th international conference on information technology and electrical engineering (ICITEE)*, pages 1–4. IEEE, 2014.
 - [50] Alexander Selivanov, Oleg Y Rogov, Daniil Chesakov, Artem Shelmanov, Irina Fedulova, and Dmitry V Dylov. Medical image captioning via generative pretrained transformers. *Scientific Reports*, 13(1):4171, 2023.
 - [51] Haixia Liu and Tim Brailsford. Reproducing “show, attend and tell: Neural image caption generation with visual attention”. In *Journal of Physics: Conference Series*, volume 2589, page 012012. IOP Publishing, 2023.
 - [52] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.
 - [53] Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016.
 - [54] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317, 2019.
 - [55] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6–12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014.

- [56] 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.
- [57] Wenkang Qin, Rui Xu, Peixiang Huang, Xiaomin Wu, Heyu Zhang, and Lin Luo. What a whole slide image can tell? subtype-guided masked transformer for pathological image captioning. *arXiv:2310.20607*, 2023.
- [58] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.
- [59] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders. *arXiv:1511.05644*, 2015.
- [60] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.
- [61] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [62] Tiancheng Lin, Zhimiao Yu, Hongyu Hu, Yi Xu, and Chang-Wen Chen. Interventional bag multi-instance learning on whole-slide pathological images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19830–19839, 2023.
- [63] Xueqin Zhang, Chang Liu, Huitong Zhu, Tianqi Wang, Zunguo Du, and Weihong Ding. A universal multiple instance learning framework for whole slide image analysis. *Computers in Biology and Medicine*, 178:108714, 2024.
- [64] Le Hou, Dimitris Samaras, Tahsin M Kurc, Yi Gao, James E Davis, and Joel H Saltz. Patch-based convolutional neural network for whole slide tissue image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2424–2433, 2016.
- [65] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017.

- [66] Xinggang Wang, Yongluan Yan, Peng Tang, Xiang Bai, and Wenyu Liu. Revisiting multiple instance neural networks. *Pattern Recognition*, 74:15–24, 2018.
- [67] Yun Liu, Krishna Gadepalli, Mohammad Norouzi, George E Dahl, Timo Kohlberger, Aleksey Boyko, Subhashini Venugopalan, Aleksei Timofeev, Philip Q Nelson, Greg S Corrado, et al. Detecting cancer metastases on gigapixel pathology images. *arXiv:1703.02442*, 2017.
- [68] Jianmo Ni, Gustavo Hernandez Abrego, Noah Constant, Ji Ma, Keith B Hall, Daniel Cer, and Yinfei Yang. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. *arXiv:2108.08877*, 2021.
- [69] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9, 2019.
- [70] Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805*, 2018.
- [71] Zhanwen Cheng, Feijiang Li, Jieting Wang, and Yuhua Qian. Deep embedding clustering driven by sample stability. *arXiv:2401.15989*, 2024.
- [72] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.
- [73] Matt J Kusner and José Miguel Hernández-Lobato. Gans for sequences of discrete elements with the gumbel-softmax distribution. *arXiv:1611.04051*, 2016.
- [74] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, Yoshua Bengio, et al. Graph attention networks. *stat*, 1050(20):10–48550, 2017.
- [75] Andrew L Maas, Awni Y Hannun, Andrew Y Ng, et al. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, page 3. Atlanta, GA, 2013.
- [76] Fabio A Spanhol, Luiz S Oliveira, Caroline Petitjean, and Laurent Heutte. A dataset for breast cancer histopathological image classification. *IEEE Transactions on Biomedical Engineering*, 63(7):1455–1462, 2015.
- [77] Amalia Luque, Javier Romero-Lemos, Alejandro Carrasco, and Luis Gonzalez-Abril. Temporally-aware algorithms for the classification of anuran sounds. *PeerJ*, 6:e4732, 2018.

- [78] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.
- [79] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380, 2014.
- [80] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.
- [81] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.
- [82] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32, 2019.
- [83] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.
- [84] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on attention for image captioning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4634–4643, 2019.
- [85] Brett Koonce. Resnet 34. In *Convolutional neural networks with swift for tensorflow: image recognition and dataset categorization*, pages 51–61. Springer, 2021.
- [86] Zijun Zhang. Improved adam optimizer for deep neural networks. In *2018 IEEE/ACM 26th international symposium on quality of service (IWQoS)*, pages 1–2. Ieee, 2018.
- [87] Minjie Wang, Da Zheng, Zihao Ye, Quan Gan, Mufei Li, Xiang Song, Jinjing Zhou, Chao Ma, Lingfan Yu, Yu Gai, et al. Deep graph library: A graph-centric, highly-performant package for graph neural networks. *arXiv:1909.01315*, 2019.