Seminar

Advanced AI for Histopathological Whole Slide Image Classification and Captioning

Presented by

S M Taslim Uddin Raju

MASc Candidate

Centre for Pattern Analysis and Machine Intelligence

Department of Electrical and Computer Engineering

University of Waterloo

Outline

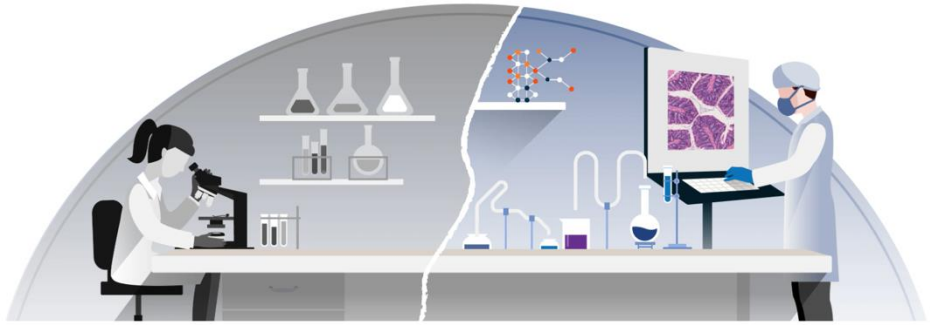
- Introduction
- Motivation
- Literature Review
- Research Gaps/Challenges
- Research Objectives
- Research Contributions
- Experimental Results and Discussions
- Conclusions and Future Works

UNIVERSITY OF WATERLOO

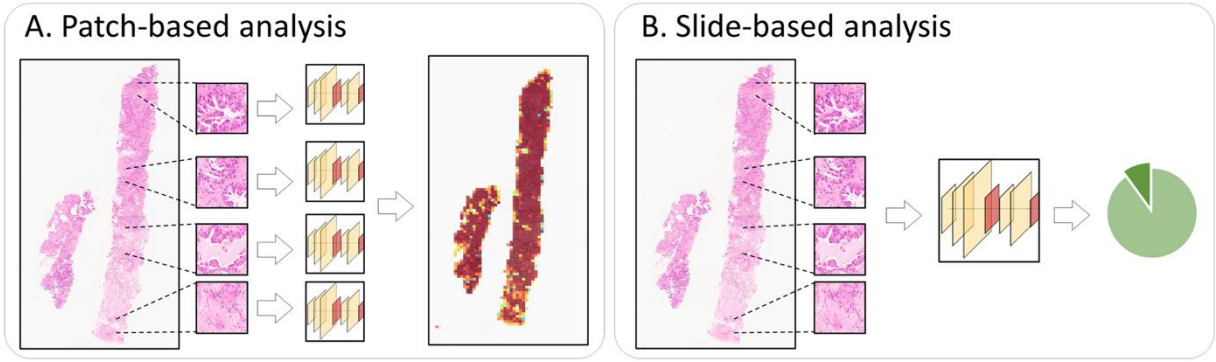
Introduction

❖Background

- ✓ Histopathology is microscopic examination of tissue, serving as cancer diagnosis, and treatment decisions.
- ✓ The process involves studying the size, shape, and patterns in cells and tissues from a patient's clinical records.
- ✓ Histopathology can be analyzed either patch-wise or slide-based.
- ✓ Histopathological captions are extracted from diagnostic reports and paired with image patches.
- ✓ Automatic diagnostic reports generation from whole slide images (WSIs) would reduce pathologists' workload.



Digital pathology for image analysis.



(A) Patch-wise and (B) Slide-based histopathological analyses.

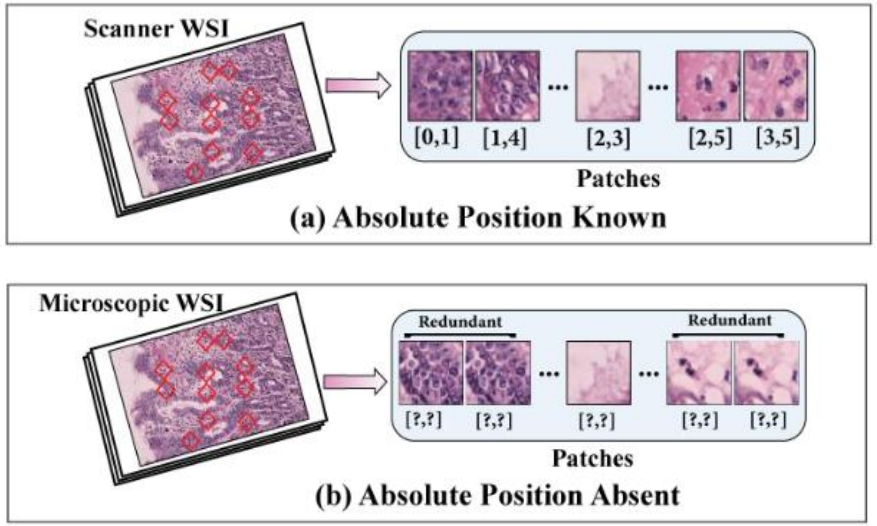
1. *https://blog.crownbio.com/digital-pathology*
2. *Scientifc Reports 12 (2022) 19075*

UNIVERSITY OF WATERLOO

Introduction

❖Problem Statements

- ✓ Histopathology WSIs have limitations due to their large size, complicating computational analysis.
- ✓ MIL divided the WSI into into independence patches for classification, neglecting viral tissue context and spatial interactions.
- ✓ Microscopic WSIs save cost and memory but lack positional data and include redundant patches from subjective captures.
- ✓ Sequential models such as RNNs/LSTMs face vanishing gradients, limiting long-range dependency capture in pathology data.



(a) Scanner WSI: Precise patch positions are known and (b) Microscopic WSI: Lacks position data, with redundant patches from subjective captures.

Motivation

- **Concise Descriptions:** Automated pathological captioning provides text summaries of large WSIs, allowing pathologists to focus on critical features.
- **Enhanced Accuracy:** Improve diagnostic consistency and support computer-aided diagnosis.
- **Role of LLMs & ViTs:** Biomedical language models excel at medical text generation, while Vision Transformers offer robust visual representations.
- **Multimodal Integration:** Combining ViTs with biomedical language models enables more precise captioning and classification for improved pathology interpretation.

UNIVERSITY OF WATERLOO

Literature Review

- **Multiple Instance Learning in WSI (Attention-based Aggregation Strategy)**
  - ✓ ABMIL [1]
    - Utilizing a learnable neural network to enhance the contribution of each instance via trainable attention weights
  - ✓ DSMIL [2]
    - Analysis of WSI using multi-scale features and features are extracted from patches through a self-supervised contrastive learning approach
  - ✓ TransMIL [3]
    - Leveraged the self-attention technique, utilizing the output data of a transformer network to encode the mutual correlations among instances
  - ✓ DTFD-MIL [4]
    - Determine probability of the instance within the structure of attention-based MIL and employed to assist in generating and analyzing the image features

[1]. *International conference on machine learning*, (2018):2127–2136.

[2]. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, (2021) 14318–14328*

[3]. *Advances in Neural Information Processing Systems*, 34(2021):2136–2147.

[4]. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, (2022) 18802–1881*

UNIVERSITY OF WATERLOO

Literature Review (Cont'd)

- **Multiple Instance Learning in WSI (Graph-based Aggregation Strategy)**
  - ✓ DeepGraphSurv [1]
    - Developed a graph convolutional neural network, and combined local patch features with global topological information through spectral graph convolution
  - ✓ DAS-MIL [2]
    - Proposed a novel graph-based multi-instance learning approach and integrated with self-knowledge distillation to improve information flow across multiple resolutions
  - ✓ GDS-MIL [3]
    - Integrated a graph attention networks with MIL to enhance the prognosis indicator (the Platinum-Free Interval) from WSIs

[1]. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, (2018): *172–182*.

[2]. *International conference on medical image computing and computer-assisted intervention, (2023) 248–258.*

[3]. *International Conference on Image Analysis and Processing*, (2023):550–562.

UNIVERSITY OF WATERLOO

Literature Review (Cont'd)

- **The most recent works for caption generation are demonstrated here**

  - ✓ Tsuneki et al. [1]
    - Developed an automated captioning system using high-resolution WSIs
    - EfficientNetB3 and DenseNet121 models were pre-trained, and an RNN-based decoder was used to generate captions
  - ✓ Qin et al. [2]

    Developed a subtype-guided masked transformer network to generate captions using transformer
  - ✓ Zhou et al. [3]
    - Developed multimodal multi-task MIL system called PathM3 [15] was proposed for WSI to classify and generate captions
    - Used a query-based transformer to accurately correlate WSIs with diagnostic texts

[1]. *International Conference on Medical Imaging with Deep Learning*, (2022): *1235–1250.*

[2]. a*rXiv:(2023):2310–20607.*

[3]. a*rXiv:(2024):2403–08967.*

UNIVERSITY OF WATERLOO

Research Gaps/Challenges

- ✓ Most MIL models overlook long-distance dependencies by neglecting patch interactions and spatial positions
- ✓ A few of the graph-based MIL tried to overcome these limitations by modeling the spatial relationships between patches in microscopic images
- ✓ Redundancy in microscopic images leads to excessively dense and repetitive graph connections, reduces the models' ability
- ✓ RNNs and LSTMs struggle with vanishing gradients while LLMs have their advanced capability to process and understand complex text

UNIVERSITY OF WATERLOO

Research Objectives

- ✓ Designed an efficient feature extractor which captures the complex attributes of tissues in histopathological patches
- ✓ Employed a clustering method to reduce the redundancy and established the spatial relationship among the patches
- ✓ Proposed a caption-generator model which takes a sequence of patch-embeddings to generate a diagnostic report for the patient

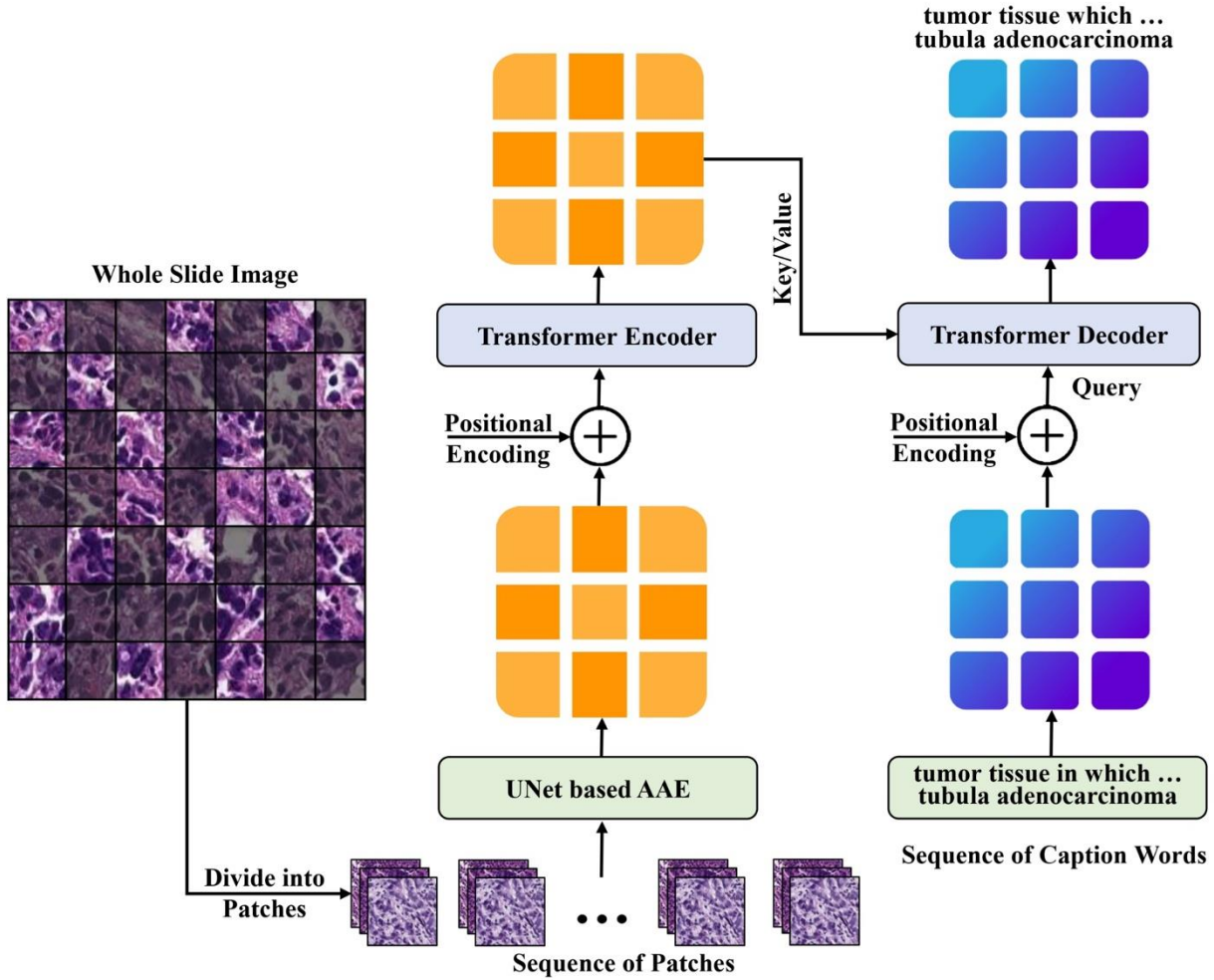
UNIVERSITY OF WATERLOO

Research Contributions

- ✓ Designed the **TransUAAE-CapGen Architecture,** consists of hybrid UNet-based Adversarial Autoencoder (AAE) and a transformer
- ✓ Developed UNet based AAE to extracts complex tissue properties and transformer to generate the caption
- ✓ Proposed the **GNN-ViTCap Architecture** for simultaneous classification and caption generation
- ✓ Applied **deep embedded clustering** for removal of redundancy, and **graph neural networks** to capture spatial relationships
- ✓ Integrated visual features into language models, combining visual and textual modalities to generate context-aware captions

UNIVERSITY OF WATERLOO

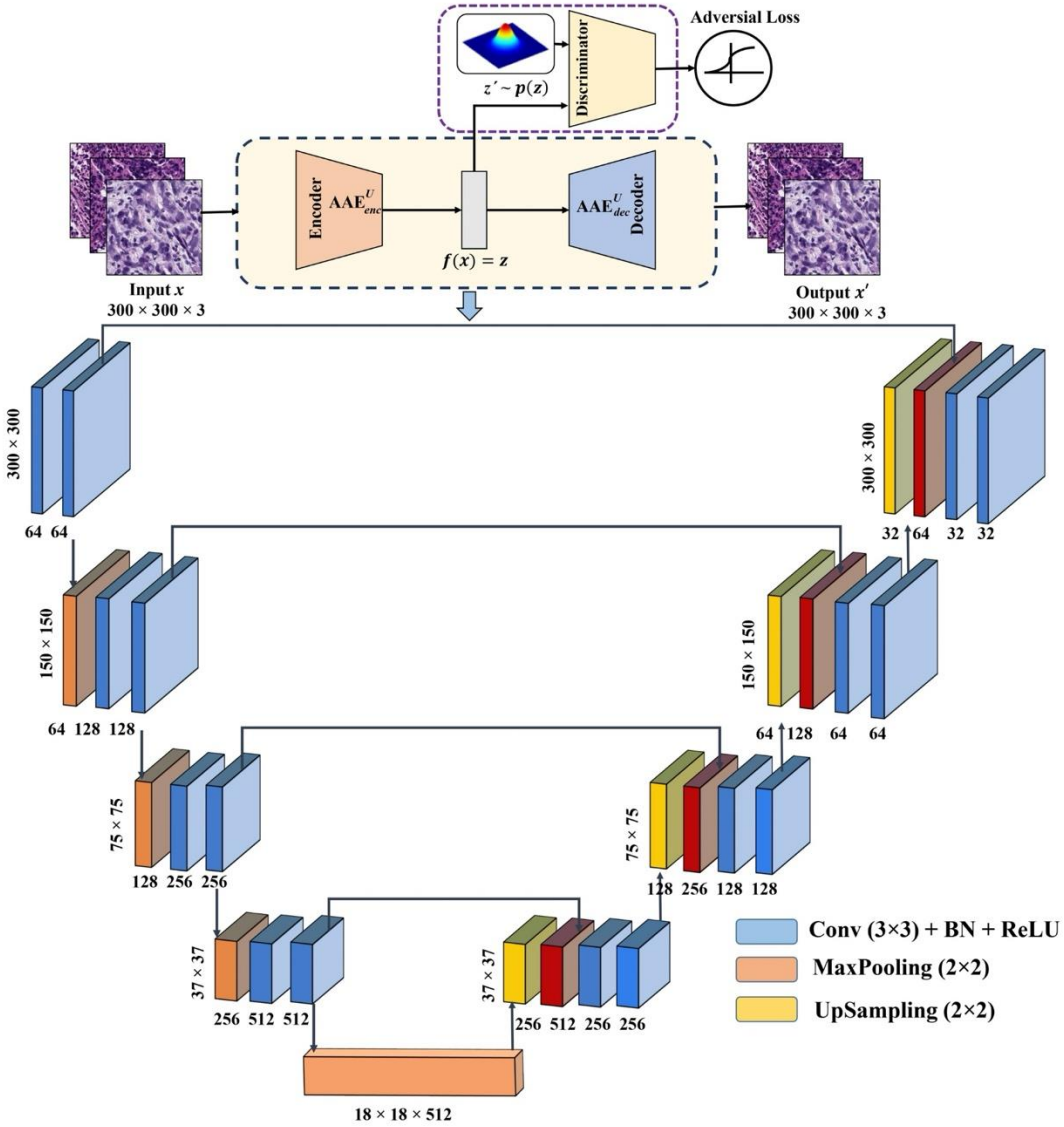
TransUAAE-CapGen Architecture: WSI Caption Generation



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

UNet-Based AAE Architecture

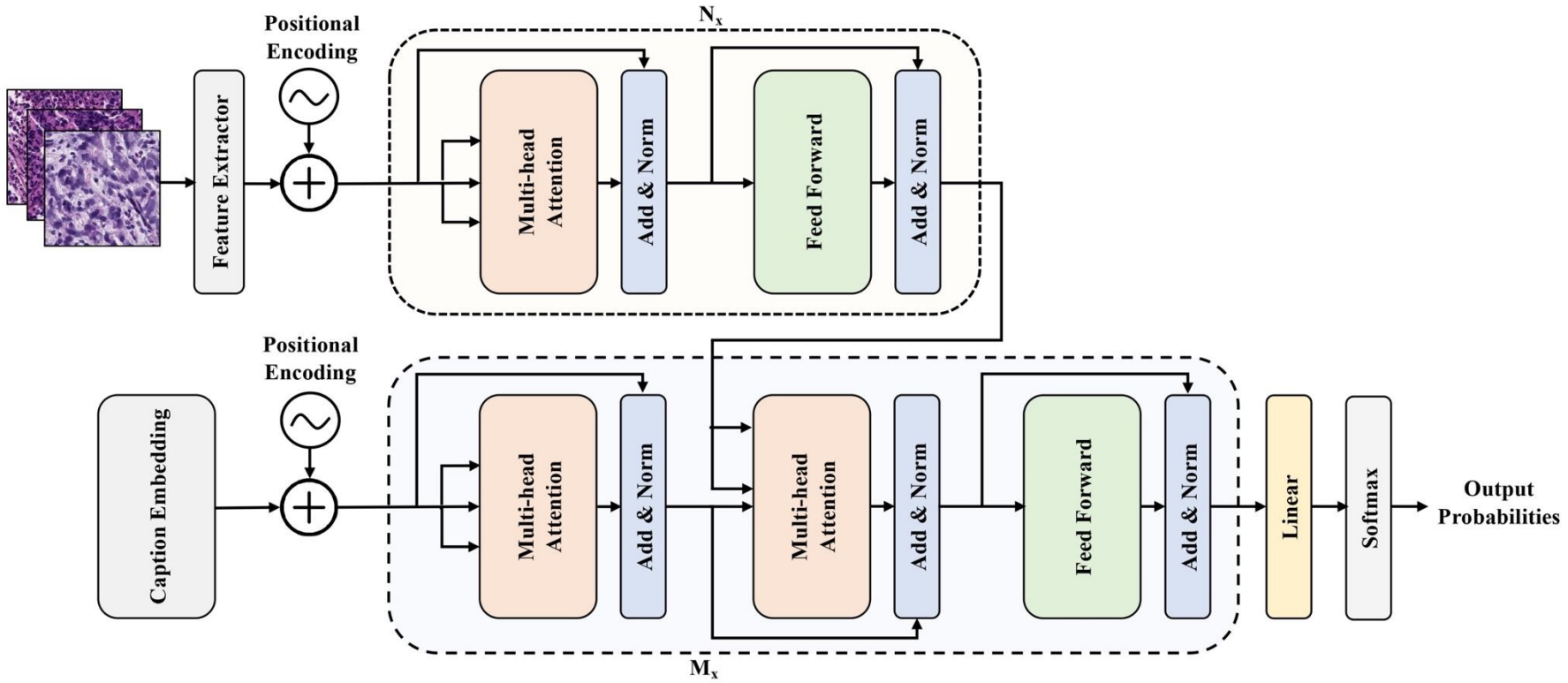
- AAE combined with UNet, it captures both local and global features for improved generalization to unseen data



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

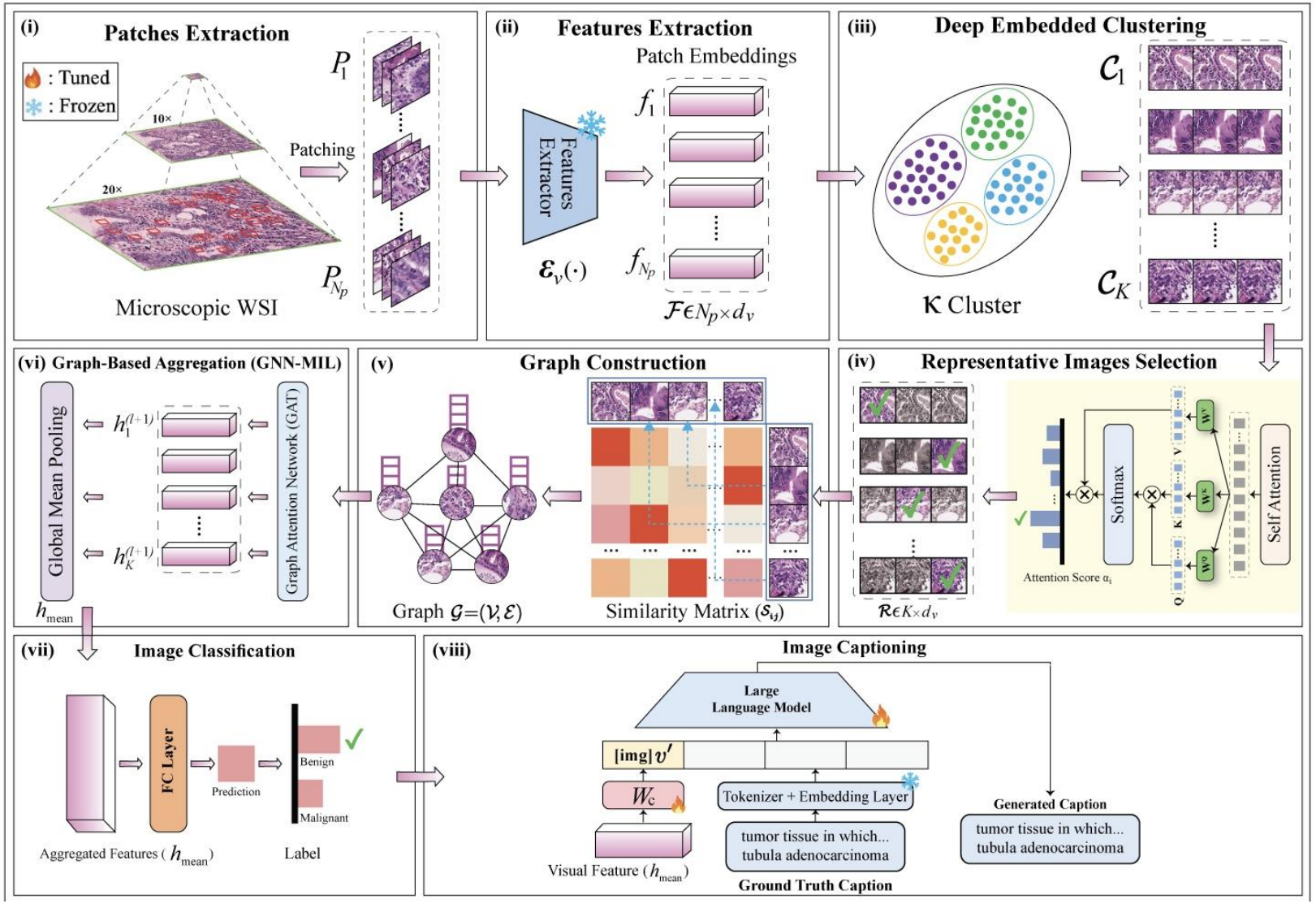
UNIVERSITY OF WATERLOO

Caption Generator through Transformer



Architecture of the transformer model for caption generation.

GNN-ViTCap Architecture: WSI Classification and Captioning



Overview of the GNN-ViTCap framework for microscopic whole slide images classification and captioning.

UNIVERSITY OF WATERLOO

Results and Analysis

- Dataset Collection and Description
- Performance Indices
- Simulation Results

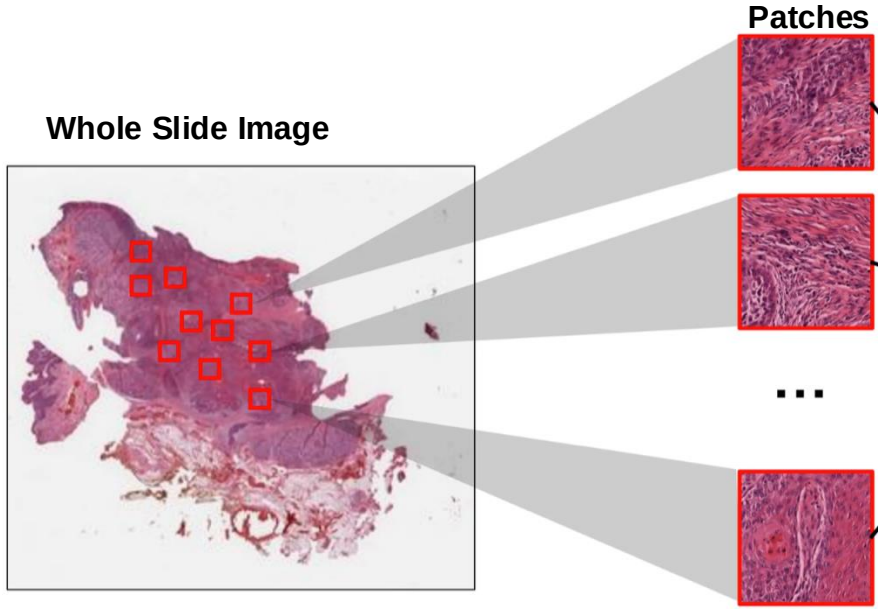
Dataset Collection and Description

- **PatchGastric Dataset**

✓ Stomach adenocarcinoma endoscopic biopsy samples images, paired with histopathological captions
✓ 991 whole slide images with **262,777** extracted patches
✓ Each patch is **300 × 300** pixels in dimension
✓ Patches are aligned with corresponding histopathological captions extracted from diagnostic reports

- **BreakHis Dataset**

✓ 7,909 microscopic histopathology biopsy images from 82 patients
✓ Image is classified into benign and malignant tumor categories
✓ Each patch is 700 × 460 pixels in dimension of pixels.



Caption: atypical epithelium cells with and proliferating atypical forming solid tubular observed is observed moderately Differentiated adenocarcinoma solid

International Conference on Medical Imaging with Deep Learning, (2022): *1235–1250.*
IEEE Transactions on Biomedical Engineering, 63(2015): 455–1462.
Proceedings of Machine Learning Research (2022):1235-1250.

UNIVERSITY OF WATERLOO

Performance Indices

- Several performance measures are adopted to evaluate the proposed architectures for classification.

✓ $Precision = \frac{1}{N}\frac{TP}{TP+FP}$

✓ $Recall = \frac{1}{N}\frac{TP}{TP+FN}$

✓ $F1 - Score = \frac{1}{N}\frac{2TP}{2TP+FP+FN}$

✓ $AUC = \int_0^1 TPR(t)dt$

Confusion Matrix

	Actual	
Predicted	**TP**	**FP**
	FN	**TN**

- Here, the correctly recognized samples are denoted as True Positives (*TP*), and True Negatives (*TN*)
- The incorrectly classified samples are known as False Positives (*FP*), and False Negatives (*FN*)
- *N* represents the total number of data samples
- TRP represents true positive rate

UNIVERSITY OF WATERLOO

Performance Indices (Cont'd)

- Statistical measures to evaluate the performance of the proposed architecture for captioning

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

✓ $METEOR = F_{\mu}\,(1-\rho)$

✓ $ROUGE - N = \frac{\sum_{S \in AS} \sum_{g_n \in S} MatchCount(g_n)}{\sum_{S \in AS} \sum_{g_n \in S} TotalCount(g_n)}$

✓ $CIDEr_n(c_j, S_j) = \frac{1}{n} \sum \frac{h_n(c_j).h_n(s_j k)}{||h_n(c_j)||.||h_n(s_j k)||}$

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

- $P(i)$represents the precision for each $n-$gram size (unigrams, bigrams, trigrams, etc.).
- F_{μ} is harmonic mean of precision and recall and ρ denotes the penalty.
- $MatchedCount(g_n)$ represents the maximum number of $n-$grams.
- $h_n(c_j)$ is a vector of all $n-$grams of length n in the candidate caption, and $||h_n(c_j)||$ is its magnitude.
- $c_j$is the candidate captions, S_j is the set of actual captions.

UNIVERSITY OF WATERLOO

Results: TransUAAE-CapGen Architecture for Captioning

- **Quantitative Results**

Performance metric (%) of our proposed TransUAAE-CapGen 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
TransUAAE-CapGen	$\boldsymbol{P \times 18 \times 18 \times 512}$	**92.1**	**90.3**	**88.3**	**86.8**	**59.6**	**89.3**	**7.72**

Comparison of our proposed method with existing methods using the PatchGastric dataset

Model	BLEU@4(%)	METEOR(%)	ROUGE(%)	CIDEr
LSTM+EfficientNetB3_AvgP [1]	28.3	30.54	49.73	2.31
LSTM+EfficientNetB3_MaxP [1]	32.4	27.75	46.49	1.62
LSTM+DenseNet121_AvgP [1]	27.2	28.21	46.95	1.54
LSTM+DenseNet121_MaxP [1]	32.3	26.85	45.47	1.48
PathM3 [2]	52.0	39.4	-	-
SGMT [3]	55.11	43.17	69.68	4.83
TransUAAE-CapGen	**86.8**	**59.6**	**89.3**	**7.72**

[1]. *International Conference on Medical Imaging with Deep Learning*, (2022): *1235–1250.*

[2]. *arXiv:(2023):2310–20607.*

[3]. *arXiv:(2024):2403–08967.*

UNIVERSITY OF WATERLOO

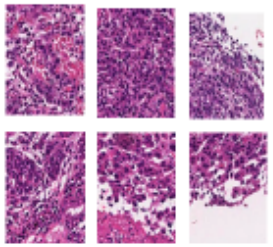
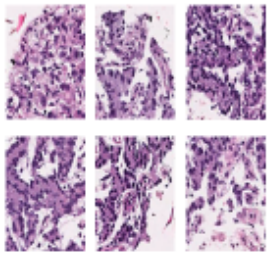
Results: TransUAAE-CapGen Architecture for Captioning

- **Qualitative Results**

✓ Both WSI images were randomly selected from the test set

✓ The TransUAAE-CapGen model effectively generates captions for histopathological images closely matching reference captions

✓ The model shows proficiency in capturing detailed pathological features comparable to expert descriptions

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

UNIVERSITY OF WATERLOO

Results: GNN-ViTCap Architecture for Classification

- **Research 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?***

UNIVERSITY OF WATERLOO

Results: GNN-ViTCap Architecture for Classification (Cont'd)

- **Quantitative Results**

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

Model	Precision	Recall	F1−Score	AUC
ABMIL [1]	0.835	0.922	0.900	0.871
DSMIL [2]	0.872	0.842	0.856	0.869
TransMIL [3]	0.865	0.908	0.886	0.862
DTFD [4]	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**

[1]. *International conference on machine learning*, (2018):2127–2136.

[2]. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, (2021) 14318–14328*

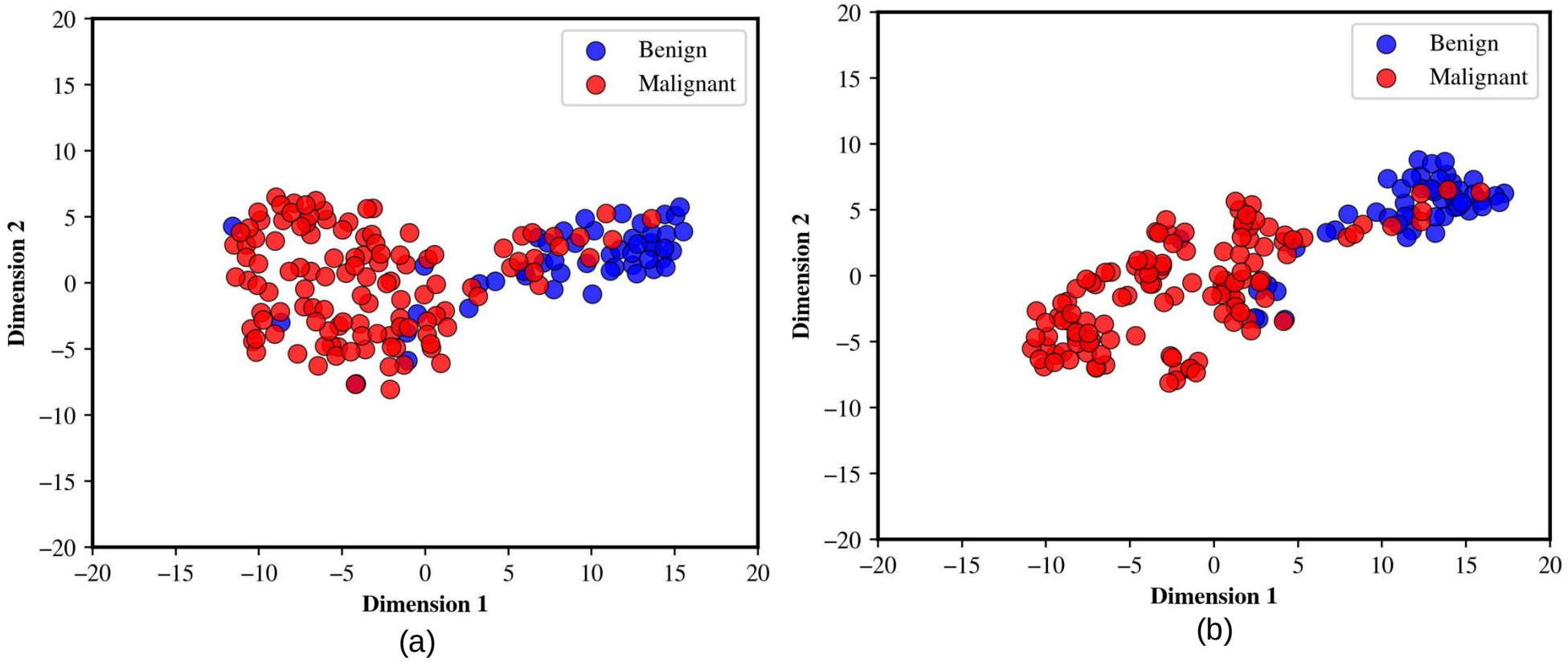
[3]. *Advances in Neural Information Processing Systems*, 34(2021):2136–2147.

[4]. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, (2022) 18802–1881*

UNIVERSITY OF WATERLOO

Results: GNN-ViTCap Architecture for Classification (Cont'd)

- **Interpretability: t-SNE feature visualization**



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

Results: GNN-ViTCap Architecture for Captioning (Cont'd)

- **Quantitative Results**

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 [1]	EfficientNetB3_AvgP	LSTM	-	-	-	28.3	30.54	49.73	2.31
	EfficientNetB3_MaxP		-	-	-	32.4	27.75	46.49	1.62
	DenseNet121_AvgP		-	-	-	27.2	28.21	46.95	1.54
	DenseNet121_MaxP		-	-	-	32.3	26.85	45.47	1.48
PathM3 [2]	ViT-g/14	Flan-T5	-	-	-	52.0	39.4	-	-
SGMT [3]	CNN	Transformer	-	-	-	55.11	43.17	69.68	4.83
GNN-ViTCap	**ViT-B/16**	BioGPT	0.802	0.748	0.713	0.686	0.485	0.766	5.72
		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**

[1]. *International Conference on Medical Imaging with Deep Learning*, (2022): *1235–1250.*

[2]. *arXiv:(2023):2310–20607.*

[3]. *arXiv:(2024):2403–08967.*

UNIVERSITY OF WATERLOO

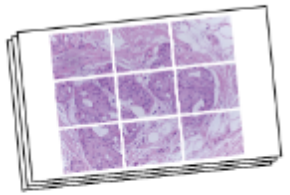
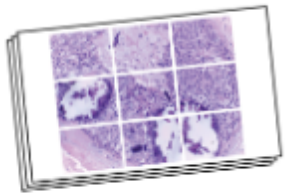
Results: GNN-ViTCap Architecture for Captioning (Cont'd)

- **Qualitative Results**

✓ Again two WSI images were randomly selected from the test set

✓ The GNN-ViTCap model effectively generates captions for histopathological images closely matching reference captions

✓ The GNN-MIL along with LLMs show proficiency in capturing detailed pathological features comparable to expert descriptions

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 submucosa can be irregular in the epithelium well differentiated tubular adenocarcinoma
	in the superficial epithelium tumor tissue that invades by forming medium sized to small irregular ducts is observed moderately differentiated adenocarcinoma	on the superficial epithelium tumor tissue that infiltrates by forming medium sized to small irregular ducts is observed moderately differentiated tubular

UNIVERSITY OF WATERLOO

Future Works

- **Adaptive Clustering for improvement of GNN-ViTCap architecture**
  - ✓ Dynamically determine optimal cluster numbers
  - ✓ Minimizes information loss and ensures efficient analysis
- **Perform parameter-efficient fine-tuning for Large Language Models**
  - ✓ Minimize computational costs while maintaining model performance and efficiency.
- **Addressing hyperparameter sensitivity and reducing the risk of model collapse or learning instability**

UNIVERSITY OF WATERLOO

Conclusions

- We developed a TransUAAE-CapGen architecture to generate the caption from WSI patches.
- We introduced a novel GNN-ViTCap approach to classify WSI as benign and malignant and generate the caption
- We applied attention-based deep embedded clustering for removal of redundancy, and graph neural networks to capture spatial relationships
- We compared the model performances with SOTA approaches and experiments show the model aids healthcare professionals with accurate WSI captions.

UNIVERSITY OF WATERLOO

Publications

International Conferences

1. **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 (SMC)*, IEEE, Sarawak, 6-10 Oct. 2024.
2. **S M Taslim Uddin Raju**, 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]**
3. 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 International Conference on Systems, Man, and Cybernetics (SMC)*, 2025 **[Submitted]**

Book Chapters

1. **S M Taslim Uddin Raju**, Md. Milon Islam, Sheikh Nooruddin, Fakhri Karray and Ghulam Muhammad, “Internet of Health Things: An Introduction,” Book: *Blockchain and Digital Twins for the Internet of Medical Things in Smart Hospitals*, Elsevier, 2024.

UNIVERSITY OF WATERLOO

Thank you !

UNIVERSITY OF WATERLOO

Supporting Slides

References

[1] N C. Frascarelli, N. Fusco, and G. Vago, “Artificial intelligence in diagnostic and predictive pathology,” in Artificial Intelligence for Medicine. Elsevier, 2024, pp. 81–90.

[2] L. Xu, Q. Tang, J. Lv, B. Zheng, X. Zeng, and W. Li, “Deep image captioning: A review of methods, trends and future challenges,” Neurocomputing, p. 126287, 2023

[3] Y. Sun, H. Wu, C. Zhu, S. Zheng, Q. Chen, K. Zhang, Y. Zhang, X. Lan, M. Zheng, J. Li et al., “Pathmmu: A massive multimodal expert-level benchmark for understanding and reasoning in pathology,” arXiv preprint arXiv:2401.16355, 2024

[4] N. Dimitriou, O. Arandjelovi´c, and P. D. Caie, “Deep learning for whole slide image analysis: an overview,” Frontiers in medicine, vol. 6, p. 264, 2019

[5] S. Elbedwehy, T. Medhat, T. Hamza, and M. F. Alrahmawy, “Enhanced descriptive captioning model for histopathological patches,” Multimedia Tools and Applications, vol. 83, pp. 36 645–36 664, 2024

[6] T. Ghandi, H. Pourreza, and H. Mahyar, “Deep learning approaches on image captioning: A review,” ACM Computing Surveys, vol. 56, no. 3, pp. 1–39, 2023

[7] M. Tsuneki and F. Kanavati, “Inference of captions from histopathological patches,” in International Conference on Medical Imaging with Deep Learning. PMLR, 2022, pp. 1235–1250

[8] MA. K. Allada, Y. Wang, V. Jindal, M. Babee, H. R. Tizhoosh, and M. 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). IEEE, 2021, pp. 2378–2381

[9] A. Selivanov, O. Y. Rogov, D. Chesakov, A. Shelmanov, I. Fedulova, and D. V. Dylov, “Medical image captioning via generative pretrained transformers,” Scientific Reports, vol. 13, no. 1, p. 4171, 2023

UNIVERSITY OF WATERLOO

References (Cont'd)

[10] A. H. Song, G. Jaume, D. F. Williamson, M. Y. Lu, A. Vaidya, T. R. Miller, and F. Mahmood, "Artificial intelligence for digital and computational pathology," Nature Reviews Bioengineering, vol. 1, no. 12, pp. 930–949, 2023

[11] Ilse, J. Tomczak, and M. Welling, "Attention-based deep multiple nstance learning," in International conference on machine learning. PMLR, 2018, pp. 2127–2136.

[12] Z. Shao, H. Bian, Y. Chen, Y. Wang, J. Zhang, X. Ji et al., "Transmil: Transformer based correlated multiple instance learning for whole slide image classification," Advances in neural information processing systems, vol. 34, pp. 2136–2147, 2021

[13] Liyakathunisa Syed, Saima Jabeen, S Manimala, and AbdullahAlsaeedi. Smart healthcare framework for ambient assisted living using iomt and big data analytics techniques. *Future Generation Computer Systems*, 101:136–151, 2019

[14] Elbedwehy, T. Medhat, T. Hamza, and M. F. Alrahmawy, "Enhanced descriptive captioning model for histopathological patches," Multimedia Tools and Applications, vol. 83, no. 12, pp. 36 645–36 664, 2024

[15] A. L. Maas, A. Y. Hannun, A. Y. Ng et al., "Rectifier nonlinearities improve neural network acoustic models," in Proc. icml, vol. 30, no. 1. Atlanta, GA, 2013, p. 3

[16] Nada Y Philip, Joel JPC Rodrigues, Hong gang Wang, Simon James Fong, and Jia Chen. Internet of things for in-home health monitoring systems: current advances, challenges and future directions. *IEEE Journal on Selected Areas in Communications*, 39(2):300–310, 2021.

[17] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax," arXiv:1611.01144, 2016

[18] Jun-Ho Choi and Jong-Seok Lee. Embracenet: A robust deep learning architecture for multimodal classification. Information Fusion, 51:259–270, 2019.

UNIVERSITY OF WATERLOO

Proposed TransUAAE-CapGen Architecture

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 (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

UNIVERSITY OF WATERLOO

UNet-Based AAE Architecture (Cont'd)

- AAE uses adversarial training to align its latent space with a prior distribution, and when combined with UNet, it captures both local and global features for improved generalization to unseen data

✓ The reconstruction loss L_{recon} for the UNet model is computed as the MSE between the input image x and its reconstruction x' (1).

$$L_{recon} = MSE\ (x, x') = \frac{1}{n}\left|\left|x_i\ - x_i'\right|\right|_2^2$$

✓ The adversarial loss of the UNet-based AAE architecture can be calculated using (2).

$$L_{adv} = E_x\left[\log\left(D_{AAE}^{U}(x)\right)\right] + E_z[\log(a - D_{AAE}^{U}(G^U(z))]$$

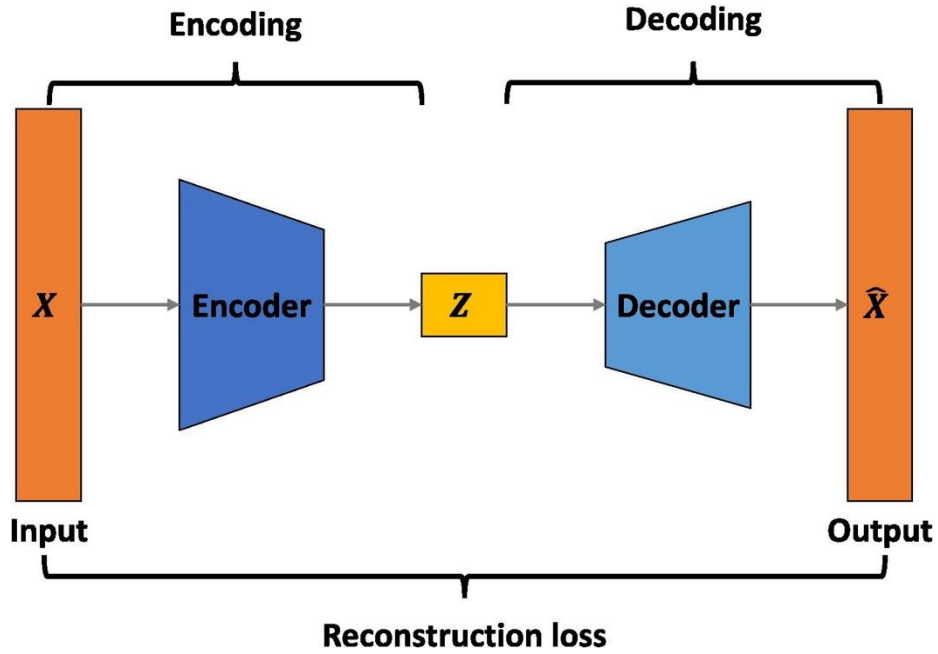


Fig. 6. Architecture of adversarial autoencoder for feature extraction and output generation.

Energy & Buildings 285 (2023): 112876

UNet-Based AAE Architecture (Cont'd)

- ✓ The **discriminator** loss is calculated as follows:

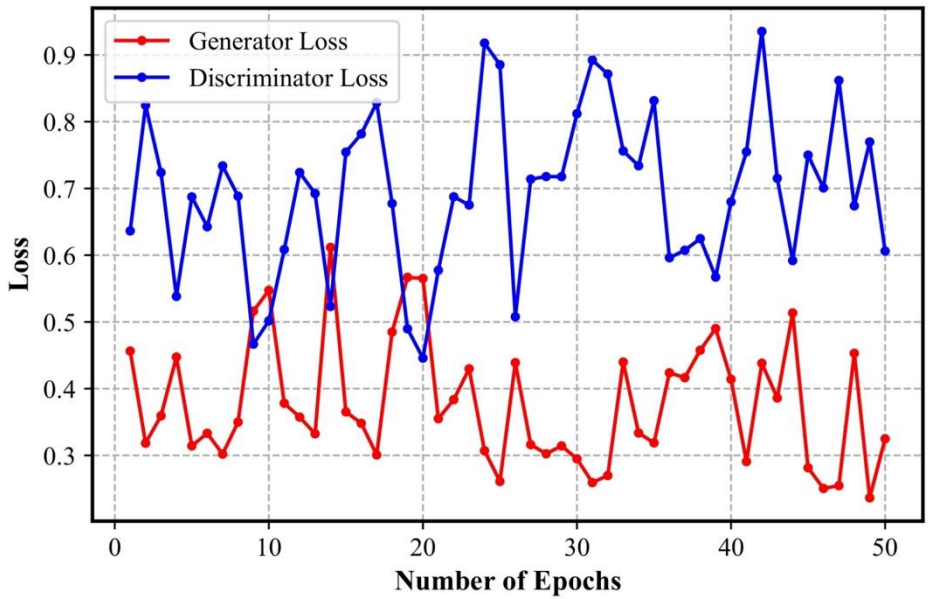
$$L_D = -\frac{1}{n}\sum_{i=1}^{n}\left[y_i.\log\left(D_{AAE}^{U}(z_i)\right) + (1-y_i).\log\left(1-D_{AAE}^{U}(z_i')\right)\right]$$

- ✓ Therefore, the **generator** loss can be calculated as:

$$L_G = \frac{1}{n}\sum_{i=1}^{n}\log\left(1-D_{AAE}^{U}\left(G^{U}(z)\right)\right)$$

- ✓ Therefore, the total loss L_{total} for the hybrid UNet-based AAE model be expressed as in (5).

$$L_{total} = \lambda.L_{recon} + (1-\lambda).L_{adv}$$



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

UNet-Based AAE Architecture (Cont'd)

- ✓ **Training Details:** The model is trained over 50 epochs using 30,000 histopathological image samples.
- ✓ **Model Architecture:** The UNet-based AAE has a latent space with dimensions of 18 × 18 × 512.
- ✓ **Feature Extraction:** Features are extracted from the images using the trained UNet-based AAE encoder.
- ✓ **Final Feature Dimension:** The final feature dimensions for each WSI with P patches are P × 18 × 18 × 512.

Caption Generator through Transformer

- The extracted feature sequences and corresponding captions are fed into a transformer model, where each patch's features are combined with textual captions.

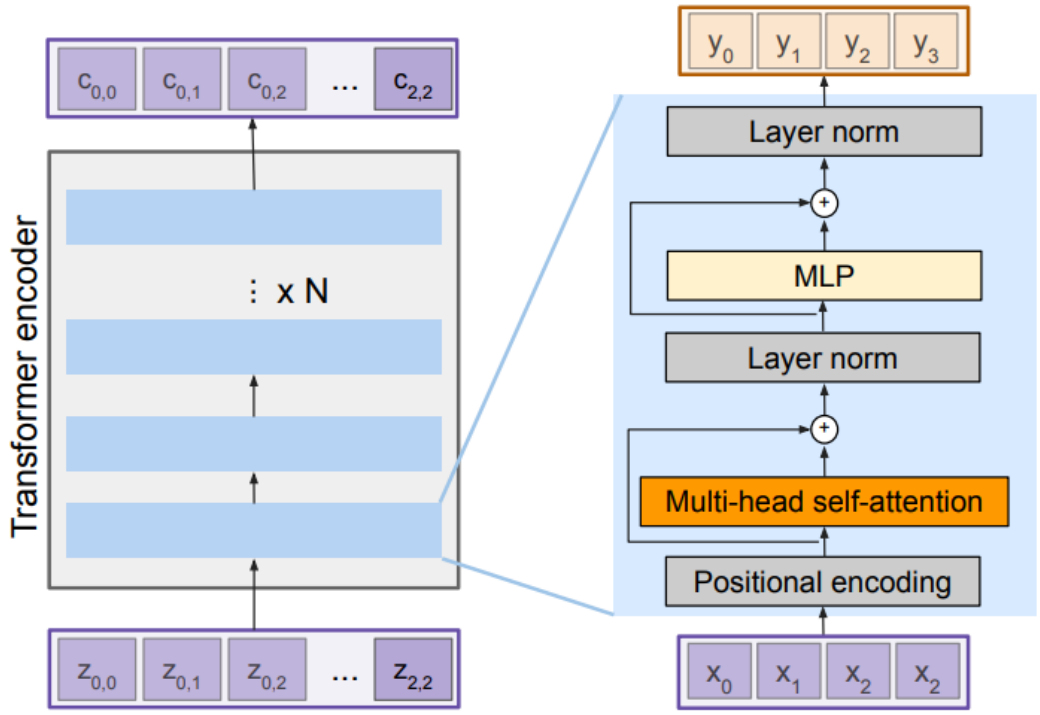
❖ **Transformer Encoder Block**:

✓ **For Inputs:** Set of extracted features **x**

✓ **Outputs**: Set of vectors y.

✓ Self-attention is the only interaction between vectors.

✓ Layer norm and MLP operate independently per vector.



Advances in Neural Information Processing Systems, 2017: 1-11

UNIVERSITY OF WATERLOO

Caption Generator through Transformer (Cont'd)

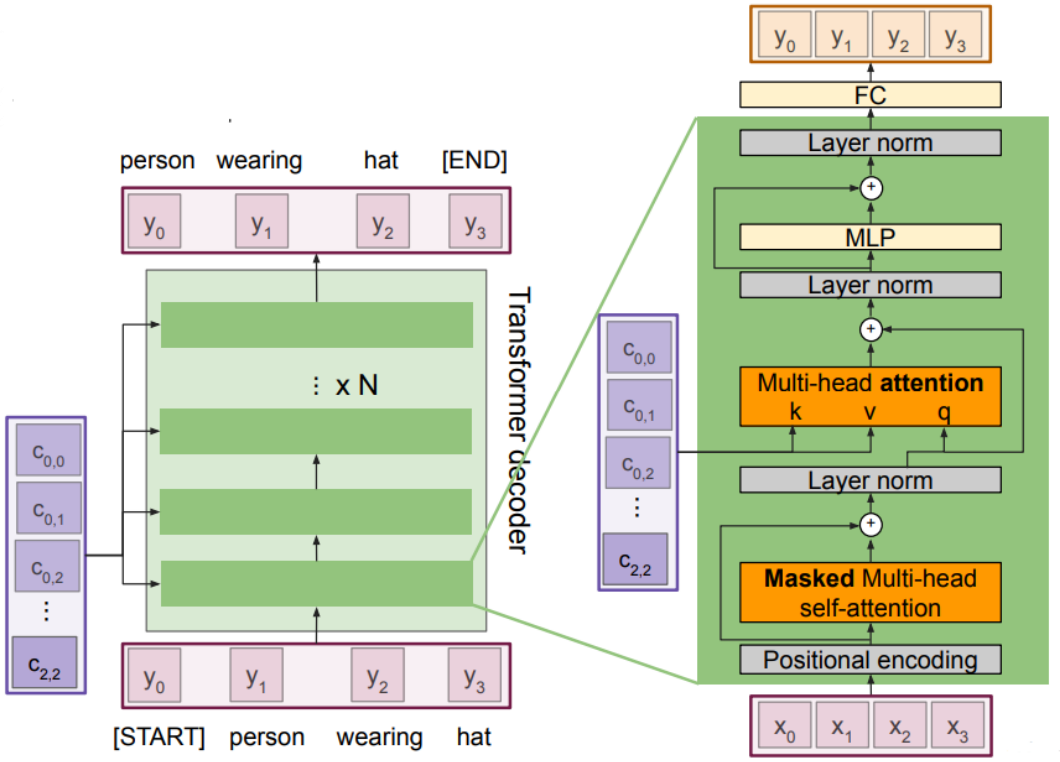
❖ **Transformer Decoder Block**:

✓ **For Inputs:** Set of vector **x** and set of context vector **c**

✓ **Outputs**: Set of vectors y.

✓ Masked Self-attention only interacts with past inputs.

✓ Multi-head attention block is NOT self-attention. It attends over encoder outputs.



Advances in Neural Information Processing Systems, 2017: 1-11

UNIVERSITY OF WATERLOO

Caption Generator through Transformer (Cont'd)

❖ Self-Attention

- ✓ Multi-head attention uses scaled dot-product attention with *query(Q)*, *key(K)*, and *value(V)* to compute attention weights as follows:

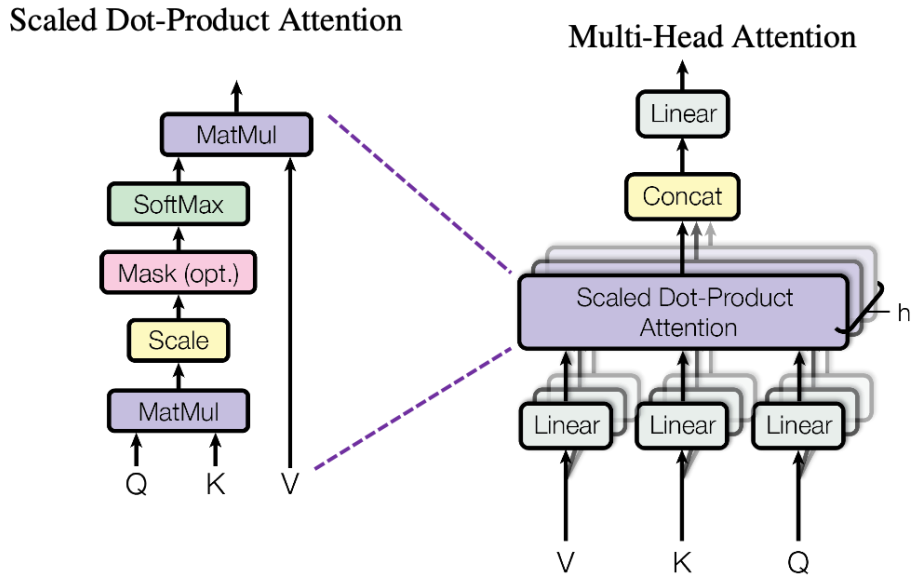
$$Attention\,(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

- ✓ For the i^{th} head, the computation of the attention output is as follows:

$$MultiHead\,(Q,K,V) = Concat(z_1,\ldots,Z_h)W^o$$

- ✓ The multi-head attention concatenates individual attention heads (Z_1 to Z_h) as (9):

$$z_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$$



UNIVERSITY OF WATERLOO

Caption Generator through Transformer (Cont'd)

❖ Loss Function

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

$$L_{TransUAAE-CapGen} = -\frac{1}{T}\sum_{t=1}^{T}\log(p(C_t|C_{<t}, x))$$

Experimental Result Analysis

❖ Experimental Environment

- ✓ We performed our experiments on a high-computing machine.
- ✓ Configuration: (Apple M3 Pro chip with 11-core CPU, 14-core GPU and 16-core Neural Engine).
- ✓ All the experiments were conducted using Python(v3.11.8) and PyTorch (v2.4.0).

❖ Hyper-parameter Settings

- ✓ We performed the experiments for 50 epochs to train the proposed architecture.
- ✓ Batch size was set to 16.
- ✓ Learning rate was set to 0.001
- ✓ Weight decay was epoch/length(train)
- ✓ Adam was used as optimizer.
- ✓ Categorical cross-entropy was used as the loss function.

Experimental Result Analysis (Cont'd)

- **Quantitative Results**

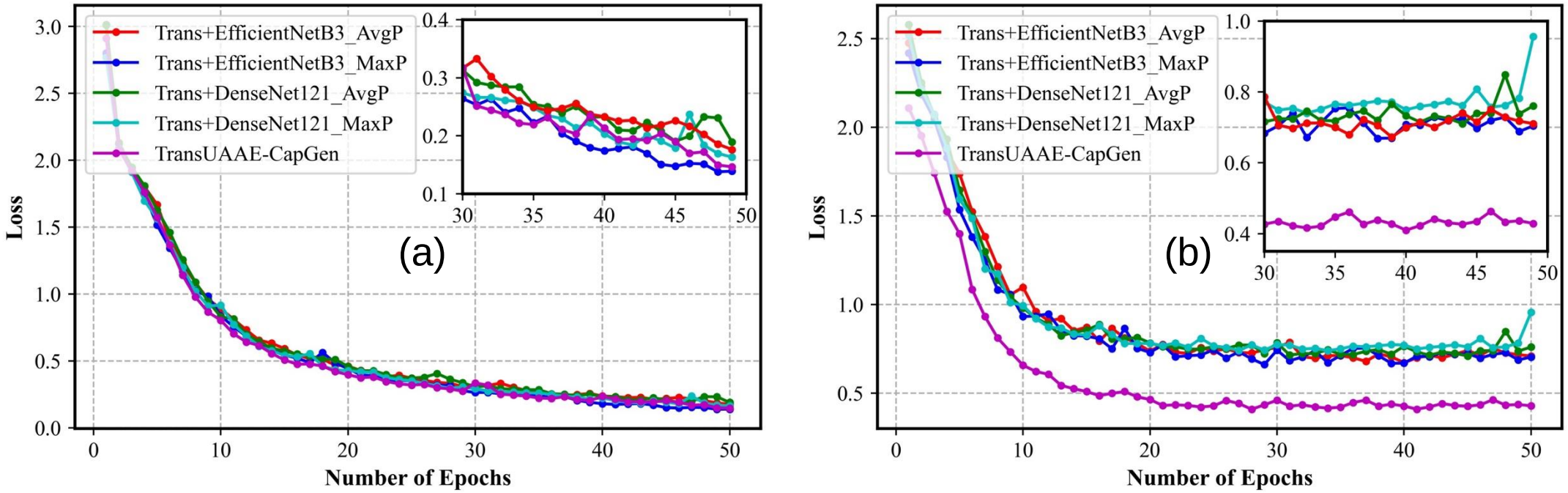


Fig. 11. Loss values of the proposed TransUAAE-CapGen approach along with two baseline feature extraction models. (a) Training (b) Validation.

UNIVERSITY OF WATERLOO

Proposed GNN-ViTCap Architecture

Algorithm 2: GNN-ViTCap architecture for classification and captioning

Input: Dataset: $\mathcal{D} = \{(X^{(s)}, Y^{(s)}, C^{(s)}) \mid s = 1, \ldots, N\}$, 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: $f_k^{(s)} = \mathcal{E}_v(p_k^{(s)}) \ \forall k = 1, \ldots, N_p$;

Concatenate features: $\mathcal{F}^{(s)} = [f_1^{(s)}; \ldots; f_{N_p}^{(s)}]$;

`/* Clustering Module */`

Initialize cluster centroids, $\mu_k^{(s)}$;

while $\left|\mathcal{L}_{Clu}^{(Curr)} - \mathcal{L}_{Clu}^{(Prev)}\right| < \epsilon$ **do**

Assign clusters using Student's t-distribution: $q_{ik}^{(s)} = \frac{\left(1+\|f_i^{(s)} - \mu_k^{(s)}\|^2\right)^{-1}}{\sum_{j=1}^{K}\left(1+\|f_i^{(s)} - \mu_j^{(s)}\|^2\right)^{-1}}$;

Compute target distribution $t_{ik}^{(s)}$;

Minimize Kullback-Leibler (KL) divergence $\mathcal{L}_{\text{Clu}}$;

Update cluster centroids $\mu_k^{(s)}$;

end

`/* Representative images selection */`

for *each cluster* $k \leftarrow 1$ **to** $\mathcal{K}$ **do**

Extract cluster features: $\mathcal{Z}_k = \{f_i^{(s)} \mid i \in C_k\}$;

Compute attention scores e_i, and weights α_i, using scaler dot attention mechanism;

Select representative embeddings $r_k^{(s)}$;

end

Aggregate representative images: $\mathcal{R}^{(s)} = [r_1^{(s)}, \ldots, r_K^{(s)}]$;

`/* Graph-Based Aggregation Module */`

Construct similarity matrix using cosine similarity: $S_{ij}^{(s)} = \langle \frac{r_i^{(s)}}{\|r_i^{(s)}\|_2}, \frac{r_j^{(s)}}{\|r_j^{(s)}\|_2} \rangle, \quad \forall i, j \in \{1, \ldots, K\}$;

Compute edge matrix using Gumbel Softmax: $\mathcal{E}_{i,j}^{(s)} = 1$, if $S_{i,j}^{(s)} = \max_{k \in \mathcal{N}(i)} \sigma_{\text{gsr}}(S_{k,j}^{(s)})$, 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 $\hat{C}_{(s)}$

UNIVERSITY OF WATERLOO

Theoretical Considerations

Multiple Instance Learning Formulation

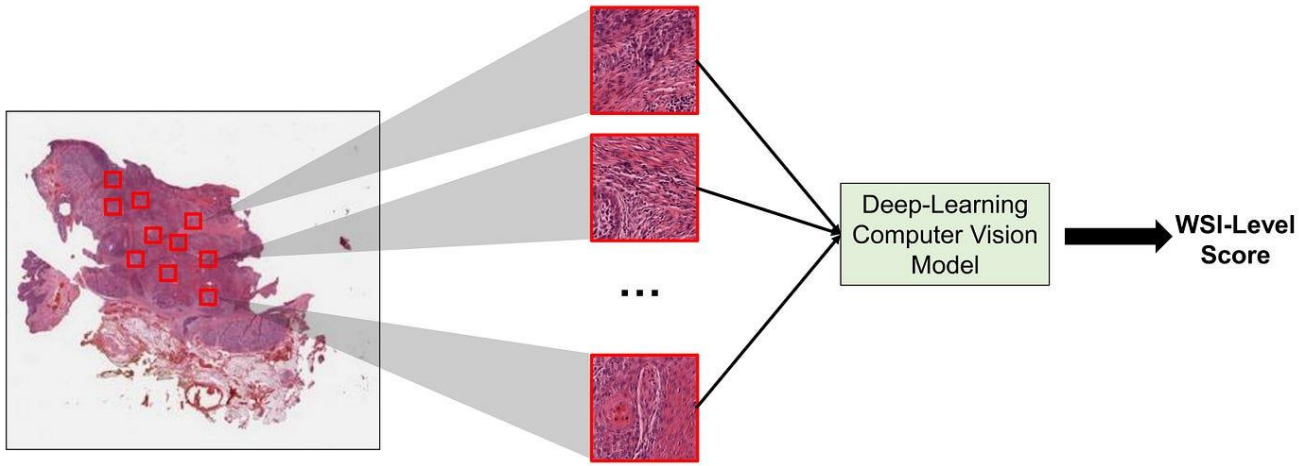
- Multiple Instance Learning organizes data into bags containing multiple instances, with labels provided only at the bag level

- WSI classification is formulated as an MIL problem, where each slide is considered a bag and its patches are instances
- According to the standard MIL assumption, the bag label Yi is described as:

$$Y_i = \begin{cases} 0 & \text{iif } \sum_{j=1}^{n_i} y_{i,j} = 0 \\ 1, & \text{otherwise} \end{cases}$$

In another approach, the bag label Yi can be determined using an aggregation function followed by a classifier as:

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



Journal of Pathology Informatics 15 (2024): 100403

Theoretical Considerations (Cont'd)

- **Vision Encoder**

- The WSI X(s) comprises a collection of *Np* patches or images, $P(s) = \left\{p_k^{(s)}\right\}_k^{Np}$, where *p(s)*, k denotes the k-th patch of patient s.
- The number of patches Np varies depending on the specific WSI and the individual characteristics of patient s.

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

- The patch-level embeddings are concatenated to represent the entire WSI of each patient *s*

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

ConvLSTM architecture

Theoretical Considerations (Cont'd)

- **Attention-based Deep Embedded Clustering**

✓ The probability of assigning each feature embedding $f_i^{(s)}$ to cluster k is determined using the Student's t-distribution as follows:

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

✓ 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 $T_{(s)} = \left\{t_{\{ik\}}^{(s)}\right\}$ is introduced based on the soft assignments $Q_{(s)} = \left\{q_{\{ik\}}^{(s)}\right\}$ to refine the clustering process.

✓ The target distribution emphasizes high-confidenceassignments and is computed as:

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

NeurIPS, 2017: 6000–6010

UNIVERSITY OF WATERLOO

Theoretical Considerations (Cont'd)

- **Attention-Based Representative Images**

✓ Let C_k denotes the set of image indices corresponding to cluster k. Therefore, each cluster k contains $N_k = |k|$images, and the feature embeddings of these images are extracted as follows:

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

✓ Then, the embeddings Z_k are mapped into query, key, and value representations using learnable linear projections:

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

✓ For each patch $i \in C_k$, an attention score is defined based on the element-wise product of its query

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

✓ Therefore, the softmax function is applied across all patches in the cluster to obtain normalized attention weight

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

Theoretical Considerations (Cont'd)

- **Attention-Based Representative Images (cont'd)**

✓ Therefore, a scalar importance score is computed for each patch using its corresponding value embedding

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

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

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

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

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

Theoretical Considerations (Cont'd)

- **Graph-Based Aggregation**
- **Constructing Graph**
  - ✓ The similarity between nodes is calculated using cosine similarity as:

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

  - ✓ where each element $S_{i,j}^{(s)}$ represents the similarity between image features $r_i^{(s)}$ and r_j^{s}. Therefore, an edge matrix $E_{i,j}^{(p)}$ is created by applying the Gumbel Softmax function. This process selects the most similar neighbors for each node, resulting in:

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

  - ✓ Therefore, the graph G(s) is defined by its set of nodes V(s) and edges E(s):

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

Theoretical Considerations (Cont'd)

- **Graph-Based Aggregation**
- **Graph Neural Network**

✓ For each layer $l = 0, 1, \ldots, L-1$, the features of node v can be updated as follows:

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

✓ The attention coefficient β_{vu}^{l} between nodes u and v at layer l is computed as:

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

✓ 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_{mean}^{s} is generated by applying global mean pooling, which aggregates all node representations

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

UNIVERSITY OF WATERLOO

Theoretical Considerations (Cont'd)

- **Visual Embedding Projection**

✓ To address this, the aggregated image embeddings h^{s}_{mean}, are transformed using a linear projection matrix Wc

✓ 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:

$$v'_{(s)} = h^{(s)}_{\text{mean}} \cdot \mathcal{W}_c \in \mathbb{R}^{d_{\text{model}}}$$

ConvLSTM architecture

Theoretical Considerations (Cont'd)

- **Loss Function**
- **Image Captioning Loss**

  - ✓ 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 Ct based on v' and the previously generated tokens $C1$ to $Ct-1$.
  - ✓ The loss for caption generation is calculated using the negative log-likelihood of the ground-truth captions:

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

  - ✓ Therefore, the total loss for the whole slide image captioning can be characterized as:

$$\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{Cap}} + \mathcal{L}_{\text{Clu}}$$

ConvLSTM architecture

Experimental Result Analysis

❖ LLMs

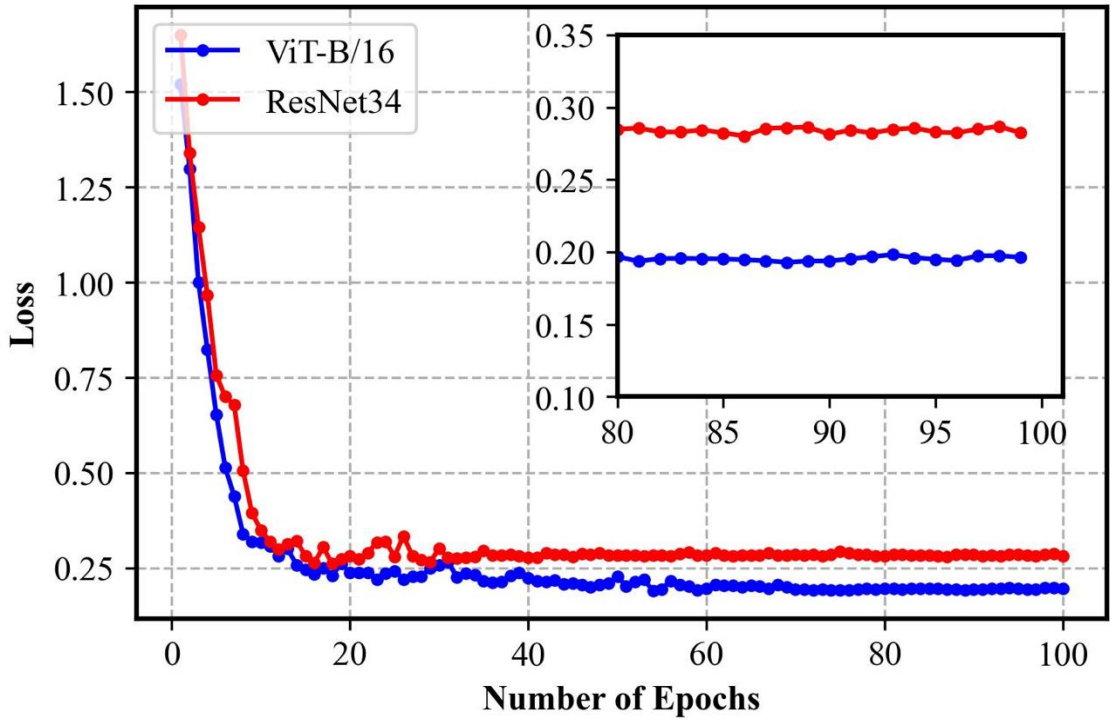
- ✓ BioMedGPT
- ✓ ClinicalT5-Base
- ✓ LLMamaV2-Chat
- ✓ BioGPT

❖ Hyper-parameter Settings

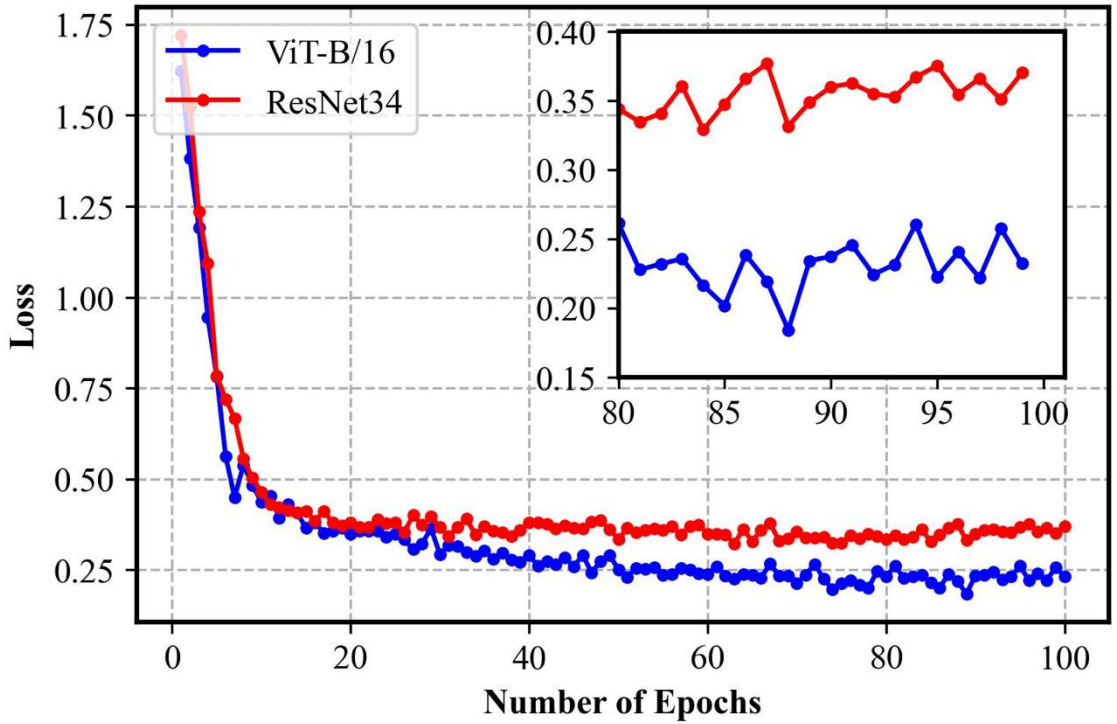
- ✓ We performed the experiments for 50 epochs to train the proposed architecture.
- ✓ Batch size was set to 16.
- ✓ Learning rate was set to 0.001
- ✓ Embedded dimension was 512
- ✓ Number of cluster k = 8 for classification and k= 50 for image captioning

Results: Loss Curve for Classification

- **Loss**



(a)

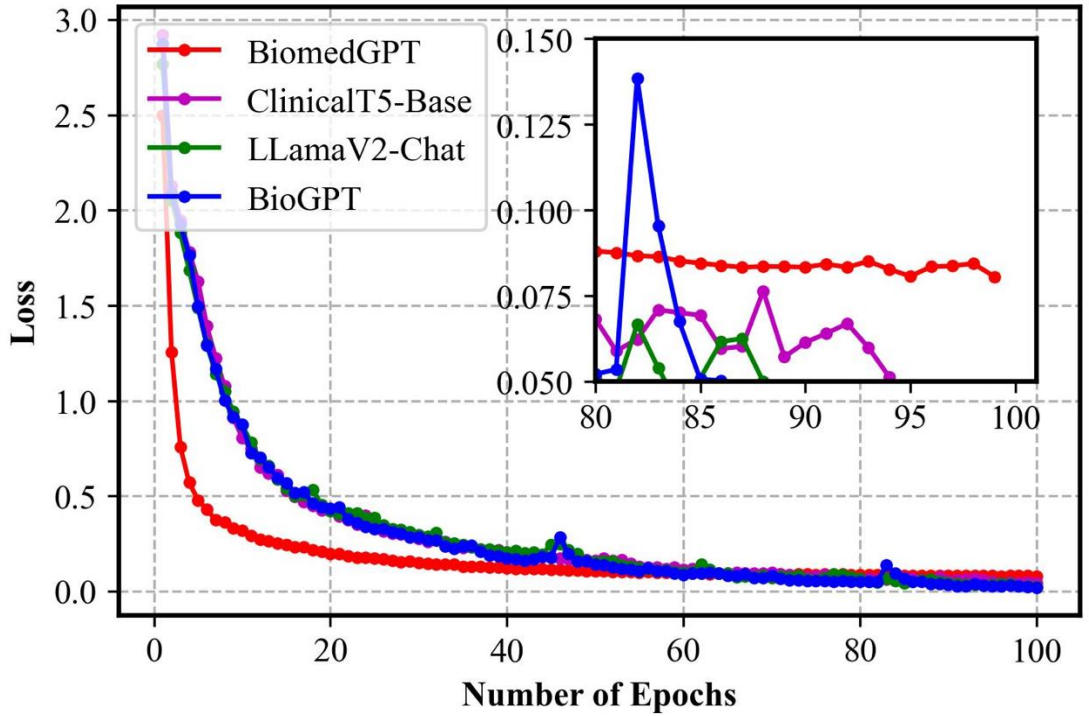


(b)

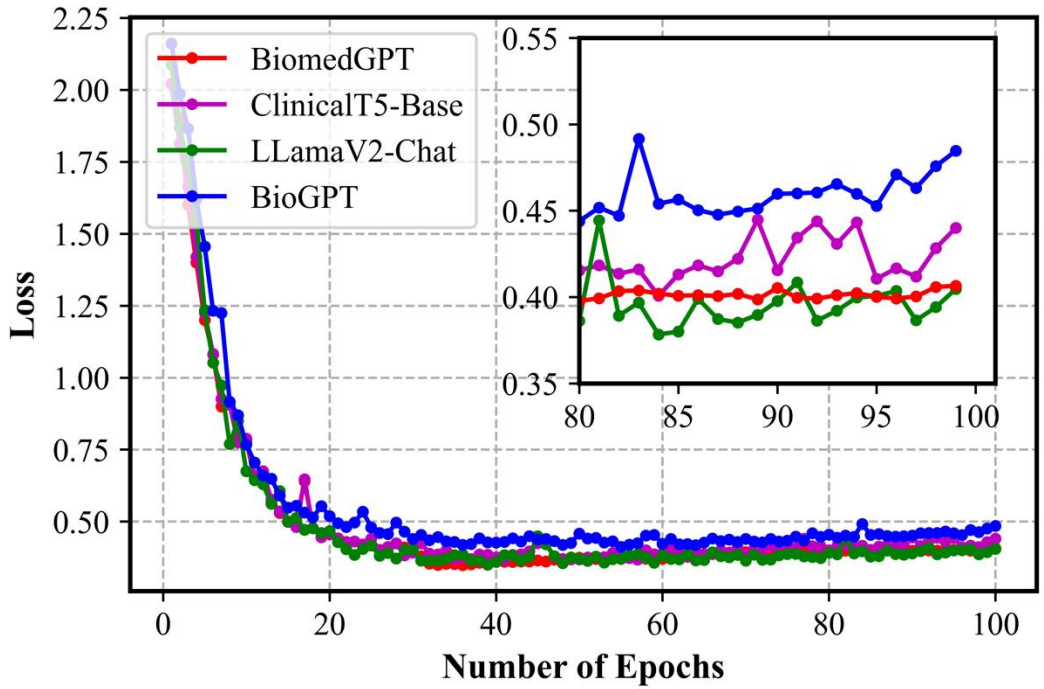
Loss curves on the BreakHis dataset using the proposed GNN-ViTCap with various feature extractors: (a) training loss, (b) validation loss.

Results: Loss Curve for Captioning

- **Loss**



(a)



(b)

Loss curves on the PatchGastric dataset using the proposed GNN-ViTCap with various LLMs: (a) training loss, (b) validation loss.

Results: Ablation Studies

- **With vs. Without Representative Images**

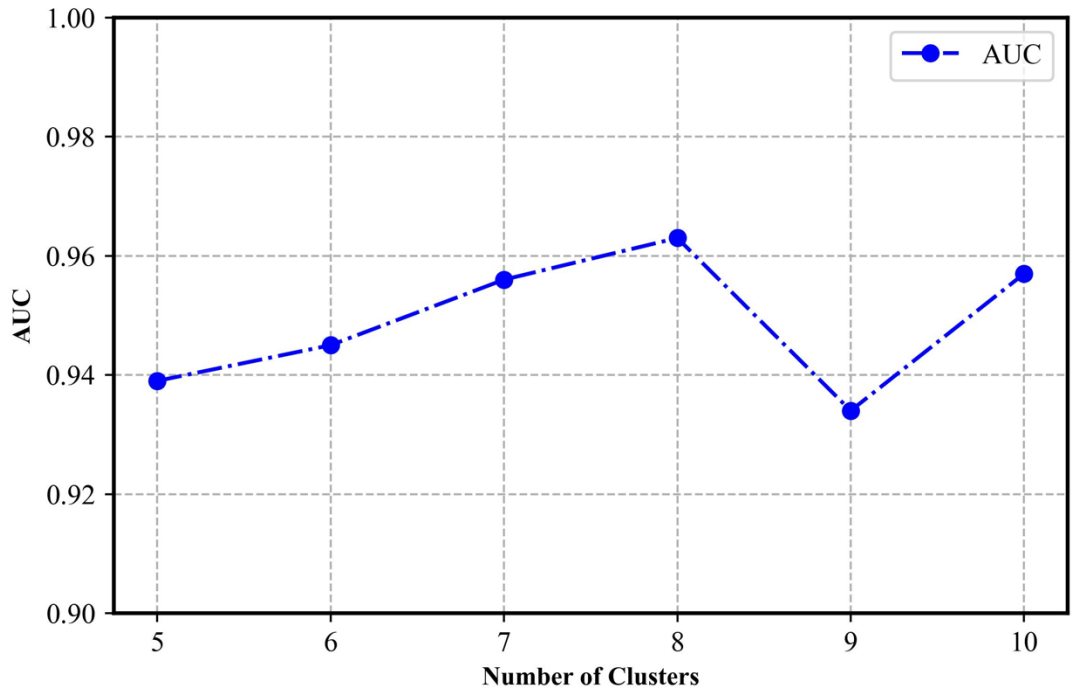
Performance evaluation with (*w/*) and without (*w/o*) representative images on the BreakHis and PatchGastric datasets.

	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

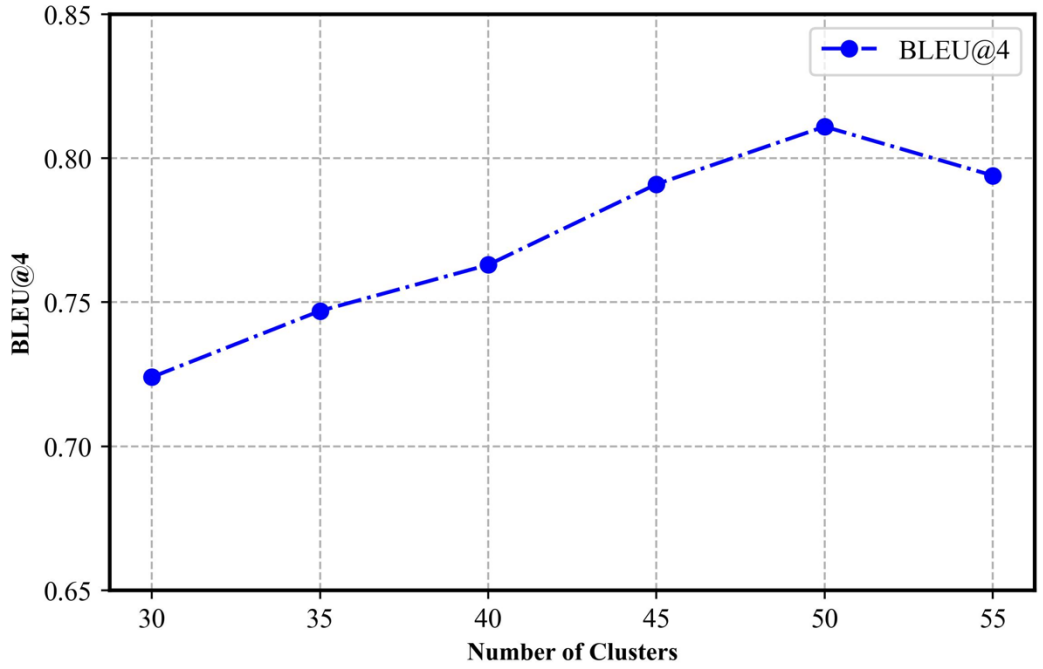
UNIVERSITY OF WATERLOO

Results: Ablation Studies (cont'd)

- **Representative Image Clusters**



(a)



(b)

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

Results: Ablation Studies (cont'd)

- **Graph Layers Effect**

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

UNIVERSITY OF WATERLOO