Available online www.jsaer.com

***Journal of Scientific and Engineering Research*, 2025, 12(1):82-87**

Research Article

ISSN: 2394-2630
CODEN(USA): JSERBR

Automated Diagnostic Text Generation from Breast Cancer Histology Images Using Large Language Models

Mohamed S.M. Geoda[1*], Tawfik Ezat Mousa[2]

[1,2]Department of Computer Technologies, Higher Institute of Science and Technology, Tobruk, Libya
*Email: Mhomadgg@gmail.com, Email: Tawfikezat@yahoo.com

Abstract: The emergence of large language models has paved the way for significant advances in medical image analysis and report generation. We used a BLIP-Image-Captioning-Base model that had already been trained to make diagnostic reports from breast cancer histopathological images in the BcHJC dataset. This dataset has 230 high-resolution microscopic images that have been labeled as normal, benign, or malignant, and each has a diagnostic text attached to it. The model was fine-tuned to learn strong links between visual features and textual descriptions. Our method produced an average BLEU score of 78.7%, which shows that the generated and reference texts were very similar. ROUGE metrics confirmed the effectiveness of the model, yielding average ROUGE-1, ROUGE-2, and ROUGE-L scores of 86.4%, 86.1%, and 86.4%, respectively. These results demonstrate the framework's ability to produce accurate, contextually relevant diagnostic texts that closely match the provided diagnostic texts.

Keywords: Breast cancer; Histology; LLMs; Diagnostics; Generative

1. Introduction

The field of histology has witnessed rapid growth, leading to an increase in case volume and complexity that surpasses the capacity of available specialists. Histologists have a lot more work to do than they can handle because of this imbalance. They have to look at and explain more and more studies in shorter and shorter amounts of time. Consequently, histologists often face extended working hours and are at heightened risk of specimen fatigue, factors that significantly contribute to diagnostic errors [1]. The challenge becomes even more pronounced during on-call periods, where urgent specimen studies demand immediate attention. To deal with these problems, there is a growing need for automated diagnostic tools that can reduce the amount of work that needs to be done, improve the accuracy of diagnoses, and speed up clinical workflows. This study focuses on the generation of multi-sentence diagnostic texts, a task of critical clinical importance that has garnered increasing attention in recent years. Existing methods in this field, which mostly use encoder-decoder architectures, are mostly based on approaches in image-video translation. However, unlike image-to-text tasks that typically produce concise, single-sentence descriptions, our research aims to generate detailed, coherent, and paragraph-length diagnostic texts. Therefore, we must develop models capable of capturing long-term dependencies and maintaining consistency across extended text outputs [2]. Large language models (LLMs) have gotten very good at many natural language processing tasks in recent years. These include logical reasoning, task execution, and using domain-specific information. These features render LLMs an attractive option for producing medical diagnostic texts. Another thing is that LLMs can reduce biases caused by certain natural samples being overrepresented in medical datasets because they have a large pre-trained knowledge base [3]. This makes diagnostic text creation more accurate and fair. Applying large language models (LLMs) to medical text generation tasks presents significant challenges due to the inherent differences between visual data

and textual representations. In this study, we propose a pre-trained model, BLIP-image-captioning-base, to address this challenge. We first process breast cancer images using an encoder to extract visual embeddings, which we then map to the feature space of the LLM. This mapping makes sure that all dimensions are the same and makes it simple to combine visual and textual data so that medical texts make sense. The framework leverages a dataset comprising breast tissue images paired with diagnostic reports collected from an El-jarrah clinic in Libya. The experimental results show that the framework does well on this dataset, which shows that it has the potential to make medical diagnostic text generation better and help doctors make decisions.

2. Literature Review

Wanga et al [4]. proposed a framework for medical report generation that highlights potential for future enhancements, including the integration of domain-specific knowledge and exploration of generalizability to other text generation tasks. The framework leverages vision-enabled large language models (LLMs), specifically Qwen1.5, to deliver competitive performance while optimizing resource efficiency. The results demonstrate that the proposed approach achieves comparable or superior performance to previous solutions on natural language generation metrics, with greater resource efficiency. The Qwen1.5 model outperformed GPT-2 and other larger LLM-based solutions in terms of the Bleu score. The study findings suggest that a well-designed, smaller LLM-based framework can effectively capture critical linguistic patterns and maintain coherence in medical report generation, outperforming larger transformer-based and LLM-based architectures in specific contexts. Li et al. [5] came up with a new framework called KARGEN. It combines large language models (LLMs) with a medical knowledge graph to make radiology report generation (R2Gen) better. The proposed framework shows cutting-edge performance on two benchmark datasets, with big gains seen across a range of evaluation metrics, some of which are clinically relevant. Notably, KARGEN did better than previous methods on almost all metrics. For example, on the MIMIC-CXR dataset, the BLEU-4 score went up by 4.5%. This framework stresses how important it is to combine disease-specific knowledge graphs with LLMs and how important it is to combine regional image features with knowledge-enhanced disease-related features to make the reports better in terms of both quality and clinical usefulness. Li et al. [6] proposed an innovative approach for automatic medical report generation (MRG) that utilizes a multimodal large language model. The model achieved an average Green score of 0.3 on the MRG task validation set and an average accuracy of 0.61 on the visual question answering (VQA) task validation set. The proposed model showed results that were similar to or even better than the baseline model in terms of both the Green score and VQA accuracy, even though it was trained on small amounts of data. These results show that the method works well and that combining Vision Transformer (ViT) with large language models (LLMs) is a good way to improve MRG and VQA performance. This integration serves as a bridge between image comprehension and language generation within the medical domain. The study also stresses how important it is to improve the model's decision-making process so that it can better handle real-life clinical situations and make sure that the results it produces are in line with the strict standards of medical practice. Liu et al. [7] developed a reward model, MRScore, aimed at aligning automated evaluations of radiology reports with human judgments through a score system. The study revealed that MRScore aligns well with expert radiologist assessments, surpassing conventional natural language generation (NLG) metrics and clinical evaluation scores. This strong association highlights the effectiveness of the suggested method, which integrates human-like evaluation criteria to improve the alignment of model outputs with human assessments. The architecture utilizes extensive training data produced by GPT, combining accepted and rejected GPT-generated reports to train large language models (LLMs) for generating MRScore as a model reward. This innovation signifies a substantial progress in AI's capacity to deliver precise and economical assessments for automated radiology report generating. Kapadnis et al. [8] suggested a new way to make radiology reports that uses a multi-modal large language model (MLLM) framework and a self-refining mechanism. This approach addresses challenges such as hallucination and enhances the accuracy of report generation. SERPENT-VLM does better than existing baselines like LlaVA-Med and BiomedGPT, reaching the highest level of performance on the X-ray and Radiology Objects in Context (ROCO) datasets. Additionally, the model demonstrates robustness against noisy images. SERPENT-VLM did much better than traditional non-LLM approaches, medical LLMs, and general-purpose vision-language models when tested using metrics like BLEU and BertScore. This is a big step forward in creating radiology reports (R2Gen).

3. Materials and Methods

The architecture we're proposing has a vision encoder that turns raw picture data into latent embeddings and a big language model that uses textual prompts to create relevant textual diagnostics that are aligned with the input images. The language module utilizes a transformer design that incorporates sequential information from tokens and image embeddings through an attention mechanism [9]. The language model and the vision encoder are first trained on a large dataset. They are then fine-tuned on the target dataset to make them work even better at creating accurate diagnostic text for tissue images.

Large Language Models

Large language models (LLMs) are a type of advanced neural network that can create text that sounds like it was written by a person and do a wide range of natural language processing (NLP) tasks [10]. For image-to-text generation tasks in this study, we utilized the blip-image-captioning-base model, a transformer-based LLM. The model leverages both visual and textual embeddings through a sophisticated attention mechanism to understand and describe the content of images. We fine-tuned the model on a specialized dataset comprising paired image-text samples to generate diagnostic text from breast tissue medical images. The BLIP model's vision encoder turns the raw image data into meaningful latent representations. The language model then combines these representations with textual prompts to make diagnostic reports that are accurate and make sense in the given context. This method makes sure that the generated text includes important pathological details that are present in the input images. It is a strong way to automate the reporting process in medical imaging. Figure 1 shows the proposed model for generating diagnostic texts for breast cancer tissue images.

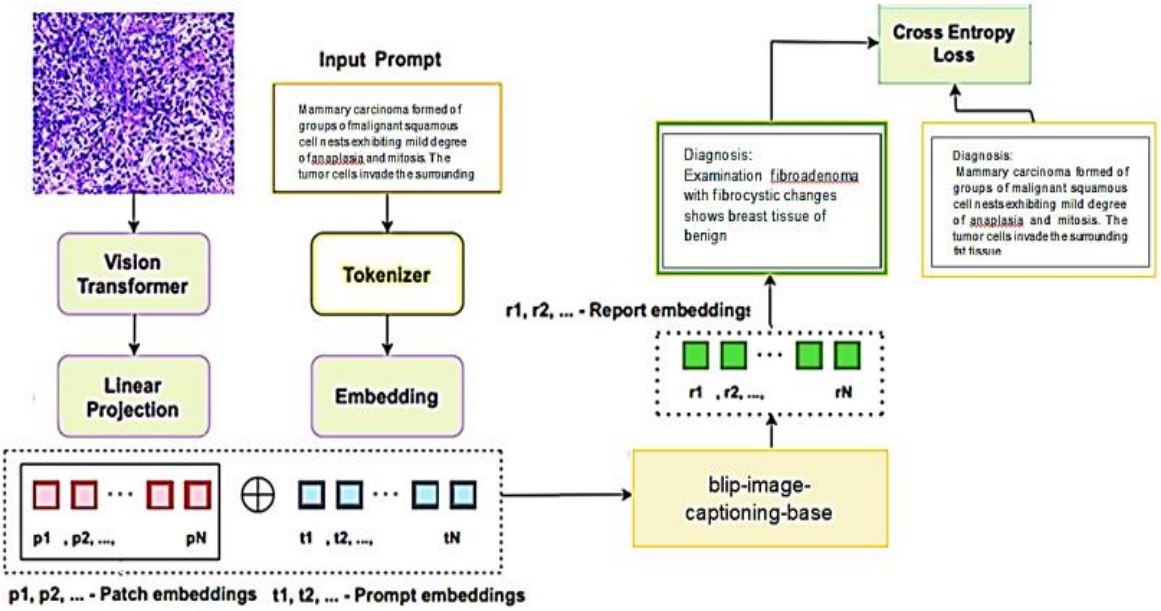

Figure 1: Model for generating diagnostic texts for breast cancer tissue images.

Database description

The Histopathology and Pathology Laboratory at El-Jarrah Clinic in Libya made the BcHJC dataset, which is made up of histopathological images of breast cancer. This dataset includes 230 high-resolution microscopic images of breast tumor tissues collected from various patients. We categorize the images into three distinct classes: normal, benign, and malignant. Textual descriptions derived from diagnostic histopathology reports annotate each image. Figure 2 presents representative samples from the BcHJC dataset, illustrating examples from the benign, malignant, and normal classes.

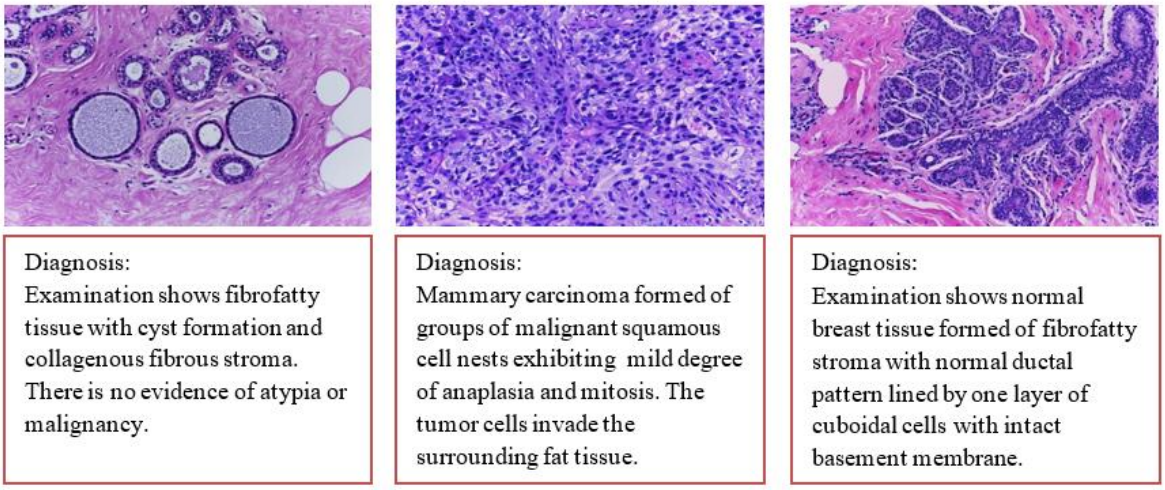

Figure 2: Representative samples from the BcHJC dataset for benign, malignant, and normal categories.

4. Results & Discussion

Implementation Details

Utilizing the Hugging Face Transformers library, we integrated a pre-trained Vision Transformer alongside a large language model to develop our framework. We performed the model optimization using the AdamW optimizer, fine-tuning the learning rate to 1e-5. We conducted training on a robust computational setup consisting of six Nvidia A800 GPUs, using a batch size of 5. The model underwent rigorous training over 10 epochs using the designated dataset, ensuring thorough learning and performance evaluation.

Results

Our proposed framework demonstrates an efficient way to generate medical reports by leveraging the generative capabilities of LLM. The proposed model was adapted, tuned, and trained on a given breast tissue image dataset and corresponding diagnostic texts, with fine-tuning of the hyperparameters to improve the model's ability to interpret medical images to produce accurate and relevant diagnostic texts that are consistent with human assessments. Despite the limited data constraints, our research yielded commendable results and exceptional performance on Natural Language Generation (NLG) metrics, especially the BLEU score, which assesses the match between the generated and reference texts based on phrase similarity. In our study, we used the average BLEU score, which achieved a high percentage, indicating that the generated texts showed 78.7% similarity to the reference texts based on word and phrase matching. The average ROUGE-1 score was used, which measures the percentage of overlapping words between the generated text and the reference text. The model achieved an accuracy of about 86.4%. The ROUGE-2 metric, which measures the average percentage of consecutive word pairs that are identical, achieved an approximate matching percentage of 86.1%. We used the average ROUGE-L score to evaluate the longest common subsequence between the generated texts and the reference texts, and obtained a percentage of 86.4%. The promising outcomes of this study underscore the efficacy of leveraging large language models (LLMs) for medical image interpretation and report generation. The fine-tuning and optimization of the BLIP-Image-Captioning-Base model on the BcHJC dataset enabled it to produce clinically relevant and contextually accurate diagnostic texts. Despite the dataset's relatively small size, the model achieved exceptional alignment with reference texts, as evidenced by the high BLEU and ROUGE scores.

A significant factor contributing to these results is the high-quality alignment between the visual data and textual annotations in the BcHJC dataset. This alignment ensures that the model learns robust correlations between histopathological features and corresponding diagnostic language, facilitating accurate and meaningful text generation. Additionally, the model's adaptability and its ability to generalize across diverse input images further validate its potential for clinical applications. Table 1 shows a summary of these results for the dataset used.

Table 1: Results of the proposed method

Metric	Definition	Result (%)
BLEU	Measures similarity of n-grams between generated and reference texts	78.7
ROUGE-1	Measures the percentage of overlapping words	86.4
ROUGE-2	Measures the percentage of matching consecutive word pairs	86.1
ROUGE-L	Measures the longest common subsequence similarity	86.4

Figure 3. shows the ability of the proposed model to predict diagnostic texts for samples compared to the original diagnostic texts.

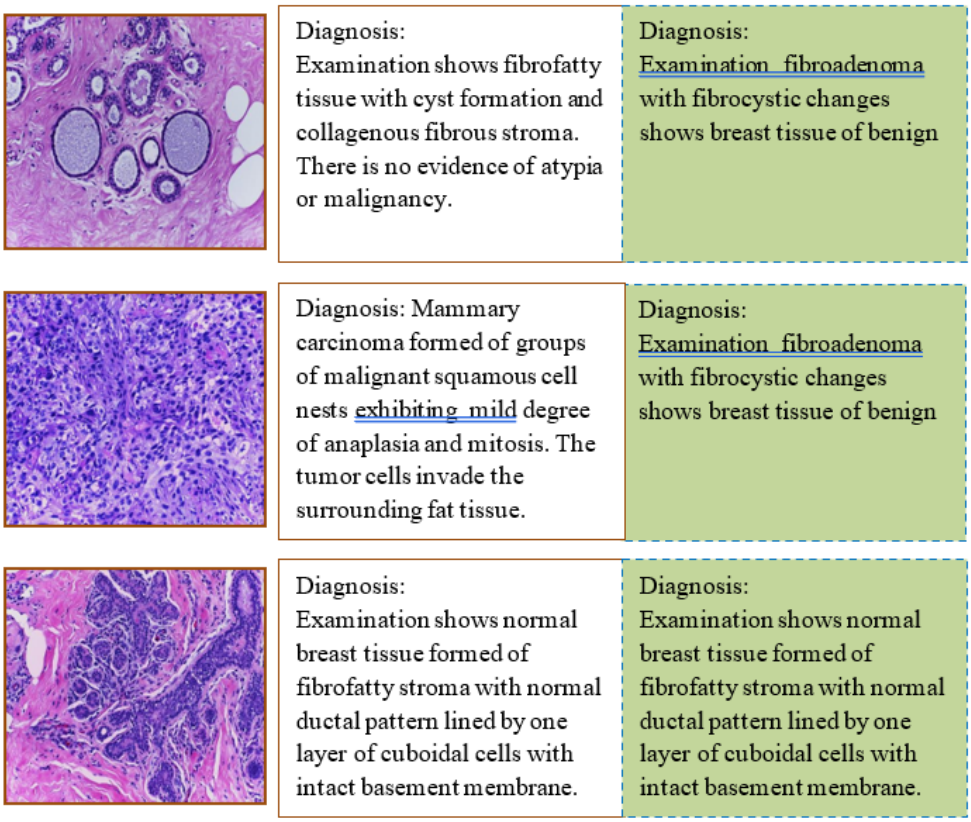

Figure 3: shows the predicted diagnostic texts for samples in comparison to the original diagnostic texts.

5. Conclusion

This study demonstrates the potential of large language models (LLMs), particularly the BLIP-Image-Captioning-Base model, to revolutionize medical report generation through the integration of advanced natural language processing and medical imaging analysis. By fine-tuning the model on the BcHJC dataset, consisting of high-resolution breast cancer histopathological images and corresponding diagnostic reports, we achieved significant performance metrics, including high BLEU and ROUGE scores. These results highlight the model's ability to produce clinically accurate, contextually relevant, and human-like diagnostic texts. The findings suggest a promising pathway for employing LLMs in medical diagnostics, with the potential to enhance diagnostic precision and reduce the workload of medical professionals. The promising results of this study open avenues for further research, including the exploration of larger and more diverse datasets to improve generalizability, the integration of multi-modal approaches for enhanced diagnostic accuracy, and the refinement of LLM architectures tailored specifically for medical applications. Moreover, the implementation of such frameworks in clinical practice could significantly enhance diagnostic efficiency and accuracy, thereby improving patient outcomes.

References

[1]. Ruiter, D. J., Roald, B., Underwood, J., Prat, J., & UEMS Section of Pathology/European Board of Pathology. (2004). Histopathology training in Europe: a lesson for other specialties?. Virchows Archiv, 444, 278-282.

[2]. Pavlopoulos, J., Kougia, V., Androutsopoulos, I., & Papamichail, D. (2022). Diagnostic captioning: a survey. Knowledge and Information Systems, 64(7), 1691-1722.

[3]. Yan, L. K., Li, M., Zhang, Y., Yin, C. H., Fei, C., Peng, B., ... & Niu, Q. (2024). Large language model benchmarks in medical tasks. arXiv preprint arXiv:2410.21348.

[4]. Hamza, A., & Kim, S. T. (2024). Resource-Efficient Medical Report Generation using Large Language Models. arXiv preprint arXiv:2410.15642.

[5]. Li, Y., Wang, Z., Liu, Y., Wang, L., Liu, L., & Zhou, L. (2024, October). Kargen: Knowledge-enhanced automated radiology report generation using large language models. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 382-392). Cham: Springer Nature Switzerland.

[6]. Li, S., Xu, B., Luo, Y., Nie, D., & Zhang, L. (2024). ViT3D Alignment of LLaMA3: 3D Medical Image Report Generation. arXiv preprint arXiv:2410.08588.

[7]. Liu, Y., Wang, Z., Li, Y., Liang, X., Liu, L., Wang, L., & Zhou, L. (2024). MRScore: Evaluating Radiology Report Generation with LLM-based Reward System. arXiv preprint arXiv:2404.17778.

[8]. Kapadnis, M. N., Patnaik, S., Nandy, A., Ray, S., Goyal, P., & Sheet, D. (2024). SERPENT-VLM: Self-Refining Radiology Report Generation Using Vision Language Models. arXiv preprint arXiv:2404.17912.

[9]. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need (2023), https://arxiv.org/abs/1706. 03762.