Image-to-Text Summarization using DeiT & GPT-2

ChitraRupa Matimpati
Assistant professor, Department of CSE
Sree Venkateswara College of Engineering
North Rajupalem,Nellore
Research Scholar JNTU Anantapuramu, Anantapuramu,India
chitrarupam@gmail.com
ORCID-0009-0007-2972-8571

Ramani Kasarapu
Professor, Dept Datascience,
Mohan babu University (Earth while Sree Vidyanikethan Engineering College),
Tirupati,Andhra Pradesh, India,
ramanidileep@yahoo.com.

Abstract— This work presents a visual summarization framework for a medical image captioning system that leverages a Data-efficient Image Transformer (DeiT) as the visual encoder and GPT-2 as the language decoder within an encoder-decoder architecture. The model is trained and evaluated on the Indiana Chest X-ray dataset to automatically generate radiology summaries from medical images. Our objective is to improve clinical text generation by summarizing visual diagnostic data into concise and semantically meaningful captions. Evaluation using standard Natural Language Processing (NLP) summarization metrics demonstrates the effectiveness of our model. Specifically, our approach achieves a BLEU score of 0.0680, ROUGE-1 of 0.2713, ROUGE-2 of 0.0981, ROUGE-L of 0.2268, and a METEOR score of 0.0940, outperforming a ViT-GPT2 baseline in all major summarization metrics. This work demonstrates the potential of integrating vision transformers with autoregressive language models for medical image summarization, laying a foundation for AI-assisted radiology reporting systems.

Keywords— Chest X-ray, Image Captioning, Vision-to-Text summarization, Transformers, DeiT, GPT-2, Medical Imaging and Deep Learning.

I. INTRODUCTION

The interpretation of chest X-ray images and the subsequent generation of accurate and comprehensive radiological reports are critical tasks in clinical practice. However, these processes are often time-consuming and require significant expertise. The increasing volume of medical imaging data necessitates the exploration of automated tools that can assist radiologists in their workflow.

Recent advancements in deep learning, particularly in the domains of computer vision and natural language processing, have paved the way for sophisticated image captioning systems. Transformer-based architectures, such as Vision Transformers (ViTs) and large language models (LLMs), have demonstrated remarkable capabilities in understanding and generating complex data.

This paper presents the implementation of a Vision Encoder-Decoder model leveraging a pre-trained

DeiT (Data-efficient Image Transformer) as the image encoder and a pre-trained GPT-2 as the language decoder for the specific task of generating findings from chest X-ray images. The study utilizes the publicly available Indiana University Chest X-ray dataset. The paper outlines the data processing steps, the model architecture, the training procedure, and the evaluation strategy employed. The potential of such a system to contribute to the field of medical image analysis and report generation is discussed.

II. Related Work

Automated image captioning has been an active area of research, with early approaches often combining Convolutional Neural Networks (CNNs) for feature extraction with Recurrent Neural Networks (RNNs) for sequence generation [1]. The introduction of attention mechanisms significantly improved the relevance and coherence of generated summary.

In the medical domain, efforts have been made to adapt these techniques for radiology report generation. Studies have explored various architectures and strategies to capture the specific nuances of medical language and the detailed observations required in radiological findings [2].

The emergence of Transformer architectures has marked a significant shift in both computer vision and natural language processing. Vision Transformers, like DeiT, have shown competitive performance in image classification and feature extraction. Large language models, such as GPT-2, have demonstrated impressive text generation capabilities [3].

The presented work builds upon these advancements by exploring the synergy of a Vision Transformer (DeiT) and a powerful language model (GPT-2) within an Encoder-Decoder framework for the task of generating findings from chest X-ray images [4].

The issue of automatically creating radiological impressions by summarising textual radiology data was initially

investigated by Zhang et al. (2018), who demonstrated that an upgraded pointer generator method obtains significant overlap between human sources [5].

Zhang et al., 2020 created a broad approach for assessing a generated summary's factual accuracy that is accomplished by employing a knowledge retrieval module to automatically factcheck it versus its reference [6].

In order to generate reports, Yin et al. [7] suggested a unique architecture that makes use of a hierarchical recurrent neural network (HRNN) with a soft attention mechanism. To create a thorough description of the trained image, they combined the subject matching strategy with the image labelling technique.

Meghana Das and Pragna Parmitha Sambodhi, explores the key challenges in medical text and image processing. It highlights a significant gap in the development of robust applications capable of detecting and interpreting prescriptions, X-rays, and clinical notes. The authors emphasize the need for advanced data technologies to enhance the accuracy and reliability of information extraction from medical images and textual documents [8].

Akash Ghosh and Nohit Singh propose an innovative encoder-decoder architecture using the BART model for summarizing multimodal clinical documents. Their approach integrates contextual multimodal attention within both the encoder and decoder. However, the study identifies a key limitation: the absence of cross-attention in the decoder, suggesting that incorporating it could significantly improve model performance [9].

Shaik Rafi and Ranjita Das explore abstractive summarization using multimodal information with attention-based sequence models. Despite leveraging attention mechanisms, the model achieved only 45% ROUGE score accuracy. The authors emphasize the need to improve performance through more advanced multimodal techniques to better capture and summarize complex clinical data [10].

Jean-Benoit Delbrouck presents a study on multimodal radiology report summarization using a Bidirectional Gated Recurrent Unit (Bi-GRU) model. The model achieved a moderate accuracy of 54%, but the research underscores the need for improved accuracy through the integration of more advanced multimodal techniques [11].

Rafi and Ranjita Das explore abstractive summarization using multimodal information with sequence-to-sequence models enhanced by attention mechanisms. Despite this, the model achieved only a 45% ROUGE score, indicating moderate performance. The study highlights the need for adopting more advanced multimodal techniques to improve summarization accuracy [12].

This approach leverages the strengths of both architectures to learn the complex mapping between visual features and descriptive language in the medical domain.

III Methodology

The proposed model follows an encoder-decoder architecture, where DEiT acts as the vision encoder and GPT-2 is used as the text decoder. This structure enables efficient extraction of medical image features and generation of descriptive medical reports.

Model Components:

- Encoder: DEiT (facebook/deit-base-distilled-patch16-224)
- Decoder: GPT-2
- Tokenizer: Tokenizer is used for text preprocessing.

A. Dataset:

The study utilizes the Indiana University Chest X-ray dataset, which contains a large collection of chest X-ray images and their corresponding radiology reports. The dataset includes two primary CSV files: indiana_projections.csv and indiana_reports.csv. The methodology involves the following data processing steps:

1. Data Loading: Loading the indiana_projections.csv and indiana_reports.csv files using the pandas library.
2. Data Merging: Linking image filenames from indiana_projections.csv with the corresponding "findings" sections from indiana_reports.csv based on the unique patient identifier (uid).
3. Data Cleaning: Filtering out entries with missing or non-textual "findings."
4. DataFrame Creation: Creating a consolidated pandas DataFrame (images_captions_df) with columns for image file paths (imgs) and their associated captions (captions).
5. Image Path Construction: Prepending the base directory of the normalized images (/kaggle/input/chest-xrays-indiana-university/images/images_normalized/) to the image filenames to create full file paths.
6. Data Splitting: Dividing the images_captions_df into training, validation, and testing sets using train_test_split from scikit-learn, with a 20% split for the initial test set and a subsequent 20% split of the training data for the validation set. A random_state of 42 is used for reproducibility.

B. Model Architecture:

The core of the system is a Vision Encoder-Decoder Transformer model implemented using the transformers library.

1. Encoder (DeiT): A pre-trained "facebook/deit-base-distilled-patch16-224" model is used as the image encoder. The AutoFeatureExtractor associated with this pre-trained model is employed to process input images into pixel values.
2. Decoder (GPT-2): A pre-trained "gpt2" model is used as the language decoder. The AutoTokenizer associated with this model is used to tokenize the captions. A padding token is explicitly set to be the same as the end-of-sequence token (eos_token).

3. Vision Encoder-Decoder Model: The VisionEncoderDecoderModel is initialized using the pre-trained DeiT encoder and GPT-2 decoder checkpoints. Key configuration parameters are set:
o decoder_start_token_id: Set to the beginning-of-sequence token ID of the GPT-2 tokenizer.
o pad_token_id: Set to the padding token ID of the GPT-2 tokenizer.
o num_beams: Set to 4 for beam search during inference.

C. Training Procedure:

A custom LoadDataset class, inheriting from torch.utils.data.Dataset, is implemented to handle the loading and preprocessing of image-caption pairs for training and evaluation. For each sample:
1. The image is loaded using PIL and converted to RGB format.
2. The image is processed using the DeiT feature extractor to obtain pixel values.
3. The corresponding caption is tokenized using the GPT-2 tokenizer with a maximum length of 128 tokens, truncation enabled for longer captions, and padding applied to ensure consistent input lengths.
4. The dataset returns a dictionary containing the processed pixel values and the tokenized labels (input IDs of the caption).
The training is performed using the Seq2SeqTrainer from the transformers library, with the following training arguments (Seq2SeqTrainingArguments):
• output_dir: "image-caption-generator" for saving model outputs.
• evaluation_strategy: "epoch" to evaluate at the end of each training epoch.
• per_device_train_batch_size: 16.
• per_device_eval_batch_size: 16.
• learning_rate: 0.00005.
• weight_decay: 0.01.
• num_train_epochs: 8.
• save_strategy: 'no' to avoid saving intermediate checkpoints.
• report_to: 'none' to disable external reporting.

D. Evaluation Metrics:

The performance of the trained model is evaluated on the test dataset using standard natural language generation metrics:
1. BLEU (Bilingual Evaluation Understudy): Calculated using sentence_bleu from the nltk.translate.bleu_score library with a smoothing function [13].
2. ROUGE (Recall-Oriented Understudy for Gisting Evaluation): ROUGE-1, ROUGE-2, and ROUGE-L scores are computed using the rouge_scorer library [14].
3. METEOR (Metric for Evaluation of Translation with Explicit Ordering): Calculated using the Meteor class from the pycocoevalcap.meteor.meteor library [15].

The evaluation process involves iterating through the test dataset, generating captions for each image using the trained model in evaluation mode (model.eval() and torch.no_grad()), and then comparing these generated captions with the ground truth captions using the aforementioned metrics. The average scores for each metric across the entire test set are reported. The generation process utilizes beam search with a beam size of 4 and a maximum output length of 17 tokens.

IV Results

The quantitative results of the evaluation, including the average BLEU, ROUGE-1, ROUGE-2, ROUGE-L, and METEOR scores obtained on the test set. Qualitative examples showcasing original chest X-ray images alongside their ground truth and generated findings would also be included to provide a visual understanding of the model's performance.

Table 1. performance metrics for image and text summarization

Metric	DEiT-GPT2	ViT-GPT2
BLEU	0.0680	0.0617
ROUGE-1	0.2713	0.2509
ROUGE-2	0.0981	0.0810
ROUGE-L	0.2268	0.2075
METEOR	0.0940	0.0854

From above table 1 displays proposed model demonstrates an improvement over ViT-GPT2 in key NLP evaluation metrics such as BLEU and ROUGE. The model is trained on the Indiana University Chest X-ray dataset and evaluated using standard NLP evaluation metrics.

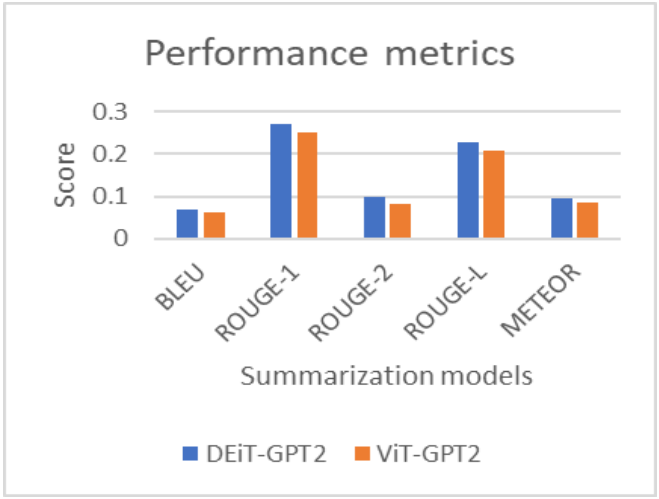

Fig 1. Comparison of different evaluation metrics for summarization of multimodal data.

Above fig 1 shows the improvement in the performance of two models such as ViT-GPT2 and DEit-GPT2, proposed model has an efficiency of them summarize the radiology report.

This paper presents a multi-modal deep learning model for automated chest X-ray report generation. The model integrates Data-efficient Image Transformer (DEiT) as an encoder and GPT-2 as a decoder, enabling accurate and efficient generation of textual summaries from radiology reports based on medical imaging data.

V Conclusion

This paper presents a practical implementation of a Vision Encoder-Decoder Transformer model for generating findings from chest X-ray images. The utilization of pre-trained DeiT and GPT-2 models demonstrates the potential of leveraging state-of-the-art deep learning techniques for medical image captioning. Future work could focus on exploring larger and more diverse datasets, experimenting with different Transformer architectures and training strategies, incorporating medical domain knowledge, and conducting thorough evaluations with clinical experts to assess the real-world utility of such systems.

IV References

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

[2] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

[3] Touvron, H., Cordier, N., Douze, M., Massa, F., Sablayrolles, R., & Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. arXiv preprint arXiv:2012.12877.

[4] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8),

[5] Yuhao Zhang, Daisy Yi Ding, Tianpei Qian, Christopher D. Manning, and Curtis P. Langlotz. 2018. Learning to summarize radiology findings. In EMNLP 2018 Workshop on Health Text Mining and Information Analysis.

[6] Yuhao Zhang, Derek Merck, Emily Tsai, Christopher D. Manning, and Curtis Langlotz. 2020. Optimizing the factual correctness of a summary: A study of summarizing radiology reports. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5108–5120, Online. Association for Computational Linguistics.

[7] Yin, C. *et al*. Automatic generation of medical imaging diagnostic report with a hierarchical recurrent neural network. In *2019*
IEEE International Conference on Data Mining (ICDM) 728–737.

[8] Meghana das, Pragna parmitha Sambodhi," Challenges of Medical Text and Image Processing", International conference on Advancements in Smart, Secure and intelligent Computing, IEEE,2022

[9] Akash Ghosh, Nohit Singh, "Towards Summarization of Multi model clinical documents" Association of computational Linguistics, volume 1 2024.

[10] Riya Mol Raji," Abstractive Text Summarization for Multimodal Data", International Conference on Computing, Communication, Security and Intelligent System, IEEEE,2022.

[11] Jean-Benoit Delbrouck," Multi-Model Radiology Report Summarization", Association with Computational Linguistics,2021.

[12] Shaik Rafi, Ranjita Das, "Abstractive Summarization using Multi-model Information", International conference on soft computing & Machine Intelligence, IEEE,2023.

[13] Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the association for computational linguistics (pp. 311-318).

[14] Lin, C. Y. (2004). Rouge: A package for automatic evaluation of summaries. In Text summarization branches out workshop at acl (pp. 1-8).

[15] Banerjee, S., & Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization (pp. 65-72).