DS@BioMed at ImageCLEFmedical Caption 2024: Enhanced Attention Mechanisms in Medical Caption Generation through Concept Detection Integration

Nhi Ngoc-Yen Nguyen[1,2], Huy Le Tu[1,2], Phuong Dieu Nguyen[1,2], Tan Nhat Do[1,2], Triet Minh Thai[3] and Thien B. Nguyen-Tat[1,2,*]

[1]University of Information Technology, Ho Chi Minh City, Vietnam
[2]Vietnam National University, Ho Chi Minh City, Vietnam
[3]Oxford University Clinical Research Unit, Ho Chi Minh City, Vietnam

Abstract

Purpose: Our study presents an enhanced approach to medical image caption generation by integrating concept detection into attention mechanisms.
Method: This method utilizes sophisticated models to identify critical concepts within medical images, which are then refined and incorporated into the caption generation process.
Results: Our concept detection task, which employed the Swin-v2 model, achieved an F1 score of 0.58944 on the validation set and 0.61998 on the private test set, securing the third position. For the caption prediction task, our BEiT+BioBart model, enhanced with concept integration and post-processing techniques, attained a BERTScore of 0.60589 on the validation set and 0.5794 on the private test set, placing ninth.
Conclusion: These results underscore the efficacy of concept-aware algorithms in generating precise and contextually appropriate medical descriptions. The findings demonstrate that our approach considerably improves the quality of medical image captions, highlighting its potential to enhance medical image interpretation and documentation, thereby contributing to improved healthcare outcomes.

Keywords

Medical Caption Generation, Multimodal Learning, Concept Detection, ImageCLEF 2024

1. Introduction

The rapid growth of deep learning techniques has profoundly influenced various sectors, notably medical imaging [1]. Among these advancements, using neural networks in radiology has garnered considerable attention due to its potential to enhance diagnostic accuracy and efficiency [2]. A particularly intriguing development in this field is the automatic generation of medical captions from radiology images [3]. This innovation aims to assist radiologists by providing preliminary interpretations and streamlining clinical documentation. Medical caption generation transforms visual information from radiological images into coherent, clinically valuable language descriptions. This process is inherently challenging due to the complexity and diversity of medical images, the need for precise and context-aware descriptions, and the necessity to incorporate domain-specific knowledge [3, 4, 5].

Traditional systems often fall short of these requirements, leading to the development of advanced attention mechanisms that can more effectively capture and interpret the intricate details found in radiological images. Recent research shows that integrating concept detection into caption generation algorithms improves performance [6, 7]. Concept detection involves identifying and categorizing critical visual elements in an image, such as anatomical structures, pathological findings, and medical devices. By incorporating these detected concepts into the caption generation process, models can produce more accurate and contextually relevant descriptions. One of the advancements in this field is the ImageCLEF campaign, an annual multimodal machine learning competition established in 2003.

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09-12, 2024, Grenoble, France
*Corresponding author.
21521231@gm.uit.edu.vn (N. N. Nguyen); 21522173@gm.uit.edu.vn (H. L. Tu); 21520091@gm.uit.edu.vn (P. D. Nguyen); 21522575@gm.uit.edu.vn (T. N. Do); triettm@oucru.org (T. M. Thai); thienntb@uit.edu.vn (T. B. Nguyen-Tat)

ImageCLEF [8] fosters advancements in multimedia processing, including computer vision, image analysis, classification, and retrieval in multilingual and multimodal contexts. In ImageCLEF 2024 [8], participants engaged in the ImageCLEFmedical Caption task [9], which included two subtasks: concept detection, aiming to identify critical elements within medical images, and caption prediction, focused on generating descriptive texts based on identified concepts. Concept detection aims to associate biomedical images with relevant medical concepts, thereby enhancing diagnostic notes by identifying key concepts that should be included in preliminary reports. Moreover, it facilitates the efficient organization and retrieval of medical images by indexing them according to related concepts. Caption prediction, or diagnostic captioning, remains a complex research challenge intended to support the diagnostic process by providing preliminary reports, rather than replacing physicians. This approach aids experienced clinicians in managing high volumes of daily medical examinations more swiftly and efficiently, while also reducing the likelihood of clinical errors among less experienced clinicians.

Our findings underscore that integrating concept detection enhances the efficacy of attention mechanisms and yields more coherent and diagnostically valuable captions. This research advances the development of intelligent technologies aimed at supporting radiologists in clinical practice, thereby elevating the standard of patient care. Section 2 provides a comprehensive review of pertinent literature. Section 3 outlines our dataset, while Section 4 describes our proposed methodology and presents experimental results. In Section 5, we discuss the conclusions drawn from our findings and outline avenues for future research. Our objective is to contribute to the fields of medical imaging and natural language processing by enhancing the capabilities of medical caption generation, thus paving the way for further advancements in automated reporting and medical data interpretation.

2. Background and RelatedWorks

2.1. Former Medical Datasets

Medical imaging has been a focal point in the application of deep learning, benefiting from the availability of comprehensive datasets. Early datasets such as the NIH ChestX-ray14 [10] provided a large collection of chest radiographs annotated with disease labels, facilitating advancements in image classification and disease detection tasks. The MIMIC-CXR dataset [11], developed by Johnson et al., further enriched the field by offering not only radiographic images but also paired radiology reports, enabling research in image-to-text generation. These datasets have been pivotal in training and validating deep learning models, providing the groundwork for more sophisticated tasks such as medical caption generation and concept detection.

2.2. Related Work Concept Detection

Concept detection in medical imaging involves identifying and categorizing essential visual elements such as anatomical structures, pathological findings, and medical devices. This task is crucial for generating accurate and contextually relevant medical captions. Early methods primarily relied on traditional machine learning techniques, which often struggled with the complexity and variability of medical images (e.g., SVMs (support vector machines), random forests, and k-nearest neighbors). However, recent advancements in deep learning, particularly CNNs (convolutional neural networks), have improved the accuracy of concept detection. Notable CNN architectures such as ResNet50 [12] and EfficientNet [13] have demonstrated substantial improvements in detecting and classifying visual elements in medical images.

Recently, Transformer-based models have been increasingly applied to concept detection due to their ability to capture long-range dependencies and contextual information. Notable examples include ViT (Vision Transformer) [14], BEiT (Bidirectional Encoder representation from Image Transformers) [15], and Swin Transformer [16]. These models provide robust feature representations and have shown promise in enhancing the accuracy and interpretability of medical image analysis.

2.3. Related Work Caption Prediction

Caption prediction, or diagnostic captioning, involves generating descriptive text that accurately summarizes the medical content of an image. This task extends beyond simple image annotation, requiring models to produce coherent and clinically meaningful narratives. Traditional captioning methods often used template-based approaches, which lacked flexibility and adaptability to different medical contexts. With the advent of deep learning, particularly sequence-to-sequence models and attention mechanisms, more sophisticated captioning systems have been developed.

For example, Jing et al. proposed a hierarchical LSTM (Long Short-Term Memory) [17] model combined with a co-attention mechanism to generate detailed radiology reports from medical images. Their model effectively captured the hierarchical structure of medical reports, producing more detailed and contextually appropriate captions [3].

The introduction of Transformer models specifically designed for the medical domain has advanced the field of medical image captioning. Transformers, particularly models like BioBERT (Bidirectional Encoder Representations from Transformers for Biomedical Text Mining) [18], have demonstrated exceptional capabilities in understanding and generating biomedical text due to their ability to handle complex medical terminology and contexts. Recent research has leveraged these models to improve medical captioning. Additionally, LLMs (large language models) such as BioGPT [19] have been explored for their potential to generate coherent and diagnostically valuable medical captions, further pushing the boundaries of automated reporting in radiology.

3. Task and Dataset Descriptions

3.1. Task Descriptions

ImageCLEF has included medical tasks annually since 2004. Since 2019, it has focused each medical task on a specific issue but combined them into a single task with multiple subtasks. Four tasks are proposed for 2024: Image Captioning, Image Question Answering for Colonoscopy Images, MEDIQA-MAGIC, Quality Control of Synthesized Medical Images Generated by GANs. In ImageCLEF 2024 [8], we engage in the Image Captioning task [9], simultaneously participating in two subtasks: Concept Detection Task and Caption Prediction Task, each crucial in the holistic process of generating informative captions for medical images.

- **Concept Detection Task:** The Concept Detection Task involves using a refined subset of the UMLS 2022 AB version for concept generation. This subset is carefully selected to enhance the accuracy of concept detection by filtering concepts based on their semantic types. Moreover, to optimize concept detection from images, a stringent exclusion criterion is applied to remove low-frequency concepts, based on insights from previous iterations.
- **Caption Prediction Task:** In the Caption Prediction Task, a series of meticulous preprocessing steps are undertaken to ensure the integrity and coherence of the captioning process. Specifically, the removal of embedded hyperlinks within captions is performed as a fundamental preprocessing step. This careful action helps maintain data cleanliness and consistency, thereby supporting subsequent analytical processes and enabling accurate caption prediction outcomes.

3.2. Dataset Information

The data for the captioning task will consist of images selected from medical literature, including annotations and related UMLS terms manually curated as metadata. For the development dataset, Radiology Objects in COntext Version 2 (ROCOv2) [20], an updated and expanded version of the Radiology Objects in COntext (ROCO) dataset [21], is used for both subtasks. As in previous versions, this dataset originates from biomedical articles in the PMC OpenAccess collection [22], with the test set comprising a set of previously unseen images.

- Training Dataset: Includes 70,108 images.

- Validation Dataset: Includes 9,972 images.
- Test Dataset: Includes 17,237 images.

4. Experiments and Results

4.1. The Proposed Approach

4.1.1. Concept Detection Methodology

We aim to extract features from images by carefully examining and testing a variety of pretrained models that fall into three main architectural paradigms, which are shown in Table 1. The list that follows summarizes the particular models that are being examined:

- **CNN-based architectures:** Microsoft/ResNet-50 [23], an archetype of conventional convolutional neural network (CNN) models, characterized by its utilization of residual blocks to mitigate the challenges associated with gradient vanishing, thereby enhancing model performance within computationally tractable bounds.
- **Transformer-based architectures:**
  - ViT (Vision Transformer) [14]: Pioneering the paradigm shift in image data processing, ViT adopts a transformative approach by encoding images into patch embeddings, followed by feature extraction using a Transformer encoder, reminiscent of text data processing methodologies.
  - DeiT (Data-efficient Image Transformers) [24]: An evolution of ViT, DeiT emphasizes data efficiency, facilitating training with reduced data volumes while preserving commendable performance metrics.
  - Swin-v2 (Shifted Window Transformer v2) [25]: Distinguished by its innovative utilization of self-attention mechanisms within shifted windows, Swin-v2 ameliorates computational complexity and augments performance across a spectrum of tasks, including image classification and segmentation.
  - BEiT (Bidirectional Encoder representation from Image Transformers) [26]: At the confluence of Transformer and BERT architectures, BEiT excels in capturing robust image features through bidirectional encoding methodologies.
  - BiomedCLIP [27]: A domain-specific adaptation of ViT tailored for biomedical applications, leveraging the CLIP architecture to enhance performance in medical domain tasks.
- **Model Ensembles:** In our ensemble framework, we leverage sophisticated fusion techniques to harness the collective predictive power of multiple models. A key method employed is weighted averaging, where predictions from each member model are aggregated based on their respective weights derived from validation performance.
  - Ensemble-2 model (Swin-v2 + BEiT): The symbiotic fusion of Swin-v2 and BEiT engenders a collaborative synergy, capitalizing on the distinctive strengths of each constituent model to surpass individual model performances.
  - Ensemble-4 model (Swin-v2 + BEiT + DeiT + ViT): Comprising a composite quartet of models, this ensemble fortifies accuracy and generalization capabilities through the combination of representatives from Transformer-based models.

Following the feature extraction step, the retrieved features pass via a linear layer and classifier, where they are transformed and classified to provide outputs that correspond to the chosen class categories. This key step emphasizes the thorough orchestration of feature transformation and classification to produce predictions specific to the required class taxonomy.

Table 1
Statistics of models for the Concept Detection subtask.

Models	Version	Detailed	# Parameters
Resnet-50	-	microsoft/resnet-50	27 122 124
BEiT	base	microsoft/beit-base-patch16-224	88 065 356
BEiT	base	microsoft/beit-base-patch16-224	88 065 356
Swin	v2	microsoft/swinv2-base-patch4-window12-192-22k	89 459 332
DeiT	base	facebook/deit-base-patch16-224	88 692 620
ViT	base	google/vit-base-patch16-224	88 692 620
BiomedCLIP	base	ikim-uk-essen/BiomedCLIP_ViT_patch16_224	88 692 620
BEIT	large	microsoft/beit-large-patch16-224	305 971 084
Ensemble-2	-	Swin-v2 + BEiT	-
Ensemble-4	-	Swin-v2 + BEiT + DeiT + ViT	-

Concept Filtering A certain process must be followed while using the BEiT (Bidirectional Encoder Representations from Image Transformers) model in order to carry out idea filtering and modify the output threshold to detect variations in the outcomes. The following are the steps to follow: On a given dataset, do inference using the BEiT model and modify the output threshold to filter the ideas or classes. Setting various threshold values and watching the ensuing outcomes allows for this modification. We may adjust and assess how different thresholds affect the model's performance using this procedure.

4.1.2. Captioning Methodology

Figure 1 depicts an overview of the proposed method for Medical Captioning task.

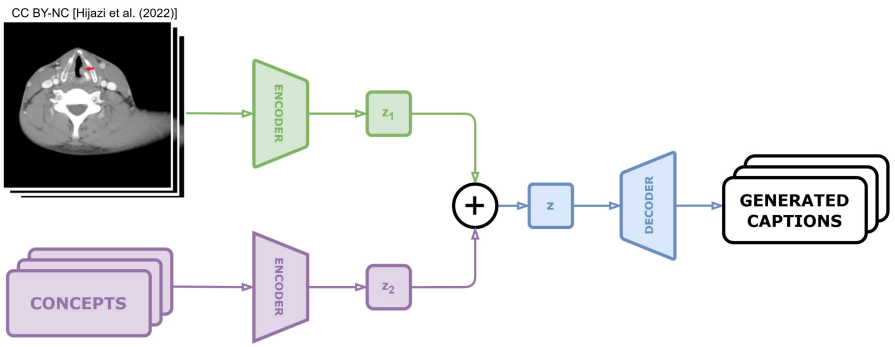

Figure 1: An overview of the multimodal architecture for Medical Caption Generation challenge.

Given the primary focus on Image Captioning in this research, the architectural design must effectively extract salient features from both the image and its corresponding text, combining them to generate the final caption. Our carefully curated multimodal fusion architecture incorporates essential components like an image encoder for pertinent feature extraction, a text encoder for eliciting semantic information from text, and a decoder to synthesize insights from the textual context. Additionally, the fusion mechanism integrates image features and output classifications from concept detection, synergistically blending them with textual input to decode and generate the caption output. The proposed approach leverages the pretrained Bidirectional Encoder Representations from Transformers (BEiT) model for image feature extraction. Boasting a symmetric Transformer architecture, BEiT can comprehend image representational features by concurrently considering both surrounding image patches and global context. With its extensive training on copious data, BEiT can be fine-tuned and achieve state-of-the-art results across several computer vision and image processing benchmarks.

To encode the input text captions, this research employs two domain-specific language models:

BioBART (Bidirectional and Auto-Regressive Transformers for Biomedical Text) and ClinicalT5 (Text-to-Text Transfer Transformer fine-tuned on clinical data).

- BioBART [28] is a version of the BART model [29] adapted and further pre-trained on biomedical text data such as medical literature, case reports, and genomic analysis documents. Leveraging its bidirectional Transformer architecture, BioBART can effectively encode both general and biomedical domain-specific text, enabling the extraction of rich semantic representations for tasks like text summarization, medical question-answering, and report generation.
- ClinicalT5 [30] is the T5 model [31] additionally fine-tuned on clinical text data including patient records and consultation reports. Harnessing its text-to-text transfer learning capability for multi-task modeling, ClinicalT5 can be applied to various natural language processing tasks in the healthcare domain, such as treatment classification, medical information extraction, and summarization of patient records.

For the process of encoding text concepts, we utilize the output from the BEiT model, which is specifically trained for the concept detection task. During this process, we apply a threshold of 0.5 to selectively retain predictions with a confidence score higher than 0.5, while discarding predictions with lower confidence scores. This discriminative process aids in capturing the semantic essence of the detected concepts, thereby facilitating their seamless integration into the multimodal fusion architecture for further processing and analysis.

4.2. Experimental Settings

Several experiments have been conducted to assess the efficacy of the proposed methodologies in addressing the ImageCLEFmedical Caption 2024 challenge. Specifically, each pre-trained vision model has been instantiated and evaluated, as detailed in Table 2, which offers a comprehensive overview of the pre-trained models employed in this study, encompassing their respective vision model designations, versions, and parameter counts for each fusion model. These experiments serve to elucidate both the potential and limitations inherent in each model with regard to the Image Captioning task, thereby facilitating the selection of the optimal approach for generating final predictions on the private test dataset of the competition.

- **Concept Detection Task:** For the concept detection subtask, the optimization criterion utilized during training is the AdamW optimizer [32]. The models are trained for 5 epochs with a batch size of 30 and an initial learning rate of 5e-5. During training, the BCEWithLogitsLoss function, which combines a Sigmoid layer and BCELoss, is applied, and a threshold value ranging from 0.45 to 0.5 is predominantly used to process the model's output. To ensure meaningful comparison results, consistent hyperparameters are maintained across all experiments.
- **Caption Prediction Task:** During the training process for the caption prediction task, the CrossEntropyLoss criterion is applied with the ignore_index parameter set to the pad token index of the tokenizer. This setup helps mitigate the influence of pad tokens on loss computation, ensuring more precise training outcomes. For optimization, the AdamW optimizer is utilized with a learning rate of 1e-4 and a weight decay rate of 0.01, chosen to balance training efficiency and model generalization [32]. To leverage the benefits of Mixed Precision Training [33], the Gradient scaler is integrated into the training pipeline. This scaler adjusts the gradient scale, enhancing training efficiency and convergence speed of the models. Additionally, the LinearScheduleWithWarmup is employed to adjust the learning rate over time during training. This scheduling mechanism requires pre-defining the number of warmup steps and total training steps to optimize the learning rate schedule effectively. During each training iteration, a batch size of 16 is utilized. Overall, these training configurations and optimizations contribute to the performance and stability of the training process, leading to superior model performance.

The hardware utilized for computation included both NVIDIA Tesla T4 and NVIDIA Tesla P100 GPUs.

4.3. Evaluation Methodology

Our evaluation consists of two tasks: Concept Detection and Caption Prediction. Each task uses specific metrics to measure performance.

- **Concept Detection Task:** We assess the performance of concept identification using Accuracy, Precision, Recall, and F1 score. These metrics measure overall correctness, positive prediction accuracy, relevant concept capture, and balanced precision and recall, respectively [34].
- **Caption Prediction Task:** We evaluate the quality and coherence of generated captions using BERTScore (Bidirectional Encoder Representations from Transformers Score) [35], BLEU (Bilingual Evaluation Understudy, 1-4) [36], ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [37], and METEOR (Metric for Evaluation of Translation with Explicit ORdering) [38]. These metrics assess semantic similarity, fluency, relevance, coherence, informativeness, and lexical/syntactic aspects.

Using this diverse set of metrics, we ensure a comprehensive understanding of the model's performance and facilitate informed decision-making for further refinement.

4.4. Experimental Results

As detailed in Table 2, the comparative evaluation of various concept detection models on the development validation set yields valuable insights into their performance across diverse evaluation metrics. Among these models, Swin-v2 emerges as the frontrunner, exhibiting the highest accuracy (0.16366), recall (0.47114), and F1 score (0.58944). This underscores Swin-v2's effectiveness in not only accurately identifying pertinent instances but also striking a harmonious balance between precision and recall, rendering it well-suited for concept detection endeavors. Ensemble methodologies, which predictions from multiple models, demonstrate promising outcomes as well. Notably, the Ensemble-2 model showcases commendable precision (0.94501) and a noteworthy F1 score (0.58581), suggesting that leveraging diverse models can augment predictive efficacy, particularly in precision-oriented tasks. While the Ensemble-4 model marginally surpasses Ensemble-2 in precision (0.94508), it exhibits a slightly lower F1 score (0.58460), implying a subtle trade-off in recall when employing additional models.

Table 2
Comparative performance of the Concept Detection method on the validation set.

Models	Accuracy	Precision	Recall	F1
Resnet-50	0.11412	0.89235	0.39643	0.51566
BEiT-B	0.15554	0.93087	0.45961	0.57662
Swin-v2	**0.16366**	0.94428	**0.47114**	**0.58944**
DeiT-B	0.15674	0.93353	0.45849	0.57641
ViT-B	0.15413	0.93477	0.45571	0.57439
BiomedCLIP	0.15975	0.94095	0.46453	0.58319
BEIT-L	0.16145	0.93669	0.46700	0.58418
Ensemble-2	0.16155	0.94501	0.46683	0.58581
Ensemble-4	0.16135	**0.94508**	0.46526	0.58460

BEiT-L and BiomedCLIP also manifest robust performance metrics. BEiT-L achieves an accuracy of 0.16145 and an F1 score of 0.58418, while BiomedCLIP demonstrates balanced performance with an accuracy of 0.15975 and an F1 score of 0.58319. These findings underscore the efficacy of these models in maintaining high precision and achieving a favorable balance with recall.

Other models such as BEiT-B, DeiT-B, and ViT-B exhibit commendable performance, albeit slightly trailing the top performers. For instance, BEiT-B records an accuracy of 0.15554 and an F1 score of 0.57662, indicating respectable yet not leading-edge performance. Similarly, DeiT-B and ViT-B attain comparable results, with DeiT-B registering an accuracy of 0.15674 and an F1 score of 0.57641, and ViT-B yielding an accuracy of 0.15413 and an F1 score of 0.57439. Conversely, ResNet-50 demonstrates

notably inferior performance across all metrics, with an accuracy of 0.11412 and an F1 score of 0.51566. This underscores its relatively limited efficacy in the concept detection task.

In summation, the Swin-v2 model emerges as the most dependable choice for concept detection owing to its superior accuracy, recall, and F1 score. Ensemble methodologies, particularly Ensemble-2, exhibit robust performance, underscoring the advantages of model amalgamation. BEiT-L and BiomedCLIP offer balanced performance, rendering them viable alternatives. Meanwhile, ResNet-50's diminished performance suggests its lesser suitability for this specific task, underscoring the strides made by newer architectural advancements.

Table 3
A comparative analysis of various configurations on the validation set, with "Process" denoting post-processing of output captions to mitigate repetition, and "Concepts" representing features potentially derived from the Concept Detection subtask.

Model	Configuration	BERTScore	BLEU1	BLEU2	BLEU3	BLEU4	ROUGE	METEOR
BEiT+BioBart	Concepts+No-Process	**0.60589**	0.03293	0.01019	0.00337	0.00040	0.10721	0.05673
BEiT+BioBart	Concepts+Process	**0.60589**	0.03293	0.01019	0.00337	0.00040	0.10721	0.05673
BEiT+Clinical-T5	Concepts+No-Process	0.45752	0.07408	0.03008	0.01244	0.00476	0.09298	0.08909
BEiT+Clinical-T5	Concepts+Process	0.57597	**0.08145**	**0.03319**	**0.01423**	**0.00519**	**0.13336**	0.09817
BEiT+Clinical-T5	No-Concepts+No-Process	0.46001	0.07501	0.03057	0.01077	0.00303	0.09711	0.09298
BEiT+Clinical-T5	No-Concepts+Process	0.57487	0.08110	0.03231	0.01137	0.00310	0.13293	**0.10086**

As detailed in Table 3, the comparative analysis of various model configurations on the validation set reveals insights into the efficacy of incorporating concepts and post-processing techniques in caption generation tasks. The models evaluated include BEiT+BioBart and BEiT+Clinical-T5, with configurations either incorporating concepts derived from the Concept Detection subtask or excluding them, and applying post-processing to mitigate repetition in the output captions. The results indicate that for the BEiT+BioBart model, the inclusion of concepts and the application of post-processing do not result in any variation in performance across all evaluated metrics, including BERTScore, BLEU (from 1 to 4), ROUGE, and METEOR. This suggests that for BEiT+BioBart, the post-processing step does not impact the model's ability to generate captions when concepts are included, maintaining consistent performance.

In contrast, the BEiT+Clinical-T5 model demonstrates a more nuanced response to the incorporation of concepts and post-processing. When concepts are included without post-processing, there is a slight decline in BERTScore compared to the configuration without concepts. However, BLEU, ROUGE, and METEOR scores show an improvement with the inclusion of concepts, highlighting the potential benefits of concept integration in enhancing the model's performance in these specific metrics. Notably, when post-processing is applied, the BEiT+Clinical-T5 model exhibits substantial improvements across all metrics, irrespective of the presence of concepts. This improvement underscores the critical role of post-processing in refining output quality, with the highest METEOR score observed in the configuration without concepts but with post-processing. Comparing the two models, BEiT+Clinical-T5 generally outperforms BEiT+BioBart in BLEU, ROUGE, and METEOR scores. This superior performance is particularly evident when post-processing is applied, suggesting that BEiT+Clinical-T5 is more responsive to post-processing enhancements. However, BEiT+BioBart achieves a higher BERTScore when concepts are included, indicating a potential strength in semantic similarity measures.

In conclusion, the analysis underscores the importance of model selection, the strategic inclusion of concepts, and the application of post-processing in optimizing caption generation performance. BEiT+Clinical-T5 emerges as a more robust model with gains from post-processing, while BEiT+BioBart maintains consistent performance with concept inclusion. These findings provide valuable insights for future research and development in automated caption generation systems, emphasizing tailored approaches for different model architectures.

As detailed in Table 4, the performance evaluation of different models on the validation and private test sets provides a comprehensive understanding of their effectiveness across various configurations and datasets. For concept detection, three configurations were assessed: Concept BEiT-B with a threshold of 0.45, Detection BEiT-B with a threshold of 0.5, and Swin-v2 with a threshold of 0.5. The results reveal

Table 4
Performance evaluation of different models on the validation set and private test set.

#	Models	Configuration	Validation set	Test set
Concept Detection	BEiT-B	Threshold_0.45	0.57662	0.61079
	BEiT-B	Threshold_0.5	-	0.60904
	Swin-v2	Threshold_0.5	0.58944	0.61998
Caption Prediction	BEiT+Clinical-T5	No-Concepts+No-Process	0.46001	0.4433
	BEiT+Clinical-T5	Concepts+No-Process	0.45752	0.4453
	BEiT+Clinical-T5	Concepts+Process	0.57597	0.558
	BEiT+BioBart	Concepts+Process	0.60589	0.5794

that the Swin-v2 model performs the best, achieving scores of 0.58944 on the validation set and 0.61998 on the private test set, suggesting superior capability in accurately detecting concepts compared to the BEiT-B models. The Concept BEiT-B model with a threshold of 0.45 also shows strong performance, though slightly lower than Swin-v2, indicating the threshold setting's impact on model efficacy.

For caption prediction, four configurations were evaluated: BEiT+Clinical-T5 without concepts and without post-processing, BEiT+Clinical-T5 with concepts and without post-processing, BEiT+Clinical-T5 with concepts and with post-processing, and BEiT+BioBart with concepts and with post-processing. The BEiT+Clinical-T5 model without concepts and post-processing scored 0.46001 on the validation set and 0.4433 on the private test set, while adding concepts slightly improved the private test set score to 0.4453. However, the most considerable performance boost was observed when post-processing was applied to the BEiT+Clinical-T5 model with concepts, raising the scores to 0.57597 on the validation set and 0.558 on the private test set. This highlights the substantial role of post-processing in enhancing model performance.

Moreover, the BEiT+BioBart model with concepts and post-processing achieved the highest scores among all configurations, with 0.60589 on the validation set and 0.5794 on the private test set. This underscores the effectiveness of combining concepts with post-processing in the BioBart architecture, suggesting that such integration can improve caption generation quality. Overall, the analysis emphasizes the critical influence of model configuration, the integration of concepts, and the application of post-processing on the performance outcomes. The superior performance of the Swin-v2 model for concept detection and the BEiT+BioBart model for caption prediction indicates that different models may excel in specific sub-tasks, advocating for a nuanced approach in model selection and optimization based on the task requirements and dataset characteristics.

4.5. Error Analysis

Table 5
Example outputs of caption prediction from different models.

Ground Truth	BEiT + Clinical-T5	BEiT + BioBART	Predicted Concepts
Axial contrasted CT image of larynx, showing left sided glottic versus supraglottic mass.	CT scan showing mass lesion (arrow)	CT scan showing left renal mass	Magnetic Resonance Imaging
Chest X-ray face (solitary pulmonary nodule of the heart-phrenic angle).	Chest X-ray showing opacification (arrow) chest	Chest X-ray showing bilateral infiltrates	X-Ray Computed Tomography

As detailed in Table 5, when employing the BEiT model in conjunction with ClinicalT5 for medical image analysis, several notable errors have been observed across various dimensions. These errors include incorrect identification of regions or image types, omissions in providing specific details, and inaccuracies in context, thereby impacting the overall reliability of the model's results. The model occasionally encounters difficulties in accurately identifying regions of interest within the images.

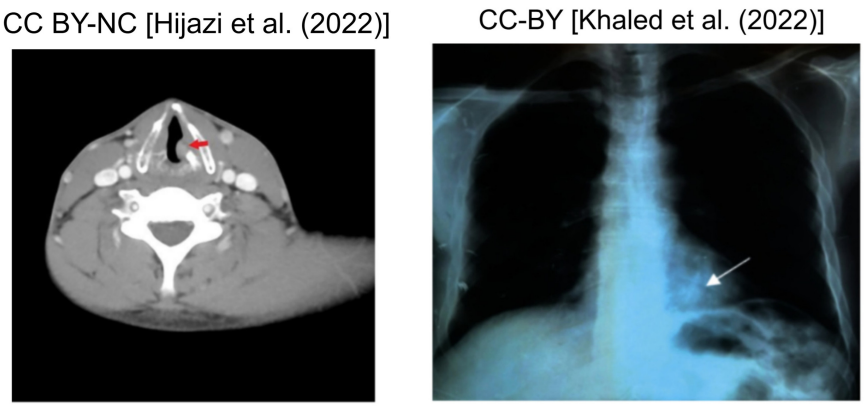

Figure 2: Example images of caption prediction. The images are arranged in sequential order in Table 5.
a) `ImageCLEFmedical_Caption_2024_valid_009001` is an example of Table 5.
b) `ImageCLEFmedical_Caption_2024_valid_009698` is an example of Table 5.

For instance, it might misinterpret an anteroposterior X-ray of the pelvis as indicating bilateral tibial fractures. Similarly, it might incorrectly classify a cross-sectional, contrast-enhanced CT scan of the larynx as a left renal tumor.

Omissions in providing specific details have become evident in the model's predictions. The model often fails to provide the complex details necessary for comprehensive clinical interpretation. For example, it may overlook critical features such as the eccentric position of a metallic head in an X-ray or the presence of stratified bile in a CT scan. Moreover, contextual inaccuracies are common, leading to misleading or entirely incorrect descriptions. The model sometimes struggles to grasp the broader context of medical images, resulting in descriptions that do not align appropriately with the actual content of the images. Similarly, when utilizing the BEiT model in combination with BioBART, analogous errors have been observed across various aspects. These include incorrect identification of regions or image types, omissions in providing specific details, and contextual inaccuracies. Comparing BEiT with ClinicalT5 and BEiT with BioBART, although both models exhibit similar error patterns, there are minor differences in their performance. BEiT combined with ClinicalT5 demonstrates slightly better performance in certain aspects, such as providing more accurate descriptions and better contextual understanding. Conversely, BEiT combined with BioBART shows a slight advantage in specific scenarios, particularly in identifying anatomical structures or image types. However, both models have room for improvement, highlighting ongoing challenges in developing robust and reliable automated methods for medical image analysis. In both models, conceptual errors frequently occur, indicating a mismatch between the predicted concept and the actual content of the medical images. These errors underscore the challenges in accurately interpreting and classifying medical images based on their content.

To enhance the accuracy of medical image analysis models, a range of strategies must be employed to improve data quality, model architecture, and training processes. Firstly, the use of high-quality, well-annotated datasets is crucial. Combining this with data augmentation techniques such as rotation, zooming, flipping, and color adjustment can help increase the size and diversity of the training dataset, thereby enhancing the model's generalization capabilities. In terms of model architecture, employing models pre-trained on domain-specific datasets or state-of-the-art (SOTA) models that achieve superior results is essential. Furthermore, incorporating additional feature extraction from image data, such as bounding-boxes, segmentation, or advanced features, can help the model better understand the structure and context of the images. Finally, regularly testing and re-evaluating the model using diverse datasets will help in early detection of errors and timely adjustment of the model, ensuring the reliability and accuracy of medical image analysis results.

5. Conclusion and Future Works

In this study, an enhanced approach to medical caption generation was introduced by integrating concept detection into attention mechanisms. The method improved performance metrics, with the Swin-v2 model achieving an F1 score of 0.58944 on the validation set and 0.61998 on the private test set, earning 3rd place in concept detection. For caption prediction, the BEiT+BioBart model, augmented with concept integration and post-processing, achieved a BERTScore of 0.60589 on the validation set and 0.5794 on the private test set, securing 9th place. These results underscore the effectiveness of concept-aware systems in generating precise and contextually relevant medical descriptions.

Future work will focus on enhancing model performance through several avenues unrelated to data expansion. First, optimizing model architectures and training protocols can further improve accuracy and efficiency. Second, incorporating more advanced attention mechanisms and fine-tuning hyperparameters may yield better contextual understanding and caption quality. Third, integrating explainability techniques will ensure that model predictions are interpretable and trustworthy for healthcare professionals. Additionally, exploring transfer learning and domain adaptation techniques could enhance model performance across various medical imaging modalities. Furthermore, leveraging large language models (LLMs) such as GPT-3 and BioGPT for their potential to generate coherent and diagnostically valuable medical captions will be explored [39] [19]. Finally, developing robust post-processing algorithms to further refine generated captions, ensuring they meet clinical standards, is planned. These efforts aim to advance the capabilities of medical image analysis and automated reporting systems, contributing to more sophisticated and reliable tools for the healthcare industry.

Acknowledgment

This research is funded by University of Information Technology-Vietnam National University HoChiMinh City under grant number D4-2024-01.

References

[1] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, C. I. SÃąnchez, A survey on deep learning in medical image analysis, Medical Image Analysis 42 (2017) 60–88. URL: https://www.sciencedirect.com/science/article/pii/S1361841517301135. doi:https://doi.org/10.1016/j.media.2017.07.005.

[2] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, S. Thrun, Dermatologist-level classification of skin cancer with deep neural networks, Nature 542 (2017) 115–118. doi:10.1038/nature21056.

[3] B. Jing, P. Xie, E. Xing, On the automatic generation of medical imaging reports, in: I. Gurevych, Y. Miyao (Eds.), Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 2577–2586. URL: https://aclanthology.org/P18-1240. doi:10.18653/v1/P18-1240.

[4] C. Y. Li, X. Liang, Z. Hu, E. P. Xing, Knowledge-driven encode, retrieve, paraphrase for medical image report generation, in: Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, AAAI'19/IAAI'19/EAAI'19, AAAI Press, 2019. URL: https://doi.org/10.1609/aaai.v33i01.33016666. doi:10.1609/aaai.v33i01.33016666.

[5] H.-C. Shin, K. Roberts, L. Lu, D. Demner-Fushman, J. Yao, R. M. Summers, Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2497–2506. doi:10.1109/CVPR.2016.274.

[6] B. Yang, M. Cao, Y. Zou, Concept-aware video captioning: Describing videos with effective prior information, IEEE Transactions on Image Processing 32 (2023) 5366–5378. doi:10.1109/TIP.2023.3307969.

[7] T. Wang, W. Chen, Y. Tian, Y. Song, Z. Mao, Improving image captioning via predicting structured concepts, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 360–370. URL: https://aclanthology.org/2023.emnlp-main.25. doi:10.18653/v1/2023.emnlp-main.25.

[8] B. Ionescu, H. Müller, A. Drăgulinescu, J. Rückert, A. Ben Abacha, A. García Seco de Herrera, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, C. S. Schmidt, T. M. Pakull, H. Damm, B. Bracke, C. M. Friedrich, A. Andrei, Y. Prokopchuk, D. Karpenka, A. Radzhabov, V. Kovalev, C. Macaire, D. Schwab, B. Lecouteux, E. Esperança-Rodier, W. Yim, Y. Fu, Z. Sun, M. Yetisgen, F. Xia, S. A. Hicks, M. A. Riegler, V. Thambawita, A. Storås, P. Halvorsen, M. Heinrich, J. Kiesel, M. Potthast, B. Stein, Overview of ImageCLEF 2024: Multimedia retrieval in medical applications, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the 15th International Conference of the CLEF Association (CLEF 2024), Springer Lecture Notes in Computer Science LNCS, Grenoble, France, 2024.

[9] J. Rückert, A. Ben Abacha, A. G. Seco de Herrera, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, B. Bracke, H. Damm, T. M. G. Pakull, C. S. Schmidt, H. Müller, C. M. Friedrich, Overview of ImageCLEFmedical 2024 – Caption Prediction and Concept Detection, in: CLEF2024 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Grenoble, France, 2024.

[10] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, R. M. Summers, Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3462–3471. doi:10.1109/CVPR.2017.369.

[11] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, H. E. Stanley, Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals, Circulation 101 (2000) E215–E220. doi:10.1161/01.CIR.101.23.e215.

[12] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778. doi:10.1109/CVPR.2016.90.

[13] M. Tan, Q. V. Le, EfficientNet: Rethinking model scaling for convolutional neural networks, in: ICML, 2019, pp. 6105 – 6114.

[14] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, International Conference on Learning Representations abs/2010.11929 (2020).

[15] H. Bao, L. Dong, F. Wei, BEit: BERT pre-training of image transformers, International Conference on Learning Representations abs/2106.08254 (2021).

[16] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 10012–10022.

[17] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (1997) 1735âĂŞ1780. doi:10.1162/neco.1997.9.8.1735.

[18] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, BioBERT: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (2020) 1234–1240.

[19] R. Luo, K. Sun, Y. Cheng, Y. Zhang, Y. Xu, Y. Li, N. Zhang, B. Bi, X. Zhao, H. Wang, et al., BioGPT: generative pre-trained transformer for biomedical text generation and mining, Briefings in Bioinformatics (2022).

[20] J. Rückert, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, C. S. Schmidt, S. Koitka, O. Pelka, A. B. Abacha, A. G. S. de Herrera, H. Müller, P. A. Horn, F. Nensa, C. M. Friedrich, ROCOv2: Radiology

Objects in COntext version 2, an updated multimodal image dataset, Scientific Data (2024). URL: https://arxiv.org/abs/2405.10004v1. doi:10.1038/s41597-024-03496-6.

[21] O. Pelka, S. Koitka, J. Rückert, F. Nensa, C. M. Friedrich, Radiology objects in context (roco): a multimodal image dataset, in: Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis: 7th Joint International Workshop, CVII-STENT 2018 and Third International Workshop, LABELS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 3, Springer, 2018, pp. 180–189.

[22] National Library of Medicine, PMC open access subset, https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/, 2003. [cited 2024 May 30].

[23] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[24] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, H. Jegou, Training data-efficient image transformers & distillation through attention, in: M. Meila, T. Zhang (Eds.), Proceedings of the 38th International Conference on Machine Learning, volume 139 of *Proceedings of Machine Learning Research*, PMLR, 2021, pp. 10347–10357. URL: https://proceedings.mlr.press/v139/touvron21a.html.

[25] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, F. Wei, B. Guo, Swin transformer v2: Scaling up capacity and resolution, in: Computer Vision and Pattern Recognition, 2021, pp. 11999 – 12009. doi:10.1109/cvpr52688.2022.01170.

[26] H. Bao, L. Dong, F. Wei, BEit: BERT pre-training of image transformers, International Conference on Learning Representations abs/2106.08254 (2021).

[27] S. Zhang, Y. Xu, N. Usuyama, H. Xu, J. Bagga, R. Tinn, S. Preston, R. Rao, M. Wei, N. Valluri, C. Wong, A. Tupini, Y. Wang, M. Mazzola, S. Shukla, L. Liden, J. Gao, M. P. Lungren, T. Naumann, S. Wang, H. Poon, BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs, 2024. arXiv:2303.00915.

[28] H. Yuan, Z. Yuan, R. Gan, J. Zhang, Y. Xie, S. Yu, BioBART: Pretraining and evaluation of a biomedical generative language model, in: Proceedings of the 21st Workshop on Biomedical Language Processing, 2022. doi:10.18653/v1/2022.bionlp-1.9.

[29] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 7871–7880. URL: https://aclanthology.org/2020.acl-main.703. doi:10.18653/v1/2020.acl-main.703.

[30] Q. Lu, D. Dou, T. Nguyen, ClinicalT5: A generative language model for clinical text, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2022, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 5436–5443. URL: https://aclanthology.org/2022.findings-emnlp.398. doi:10.18653/v1/2022.findings-emnlp.398.

[31] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, Journal of Machine Learning Research 21 (2020) 1–67.

[32] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, International Conference on Learning Representations abs/1412.6980 (2014).

[33] P. Micikevicius, S. Narang, J. Alben, G. Diamos, E. Elsen, D. Garcia, et al., Mixed precision training, in: International Conference on Learning Representations, 2018.

[34] M. Sokolova, G. Lapalme, A systematic analysis of performance measures for classification tasks, Information Processing & Management 45 (2009) 427–437.

[35] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, BERTScore: Evaluating text generation with BERT, International Conference on Learning Representations abs/1904.09675 (2019).

[36] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: P. Isabelle, E. Charniak, D. Lin (Eds.), Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics,

Philadelphia, Pennsylvania, USA, 2002, pp. 311–318. URL: https://aclanthology.org/P02-1040. doi:10.3115/1073083.1073135.

[37] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 74–81. URL: https://aclanthology.org/W04-1013.

[38] S. Banerjee, A. Lavie, METEOR: An automatic metric for MT evaluation with improved correlation with human judgments, in: J. Goldstein, A. Lavie, C.-Y. Lin, C. Voss (Eds.), Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Association for Computational Linguistics, Ann Arbor, Michigan, 2005, pp. 65–72. URL: https://aclanthology.org/W05-0909.

[39] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, in: Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20, Curran Associates Inc., Red Hook, NY, USA, 2020.