UIT-DarkCow team at ImageCLEFmedical Caption 2024: Diagnostic Captioning for Radiology Images Efficiency with Transformer Models

Quan Van Nguyen[1,2], Huy Quang Pham[1,2], Dan Quang Tran[1,2], Thang Kien-Bao Nguyen[1,2], Nhat-Hao Nguyen-Dang[1,2] 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*

Abstract

Purpose: This study focuses on the development of automated text generation from radiology images, termed diagnostic captioning, to assist medical professionals in reducing clinical errors and improving productivity. The aim is to provide tools that enhance report quality and efficiency, which can significantly impact both clinical practice and deep learning research in the biomedical field.
Methods: In our participation in the ImageCLEFmedical2024 Caption evaluation campaign, we explored caption prediction tasks using advanced Transformer-based models. We developed methods incorporating Transformer encoder-decoder and Query Transformer architectures. These models were trained and evaluated to generate diagnostic captions from radiology images.
Results: Experimental evaluations demonstrated the effectiveness of our models, with the VisionDiagnostor-BioBART model achieving the highest BERTScore of 0.6267. This performance contributed to our team, DarkCow, achieving third place on the leaderboard. Our source code is public at **this link**.
Conclusion: Our diagnostic captioning models show great promise in aiding medical professionals by generating high-quality reports efficiently. This approach can facilitate better data processing and performance optimization in medical imaging departments, ultimately benefiting healthcare delivery.

Keywords

ImageCLEF, Computer Vision, Diagnostic Captioning, Image Captioning, Image Understanding, Radiology Images, Transformer Models, Encoder-Decoder, Query Transformer

1. Introduction

Machine learning, especially Deep Learning, is creating breakthroughs in many different fields, and its impact on biomedicine is remarkable. With the exponential growth of biomedical data, researchers are exploring its potential in biomedical engineering, advanced computing, imaging systems, and biomedical data mining algorithms based on machine learning [1]. One important area is Diagnostic Captioning. Diagnostic Captioning is the process of automatically generating diagnostic text based on a set of medical images collected during a medical examination. It can assist less experienced physicians by minimizing clinical errors and helping experienced physicians generate diagnostic reports faster [2].

ImageCLEF is an annual multimodal machine learning campaign, part of the Cross-Language Evaluation Forum (CLEF), which has been running since 2003. It encourages breakthroughs in research and development of processing systems. Advanced multimedia processing in computer vision, image analysis, classification and retrieval in a multilingual, multimodal context. This year, one of ImageCLEF's four main missions is ImageCLEFMedical, which includes a series of challenges from annotating images to creating synthetic images and answering questions. In ImageCLEF 2024 [3], we took part in the

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

[*] Corresponding author.

21521333@gm.uit.edu.vn (Q. V. Nguyen); 21522163@gm.uit.edu.vn (H. Q. Pham); 21521917@gm.uit.edu.vn (D. Q. Tran); 21521432@gm.uit.edu.vn (T. K. Nguyen); 20520490@gm.uit.edu.vn (N. Nguyen-Dang); thienntb@uit.edu.vn (T. B. Nguyen-Tat)
0009-0000-3604-4679 (Q. V. Nguyen); 0009-0006-5815-4469 (H. Q. Pham); 0009-0003-8806-5289 (D. Q. Tran); 0009-0009-6456-4247 (T. K. Nguyen); 0009-0007-6405-7603 (N. Nguyen-Dang); 0000-0002-4809-7126 (T. B. Nguyen-Tat)

ImageCLEFmedical Caption task [4]. As in previous years, this task comprised two subtasks: concept detection and caption prediction.

Concept detection aims to associate biomedical images with related medical concepts while captioning prediction focuses on automatically generating preliminary diagnostic reports that accurately describe medical conditions and structures and anatomy shown in images. Concept detection also supports diagnostic notes by identifying key concepts that should be included in the preliminary report. Additionally, it can be used to index medical images according to related concepts, facilitating more efficient organization and retrieval.

Captioning prediction, in other words, diagnostic captioning, remains a challenging research problem, designed to support the diagnostic process by providing a preliminary report rather than replacing the physicians and human factors involved [2]. It is designed as a tool to assist in generating an initial diagnostic report of a patient's condition, helping doctors focus on important areas of the image [5] and assisting them in making diagnoses. Guess more accurately quickly [6]. This approach can increase the efficiency of experienced clinicians, allowing them to handle high volumes of daily medical examinations more quickly and efficiently. For less experienced clinicians, automated annotation can help reduce the likelihood of clinical errors[7].

1.1. DarkCow Team Contributions

In this paper, we presented the experiments and the systems that were submitted by our DarkCow team in this year's caption prediction task, which helped us secure third place on the leaderboard (see Table 1). Our new approaches build on the rapid development of deep learning techniques, especially the Transformer [8] encoder-decoder architecture and the Query Transformer [9] for Large Language Model [10]. We leveraged the Vision Transformer (ViT) to extract visual features from radiology images. To optimize the use of information, we also used VinVL [11] to extract features of objects in the images. Our first approach is based on encoder-decoder architecture to generate image captions. In the second approach, we leveraged Query Transformer to help LLM understand images. We also conducted experiments with image pre-processing, caption length, and object features to analyze the impact of those aspects.

Table 1
Caption prediction task scores, rankings are based on BERTScore

Team	ID	BERTScore	ROUGE	BLEU-1	BLEURT	METEOR	CIDEr	CLIPScore	RefCLIPScore	ClinicalBLEURT	MedBERTScore
pclmed	634	0.629913	0.272626	0.268994	0.337626	0.113264	0.268133	0.823614	0.817610	0.466557	0.632318
CS_Morgan	429	0.628059	0.250801	0.209298	0.317385	0.092682	0.245029	0.821262	0.815534	0.455942	0.632664
DarkCow	**220**	**0.626720**	**0.245228**	**0.195044**	**0.306005**	**0.088897**	**0.224250**	**0.818440**	**0.811700**	**0.456199**	**0.629189**
auebnlpgroup	630	0.621112	0.204883	0.111034	0.289907	0.068022	0.176923	0.804067	0.798684	0.486560	0.626134
2Q2T	643	0.617814	0.247755	0.221252	0.313942	0.098590	0.220037	0.827074	0.813756	0.475908	0.622447
MICLab	678	0.612850	0.213525	0.185269	0.306743	0.077181	0.158239	0.815925	0.804924	0.445257	0.617195
DLNU_CCSE	674	0.606578	0.217857	0.151179	0.283133	0.070419	0.168765	0.796707	0.790424	0.475625	0.612954
Kaprov	559	0.596362	0.190497	0.169726	0.295109	0.060896	0.107017	0.792183	0.787201	0.439971	0.608924
DS@BioMed	571	0.579438	0.103095	0.012144	0.220211	0.035335	0.071529	0.775566	0.774823	0.529529	0.580388
DBS-HHU	637	0.576891	0.153103	0.149275	0.270965	0.055929	0.064361	0.784199	0.774985	0.476634	0.588744
KDE-medical-caption	557	0.567329	0.132496	0.106025	0.256576	0.038628	0.038404	0.765059	0.760958	0.502234	0.569659

This paper is organized explicitly as follows: Section 2 presents an overview of studies related to our research field. In Section 3, we introduce the data process and some detailed analysis of our dataset. Next, Section 4 introduces some image pre-processing techniques. Section 5 details the design of the proposed methods and evaluation metric. Section 6 Present experimental results based on the proposed method. Section 7 discusses some impact. Finally, Section 8 summarizes the research and suggests future directions.

2. Background and Related Works

2.1. Radiology Techniques

With the continuous advancement of imaging technology, medical imaging diagnosis has evolved from a supplementary examination tool to the most important clinical diagnostic and differential diagnostic method in modern medicine. Radiology techniques are used to scan images within the body, which are then interpreted and reported by radiologists to specialists [12]. With advancements in imaging technology, various imaging diagnostic methods have been developed, each with its own advantages and limitations. For example, X-ray imaging [13] offers non-invasive, quick, and painless imaging, but it involves exposure to ionizing radiation, which increases the risk of developing cancer later in life. On the other hand, MRI imaging [14] provides non-ionizing radiation and high spatial resolution, but it has relatively low sensitivity and longer scanning times, etc.

2.2. Former Medical Image Captioning Datasets

Medical imaging diagnosis today plays an incredibly important role in both the healthcare and information technology sectors. It not only aids in diagnosis and increases understanding of diseases but also holds immense potential in improving healthcare delivery and enhancing quality of life. The application of deep learning in medical image captioning in an era where AI is ubiquitous is evident; it automates the annotation process and significantly accelerates image analysis. Several datasets have been created to facilitate the training of medical image captioning tasks such as ROCO [15], PadChest [16], MIMIC-CXR [17], IU X-Ray [18], and MedICaT [19].

2.3. Related Work Methods

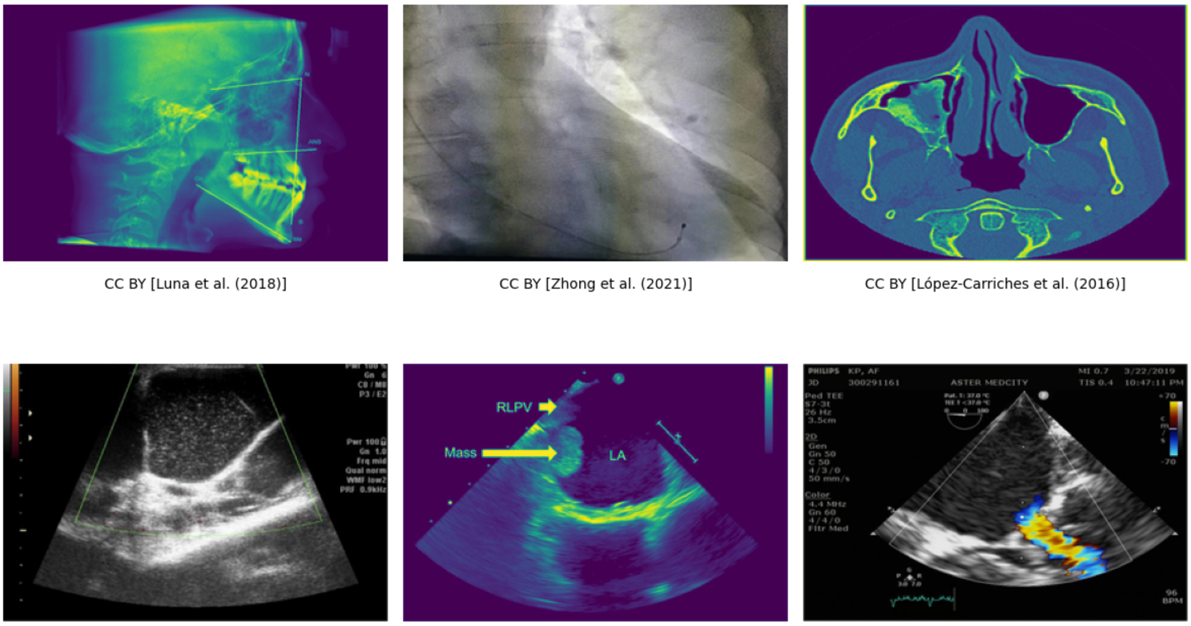
For the task of medical image captioning, various methods have been developed, with pioneering work in applying the CNN-RNN encoder-decoder approach to generate captions from medical images conducted by Shin et al. [5]. They utilized either the Network-in-Network or GoogLeNet architectures as encoding models, followed by LSTM [20] or GRU [21] as the decoding RNN to translate the encoded images into descriptive captions. In the process of translating images into biomedical text, MDNET [22] made a notable advancement by incorporating an attention mechanism. This model employs RESNET for image encoding, extending its skip connections to mitigate gradient vanishing.

In recent studies by Wang et al. [23], Kougia et al. [24], and Li et al. [25], a fusion of generative models and retrieval systems for Medical Image Captioning (MIC) has been explored. For instance, Wang et al. [23] proposed an approach that alternates between template retrieval and sentence generation for rare abnormal descriptions. This method relies on a contextual relational-topic encoder derived from visual and textual features, facilitating semantic consistency through hybrid knowledge co-reasoning. Additionally, Kougia et al. [24] from AUEB NLP group presented various systems for the Image-CLEFmed 2019 Caption task. One approach utilized a retrieval-based model that leverages visual features to retrieve the most similar images based on cosine similarity, combining their concepts to predict relevant captions. Another system incorporated CheXNet [26] with enhanced classification labels, employing a CNN encoder and a feed-forward neural network (FFNN) for multi-label classification. They also suggested an ensemble model by combining these systems, computing scores for returned concepts and merging them with image similarity scores to select the most relevant concepts.

Large language models (LLMs) have catalyzed significant progress in medical question answering; Med-PaLM [27] was the first model to exceed a “passing” score in US Medical Licensing Examination (USMLE). However, this and other prior work suggested significant room for improvement, especially when models’ answers were compared to clinicians’ answers. Med-PaLM 2 [28] bridges these gaps by leveraging a combination of base LLM improvements, medical domain finetuning, and prompting strategies including a novel ensemble refinement approach.

3. Dataset

Thanks to AUEB NLP Group for providing an excellent analysis of the dataset in the study of Kaliosis et al. [29]. When comparing ImageCLEFmedical2023 data with ImageCLEFmedical2024, we found no significant differences in the task of caption prediction. Therefore, we decided to reapply to analyze the dataset in this section.



CC BY [Luna et al. (2018)]

CC BY [Zhong et al. (2021)]

CC BY [López-Carriches et al. (2016)]

CC BY [Rao et al. (2016)]

CC BY-NC [Hong et al. (2014)]

CC BY [Samuel et al. (2020)]

Figure 1: Several images from the ImageCLEFmedical2024 dataset.

This year's ImageCLEFmedical Caption task provided a dataset that includes 70,108 radiology images in the training set, each annotated with medical concepts using UMLS terms and diagnostic captions. The organizers initially divided the dataset into training and validation subsets [30]. Building on previous campaigns, this year's dataset is an updated and expanded version of the Radiology Objects in Context (ROCO) dataset, which is sourced from a variety of biomedical studies in the PubMed Central OpenAccess (PMC OA) subset. The dataset used for the caption prediction task includes images from different modalities, such as X-ray and Computed Tomography (CT), although specific details about the image types were not provided. The goal of the caption prediction task is to generate open-ended diagnostic texts for the medical images (see Figure 1).

Table 2
The ten most common words and their frequencies in the ImageCLEFmedical2024 train set.

Most common words (excluding stop-words)										
Word	**showing**	**right**	**left**	**ct**	**image**	**chest**	**scan**	**computed**	**tomography**	**shows**
Occurrences	22,519	18,258	18,136	15,167	10,245	10,082	9,296	9,273	8,969	8,600

Table 3
The five most common captions found in the ImageCLEFmedical2024 train set alongside the number of images they are associated with.

Most common captions		
Position	**Caption**	**Occurences**
1	Initial panoramic radiograph	40
2	Final panoramic radiograph	37
3	Chest X-ray	20
4	Chest radiograph	17
5	Preoperative CT scan.	9

In the Caption prediction sub-task, each image has a diagnostic caption describing the described medical condition. There are a total of 69,743 captions in the training dataset and 9,959 captions in the validation dataset, one for each image. Similar to last year's campaign, the majority of captions (99.47%, or 69,743 out of 70,108) were unique. This is a notable difference from previous versions of the quest, where the uniqueness percentage was much lower. As a result, traditional retrieval methods based on nearest neighbor search are less efficient this year, including variants with a weighting mechanism based on the cosine similarity of the retrieved images. Therefore, more complex methods of creating subtitles are needed.

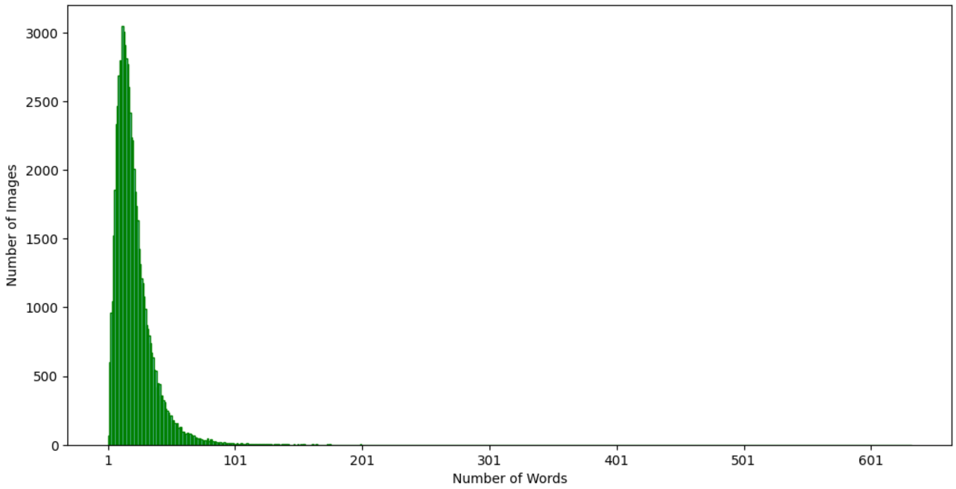


Figure 2: Distribution of caption lengths in the training set.

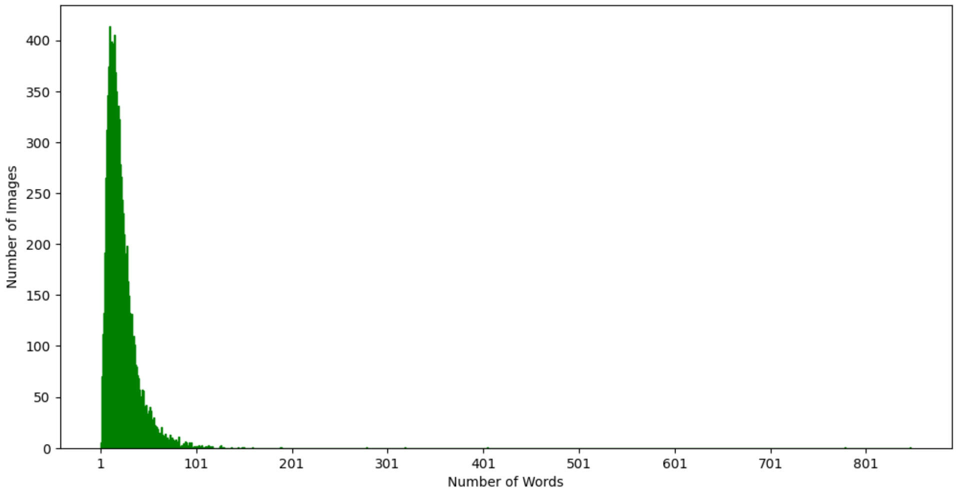


Figure 3: Distribution of caption lengths in the valid set. [30]

We observed that the maximum number of words in a single caption is 848 (occurred once), while the minimum is 1 (encountered 1 time). The average caption length is 20.84 words. These statistics apply to the entire dataset (training set and valid set). The five most common captions, as well as the ten most popular words (excluding stopwords), can be found in Tables 3 and 2, respectively. In Figure 2 and Figure 3, we present a distribution caption length of the training and valid sets, both indicating that the majority of captions contain fewer than 100 words.

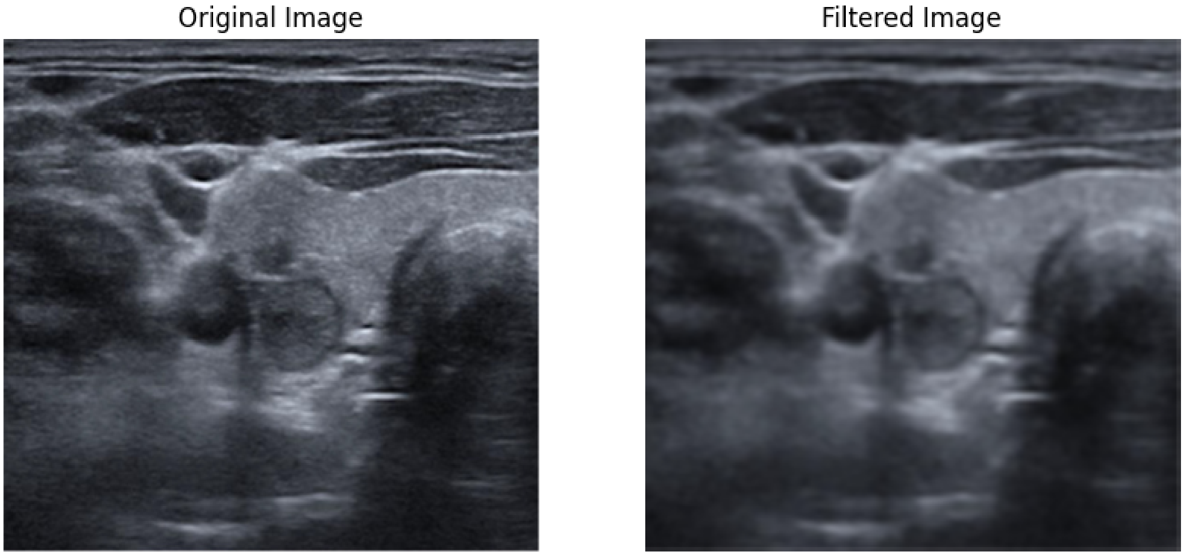


Figure 4: Application of Gaussian filter.

4. Image Pre-processing

4.1. Denoising

Denoising is crucial in enhancing image quality by reducing the noise while preserving the important details. Noise in medical images can come from various sources, such as sensor imperfections, poor scan conditions, or inherent patient movements during image acquisition.

The smoothness of images is controlled through the utilization of a Gaussian filter with a fixed kernel size. The Gaussian filter operates by smoothing images using a technique called convolution. It employs a Gaussian kernel - a matrix based on the Gaussian function to adjust pixel values. This kernel is applied over each pixel in the image, averaging the pixel values in its vicinity, weighted by their distance from the central pixel. The standard deviation σ of the Gaussian determines the amount of blurring: a larger σ results in more blurring, smoothing out more details and noise. This process helps in reducing noise and is often used as a preparatory step in image processing tasks to enhance image quality without losing critical structural details (see Figure 4).

The 2-D Gaussian function is given by:

$$G(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{1}$$

Medical image enhancement is one of the most widely used medical image processing techniques in medical domain. Its purpose is to improve the visual effect of the image and facilitate the analysis and understanding of the image by humans or machines. The Laplace transform and the Sobel gradient operator are two common ways of performing edge detection, image sharpening, and enabling the enhancement of the image (see Figure 5).

Step 1 Laplace Transform: Apply the Laplace transform to enhance contrast by emphasizing areas of rapid intensity change in the original image.

Step 2 Sobel Operator: Use the Sobel operator to enhance the edges of the image. This step also helps to smooth out noise, making the edges clearer and more cohesive.

Step 3 Smoothing: Smooth the image processed by the Sobel operator using a 3x3 mean filter. This step increases the contrast of the edges against the background.

Step 4 Dot Product: Intensify the contrast by performing a dot product of the smoothed image with the result from the Laplace transform from step 1.

Step 5 Addition for Final Sharpening: Enhance the sharpness and visibility of detail by adding the result of the dot product back to the original image.

Step 6 Histogram Equalization: Apply histogram equalization to distribute the histogram of the image uniformly, improving the overall contrast and making fine details more visible.

4.2. Image Enhancement

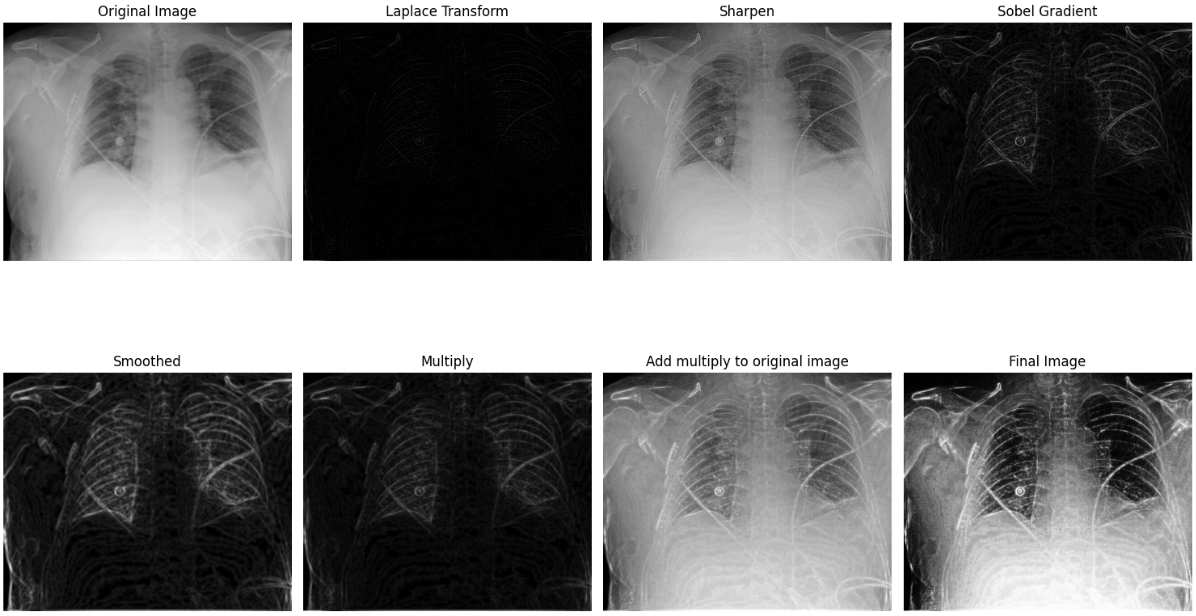


Figure 5: The image after a series of processing.

5. Proposed Method

5.1. Encoder-Decoder Approach

5.1.1. Features Embedding

We propose VisionDiagnostor centers around the implementation of Transformer encoder-decoder approach and deployed to evaluate methods having **ClinicalT5** [31] and **BioBART** [32] as encoder-decoder module (see Figure 6). **ClinicalT5**, based on the **T5** [33] architecture, and **BioBART**, a variant of the **BART** [34] architecture, have both been pre-trained on large of biomedical text data. These models stand out as the preeminent and potent pre-trained language models for the medical domain, ensuring the efficacy and robustness of our proposed method.

Object features: To extract object features in an image, we used the VinVL model to extract object features $R = \{r_1, r_2, ..., r_k\}$ from an image, with each r_i being a 2048-dimensional vector. Bounding box coordinates are normalized as $b_i = \left[\frac{x_i^{min}}{w}, \frac{y_i^{min}}{h}, \frac{x_i^{max}}{w}, \frac{y_i^{max}}{h}\right]$, forming $B_{\text{obj}} = \{b_1, b_2, ..., b_k\}$.

Final object features V_{obj} are computed by projecting R and B_{obj} to the language model dimension and summing the results:

$$V_{\text{obj}} = R' + B'_{\text{obj}} \tag{2}$$

We use ViT for visual feature extraction due to its ability to capture global information through its attention mechanism. By freezing ViT and projecting the last hidden state to match the language model's dimension, we obtain visual features V.

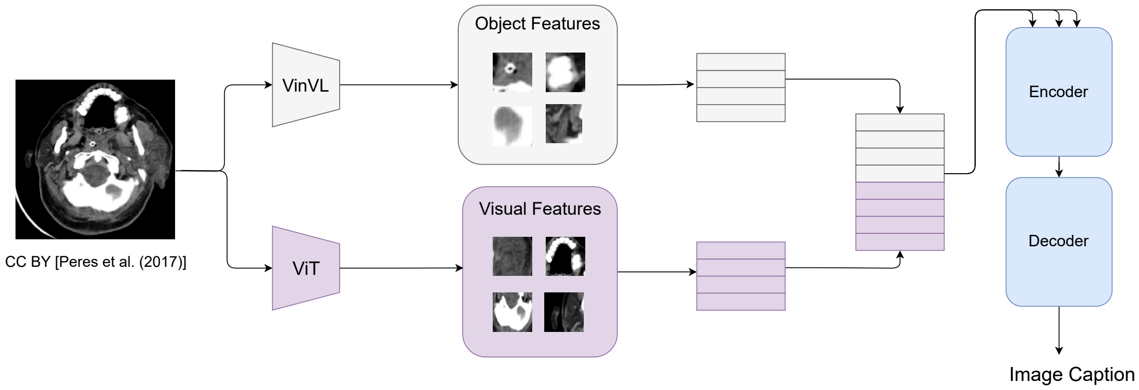


Figure 6: Overview of VisionDiagnostor.

The input embedding to the encoder-decoder module is:

$$\text{Input} = \text{Concat}(V, V_{\text{obj}}) \tag{3}$$

Where V are the visual features from ViT, and V_{obj} are the VinVL region object features. The $\text{Concat}(\cdot)$ function concatenates these features.

5.1.2. Encoder-Decoder Module

In this task, we employed the Transformer encoder-decoder architecture, which is used in ClinicalT5 [31] and BioBART [32] for the encoder-decoder module of VisionDiagnostor. The encoder receives the input features and then passes them to the decoder to generate the output sentence. In the decoder, attention mechanisms are employed, directing focus to both the output of the encoder and the input of the decoder.

Encoder

Multi-Head Attention:

$$\text{Attention}^{(\text{Enc})}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{4}$$

where Q, K, and V are the query, key, and value matrices, and d_k is the dimensionality of the key vectors.

Encoder Feed-Forward Network:

$$\text{FFN}^{(\text{Enc})}(x) = \text{ReLU}(xW_1^{(\text{Enc})} + b_1^{(\text{Enc})})W_2^{(\text{Enc})} + b_2^{(\text{Enc})} \tag{5}$$

where $W_1^{(\text{Enc})}$, $W_2^{(\text{Enc})}$, $b_1^{(\text{Enc})}$, and $b_2^{(\text{Enc})}$ are learnable parameters.

Encoder Layer Normalization:

$$\text{LayerNorm}^{(\text{Enc})}(x) = \text{LN}^{(\text{Enc})}(x + \text{LayerNorm}^{(\text{Enc})}(x)) \tag{6}$$

where $\text{LN}^{(\text{Enc})}$ is the layer normalization function.

Decoder

Decoder Self-Attention:

$$\text{Attention}^{(\text{Dec})}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V \tag{7}$$

where Q, K, and V are the query, key, and value matrices, and d_k is the dimensionality of the key vectors.

Decoder-Encoder Cross-Attention:

$$\text{Attention}^{(\text{Dec})}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V \tag{8}$$

where Q comes from the decoder and K, V come from the encoder.

Decoder Feed-Forward Network:

$$\text{FFN}^{(\text{Dec})}(x) = \text{ReLU}(xW_1^{(\text{Dec})} + b_1^{(\text{Dec})})W_2^{(\text{Dec})} + b_2^{(\text{Dec})} \tag{9}$$

where $W_1^{(\text{Dec})}$, $W_2^{(\text{Dec})}$, $b_1^{(\text{Dec})}$, and $b_2^{(\text{Dec})}$ are learnable parameters.

Decoder Layer Normalization:

$$\text{LayerNorm}^{(\text{Dec})}(x) = \text{LN}^{(\text{Dec})}(x + \text{LayerNorm}^{(\text{Dec})}(x)) \tag{10}$$

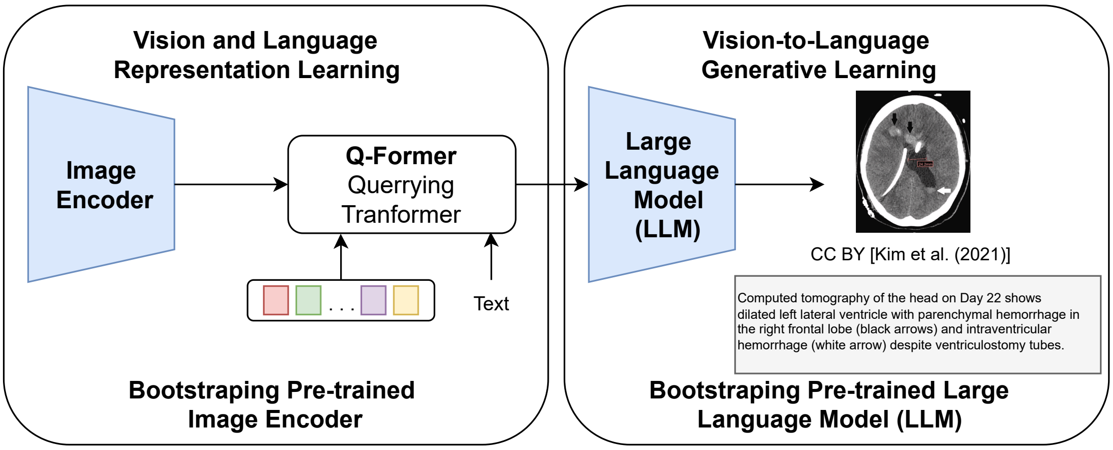
where $\text{LN}^{(\text{Dec})}$ is the layer normalization function.

5.2. Query Transformer Approach

Inspired by the BLIP2 architecture [35], we leveraged the Query Transformer (Q-Former) module, which serves as the trainable intermediary between a fixed image encoder and a fixed Large Language Model. It extracts a consistent number of output features from the image encoder, irrespective of the input image resolution. Q-Former comprises two transformer submodules that share self-attention layers: an image transformer for visual feature extraction from the fixed image encoder and a text transformer acting as both an encoder and decoder.

We initialize a set number of learnable query embeddings as input to the image transformer. These queries engage in self-attention and cross-attention interactions with each other and the frozen image features. Additionally, they can interact with the text through self-attention layers, with different attention masks applied based on the pre-training task.

In our experiments, we employ 64 queries, each with a dimensionality of 768, matching the hidden dimension of Q-Former. We utilize VIT-huge [36] as the frozen image encoder and BioMistral-7B [37] as the frozen LLM for caption generation, and we call it VisionDiagnostor-Q-BioMistral which is depicted in Figure 7. This bottleneck architecture, combined with our pre-training objectives, compels the queries to extract visual information most pertinent to the accompanying text.



Computed tomography of the head on Day 22 shows dilated left lateral ventricle with parenchymal hemorrhage in the right frontal lobe (black arrows) and intraventricular hemorrhage (white arrow) despite ventriculostomy tubes.

Figure 7: Overview of VisionDiagnostor-Q.

5.3. Evaluation Metrics

5.3.1. BERTScore

BERTScore is computed as proposed by Zhang et al. [38], where the cosine similarity of each hypothesis token j with each token i in the reference sentence is calculated using contextualized embeddings. Instead of using a time-consuming best-case matching approach, a greedy matching strategy is employed. The F1 measure is then calculated as follows:

$$R_{\text{BERT}} = \frac{1}{|\mathbf{r}|} \sum_{i \in \mathbf{r}} \max_{j \in \mathbf{p}} \cos(\vec{i}, \vec{j}), \tag{11}$$

$$P_{\text{BERT}} = \frac{1}{|\mathbf{p}|} \sum_{j \in \mathbf{p}} \max_{i \in \mathbf{r}} \cos(\vec{i}, \vec{j}), \tag{12}$$

$$\text{BERTScore} = F_{\text{BERT}} = \frac{2 \cdot P_{\text{BERT}} \cdot R_{\text{BERT}}}{P_{\text{BERT}} + R_{\text{BERT}}}. \tag{13}$$

The BERTScore correlates better with human judgments for the tasks of image captioning and machine translation.

5.3.2. Other Metrics

In addition to BERTScore, the competition also uses many other metrics such as ROUGE [39], BLEU-1 [40], BLEURT [41], METEOR [42], CIDEr [43], CLIPScore [44], RefCLIPScore [44], ClinicBLEURT [45] and MedBERTScore [46]. Applying a variety of these metrics helps us have a more accurate and general view of the model performance of participating teams. Each measure has its own advantages and provides a different perspective on text quality that makes it relevant in a medical context. This multi-dimensional evaluation helps identify outstanding models and gain an objective view of the competition.

6. Experiment Results

6.1. Experimental Configuration

All our proposed methods were trained and fine-tuned using the Adam optimization [47]. We utilized an A100-GPU setup with 80GB of memory to train models, taking 10 hours on average for each method.

We set the learning rate to 3e-05, dropout is set at 0.2, batch size is 32, and the training process is terminated after 3 epochs of not finding any reduction in the valid loss.

6.2. Main Result

Table 4
Performance comparison of different models on test set, VD stands for VisionDiagnostor.

Model	Model Size	BERTScore	ROUGE	BLEU-1	BLEURT	METEOR	CIDEr	CLIPScore	RefCLIPScore	ClinicalBLEURT	MedBERTScore
VD-Q-BioMistral	8B	0.6200	0.2139	0.1685	0.2913	0.0751	0.1585	0.8132	0.8014	**0.4597**	0.6233
VD-ClinicalT5	310M	0.5994	0.2363	**0.2323**	0.2954	**0.0989**	0.1442	**0.8244**	0.8100	**0.4597**	0.6016
VD-BioBART	227M	**0.6267**	**0.2452**	0.1950	**0.3060**	0.0889	**0.2243**	0.8184	**0.8117**	0.4562	**0.6292**

Table 4 presents a comprehensive of the results on the test set achieved by individual models, showcasing their BERTScore and other metrics. The findings underscore significant disparities in performance among the various models, providing valuable insights into their respective strengths and weaknesses. Notably, within the baseline models, VisionDiagnostor-BioBART stands out as the top performer, showcasing an impressive BERTScore of 0.6267 and almost all other metrics with the smallest size at 227M parameters. Moreover, Table 4 demonstrates that using large-scale pre-trained models in VisionDiagnostor-Q-BioMistral with a very large size (8B) does not result in significant performance improvement in this task.

7. Result Analysis

In this section, we conduct a subjective analysis of the **valid set** due to the limited number of submissions in the competition. This means that instead of using the test set for objective evaluation, we used the **valid set** to analyze the results our proposed methods achieved.

7.1. Impact of Image Pre-processing

Table 5
Results of models with image pre-processing in valid set. △ indicates the increase (↑) or decrease (↓) and compares without pre-processing (*).

Model	**BERTScore**
VisionDiagnostor-Q-BioMistral	0.6841*
△	↓ 0.0101
VisionDiagnostor-ClinicalT5	0.7071*
△	↓ 0.0166
VisionDiagnostor-BioBART	0.7165*
△	↑ 0.0198

Table 5 presents the results comparing the performance of the models with and without image pre-processing on the validation dataset, evaluated using BERTScore. Specifically, for the VisionDiagnostor-Q-BioMistral model, BERTScore decreased from 0.6841 to 0.6740 after applying pre-processing, corresponding to a decrease of 0.0101. Similarly, VisionDiagnostor-ClinicalT5 also saw a decrease in performance from 0.7071 to 0.6905, a decrease of 0.0166. In contrast, VisionDiagnostor-BioBART is the only model with an improvement with BERTScore increasing from 0.7165 to 0.7363, an increase of 0.0198.

Overall, applying image pre-processing does not appear to yield significant improvement for most models. Even for two of the three models (VisionDiagnostor-Q-BioMistral and VisionDiagnostor-ClinicalT5), image pre-processing degrades performance. The reason may be because of the input images are of good quality and have almost no noise. Some images also have clear instructions, such as arrows pointing to the relevant caption of the image (see Figure 1 in Section 3), making it easy for the model to understand and process the content without additional pre-processing.

7.2. Impact of Caption Length

Table 6
Group of caption length in valid set.

Group	Length (n)	Samples
Short	$n \leq 20$	5,520
Medium	$20 < n \leq 25$	2,179
Long	$25 < n \leq 30$	1,339
Very long	$n > 30$	934

The details of the test set based on different groups of lengths are in Table 6. Classification is done as follows:

- Short caption: These are captions shorter than 21 words.
- Medium caption: This group includes captions from 21 to 25 words.
- Long caption: Captions in this group from 26 to 30 words.
- Very long caption: This group contains captions longer than 30 words.

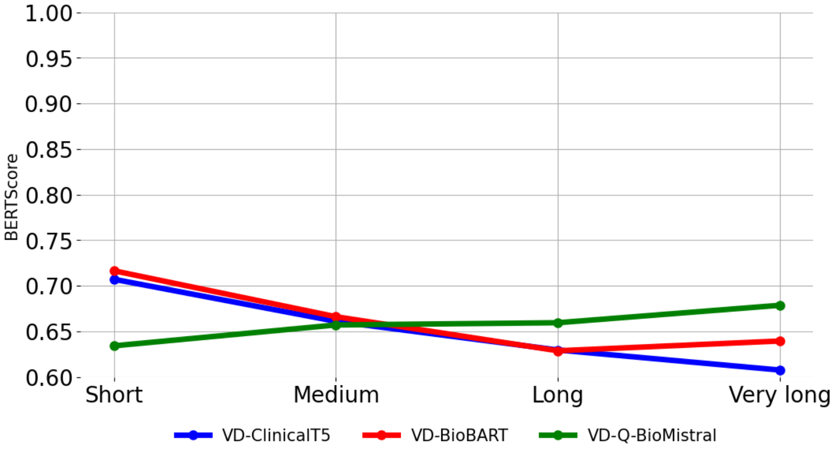


Figure 8: The results of models based on caption length.

The illustration from Figure 8 is an important step in gaining insight into the model's performance for different caption lengths. The results show that the length of the caption plays an important role in influencing model performance.

Specifically, the two models VisionDiagnostor-ClinicalT5 and VisionDiagnostor-BioBART based on the encoder-decoder method have similar trends, both showing a gradual decrease in BERTScore as the caption length increases. This may indicate a limitation in handling longer captions with this method.

It is worth noting that the VisionDiagnostor-Q-BioMistral model represents a different case, with performance increasing as the caption length increases. This may imply that this model is capable of handling longer captions more efficiently than other models, possibly due to its complexity and magnitude.

7.3. Impact of Object Features

According to papers from competing teams in previous years [48], [49], [29], the most popular image feature extraction methods today have two main directions: convolutional neural networks (CNN) and Vision transformers (ViT). Studies and demonstrations have shown that ViT often gives better results than CNN in the task of image captioning. ViT is capable of capturing long-term and global relationships in images more effectively, leading to the creation of richer and more accurate captions. However, to improve the quality of feature extraction further, we used the VinVL model. VinVL takes advantage of the power of the ability to detect and represent objects in images in detail. This allows the model to gain a deeper understanding of the context and elements in the image, thereby creating more accurate captions.

Table 7
Results of models with image pre-processing in valid set. △ indicates the increase (↑) or decrease (↓) and compares with models using object features (*).

Model	BERTScore
VisionDiagnostor-ClinicalT5	0.7071*
△	↓ 0.0239
VisionDiagnostor-BioBART	0.7165*
△	↓ 0.0321

Table 7 presents the results of the models when not using object features in the valid set. The figures show that not using object features significantly reduced the performance of the models.

Specifically, the VisionDiagnostor-ClinicalT5 model has a BERTScore of 0.7071 when using object features. However, when not using object features, the performance of this model drops by 0.0239. Similarly, the VisionDiagnostor-BioBART model also shows a significant decrease when not using object features, with BERTScore decreasing from 0.7165 to 0.6844, corresponding to a decrease of 0.0321.

These results indicate that using object features has an important effect in improving model performance. Object features can provide detailed and characteristic information about objects in images, helping models understand and describe images more accurately. Removing object features results in the loss of important information, reducing the model's ability to produce accurate and detailed captions, which in turn reduces BERTScore significantly.

8. Conclusion and Future Works

In this paper, we have proposed three different models to solve the task of medical image captioning, in other words medical image diagnosis, including VisionDiagnostor-ClinicalT5 and VisionDiagnostor-BioBART based on encoder-decoder architecture, VisionDiagnostor-Q-BioMistral based on BLIP2 architecture with Query Transformer which leveraging the power of Large Language Models (LLM).

Our results show that the VisionDiagnostor-BioBART model achieved third place on the leaderboard, with the highest BERTScore of 0.6267, despite being the smallest in size with only 227M parameters. Additionally, we performed analysis of the results to gain a deeper understanding of the factors that influence the performance of the models, including image pre-processing, caption length, and object features. These analyses have provided the comprehensive insight needed to shape and improve future methods and models for this task.

In future works, our objective is to delve deeper into the applications of other biomedical large language models (LLMs) BioMedLM [50], BioGPT [51], especially focusing on enhancing their capabilities to generate precise captions that are context-sensitive. This development will be pursued through methods like instruction tuning and better alignment of the models with specific user requirements. In addition, we plan to explore the integration of dense retrieval techniques into the biomedical image captioning process [52]. By adopting frameworks akin to Retrieval Augmented Generation, we intend to supplement the LLMs with an external, non-parametric memory using a FAISS index [53], thereby enriching their reasoning capabilities. Another area of interest will be investigating the interconnections between these approaches. We also anticipate evaluating the qualitative variations in the captions generated through these different methodologies to ascertain their efficacy and practicality in real-world applications.

Acknowledgment

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

References

[1] C. E. Lawson, J. M. Martí, T. Radivojevic, S. V. R. Jonnalagadda, R. Gentz, N. J. Hillson, S. Peisert, J. Kim, B. A. Simmons, C. J. Petzold, et al., Machine learning for metabolic engineering: A review, Metabolic Engineering 63 (2021) 34–60.

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

[3] 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. G. 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.

[4] 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.

[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: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2497–2506.

[6] G. Moschovis, Medical image captioning based on deep architectures, 2022.

[7] J. Pavlopoulos, V. Kougia, I. Androutsopoulos, A survey on biomedical image captioning, in: Proceedings of the second workshop on shortcomings in vision and language, 2019, pp. 26–36.

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

[9] Y. Wang, X. Zhang, T. Yang, J. Sun, Anchor detr: Query design for transformer-based detector, in: Proceedings of the AAAI conference on artificial intelligence, volume 36, 2022, pp. 2567–2575.

[10] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, et al., A survey of large language models, arXiv preprint arXiv:2303.18223 (2023).

[11] P. Zhang, X. Li, X. Hu, J. Yang, L. Zhang, L. Wang, Y. Choi, J. Gao, Vinvl: Revisiting visual representations in vision-language models, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 5579–5588.

[12] H. Kasban, M. El-Bendary, D. Salama, A comparative study of medical imaging techniques, International Journal of Information Science and Intelligent System 4 (2015) 37–58.

[13] J. A. Seibert, J. M. Boone, X-ray imaging physics for nuclear medicine technologists. part 2: X-ray interactions and image formation, Journal of nuclear medicine technology 33 (2005) 3–18.

[14] M. S. Atkins, B. T. Mackiewich, Fully automatic segmentation of the brain in mri, IEEE transactions on medical imaging 17 (1998) 98–107.

[15] 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.

[16] A. Bustos, A. Pertusa, J.-M. Salinas, M. De La Iglesia-Vaya, Padchest: A large chest x-ray image dataset with multi-label annotated reports, Medical image analysis 66 (2020) 101797.

[17] A. E. Johnson, T. J. Pollard, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, Y. Peng, Z. Lu, R. G. Mark, S. J. Berkowitz, S. Horng, Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs, arXiv preprint arXiv:1901.07042 (2019).

[18] V. Wijerathna, H. Raveen, S. Abeygunawardhana, T. D. Ambegoda, Chest x-ray caption generation with chexnet, in: 2022 Moratuwa Engineering Research Conference (MERCon), IEEE, 2022, pp. 1–6.

[19] S. Subramanian, L. L. Wang, B. Bogin, S. Mehta, M. van Zuylen, S. Parasa, S. Singh, M. Gardner, H. Hajishirzi, Medicat: A dataset of medical images, captions, and textual references, Findings of the Association for Computational Linguistics: EMNLP (2020).

[20] A. Graves, A. Graves, Long short-term memory, Supervised sequence labelling with recurrent neural networks (2012) 37–45.

[21] R. Dey, F. M. Salem, Gate-variants of gated recurrent unit (gru) neural networks, in: 2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS), IEEE, 2017, pp. 1597–1600.

[22] Z. Zhang, Y. Xie, F. Xing, M. McGough, L. Yang, Mdnet: A semantically and visually interpretable medical image diagnosis network, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 6428–6436.

[23] X. Wang, Z. Guo, C. Xu, L. Sun, J. Li, Imagesem group at imageclefmed caption 2021 task: Exploring the clinical significance of the textual descriptions derived from medical images, in: Conference and Labs of the Evaluation Forum, 2021. URL: https://api.semanticscholar.org/CorpusID:237298727.

[24] V. Kougia, J. Pavlopoulos, I. Androutsopoulos, AUEB NLP group at imageclefmed caption 2019, in: L. Cappellato, N. Ferro, D. E. Losada, H. Müller (Eds.), Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9-12, 2019, volume 2380 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2019. URL: https://ceur-ws.org/Vol-2380/paper_136.pdf.

[25] Y. Li, X. Liang, Z. Hu, E. P. Xing, Hybrid retrieval-generation reinforced agent for medical image report generation, in: S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 31, Curran Associates, Inc., 2018.

[26] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, et al., Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning, arXiv preprint arXiv:1711.05225 (2017).

[27] K. Singhal, S. Azizi, T. Tu, et al., Large language models encode clinical knowledge, Nature 620 (2023) 172–180.

[28] K. Singhal, T. Tu, J. Gottweis, R. Sayres, et al., Towards expert-level medical question answering with large language models, arXiv 2305.09617 (2023).

[29] P. Kaliosis, G. Moschovis, F. Charalambakos, J. Pavlopoulos, I. Androutsopoulos, Aueb nlp group at imageclefmedical caption 2023, in: CLEF2023 Working Notes, CEUR Workshop Proceedings, CEUR-WS. org, Thessaloniki, Greece, 2023.

[30] 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, 2024. URL: https://arxiv.org/abs/2405.10004v1. arXiv:2405.10004.

[31] E. Lehman, A. Johnson, Clinical-t5: Large language models built using mimic clinical text, PhysioNet (2023).

[32] H. Yuan, Z. Yuan, R. Gan, J. Zhang, Y. Xie, S. Yu, Biobart: Pretraining and evaluation of a biomedical generative language model, arXiv preprint arXiv:2204.03905 (2022).

[33] 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.

[34] 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, arXiv preprint arXiv:1910.13461 (2019).

[35] J. Li, D. Li, S. Savarese, S. Hoi, Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, in: International conference on machine learning, PMLR, 2023, pp. 19730–19742.

[36] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, R. Girshick, Masked autoencoders are scalable vision learners, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 16000–16009.

[37] Y. Labrak, A. Bazoge, E. Morin, P.-A. Gourraud, M. Rouvier, R. Dufour, Biomistral: A collection of open-source pretrained large language models for medical domains, arXiv preprint arXiv:2402.10373 (2024).

[38] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert, arXiv preprint arXiv:1904.09675 (2019).

[39] C.-Y. Lin, Rouge: A package for automatic evaluation of summaries, in: Text summarization branches out, 2004, pp. 74–81.

[40] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318.

[41] T. Sellam, D. Das, A. Parikh, Bleurt: Learning robust metrics for text generation, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 7881–7892.

[42] S. Banerjee, A. Lavie, 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, 2005, pp. 65–72.

[43] R. Vedantam, C. Lawrence Zitnick, D. Parikh, Cider: Consensus-based image description evaluation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 4566–4575.

[44] J. Hessel, A. Holtzman, M. Forbes, R. Le Bras, Y. Choi, Clipscore: A reference-free evaluation metric for image captioning, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, pp. 7514–7528.

[45] K. Huang, J. Altosaar, R. Ranganath, Clinicalbert: Modeling clinical notes and predicting hospital readmission, arXiv preprint arXiv:1904.05342 (2019).

[46] A. B. Abacha, W.-w. Yim, G. Michalopoulos, T. Lin, An investigation of evaluation methods in automatic medical note generation, in: Findings of the Association for Computational Linguistics: ACL 2023, 2023, pp. 2575–2588.

[47] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).

[48] A. Nicolson, J. Dowling, B. Koopman, A concise model for medical image captioning, in: CLEF2023 Working Notes, CEUR Workshop Proceedings, CEUR-WS. org, Thessaloniki, Greece, 2023.

[49] W. Zhou, Z. Ye, Y. Yang, S. Wang, H. Huang, R. Wang, D. Yang, Transferring pre-trained large language-image model for medical image captioning, in: CLEF2023 Working Notes, CEUR Workshop Proceedings, CEUR-WS. org, Thessaloniki, Greece, 2023.

[50] E. Bolton, A. Venigalla, M. Yasunaga, D. Hall, B. Xiong, T. Lee, R. Daneshjou, J. Frankle, P. Liang, M. Carbin, et al., Biomedlm: A 2.7 b parameter language model trained on biomedical text, arXiv preprint arXiv:2403.18421 (2024).

[51] R. Luo, L. Sun, Y. Xia, T. Qin, S. Zhang, H. Poon, T.-Y. Liu, Biogpt: generative pre-trained transformer for biomedical text generation and mining, Briefings in bioinformatics 23 (2022) bbac409.

[52] G. Moschovis, E. Fransén, Neuraldynamicslab at imageclefmedical 2022., in: CLEF (Working Notes), 2022, pp. 1487–1504.

[53] J. Johnson, M. Douze, H. Jégou, Billion-scale similarity search with gpus, IEEE Transactions on Big Data 7 (2019) 535–547.