Medical Image Captioning with Medical Pre-Trained Transformers

Diego Esteban Quintero Rey

Ingeniería de Sistemas y Computación

Visión de Máquina – 2023-2

ABSTRACT

This work proposes and implements an encoder-decoder architecture with an attention mechanism for generating medical captions from chest X-ray images. The encoder utilizes the MedViT model, a robust CNN-Transformer hybrid pre-trained on medical images, while the decoder employs GPT-2, fine-tuned on PubMed articles. The study evaluates the model's performance across three dataset versions: original images, images with histogram equalization, and images with both histogram equalization and adaptive masking. The Indiana University Chest X-Ray Collection dataset is used, consisting of 3305 frontal X-ray images paired with their corresponding medical reports. The dataset is pre-processed by resizing images, applying grayscale, and creating additional datasets with histogram equalization and adaptive masking. The MedViT model serves as the encoder, and features are transformed into a three-channel image before passing through a standard vision transformer. Experiments involve training with GPT-2 as the decoder and evaluating performance through metrics such as ROUGE, BLEU, and learning curves. Preliminary experiments highlight the influence of MedViT on generating relevant medical captions, leading to the main experiments. Results indicate no significant advantage from histogram equalization or adaptive masking. However, MedViT demonstrates superiority, emphasizing the need for understanding model parameters and further exploration. The investigation uncovers challenges in generalization and model exploration, with repeated captions indicating potential limitations. Recommendations include exploring alternative parameter configurations, extending training epochs, and addressing issues with the HuggingFace library. Despite the absence of clear advantages in some aspects, the study contributes insights into the complexities of utilizing pre-trained models for medical image captioning. Further refinement and exploration are necessary for advancing the model's capabilities and ensuring optimal implementation.

INTRODUCTION

Today, with the advances in digital health technology, hospitals and imaging centers produce an increasing number of medical images of different types. However, manually analyzing and understanding the content of medical images can take a lot of time and may require extensive medical experience and expertise. This raises the crucial need to develop automatic tools to discover and extract the relevant information from images which will then be used to comprehend the content of these images and deliver accurate descriptions, which gave birth to (medical) image captioning [1]. In the medical domain, image captioning consists in generating medical reports to highlight the most important clinical findings observed in the image [1].

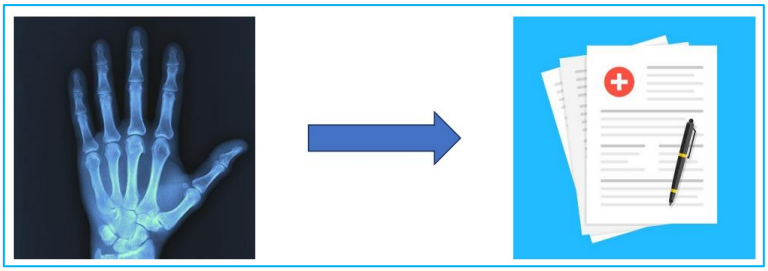

Figure 1 - A schematic of the medical image captioning task

Medical image captioning plays a pivotal role in modern healthcare, offering a multitude of benefits that significantly impact patient care, clinical practice, and healthcare systems. Some of their benefits include:

- *Clinical Decision Support:* Automated generation of medical reports and image captions can serve as valuable tools for clinicians and radiologists. They help in interpreting complex medical images, aiding in clinical decision-making, and improving the accuracy of diagnoses.
- *Efficiency and Productivity:* These technologies have the potential to enhance the efficiency of healthcare workflows. By automating the generation of reports and captions, healthcare professionals can save time and focus on more critical tasks, such as patient care and treatment planning.
- *Interpretability:* Medical image captioning and report generation can improve the interpretability of medical images for both healthcare professionals and patients. Natural language descriptions make it easier to understand and communicate findings, even for individuals without specialized medical training.
- *Standardization:* Automated reports and captions can help standardize the documentation of medical findings. This consistency in reporting can facilitate communication among healthcare providers, reduce the risk of errors, and ensure that essential information is not overlooked.
- *Education and Training:* These technologies can be valuable for medical education and training. They can serve as teaching tools to help students and new healthcare professionals learn to interpret medical images and understand common findings and abnormalities.
- *Scalability:* As the volume of medical imaging data continues to grow, automated captioning and report generation can scale to handle this increasing workload. This scalability is crucial for handling the demands of modern healthcare systems.

In recent years, the field of medical image captioning has witnessed significant advancements, driven by the intersection of computer vision and natural language processing (NLP) technologies [1]. This progress has been spurred by the recognition of the potential benefits of automatic report generation from medical imaging data. Several prototypes and models have been proposed to address this task. Notably, the early architectures addressing this problem were based on CNN-RNN models, as described in [2]. These models showed promise but were primarily effective for single-pathology tasks, highlighting the need for more sophisticated approaches.

The generation of medical image reports entails specific challenges not encountered in generic image captioning. In medical image captioning, the goal is not merely to describe objects and their relationships but to provide accurate clinical findings in a structured manner. Template-based methods have been proposed to address this issue, but they often result in fixed phrases and terminologies, making report generation time-consuming and non-trivial. Despite these challenges, there is considerable promise in the field of medical image captioning. Researchers have proposed hybrid approaches that combine generative and retrieval-based models, aiming to address some of the limitations in existing methods [3], however in this work the focus will be on **generative models**.

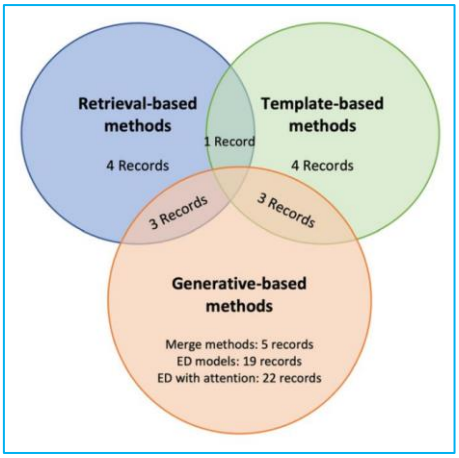

Figure 2 - Image captioning approaches classification [3]

Retrieval-based methods leverage a database of image captions. Upon receiving a new image, these approaches generate a caption by amalgamating the captions associated with images that exhibit the highest similarity to the input image. In contrast, template-based captioning operates with a repository of syntactic templates, such as those specifying a verb, an object, an action, and a complement. When an image is input, one or more object detection algorithms are employed to extract entities and their attributes. Subsequently, these extracted components are integrated into the template structure to construct the final caption. While this approach excels at producing grammatically accurate captions, it may lack adaptability and broad applicability. To address these limitations, generative models come into play. They employ an encoder-decoder architecture, uniting visual and natural language models to autonomously generate captions, thus enhancing flexibility and generalization.

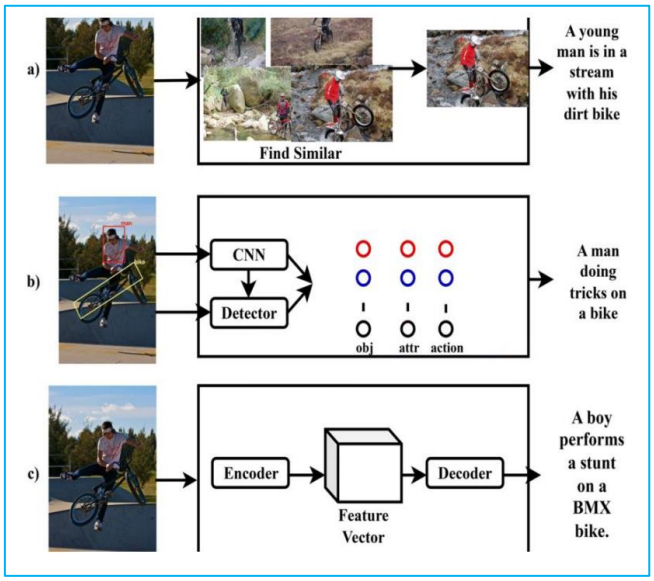

Figure 3 - A schematic of each of the image captioning approaches classifications: a) retrieval-based, b) template-based, c) generative models [17]

Generative deep learning models have excelled in image captioning competitions, NLP tasks, and medical computer vision analyses, but their black-box nature makes it challenging to understand how they make decisions. This lack of transparency becomes problematic, especially in the medical field, where it's crucial to know which features influenced a model's decision, particularly in cases of diagnostic failures. Health authorities often demand thorough scrutiny and complete transparency when adopting new AI technology, such as AI-based cancer diagnosis systems, to provide evidence and causal explanations for the system's decisions. Explainability becomes necessary when certain aspects of the

system can't be encoded into the model or when there's a disconnect between the model's output and stakeholders' expectations [1].

The field of medical image captioning offers promising opportunities for further research and improvement as suggested in [1] and [4]:

- Multi-Modal Integration: Integrating visual and textual information for a more comprehensive understanding of medical images.
- Reinforcement Learning: Exploring reinforcement learning to optimize evaluation metrics and enhance caption diversity.
- Transfer Learning: Leveraging pre-trained models from large-scale image-text datasets to improve captioning performance.
- Multilingual Captioning: Extending captioning models to generate captions in multiple languages for cross-lingual understanding.
- Improved Evaluation Metrics: Developing more sophisticated evaluation metrics that better align with human judgment and preferences.
- Interpretability: Enhancing explainability through explicit reasoning steps, attention localization, and advanced visualization techniques.

STATE OF THE ART

In this section the focus will be only on generative models for image captioning. First, there is a description of the most important types of generative models for this task according to [3] and then some specific state-of-the-art models that use or combine these approaches are mentioned:

Encoder-Decoder Models

These models typically consist of an image encoder (usually a CNN) that produces a fixed-length vector representation of the image, and a text decoder (often an RNN like LSTM) that generates descriptive captions. The key approaches and adaptations mentioned include:

- *Show-And-Tell Model:* This model, inspired by deep recurrent models for machine translation, utilizes CNN for image encoding and an LSTM for generating descriptions. It has been widely adopted in image captioning tasks, including medical image captioning [4].
- *Integration of Attributes:* Some models integrate high-level attributes with CNN-LSTM encoder-decoder architectures for image captioning. These attributes are learned and contribute to improving caption quality.
- *Multi-Modal Embeddings*: In certain cases, models combine textual and visual features through multi-modal embeddings and employ RNN decoders to generate sequences of words, such as for visual question answering.
- *Hierarchical Decoders*: Some models implement decoders in a multi-stage hierarchical manner to translate image features into text, enhancing the captioning process.
- *Specialized Architectures for Medical Images*: Several adaptations of encoder-decoder models are designed specifically for medical image captioning. These adaptations may include extensions of the Show-And-Tell model, the use of specialized features, and the integration of medical domain knowledge.

Encoder-Decoder with Attention Models

These models utilize attention mechanisms to focus on specific areas of interest within the images, enhancing the quality and relevance of generated captions. Several approaches include:

- *Show Attend and Tell (SAT)*: It's a model for image captioning that combines a convolutional neural network (CNN) for image feature extraction with an attention mechanism and a long short-term memory (LSTM) network for generating descriptive captions. The attention mechanism allows the model to focus on different parts of the image as it generates each word of the caption, resulting in more contextually relevant and detailed captions for images. It was introduced as an improvement to the traditional encoder-decoder models for image captioning.
- *Combined Bottom-Up and Top-Down Attention*: It integrates both bottom-up and top-down attention mechanisms. It calculates attention at the level of objects and salient image regions. Faster R-CNN is used for bottom-up attention, and attention distribution is predicted over image regions using task-specific context.
- *Context Level Visual and Textual Attention*: It includes a multi-attention encoder-decoder with a teacher forcing strategy. It combines context-level visual attention and textual attention from different views of X-ray images, facilitating the learning of heterogeneous semantic patterns.
- *Multimodal Integration*: Some studies combine healthcare data from various sources to improve clinical decisions. Models like the multitask CNN-RNN model with attention (Rodin et al., 2019) combine the analysis of images with recorded patient information, describing pathologies, their locations, and severity in medical reports.
- *Co-Attention and Hierarchical LSTM*: It uses a co-attention and hierarchical LSTM to focus on abnormal findings. They combine feature differences between normal and abnormal cases with visual and textual information for diagnosis generation.
- *Adaptive Multi-Modal Attention*: It proposes an adaptive multi-modal attention network to describe local properties in ultrasound images. This approach generates captions based on stored memories in the LSTM decoder, combining visual features with semantic features.
- *Adversarial Autoencoders*: This model justifies medical image diagnoses by leveraging textual and visual evidence from the nearest alternative diagnosis. The model creates an intermediate space bridging text and images, employing a text-to-image Adversarial Regularized Auto-encoder (ARAE) model. During inference, mapping from the visual input to this intermediate space is performed using a CNN with an attention mechanism, and the decoder generates the diagnosis.
- *Feature Transfer and Fusion*: This model addresses the challenge of limited annotated radiology report datasets by proposing a method to transfer visual representations learned on small datasets for report generation. They utilize an encoder-decoder model with an attention mechanism, focusing on feature transfer and fusion models for thoracic disease classification.
- *Visual Transformers as Encoders*: This approach uses a strategy where images are divided into patches. These patches are processed by a visual image transformer (ViT), serving as an image encoder. Captions are then generated using a self-attention based PubMedBERT model as the decoder.
- *CNN-GRU Hybrid Model*: it combines a CNN encoder model with an attention-based GRU language generator model for the caption prediction task in medical imaging.

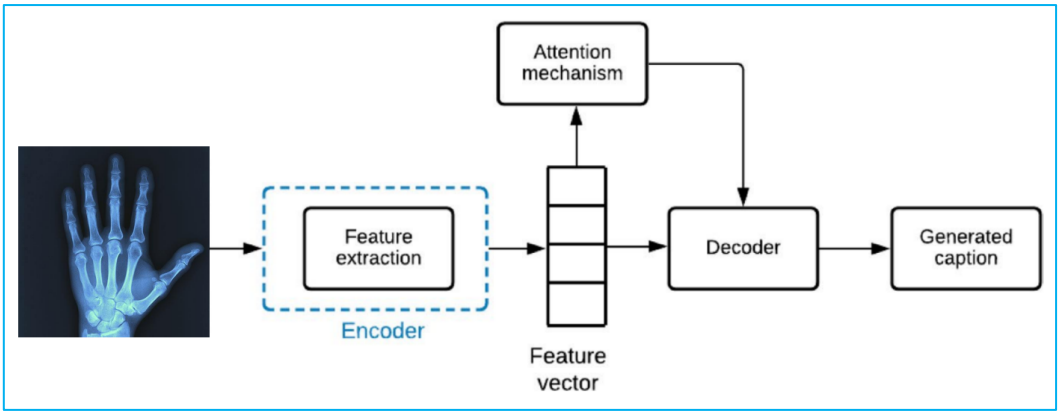

Figure 4 - Encoder-Decoder with Attention Architecture for Image Captioning [3]

Merge Models

Merge models utilize Convolutional Neural Networks (CNNs) to extract visual features from images and Recurrent Neural Networks (RNNs) to learn textual features from associated text. These two types of features, visual and textual, are then merged or fused together to create a joint representation vector. This joint representation vector encodes significant features from both modalities in the same embedding space. Once the joint embeddings are obtained, a decoder component is used to generate new captions based on these embeddings. The decoder leverages the merged features to create coherent and contextually relevant captions. Merge models can be considered as a variant of encoder-decoder-based models but with an explicit feature fusion module. While encoder-decoder models focus on encoding and decoding sequences of data, merge models emphasize the fusion of features from different modalities [3].

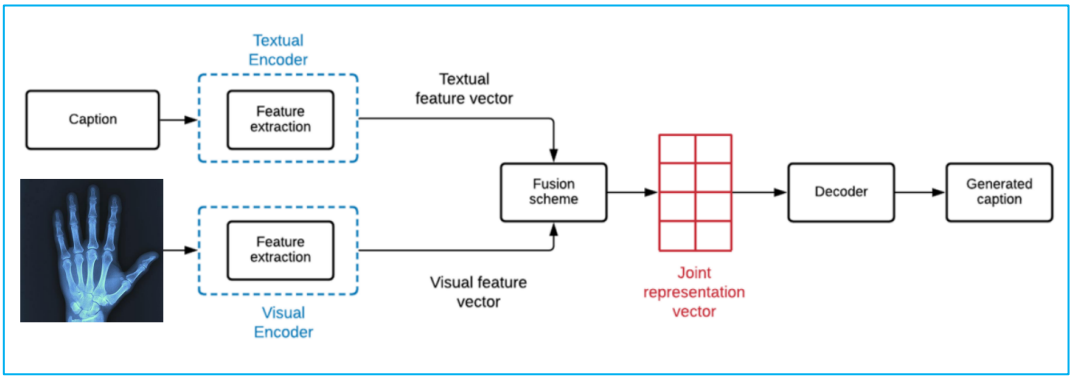

Figure 5 Merge Models Architecture for Image Captioning [3]

Some state-of-the-art models that combine these approaches are listed below:

- *SAT GPT-3 merge model*: The first approach model presented in [2] uses SAT encoder decoder with attention model and GPT-3 models separately to generate the scores for the next word in a sequence. Then both scores are fed to a feed forward neural network to get the actual word that is to be included in the generated report.
- *SAT GPT-3 stacked model*: The second approach model presented in [2] uses SAT encoder decoder with attention model which takes an image as input and generates a report using both CNN and LSTM neural networks. The generated text is then fed to a GPT-3 model to continue generating the text.
- *Clinically Coherent Reward Mechanism*: The model first encodes input images into image embedding maps, which are then used as input to a sentence decoder that generates topics for

sentences in a recurrent manner. The word decoder then generates the final caption sequence from the topic, with attention on the original images [8].

- *MLTL-LSTM Framework:* is a multi-step approach for enhancing medical image captioning. First, it establishes a foundational MLTL (Multi-Level Transfer Learning) framework, which includes a Generalized Feature Extraction (GFE) network to train on readily available non-medical images, an Inter-tune MedCNN to work on medical images related to the target domain, and a Target Fine-tune MedCNN to classify the limited medical datasets. Then, it incorporates an LSTM (Long Short-Term Memory) model for generating detailed captions. Additionally, the framework involves further refinement through a fine-tuned multi-input CNN model and feature extraction techniques to improve the quality and relevance of captions for medical images [9].
- *BERTHop*: It uses a PixelHop++ model followed by a "PCA and concatenation" block to generate Q feature vectors. These features along with language embedding are fed to the transformer that is initialized with BlueBERT, a transformer specialized in the medical domain [7].
- *Dual Curriculum Learning for Image Captioning*: The model consists of two branches: one for processing image data, utilizing a fine-tuned VGG-16 and fully connected layers, and the other for processing text, embedding tokenized words using Word2vec embeddings and passing them through a recurrent neural network. The outputs from both branches are combined and employed to predict the next word in the caption generation sequence [10].

PROPOSAL

This work entails implementing and encoder-decoder architecture with an attention mechanism. For this, the medical domain pre-trained visual transformer of [6] as the encoder and a medical domain pre-trained transformer for the decoder as suggested in [7]. The performance of the model is assessed across three versions of the same dataset: original images, images with histogram equalization, and images with both histogram equalization and the adaptive masking technique introduced in [5].

DATASET

The Indiana University Chest X-Ray Collection IU X-Ray [11]: It's a dataset that consists of chest X-ray images and their corresponding medical reports. These images and reports were obtained from the Open Access Biomedical Image Search Engine (OpenI). The dataset comprises both frontal and lateral images; however, only the frontal images are utilized in this study. Specifically, there are a total of 3822 frontal images in the dataset, with 3305 accompanied by reports. Consequently, the dataset for this study is comprised of 3305 frontal images, each paired with its corresponding report. Each report is generated by concatenating the findings and impressions of texts corresponding to each sample. The decision to exclusively utilize frontal images was made to facilitate a direct comparison of performance with the dataset incorporating the adaptive masking technique, as proposed in [5]. This masking method is specifically designed for frontal images, hence the focus on this subset of the dataset for a more meaningful evaluation.

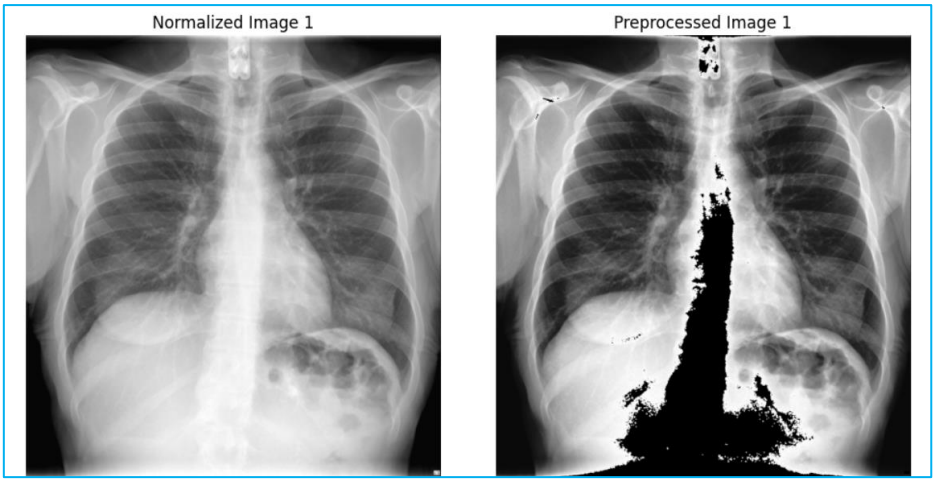

Figure 5 – Original image (left) and pre-processed image after applying histogram equalization and adaptative masking (right)

PRE-PROCESSING

The images were resized to 224×224 pixels and converted to grayscale. Two additional image datasets were created by first applying histogram equalization, and then applying both histogram equalization and the adaptive masking method proposed in [5]. The adaptive masking employs a threshold defined as $\theta = \min + 0.9 \cdot (\max - \min)$ to remove the diaphragm from these images, with a morphological closure in between to ensure proper refinement of the segmentation. The three datasets were persisted in HDF5 files and uploaded to [12].

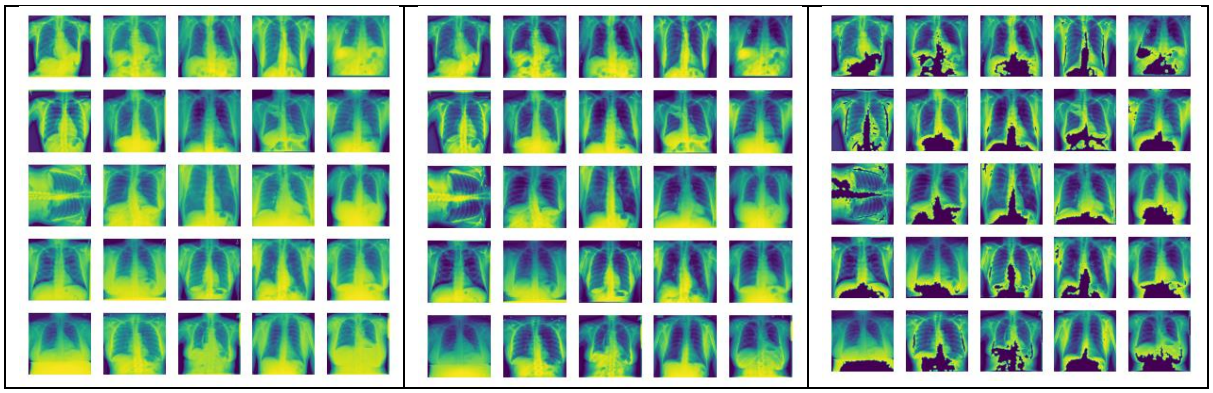

Figure 6 – A set of samples from the image dataset: original images (left), images with histogram equalization (center), and images with both histogram equalization and adaptive masking (right).

For the text pre-processing, only two functions were applied. One that removes a specific kind of substring encountered in the reports ("XXXX") and extra blank spaces. Further text pre-processing is handled by the tokenizer.

ENCODER

For the encoder, the MedViT model proposed in [6] was used. It's a robust and efficient CNN-Transformer hybrid model equipped with the locality of CNNs and the global connectivity of vision transformers. This model was pre-trained on the MedMNIST-2D datasets [13] for multiclass classification.

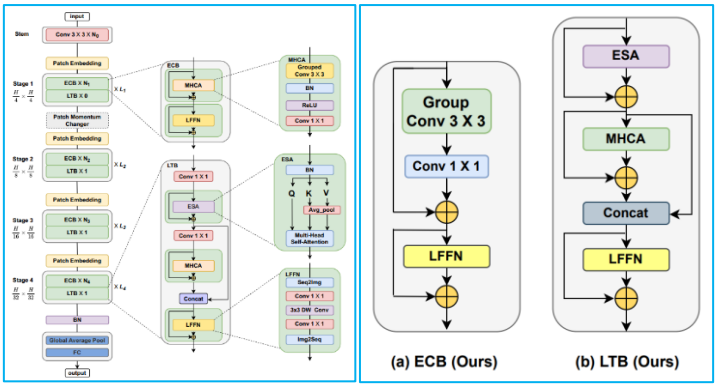

Figure 6 – MedViT architecture (left) and MedViT Efficient Convolution Block (ECB) and Local Transformer Block (LTB) units (right)

The datasets employed by MedViT encompass various medical image collections from diverse domains, including ChestMNIST, a dataset resembling the one used in the study, containing exclusively frontal images [11]. This dataset comprises 112120 frontal X-ray images from a total of 32717 patients, representing 14 different disease classes [6].

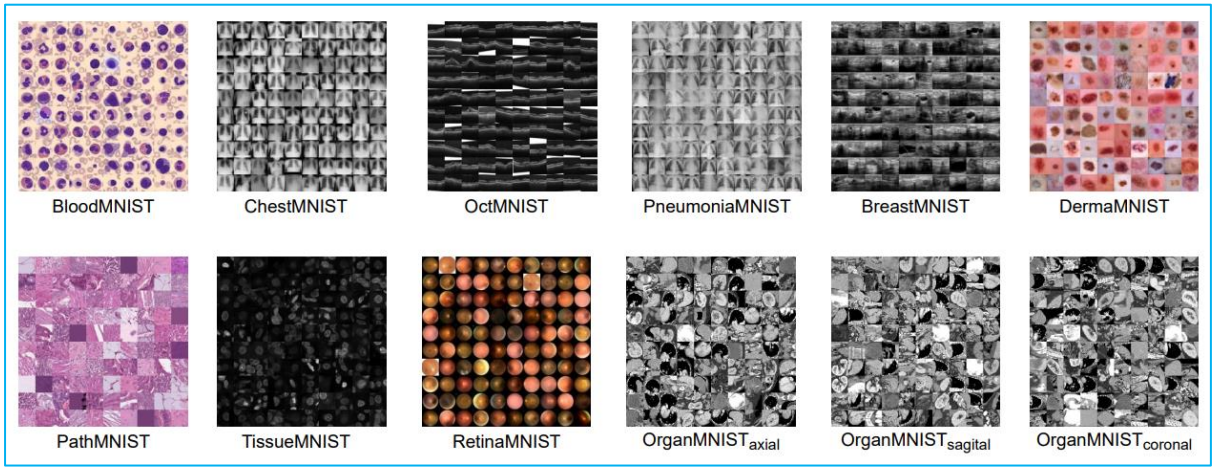

Figure 7 - MedMNIST-2D Classification Datasets

The MedViT model was downloaded from the original repository [14], opting for the pre-trained large version, which was initialized and employed. A custom forward pass method was implemented [12], disabling gradient computation and removing the classification head. The features extracted for each image were vectors of size 1024. Images were processed in batches of 400 samples during inference to prevent exceeding the RAM consumption limits of Google Colab. A total of 9 files per dataset were persisted and subsequently uploaded to [12].

These features were compiled into a three-channel image of size 224×224×3 by replicating the same 1024=224×224 matrix three times. Then, this composite image is passed to a standard vision transformer [16], implementing transfer learning in doing this. The main reason behind introducing this supplementary vision transformer is to establish a more convenient connection between the encoder and the decoder via HuggingFace. This decision was driven by the unavailability of the MedViT model in the HuggingFace library.

DECODER

For the decoder, the captions corresponding to each of the images were tokenized using GPT-2 specifying a maximum length of $\lfloor 40 \times 1.5 \rfloor$ and passed to this same GPT-2 model to generate the captions. The maximum length was set experimentally given the fact the average number of words for the reports of the constructed dataset is approximately 38.69. However, better alternatives to set this parameter should be considered in future works.

The three image datasets are trained using GPT-2 as the decoder. The best performant dataset was chosen to do a final training but this time using gpt-2-pubmed-medium model as the decoder. This one is a GPT-2 (355M model) finetuned on 0.5M PubMed abstracts [18]. To use this model, it's necessary to clone the repository and add this key to the configuration JSON file: "model_type": "gpt2".

The architecture was connected using the Vision Encoder Decoder Model class from HuggingFace.

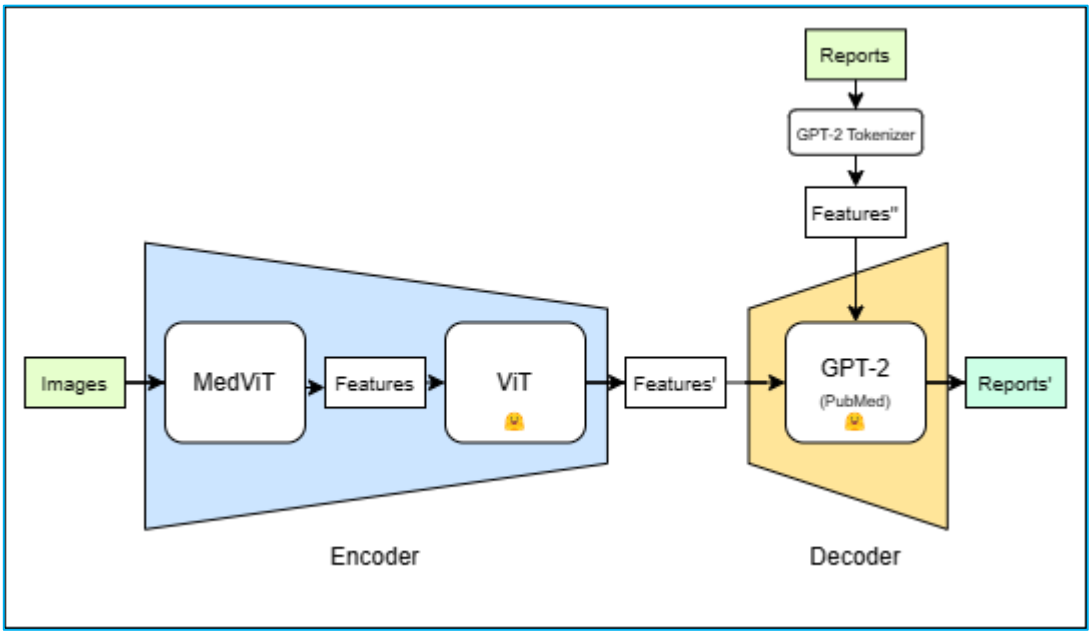

Figure 8 - Proposed model architecture

EXPERIMENTS

An initial experiment was conducted using a subset of 297 images, with 20% allocated for validation, and training over 10 epochs. This experiment excluded the MedViT model as an encoder, instead passing the images directly to the standard vision transformer. Subsequently, the same experiment was repeated, but with MedViT serving as the encoder before the vision transformer. The aim was to assess whether there were differences in the captions generated by the two approaches and to evaluate if the medical pre-trained vision transformer conferred an advantage in the quality of the generated text. Another experiment, using the same number of images, employed MedViT as the encoder but with a reduced number of epochs, solely for the purpose of comparing the quality of the generated captions.

Table 1 - Preliminary Experiments to assess quality of captions

Exp.	Dataset (Size)	Encoder	Decoder	Epochs
1	Original (297)	ViT	GPT-2	10
2	Original (297)	MedViT + ViT	GPT-2	2
3	Original (297)	MedViT + ViT	GPT-2	10

Table 2 - Actual experiments on the whole constructed dataset

Exp.	Dataset (Size)	Encoder	Decoder	Epochs

4	Original (3305)	MedViT + ViT	GPT-2	10
5	Equalized (3305)	MedViT + ViT	GPT-2	10
6	Equalized + Adaptative Masking (3305)	MedViT + ViT	GPT-2	10
7	Equalized + Adaptative Masking (3305)	MedViT + ViT	GPT-2 (PubMed)	10

Following this preliminary phase, the main experiments were executed. The three constructed datasets were trained using the proposed encoder-decoder with attention architecture, with GPT-2 employed as the decoder. The dataset exhibiting the best performance was then utilized in a final experiment, where the gpt-2-pubmed-medium model replaced the bare GPT-2 model as the decoder. For these experiments, the dataset was split differently: 90% of the 3305 samples were allocated for training, and 10% for testing. During training evaluation (validation), only 1% of the training samples were used. The choice of allocating 10% for testing, instead of the typical 20%, was influenced by the relatively small size of the dataset compared to those used in similar studies (e.g., [2]). This partitioning decision aimed to augment the size of the training samples to some extent. The use of 1% for validation, though unconventional, was intended to reduce training time, especially given the relatively high inference cost of this model. It's worth noting that in [2], the authors allocate 0.25% for testing (not validation), implying that 1% for validation may be deemed acceptable. The purpose of performing validation in this work was to obtain an overview of metric evolution during training, rather than conducting an in-depth analysis of generalization capabilities and overfitting.

The complete implementation, including model parameters and results, can be found in the notebooks uploaded to [12].

RESULTS

Table 3 presents examples of generated captions from the preliminary experiments. In Experiment 1, the captions were found to be unrelated to the medical domain. In Experiment 2, where the MedViT transformer was employed and trained for only 2 epochs, the generated caption exhibited content relevant to the medical domain. This observation suggests that the use of the medical pre-trained visual transformer contributes to generating captions aligned with the project's task. However, the generated caption in Experiment 2 resembled content from a news article in a newspaper or magazine. Encouragingly, extending the training to 10 epochs resulted in a caption that more closely resembled an authentic medical report, incorporating vocabulary such as "pneumothorax" or "pleural effusion" commonly found in the dataset reports. This indicates that leveraging the medical pre-trained vision transformer does indeed enhance the generation of somewhat relevant captions, providing a positive signal to proceed with the full-scale experiments on the entire dataset.

Table 3 - Examples of generated captions from preliminary experiments

Exp.	Example Generated Caption
1	`I'm not going to lie to you," he said. "I'm just going to tell you what I think. I'm going to be honest with you. I don't think you're going to believe me." "I don't know what you're talking about," she replied. "You're just saying that you think I'm crazy. You don't believe me`
2	`"We are very pleased with the results of the study," said lead author Dr. Michael J. Schoenfeld, MD, professor of medicine at the University of California, San Francisco School of Medicine. "This is the first time that we have seen a significant increase in the incidence of acute myoca`
3	`The heart is normal in size and contour. The mediastinum is unremarkable. There is no pneumothorax or pleural effusion. No acute cardiopulmonary abnormality. No evidence of acute disease. 1. No focal airspace disease. 2. Mild pulmonary edema. 3. Mild nodule nodule disease.`

Table 4 illustrates the learning curves during the training process for the final four experiments. Experiments 4, 5 and 6 exhibit similar results, indicating that there doesn't appear to be a substantial improvement when applying histogram equalization and/or adaptive masking, at least based on an initial observation of the loss curves.

In contrast, the Experiment 7 demonstrates a notable achievement as the curve reaches a lower training loss value, falling below 0.5. However, it's important to note that this experiment employed nearly three times the number of steps compared to Experiments 4, 5 and 6. This extended duration could be attributed to either the depth of the medical pre-trained transformer used or a reduction in batch size from 8 to 3, implemented to prevent exceeding the RAM consumption limits of Google Colab.

If the prolonged training is due to the transformer's depth, it's noteworthy that around steps 3000-3500, the training loss falls within the range of 0.5-1.0, like the lowest values reached by the Experiments 4, 5 and 6. Alternatively, if the extended training time is a result of the reduced batch size, it suggests that batch size might be a parameter worth exploring with different values to identify the range that produces optimal results for the model.

Table 4 - Learning Curves of the final four experiments

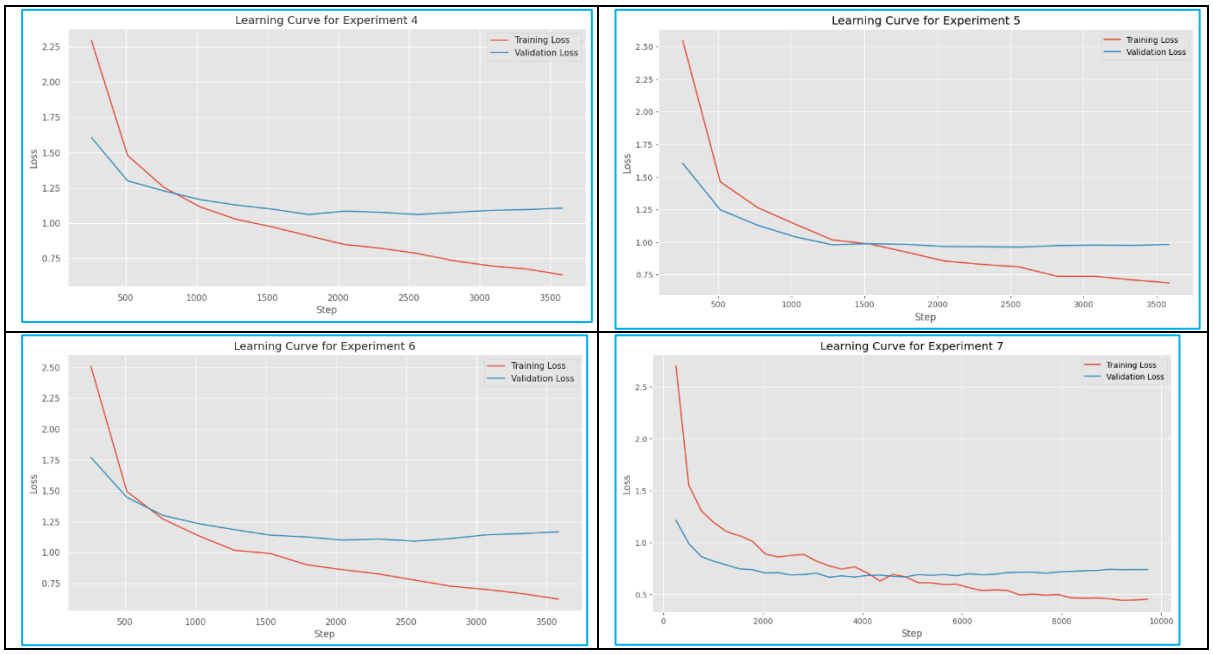

Table 5 shows the ROUGE-L metric results for the final four experiments evaluated on the test dataset. This metric provides insights into how well the generated report captures the key information and overall meaning of the original text. Precision measures the proportion of n-grams (sequences of n words) that are correctly identified in the generated report. In other words, it indicates how accurate the summary is in terms of including relevant information. A higher precision score suggests that the summary is more concise and focused on the essential elements of the text. Recall measures the proportion of n-grams in the reference report that are correctly identified in the generated report. In other words, it indicates how complete the summary is in terms of capturing all the important information from the original text. A higher recall score suggests that the summary is more comprehensive and covers a broader range of relevant information. F-measure is a harmonic mean of precision and recall.

According to these results, the medical pre-trained transformer fine tuned on PubMed articles doesn't seem to have an advantage over the other experiments. In particular, the experiment that shows slightly better results is Experiment 6 one that used the images dataset pre-processed with histogram

equalization and the adaptative masking. However, the difference doesn't seem to be that significant. It is observed for example that the precision was slightly smaller in Experiment 6, but with an increase in the recall, it made the F-Measure increase as well.

Table 5 - ROUGE-L metric results for the final four experiments

#	ROUGE-L Precision	ROUGE-L Recall	ROUGE-L F-Measure
4	0.18930	0.20868	0.19831
5	0.18625	0.21916	0.20123
6	**0.18402**	**0.23543**	**0.20646**
7	0.18140	0.23156	0.20334

Table 6 displays the BLEU metric results evaluated on the test dataset up to 4-grams, measuring the accuracy of the machine translation system in predicting the correct n-grams. An n-gram refers to a sequence of n words from a sentence. As anticipated, the results for 4-grams are smaller than those for 1, 2, and 3-grams.

There is no definitive conclusion on which experiment yields better results, as they exhibit considerable similarity. However, what is evident is that, once again, the medical pre-trained transformer does not appear to provide a distinct advantage over the GPT-2 model. Additionally, there is no apparent enhancement in the preprocessing of the datasets, indicating that histogram equalization and adaptive masking do not seem to contribute to improvement under these experimental conditions, at least concerning this metric.

Table 6 - BLEU metric results for the final four experiments

#	BLEU-1	BLEU-2	BLEU-3	BLEU-4
4	**0.29448**	0.09582	0.04283	0.02194
5	0.29150	0.09761	**0.04890**	**0.02785**
6	0.28563	**0.09765**	0.04639	0.02455
7	0.28250	0.09556	0.04704	0.02682

When it comes to the quality of the generated captions, as the metrics show, they are very similar among Experiments 4, 5, 6, and 7. Table 7 shows the generated captions for the real caption presented below. In green, it's highlighted those words that appear in the real caption.

Real Caption: `Heart size normal. Lungs are clear. are normal. No pneumonia, effusions, edema, pneumothorax, adenopathy, nodules or masses. Normal chest`

Table 7 - Generated captions for one example of experiments 4, 5, 6, and 7

Exp.	Generated Caption
4	`The heart is normal in size. The mediastinum is unremarkable. The lungs are clear. No acute`
5	`The cardiomediastinal silhouette is within normal limits for size and contour. The lungs are normally inflated without evidence`
6	`The heart and lungs have in the interval. Both lungs are clear and expanded. Heart and mediastinum normal. No active disease`
7	`The lungs are clear bilaterally. Specifically, no evidence of focal consolidation, pneumothorax, or pleural effusion.. Card`

Some things to notice is that the generated captions are cut without finishing the sentences. However, this may be consequence of the way the maximum length of the generated sequences was defined.

It was observed that in all experiments, there are some captions that are generated very repeatedly, such as the one presented below. This suggests that the model is lacking some degree of generalization or exploration in the generation of captions.

Repeated Caption: `The lungs are clear bilaterally. Specifically, no evidence of focal consolidation, pneumothorax, or pleural effusion.. Cardio mediastinal silhouette is unremarkable. Visualized osseous structures of the thorax are without acute abnormality.`

Table 8 shows a repeated caption that was generated by experiment 7. The caption is generated around 80 out of 331 times (test dataset size). For some real captions it may be quite similar, but for others there are a few matches. This may be an explanation to the low results for both the ROUGE and BLEU metrics.

Table 8 – Repeated generated caption

Real	Generated
Stable calcified hilar and granulomas. Lungs are clear bilaterally. There is no focal consolidation, pleural effusion, or pneumothoraces. Cardiomediastinal silhouette is within normal limits. are unremarkable. No acute cardiopulmonary abnormality.	The lungs are clear bilaterally. Specifically, no evidence of focal consolidation, pneumothorax, or pleural effusion.. Cardio mediastinal silhouette is unremarkable. Visualized osseous structures of the
Heart size mildly enlarged, stable mediastinal and hilar contours. No focal alveolar consolidation, no definite pleural effusion seen. No typical findings of pulmonary edema. No pneumothorax. Mild cardiomegaly, no acute pulmonary findings	The lungs are clear bilaterally. Specifically, no evidence of focal consolidation, pneumothorax, or pleural effusion.. Cardio mediastinal silhouette is unremarkable. Visualized osseous structures of the thorax are

All the generated captions in the test dataset can be seen in the notebooks uploaded to [12].

CONCLUSIONS

The investigation revealed no distinct advantage in the use of histogram equalization or histogram equalization with adaptive masking for enhancing the quality of generated captions and the metrics associated with the training and evaluation of the model architecture. Similarly, no clear advantage emerged from utilizing the medical pre-trained transformer fine-tuned on PubMed articles' abstracts.

Conversely, a noticeable advantage was observed when employing the MedViT medical pre-trained visual transformer, as confirmed in the preliminary experiments. To advance our understanding and refine the model architecture, dedicated exploration is imperative. A comprehensive study and comprehension of the model components are necessary, particularly delving into the meaning of parameters and the inner architecture of the models used. Clear comprehension of the mechanisms behind tokenizers and their correct setup, along with additional experimentation involving different parameter configurations (e.g., varying batch size or the maximum length of generated captions), is crucial. Particular attention should be given to investigating alternatives for defining the maximum length of generated captions.

An oversight was noted in not using a fixed test dataset for the four final experiments, complicating the comparison of generated caption quality among models. However, the clarity obtained from metrics results and captions suggests no major differences between the models in this regard. Consideration should be given to training the model for more than 10 epochs, as this may potentially yield improved results. Also, using the large version of the PubMed fine tuned model, may be beneficial.

Further study is warranted, especially concerning the HuggingFace library, to address potential deprecations or identify easier implementation methods for certain constructs used in the project. This effort will contribute to optimizing the utilization of the library and ensuring alignment with current best practices.

REFERENCES

[1] Beddiar, R., & Oussalah, M. (2023). Explainability in medical image captioning. In J. Benois-Pineau, R. Bourqui, D. Petkovic, & G. Quénot (Eds.), Explainable deep learning AI (pp. 239-261). Academic Press. https://doi.org/10.1016/B978-0-32-396098-4.00018-1

[2] Selivanov, A., Rogov, O.Y., Chesakov, D. et al. Medical image captioning via generative pretrained transformers. Sci Rep 13, 4171 (2023). https://doi.org/10.1038/s41598-023-31223-5

[3] Beddiar DR, Oussalah M, Seppänen T. Automatic captioning for medical imaging (MIC): a rapid review of literature. Artif Intell Rev. 2023;56(5):4019-4076. doi: 10.1007/s10462-022-10270-w. Epub 2022 Sep 17. PMID: 36160365; PMCID: PMC9483422.

[4] Thakare, Y. A., & Walse, K. H. (2023). A review of Deep learning image captioning approaches. Journal of Integrated Science and Technology, 12(1), 712. Retrieved from https://pubs.thesciencein.org/journal/index.php/jist/article/view/a712

[5] Giełczyk A, Marciniak A, Tarczewska M, Lutowski Z. Pre-processing methods in chest X-ray image classification. PLoS One. 2022 Apr 5;17(4):e0265949. doi: 10.1371/journal.pone.0265949. PMID: 35381050; PMCID: PMC8982897.

[6] Nejati Manzari, O., Ahmadabadi, H., Kashiani, H., Shokouhi, S. B., & Ayatollahi, A. (2023). MedViT: A robust vision transformer for generalized medical image classification. Computers in Biology and Medicine, 157, 106791. https://doi.org/10.1016/j.compbiomed.2023.106791.

[7] Monajatipoor M, Rouhsedaghat M, Li LH, Kuo CJ, Chien A, Chang KW. BERTHop: An Effective Vision-and-Language Model for Chest X-ray Disease Diagnosis. Med Image Comput Comput Assist Interv. 2022 Sep;13435:725-734. doi: 10.1007/978-3-031-16443-9_69. Epub 2022 Sep 16. PMID: 37093922; PMCID: PMC10120542.

[8] Liu, G., Hsu, T. M. H., McDermott, M., Boag, W., Weng, W. H., Szolovits, P., & Ghassemi, M. (2019, October). Clinically accurate chest x-ray report generation. In Machine Learning for Healthcare Conference (pp. 249-269). PMLR.

[9] Aswiga RV, Shanthi AP. A Multilevel Transfer Learning Technique and LSTM Framework for Generating Medical Captions for Limited CT and DBT Images. J Digit Imaging. 2022 Jun;35(3):564-580. doi: 10.1007/s10278-021-00567-7. Epub 2022 Feb 25. PMID: 35217942; PMCID: PMC9156604.

[10] Alsharid M, El-Bouri R, Sharma H, Drukker L, Papageorghiou AT, Noble JA. A Course-Focused Dual Curriculum For Image Captioning. Proc IEEE Int Symp Biomed Imaging. 2021 Apr;2021:716-720. doi: 10.1109/ISBI48211.2021.9434055. Epub 2021 May 25. PMID: 34413932; PMCID: PMC7611521.

[11] RADDAR. "Chest X-rays (Indiana University)." Kaggle, https://www.kaggle.com/datasets/raddar/chest-xrays-indiana-university. Accessed 31 Oct. 2023.

[12] este6an13 (2023). Transformers Image Captioning. GitHub. https://github.com/este6an13/transformers-image-captioning

[13] Yang, J., Shi, R., Wei, D. et al. MedMNIST v2 - A large-scale lightweight benchmark for 2D and 3D biomedical image classification. Sci Data 10, 41 (2023). https://doi.org/10.1038/s41597-022-01721-8

[14] Omid-Nejati (2023). MedViT: A robust vision transformer for generalized medical image classification. GitHub. https://github.com/Omid-Nejati/MedViT.

[16] Dosovitskiy, A. et al. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. Hugging Face. https://huggingface.co/google/vit-base-patch16-224.

[17] Verma, A., Yadav, A.K., Kumar, M. et al. Automatic image caption generation using deep learning. Multimed Tools Appl (2023). https://doi.org/10.1007/s11042-023-15555-y

[18] Papanikolaou, Y., & Pierleoni, A. (2020). DARE: Data Augmented Relation Extraction with GPT-2. arXiv preprint arXiv:2004.13845. Retrieved from [Hugging Face] https://huggingface.co/healx/gpt-2-pubmed-medium