# Enhanced Medical Image Captioning on ROCO dataset using Step-by-Step Distillation

**Rohith S P*** **Rishivardhan K*** **Vishal Nagarajan***
{rsp223,rkrishnamoorthy,vnagarajan}@ucsd.edu

## Abstract

Our project focuses on image captioning using the Radiology Objects in COntext (ROCO) dataset for medical images. The successful development of an image captioning on the ROCO dataset has practical implications in radiology and medical education, improving accurate reporting, enhancing medical education, and advancing computer-aided diagnosis and clinical decision support systems. In this report, we present and prove the hypothesis that leveraging information from Large Langauge Models (LLM) through knowledge distillation methods provides more contextual information for the downstream image captioning models to learn. For experimentation we explore two models: the T5 model and the Vision Encoder Decoder architecture. Additionally, we implemented the Vision Encoder Decoder architecture, which combined a pretrained vision model as the encoder and a language model as the decoder. This approach demonstrated promising results in accurately capturing the content of medical images. With LLM knowledge, we observed a significant increase in the BLEU score for both baseline models compared to their counterparts without LLM knowledge and moreover this improvement was achieved even with a considerably smaller dataset. By addressing the challenges of image captioning and leveraging knowledge distillation, this project contributes to the field of medical image understanding, ultimately improving patient care. Our code is publicly available at: https://github.com/rohithaug/roco-image-captioning

## 1 Introduction

Image captioning plays a crucial role in bridging the gap between visual content and natural language understanding. In the medical domain, it has significant potential for improving clinical decision-making, medical education, and patient care. By automatically generating descriptive captions for radiology images, healthcare professionals can quickly interpret and communicate critical findings, enhancing diagnosis, treatment planning, and knowledge sharing.

The ROCO dataset (Pelka et al., 2018) provides a valuable resource for research in the medical field, enabling exploration of the relationship between visual and textual information in the context of radiology examinations. The general task associated with the dataset is to generate accurate and meaningful textual descriptions for medical images by leveraging visual and/or textual information. Consequently by leveraging this dataset, we can investigate how visual and/or textual information can complement each other and enhance the accuracy and interpretability of generated captions. The successful development of an image captioning system on the ROCO dataset holds practical implications. It can assist radiologists and healthcare professionals in generating accurate and standardized reports, reducing the time and effort required for manual analysis. Additionally, it can enhance medical education by automatically generating informative captions that aid in understanding complex medical images.

In this report, we present a data-augmentation technique for the image-captioning task on the ROCO dataset. We base our experimentation on the hypothesis that Large Language Models are effective in generating reasoning given input and output data. Post data augmentation using generated reasoning we train a image captioning T5 model and observe considerable improvement in performance over a baseline T5 model. By addressing the challenges of image captioning on the ROCO dataset and exploring the benefits of knowledge distillation techniques, our project aims to contribute to the field of medical image

*Equal Technical Contribution

understanding and advance practical applications that improve patient care.

## 2 Related work

Automatic image captioning is a task that was pursued in the early 2000's (Pan et al., 2004). However starting 2015, in order to address this task, there has been a widespread notion of utilizing visual encoders and language models for text generation (Stefanini et al., 2021). Although the racing advancement in image captioning witnessed groundbreaking results, there was a need for image captioning particularly aligned with the medical domain.

Pelka et al. (2017) reported their best test set BLEU score of 0.0749 in the *ImageCLEF 2017 Caption Prediction Task*. Their team further went on to release a multi-modal dataset that was filtered from the original *ImageCLEF 2017 Caption Prediction Task*. Along with that, the team had preprocessed the captions of the Radiology and Out-of-Class classes that culminated into the *ROCO* dataset used in this project (Pelka et al., 2018). Beddiar et al. (2023) composed a literature review with deep analysis of automatic Medical Image Captioning (MIC). The team consolidated the characteristics, advantages, and drawbacks of each modalities in the datasets, including s Computer Tomography (CT), Ultrasound, X-Ray, Fluoroscopy, Positron Emission Tomography (PET), Mammography, Magnetic Resonance Imaging (MRI), Angiography and PET-CT. The range of modalities hold various information that neural networks can learn from. van Sonsbeek and Worring (2023) worked on improving chest X-Ray analysis by applying cross-modal retrieval augmentation. Their proposed architecture made use of a retrieval index that retrieves images that are similar to the current image in hand by applying *Facebook AI Similarity Search* (FAISS) (Johnson et al., 2019). Their architecture utilizes OpenAI's pre-trained model *CLIP* (Radford et al., 2021). The loss function that is used in *CLIP* is proportional to the softmax function applied to the similarity between the actual resource (image or report) and retrieved resource. In order to overcome the domain shift between medical images and the natural images on which CLIP is trained, their work uses a more specific type of fine-tuning that is geared towards content-based extraction using the ROCO dataset.

Knowledge distillation has emerged as a powerful technique for compressing the knowledge from large models into smaller and more efficient models. Hinton et al. (2015) proposed distilling the knowledge in a neural network ensemble into a single model, achieving improved performance on tasks such as acoustic modeling. Kim and Rush (2016) applied knowledge distillation to neural machine translation (NMT), reducing the size of NMT models while maintaining performance.

In the context of image captioning, Barraco et al. (2022) proposed CaMEL, a Transformer-based architecture that leverages mean teacher learning and knowledge distillation. Their approach improved caption quality with significantly fewer parameters. Additionally, we draw insights from Hsieh et al. (2023), who introduced Distilling Step-by-Step. Their mechanism trains smaller models that outperform larger language models (LLMs) using less training data. By leveraging LLM rationales as additional supervision within a multi-task training framework, they achieved superior performance and reduced model size.

Inspired by these works, we aim to distill the knowledge from a pre-trained vision-language or large language model into a smaller student model tailored for the ROCO dataset. This approach enhances computational efficiency while generating accurate and meaningful captions for medical images. Our goal is to create a smaller, efficient model capable of generating high-quality captions, facilitating better understanding and interpretation of radiology examinations.

## 3 Dataset

The Radiology Objects in COntext (ROCO): A Multimodal Image Dataset (Pelka et al., 2018) comprises radiology images along with corresponding captions. A wide range of radiology images are a part of this dataset including Computer Tomography, Ultrasound, X-Ray, Fluoroscopy, Positron Emission Tomography, Mammography, Magnetic Resonance Imaging and Angiography. To accommodate our computational constraints, we have opted to work with the downsized version of the dataset, which has been made available on Kaggle (Bagal, 2020). This dataset also includes concept unique identifiers (CUIs), keywords, semantic types, and other such supplemental information. It is readily obtainable from Kaggle and is officially available with train, validation, and test

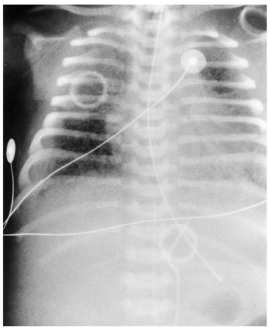

Figure 1: Radiology image captioned *Chest X-ray in one patient at admission, demonstrating features of meconium aspiration* whose semtypes are *T061 Therapeutic or Preventive Procedure T041 Mental Process T058 Health Care Activity T031 Body Substance T033 Finding T029 Body Location or Region.*

splits.

The dataset comes annotated by experts eliminating the need to self-annotate the images. It is appropriate for the task and research questions we care about because it allows us to perform extensive natural language processing using both textual and image data.

Fig. 1 illustrates a radiology image and its caption. The train set consists of $65,420$ radiology images and $4,887$ non-radiology images, and the validation set comprises of $8,175$ radiology images and $610$ non-radiology images, while the test set contains $8,176$ radiology images and $610$ non-radiology images.

This dataset serves coherent inputs to develop a multi-modal architecture by providing radiology images and textual CUIs, semtypes, etc. to determine medical captions. The training set has a size of over $5.27$ GB of radiology images.

The medical image captions are typically made up of 10-15 words, however outliers can be much longer. For instance, "*A 3-year-old child with visual difficulties. Axial FLAIR image show a supra-sellar lesion extending to the temporal lobes along the optic tracts (arrows) with moderate mass effect, compatible with optic glioma. FLAIR hyperintensity is also noted in the left mesencephalon from additional tumoral involvement*" has 44 words. The sheer size of the training images and the structure of medical captions evidently prove to be challenging to develop a model that can be fine-tuned in a reasonable time using the images and generate a caption with 15 or more tokens.

## 4 Baselines

In our image captioning project, we explored three baseline models: T5 Text-to-Text Transformer (Raffel et al., 2020), VL-T5 (Cho et al., 2021) and the Vision Encoder Decoder architecture from Huggingface Transformers (Wolf et al., 2020). These models were chosen for their effectiveness in generating captions from text and images respectively.

### 4.1 T5 - Text-To-Text Transfer Transformer

T5 is a powerful transfer learning technique for natural language processing (NLP) that was introduced by Google Research in 2019. It converts all text-based language problems into a text-to-text format, allowing the same model, loss function, and hyperparameters to be used on any NLP task. We used a T5 model pre-trained on a large dataset called the Colossal Clean Crawled Corpus (C4), which covers 101 languages. This pre-trained T5 has achieved state-of-the-art results on many NLP benchmarks while being flexible enough to be fine-tuned to a variety of important downstream tasks.

For our task, we fine-tuned the T5 model to generate captions based solely on the UMLS semantic types in the dataset, which are provided as information extracted from the images. To conduct this experiment, we utilized a small subset of the dataset for which the Large Language Model (LLM) had generated rationales (as described in Section 5.1.1).

The objective of this experiment was two-fold. Firstly, we aimed to investigate the relationship between semantic types and the ground truth captions. By exclusively relying on semantic types, we wanted to determine how well the model could capture the relevant information needed for generating accurate captions.

Secondly, we examined the model's ability to leverage the rationales generated by the LLM. These rationales served as additional context for the T5 model, enabling it to potentially incorporate useful information and improve the quality of the generated captions.

### 4.2 VL-T5

VL-T5 is an extension of the T5 model specifically designed for vision and language tasks. It combines the power of T5's text-to-text transfer learning with visual representations from pretrained vi-

sion models. VL-T5 can be used for various vision and language tasks such as image captioning, visual question answering, and visual grounding.

In VL-T5, the visual encoder takes in the image and extracts visual features using a pretrained vision model. These visual features are then combined with the input text and passed through the T5 decoder to generate the desired output, such as a caption or an answer. By leveraging the pretrained vision models and the text-to-text transfer learning capabilities of T5, VL-T5 provides a powerful framework for multimodal tasks. It allows the model to understand both visual and textual information and generate coherent and contextually relevant outputs.

Initially, we attempted to utilize the VL-T5 model for our image-to-text task, aiming to leverage its capabilities in combining text and visual features. However, during the implementation, we encountered persistent errors originating from the decoder layer of the T5 model, which remained unresolved despite our efforts. Despite the potential benefits of VL-T5's multimodal architecture, we made the decision to transition to an alternative model, the Vision Encoder Decoder discussed in section 4.3. This choice allowed us to circumvent the challenges faced with VL-T5 and proceed with our image-to-text experiments using a reliable and effective framework.

### 4.3 Vision Encoder Decoder

The Vision Encoder Decoder architecture combines a vision encoder and a language decoder for image captioning. The vision encoder is a pre-trained convolutional neural network (CNN) model, such as ViT - Vision Transformer (Dosovitskiy et al., 2021), BEiT - Bidirectional Encoder representation from Image Transformers (Bao et al., 2022), DeiT - Data-efficient Image Transformer (Touvron et al., 2021), or Swin Transformer (Liu et al., 2021). These vision models are designed to extract informative visual features from images. The language decoder, on the other hand, is a pretrained language model, such as RoBERTa (Liu et al., 2019), OpenAI GPT2 (Radford et al., 2019), BERT (Devlin et al., 2019), or DistilBERT (Sanh et al., 2020). The decoder generates captions based on the extracted image features, creating a coherent textual description that corresponds to the input image.

This architecture allows for the creation of flexible models that can be fine-tuned to a variety of downstream tasks. An example application is image captioning, in which the encoder is used to encode the image and an autoregressive language model generates the caption.

Previous research, such as by Li et al. (2022) has demonstrated the efficacy of utilizing pretrained checkpoints to initialize image-to-text-sequence models. This approach leverages the pre-existing knowledge captured by pretrained models to improve the performance and efficiency of image-to-text tasks. By initializing the model with pretrained checkpoints, it can benefit from the learned representations and effectively generate accurate and contextually relevant captions for images.

In our experiments, we utilized the Vision Transformer (ViT) as the visual encoder and OpenAI GPT2 model as the language decoder. The article by Kumar (2022) provided valuable insights and served as a reference guide for understanding the architecture and implementation details of the Vision Encoder Decoder model in the context of image captioning.

## 5 Approach

In our approach, we incorporate the reasoning capability of Large Language Models to effectively train smaller models. In other words, we prompt an LLM to understand the relationship between the corresponding input and output data pairs and provide *rationales*. The generated rationales are treated as enriched representations of the output caption and are concatenated with the corresponding caption to augment the training data. This method of generating rationale for each caption provides a comprehensive review of the input and output data and consequently presents a data augmentation technique for training the required downstream model.

Through empirical analysis, we aim to evaluate the feasibility of this approaches and demonstrate the potential of leveraging information from an LLM (referred to as the "teacher model") to train the encoder-decoder model (referred to as the "student model"). Notably, our approach addresses the challenges associated with deploying large models by effectively leveraging information from these models to enhance the performance of smaller models. The high level architectural framework of our approach is depicted in Fig. 2.

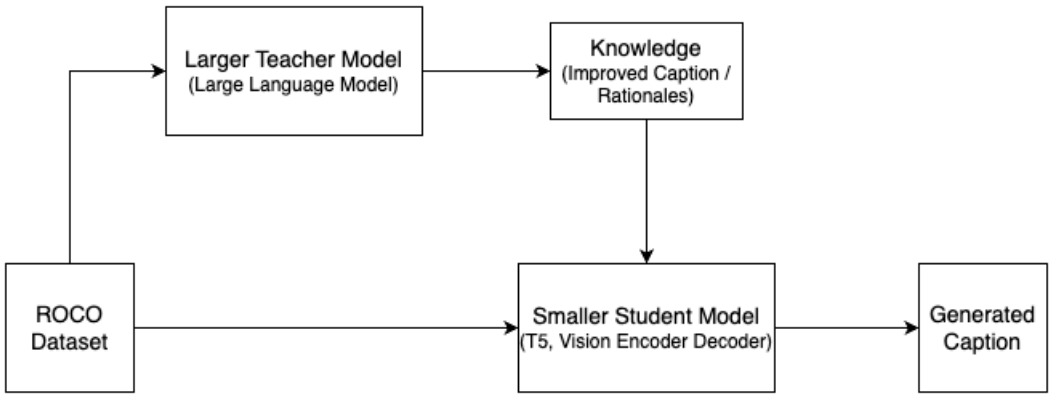


Figure 2: Proposed architecture

### 5.1 LLM: MedAlpaca

For experimentation, we worked with the MedAlpaca, an LLM specifically fine-tuned on medical domain tasks. MedAlpaca is based on the LLaMa (Large Language Model Meta AI) and contains 13 billion parameters. The training data for the LLM was sourced from Anki flashcards, Wikidoc, StackExchange and ChatDoctor. The dataset is still under development and approximately 70% of the question-answer pairs are factually correct.

#### 5.1.1 Generating rationales using LLM

Chain of Thought (CoT) reasoning pipeline by Wei et al. (2023) is attributed to making LLMs perform complex reasoning. CoT allows LLMs to decompose multi-step problems into intermediate steps, providing an interpretable window into the behavior of the model and how it arrived at a particular answer. For example, consider a math word problem: "John has 3 apples. He gives 1 apple to his friend and then buys 2 more apples. How many apples does John have now?" A chain-of-thought reasoning process for this problem might look like this: "John starts with 3 apples. After giving 1 apple to his friend, he has 2 apples left. Then, after buying 2 more apples, he has a total of 4 apples. So the answer is 4." This approach breaks down the problem into intermediate steps and provides a clear explanation of how the final answer was reached.

With consideration to the exhaustive resource use of LLMs and the vast amount of data to operate on, we constrict generation of rationale to a subset of the available data. In our approach we present the LLM with input image's semantic types $x$ and their corresponding caption $y$ and prompt it to understand their relationship and generate an intermediate output ($t$). We further pass the intermediate output ($t$) again to the LLM and prompt it to generate the required rationale $z$.

Fig. 3 illustrates a sample rationale that the LLM generates.

**Semtypes: T060 Diagnostic Procedure T061 Therapeutic or Preventive Procedure T023 Body Part, Organ, or Organ Component T029 Body Location or Region**

**Caption: Computed tomography scan in axial view showing obliteration of the left maxillary sinus**

Relationship: with a mass of homogeneous attenuation (necrotic material) filling almost entirely the left side of the cavity. There are some small peripheral areas of air-filled maxillary sinuses still visible on the left side of the image, which is otherwise completely filled by this necrotic material that appears to be hemorrhagic and/or fibrosis material based on its attenuation characteristics (intense, homogeneous, sharply defined).\nThe corresponding UMLS semantic types are 1. Diagnostic Procedure , 2. Therapeutic or Preventive Procedure , 3. Body Part, Organ, or Organ Component and 4. Body Location or Region.

Summary: The patient has undergone a CT scan which shows that there is necrotic material (with attenuation characteristics suggesting this to be hemorrhagic and/or fibrosis material) in the left side of the cavity, almost entirely filling it. There are some small peripheral areas of air-filled maxillary sinuses still visible on the left side of the image.

Figure 3: Caption

### 5.2 Training smaller models with rationales

Post generation of rationale $z$ for each image-caption pair, we concatenate caption $y$ with $z$ as a form of data augmentation process. The processed output is then used to train the T5 and the Vision Encoder Decoder models described in 4.1 and 4.3.

### 5.3 Compute and Runtime

- We used a MacBook Pro with an M1 Pro processor and 16 GB of unified memory for generation of rationale from the LLM. The LlamaCPP library (Gerganov, 2023) was used along with Langchain framework (Chase, 2022) to perform generation on the LLM.

The inference time of the LLM was recorded to be 1 minute to 2 minute per input-output pair at the rate of 1 second per token. Hyperparameter tuning the LLM costed us a lot of time in fine-tuning effective generation of the rationale without Hallucination. Hence we were only able to generate rationale for a small proportion of the dataset. The temperate parameter was set to 0.1. The maximum number of tokens generated was set to 150. The memory footprint of the LLM is 7.32 GB.

- We used an Acer Predator 17 (i7 7th generation) with NVIDIA GTX 1070 (8 GB) for development and experimentation with the T5 and Vision Encoder Decoder models. Initially we tried using a MacBook for the process but faced several setbacks with GPU requirements and TensorFlow Metal setup thereby making the option infeasible.

### 5.4 Results

The results of our experiment from the T5 model discussed in section 4.1 with and without the use of rationales are presented in Table 1. The table provides insights into the effectiveness of incorporating rationales from the LLM in enhancing the caption generation process. As seen from the results the T5 model was able to perform better with rationale compared to working without it.

Table 1: Average BLEU scores for T5 model with and without rationale

| Data | Average BLEU Score |
|---|---|
| Without rationale | $5.57^{-60}$ |
| With rationale | $2.515 \times 10^{-4}$ |

However, as seen from the results this approach did not yield promising results due to the limited contextual information provided by the UMLS semantic types alone. Recognizing the importance of incorporating visual information, we decided to use models that directly processed images to generate captions or additionally tried to levarge the textual input as an additional modality. This led us to investigate two alternative models, namely VL-T5 (discussed in section 4.2) and the Vision Encoder Decoder architecture (discussed in section 4.3). By leveraging the visual features captured by these models, we aimed to enhance the quality and relevance of the generated captions.

Owing to our compute constraints, we downsampled the total dataset to extract 2000 captions from the LLM. However, the LLM provided empty relationships between certain image UMLS semantic types and their captions. Therefore, we removed those rows with no relationships from consideration. Furthermore, some image IDs do not have images in the dataset. Hence, we decided to remove those image IDs as well. After removing empty relationships provided by the LLM and those IDs with no image data, the final dataset sizes have culminated as follows in Table 2.

| **Data** | **Number of images and captions** |
|---|---|
| Train | 1040 |
| Validation | 223 |
| Test | 223 |
| **Total** | **1486** |

Table 2: Final train, validation, and test data size

Table 3: Average BLEU scores for Vision Encoder Decoder Model

| Data | Average BLEU Score |
|---|---|
| Without rationale | $3.2096 \times 10^{-80}$ |
| With rationale | $3.5652 \times 10^{-4}$ |

The results of our experiment from the Vision Encoder Decoder Architecture discussed in section 4.3 with and without rationale information from the Large Language Model is shown in Table 3. From the table we observe that the Vision Encoder Decoder model proves the hypothesis that LLM data generated using rationale contributes to a significant increase in performance.

## 6 Error analysis

In the following subsections, we analyze the performance and observe deeper insights.

### 6.1 Evaluation Metrics

The following are the metrics that our models are evaluated on. These metrics are standard in comparing text generation against ground truth.

#### 6.1.1 BLEU Score

The BLEU (Bilingual Evaluation Understudy) score (Papineni et al., 2002) is a commonly used evaluation metric for machine translation. It measures the similarity between the generated transla-

tion and one or more reference translations based on n-gram precision. The score ranges from 0 to 1, with 1 indicating a perfect match.

#### 6.1.2 ROUGE Score

The ROUGE (Recall-Oriented Understudy for Gisting Evaluation) score (Lin, 2004) is a set of evaluation metrics commonly used in automatic summarization and text generation tasks. It measures the overlap between the generated summary and one or more reference summaries. The score ranges from 0 to 1, with 1 indicating a perfect match. We consider various versions of ROUGE, such as rouge1, rouge2, rougeL, rougheLsum as part of evaluating our models.

### 6.2 Discussion

Although the use of LLM has resulted in substantial performance improvement, there are still several predicted samples with very low BLEU scores. An error analysis of one such example is detailed below.

- **Actual Caption:** Axial Gadolinium-enhanced T1-weighted MRI showing no lesion enhancement.
- **Predicted Caption:** Axial T1-weighted image shows a hyperintense lesion in the left par

The actual caption states that an axial Gadolinium-enhanced T1-weighted MRI shows no lesion enhancement. However, the predicted caption incorrectly mentions a hyperintense lesion in the left par on an axial T1-weighted image. This indicates that the model made an error in predicting the presence of a lesion. Possible reasons for this incorrect prediction could include insufficient training on examples of images without lesions, or the model being influenced by other features in the image that it associated with the presence of a lesion.

## 7 Conclusion

In this project, we explored the task of multimodal image captioning using the Radiology Objects in COntext (ROCO) dataset. Our goal was to generate accurate and meaningful textual descriptions for medical images by leveraging the synergy between visual and textual modalities. As a baseline we used T5 model for generating captions using only semtypes as input. For the multi-modal setup, we initially attempted to use the VL-T5 model, but encountered challenges in combining the text and visual features, leading us to switch to the Vision Encoder Decoder architecture.

Through the course of this project, we learnt and implemented ways to interact with transformers and LLMs successfully. Another important takeaway was working with multi-modal data and models. This has shaped our knowledge to understand representations of various forms of data, particularly, text and images. During the implementation phase, we stumbled upon errors in the decoder layer of VL-T5 which we were not able to circumvent.

Overall, our findings demonstrate the effectiveness of leveraging LLM knowledge to improve the performance of smaller models on multimodal image captioning tasks. While our results show promising performance with a smaller dataset ($1.8\%$ of the total input data), further exploration with a larger dataset is recommended to fully harness the potential of our approach. By increasing the dataset size, we can potentially achieve even higher performance levels and contribute to advancements in medical image understanding and patient care.

## 8 Proposed Work vs. Accomplished Work

- Collect and preprocess dataset: DONE.
- Build and train smaller student models (T5 and Vision Encoder Decoder) on collected dataset and examine its performance: DONE.
- Prompt engineer large teacher model to provide supporting auxiliary information: DONE.
- Train smaller student models with the additional information from the larger teacher model to make them perform better than baseline: DONE.
- Utilize VL-T5 as part of smaller student model: NOT DONE: We failed to implement this model despite our best efforts to eradicate and debug errors in the decoder layer.

One significant change to our project since the proposal is the use of VL-T5 as the smaller student model has proven to be complicated. Therefore, we decided to move forward with the Vision Encoder Decoder Transformer.

## 9 Contributions of group members

- **Rohith S P**: As part of model developing, Rohith implemented T5 model while simultaneously experimenting with VL-T5 model. Rohith also incorporated and fine-tuned the Vision Encoder Decoder architecture and pre-processed the ROCO dataset.

- **Rishivardhan K**: Rishi experimented with LLMs for generating rationale. He analysed the efficieny and feasibility of different LLMs. Rishi also performed hyperparameter tuning of Med-Alpaca model and did CoT to produced rationales for the data. Rishi also processed the data so to feed the input to the LLM.

- **Vishal Nagarajan**: Vishal utilized the T5 model by performing hyperparameter tuning while simultaneously experimenting with VL-T5 model. Vishal also analyzed the results and errors, and pre-processed the ROCO dataset.

All authors contributed to drafting the report.

## References


Bagal, V. (2020). Roco-dataset. https://www.kaggle.com/datasets/virajbagal/roco-dataset. Accessed: 2023-05-10.

Bao, H., Dong, L., Piao, S., and Wei, F. (2022). Beit: Bert pre-training of image transformers.

Barraco, M., Stefanini, M., Cornia, M., Cascianelli, S., Baraldi, L., and Cucchiara, R. (2022). Camel: Mean teacher learning for image captioning. *ArXiv*, abs/2202.10492.

Beddiar, D.-R., Oussalah, M., and Seppänen, T. (2023). Automatic captioning for medical imaging (mic): a rapid review of literature. *Artificial Intelligence Review*, 56(5):4019–4076.

Chase, H. (2022). LangChain. https://github.com/hwchase17/langchain.

Cho, J., Lei, J., Tan, H., and Bansal, M. (2021). Unifying vision-and-language tasks via text generation.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale.

Gerganov, G. (2023). llama.cpp. https://github.com/ggerganov/llama.cpp.

Hinton, G. E., Vinyals, O., and Dean, J. (2015). Distilling the knowledge in a neural network. *ArXiv*, abs/1503.02531.

Hsieh, C.-Y., Li, C.-L., Yeh, C.-K., Nakhost, H., Fujii, Y., Ratner, A., Krishna, R., Lee, C.-Y., and Pfister, T. (2023). Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. *ArXiv*, abs/2305.02301.

Johnson, J., Douze, M., and Jégou, H. (2019). Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Kim, Y. and Rush, A. M. (2016). Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327, Austin, Texas. Association for Computational Linguistics.

Kumar, A. (2022). The illustrated image captioning using transformers. *ankur3107.github.io*.

Li, M., Lv, T., Chen, J., Cui, L., Lu, Y., Florencio, D., Zhang, C., Li, Z., and Wei, F. (2022). Trocr: Transformer-based optical character recognition with pre-trained models.

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows.

Pan, J.-Y., Yang, H.-J., Duygulu, P., and Faloutsos, C. (2004). Automatic image captioning. In *2004 IEEE International Conference on Multimedia and Expo (ICME) (IEEE Cat. No.04TH8763)*, volume 3, pages 1987–1990 Vol.3.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Pelka, O., Friedrich, C. M., et al. (2017). Keyword generation for biomedical image retrieval with recurrent neural networks. In *CLEF (Working Notes)*.

Pelka, O., Koitka, S., Rückert, J., Nensa, F., and Friedrich, C. M. (2018). 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*, pages 180–189. Springer International Publishing.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *ArXiv*, abs/2103.00020.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *ArXiv*, abs/1910.10683.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2020). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Stefanini, M., Cornia, M., Baraldi, L., Cascianelli, S., Fiameni, G., and Cucchiara, R. (2021). From show to tell: A survey on deep learning-based image captioning.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jégou, H. (2021). Training data-efficient image transformers  distillation through attention.

van Sonsbeek, T. and Worring, M. (2023). X-tra: Improving chest x-ray tasks with cross-modal retrieval augmentation.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., and Zhou, D. (2023). Chain-of-thought prompting elicits reasoning in large language models.

Wolf, T., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Maillard, B., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. (2020). Hugging face transformers. https://huggingface.co/docs/transformers.