CSC490: Comparing General Purpose Image Captioning Model against Fine-tuned Model for Medical Domain

Amane Takeuchi
Department of Computer Science
University of Toronto
Toronto, ON
amane.takeuchi@mail.utoronto.ca

Khanh Vo
Department of Computer Science
University of Toronto
Toronto, ON
khanh.vo.utoronto.ca

Danyal Ilyas
Department of Computer Science
University of Toronto
Toronto, ON
danyal.ilyas@mail.utoronto.ca

Abstract

As computer vision and Natural Language Processing (NLP) has progressed in recent years, the demand for domain-specific Machine Learning models is increasing, especially in the healthcare industry. Being able to describe medical images well is not just important for doctors but also for training new ones in treating patients. Hence, our goal is to build a machine learning model that can potentially assist professionals in the medical field by extracting helpful information from healthcare images. Therefore, we've attempted to develop a fine-tuned domain-specific image captioning model for the healthcare dataset MedICaT using the BLIP model. Rather than embarking on an extensive project, our focus is to explore the potential of fine-tuning the existing image captioning model, testing and comparing our versions of the model developed in-house with the original BLIP model. In short, this study aims to provide insights into the adaptability and performance of the BLIP model in a healthcare-specific image captioning scenario.

1 Introduction

This report will outline a comprehensive exploration into the refinement and optimization of the BLIP language model for a specialized medical image captioning task. The primary goal is to address two pivotal research questions within the realm of image captioning. Particularly, our study aims to ascertain whether a fine-tuned image captioning model can outperform a general-purpose image captioning model. Moreover, it delves into the exploration of the fine-tuned model's potential to generate output suitable for application in the medical industry. To achieve these goals, we leverage the BLIP model, renowned for its state-of-the-art performance in various vision-language tasks. According to Li et al. (2022), BLIP, also known as Bootstrapping Language-Image Pre-training, is characterized by a multimodal mixture of encoder and decoder. In specific, its architecture consists of three key elements (Figure 1). Firstly, there is a text encoder that undergoes training using image-text contrastive (ITC) learning. Secondly, an image-grounded text encoder is employed, incorporating cross-attention layers to capture the interactions between vision and language. This encoder is trained with an image-text matching (ITM) loss. Lastly, the image-grounded text decoder replaces the bi-directional self-attention layers with causal self-attention layers. Additionally, it shares the same cross-attention layers and feed forward networks as the encoder (Li et al., 2022).

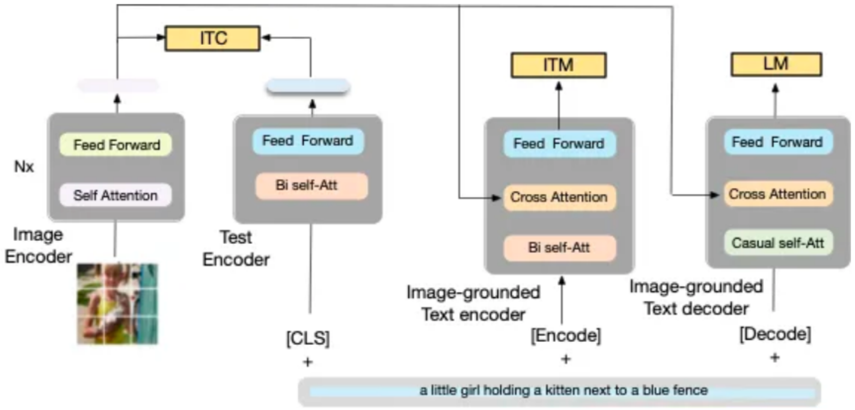


Figure 1: General overview of BLIP architecture (Li et al. 2022)

Using BLIP as the foundation, two distinct models were developed for our research – Fine-Tuned 1 (FT-1) and Fine-Tuned 2 (FT-2), which will be subsequently benchmarked against the performance of the original BLIP model prior to fine-tuning. The training of these models will be conducted using the MedICaT dataset, acting as the fundamental groundwork for assessing how well the BLIP model adapts to the intricate challenges posed by medical image captioning tasks.

The structure of this report is organized to provide a detailed overview of our research methodology. The initial sections encompass data preparation, including a comprehensive dataset overview, data cleaning procedures, and subset selection. The procedural sections outline major implementation steps, training procedures, and emphasizes the novelty of our approach. Evaluation methods, covering summary results, BLEU scores, ROUGE scores, and BERT-F1 scores, will be discussed in detail. Additionally, the report addresses the study's limitations and identifies areas for further improvement. Through this systematic approach, we aim to illuminate the effectiveness and adaptability of the BLIP model in the specialized domain of medical image captioning.

2 Dataset

2.1 Dataset Overview

Attaining a voluminous scale with a total size of 112.85 GB, the MedICaT dataset is a rich compilation of various medical images, captions, subfigure-subcaption annotations, and inline textual references. In specific, the dataset contains a comprehensive collection of 217,060 figures extracted from 131,410 open-access papers and 7,507 subcaption/subfigure annotations specifically associated with 2,069 compound figures (Subramanian et al., 2020). Additionally, it includes inline references for approximately 25,000 figures within the ROCO dataset, adding a layer of contextual information to the images (Subramanian et al., 2020). It draws its source from open access articles in PubMed Central, ensuring the authenticity of the data, and its high quality as well as corresponding reference text is derived from S2ORC.

MedICaT exhibits a broad range of medical figures, covering essential categories vital for comprehensive medical research. Noteworthy categories include diagnostic imaging, featuring radiographs, MRIs (Magnetic Resonance Imaging), CT Scans (Computed Tomography), and Ultrasound Images. Access to such a large and well-annotated dataset is crucial for training an accurate and robust image captioning model.

To access the resource, users are required to undergo a form submission process, strictly adhering to non-commercial usage guidelines.

2.2 Data Cleaning

The initial dataset presented a challenge to use in terms of scale and complexity. In order to refine the dataset and tailor it to a more focused area of study, the decision was made to concentrate on enhancing the clarity of chest images, a pivotal domain in medical diagnostics. To achieve this focus, a filtering process was utilized, wherein figures were selectively retained based on keyword filters applied to captions containing the term ’chest.’ As a consequence, the size of the MedICaT dataset was trimmed down to less than 10 gigabytes, with a targeted selection of approximately 16,000 figures.

Another crucial aspect of the pre-processing phase involved refining the textual components of the dataset. Metadata, initially encoded in JSONL format, was converted to CSV, so that it would be easier for us to manipulate the information. Next, text descriptions accompanying each figure underwent examination and cleanup. In specific, unnecessary terms such as "Fig.," "figure," and "Figure A" were removed, streamlining the textual content and contributing to a more standardized and coherent dataset. Furthermore, to enhance readability and promote consistency, random punctuation marks were systematically removed from the text descriptions. This step not only contributed to a more refined and uniform dataset but also facilitated subsequent natural language processing tasks by eliminating potential sources of noise and variability.

3 Method

3.1 Major Implementation Decisions

Various hyperparameters and architecture choices were modified during the training of the model in order to reach optimal validation metrics. This included tuning various hyperparameters such as Batch Size and learning Rate, as well as experimenting with different loss functions.

3.1.1 Batch Size

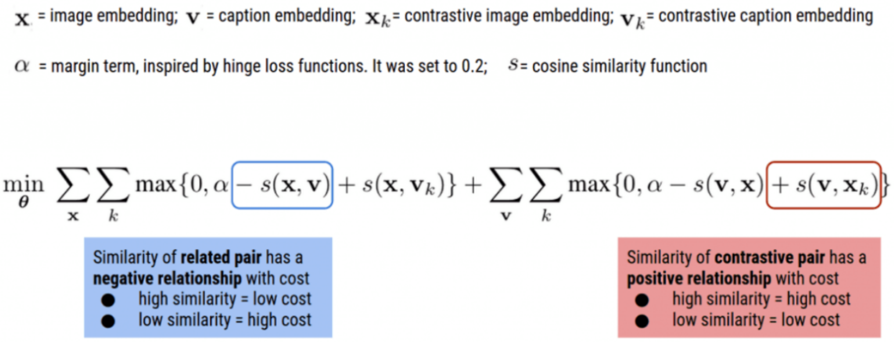
GPU constraints limited batch size to 4 in initial training, but this led to very noisy gradient descent convergence as the model could not progress along the loss landscape and reach any sort of convergence. As a result, a technique called gradient accumulation was implemented based on the paper “End-to-end Multiple Instance Learning with Gradient Accumulation”(Kigma, 2017). This novel approach lets the gradient accumulate while running 4 mini-batches of size 4, then taking an average of the gradients and finally performing backpropagation. This allowed the model to simulate a batch size of 16 while only ever putting 4 images in memory during a given time to prevent memory leakage. Gradient accumulation also has other benefits, such as improved stability in optimization, as discussed in “End-to-end Multiple Instance Learning with Gradient Accumulation”(Kigma, 2017).

3.1.2 Learning Rate

Initial choices of fixed learning rate saw poor gradient descent convergence and slow training times. As a result, well-researched learning rate schedulers such as **Adaptive Moment Estimation (ADAM)** were chosen to adaptively adjust the learning as the training loop progresses in order to enhance gradient descent convergence. ADAM is based on the paper “End-to-end Multiple Instance Learning with Gradient Accumulation”(Andersson et al., 2022). It works by adjusting the learning rate for each parameter based on a moving average of the previous gradients. We found this produced much better loss curve behavior, as shown in Figure 3.

3.1.3 Choice of Loss Function

The optimal choice of loss function for this task was determined to be a contrastive loss for image captioning, which the equation summarizes in Figure 2. This choice of loss function was implemented after the initial use of cross-entropy loss because cross-entropy maximizes the likelihood of the next word given the previous ones, which resulted in our model not learning contentful enough embedding. Contrastive loss emphasizes distinguishing between negative and positive pairs of captions and their associated images, allowing us to learn much richer word embeddings.



$\mathbf{x}$ = image embedding; $\mathbf{v}$ = caption embedding; $\mathbf{x}_k$= contrastive image embedding; $\mathbf{v}_k$= contrastive caption embedding

α = margin term, inspired by hinge loss functions. It was set to 0.2; s= cosine similarity function

$$\min_{\boldsymbol{\theta}} \sum_{\mathbf{x}}\sum_{k} \max\{0, \alpha \boxed{- s(\mathbf{x},\mathbf{v})} + s(\mathbf{x},\mathbf{v}_k)\} + \sum_{\mathbf{v}}\sum_{k} \max\{0, \alpha - s(\mathbf{v},\mathbf{x}) \boxed{+ s(\mathbf{v},\mathbf{x}_k)}\}$$

Similarity of **related pair** has a **negative relationship** with cost

- high similarity = low cost
- low similarity = high cost

Similarity of **contrastive pair** has a **positive relationship** with cost

- high similarity = high cost
- low similarity = low cost

Figure 2: A diagram illustrating contrastive loss for image captioning (Kiros, 2015)

3.2 Training Model

The initial implementation of gradient descent saw the usage of 1,800 images. Initial implementations led to suboptimal loss curves. But upon fine-tuning and implementation of the model architecture choices mentioned above, the following iterations vs training and validation loss(seen in figure 3) show the model is able to reach convergence and prevent overfitting.

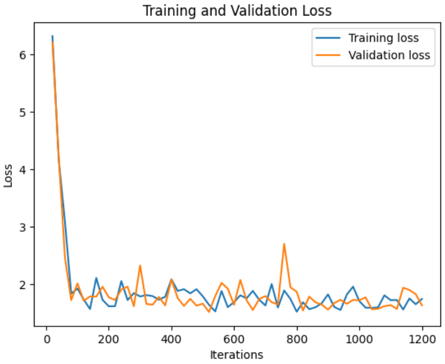


Figure 3: A diagram illustrating training/validation loss curve on the first 1.8k images

From Figure 3, we can see that there is not a wide discrepancy between validation loss and training loss. This indicates that the mentioned design choice resulted in a model that can generalize well to unseen data. To further test this claim, an additional 1000 images were used to train the model further and measure the effect on validation loss.

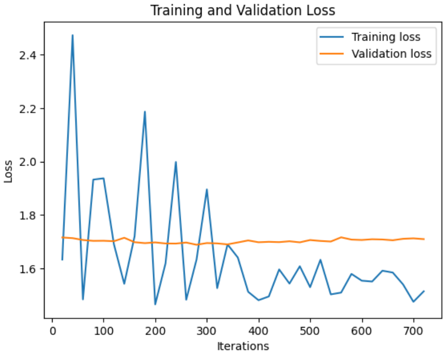


Figure 4: A diagram illustrating the training/validation loss curve on an additional 1k images

As can be seen in Figure 4, increasing the training set size has minimal impact on validation loss, indicating the model has reached optimal convergence and can generalize fairly well to unseen data.

3.3 Novelty of Our Approach

Our implementation of techniques such as gradient accumulation, use of ADAM optimizer, the choice of Contrastive Loss for Image Captioning, and our overall approach of fine-tuning a general-purpose vision transformer model summarize the novelty of our approach. We chose to implement gradient accumulation instead of regular stochastic gradient descent to train our model efficiently despite the limited memory constraint of our available hardware. This approach allowed us to train our model with much higher efficiency as well as without compromising on dataset size. The use of the ADAM optimizer was also a more novel approach as opposed to a traditional fixed learning rate and helped our model learn more efficiently. Additionally, implementing contrastive loss for image captioning, which focuses on learning features between classes, is related explicitly to identifying various medical conditions in healthcare images. By using this more novel loss function as opposed to its counterpart cross-entropy loss, we enhanced the accuracy of generated captions. We ensured that our model effectively captured and communicated critical details in medical imaging and captioning. Our overall unique approach involved tuning a general-purpose vision transformer into a medically accurate neural network capable of diagnosing X-rays. A relatively novel method for addressing this unique problem.

3.4 Evaluation Method

Before training the model, we randomly selected 300 chest images from the MedICaT dataset, separate from the training and validation datasets. We conducted an evaluation on three different models: BLIP (without fine-tuning), **Fine-Tuned 1 (FT-1)** and **Fine-Tuned 2 (FT-2)** models. The FT-1 model is trained on 1.8k images, whereas the FT-2 model is trained on 2.8k images. For each model, we generated captions for the testing dataset and compared the reference text against the generated text. We used five major metrics to evaluate two captions: BLEU, ROUGE-1, ROUGE-2, ROUGE-l, and BERT-F1 scores. The **BLEU (Bilingual Evaluation Understudy)** score is a commonly used NLP evaluation metric to assess the similarity between machine-generated and reference sentences by comparing n-grams adjacent sequences of n-words. This score is calculated based on the precision of n-grams and multiplied by a penalty term, which penalizes the precision for generating a shorter sentence than a reference sentence. (Santhosh, 2023)

$$\textbf{BLEU} = \exp(\sum \text{precision of n-grams}) \times \text{brevity penalty}$$

On the other hand, the **ROUGE (Recall-Oriented Understudy for Gisting Evaluation**) score is calculated based on the similarity of generated and reference text by comparing the overlapping n-grams. (Santhosh) In general, for the ROUGE-n score, we have the following formula.

$$\textbf{ROUGE-n} = \sum(\text{recall of n-grams})$$

Thus, ROUGE-1 measures the overlap of single words, whereas ROUGE-2 measures the overlap of two consecutive word sequences. Meanwhile, the ROUGE-I compares the longest common subsequence of generated and reference captions. Finally, unlike the previous scores, the **BERT (Bidirectional Encoder Representations from Transformers)** score could measure the similarity of generated and target text by comparing their semantic meaning. The calculation of this score involves using the existing **Large Language Model (LLM)**, BERT (Figure 5). We used the BERT-F1 score in the evaluation to examine the fined-tuned BLIP model, which ranges from -1 to 1. (Mishra, 2022)

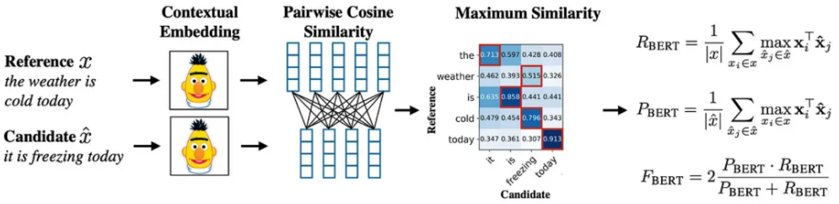

Contextual Embedding
Pairwise Cosine Similarity
Maximum Similarity
Reference x
the weather is cold today
Candidate $\hat{x}$
it is freezing today
Reference
the
weather
is
cold
today
0.713 0.597 0.428 0.408
0.462 0.393 0.515 0.326
0.635 0.858 0.441 0.441
0.479 0.454 0.796 0.343
0.347 0.361 0.307 0.913
it
is
freezing
today
Candidate
$R_{\text{BERT}} = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} \mathbf{x}_i^\top \hat{\mathbf{x}}_j$
$P_{\text{BERT}} = \frac{1}{|\hat{x}|} \sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} \mathbf{x}_i^\top \hat{\mathbf{x}}_j$
$F_{\text{BERT}} = 2\frac{P_{\text{BERT}} \cdot R_{\text{BERT}}}{P_{\text{BERT}} + R_{\text{BERT}}}$

Figure 5: A diagram illustrating the derivation of the BERT-F1 score (Mishra, 2022)

4 Results and Discussion

4.1 Summary Results

Overall, the Fine-Tuned 2 model outperformed the BLIP and Fine-Tuned 1 models for all the evaluation metrics. (Table 1) However, also note that the standard deviation of the FT-2 is relatively higher than the FT-1, indicating that the more trained model generally performs better than the less trained model, but the performance tends to be inconsistent.

Table 1: Average and standard deviation of five metrics between different models

	BLEU		**BERT-F1**		**ROUGE-1**		**ROUGE-2**		**ROUGE-l**	
Model Name	Avg.	S.d	Avg.	S.d	Avg.	S.d	Avg.	S.d	Avg.	S.d
BLIP	0.000173	0.00300	0.787	0.0412	0.120	0.0886	0.0169	0.0345	0.107	0.0803
FT-1	0.0116	0.0534	0.834	0.0370	0.220	0.119	0.0672	0.0912	0.195	0.113
FT-2	0.0186	0.0594	0.840	0.0411	0.241	0.125	0.0796	0.0967	0.211	0.125

4.2 Discussion

4.2.1 BLEU Score

Out of all other metrics, the BLEU score improved the most from a not fine-tuned model to a fine-tuned model. Notice that we could only attain 0.000173 for the unfine-tuned BLIP model while the FT-2 model achieved 0.0186. The possible reason why there was a significant increase in BLEU score is because the unfine-tuned BLIP tends to output a short caption. As the model trained, it started to produce longer text, possibly contributing to improving the BLEU score. However, in the machine-translation task of NLP, a well-trained model attains around 0.6 to 0.7 BLEU score. Although we are evaluating the model on an image captioning task, attaining 0.0186 as the highest BLEU score is sufficient to conclude the model can generate a reasonable caption for a given medical image.

4.2.2 ROUGE Score

For ROUGE scores, the model FT-2 model achieved significantly better results than the unfine-tuned BLIP model. Further, as the ROUGE-1 score increased from 0.120 to 0.24, this indicates that the FT-2 model learned to produce text containing words similar to the reference text. However, although the FT-2 has a much higher ROUGE-2 score than the BLIP model, the FT-2 could only attain a 0.0796 ROUGE-2 score. This suggests that the FT-2 model is suffering from producing text with overlapping two consecutive words as the reference text.

4.2.3 BERT-F1 Score

For the BERT-F1 score, the unfine-tuned BLIP model has already attained a relatively high score of 0.787. After fine-tuning the model, the FT-2 model achieved a 0.840 BERT-F1 score. However, by carefully examining the generated caption from the unfine-tuned model and the reference text, it's hard to say that they share a similar contextual meaning. (Appendix A) Even though we observed an improvement in this score, this score might not accurately compare the semantic meaning of two sentences for the image-captioning task.

5 Conclusion

Through a thorough evaluation of the fine-tuned BLIP model, we observed that the fine-tuned version of the model achieved significantly better results than the unfine-tuned version. The FT-2 model, which was trained with most training datasets, attained the highest scores for all the evaluation metrics we used. Thus, we can conclude that fine-tuned imaged captioning model BLIP can achieve better results than the general purpose BLIP for the healthcare domain. However, concluding that the fine-tuned model can produce reasonable captions for medical images is hard. This is mainly because the model is suffering a lot for BLEU score. Further, even for the ROUGE-1 score, though we observed the improvements from the unfine-tuned to fine-tuned model, the FT-2 only achieved a 0.241 ROUGE-1 score. This is not considered high, and these scores clearly indicate that the generated text is not of a reasonable quality to be reliable. Even though fine-tuning BLIP has excellent potential to be utilized at the industry level, given that the model has significantly improved from fine-tuning, the current version of the model is far from being applied to the medical industry.

6 Limitation

While conducting this project, we noticed three main limitations: MedICaT dataset quality, correctness of evaluation metrics and computational capacity. For the MedICaT dataset, we had two main issues. First, some texts associated with medical images were not descriptive enough to understand what the images represented. This was inevitable as almost all of the image-text pairs were extracted from healthcare articles and books, and these pairs were extracted without having the context of how they were described in the articles or books. As a result, many of these captions for medical images gave limited background information. Second, although in the data cleaning and pre-processing phase, we selected images associated with the word “chest,” the dataset clearly included some images that were irrelevant to the term “chest.” Although we aimed to fine-tune the model on chest X-rays, the training dataset included some irrelevant images. This might have added noise to the training and negatively impacted the model.

Furthermore, the metrics we used to evaluate the model do not perfectly assess its performance. For example, for the BLEU score, even if a generated caption shares the exact same semantic meanings as the reference caption, it might output a low score if two sentences don’t share similar wordings. The ROUGE score also works simlarilary as the BLEU score. As discussed in section 4.2.3, the BERT-F1 score is also not perfect at assessing the semantic meaning of two sentences, as poorly generated captions by the unfine-tuned BLIP model attained a relatively high BERT-F1 score.

Finally, after training the model, we noticed limitations in computational capability. As training the model on a large set of images was very time-consuming, we limited the number of training images. However, in the evaluation phase of the project, we noticed that the model with more training images achieved higher results. Thus, the model performance could have increased if we trained the model with more images. However, this was not possible considering the current capability of the compute.

References

[1] Andersson, A., Koriakina, N., Sladoje, N., & Lindblad, J. (2022). End-to-end Multiple Instance Learning with Gradient Accumulation. arXiv [Cs.CV]. Retrieved from http://arxiv.org/abs/2203.03981

[2] Li, J., Li, D., Xiong, C., & Hoi, S. (2022). BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation (arXiv:2201.12086). arXiv. https://doi.org/10.48550/arXiv.2201.12086

[3] Mishra, P. (2022, May 23). BERTScore: Evaluating text generation with bert. Medium. https://towardsdatascience.com/bertscore-evaluating-text-generation-with-bert-beb7b3431300

[4] Kingma, D. P., & Ba, J. (2017). Adam: A Method for Stochastic Optimization. arXiv [Cs.LG]. Retrieved from http://arxiv.org/abs/1412.6980

[5] Kiros, R. (2015). Neural Image Captioning for Mortals – Indico Data — indicodata.ai. Retrieved from https://indicodata.ai/blog/neural-image-captioning-for-mortals/

[6] Santhosh, S. (2023, April 16). Understanding Bleu and Rouge score for NLP evaluation. Medium. https://medium.com/@sthanikamsanthosh1994/understanding-bleu-and-rouge-score-for-nlp-evaluation-1ab334ecadcb#:~:text=In%20the%20field%20of%20NLP%20evaluation%2C%20BLEU%20and%20ROUGE%20scores,used%20for%20text%20summarization%20tasks.

[7] Subramanian, S., Wang, L. L., Mehta, S., Bogin, B., van Zuylen, M., Parasa, S., Singh, S., Gardner, M., & Hajishirzi, H. (2020). MedICaT: A Dataset of Medical Images, Captions, and Textual References (arXiv:2010.06000). arXiv. https://doi.org/10.48550/arXiv.2010.06000

Contributions

Danyal

- Fine Tuning of the learning rate
- Built training loop
- Conducting model training
- Implementing gradient accumulation
- Built functionality for saving and importing of trained models
- Writing Procedure and Architecture choices part of report
- Determined and Implemented Contrastive Loss for Image Captioning

Khanh

- Data exploration/cleaning
- Writing the Report on Problem and Dataset Decision
- Writing the Abstract, Introduction, Dataset Overview, and Preprocessing Step of the final report
- Made the progress presentation and the final poster

Amane

- Requested dataset access
- Writing the progress report
- Writing Evaluation Method, Results and Discussion, Conclusion, Limitation and Reference sections of the final report
- Initial data exploration/cleaning
- Conducting model training
- Evaluation of the model
- Formatting the final report and generating figures and plots

Overall, we worked collaboratively throughout the project and tried to distribute tasks evenly. For any decision-making process, we discussed and tried to incorporate everyone's opinion as much as possible.

Appendix

A. Example of generated caption by unfine-tuned BLIP

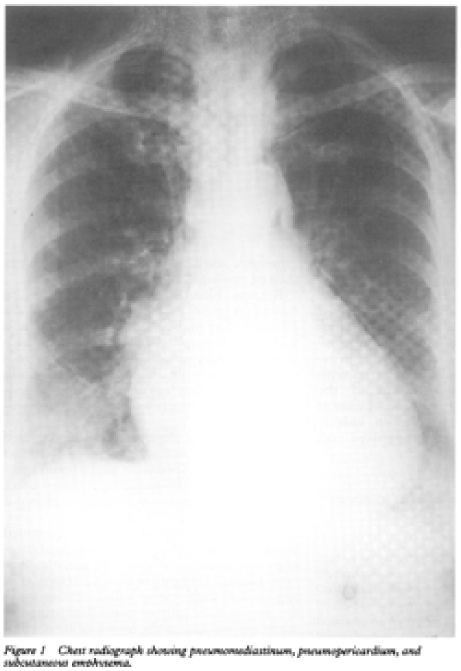


Figure 1 Chest radiograph showing pneumomediastinum, pneumopericardium, and subcutaneous emphysema.

Figure 6:
Reference caption: Chest radiograph showing pneumomediastinum pneumopericardium and subcutaneous emphysema
Generated caption: a chest with a large, open lung