Fine-tuning Image-Captioning Models for Chest X-ray Interpretation

Amr MOHAMED - Thu DOAN

ING3 - IA - Group 2

Deep Learning

17/01/2024

Table of contents:

Introduction

Introduction

- Recent advancements in image captioning has not been exhaustively applied to medical imaging.
- Leveraging advanced **image captioning** techniques to interpret and describe complex medical images can **help healthcare practitioners** better diagnose and interpret medical images, speed up the diagnostic process, treatment planning, and overall patient management.
- Our goal is to develop a **robust model** capable of **generating precise and informative captions for radiological images**, aimed at improving diagnostic processes.

Methods

Methods: Dataset

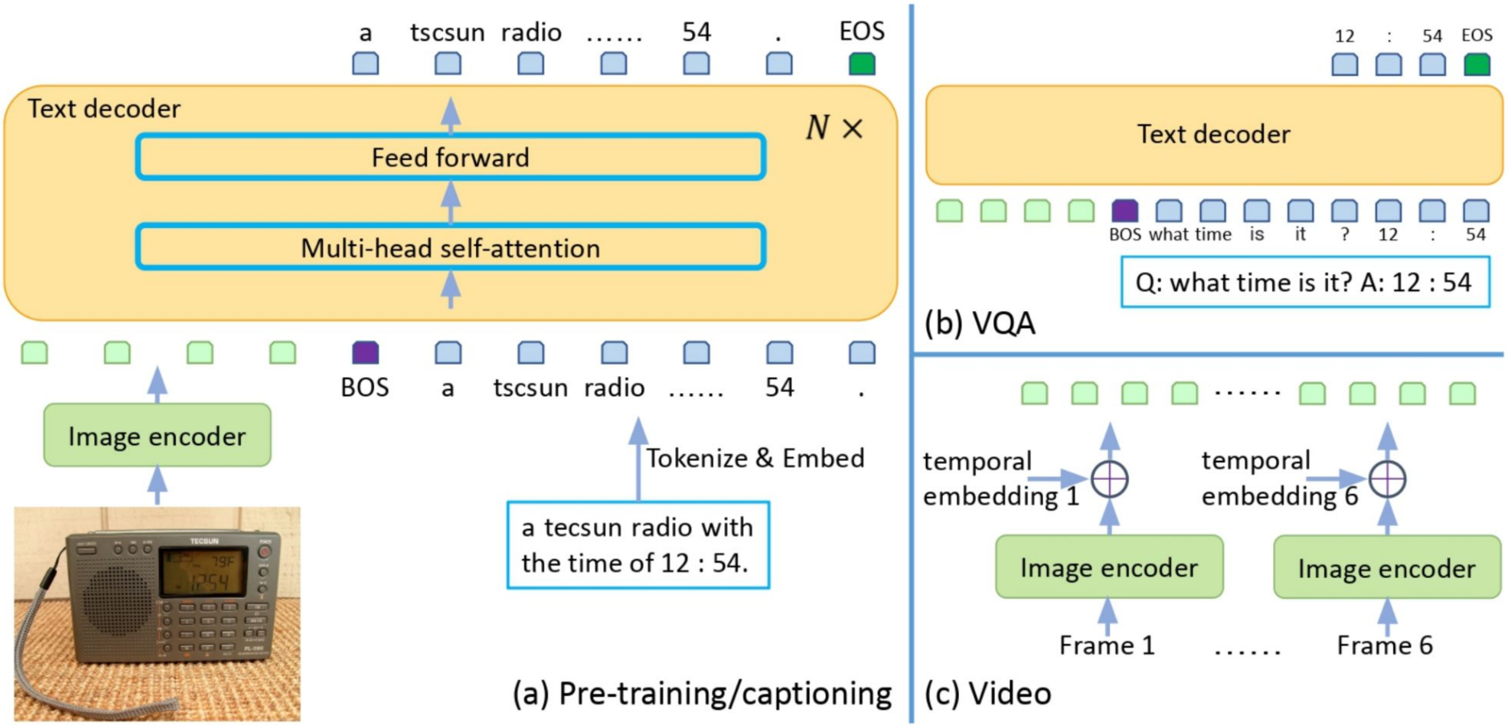
ROCO Dataset: Overview

- **Purpose**: Designed for image captioning in medical imaging.
- **Content**: Includes a variety of radiological images (X-rays, MRI, CT scans) from medical literature divided into ~**65k for training**, **~ 8.2k for testing**, ~ **8.2k for validation**
- **Annotations**: Accompanied by **descriptive text for each image**, providing detailed insights into medical conditions and imaging techniques.

Methods: Data preprocessing

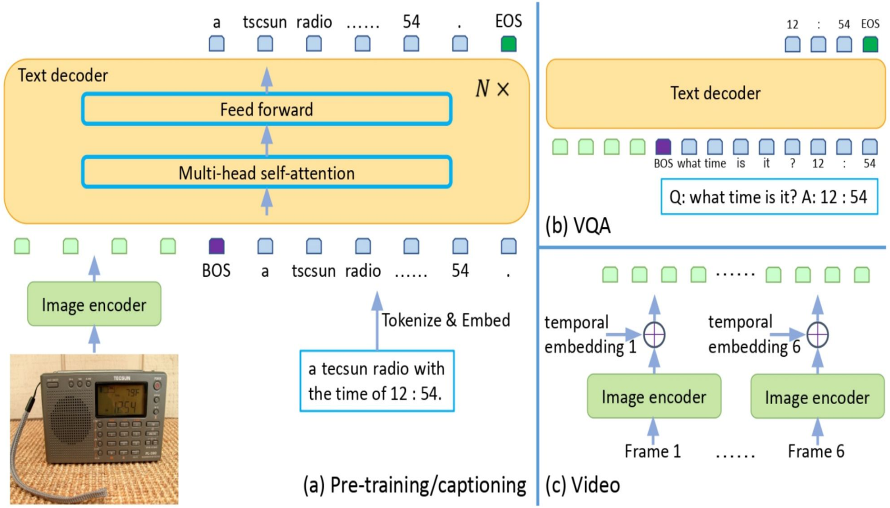
1. **Data Filtering**: **Selected chest x-ray images only** resulted in a reduction of the data size to:
   a. Training images: ~1.7k
   b. Testing images: ~200
   c. Validation images: ~200
2. **Images preprocessing: (Adjusted to BLIP Configurations for Fine-tuning)**
   a. Images:
      i. **Resized** to **384 x 384 x 3**
      ii. **Rescaled** by a factor of **1/255**
      iii. **Normalized** by their **mean**

Model: GIT (Generative Image2Text), base-sized



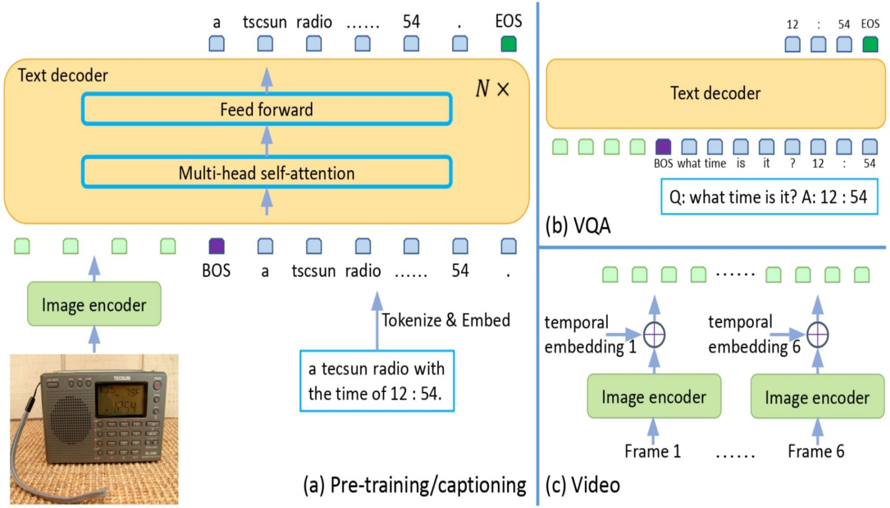
Model: GIT (Generative Image2Text), base-sized

- GIT is a **Transformer decoder** conditioned on both CLIP image tokens and text tokens. The model is trained using "teacher forcing" on a lot of (image, text) pairs.
- The goal for the model is simply to predict the next text token, giving the image tokens and previous text tokens.



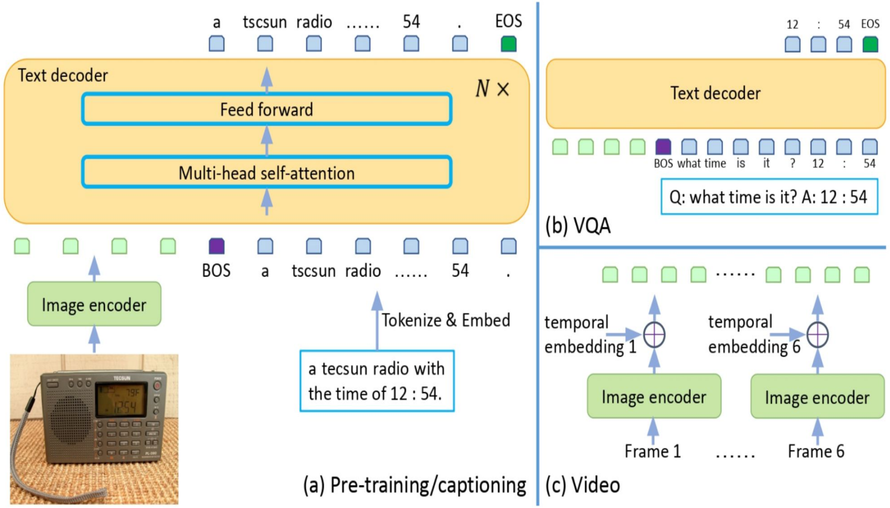
Model: GIT (Generative Image2Text), base-sized

- **Image Encoder**:
  - **Base Model**: The image encoder is initially pre-trained with contrastive tasks
  - **Process**: It takes a raw image and outputs a compact 2D feature map. This map is then flattened into a list of features.
  - **Projection**: These features are projected into 'D' dimensions through a linear layer and a layernorm layer.
  - **Purpose**: The projected features serve as the input for the text decoder.



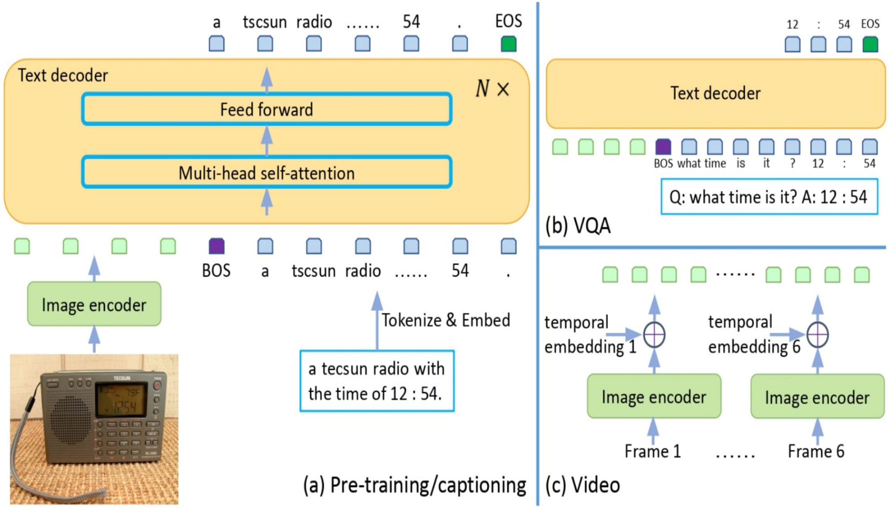
Model: GIT (Generative Image2Text), base-sized

- **Text Decoder**:
  - **Structure**: Comprised of a transformer module with multiple blocks, each containing a self-attention layer and a feed-forward layer.
  - **Process**: Text is tokenized, embedded into 'D' dimensions, added with positional encoding and a layernorm layer. These text embeddings are concatenated with image features for the transformer module's input.
  - **Decoding**: Begins with a [BOS] token and is decoded auto-regressively until an [EOS] token or a maximum step is reached. It employs a seq2seq attention mask.



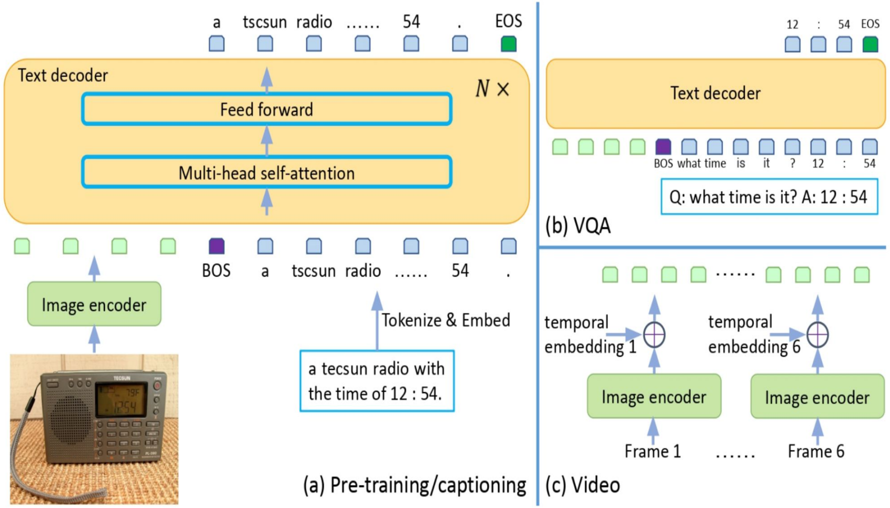
Model: GIT (Generative Image2Text), base-sized

- **Initialization and Training Approach:**
  - Text Decoder Initialization: Unlike the image encoder, the text decoder is randomly initialized, which has been shown to perform comparably to BERT initialization in experiments.
  - Training: All parameters, including those in the GIT (presumably a model name), are updated for better fitting VL tasks, differing from approaches like Flamingo, where the decoder is pre-trained and frozen.



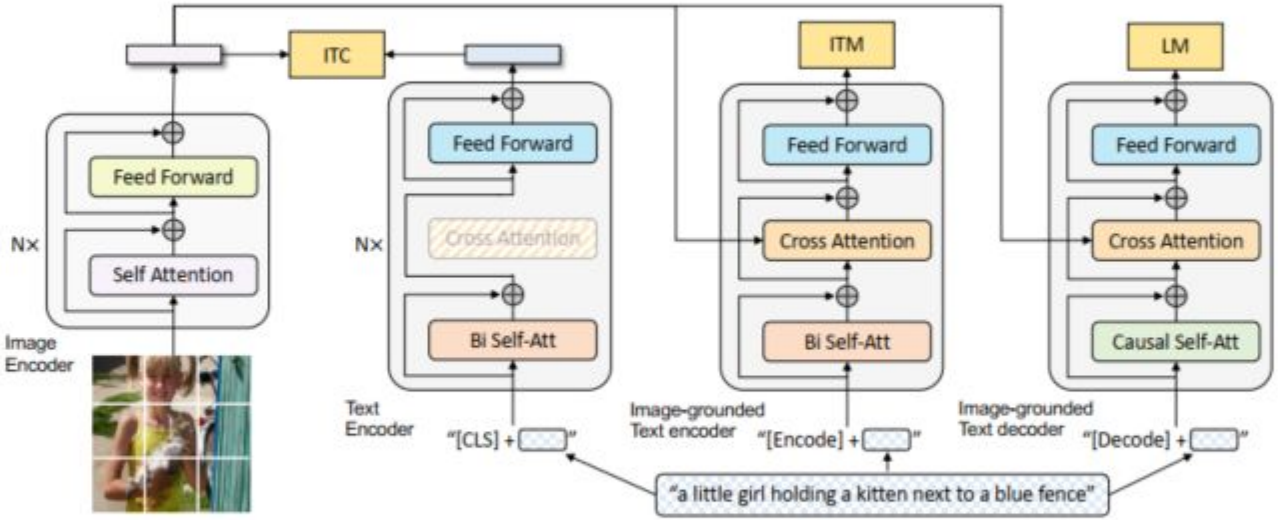
Model: GIT (Generative Image2Text), base-sized

- **Alternative Architectures**:
  - **Cross-Attention-Based Decoder**: An alternative to the self-attention-based decoder, which shows better performance in small-scale settings.
  - **Empirical Findings**: With large-scale pre-training, the self-attention-based decoder is more effective. This is attributed to the decoder's ability to process both image and text effectively, and the self-attention mechanism updates image tokens more aptly for text generation.



Methods: Model

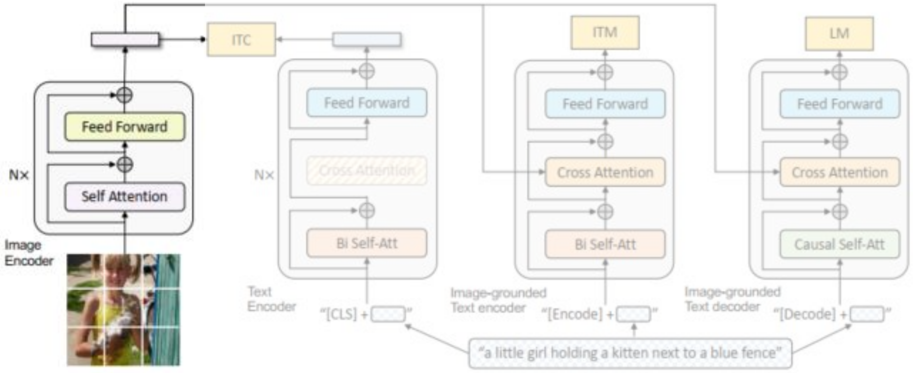
Salesforce Blip (Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation)

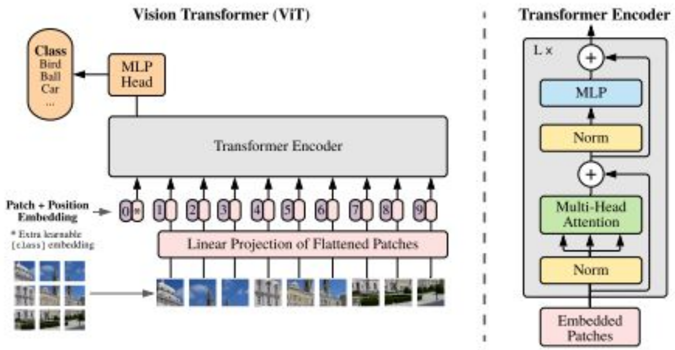


Methods: Model

Salesforce BLIP: Image Encoder

BLIP uses Vision Transformer (ViT) to divide an input image into **patches** and **encode them as a sequence of embeddings**



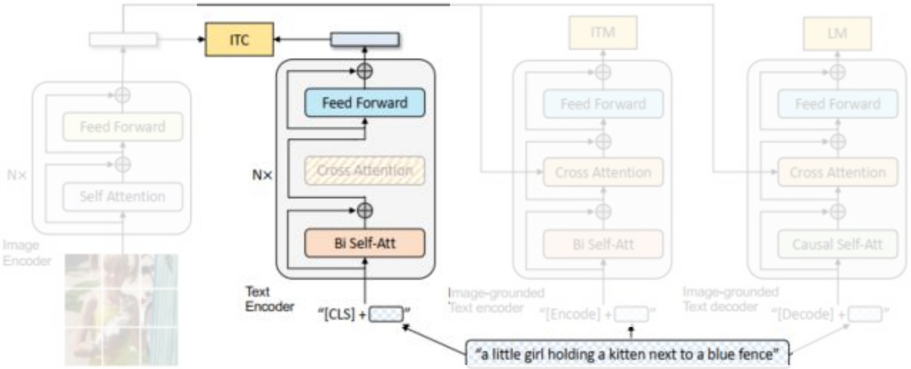


[Image source: An Image is Worth 16x16 Words: Transformers for Image Recognition at](#)

Methods: Model

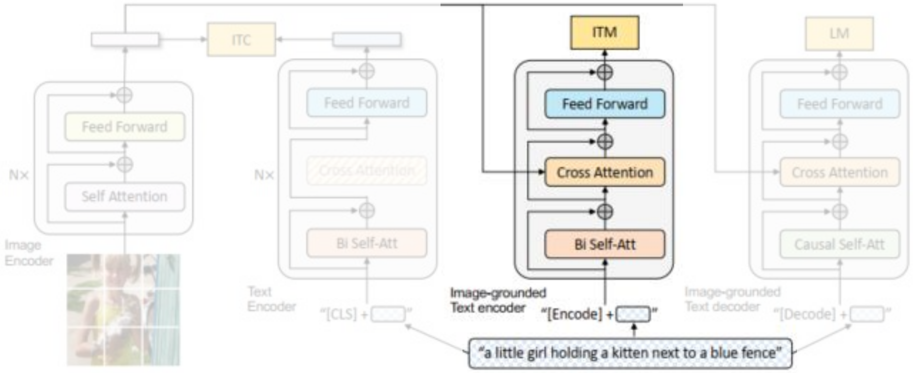
Salesforce BLIP : Text Encoder

- The text encoder separately **encodes image and text**.
- It has the architecture of **BERT**
- A **[CLS] token** is appended to the beginning of the text input to **summarize the sentence**
- **Image-Text Contrastive Loss** (ITC) is the loss function for this part of the model.
- It **aims** to **align** the **feature space** of the **visual** transformer and the **text** transformer by encouraging **positive image-text pairs** to have **similar representations** in contrast to the **negative** pairs.



Methods: Model

Salesforce Blip: Image-grounded Text Encoder

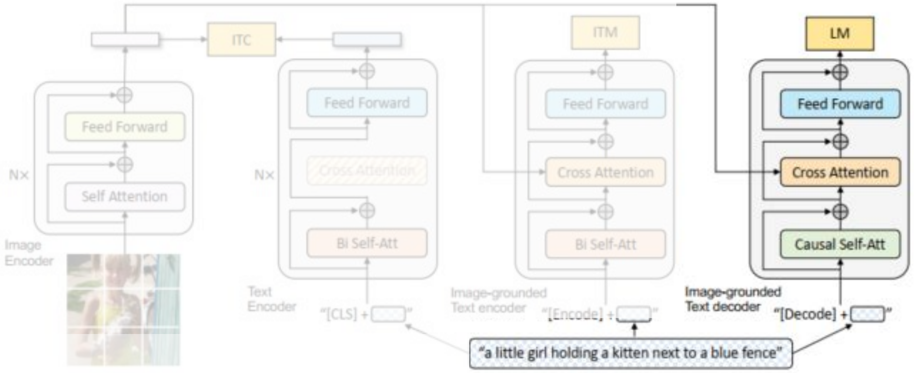


- Injects visual information by inserting **one additional cross-attention (CA)** layer between the self-attention (SA) layer and the feed forward network (FFN) **for each transformer block of the text encoder.**
- A task-specific **[Encode] token is appended to the text**, and the output **embedding of [Encode]** is used as the **multimodal representation** of the image-text pair.
- **Image-Text Matching Loss (ITM)** is minimized, which aims to **learn image-text multimodal representation** that **captures** the **fine-grained alignment between vision and language**. ITM is a **binary classification task**, where the model uses an ITM head (a linear layer) to predict whether an image-text pair is positive matched or not

Methods: Model

Salesforce BLIP: Image-grounded text decoder

- **Replaces the bidirectional self-attention** layers in the image-grounded text encoder **with causal self-attention** layers to motivate the **autoregressive generation of captions**



- **Language Modeling Loss (LM)** aims to **generate textual descriptions given an image**. It **optimizes a cross entropy loss** which trains the model to maximize the likelihood of the text in an autoregressive manner.

Methods: Fine tuning Plan

For the fine-tuning of the model on our custom dataset, we set the following parameters:

- **Learning Rate**: initially set to **5e-5.**
- **Optimizer** : **AdamW**
- **Weight decay**: **1e-08**
- **Number of Training Epochs**: **10** epochs
- **Loss function**: **Cross Entropy** loss
- **GPU efficient parameters:**
  - Mixed Precision Training (**fp16**): Training was performed using mixed precision to leverage lower memory GPUs efficiently.
  - **Per-device Train Batch Size:** Each training batch consisted of **8 samples**
  - **Per-device Eval Batch Size**: Each evaluation batch consisted of **2 samples**
  - **Gradient Accumulation Steps**: Gradients was **accumulated over 2 steps before performing a backward pass.**

Evaluation metrics

- **BLEU (Bilingual Evaluation Understudy):** Evaluates quality of machine-translated text against reference by measuring overlap in n-grams (word sequences) between machine generated text and reference texts.

$$P_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n-gram \in C} Conut_{clip}(n - gram)}{\sum_{\mathscr{C}' \in \{Candidates\}} \sum_{n-gram' \in \mathscr{C}'} Conut(n - gram')}$$

$$BLEU = BP \times \exp\left(\sum_{n=1}^{N} w_n \log P_n\right)$$

$$BP = \begin{cases} 1 & if\ c < r \\ e^{1-\mathrm{r}/\mathrm{c}} & if\ c > r \end{cases}.$$

Evaluation metrics

- **ROUGE** (Recall-Oriented Understudy for Gisting Evaluation)**:** Originally developed as a package for evaluation of text summaries. Recall is used to encourage detailed description.
  - **ROUGE-1**
    - Focus: Overlap of unigrams (individual words) between generated text and reference.
    - Measures: Lexical similarity on a word-by-word basis.
  - **ROUGE-2**
    - Focus: Overlap of bigrams (two consecutive words) between generated text and reference.
    - Measures: Phrase-level lexical similarity and basic structural coherence.
  - **ROUGE-L**
    - Focus: Longest Common Subsequence (LCS) between generated text and reference.
    - Measures: Sentence-level structure similarity and word order.
  - **ROUGE-Lsum**
    - Variation of ROUGE-L.
    - Focus: LCS, but applied to each sentence separately before aggregation.
    - Measures: More sensitive to sentence-level structure and coherence in multi-sentence summaries.

Evaluation metrics

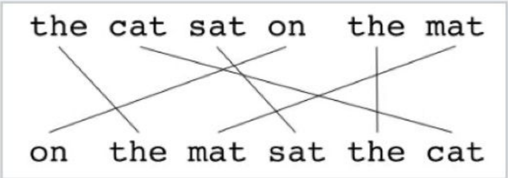
- **Meteor (Metric for Evaluation of Translation with Explicit ORdering):** It is based on an explicit word-to-word matching between the MT output being evaluated and one or more reference translations. It can also match synonyms. Calculate mapping between the candidate and reference caption. In conflict, mapping between least crosses is selected.

$$P = \frac{m}{w_t} \text{ and } R = \frac{m}{w_r}$$

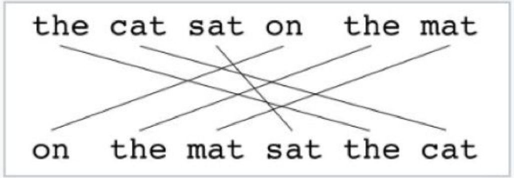
$$F_{mean} = \frac{PR}{\alpha P + (1-\alpha) R}$$

$$METERO = (1 - pen) \times F_{mean}$$

$$pen = \gamma \left(\frac{ch}{m}\right)^{\theta}$$



Example alignment (a).



Example alignment (b).

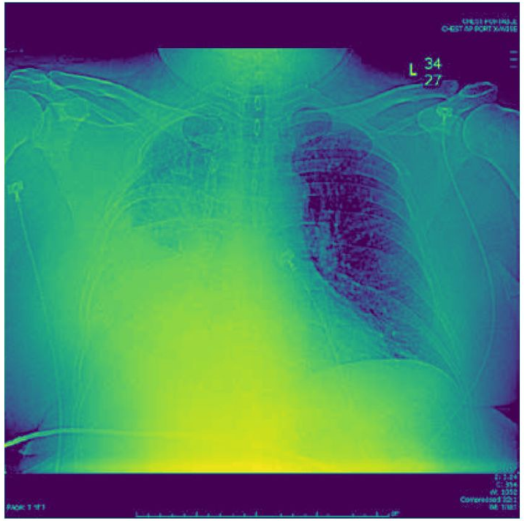
Results

Results: Evaluation Metrics Based

Model	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum	BLEU	Meteor
Git-base	0.1	0.01	0.8	0.8	0	0.05
Git-base-ft	0.35	0.22	0.33	0.34	0	0.14
Blip-base	0.2	0.06	0.19	0.19	0	0.08
Blip-base-ft	0.34	0.22	0.33	0.35	0.08	0.12

Comparison of the different evaluation metrics used for each of the models

Results: Visualizing X-rays along with the generated captions



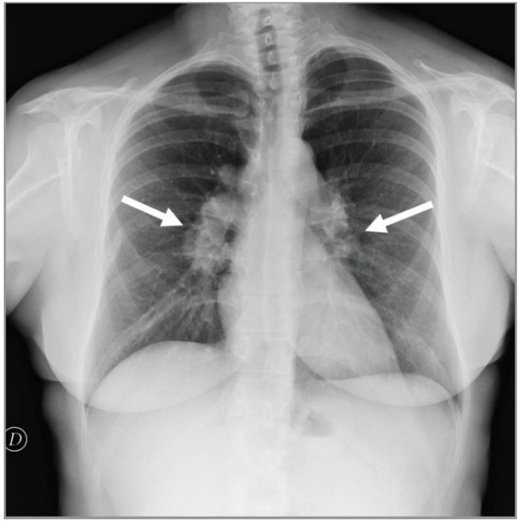
Original caption: coronary chest x-ray computed tomography (mediastinal window) showing massive pericardial effusion with an increased pericardial thickness (arrowheads)

Git base: what is the name of a pregnant woman?

Git fine-tuned: chest x - ray showing a large left - sided pneumothorax.',

Blip base: ct scan of the brain'

Blip fine-tuned: Chest - x ray showing pneumomediastinum and pneumothorax. arrow



Original caption: chest x-ray: multiple **bilateral opacities** and reticular pattern in **both thoracic fields**

Git base: a black and white image of human skeleton with a broken chest

Git fine-tuned: chest x - ray showing a large mass in the right hemithorax

Blip base: a chest with a chest with a chest

Blip fine-tuned: Chest - x ray showing a large right - sided mass with a mass - like **opacity** at the right **mid and lower lung zones**

Discussion

Discussion

- Blip **outperformed** the rest in semantic based metrics
- The model has to be used within the context of **providing suggestions**, rather than **making final decisions**
- For further advancements, the engagement of a **domain specialist** will be **crucial**. This expert will be responsible for carefully choosing examples, assessing the model's predictions for accuracy, and confirming the diversity of the data used.

Discussion

Limitations and future work

- Limitations were centred around the lack of **domain knowledge** to validate the **model predictions**
- The lack of access to **high-powered GPUs** to Fine-tune such complex architecture models
- The **data available** was **limited**, a deficiency in the expertise needed to ensure the diversity of this data.
- Future efforts will focus on developing more **comprehensive and advanced fine-tuning strategies** to allow the model to excel in diagnostics, so that we can gain confidence in the **model's predictive capabilities.**

References

- GIT: A Generative Image-to-text Transformer for Vision and Language
- An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale
- BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation
- Learning to Evaluate Image Captioning

Thank You!