

MEDICAL REPORT GENERATION USING CHEST X RAY IMAGES



Members :
Anand Raj
Shanun Randev
Abhishek Chiffon Muthuraja
Saniya Samir Shinde



PROBLEM STATEMENT

- **Objective:** Develop an AI-driven system to automatically generate clinically accurate radiology reports from chest X-ray images.
- **Key Goals:**
 - Enhance diagnostic efficiency and accuracy.
 - Reduce radiologists' workload by automating report generation.



DATA SOURCE

- The dataset is a public dataset from the MIMIC CXR Dataset. It comprises of **15,000** chest X-ray images and their associated radiology reports.
- The dataset has the following key features :
 - **Image File Path**: Location or link to the corresponding chest X-ray image.
 - **Findings**: A textual description of abnormalities or observations made by the radiologist.
 - **Impression**: A concise summary of the radiologist's primary conclusions.

BioGPT

Developer: Microsoft Research.

Model Size:

- **BioGPT Base:** ~ 347 million parameters.

Architecture:

- Based on GPT-2.
Vocabulary size: 42384 tokens.

Applications:

- Summarizing biomedical literature.
- Medical writing assistance.
- Information extraction from biomedical texts.

Special Features:

- Pre-trained on extensive biomedical literature.
- Outperforms other models in biomedical NLP tasks.

Our Model configuration

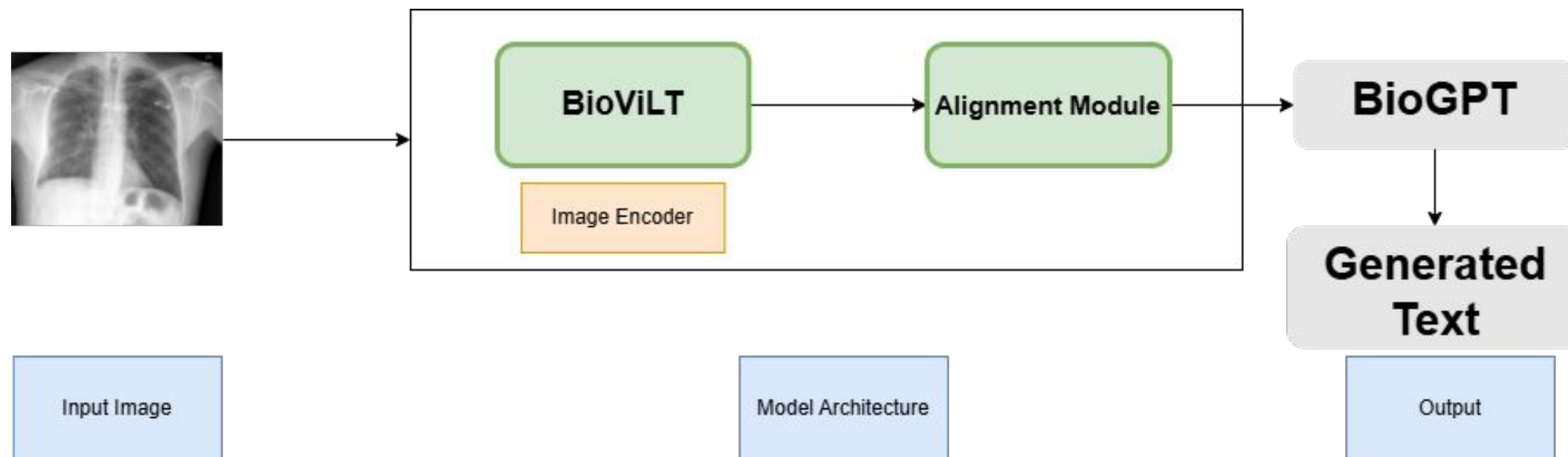
Parameter Efficient Fine Tuning

- LoRA (Low-Rank Adaptation) for efficient fine-tuning of "q_proj", "v_proj", "k_proj", "out_proj" (specific to attention mechanisms within the Transformer architecture)
- Rank : 16
- lora_alpha : 32 (scaling factor of LoRa)
- dropout : 0.1 (prevents overfitting)

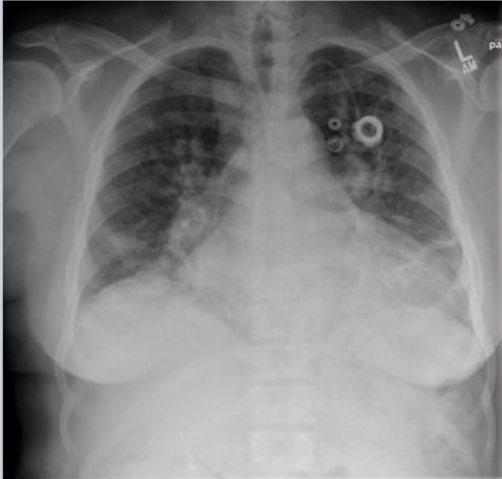
Generation Parameters :

- Configured to control the length, diversity, and coherence of the generated reports (e.g., max_length=150, temperature=0.8, top_k=30, top_p=0.85)

BioViLT with BioGPT Base Model



Results: Base Model

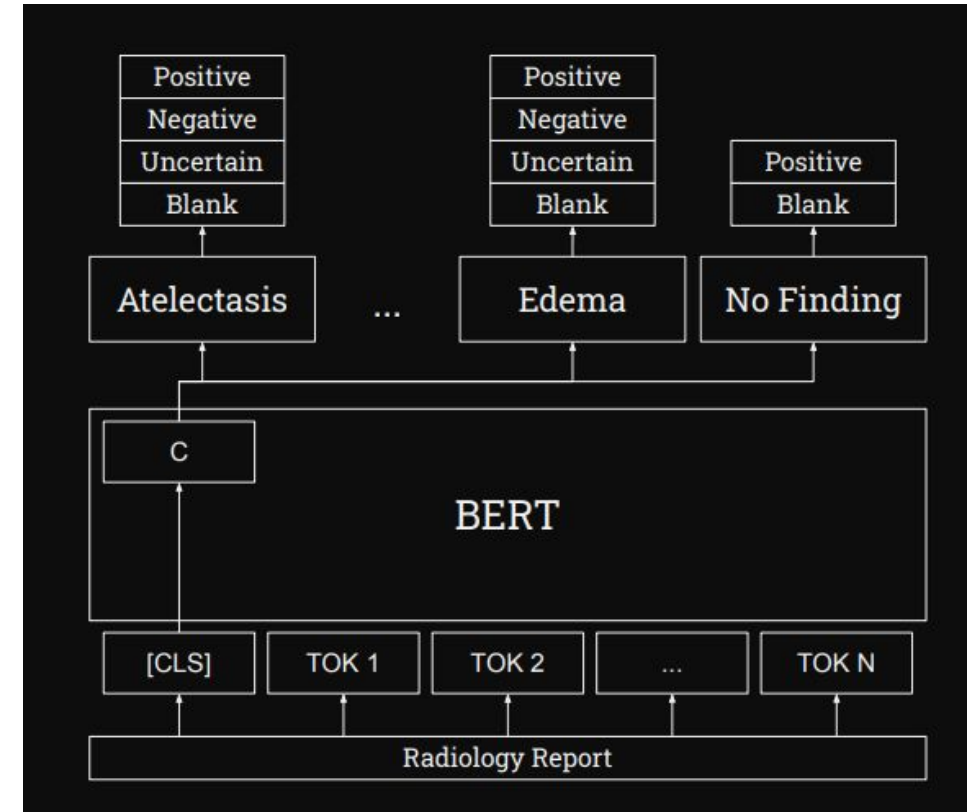
Chest X Ray Image	Ground Truth	Prediction
	<p>Report:</p> <p>Bilateral patchy pulmonary opacities noted. Interval improvement in left base consolidative opacity. Pulmonary vascular congestion again noted. Stable enlarged cardio mediastinal silhouette. Stable left XXXX. No evidence of pneumothorax. No large pleural effusions.</p>	<p>Report:</p> <p>Sternotomy and repair of a right-sided chest wall pericardiac loop in association with congestive heart failure is unchanged, but on the opposite side of this. The lungs are well expanded without evidence for acute or chronic consolidation. No pleural or pneumothorax found. No new areas of focal airspace disease. There has been interval resolution of left basilar atelectasis.</p>

Chexbert

Report labeling task is to extract presence of one or more text observations from radiology report.

Model Architecture :

- This is a modification of BERT base architecture with 14 linear heads in total. 12 heads correspond to medical abnormalities, 1 to medical support devices, 1 to no finding. Each radiology report is tokenized and the maximum number of tokens are capped at 512.
- Final layer's hidden state corresponding to CLS token is fed as input into each of the linear heads.



ChexNet

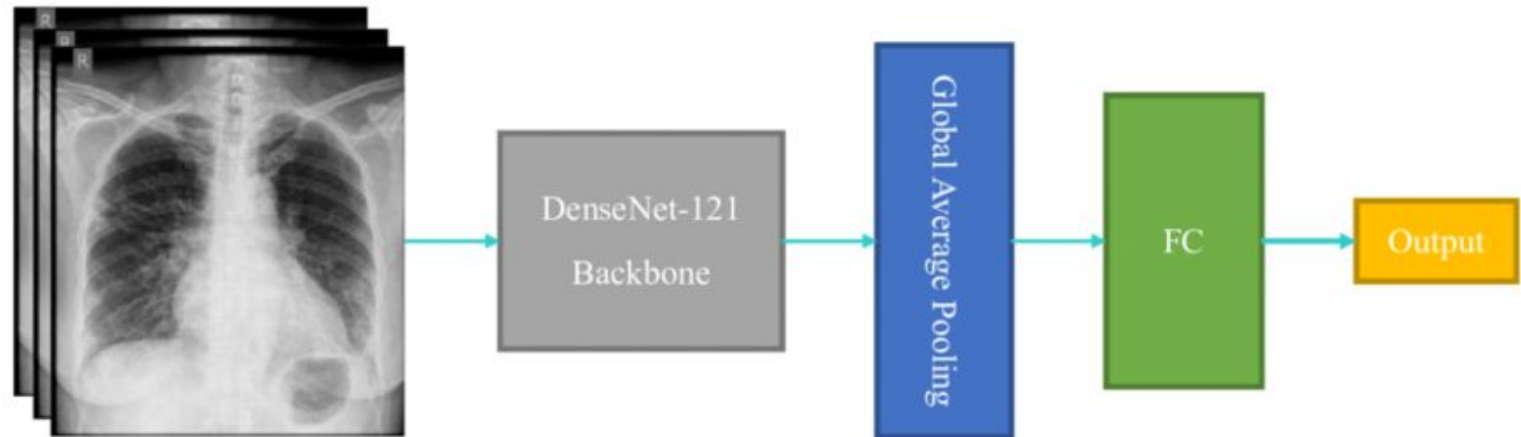
- Multi label image classification. DensenetNet Pretrained on chest xrays
- Labels passed as input to the language model.

Custom Classifier Head

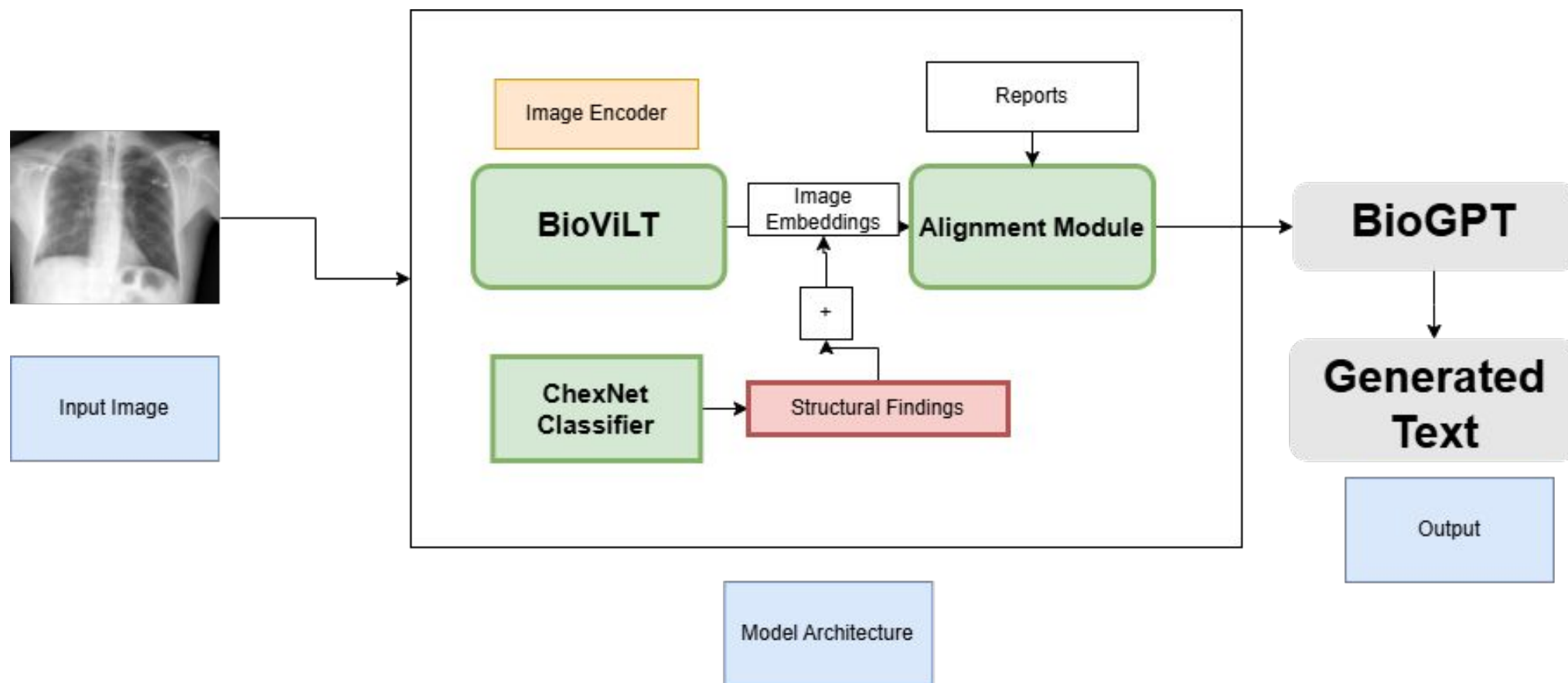
- 1024,288 parameters trained
- Additional 512-unit hidden layer
- Dropout (0.3) for regularization
- Sigmoid activation for multi-label output

Selective Layer Freezing


- First dense blocks frozen
- Trained last 2 dense blocks



BioGPT + Labels Model



Result: BioGPT + Labels Model

Chest X Ray Image	Ground Truth	Prediction
	<p>Clinical Findings: Opacity</p> <p>Report: there is a large right hilar opacity, xxxx in the posterior segment of the lung. the heart size and mediastinal contour are normal. no pneumothorax or pleural effusions. this appears to be hyperinflated no focal airspace consolidations.</p>	<p>Clinical Findings: Opacity</p> <p>Report: there is a rounded opacity in the right lower zone measuring 2.0 cm which is xxxx to be in the posterobasal segment. there is of uncertain etiology but would benefit from followup at xxxx some concern for neoplasm. a xxxx is recommended. no airspace disease, effusion or cavitary nodule. normal heart size and mediastinum. visualized xxxx of the chest xxxx are within normal limits.</p>

Biomedclip

Core Architecture Components:

- Vision Encoder: ViT-Base/16
 - Transformer layers with self-attention
 - Output dimension: 512
- Text Encoder: PubMedBERT-base
 - Specialized BERT model trained on biomedical text
 - Output dimension: 512

Key Differences from Standard CLIP:

- Domain-specific (medical) model
- Uses PubMedBERT instead of standard text encoder

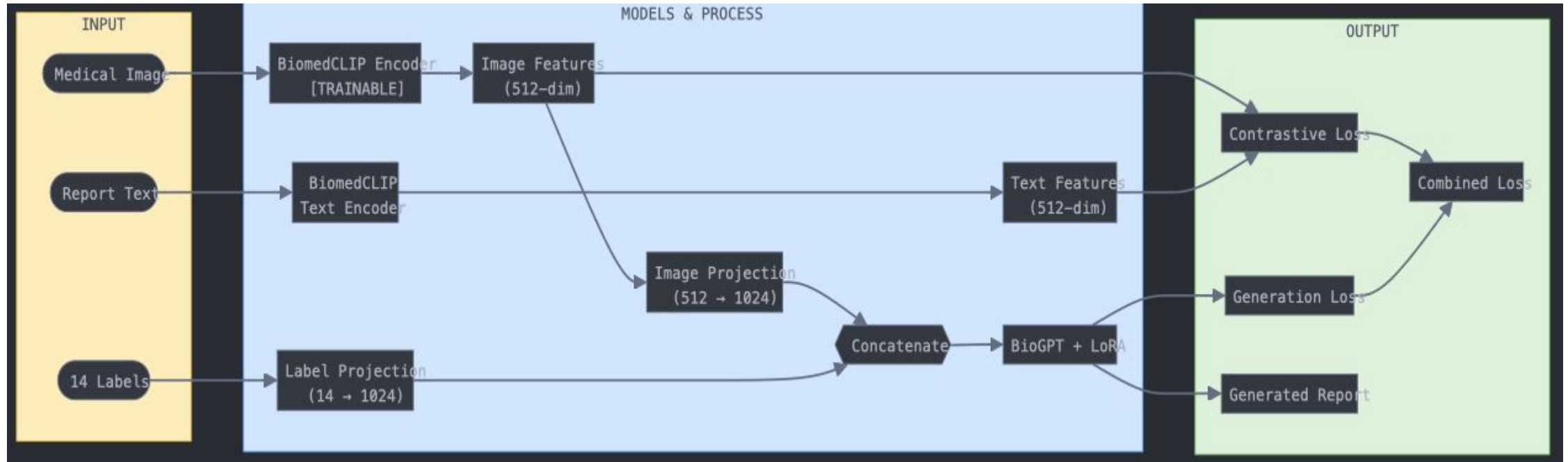
Our approaches

- Trained biomed image encoder
- Biomed as image encoder to create text embedding

We only trained:

1. The Report Generator (with LoRA)
2. The input projection layer
3. The label projection layer

Our Biomed with BioGpt Implementation



BiomedCLIP Stage:

- Pass images through CLIP image encoder (trainable)
- Pass findings text through PubMedBERT text encoder
- Normalize both feature vectors
- Calculate similarity matrix (logits)
- Compute contrastive loss

Report Generator Stage:

- Project CLIP's image features (512) to BioGPT's hidden size
- Project label vector (14) to BioGPT's hidden size
- Concatenate projected image and label embeddings
- Get token embeddings for target text
- Concatenate all embeddings: [img_proj; label_proj; token_embeddings]
- Create attention mask for the full sequence
- Pass through BioGPT with LoRA

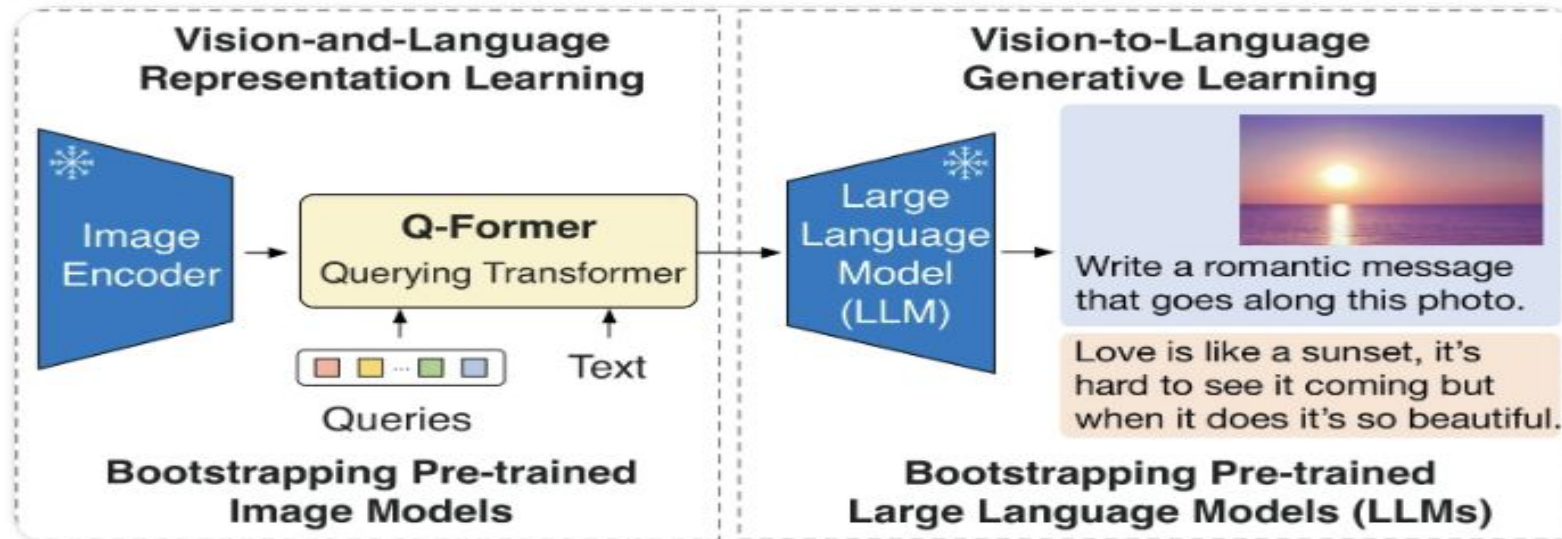
BLIP2

BLIP2 Overview:

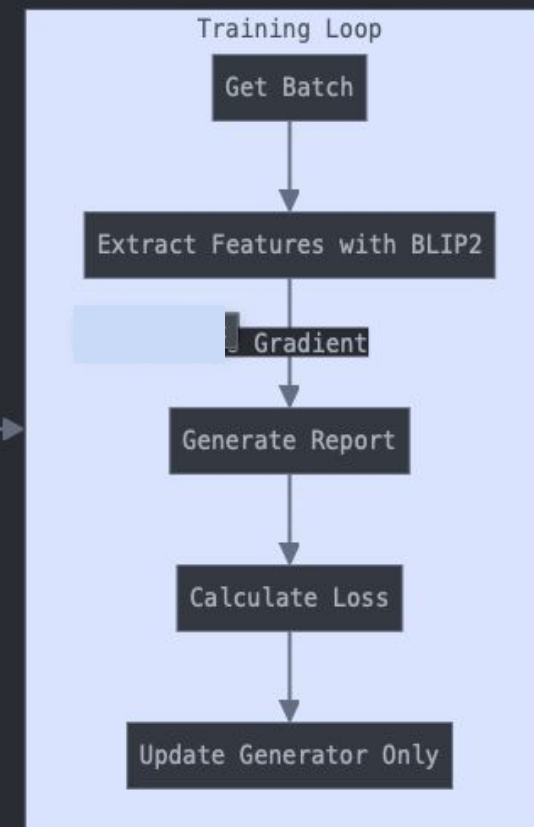
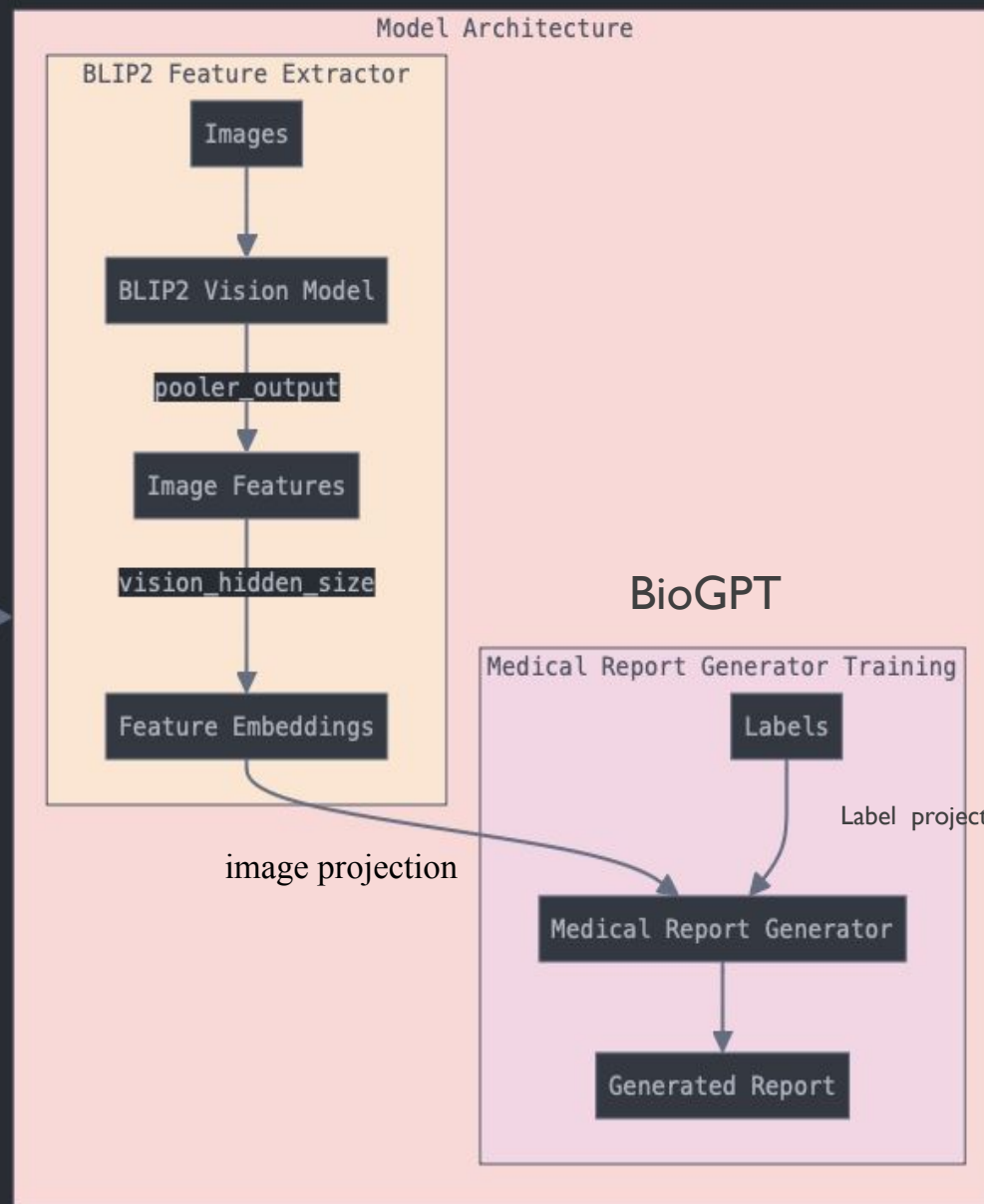
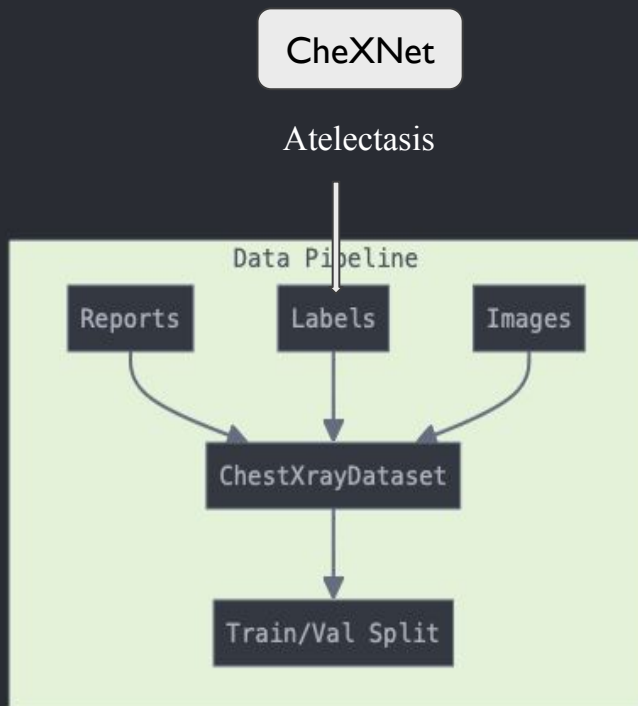
- BLIP2 is a vision-language model developed by Salesforce Research
- Uses a pretrained vision encoder (ViT) to process images
- Main innovation is Q-Former architecture that bridges visual embeddings and chexnet labels
- Acts as a learnable interface between frozen encoders
- Enables zero-shot transfer between different tasks
- Unlike CLIP or traditional transformers where tokens directly represent image/text, these queries learn to be optimal "information extractors"

In our implementation:

- We use BLIP2 (specifically "Salesforce/blip2-opt-2.7b") purely as a feature extractor
- We extract image features using its vision model.

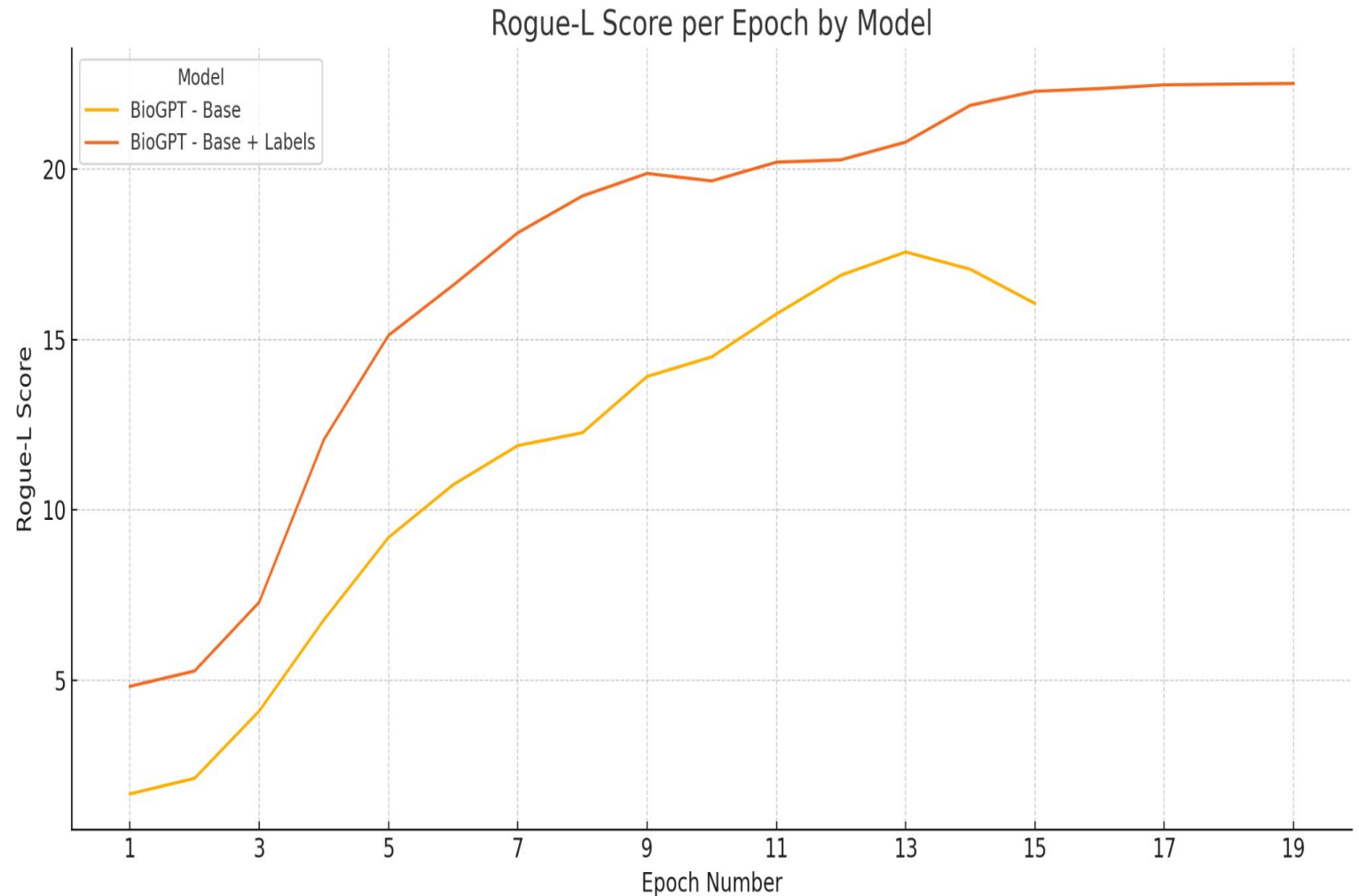


Our Blip2 with Biogpt Model

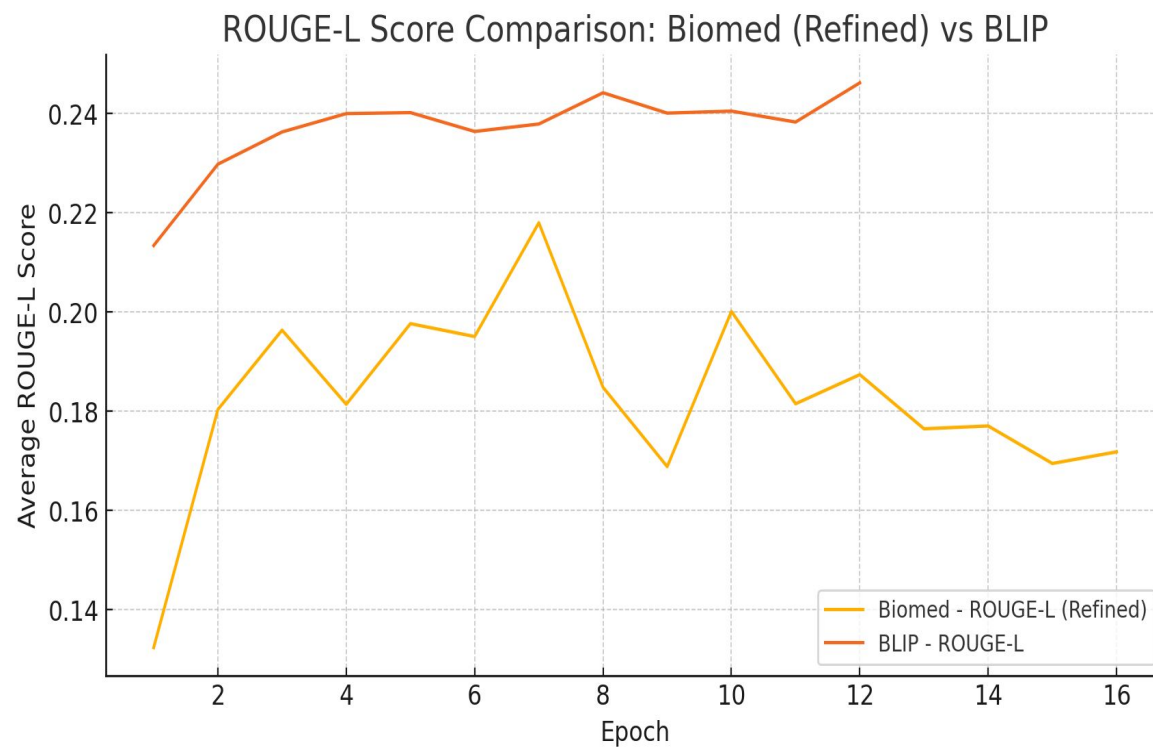


Model comparison

- The snapshot describes the Rouge L score for 2 configurations.
- The Rouge-L score focuses on the longest common subsequence between the generated finding and actual finding text, which capture sentence level structure and ensures the order of words is considered while evaluation.
- The model with integration of BioGPT and chexnet(predicted) labels have a higher Rouge score than using a BioGPT with image embeddings.



Biomed vs blip



Challenges

- Limited computation power
- Smaller models fail to capture the findings, larger models are required.
- Using Image features standalone isn't enough to generate accurate reports.
- Clinical findings extraction is done by a model which introduces errors.
- Huge mimic dataset image files sizes:
- Downloading through scripts took days in ec2 machine and transferring was hard.



Thank you