Beni Suef University
Faculty of computers and artificial intelligence

Advancing Medical Diagnostics with Vision-Language Models: A Framework for Image Captioning.

Supervised by:
Assoc. Prof. Ibrahim Eldsoky
Dr. Noha Yahia

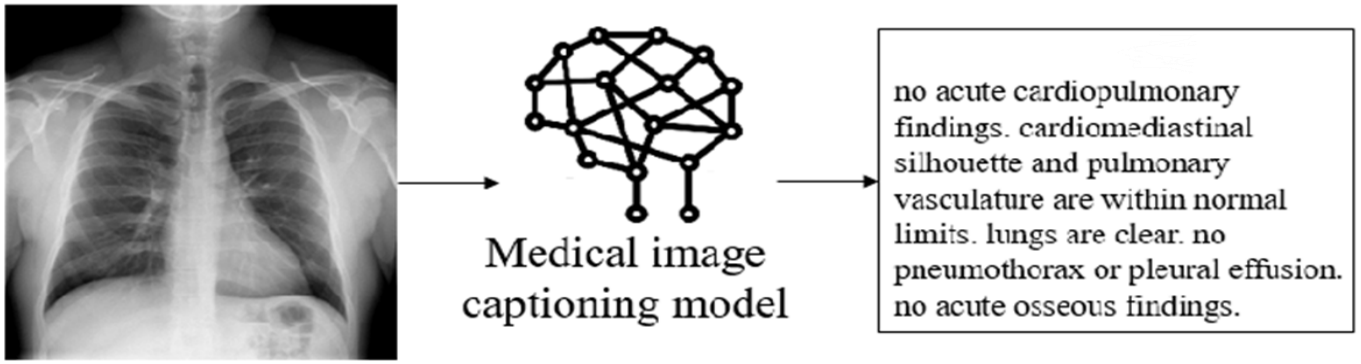
Presented by:
Wafaa Abdullah

Table of contents

Introduction

Introduction

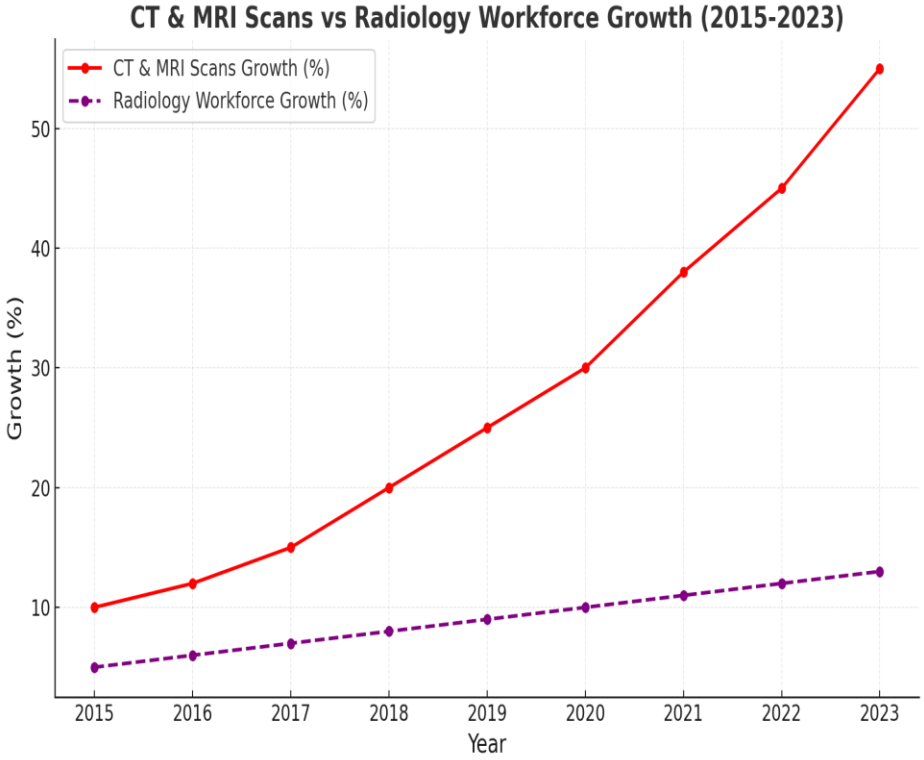
- Medical image captioning represents the content of an input image in a natural language by using various machine and deep learning models. It initially extracts the content information and, afterward, it provides descriptive sentences.



Radiology Statistics

- **45 million** diagnostic imaging studies were conducted in England in 2023.

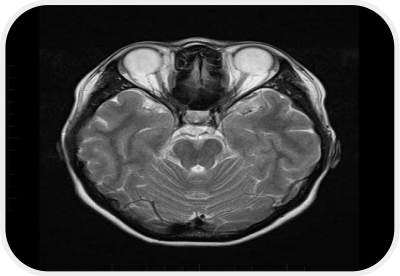
- **4.2 billion** imaging studies are conducted globally each year, and this number is steadily increasing.



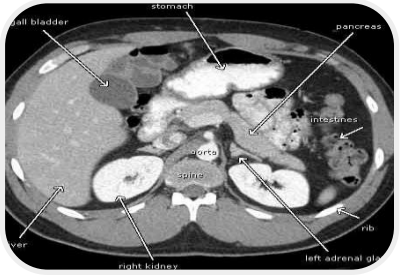
The Role of AI in Radiology

- **Challenges in Radiology:**
  - Expanding case burden: Growing complexity of imaging cases.
  - Aging populations: Leading to higher demand for imaging services.
  - Workforce shortages: Increasing the strain on radiologists and physicians.
- **Objectives for AI:**
  - AI can improve efficiency and accuracy in imaging interpretation.
  - Offers solutions to handle increased demand and mitigate workforce challenges.

Medical Imaging



MRI



CT

Ultrasound

X-ray

Overview of medical imaging

- **Ultrasound:** Real-time imaging for organs and blood flow.
- **X-rays:** Good contrast between bone and air.
- **CT scans:** 3d X-ray provide precise views of bones, organs, and tissues.
- **MRI:** Excellent for soft tissue imaging like brain and muscles.

Basic radiology workflow

1. Patient sent for radiology scan.

2. Scan is taken by physicians

3. Radiologist examines the image and describes their findings

4. Resulting report goes back to referring clinician to inform care and treatment

Problem statement

Problem statement (Research gap)

- **Inefficiencies in current manual medical image captioning methods.**
  - Labor intensive
  - High cost
  - Human error
  - Slow turnaround time

- **Need to enhance the model evaluation performance metrics (BLEU – ROUGE).**

BLUE: Bilingual Evaluation Understudy
ROUGE: Recall-Oriented Understudy for Gisting Evaluation.

Literature review

Year	Paper title	Imaging Modalities	Dataset	Method	Results
2020	Automatic Radiology Report Generation Based on Medical Images	Chest X-ray	MIMIC-CXR	Generative encoder-decoder model	BLEU-1: 0.529 ROUGE-L: 0.453
2022	AlignTransoformer to alleviate the data bias problem and model the very long sequence for medical report generation	Chest X-ray	IU X-Ray, MIMIC-CXR	Hierarchical transformer with multi-stage attention for caption generation.	BLEU-1: 0.378 MIMIC-CXR BLEU-1: 0.484 IU-Xray
2022	Chest X-ray caption generation with cheXNet	Chest X-ray	IU X-Ray	LXMERT(Learning Cross-Modality Encoder Representations from transformers).	BELU1: 0.498 ROUGE-L: 0.379

Year	Paper title	Imaging Modalities	Dataset	Method	Results
2023	A Concise Model for Medical Image Captioning.	Chest X-ray	MIMIC-CXR dataset	Encoder-decoder model (Show, Attend, and Tell)	BERTScore: 0.643
2023	Medical Image Captioning Using Optimized Deep Learning Model.	Chest X-ray	MIMIC-CXR dataset	Deep learning-based Show, Attend, and Tell model with encoder-decoder architecture, optimized using SPEA-II.	BERTScore: 0.697
2023	Customizing General-Purpose Foundation Models for Medical Report Generation.	Chest X-ray	MIMIC-CXR dataset	Customizing general-purpose pre-trained models (FMs) for medical report generation using ChatGLM-6B and the P-tuning technique, following the BLIP-2 approach.	Achieved a 0.9% improvement under the ROUGE-1 metric.

Year	Paper title	Imaging Modalities	Dataset	Method	Results
2024	Towards a Holistic Framework for Multimodal Large Language Models in Three-dimensional Brain CT Report Generation.	CT (3D brain imaging)	CQ500	3D Brain CT dataset with Clinical Visual Instruction Tuning (CVIT) to fine-tune BrainGPT.	91% accuracy in generating brain CT reports.
2024	Concept-Aware Medical Caption Generation with Enhanced Attention Mechanisms.	CT, MRI	ROCO	Integrated concept detection into attention mechanisms using Swin-V2 and BEiT+BioBart models.	F1 score: 0.61998 BERTScore: 0.5794
2024	Medical Image Interpretation with Large Multimodal Models.	CT, X-ray, MRI	MedTrinity-25M)	Experimentation with variants of LLaVA models for medical images, LMM, MoonDream2, IDEFICS 9B, visionGPT2, CNN-Transformer.	BERTScore: 0.63 ROUGE: 0.25

Year	Paper title	Imaging Modalities	Dataset	Method	Results
2024	Vision-Language Model for Generating Textual Descriptions From Clinical Images.	Chest x-ray	MIMIC-CXR IU X-RAY.	a radiology report generation model named ClinicalBLIP.	METEOR score of 0.570 on IU X-RAY. 0.365 on MIMIC-CXR.
2024	UIT-DarkCow team at ImageCLEFmedical Caption 2024: Diagnostic Captioning for Radiology Images Efficiency with Transformer Models.	X-ray, CT	ImageCLEF medical 2024	Proposed three models: VisionDiagnostor-ClinicalT5 and VisionDiagnostorBioBART (encoder-decoder architecture), VisionDiagnostor-Q-BioMistral based on BLIP2 architecture with Query Transformer using Large Language Models (LLM).	VisionDiagnostor-BioBART achieved third place on the leaderboard with the highest BERTScore of 0.6267.

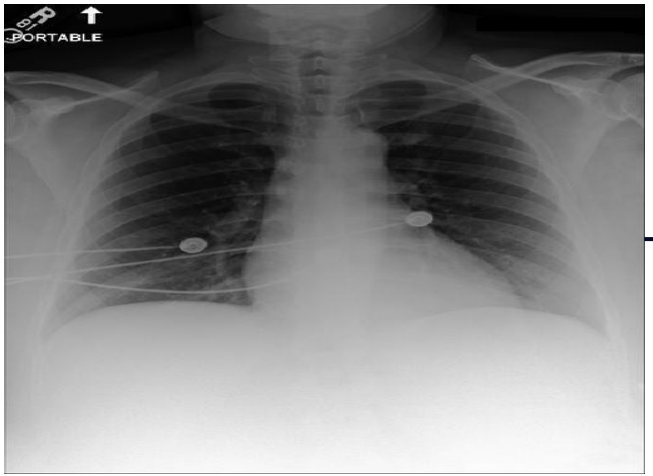
Datasets

Datasets

Dataset name	Size	Modality
MIMIC-CXR	200k images with paired reports.	Chest X-rays
ROCO	81k image-text pairs	X-ray, MRI, CT, etc
Medtrinity_brain	180k image-text pairs	Brain CT

MIMIC-CXR

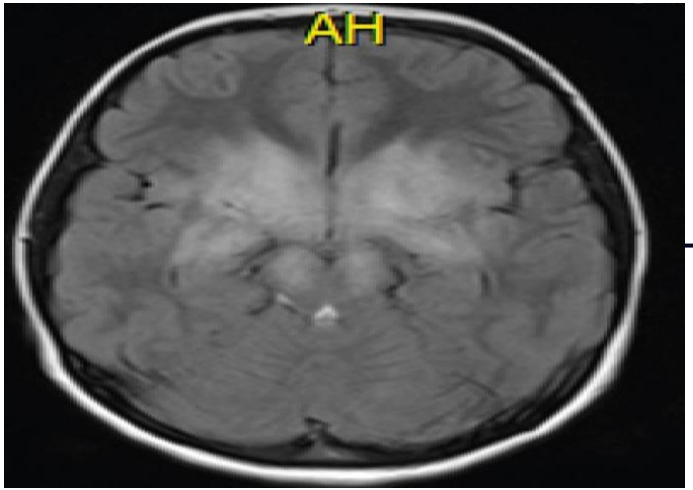
- **The MIMIC Chest X-ray (MIMIC-CXR)** dataset is a large publicly available dataset of chest radiographs in DICOM format with free-text radiology reports. This dataset consists of 377,110 images corresponding to 227,835 radiographic studies.



Heart size is normal. The mediastinal and hilar contours are normal. The pulmonary vasculature is normal. Lungs are clear. No pleural effusion or pneumothorax is seen. There are no acute osseous abnormalities.

ROCO

- **Radiology Objects in COntext (ROCO)** dataset contains over 81k radiology images with several medical imaging modalities including CT, Ultrasound, X-Ray, Mammography, MRI, Angiography. All images in ROCO have corresponding caption .



Brain MRI Flaire image showing hyperintensities in basal ganglias.

MedTrinity-25M

- **MedTrinity-25M** is a large multimodal medical dataset across 10 modalities (e.g., X-ray, CT, MRI). It includes detailed global and local annotations like disease types, bounding boxes, and segmentation masks. Using an automated pipeline, it generates image-text triplets without paired data. Sourced from 90+ origins, it supports tasks like captioning, report generation, and segmentation, making it ideal for training advanced medical AI models.
- **MedTrinity-brain** another version from MedTrinity-25 with 180k brain CT and caption.

This CT scan of the brain shows a region of interest located centrally and in the upper-middle area, indicative of an intracranial hemorrhage, subdural hematoma, or blebs, which are characterized by an abnormal density compared to the surrounding brain tissue.

Methodology

Key components of medical image captioning

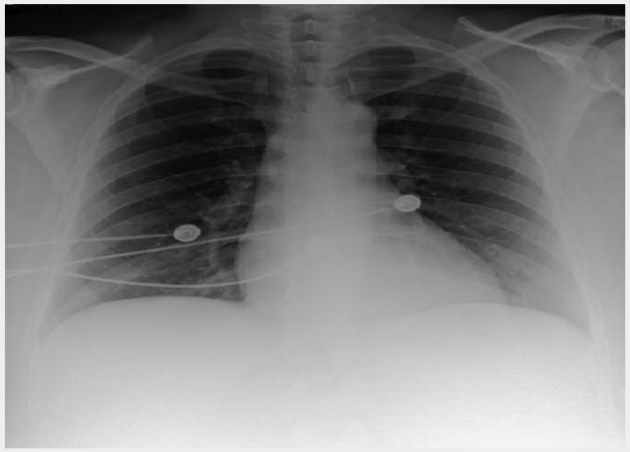
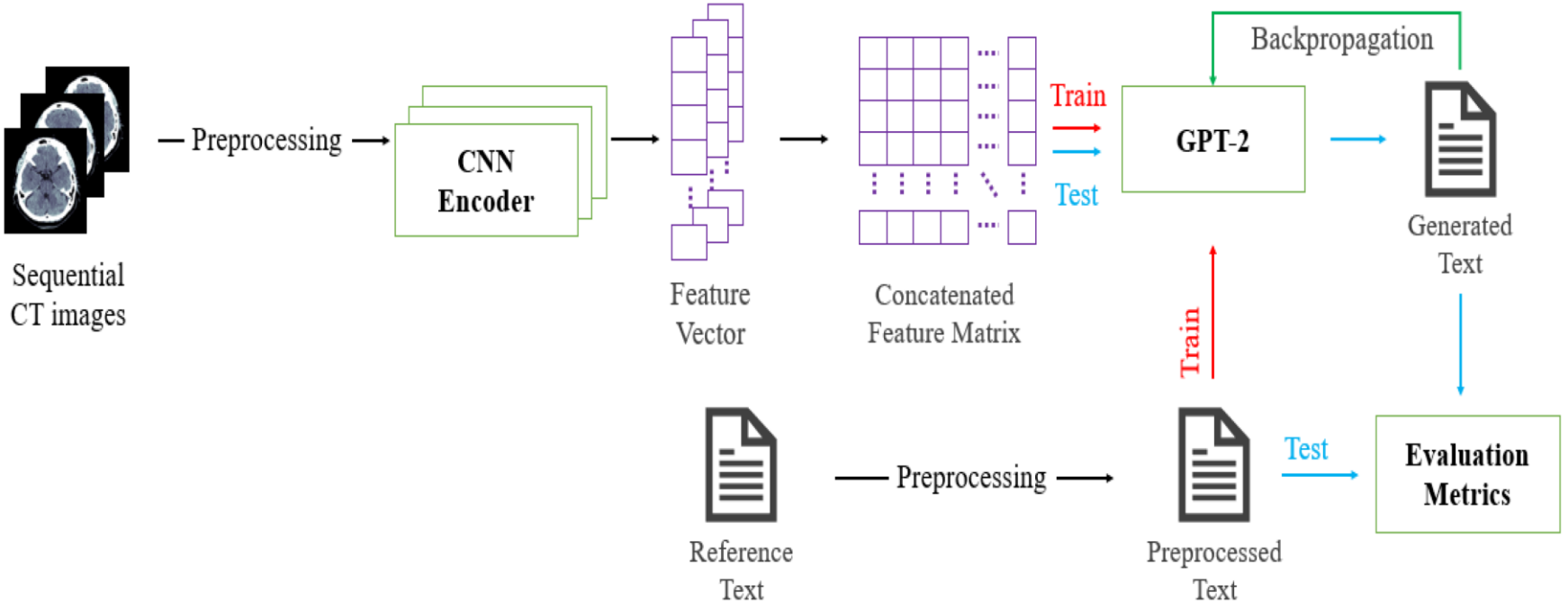


Image understanding

Heart size is normal. The mediastinal and hilar contours are normal. The pulmonary vasculature is normal. Lungs are clear. No pleural effusion or pneumothorax is seen. There are no acute osseous abnormalities.

Language generation

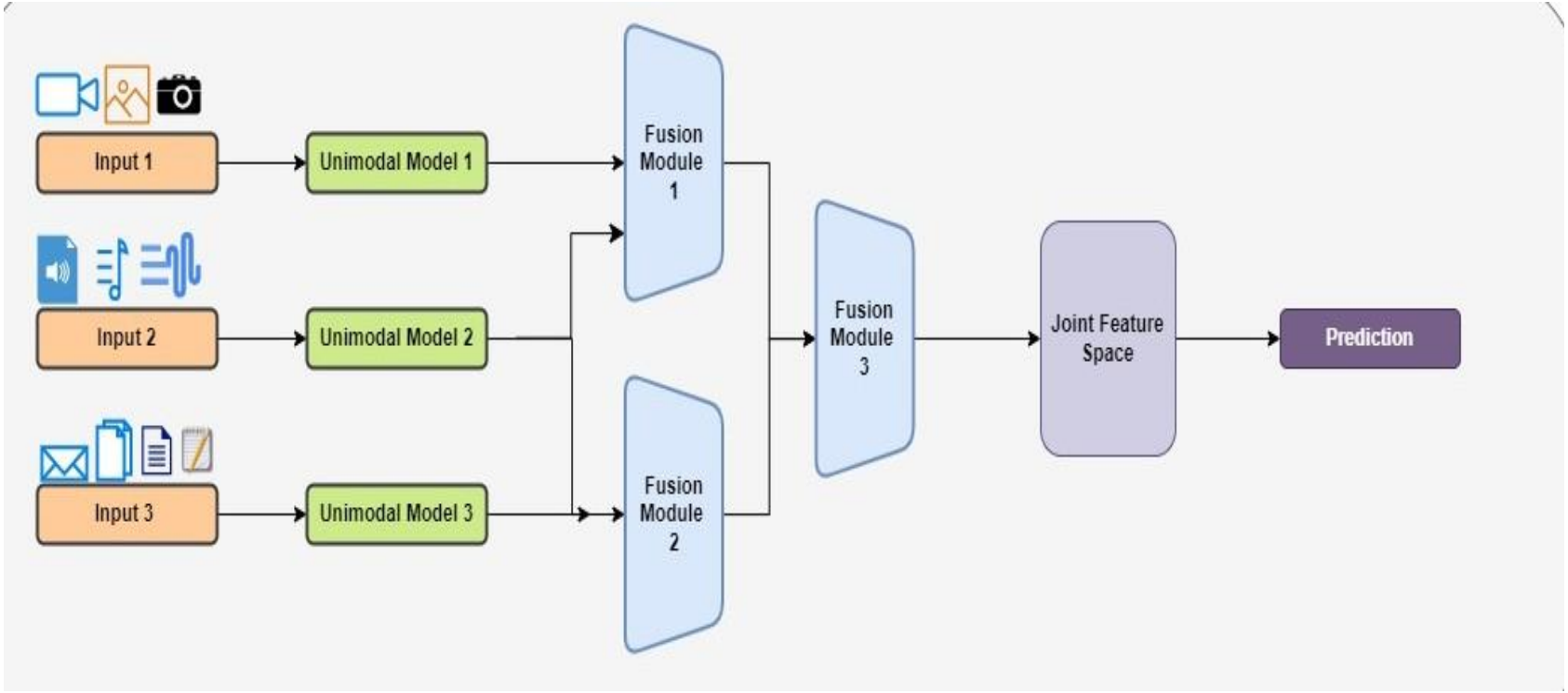
CNN-Transformer



Multimodal

- A multimodal model, is a model that can be used to perform multimodal tasks by processing data coming from multiple modalities at the same time.
- Multimodal models are trained to integrate and process data from sources like images, videos, text, audio etc. The process of combining these modalities begins with multiple unimodal models. The outputs of these unimodal models (encoded data) are then fused using a strategy by the fusion module.
- The overall task of the fusion module is to make a combined representation of the encoded data from the unimodal models. Finally, a classification network takes up the fused representation to make predictions.

Multimodal


Input 1
Unimodal Model 1
Input 2
Unimodal Model 2
Input 3
Unimodal Model 3
Fusion Module 1
Fusion Module 2
Fusion Module 3
Joint Feature Space
Prediction

Vision-language models

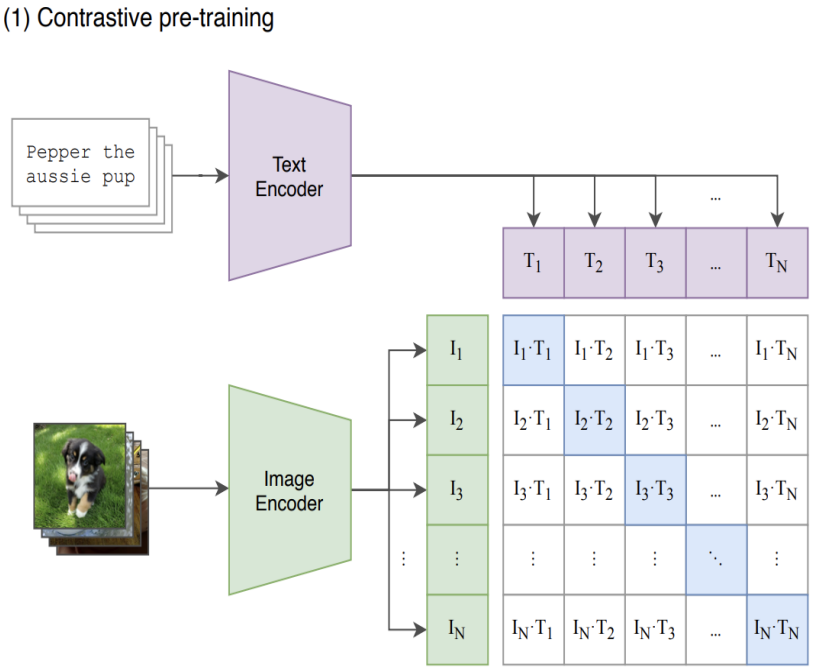
- Vision language models are broadly defined as multimodal models that can learn from images and text. They are a type of generative models that take image and text inputs, and generate text outputs.

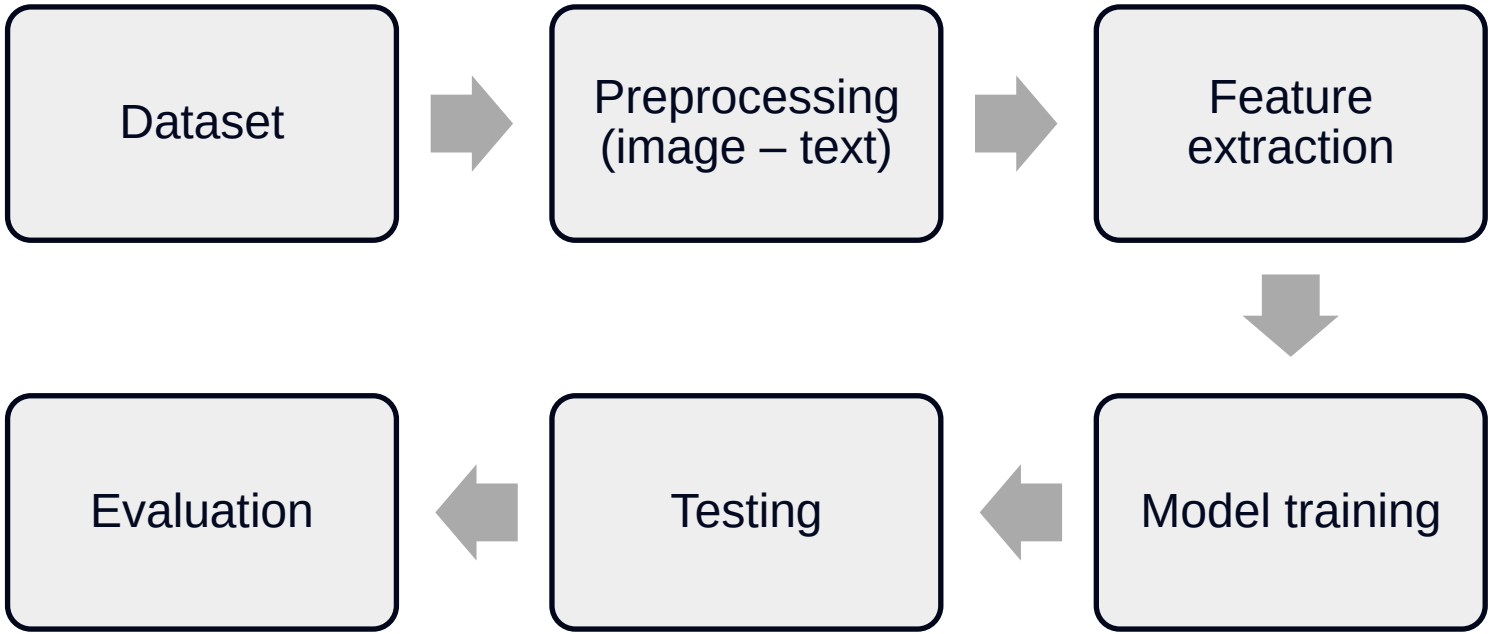
- Large vision language models have good zero-shot capabilities, generalize well, and can work with many types of images, including documents, web pages, and more. The use cases include chatting about images, image recognition via instructions, visual question answering, document understanding, image captioning, and others.

Vision-language (CLIP)

- Learn an image encoder and a text encoder.
- CLIP(Contrastive Language-Image Pretraining):

1. Use paired image-text dataset.
2. Embed text and image in the same space.
3. Contrastive objective: high similarity between true pairs low similarity between non-pairs.




Dataset
Preprocessing (image – text)
Feature extraction
Model training
Testing
Evaluation

Proposed approach

- **Dataset Selection and preprocessing:**
  - Cleaning the data
  - Normalize images
  - Splitting data
- **Feature extraction:**
  - Encoder
  - Concatenated feature matrix.
- **Model training:**
  - Input the concatenated feature matrix into the model with preprocessed reference text as the target output for training.
- **Model testing**
- **Evaluation:**
  - Evaluation metrics

References

1. Kumar, A., & Jha, S. (2023). *A novel approach to medical image captioning using multimodal learning*. *Scientific Reports, 13*, 31223. https://doi.org/10.1038/s41598-023-31223-5
2. *Zhu, Z., & Liu, J.* (2024). *A Unified Framework for Medical Image Captioning Using Vision-Language Models.* arXiv preprint arXiv:2406.00391. Retrieved from https://arxiv.org/pdf/2406.00391
3. Xie, Y. (2024). MedTrinity-25M: A Large-Scale Multimodal Dataset for Medical Image Captioning and Report Generation. Retrieved from https://yunfeixie233.github.io/MedTrinity-25M/
4. Hugging Face. (n.d.). Multimodal Models: Introduction. Retrieved from https://huggingface.co/learn/computer-vision-course/en/unit4/multimodal-models/pre-intro
5. Wang, Y., & Li, X. (2024). Towards a Holistic Framework for Multimodal Large Language Models in Three-dimensional Brain CT Report Generation. arXiv preprint arXiv:2407.02235v1. Retrieved from https://arxiv.org/pdf/2407.02235v1
6. Reale-Nosei, G., Amador-Domínguez, E., & Serrano, E. (2024). From vision to text: A comprehensive review of natural image captioning in medical diagnosis and radiology report generation. Journal of Artificial Intelligence in Medicine, 35(2), 58-72.
7. Ji J, Hou Y, Chen X, Pan Y, Xiang Y Vision-Language Model for Generating Textual Descriptions From Clinical Images: Model Development and Validation Study JMIR Form Res 2024;8:e32690.

References

8. Hoque, M., Hasan, M.R., Emon, M.I.S., Khalifa, F. and Rahman, M.M., 2024. Medical image interpretation with large multimodal models. In CLEF2024 Working Notes, CEUR Workshop Proceedings, CEUR-WS. org, Grenoble, France.
9. Yang, B., Raza, A., Zou, Y. and Zhang, T., 2023. Customizing general-purpose foundation models for medical report generation. arXiv preprint arXiv:2306.05642.
10. 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.
11. Beddiar, DR., Oussalah, M. & Seppänen, T. Automatic captioning for medical imaging (MIC): a rapid review of literature. Artif Intell Rev 56, 4019–4076 (2023). https://doi.org/10.1007/s10462-022-10270-w.

Thank you