MMG-CLIP: Automated Mammography Reporting through Image-to-Text Translation

Abdelrahman Habib [1,2,3] Santiago Pires [4] Jaap Kroes [4]

[1]Universitat de Girona, Spain [2]University of Bourgogne, France [3]Università degli studi di Cassino e del Lazio Meridionale, Italy [4]ScreenPoint Medical, Netherlands

With the support of the Erasmus+ Programme of the European Union

Introduction and Objectives

Recently medical image-text datasets have gained growing utilization in the development of deep learning applications, including automated radiology report generation models. In this work, we tackle the task of automated mammography report generation following Breast Imaging Reporting & Data System (BI-RADS) guidelines. We utilize an image-label and exam reports datasets, along with text prompting techniques, to generate a well-structured text report that supports training.

Methodology

- Motivated by *Contrastive Language-Image Pre-training (CLIP)* [1], *MMG-CLIP* jointly trains an image and text encoders to maximize the cosine similarity of real image-text pairs in each batch, and minimize the cosine similarity of the incorrect pairs.
- Network performance was evaluated on several downstream classification tasks based on BI-RADS guidelines (i.e. malignancy, mass shapes, mass margins).
- During evaluation, label specific prompts are constructed as text input for each BI-RADS classification task, and a full exam (4 views) or a single view is used as a visual input.

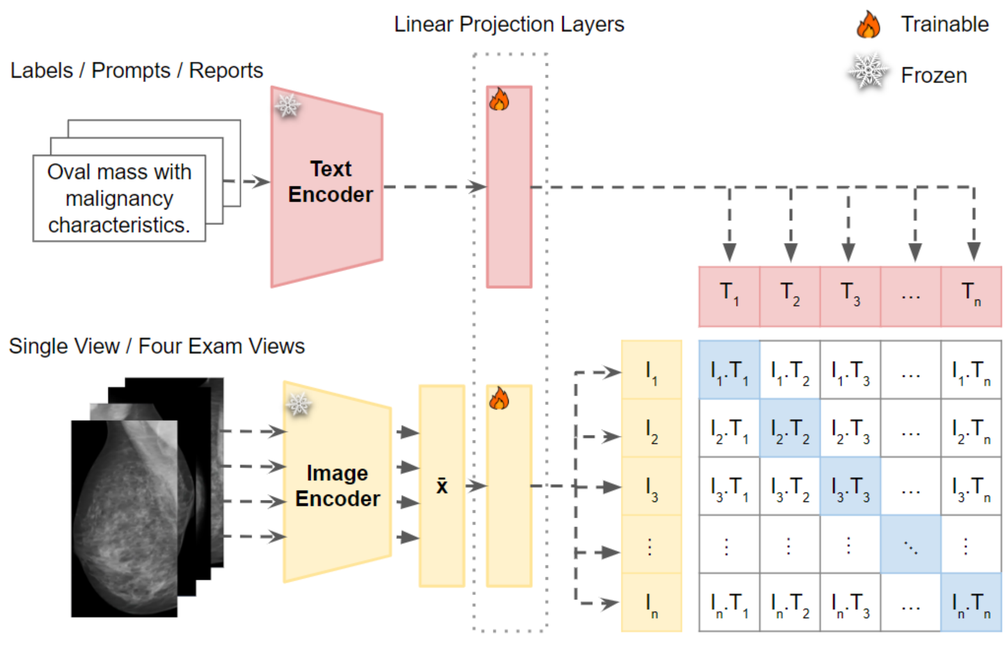

Figure 1. MMG-CLIP training architecture.

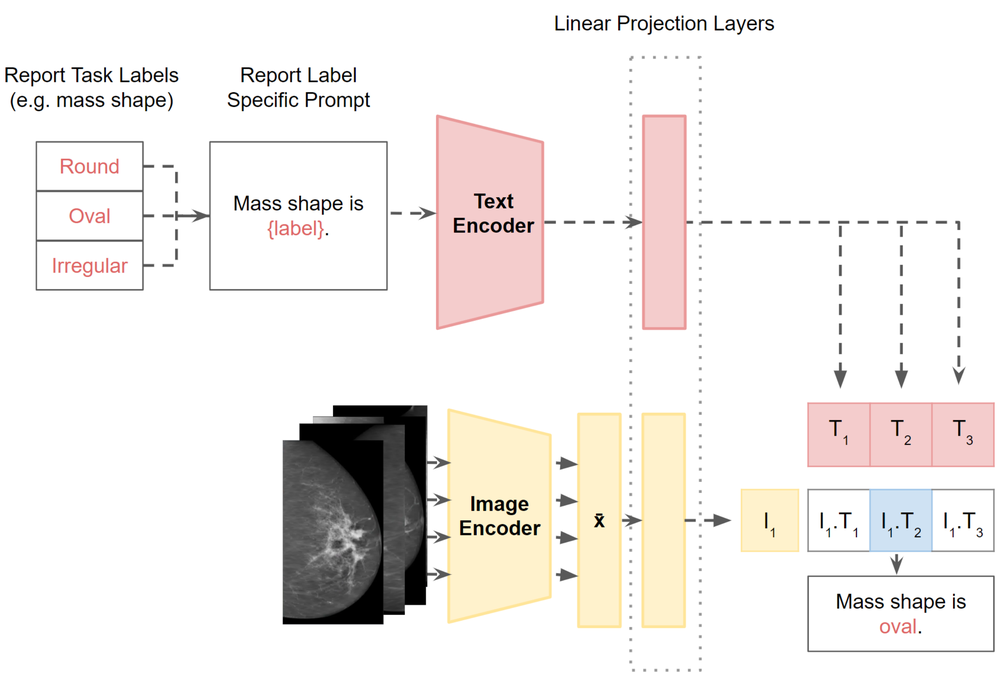

Figure 2. Report task zero-shot prediction.

Experiments and Results

Several experiments were conducted using different datasets, and evaluated using zero-shot classification settings as described in Table 1. All experiments results are summarized in Figure 3.

Experiment Name	Description
Image-Label	Training with images and labels (for example “benign” or “malignant” labels).
Image-Prompts	Training with images and variations of generated text reports based on prompt templates (generated from annotations).
Exam-Reports	Training with all images from an exam and radiology reports (clinical report as provided by source).
Exam-Reports + Prompts	Training with all images from an exam, radiology reports and generated text based on annotations.
Exam-Prompts	Training with all images from an exam and generated text based on annotations.

Table 1. Experiments description and the datasets used in each of them.

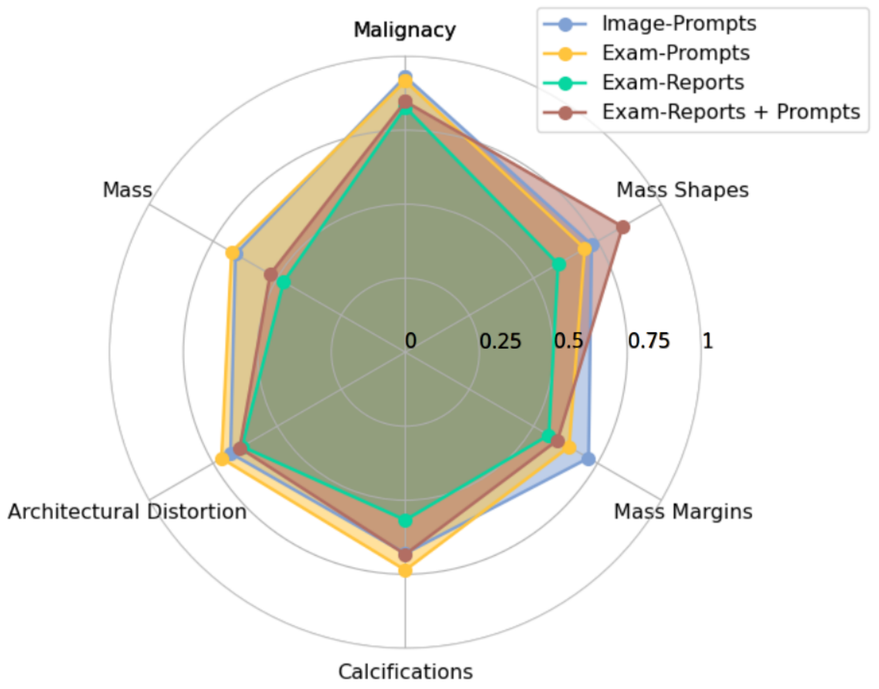

Figure 3. Experiments performance comparison using Area Under ROC (AUROC) on multiple downstream tasks.

The report generation is performed using the best model of exam-prompts experiment, where reports such as those in Figure 5 are generated as a series of zero-shot classification tasks as demonestrated in the report generation pipeline in Figure 4.

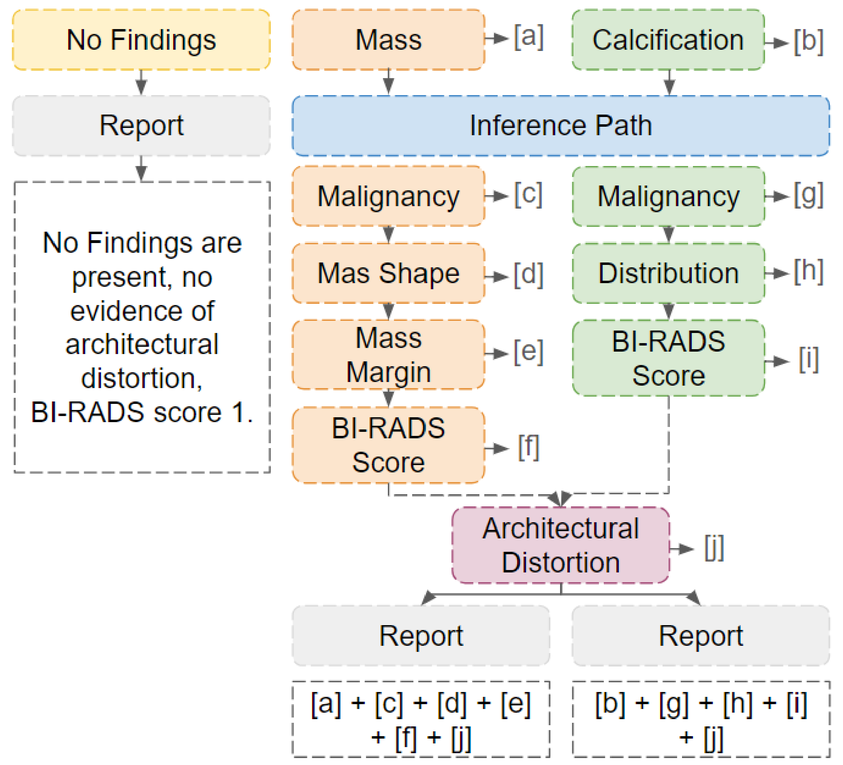

Figure 4. Report generation pipeline. Symbol *[letter]* represent the inference task output, and + represent output formatting and concatenation.

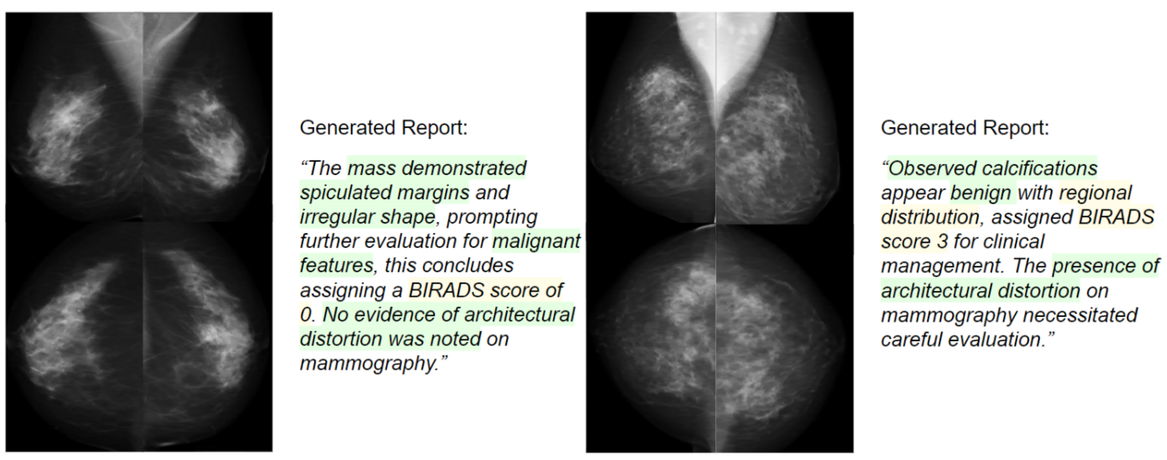

Figure 5. Exam level generated report revealing a malignant mass on the left figure and calcifications on the right figure. Green highlighted text represent correct predictions, while yellow represents unknown labels.

Conclusion

In this work, we proposed an image-text contrastive learning framework named *MMG-CLIP* as well as a report generation BI-RADS specific pipeline for mammography X-Ray 2D images. Our experiments results shows the network zero-shot capability of the learned representations for various downstream classification tasks.

References

[1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

Access the thesis by scanning the QR code: