Explainable DL models in lung cancer using CLIP

Abstract Proposal:

Lung cancer is one of the leading causes of mortality worldwide, and its early detection is crucial for improving survival rates. With advancements in artificial intelligence and deep learning (DL), models that have been developed are able to identify patterns in medical images with increasing performance. However, the lack of explainability in these models makes their adoption difficult in clinical settings, as healthcare professionals need to understand the reasoning behind the model's decisions.

This study focuses on the use of explainable deep learning models for lung cancer detection, taking advantage of the CLIP**[6][7]** (Contrastive Language-Image Pretraining) model. Developed by OpenAI, CLIP is a multimodal model that combines images and text to generate shared representations, enabling better interpretation of model decisions and providing more intuitive explanations. This research explores how this technique can enhance the interpretability of AI-assisted diagnostic systems, facilitating their integration into medical practice.

The expected outcomes include a deeper understanding of the key features identified by the model in lung cancer images and an analysis of its effectiveness compared to other deep learning approaches. This study aims to contribute to the development of more transparent and reliable systems for automated medical diagnosis.

Introduction

Lung cancer detection using deep learning (DL) has shown remarkable advancements in recent years, leveraging powerful models that extract meaningful patterns from medical images. However, one of the major challenges in medical AI applications remains the lack of interpretability, as deep learning models often operate as *black boxes*, making it difficult for clinicians to trust and validate their decisions. To address this, explainable deep learning (XDL) approaches are essential to bridge the gap between AI-generated predictions and medical reasoning.

This project focuses on leveraging CLIP to enhance explainability in lung cancer detection models. CLIP is a multimodal model capable of learning a shared latent space where images and textual descriptions are aligned**[8]**. While traditional deep learning models rely solely on image-based features, CLIP introduces a language-guided approach, enabling the association of lung cancer imaging data with medical reports. This allows for a more intuitive and interpretable decision-making process, aligning with the need for transparency in medical AI applications.

To achieve this, the project builds upon an existing lung cancer detection model that has already been trained to extract the most relevant visual features from radiological images. The primary goal is to extend this by effectively extracting and structuring features from medical reports, ensuring that textual and visual embeddings are properly aligned within a shared latent space. This is crucial for

making CLIP fully functional in this specific medical application, as it ensures that the model learns the correct relationships between medical images and their corresponding diagnostic descriptions.

While CLIP is the main technique explored, alternative approaches like BLIP**[10]** will also be analyzed from a theoretical perspective. This comparative analysis will provide insights into why CLIP stands out as the most suitable choice for this task.

State of the Art (SOTA)

Medical Image Captioning and Report Generation

Nowadays, the automatic generation of medical reports from radiological images has emerged as a crucial area in medical AI **[1][2]**, with the purpose to bridge the gap between computer vision and natural language processing (NLP). The main goal is to translate visual medical data into accurate textual descriptions, enhancing diagnostic efficiency and reducing the workload of radiologists and clinicians. Retrieval-based and template-based approaches have been traditionally used but, in the last years, gradually replaced by deep learning (DL) - based generative models**[5][9]**, which allow for more flexible context-aware, and detailed medical report generation.

Thus, delving into deep learning-based methods for medical image captioning, modern approaches employ deep learning architectures to automate diagnostic report generation. These models generally adhere to an encoder-decoder paradigm**[9]**, categorized into three main types:

- CNN-Based Encoders: these use Convolutional Neural Networks (CNNs) like ResNet, EfficientNet, and Vision Transformers (ViTs) to extract meaningful visual features from medical images.
- RNN-Based Decoders: recurrent models such as RNNs, LSTMs, and Transformer-based architectures process these visual features to generate coherent textual descriptions.
- Attention Mechanisms: these enhance caption relevance by directing focus to the most critical image regions during text generation**[3].**

Despite these advancements, when deploying deep learning in medical AI, two critical aspects must be considered:

1. Interpretability: refers to how easily a human can understand the model's logic behind its predictions. A highly interpretable model, such as linear regression, allows direct cause-effect relationships to be identified.

2. Explainability: refers to techniques that provide insights into a model’s decision-making process, even if the model itself remains a "black box." Explainability tools like Grad-CAM, SHAP, LIME, and attention maps help highlight which image regions or text features influenced a model’s prediction.

As mentioned, explainability is crucial actually because deep learning models must justify their decisions to gain tust from healthcare professionals. This is where CLIP and similar multimodal approaches play a vital role.

CLIP: A Game Changer in Medical AI

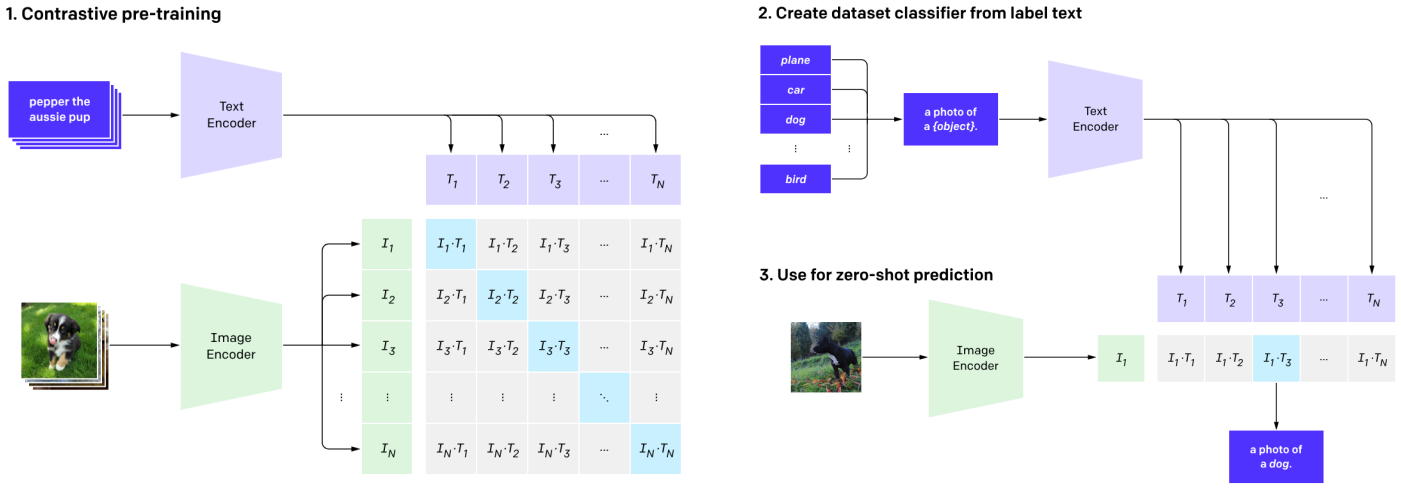
CLIP, developed by OpenAI, is a vision-language model designed to understand both images and text by jointly training on image-text pairs. CLIP is pre-trained on vast amounts of image-text pairs from the internet using contrastive learning**[3]**.

The way it works can be simplified to 3 steps:

1. Feature Extraction: CLIP encodes images and text separately using a VIsion Transformer (ViT) for images and a Transformer-based text encoder.

2. Latent Space Alignment: Both encoders project their outputs into a shared embedding space, where semantically related images and texts are pulled closer together**[4]**.

3. Zero-Shot Learning: Unlike traditional models that require fine-tuning on domain-specific datasets, CLIP is capable to classify images and generate text descriptions without retraining, making it highly adaptable to medical imaging tasks such as lung cancer detection, hematology and pathology (PLIP and BiomedCLIP), radiology report generation.



Objectives

The primary objective of this project is to enhance the explainability and interpretability of deep learning models for lung cancer detection by integrating medical image and diagnostic text embeddings into a shared latent space using CLIP-based architectures. This approach aims to create a more transparent and clinically relevant AI system that effectively aligns radiological images with textual medical diagnoses through contrastive learning.

A critical aspect of the project is modifying or adapting an existing CLIP architecture to integrate both the image encoder and the text encoder into a joint embedding space. Through contrastive learning, the model will be trained to maximize the similarity between images and their corresponding medical reports while increasing the distance between unrelated image-text pairs. To optimize this process, the contrastive loss function will be fine-tuned, ensuring that the alignment between visual and textual modalities is precise and meaningful in a medical context.

By successfully implementing this multimodal learning approach, the project aims to develop a model capable of generating a more explainable and clinically relevant decision-making process. Integrating text and image features will provide a richer representation of lung cancer diagnosis, enabling AI-driven predictions that align more closely with medical expert reasoning.

Methodology

First phase of the project includes cleaning, tokenization, and transforming text data into structured embeddings that can be aligned with precomputed image features. Next, the model architecture is constructed and adapted to the task. A transformer-based text encoder is designed to extract embeddings from diagnostic reports, which are then projected into a shared latent space alongside the precomputed image embeddings. In the training and experimental setup phase, the model undergoes fine-tuning and optimization. Contrastive learning is applied to strengthen associations between medical images and corresponding diagnoses.

Finally, the evaluation and result analysis phase assesses the model's effectiveness using image-to-text and text-to-image retrieval accuracy, contrastive loss evaluation, and explainability assessments**[11]**.

High-Level Overview of Model Architecture Proposal

Explanation:

- Class TextEncoder (with a transformer-based language model):
    - Input: batch of raw diagnostic report
    - Tokenization
    - Transformer: tokens into Bert (explore more) and extract context-aware embeddings (usage of CLS token representation as summary of the entire batch)
    - Projection: into latent space by applying a fc layer to map transformer output to a 512-dimensional latent space
    - Output: (batch_size, 512) tensor, where each row is the embedding of a medical report

- Clip Class (value if a pre-trained CLIP is profitable) :
    - Input: image features (batch_size,1024) tensor, texts (batch of diagnostic reports)
    - Extract text features: use textEncoder (we get (batch_size,512) text embeddings)
    - Project img features: linear layer to map img features to 512 dimensions
    - Apply contrastive learning (temperature): scales similarity scores for stable training.
    - Output: (batch_size, batch_size) similarity matrix, each value represents how well an images matches text.

- Contrastive Loss:
    - Input: logits (similarity matrix)
    - Create ground truth labels: matching pairs (img, text) as correct
    - Applies Cross-Entropy Loss: force matching pairs to have high similarity and incorrect pairs to have low similarity.
    - Adds loss for both directions: both image-to-text and text-to-image learning.
    - Output: scalar loss used for model optimization

- Train model:

- Input: img_features, a batch of diagnostic reports, optimizer
- Compute similarity scores
- Compute contrastive loss
- Backpropagation
- Update model weights

References

[1] H. Sharma and D. Padha, "A Comprehensive Survey on Image Captioning: From Handcrafted to Deep Learning-Based Techniques, a Taxonomy and Open Research Issues," *Artificial Intelligence Review*, vol. 56, pp. 13619–13661, 2023, doi:10.1007/s10462-023-10488-2.

[2] I. Hartsock and G. Rasool, "Vision-Language Models for Medical Report Generation and Visual Question Answering: A Review," *arXiv preprint*, arXiv:2403.02469, 2024.

[3] A. Yan, Z. He, X. Lu, J. Du, E. Chang, A. Gentili, J. McAuley, and C.-N. Hsu, "Weakly Supervised Contrastive Learning for Chest X-Ray Report Generation," *arXiv preprint*, arXiv:2109.12242, 2021.

[4] S. Yang, X. Wu, S. Ge, Z. Zheng, S. K. Zhou, and L. Xiao, "Radiology Report Generation with a Learned Knowledge Base and Multi-Modal Alignment," *Medical Image Analysis*, vol. 86, p. 102798, 2023, doi:10.1016/j.media.2023.102798.

[5] Y. Yang, J. Yu, J. Zhang, W. Han, H. Jiang, and Q. Huang, "Joint Embedding of Deep Visual and Semantic Features for Medical Image Report Generation," *IEEE Transactions on Multimedia*, vol. 25, pp. 167–178, 2021, doi:10.1109/TMM.2021.3063241.

[6] OpenAI, "CLIP: Connecting Vision and Language Through Contrastive Learning," *OpenAI Blog*, available at https://openai.com/index/clip/, 2021.

[7] OpenAI, *OpenAI GitHub Repository*, available at https://github.com/openai, 2024.

[8] Y. Zhang, H. Jiang, Y. Miura, C. D. Manning, and C. P. Langlotz, "Contrastive Learning of Medical Visual Representations from Paired Images and Text," in *Machine Learning for Healthcare Conference*, PMLR, pp. 2–25, 2022.

[9] Z. M. Ziegler, L. Melas-Kyriazi, S. Gehrmann, and A. M. Rush, "Encoder-Agnostic Adaptation for Conditional Language Generation," *arXiv preprint*, arXiv:1908.06938, 2019.

[10] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping Language-Image Pre-Training for Unified Vision-Language Understanding and Generation," in *International Conference on Machine Learning*, PMLR, pp. 12888–12900, 2022.

[11] G. Torres, C. Sanchez, and D. Gil, "Learning Networks Hyper-Parameter Using Multi-Objective Optimization of Statistical Performance Metrics," in *2022 24th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*, IEEE, pp. 157–164, 2022, doi:10.1109/SYNASC57785.2022.00032.