Language Augmentation in CLIP for Improved Anatomy Detection on Multi-modal Medical Images

Mansi Kakkar*, Dattesh Shanbhag†, Chandan Aladahalli† and Gurunath Reddy M†

*Indian Institute of Technology Madras, †GE HealthCare

mansikakkar97@gmail.com, {Dattesh.Shanbhag, Chandan.Aladahalli, GurunathReddy.M}@gehealthcare.com

Abstract—**Vision-language models have emerged as a powerful tool for previously challenging multi-modal classification problem in the medical domain. This development has led to the exploration of automated image description generation for multi-modal clinical scans, particularly for radiology report generation. Existing research has focused on clinical descriptions for specific modalities or body regions, leaving a gap for a model providing entire-body multi-modal descriptions. In this paper, we address this gap by automating the generation of standardized body station(s) and list of organ(s) across the whole body in multi-modal MR and CT radiological images. Leveraging the versatility of the Contrastive Language-Image Pre-training (CLIP), we refine and augment the existing approach through multiple experiments, including baseline model fine-tuning, adding station(s) as a superset for better correlation between organs, along with image and language augmentations. Our proposed approach demonstrates 47.6% performance improvement over baseline PubMedCLIP.**

I. INTRODUCTION

Over the last decade, deep learning models from CNNs to Vision Transformers (ViT) [1] have gained prominence in aiding clinicians in performing their medical imaging studies efficiently through consistent image acquisition, reconstruction and AI-assisted reporting. Recent advances in Large Language Models (LLMs) have led to the integration of vision and text encoders, giving rise to Vision-language Models (VLMs), incorporating semantic descriptions into medical image analysis, and correlating textual information with image features. State-of-the-art models such as CLIP [2], ALIGN [3], BASIC [4], and LiT [5] have shown remarkable results in cross-modal search using zero-shot classification and image-to-text and text-to-image applications over the multi-modal datasets [6], [7]. In the medical context, where images often require descriptive textual interpretation, VLMs become crucial for relating visual features to clinical findings, supporting prognosis and diagnosis by leveraging multi-modal data.

This work explores VLMs to describe organs and anatomical regions in multi-modal radiology images. This is crucial to describe the anatomical context for image acquisition and reporting. Given challenges in semantically segmenting diverse anatomies across each image, a vision encoder-only approach may not be ideal due to wide patient pose, orientation, and coverage variations. Text-based descriptors for organs and regions is much more attractive due to simplified labelling and the potential for capturing detailed hierarchical information. We utilize CLIP for its proven effectiveness in addressing multi-modal challenges, specifically within the realm of medical

data [8]. For clinical applications, CLIP has been refined to PubMedCLIP using multi-modal medical image-text pairs, demonstrating organ-specific vision and language embeddings [9]. Since PubMedCLIP has been trained over images obtained from publications, it typically fails to provide good results on pristine medical image data.

In this work, we evaluated a methodology to fine-tune PubMedCLIP model over clinical imaging data for labelling organs and anatomical regions (stations). Further, this paper demonstrates the importance of text prompts [2], [10] drawing inspiration from LLMs, where manipulating the textual modality is a common practice. We use VLMs inherent data manipulation capability to our advantage for multi-modal classification. Our key contributions include, (1) analysing PubMedCLIP model over multi-modal, multi-label classification on pristine clinical single slice datasets for anatomy description; (2) fine-tuning PubMedCLIP on approximately 4000 clinical scans to achieve enhanced performance for multi-label anatomy detection; (3) showcasing enhanced model performance using data augmentations for both images and text phrases.

II. METHODS

The proposed multi-modal anatomy detection framework is shown in Fig. 1.

A. Base model

In this study, we have used PubMedCLIP as our base model which is CLIP fine-tuned over ROCO (Radiology Objects in Context) dataset [11]. ROCO dataset comprises approximately 82K radiology images and captions from diverse modalities, typically image-caption pairs from PubMed. ROCO dataset has provided promising outcomes in clinical use cases [6], [11]. Our experiments have demonstrated the ineffectiveness of PubMedCLIP describing organs, possibly due to the presence of various artefacts like figures, portraits, digital arts, and illustrations in the ROCO dataset, unrelated to clinical medical scans. Moreover, MR and CT scans in ROCO dataset have highlighted or marked artefacts, and an imbalance towards some organs. Due to these limitations, we further fine-tune PubMedCLIP with additional clean multi-modal clinical images and labels.

B. Data description

We fine-tuned the base model for our study using data from various sources, including in-house and clinical open-source
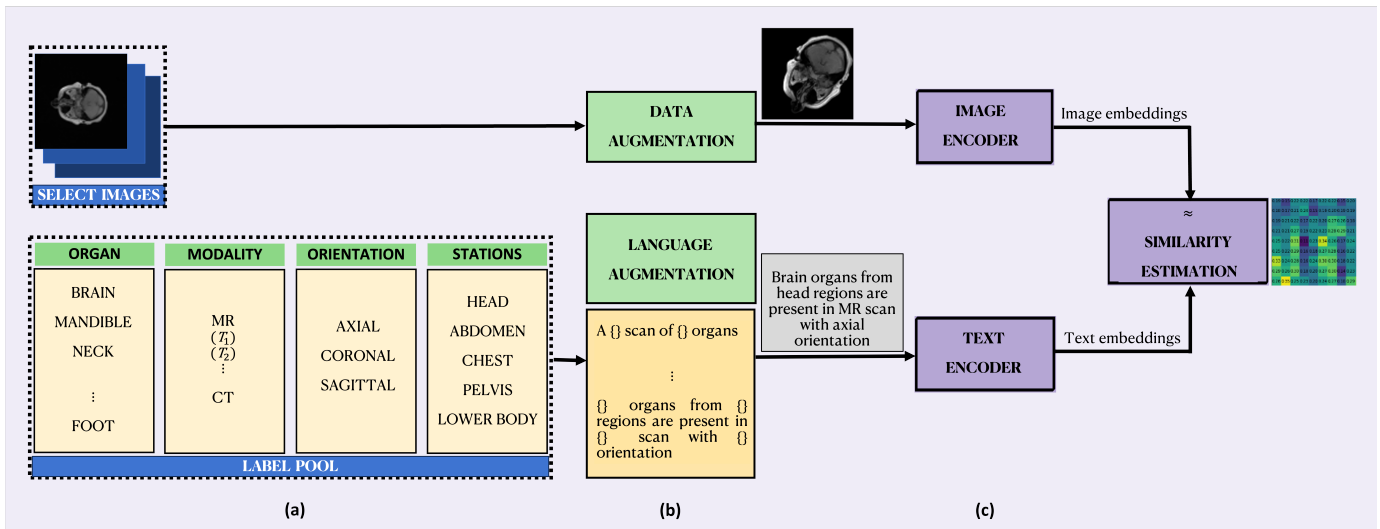
Fig. 1. Pipeline of our approach for anatomy classification. (a) Dataset creation – multi-modal anatomy dataset creation with label pools for detailed caption of images, (b) pre-processing – image augmentation and language augmentation, involving set of labels from label pool passing through a manual prompt phrasing system providing 10 different sets of prompts for each image, and (c) baseline model – PubMedCLIP model with ViT-B/32 as vision encoder and text tokenizer as text encoder will be fine-tuned over these images to give us our proposed model for anatomy detection.

datasets (TotalSegmentator [12]). The data includes modalities (MR and CT), orientations (axial, coronal, and sagittal), and covers different organs. MR includes data across protocols: T1, T2, FLAIR, DWI, ADC and STIR. The images were taken from five different stations of the human body: head, chest, abdomen, pelvis, and lower body. The image labels have been distributed over 20 organ labels–brain, mandible, neck, shoulder, humerus, elbow, forearm, wrist, hand, lungs, heart, liver, kidneys, intestine, pelvic bone, thigh, knee, leg, ankle, and foot. We included additional labels, such as modality (including protocols), orientation, and stations/regions along with the organs, to generate text captions similar to the ROCO dataset. The overall training dataset contains single-slice images, captions, and organ labels. The total training dataset size was 4994 images, comprising 3995 for model training and 999 validation images. We used two datasets to test the approach: Set #1 comprising 262 MR and CT in-house images and Set #2 comprising of open-source visible human project [13] with 650 CT images.

C. Augmentation and pre-processing of images

To mimic the data variety encountered in clinical practices, we include data from different modalities and apply augmentations. We have used these augmentations: 1) histogram manipulations (CLAHE, contrast enhancement using PIL library [14] and gamma correction), 2) rotation (-180 to $180°$) and 3) translation (-100 to 100 pixels). Using combinations of these augmentations, we generate 10 different images for each image. CLIP model itself performs some data augmentations involving random resizing, random noise, gaussian blur, colour jittering, horizontal and vertical flips and arbitrary rotation. Our augmentations yielded better results than baseline methods, further discussed in the results section.

D. Language augmentation for training

Augmentation on text prompt phrases and fine-tuning make the model task-specific, generalized for different modalities, and more robust to distribution shift. A complete prompt will describe modality, orientation, station and anatomy. Complete prompt diversity is obtained by shuffling the entities. To enhance model's robustness, we augmented text prompts with incomplete information, such as missing station or organ details. See Table I for examples.

E. Fine tuning

We experimented with fine-tuning PubMedCLIP, with encoders from two architectures: ResNet50 [15] and ViT–B/32 [1]. We proceeded further with ViT–B/32 since it provided comparatively better performance. We retained the baseline text tokenizer, and trimmed longer captions while zero padding the shorter ones, according to CLIP's maximum acceptable text length of 76. Moreover, we used the baseline loss function,

$$L = \lambda H(\hat{y}_{vision}, Y) + (1 - \lambda)H(\hat{y}_{text}, Y), \qquad (1)$$

where H is the cross entropy loss, Y is the set of labels, and $\lambda = 0.5$ for equal weightage to vision and text losses. Further,

TABLE I
EXAMPLES OF PROMPT TEXTS GIVEN AS CAPTIONS FOR IMAGES FOR
TRAINING AND VALIDATION DATASETS

Prompts	
1.	A {*orientation*} oriented {*modality*} image of {*organ*} organs belong to {*station*} region
	E.g., A sagittal oriented MR T2 image of knee organs belong to lower body region
2.	An image of {*orientation*} {*modality*} scan consisting of {*organ*} organs
	E.g., An image of axial CT scan consisting of liver, intestine organs

Experiments	Augmentations
Exp 1. (PMC) Tested baseline model (PubMedCLIP with ViT-B/32 vision encoder).	No
Exp 2. (PMC-M) Fine-tuned baseline model over multi-modal anatomy dataset of 3995 training images and corresponding captions involving organ(s), modality, and orientation labels (see Section II-B).	No
Exp 3. (PMC-MS) Revised captions for better region segregation by including station labels. For eg., labels like brain and knee do not get paired in an axial scan.	No
Exp 4. (PMC-MSA) Performed augmentations over training images and captions as described in Section II-C, and Section II-D, respectively.	Yes

we explored multiple training approaches including expanding the input training data. The resulting performance is discussed in the experiments section.

F. Zero–shot prediction

After fine-tuning the model on our multi-modal anatomy dataset, we set up a zero-shot classifier, independently for 5 station labels and 20 organ labels. Logits are computed between fixed dimensional text and image feature vectors, obtained by sending text prompt to the text encoder and image to the image encoder, respectively, as shown in Fig. 1. Zero-shot classification is accomplished by using

$$logits_m = [I_m \cdot T_1, \ldots, I_m \cdot T_k, \ldots, I_m \cdot T_{20}], \quad (2)$$

$$\hat{y}_m^1 = \arg\max \text{softmax}(logits_m), \quad (3)$$

to get the top prediction, where I_m is individual image embedding, $T_1, \ldots, T_{20}$ are text embedding for 20 labels prompts, and $\hat{y}_m^1$ is the top prediction. A similar computation is done for stations. To ensure computational accuracy, we verify if the predicted label exists within the set of target labels by

$$\text{accuracy} = \frac{\sum_{i=1}^{N} \# \text{ correct matches for image } i}{N}, \quad (4)$$

where N is the size of the total test dataset.

III. EXPERIMENTS

A. Annotation and Implementation

We annotated the datasets in-house, consisting of labels for organs, modality, orientation and stations. The annotations were reviewed with a trained radiologist and a clinician. The training was performed on NVIDIA DGX A100 Tensor Core GPU. We conducted four experiments using different configurations: (1) PubMedCLIP (PMC), (2) PMC fine-tuned over a multi-modal anatomy dataset (PMC-M), (3) PMC-M with additional stations (PMC-MS), and (4) PMC-MS with image and text augmentations (PMC-MSA). The details are provided in Table II.

B. Evaluation

All experiments are assessed through zero-shot classification, analyzing the top accuracy. Further, we visualize the performance of models PMC, PMC-M, and PMC-MSA, through a One-vs-the-Rest (OvR) AUC-ROC curve using *Scikit-Learn*.

IV. RESULTS AND ANALYSIS

The results for all experiments are given in Table III. Our proposed approach gives an overall average enhancement of 47.6% for organ detection, and 27% for station detection in comparison to the baseline model. After fine-tuning the baseline model over multi-modal anatomy dataset, the organ prediction performance is improved in PMC-M, but gave a poor performance for station prediction due to its reliance on the ROCO dataset for station information. In PMC-MS, providing detailed station information improved station prediction, but organ prediction accuracy dropped, primarily attributed to confusion between station and organ (no strong correlation observed). Further, in PMC-MSA, this confusion is reduced by data augmentation and text prompt diversity, enhancing both organ and station prediction performance. Fig. 2 shows the OvR AUC-ROC curve for Set #2 for different organs across the models. Our proposed model performs well for all organs, but due to data imbalance, it gives comparatively low performance for humerus. Furthermore, a comparison among PMC, PMC-MS, and PMC-MSA is performed and the result is shown in Fig. 3. Given two sets of prompts– one with correct organ-station correlation and the other with a contradiction, our approach outperforms the baseline model. Notably, for the correct and the false prompts–our approach predicts 90.7% and 9.3% scores, respectively, compared to the baseline scores of 53.1% and 46.9%.

V. CONCLUSION

Our research demonstrates the effectiveness of utilizing station data to establish correlations between organs, resulting in a notable improvement in accuracy. Along with it, the integration of image and text prompt augmentations significantly improves model performance. The integration of CLIP, coupled with strategic data augmentations, has notably enhanced

TABLE III
ZERO-SHOT ACCURACY COMPARISON AMONG ALL EXPERIMENTS. SET #1–IN-HOUSE MR AND CT TEST DATASET AND SET #2–OPEN SOURCE CT DATASET

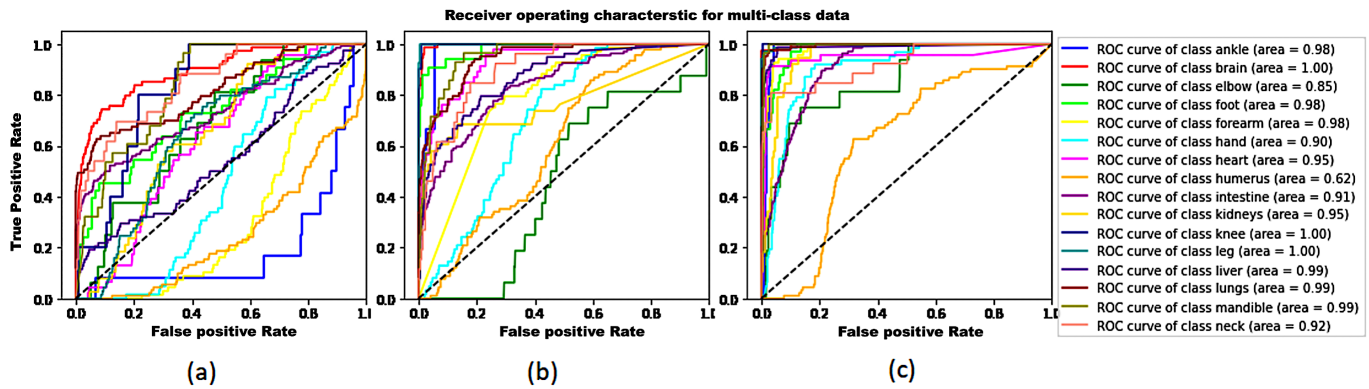| Label Type	Dataset	PMC [Baseline]	PMC-M	PMC-MS	PMC-MSA [proposed approach]
Organ labels	Set #1 (MR)	54%	62%	57%	**91.62%**
	Set #1 (CT)	33%	70.8%	63%	**81.25%**
	Set #2 (CT)	24%	58.5%	47%	**81%**
Station labels	Set #1 (MR+CT)	52%	49%	65%	**75%**
	Set #2 (CT)	45%	41%	60%	**76%**

Fig. 2. AUC-ROC curves for test dataset (visible human project) across different models: (a) result for PMC (baseline), (b) result for PMC-M (fine-tuning over mutli-modal anatomy dataset), and (c) result for PMC-MSA (model with text and image data augmentations). The AUC values are for proposed approach. We receive good AUC values for all organs except humerus (AUC = 0.62), probably due to imbalance towards other limbs as compared to humerus
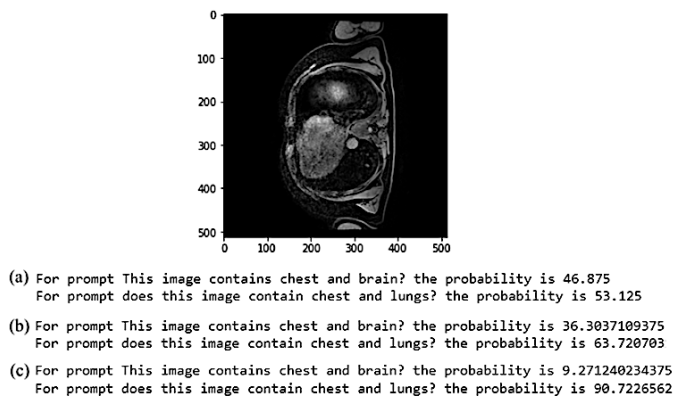

Fig. 3. Examples of performance for different models: (a) result for PMC (baseline), (b) result for PMC-MS, and (c) result for PMC-MSA (proposed approach). Showcasing our approach outperforming the baseline

accuracy and reliability in multi-modal organ-related studies, making a step forward in the realm of medical imaging.

We have made efforts to address the data imbalance in the ROCO dataset. Still, there is room for improvement, particularly when assessing the limbs' performance compared to other organs. As a future step, we could learn class-specific text tokens to eliminate manual text augmentations to further improve the model accuracy.

ACKNOWLEDGMENT

REFERENCES

[1] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *International Conference on Learning Representations*, 2021.

[2] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, et al., "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*, PMLR, pp. 8748–8763, 2021.

[3] C. Jia, Y. Yang, Y. Xia, Yi-Ting Chen, Z. Parekh, H. Pham, et al., "Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision," in *Proceedings of the 38th International Conference on Machine Learning, PMLR*, vol. abs/2102.05918, pp. 4904-4916., 2021.

[4] H. Pham, Z. Dai, G. Ghiasi, K. Kawaguchi, H. Liu, A. W. Yu, et al. "Combined scaling for zero-shot transfer learning," *Neurocomputing*, Volume 555, 2023.

[5] X. Zhai, X. Wang, B. Mustafa, A. Steiner, D. Keysers, A. Kolesnikov, et al., "LiT: Zero-Shot Transfer with Locked-image Text Tuning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18123-18133, 2022.

[6] A. G. Barreto, J. M. de Oliveira, F. N. B. Gois, P. C. Cortez, and V. H. C. de Albuquerque, "A New Generative Model for Textual Descriptions of Medical Images Using Transformers Enhanced with Convolutional Neural Networks," *Bioengineering* 10, no. 9: 1098, 2023.

[7] M. Wortsman, G. Ilharco, J. W. Kim, M. Li, S. Kornblith, R. Roelofs, et al., "Robust fine-tuning of zero-shot models," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7949-7961, 2022.

[8] Z. Zhao, Y. Liu, H. Wu, Y. Li, S. Wang, L.Teng, et al. "CLIP in Medical Imaging: A Comprehensive Survey," *arXiv preprint arXiv:2312.07353*, 2023.

[9] S. Eslami, C. Meinel, and G. de Melo, "PubMedCLIP: How Much Does CLIP Benefit Visual Question Answering in the Medical Domain?" in *Findings of the Association for Computational Linguistics :EACL 2023*, pp. 1181-1193, 2023.

[10] J. Gu, Z. Han, S. Chen, A. Beirami, B. He, G. Zhang, et al., "A systematic survey of prompt engineering on vision-language foundation models," *arXiv preprint arXiv:2307.12980*, 2023.

[11] O. Pelka, S. Koitka, J. Rückert, F. Nensa, and C. M. Friedrich, "Radiology Objects in COntext (ROCO): A Multimodal Image Dataset," in *MICCAI Workshop on Large-scale Annotation of Biomedical Data and Expert Label Synthesis (LABELS)*, Lecture Notes on Computer Science, vol. 11043, pp. 180-189, Springer Cham, 2018.

[12] J. Wasserthal, H. Breit, M. T. Meyer, M. Pradella, D. Hinck, A. E. Sauter, et al., "TotalSegmentator: Robust Segmentation of 104 Anatomic Structures in CT Images," *Radiology: Artificial Intelligence*, vol. 5, no. 5, 2023.

[13] M. J. Ackerman, "The Visible Human Project," *Information services & use* vol. 42(1): 129-136, 2022.

[14] Pillow (PIL Fork), "Pillow Documentation (Version 8.4.0)," 2021. Retrieved from https://pillow.readthedocs.io/en/stable/ (accessed on Sep. 3, 2023).

[15] K He, X. Zhang, S. Ren, and J.Sun, "Deep Residual Learning for Image Recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770-778, 2015.