Cross-Modal Learning with CLIP for Robust Anomaly Detection in Medical Images

Swetha Krishnan
swethakrishn@umass.edu

Prakriti Shetty
psshetty@umass.edu

Debarshiya Chandra
dchandra@umass.edu

Abstract

Anomaly detection plays a crucial role across various fields, especially in domains requiring fine-grained visual distinctions, such as industrial quality control and medical imaging. Adapting vision-language models (VLMs) like CLIP[7] to the medical domain, however, presents unique challenges due to the high variability and subtlety of anomalies, alongside a scarcity of labeled anomalous examples for training. Leveraging CLIP's zero-shot capability, this study examines its direct applicability to medical anomaly detection with minimal architectural modifications. We first evaluate the baseline performance of CLIP and then propose to explore cross-modal enhancements to improve its classification accuracy in medical contexts. Our findings provide insights into the limitations and possibilities for advancing anomaly detection through VLMs, particularly in complex, domain-adapted tasks where traditional supervised approaches may falter.

1. Introduction

Anomaly detection remains a cornerstone in various critical domains, including industrial quality control and medical diagnostics, where the identification of atypical samples is paramount. However, in medical imaging, the task of identifying anomalous samples becomes markedly complex. Anomalies in medical data are often subtle and fine-grained, demanding a level of precision that standard models struggle to achieve. Additionally, labeled anomalous samples are rarely available in sufficient quantity, underscoring the critical need for robust, domain-adaptable models capable of operating effectively under zero- and few-shot settings.

This complexity arises from the subtlety and variability of anomalies, coupled with the scarcity of labeled anomaly images for training. Current anomaly detection and segmentation approaches primarily rely on reconstruction-based techniques or Vision-Language Model (VLM) methods. While VLMs like CLIP[7] have demonstrated strong zero-shot performance on global image classification tasks, they struggle with fine-grained tasks such as anomaly detection. In the industrial domain, zero-shot anomaly detection and segmentation becomes a multi-class approach, where a significant number of seen classes (e.g. nail, PCB) and its respective pixel-level anomaly segmentation masks are available for training, and are evaluated on unseen classes (e.g. capsule, wires). However, in the case of medical images, zero-shot anomaly detection and segmentation becomes a single-class approach, where only "normal" images are available for training, and the unseen class is the anomaly. Thus, addressing challenges in the zero-shot setting for the medical domain is still an important area of research [2], i.e., the subtle differences of anomalies present within medical domains make it a critical problem to tackle.

In response to these challenges, our study explores the potential of CLIP's zero-shot capabilities for anomaly detection and segmentation in medical imaging, focusing on a minimally modified, cross-modal adaptation. We present an evaluation of CLIP's baseline performance on this task and propose enhancements that leverage cross-modal learning to better capture fine-grained anomalies in the medical domain. Through this approach, we aim to advance the capabilities of VLMs in anomaly detection, pushing the boundaries of State-of-the-Art (SoTA) methods and addressing the pressing demands of medical applications requiring exceptional precision and adaptability.

2. Related work

Zero-Shot Anomaly Detection (ZSAD). The task of multi-class ZSAD has been largely explored with the MVTechAD[1] and VisA[11] datasets in the industrial domain. WinCLIP[5] is an excellent example of harnessing CLIP's abilities for ZSAD without the need for fine-tuning of any sort. The sliding-window approach, combined with multiple text prompts, allows CLIP's visual and textual relations to be well-utilized for both global and pixel-level detection.

Using the fine-tuning approach, methods have explored incorporating learnable text prompts ([3], [10]) or improving upon visual embeddings ([4]). With these methods, the model is fine-tuned on identifying anomalies in a multi-

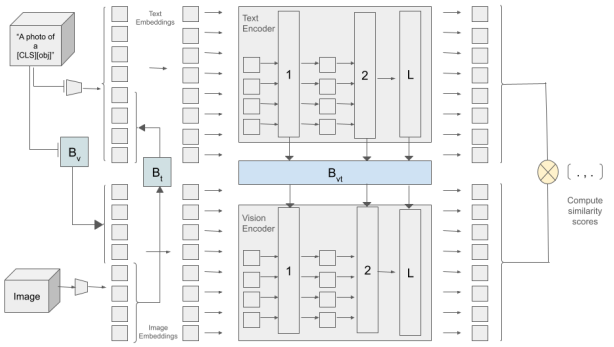

Figure 1. Proposed method, adopted from [8] for this task of medical ZSAD. Cross-modal interactions allow for dynamic learning of image and text representations.

class approach, and can be used to evaluate on medical images as an out-of-domain adaptation. However, due to the complexities of the anomalies present in medical images, such methods still fail to be robust across various medical imaging categories. In the medical domain, prompt engineering of synthesis of expert medical prompts [9] have also been explored. Other methods explore incorporating cross-modal learning in fine-tuning CLIP for downstream tasks.

Multi-modal learning. It is interesting to note that adapting representations in only one branch of CLIP may not provide the required dynamics to adjust both representation spaces for fine-grained downstream tasks. Certain methods ([6], [8]) explore modeling such cross-modal interactions between the intermediate visual and textual transformer layers, but for different tasks entirely. The main motivation is to harness information exchange at every step using CLIP for fine-grained downstream tasks. However, the potential of bi-directional interleaved learning has not yet been explored for medical ZSAD.

3. Method

Anomaly Classification (AC). Given an input photo $x \in X$, where $x \in \mathbb{R}^{\mathrm{H\,x\,W}}$, our objective is to identify the image as a normal image or an anomalous image $\mathbf{X} \rightarrow \{-, +\}$, where + indicates an anomalous image. Text prompts for normal/anomalous images are represented as S: `"a photo of a [CLS] [obj]"`. `[CLS]` denotes the normal/anomalous class, and `[obj]` denotes the particular object, ex. "Brain MRI scan". Despite the lack of anomalous (or positive) samples in practice ($\mathcal{D} := \{(x_i, -)\}_{i=1}^{K}$), the task is identified as a binary classification problem since it significantly outperforms one-class scenarios, as tested in [5].

Let the ViT-based image encoder of CLIP be denoted by $\mathcal{F}(.)$ and transformer-based text encoder by $\mathcal{G}(.)$, both with $\mathcal{L}$ encoder layers. The output visual and text prompt embeddings would thus be represented as $\mathcal{F}(x)$ and $\mathcal{G}(s)$ respectively. The text encoder tokenizes the $\mathcal{J}$-word-long text inputs, and projects them into word embeddings $\mathbf{W}0 = [w^1 0, w^2 0, \cdots, w^{\mathcal{M}} 0] \in \mathbb{R}^{\mathcal{M} \times d_t}$, where $[.,.]$ denotes stacking and concatenation, $\mathcal{M}$ refers the number of embedding tokens and d_t is the dimension of text tokens. We obtain the l-th layer embedding of $\mathcal{F}$, denoted as $\mathcal{F}_l$, as follows,

$$[\mathbf{W}_l] = \mathcal{F}_l(\mathbf{W}_{l-1}) \in \mathbb{R}^{\mathcal{M} \times d_t} \qquad l = 1, 2, \cdots, \mathcal{L} \quad (1)$$

The image input x is partitioned into fixed-size patches and encoded through $\mathcal{G}$. Each patch undergoes projection to generate initial patch embeddings ($\mathbf{E}_0$), along with a learnable `cls` token $\mathbf{c}_0$, denoted as $[\mathbf{c}_0, \mathbf{E}_0] \in \mathbb{R}^{1+\mathcal{N} \times d_v}$, where $\mathcal{N}$ denotes the number of patches and d_v is the dimension of patch tokens. Henceforth, the output embedded tokens of the l-th layer of $\mathcal{G}$ can be expressed as

$$[\mathbf{c}_l, \mathbf{E}_l] = \mathcal{G}^l([\mathbf{c}_{l-1}, \mathbf{E}_{l-1}]) \in \mathbb{R}^{1+\mathcal{N} \times d_v} \quad (2)$$

Bi-directional cross-modal feature association. Taking inspiration from [8]'s architecture, we propose to utilize a bi-directional scheme for feature sharing between our text and visual modalities. Specifically, there is a *visual-to-textual mapping* block $\mathcal{B}_t$ which generates m learnable prompt tokens, denoted by $\mathcal{T}_{1:m}$, and collectively as ($\mathbf{T}$), from the $\mathcal{N}$ visual patch embeddings $\mathbf{E}_0$, which operates across all layers of $\mathcal{F}$. Next, to incorporate textual information within the visual encoder of CLIP, the initial tokenized text input excluding the [CLS] token, denoted as $\mathcal{W}'$ comprising of $(\mathcal{J}-1)$ tokens are employed as *semantic domain knowledge* ($\mathbf{V}^{\mathtt{tg}}$), which is then mirrored by an equivalent stack of $(\mathcal{J}-1)$ learnable tokens ($\mathcal{V}^{\mathtt{tg}} 1 : \mathcal{J}-1$) via a *textual-to-visual mapping block* ($\mathcal{B}_v$), which operates across all layers of $\mathcal{G}$.

Precisely, each layer in $\mathcal{F}$ receives ($\mathbf{T}$) in the corresponding input space. In the first layer, $\mathbf{T}$ (aka $\mathbf{T}_0$) replaces m tokens of $\mathbf{W}_0$, and the input embedding for the first layer of $\mathcal{F}$ becomes $[\mathbf{T}_0; \mathbf{W}_0]$. Here, $[a; b]$ denotes stacking after replacing a similar number of tokens of b with all of the tokens of a. Consequently, for the l-th layer, $\mathbf{T}_l = \mathbf{T} = \mathcal{B}_t(\mathbf{E}_0)$.

Finally, the output operation for the l-th layer can be expressed as

$$[\,\mathbf{W}_l] = \mathcal{F}^l([\mathbf{T}_{l-1}; \mathbf{W}_{l-1}]) \in \mathbb{R}^{\mathcal{M} \times d_t} \quad (3)$$

For any given $(l+1)^{th}$ layer within $\mathcal{G}$, the process involving $\mathcal{B}_v$ can be succinctly described as: $\mathbf{V}_l^{\mathtt{tg}} = \mathbf{V}^{\mathtt{tg}} = \mathcal{B}_v(\mathcal{W}')$, effectively embedding semantic textual insights into the visual domain for enhanced model comprehension.

The flow of information for the text to the visual domain is further reinforced by token sharing from each layer of $\mathcal{F}$ to the respective layer input of $\mathcal{G}$, i.e the output ($\mathbf{W}_l$) of the l-th layer of $\mathcal{F}$ is transferred to a *vision-text conjunction block* ($\mathcal{B}_{vt}$), which then generates scale-specific inputs

($\mathbf{V}^{\mathtt{ms}}_{l-1}$) for the corresponding l-th layer of $\mathcal{G}$ consisting of n learnable tokens ($\mathcal{V}^{\mathtt{ms}}_{1:n}$). For the l-th layer of $\mathcal{G}$, $\mathbf{V}^{\mathtt{ms}}$ can be defined as

$$\begin{aligned}\mathbf{V}^{\mathtt{ms}}_{l-1} &= \{\mathcal{V}^{\mathtt{ms}}_{k_{l-1}} \in \mathbb{R}^{d_v}\}^{n}_{k=1} \\ &= \mathcal{B}_{vt}(\mathcal{F}^{l}([\mathbf{T}_{l-1}; \mathbf{W}_{l-1}])) \in \mathbb{R}^{n \times d_v}\end{aligned} \tag{4}$$

Finally, we concatenate the generated prompt tokens of $\mathbf{V}^{\mathtt{tg}}$ and $\mathbf{V}^{\mathtt{ms}}$ for each of the layers, expressed as a common *visual prompt* ($\mathbf{V}$) i.e. for inputting to the l-th layer of $\mathcal{G}$: $\mathbf{V}_{l-1} = [\mathbf{V}^{\mathtt{tg}}_{l-1}, \mathbf{V}^{\mathtt{ms}}_{l-1}]$. Hence, the processing at the l-th layer of $\mathcal{G}$ is mentioned as,

$$[\mathbf{c}_l, \ ; \mathbf{E}_l] = \mathcal{G}^{l}([\mathbf{c}_{l-1}, \mathbf{V}_{l-1}; \mathbf{E}_{l-1}]) \in \mathbb{R}^{1+\mathcal{N} \times d_v} \tag{5}$$

Classification Loss. Given an image x and a particular class c, the probability of x belonging to c is given as

$$p(y = c|x) = \frac{exp(\langle \mathcal{F}(x), \mathcal{G}(s_c) \rangle / \tau)}{\sum_{i}^{C} exp(\langle \mathcal{F}(x), \mathcal{G}(s_i) \rangle) / \tau)} \tag{6}$$

where τ denotes the temperature hyperparameter, and $\langle ., . \rangle$ denotes the cosine similarity computation between the image and text embeddings.

4. Experiments and Implementation

Benchmarks for Medical Anomaly Detection (BMAD). This benchmark includes six restructured datasets spanning five medical fields: brain MRI, liver CT, retinal OCT, chest X-ray, and digital histopathology. This well-curated and standardized medical benchmark, along with its organized code base, facilitates thorough comparisons between recent anomaly detection methods. The data consists of both images and text labels to identify normal and anomalous images; the train data for all the categories only contains normal images to facilitate zero-shot anomaly detection.

Accuracy vs. AUROC. In a multi-class domain, the performace of an anomaly classification problem would be observed with accuracy directly. However, in the case of a binary classification problem, the Area Under Receiver Operators Curve (AUROC) presents a refined metric for measuring performance, as shown in Table 1, and is a widely-used metric for anomaly detection.

Implementation. We use the publicly available CLIP model `VIT-L/14@336px` as our backbone. Fine-tuning was performed on the last 2 layers of the CLIP model using a batch size of 8 on a single NVIDIA A100 80GB. All training experiments were run with 10 epochs with the learning rate as 1e-5 and the Adam optimizer.

5. Preliminary Results

Results are summarized in Table 1. Our first set of evaluations brought out the capabilities of vanilla CLIP (without any fine-tuning) with simple prompts to adapt to the task of anomaly detection on various categories in the medical domain [Results 1-4 in Table 1]. We observed an average accuracy of 61%, and average AUROC of 66%.

Table 1. Performance results of CLIP with and without fine-tuning for Anomaly Detection, sorted by decreasing Image-level AUROC (%). An AUROC of 50% indicates random choice by a model.

	Fine-tuning	Test Data	Acc.↑	AUROC↑
CLIP	-	Retinal OCT	63.21	75.95
	-	Pathology	39.66	68.92
	-	Liver CT	57.60	61.59
	-	Brain MRI	82.23	57.54
CLIP$^{\text{FT}}$	Brain MRI	Brain MRI	82.83	66.13
	Pathology	Pathology	50.33	66.04
	Pathology	Brain MRI	84.52	64.31
	Brain MRI	Liver CT	65.97	57.21

Next, we fine-tuned the last two layers of the CLIP model, and performed zero-shot detection by training and testing on the train and test splits of images belonging to the same domain [Results 5-6 in Table 1]. Here, we observed an average accuracy of 66.58%, and average AUROC of 66.08%, which is slightly higher than the vanilla CLIP results. We come to the conclusion that finetuning CLIP helped us garner better results in the zero-shot medical anomaly detection task.

Finally, we measured robustness by performing zero-shot detection on OOD (out-of-domain) distributions. Here, we observed an average accuracy of 75.25%, and average AUROC of 60.76%, which is significantly lower than the fine-tuned as well as vanilla CLIP results. This is of particular importance because we can infer that a simple finetuning the CLIP model does not make it robust to OOD distributions, which is why a more involved parameter-sharing and association is required for the medical anomaly detection task.

Compiling our inferences from each of the tasks we explored using our baselines, we came to the following preliminary conclusions:

- Vanilla CLIP is not adapted well to identifying the nuances associated with the images from the medical domain, and corresponding does not yield very good results for medical image anomaly detection.
- A simple two-layer finetuning methodology works better for in-domain zero shot detection, but the performance drastically degrades in OOD setups. In such a scenario, there is a proven need for a more involved associated between image and text features for CLIP to learn better representations and hence aid in our task of robust medical image anomaly detection.

Statement of Contribution

Swetha explored the literature review, developed the code and conducted the preliminary experiments. Prakriti conducted a literature review, contributed to the methodology and compiled the preliminary results. Debarshiya rendered the architecture diagram, contributed to the abstract & datasets sections, and aided with the preliminary experiments.

References

[1] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad – a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[2] Yu Cai, Weiwen Zhang, Hao Chen, and Kwang-Ting Cheng. MedIAnomaly: A comparative study of anomaly detection in medical images, 2024. 1

[3] Xuhai Chen, Yue Han, and Jiangning Zhang. April-gan: A zero-/few-shot anomaly classification and segmentation method for cvpr 2023 vand workshop challenge tracks 12: 1st place on zero-shot ad and 4th place on few-shot ad, 2023. 1

[4] Chaoqin Huang, Aofan Jiang, Jinghao Feng, Ya Zhang, Xinchao Wang, and Yanfeng Wang. Adapting Visual-Language Models for Generalizable Anomaly Detection in Medical Images. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11375–11385, Seattle, WA, USA, 2024. IEEE. 1

[5] Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. Winclip: Zero-/few-shot anomaly classification and segmentation, 2023. 1, 2

[6] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning, 2023. 2

[7] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 1

[8] Mainak Singha, Ankit Jha, Divyam Gupta, Pranav Singla, and Biplab Banerjee. Elevating all zero-shot sketch-based image retrieval through multimodal prompt learning, 2024. 2

[9] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. MedCLIP: Contrastive Learning from Unpaired Medical Images and Text, 2022. arXiv:2210.10163. 2

[10] Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection, 2024. 1

[11] Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation, 2022. 1