https://doi.org/10.2352/EI.2026.38.12.GENAI-175

Vision-Language Learning for Wireless Capsule Endoscopy: Diagnostic Captioning with CLIP

Lu Xu, Anuja Vats, Marius Pedersen and Kiran Raja.
Colourlab, Department of Computer Science, Norwegian University of Science and Technology

Abstract

Wireless Capsule Endoscopy (WCE) is a minimally invasive diagnostic tool for examining the gastrointestinal tract, but the interpretation of large amounts of WCE image data demands extensive manual efforts and expert knowledge. Deep learning offers a promising approach to automate WCE data analysis, but training robust models is hindered by the scarcity of large-scale, high-quality labeled data in the WCE domain. This study explores the use of Contrastive Language-Image Pre-training (CLIP), a vision-language model pre-trained on extensive image-text pairs, to address these challenges in deep learning for WCE. We focus on caption retrieval and pathology classification tasks, using the CAPTIV8 dataset, a multi-modal WCE dataset containing image-diagnostic text pairs. After customizing the dataset for deep learning tasks, we conducted experiments comparing CLIP with state-of-the-art vision models. The results demonstrated that CLIP performs better than vision-only models, particularly in small-sample regimes such as one-shot and few-shot setups. By replacing the original CLIP loss with a KL-divergence loss, we further enhanced the model's ability to handle multiple positive pairs in a mini-batch during the training, to further attune learning for this specific medical domain.

Introduction

Wireless Capsule Endoscopy (WCE) is a minimally invasive diagnostic tool designed to examine the gastrointestinal tract. A small, swallowable capsule equipped with a camera, light source, and wireless transmitter captures high-resolution images as it traverses the digestive system. These images are transmitted to an external receiver for analysis. WCE offers detailed visualization of the gastrointestinal tract without requiring sedation or invasive instrumentation. However, analyzing WCE videos remains labor-intensive and time-consuming, as each recording spans hours and requires expert interpretation, making the process costly. Recent advancements in deep learning present a promising direction for automating the analysis of WCE data, enabling efficient and accurate diagnosis [1–5]. However, the development of robust deep learning models in the medical domain faces a key obstacle: the scarcity of accessible high-quality labeled data. Annotating medical datasets requires specialized domain knowledge, making the process labor-intensive and expensive. Moreover, collecting sufficient medical data is further constrained by privacy regulations.

To address the challenge of limited labeled data, Vision-Language Pre-training (VLP) has emerged as a promising approach. Vision-Language Models (VLMs) combine visual and textual information to learn joint representations of images and text. Unlike vision-only models that rely solely on extracting image features, VLMs incorporate accompanying textual descriptions to provide additional semantic context, enhancing the effectiveness and robustness of visual representations. VLP combines VLM with the concept of pre-training, a process where models are trained on large-scale datasets to learn generalizable representations that can be fine-tuned for specific downstream tasks. By pre-training on large-scale image-text pairs, VLPs are highly versatile for zero-shot and few-shot learning scenarios.

This study focuses on investigating the use of Contrastive Language-Image Pre-training (CLIP) [6], a state-of-the-art VLP model, on the publicly available CAPTIV8 dataset [7], a joint image-text dataset specific to the WCE domain. In this study, we focus on two key tasks: caption retrieval and pathology classification. The caption retrieval task involves identifying the most appropriate textual description for a given WCE image from a set of candidate captions. Beyond its clinical relevance, this task serves as a benchmark for evaluating the model's ability to learn and align joint visual-textual representations. Superior performance in caption retrieval indicates the model's effectiveness in capturing meaningful cross-modal relationships. The pathology classification task involves categorizing WCE images into clinically significant pathological classes, including inflammation, polyps, and angioectasia. However, current methods often struggle to fully exploit the semantic alignment between images and their textual counterparts, particularly in scenarios where multiple valid captions correspond to similar or sequential images. This limitation becomes more pronounced in WCE data, where consecutive video frames frequently depict the same or closely related pathological findings, resulting in multiple positive image-text pairs within a single batch. to address this we propose a modification to CLIP's original contrastive loss function, Instead of relying solely on pairwise contrastive objectives, we incorporate a KL-divergence based loss that explicitly accounts for multiple positive pairs within a batch. This adjustment enables the model to learn more richer and flexible joint representations.

Our experiments demonstrate that CLIP performs better than vision-only vision models, particularly in small-sample regimes (e.g., one-shot and few-shot setups). Furthermore, the KL-divergence loss we proposed improved the performance and robustness of CLIP in the WCE application.

Background

In the field of computer vision, VLP have emerged as a hot research focus. According to Chen et al. [8], the first step in VLP is to extract features from both image and text inputs. Convolutional Neural Networks (CNNs), which have historically been effective for visual feature extraction, continue to play a role in VLP. However, the advent of Transformers in natural language processing has significantly influenced computer vision method-

ologies. Transformers, originally introduced by Vaswani [9]for natural language processing tasks, revolutionized sequence modeling by introducing the self-attention mechanism, which allows for capturing global dependencies efficiently. Vision Transformer (ViT) [10] extended Transformer architectures to the computer vision domain by dividing images into small patches and treating them as tokens, akin to words in natural language processing tasks. Liu et al. [11] improve ViT by incorporating a hierarchical structure with shifted windows to achieve efficient computation and better leverage context information. The success of transformer-based methods has led to their widespread adoption in VLP for extracting image features. For text features, BERT [12], a large pre-trained language model, is widely used in VLP due to its versatility and strong contextual understanding.

To jointly learn features extracted from images and text, VLP models often employ contrastive learning, a technique that aligns representations of paired inputs in a shared latent space. Contrastive learning was originally developed for uni-modal vision models [13]. Zhang et al. [14] introduced contrastive representation learning for image-text pairs in the medical imaging domain, showcasing the potential of this technique in aligning modality specific representations. Building on these foundations, OpenAI introduced Contrastive Language-Image Pre-training (CLIP) [6] in 2021. By leveraging contrastive learning to be trained on an extensive dataset of 400 million image-text pairs sourced from publicly available internet data, CLIP has exhibited remarkable performance on zero-shot scenarios and several downstream tasks. CLIP's ability to generalize across domains without task-specific fine-tuning has made it a cornerstone for VLP research.

Given its generalizability, CLIP has been adapted for various domains and downstream tasks. A large number of studies focus on adapting and improving CLIP. Wang et al. [15] utilize CLIP in action recognition for videos where they propose using soft similarity-distributions between video0-caption pairs that are not exactly one-hot to encourage matching a caption to multiple videos and vice-versa. Similarly, Gao et al. [16] proposed SoftCLIP, which relaxed the strict constraint of one positive pair per batch in the original CLIP framework. Other studies focus on efficiently fine-tuning CLIP. Zhou et al. [17] adapted the prompt learning paradigm from the natural language domain to CLIP, introducing learnable text prompts to guide the model. Building on this, Khattak et al. [18] proposed a self-regulating prompt mechanism to mitigate overfitting during fine-tuning. Fan et al. [19] proposed a method to improve CLIP training by utilizing language generative models to rewrite textual prompts.

Adapting VLP models to the medical domain presents unique challenges, particularly the scarcity of high-quality paired data. To address this, Wang et al. [20] introduced MedCLIP, which leveraged large-scale unpaired data for training. Similarly, Zhang et al. [21] proposed BiomedCLIP, re-training CLIP on 15 million image-text pairs extracted from biomedical research articles in PubMed Central. To leverage the corresponding medical reports that contain detailed descriptions, GLoRIA [22] proposed a framework that jointly learns global and local representations for medical images by contrasting image sub-regions and words in the paired report.

Constrained by the scarcity of labeled data, deep learning research in WCE has lagged behind other medical domains. Existing studies predominantly rely on uni-modal methods, focusing primarily on CNNs. For example, Masmoudi et al. [23] utilized CNNs for ulcer classification. Similarly, Fernandes et al. [24] explored CNN-based approaches for image retrieval tasks in WCE datasets. Lafraxo et al. [25] proposed a two-stage method for detecting bleeding, combining CNN for feature extraction with gated recurrent units for classification. Despite these advancements, the performance and potential of VLP in WCE remain open research questions.

Methodology

Dataset

The CAPTIV8 dataset ([7]) is an image-text dataset specifically created for WCE. It consists of frames and short video segments extracted from WCE examination videos, from 10 patients diagnosed with ulcerative colitis. Alongside the visual data, the dataset includes alphanumeric metadata such as diagnostic summaries and histopathology reports. Experienced gastroenterologists utilized the Rapid Reader software to annotate clean and representative frames, distinguishing between normal and abnormal observations and providing corresponding descriptions of diagnostic text. In total, the dataset has 1,352 annotated frames. Examples of images and their descriptive text are shown in Figure 1.

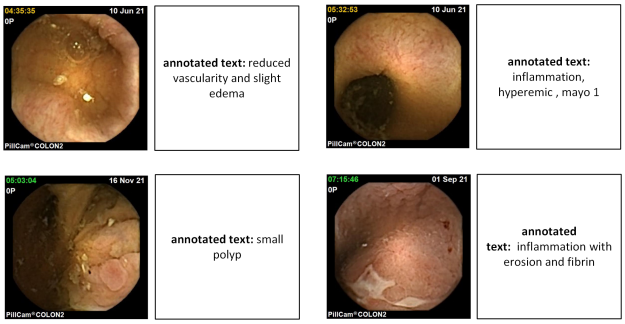


Figure 1: Examples of selected frame and corresponding descriptive text in the CAPTIV8 dataset.

Annotation Refinement for Clinical Data

The raw clinical annotations in the dataset present significant challenges for learning. These challenges manifest as: (1) varying detail granularity, where some annotations include diagnostic conclusions (e.g., "colitis") while others contain only frame-specific observations such as "inflammation and ulcer"; (2) syntactic inconsistencies in formatting, whereby the same clinical findings are expressed differently (e.g., "erosion with fibrin, reduced vascularity" versus "reduced vascularity, erosion, fibrin"); and (3) presence of rare abnormalities, providing insufficient examples for learning.

To address these limitations, we perform annotation refinement on the dataset to generate standardized caption labels and clinically meaningful category labels. The caption labels were derived through text-refinement, retaining only clinically relevant entities/terms in images for learning. These captions were then mapped to higher-level category labels under clinical guidance. This consolidated (pathophysiologically) related abnormalities into a single diagnostic category-for example, mapping reduced vascularity, fibrin deposition and edema to a single "inflammation" category. Algorithm 1 details this step. The resulting taxonomy comprises seven diagnostic categories: inflam-

mation, polyp, hemorrhoid, biopsy lesion, angioectasia, lipoma, and normal to be used during evaluation of the trained vision-language representations through classification tasks. The dataset was split into 896/105/267 samples for training/validation/testing, maintaining balanced category distribution across all partitions.

Algorithm 1 Hierarchical Refinement of Clinical Annotations

Require: Raw text T associated with WCE image
Ensure: Refined caption C and category label Y

/* Caption Label */
Initialize clinically relevant entity set $\mathscr{E} \leftarrow \{$`inflammation`, `angioectasia`, `reduced vascularity`, $\dots\}$
$L \leftarrow \emptyset$ ▷ Initialize empty list for extracted entities
for each entity $e \in \mathscr{E}$ **do**
  if e appears in T **then**
    $L \leftarrow L \cup \{e\}$ ▷ Extract entities from raw description
  end if
end for
for each entity $e \in L$ **do**
  ▷ Remove modifiers such as anatomical locations (e.g., "in the duodenum"), adjectives (e.g., "mild"), and broader diagnostic interpretations. This ensures that the labels capture only the core visual concepts.
  $e \leftarrow$ `standardized`(e)
end for
$L \leftarrow L \setminus \{$`food residue`, `bubbles`, $\dots\}$ ▷ Filter non-diagnostic entities
$C \leftarrow$ `Join`$(L, \text{delimiter} = \text{`, '})$ ▷ Generate caption label
/* Category Label */
Initialize category mapping $\mathscr{M} : \mathscr{E} \rightarrow \mathscr{Y}$ where $\mathscr{Y} = \{$`inflammation`, `polyp`, $\dots\}$
$\mathscr{M}[$`reduced vascularity`$] \leftarrow$ `inflammation`
$\mathscr{M}[$`fibrin`$] \leftarrow$ `inflammation`
$\mathscr{M}[$`erosion`$] \leftarrow$ `inflammation`
$\mathscr{M}[$`edema`$] \leftarrow$ `inflammation`
$\mathscr{M}[$`polyp`$] \leftarrow$ `polyp`
$\mathscr{M}[$`ulcer`$] \leftarrow$ `ulcer`
$\vdots$ ▷ Additional entity-category mappings
$Y \leftarrow \emptyset$ ▷ Initialize empty set for category labels
for each entity $e \in L$ **do**
  $Y \leftarrow Y \cup \{\mathscr{M}[e]\}$ ▷ Map entities to categories
end for
if $Y = \emptyset$ **then**
  $Y \leftarrow \{$`normal`$\}$ ▷ Default to normal if no abnormalities detected
end if
return C, Y ▷ Return both caption and category labels

We use CLIP [6] as our base model. The CLIP architecture comprises dual encoder pathways, an image encoder $f_I(\cdot)$ and a text encoder $f_T(\cdot)$ that project their respective inputs into a shared semantic embedding space. During training, these encoders process paired endoscopic images and their corresponding clinical caption labels in batches (Figure 2). The representations extracted by each encoder are L2-normalized and linearly projected into a shared d-dimensional latent space, enabling direct computation of cross-modal semantic similarities. The alignment between modalities is achieved through a bidirectional contrastive learning objective that optimizes the model to maximize the similarity between semantically corresponding image-text pairs while minimizing similarity between non-corresponding pairs (Eq. (1)).

For a training batch of size N, the model processes N^2 potential image-text combinations - N positive pairs (highlighted in blue in Figure 2) and $N^2 - N$ negative pairs. The contrastive objective for image-to-text alignment is formulated as:

$$\mathscr{L}_{\text{img}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\left(sim(\mathbf{I}_i,\mathbf{T}_i)/\tau\right)}{\sum_{j=1}^{N}\exp\left(sim(\mathbf{I}_i,\mathbf{T}_j)/\tau\right)}, \tag{1}$$

where $sim(\cdot,\cdot)$ is the cosine similarity. $\mathbf{I}_i$ and $\mathbf{T}_i$ denote the i-th image and text representations in the multimodal embedding space respectively. τ is a learnable temperature parameter. Similarly, the objective for text-to-image pairing is formulated as:

$$\mathscr{L}_{\text{txt}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\left(sim(\mathbf{T}_i,\mathbf{I}_i)/\tau\right)}{\sum_{j=1}^{N}\exp\left(sim(\mathbf{T}_i,\mathbf{I}_j)/\tau\right)}. \tag{2}$$

Finally, the CLIP model is optimized with the loss function of the average of these two losses:

$$\mathscr{L} = \frac{\mathscr{L}_{\text{img}} + \mathscr{L}_{\text{txt}}}{2}. \tag{3}$$

Pre-training

The visual-encoder, either Resnet-based [26] or ViT-based [10] extracts visual features, while the BERT-based [27] text-encoder processes textual input.

Handling False Negatives

A fundamental limitation of CLIP's conventional contrastive learning paradigm is its restrictive one-to-one correspondence assumption between images and textual descriptions. In the original CLIP framework, each mini-batch constructs a similarity matrix where only diagonal elements (representing paired image-text inputs) are considered positive matches, while all off-diagonal elements are treated as negative pairs. This approach is suboptimal for medical images where multiple frames can share pathological features and, consequently, share caption labels creating multiple potential positive matches for a single image or text input. Under CLIP's standard contrastive formulation, these additional positive pairs are erroneously penalized as false negatives, introducing noisy supervision signals that may impede effective multimodal alignment.

To address this challenge, we introduce a simple yet effective modification that allows multi-positive contrastive learning that accommodates the many-to-many relationships inherent in several domains including medical. Rather than enforcing a strict diagonal-only positive pair structure, our modified similarity matrix Q assigns a value of 1 to all valid image-caption pairs that exhibit true semantic correspondences, regardless of their position in the batch matrix as seen in Figure 3. Through this adaptation, the learning objective can be adapted from a 1-in-N classification problem to a distribution matching problem. The multiple positives imply that for a given image (i), multiple text descriptions may be equally valid matches, creating a non-sparse distribution over predicted similarities $Q(i,\cdot)$.

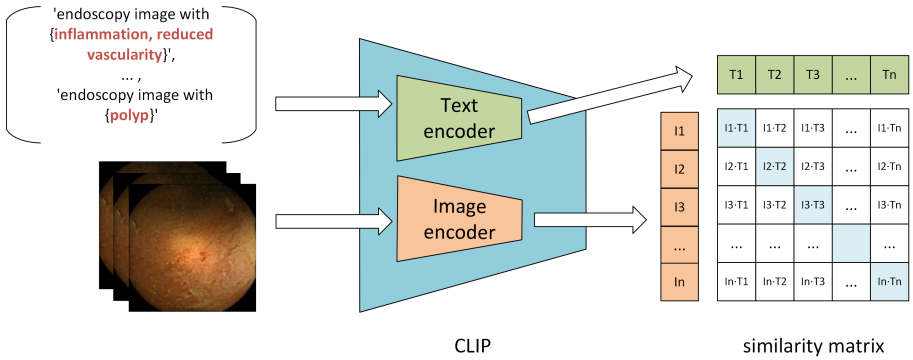


Figure 2: Diagram of CLIP architecture.

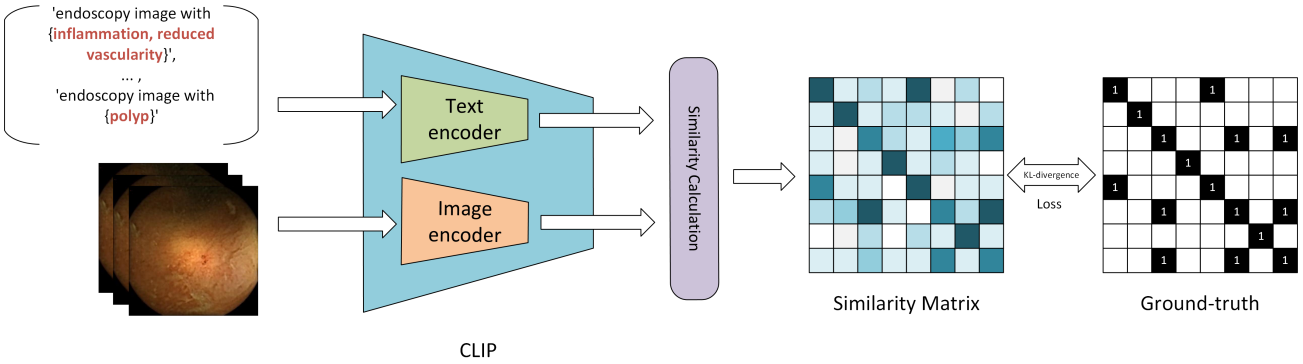


Figure 3: Diagram of replacing CLIP's original loss function with KL-divergence loss for contrastive learning.

The proposed loss for image-to-text and text-to-image are formulated as:

$$P_{\text{img}}(i,j) = \frac{\exp\left(sim(\mathbf{I}_i, \mathbf{T}_j)/\tau\right)}{\sum_{k=1}^{N} \exp\left(sim(\mathbf{I}_i, \mathbf{T}_k)/\tau\right)}, \tag{4}$$

and

$$P_{\text{txt}}(i,j) = \frac{\exp\left(sim(\mathbf{T}_i, \mathbf{I}_j)/\tau\right)}{\sum_{k=1}^{N} \exp\left(sim(\mathbf{T}_i, \mathbf{I}_k)/\tau\right)}, \tag{5}$$

Finally, the bi-directional objective used to optimize the model is:

$$\mathscr{L}_{\text{KL}} = \frac{1}{2}\Big(\frac{1}{N}\sum_{i=1}^{N} D_{\text{KL}}(Q(i,\cdot)\|P_{\text{img}}(i,\cdot)) + \frac{1}{N}\sum_{i=1}^{N} D_{\text{KL}}(Q(\cdot,i))\|P_{\text{txt}}(i,\cdot))\Big), \tag{6}$$

where target distribution $Q(i,j)$ is constructed by applying thresholds m_+ and m_- to the raw similarity scores : pairs with similarity above m_+ are treated as positives, while those below m_- are negatives. This allows multiple positives to be assigned within a single batch. The final loss is an average KL-divergence between Q and the predicted distributions in both directions,

$$L = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} Q(i,j)\max(0, m_+ - sim(I_i,T_j)) + (1-Q(i,j))\max(0, sim(I_i,T_j) - m_-(i)) \tag{7}$$

Inference: Caption-Retrieval

This task aims to identify the most relevant caption for a given image from a predefined set of candidate captions. The inference procedure is shown in Figure 4 and summarized as follows:

1. **Image encoding**: The image is passed through the image encoder to generate its embedding in the multimodal embedding space.
2. **Text encoding**: To ensure consistency with the fine-tuning stage, textual inputs are formatted using the same prompt templates employed during training. Specifically, each candidate caption is prefixed with *"endoscopy image with"* before the medical description. The resulting text is then passed through the text encoder to generate embeddings in the same multimodal space.
3. **Similarity computation**: Cosine similarity is computed between the image embedding and each candidate text embedding.
4. **Prediction**: The caption with the highest similarity to the image embedding is selected as the prediction.

Inference: Classification

While caption labels provide rich semantic descriptions and may include one or more abnormalities observed in an image, category labels consist of a single word indicating a specific pathology. As a result, they offer substantially less semantic information. Due to this difference, we do not treat category labels as textual inputs for fine-tuning CLIP. Instead, we adopt a classification-based approach, in which a set of fully connected (FC) layers is

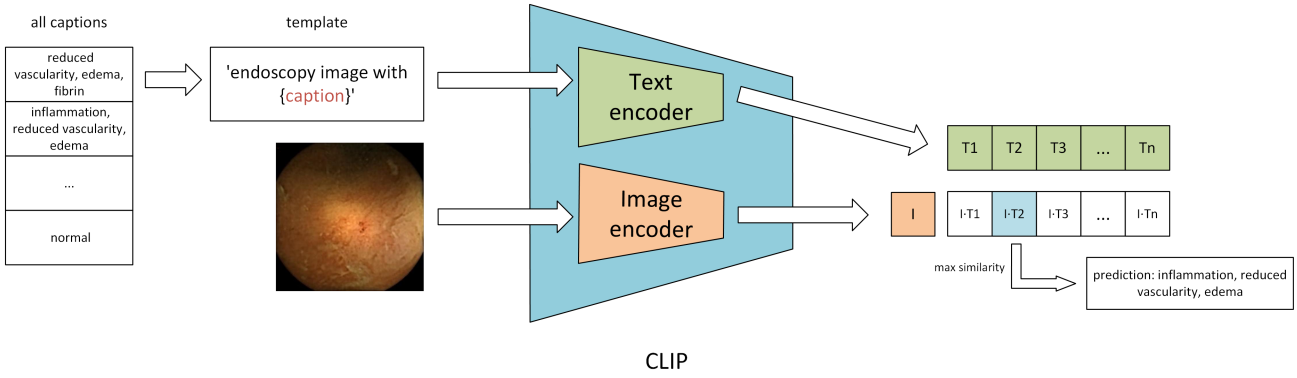


Figure 4: Diagram of the procedure of using CLIP for the inference of matching captions.

appended to the image encoder of a fine-tuned CLIP model to perform pathology classification.

In this setup, only the image encoder component of CLIP that is already fine-tuned on image-caption pairs is used. The image encoder extracts feature embeddings from input images which are then passed through the FC layers, which serve as a classifier to predict the pathology category. Figure 5 shows this pipeline.

During training, the weights of the fine-tuned image encoder are frozen, ensuring that only the FC layers are trained. Given the highly imbalanced category distribution in our dataset, we employed a weighted cross-entropy loss to optimize the model.

Results and Discussion

Training Configuration

We selected the ViT-L/14@336px CLIP model. To fine-tune the models, we use the Adam (Adaptive Moment Estimation) optimizer [28]. As suggested by Liang et al. [29], fine-tuning CLIP models benefit from a small initial learning rate. Therefore, we set the initial learning rate to 1×10^{-6}. However, when training CLIP, a larger batch size may increase the risk of including multiple images with the same caption in a mini-batch, which can adversely impact training effectiveness when using CLIP's original loss function. To determine the optimal batch size under our GPU's memory constraints, we tested different configurations. The results, presented in Table 1, indicate that a batch size of 8 offers the best trade-off. Therefore, we selected a batch size of 8 for our experiments.

Table 1: Performance comparison with fine-tuning CLIP (ViT-L/14@336px) on different batch size.

Batch Size	ACC1 [%]	ACC5 [%]
16	33.64± 3.26	70.24 ± 2.30
8	**35.26** ± 2.00	**70.42** ± 2.71
4	32.18 ± 2.14	67.51 ± 3.62

Performance Comparison

We compare our method with state-of-the-art vision-only models, including VGG19, ResNet50, ViT-L/16, and MobileNetV3. All vision-only models were initialized with weights pre-trained on ImageNet-1K, accessed via TorchVision [30].

Caption Retrieval

For the caption retrieval task, we fine-tuned CLIP on pairs of images and caption labels and employed the inference process to identify the matching captions. To provide a comparison with vision-only models, we treated caption labels as class names and reformulated the task as a standard classification problem for vision-only models. The results are presented in Table 2.

Table 2: Caption retrieval performance. ACC 1 denotes the Top-1 accuracy score, and ACC 5 denotes the top-5 accuracy score. Bold font highlight the highest average score within every setup.

Model	Setup	ACC 1 [%]	ACC 5 [%]
Zero-shot CLIP (ViT-B/16)	zero-shot	8.60	36.83
CLIP (Original Loss)	one-shot	14.34 ± 2.46	**43.45** ± 2.73
	four-shot	19.57 ± 1.24	51.05 ± 5.03
	full-set	35.26 ± 2.00	70.42 ± 2.71
CLIP (KL Loss)	one-shot	**14.98** ± 2.73	37.77 ± 4.99
	four-shot	**24.12** ± 2.50	**60.78** ± 2.06
	full-set	**39.89** ± 1.44	67.67 ± 4.58
VGG19	one-shot	7.91 ± 2.65	36.00 ± 3.68
	four-shot	9.43 ± 1.82	32.44 ± 7.97
	full-set	26.73 ± 3.15	65.87 ± 2.81
ResNet50	one-shot	11.69 ± 2.66	40.75 ± 8.86
	four-shot	18.26 ± 3.13	50.83 ± 3.83
	full-set	31.53 ± 3.44	72.44 ± 5.53
ViT-L/16	one-shot	7.22 ± 1.79	31.85 ± 3.45
	four-shot	17.51 ± 2.08	46.95 ± 5.83
	full-set	33.36 ± 1.79	**73.42** ± 1.71
MobileNetV3	one-shot	6.61 ± 3.82	29.32 ± 6.57
	four-shot	9.43 ± 3.61	42.90 ± 4.51
	full-set	29.59 ± 1.06	70.72 ± 2.00

We primarily focus on Top-1 accuracy as a key evaluation metric. From the results presented in Table 2, we can observe that CLIP models consistently perform better than vision-only models across various setups. Specifically, CLIP optimized with the original loss achieves a Top-1 accuracy of 35.26% under the full-set setup, while CLIP optimized with KL-divergence loss achieves the highest Top-1 accuracy of 39.89%, suggesting that the effectiveness of KL-divergence loss in improving contrastive learning, especially in our task where multiple positive pairs may exist within a batch.

Additionally, in small-sample setups, the advantage becomes even more pronounced. For example, under the one-shot setup, CLIP (with KL-divergence loss) achieves a Top-1 accuracy of

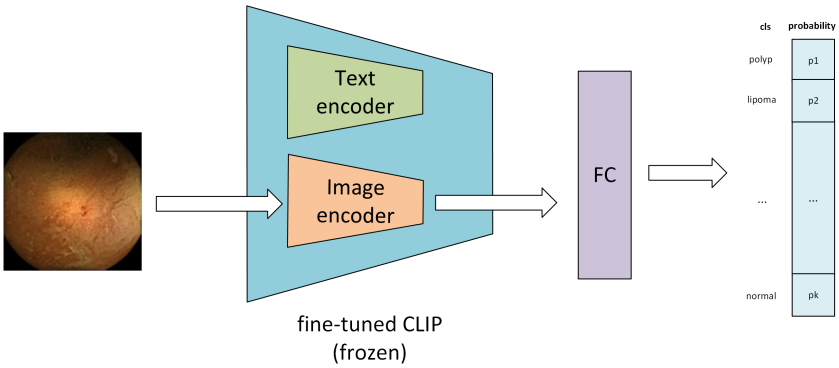


Figure 5: Diagram of adding FC layers after CLIP image encoder for classification.

14.98%, whereas vision-only models such as VGG19, ResNet50, ViT-L/16, and MobileNetV3 exhibit significantly lower performance, with Top-1 accuracies of 7.91%, 11.69%, 7.22%, and 6.61%, respectively. Similarly, in the few-shot setup, CLIP (with KL-divergence loss) achieves 24.12%, while VGG19, ResNet50, ViT-L/16, and MobileNetV3 only reach 9.43%, 18.26%, 17.51%, and 9.43%, respectively. We plot the trend in Figure 6. From the figure, it is clear that CLIP models consistently achieve higher Top-1 accuracy across all setups compared to other vision models. Some vision-only models, such as ViT-L/16, VGG19, and MobileNetV3 show a one-shot performance even lower than zero-shot CLIP (ViT-B/16 image encoder). Notably, when comparing CLIP with ViT-L/16, the performance difference is more pronounced in small-sample settings. In the one-shot setup, CLIP demonstrates a significant improvement over ViT-L/16, while the performance gap narrows in the full-set condition. This indicates that CLIP has the potential to learn meaningful features even from a limited amount of labeled data.

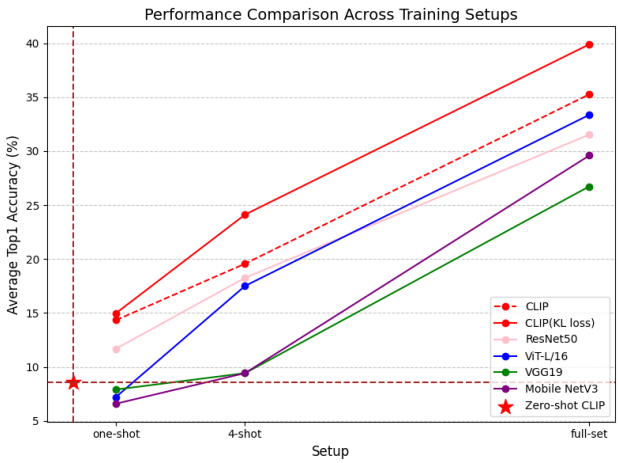


Figure 6: Models' performance comparison across setups.

However, for Top-5 accuracy, while CLIP models perform better than vision-only models in the one-shot and few-shot setups, they show a lower Top-5 accuracy in the full-set setup. This discrepancy may suggest the possible overfitting, as CLIP is a large-scale model while our dataset is comparatively small. Overfitting makes CLIP achieve higher precision in aligning image and textual representations but may struggle to detect relevant cases effectively, leading to higher Top-1 accuracy but lower Top-5 accuracy compared to vision-only models.

Pathology Classification

To fairly evaluate and compare the performance of different models on the pathology classification task, we append the same classifier composed of an FC layer after CLIP image encoders and other vision models. In this task, we only conducted experiments under full-set setup. The results are presented in Table 3

Table 3: Performance comparison of models in the pathology classification task. ACC denotes the accuracy score, while F1 represents the F1 score. The highest average score is highlighted in bold font.

Encoder	ACC [%]	F1 (macro)
Vit-L/16	66.15 ± 3.61	35.87 ± 2.87
VGG19	55.51 ± 11.29	27.14 ± 2.68
ResNet50	**67.87** ± 0.97	38.43 ± 1.82
MobileNetV3	64.58 ± 3.33	29.68 ± 2.31
CLIP (Original Loss)	64.43 ± 2.76	**42.54** ± 2.48
CLIP (KL Loss)	66.41 ± 1.22	41.49 ± 1.53

From Table 3, we can conclude that under the full-set setup, although CLIP models produce slightly lower accuracy, they consistently achieve higher F1 scores compared to vision-only models. This difference highlights the CLIP models' ability to handle class imbalance effectively. Since certain categories contain very few examples, vision-only models may struggle to learn meaningful features of minor categories from limited data, resulting in lower F1 scores. In contrast, CLIP leverages textual input, enabling them to extract richer and more robust representations even from scarce data, contributing to their higher F1 scores.

Embedding Space Analysis

From our experiments, we concluded that the KL-divergence loss better suits our tasks compared to CLIP's original loss. To provide further insight, we visualized the image embedding spaces of CLIP models fine-tuned with these two different losses, respectively. Since high-dimensional embedding space cannot be visualized directly, we applied Uniform Manifold Approximation and Projection (UMAP) [31] which is a non-linear dimensionality reduction technique that preserves both local and global structures, making it particularly effective for revealing patterns in high-dimensional embedding spaces.

The visualizations in Figure 7 compare three scenarios: the pre-trained CLIP without fine-tuning, CLIP fine-tuned with the original contrastive loss, and CLIP fine-tuned with KL-divergence loss. To ensure a fair comparison, we used the same UMAP parameters for all visualizations (n_neighbors = 5, min_dist = 0.3), and the fine-tuned models selected for visualization were the ones achieving the best performance during training.

In the pre-trained CLIP visualization, no discernible clusters are observed. Images with the same caption label appear scattered throughout the embedding space. This reflects the domain gap between CLIP's natural image-text pre-training data and WCE data.

When fine-tuned with CLIP's original contrastive loss, clusters corresponding to images with the same caption labels begin to emerge. However, the clusters remain entangled, with significant overlap between neighboring categories. This indicates that while the original loss successfully pulls images with the same caption closer together, it struggles to fully cluster them. This is because the original contrastive loss incorrectly defined negative pairs, reducing the model's ability to align all matching image-text pairs accurately.

In contrast, fine-tuning with the KL-divergence loss yields a much more structured embedding space. Images sharing the same caption labels form well-separated and coherent clusters with minimal overlap between categories. This suggests that KL-divergence loss is more effective in our tasks.

Conclusion

In this study, we investigated the application of CLIP for WCE image analysis, focusing on two key tasks: caption retrieval and pathology classification. We systematically compared CLIP with conventional uni-modal vision models and found that CLIP consistently outperformed these baselines across all experimental settings. Its advantages were particularly pronounced in low-data regimes, such as one-shot and few-shot scenarios, where traditional models exhibited notable performance degradation. Furthermore, by replacing CLIP's original contrastive loss with a KL-divergence-based objective, we enhanced the model's ability to accommodate multiple positive image–text pairs within a single batch: a common characteristic of sequential, video-derived WCE data. This led to measurable performance gains, demonstrating the effectiveness of adapting VLP models to the unique properties of medical imaging datasets.

Acknowledgments

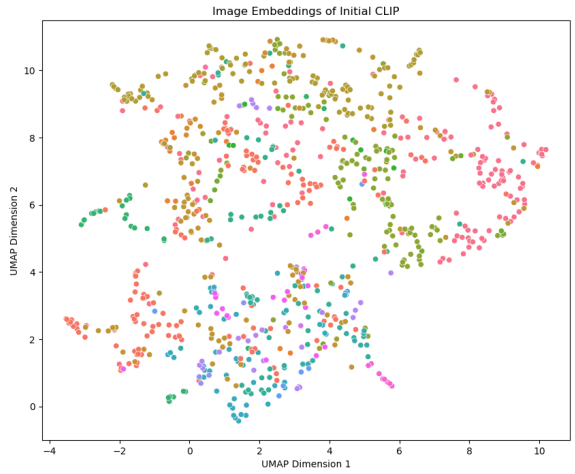
This work is supported by a grant from the Research Council of Norway "Capsnetwork" project: 322600.

References

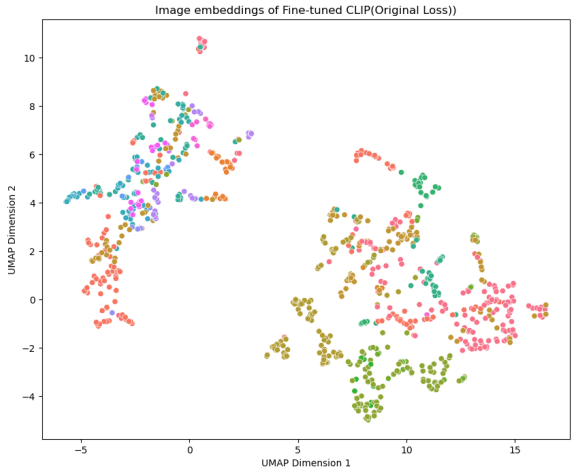
[1] Anuja Vats, Ahmed Mohammed, Marius Pedersen, and Nirmalie Wiratunga. This changes to that: Combining causal and non-causal explanations to generate disease progression in capsule endoscopy. In *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[2] Anuja Vats, Ahmed Mohammed, and Marius Pedersen. From labels to priors in capsule endoscopy: a prior guided approach for improving generalization with few labels. *Scientific Reports*, 12(1):15708, 2022.

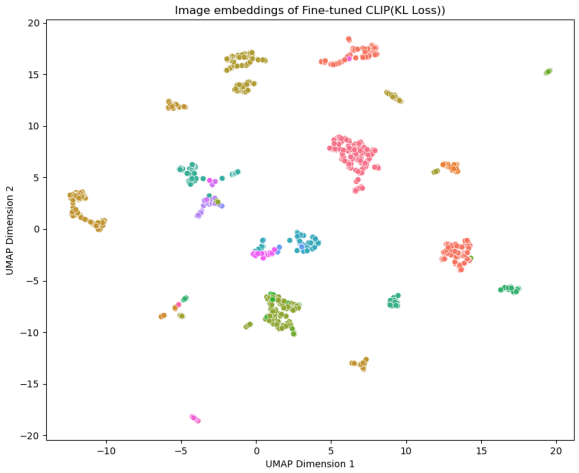
[3] Anuja Vats, Marius Pedersen, Ahmed Mohammed, and Øistein Hovde. Learning more for free-a multi task learning approach for



(a) Pre-trained CLIP



(b) Fine-tuned CLIP (Original Loss)



(c) Fine-tuned CLIP (KL Loss)

Figure 7: Visualization of image representations in the embedding space by UMAP. Different colors denote images of different caption labels.

improved pathology classification in capsule endoscopy. In *International conference on medical image computing and computer-assisted intervention*, pages 3–13. Springer, 2021.

[4] Ahmed Mohammed, Ivar Farup, Marius Pedersen, Sule Yildirim, and Øistein Hovde. Ps-devcem: Pathology-sensitive deep learning model for video capsule endoscopy based on weakly labeled data. *Computer Vision and Image Understanding*, 201:103062, 2020.

[5] Ahmed Mohammed, Sule Yildirim, Marius Pedersen, Øistein Hovde, and Faouzi Cheikh. Sparse coded handcrafted and deep features for colon capsule video summarization. In *2017 IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS)*, pages 728–733. IEEE, 2017.

[6] 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.

[7] Anuja Vats, Bilal Ahmed, Pål Anders Floor, Ahmed Mohammed, Marius Pedersen, and Øistein Hovde. Replication Data for: CAP-TIV8 : A comprehensive large scale CAPsule endoscopy dataset for Integrated diagnosis, 2024.

[8] Fei-Long Chen, Du-Zhen Zhang, Ming-Lun Han, Xiu-Yi Chen, Jing Shi, Shuang Xu, and Bo Xu. Vlp: A survey on vision-language pre-training. *Machine Intelligence Research*, 20(1):38–56, 2023.

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[11] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[13] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020.

[14] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Machine learning for healthcare conference*, pages 2–25. PMLR, 2022.

[15] Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*, 2021.

[16] Yuting Gao, Jinfeng Liu, Zihan Xu, Tong Wu, Enwei Zhang, Ke Li, Jie Yang, Wei Liu, and Xing Sun. Softclip: Softer cross-modal alignment makes clip stronger. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 1860–1868, 2024.

[17] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.

[18] Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without forgetting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15190–15200, 2023.

[19] Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. Improving clip training with language rewrites. *Advances in Neural Information Processing Systems*, 36:35544–35575, 2023.

[20] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2022, page 3876, 2022.

[21] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, et al. Large-scale domain-specific pretraining for biomedical vision-language processing. *arXiv preprint arXiv:2303.00915*, 2(3):6, 2023.

[22] Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, and Serena Yeung. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3942–3951, 2021.

[23] Youssef Masmoudi, Muhammad Ramzan, Sajid Ali Khan, and Mohammed Habib. Optimal feature extraction and ulcer classification from wce image data using deep learning. *Soft Computing*, 26(16):7979–7992, 2022.

[24] Rodrigo Fernandes, Alexandre Pessoa, José Nogueira, Anselmo Paiva, Ishak Paçal, Marta Salgado, and António Cunha. Evaluation of deep learning models in search by example using capsule endoscopy images. *Procedia Computer Science*, 239:2065–2073, 2024.

[25] Samira Lafraxo, Mohamed El Ansari, and Lahcen Koutti. Computer-aided system for bleeding detection in wce images based on cnn-gru network. *Multimedia tools and applications*, 83(7):21081–21106, 2024.

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[27] Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[28] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[29] Chia Xin Liang, Pu Tian, Caitlyn Heqi Yin, Yao Yua, Wei An-Hou, Li Ming, Tianyang Wang, Ziqian Bi, and Ming Liu. A comprehensive survey and guide to multimodal large language models in vision-language tasks. *arXiv preprint arXiv:2411.06284*, 2024.

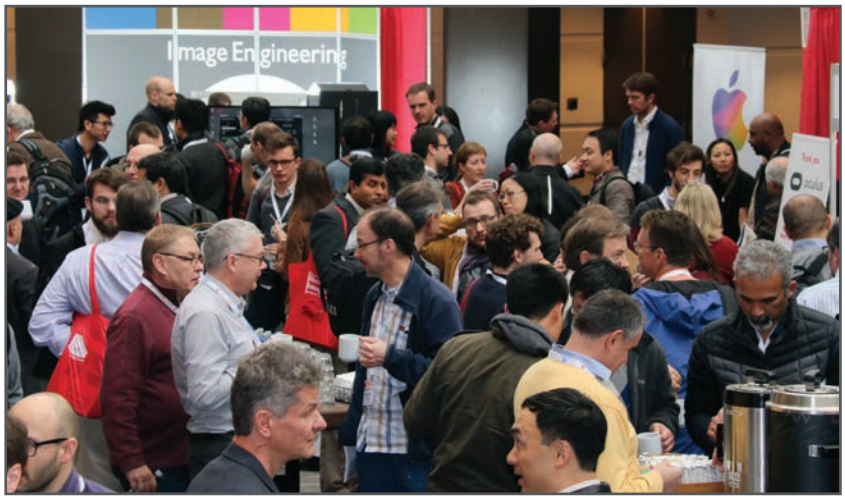
[30] TorchVision. Torchvision document, 2016. Accessed: 2024-12-15.

[31] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

JOIN US AT THE NEXT EI!

electronic IMAGING

Imaging across applications . . . Where industry and academia meet!



• SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •
• INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •

www.electronicimaging.org

ADVANCING IMAGING SCIENCE 75 1947 IS&T 2022 imaging.org