An Empirical Analysis for Zero-Shot Multi-Label Classification on COVID-19 CT Scans and Uncurated Reports

Ethan Dack[1,2], Lorenzo Brigato[1,2], Matthew McMurray[3],
Matthias Fontanellaz[1,2], Thomas Frauenfelder[3], Hanno Hoppe[2,5,6], Aristomenis Exadaktylos[7], Thomas Geiser[8], Manuela Funke-Chambour[8], Andreas Christe[3], Lukas Ebner[3], and Stavroula Mougiakakou[1,2]

[1]AI in Health and Nutrition, ARTORG Center for Biomedical Engineering Research
[2]University of Bern
[3]Department of Diagnostic, Interventional, and Pediatric Radiology, Bern University Hospital
[4]Diagnostic and interventional Radiology, University Hospital Zurich
[5]Department of Radiology, Lindenhofspital, Bern
[6]Campus Stiftung Lindenhof Bern (SLB), Bern, Switzerland
[7]Department of Emergency Medicine, Bern University Hospital
[8]Department of Pulmonary Medicine, Bern University Hospital
{ethan.dack,lorenzo.brigato,matthias.fontanellaz,
hanno.hoppe,stavroula.mougiakakou}@unibe.ch
{matthewthomas.mcmurray,aristomenis.exadaktylos,thomas.geiser,
manuela.funke-chambour,andreas.christe,lukas.ebner}@insel.ch
thomas.frauenfelder@usz.ch

Abstract

The pandemic resulted in vast repositories of unstructured data, including radiology reports, due to increased medical examinations. Previous research on automated diagnosis of COVID-19 primarily focuses on X-ray images, despite their lower precision compared to computed tomography (CT) scans. In this work, we leverage unstructured data from a hospital and harness the fine-grained details offered by CT scans to perform zero-shot multi-label classification based on contrastive visual language learning. In collaboration with human experts, we investigate the effectiveness of multiple zero-shot models that aid radiologists in detecting pulmonary embolisms and identifying intricate lung details like ground glass opacities and consolidations. Our empirical analysis provides an overview of the possible solutions to target such fine-grained tasks, so far overlooked in the medical multimodal pretraining literature. Our investigation promises future advancements in the medical image analysis community by addressing some challenges associated with unstructured data and fine-grained multi-label classification.

1. Introduction

Artificial intelligence (AI) research in the medical domain throughout the pandemic prioritized applying supervised learning methods to help hospitals diagnose or triage patients faster. Whilst models were developed at a fast pace, the majority of these were deemed not fit for clinical use [24]. Despite these setbacks, the pandemic led to the generation and collection of substantial medical imaging data, among others. AI is still considered a high-quality strategy to assist in diagnosis, severity assessment, and prognosis of long COVID-19 [22]. The large amounts of unlabelled data from the pandemic allow the possibility of exploring self-supervised learning models to be developed. One of the most popular methods is contrastive visual language pertaining (CLIP), which enables training on pairs of images and text with zero-shot capabilities [1]. CLIP eliminates the need for precisely annotated datasets and learns representation from noisy image-text pairs, potentially resulting in significant time and cost savings.

Applying self-supervised deep learning methods to open-source data has rapidly gained popularity in medical and non-medical domains [41, 38, 11, 49]. The current success of self-supervised learning, particularly contrastive,

can be attributed to preventing dimension collapse through aggressive data augmentation and negative pairs [25] and utilizing large datasets. For instance, CLIP training from scratch was enabled by access to 400 million pairs of images and texts crawled from the web. Contrastive methods have been highly valuable in downstream tasks like image classification, enabling the creation of competitive representations comparable to fully supervised networks [26, 12].

The medical domain does not have access to such large datasets since collecting data is costly and requires highly specialized human expertise [45, 43]. Despite this, there has recently been a successful application of multimodal contrastive pretraining techniques [49, 11, 37, 51]. Most of the aforementioned studies performed pretraining on X-rays [49, 11], which are easier to gather but less precise than other modalities, e.g., CT scans. Furthermore, given the smaller scale of biomedical datasets and the significant domain shift among different subdomains compared to natural images (e.g., X-rays to CT scans), it remains an open research question on how to adapt and fine-tune available pre-trained models properly [51].

This work focuses on fine-tuning pre-trained encoders via CLIP on CT images and uncurated radiology reports obtained during the pandemic. We investigate several issues deriving from fine-tuning on a different domain, such as the data preprocessing of large volumetric CT scans and long unstructured reports. Furthermore, in collaboration with expert radiologists, we establish a fine-grained multi-label classification task that evaluates disease severity and identifies the presence of five distinct characteristics commonly associated with COVID-19: pulmonary embolism, pneumonia, consolidation, infiltrates and ground glass opacities. The zero-shot classification task is particularly challenging due to the uncurated structure of the reports and the fine-grained nature of the task. To improve the correct matching across visual predictions and text targets, rather than keeping class-independent templates like standard practice [1], we design per-class templates.

We specifically focused on patients diagnosed with COVID-19, as we aim to develop a valuable tool to aid radiologists in identifying individuals at the highest risk. More broadly, we hope that our empirical analysis could be helpful for researchers involved in deploying pre-trained models on different medical imaging domains by only exploiting uncurated data such as CT scans and corresponding radiology reports.

2. Related Work

Contrastive visual language learning. Contrastive learning [30, 41] has evolved to be used in multi-modal data pipelines as popularised in CLIP [1]. When applying CLIP, we are considering image and text modalities. Recent work has extended the modalities to six [34]. In essence, we are mapping multi-modal data to a unified latent space, where we calculate the similarities between different elements and learn models to map data across different modalities effectively. When looking at how to calculate how similar different modalities are, we first look at the contrastive loss presented in Radford *et al*. [1], which Sohn originally influenced [39] and then was further used in contrastive representation learning in Oord *et al*. [31]. An early attempt in CLIP radiology presented ConVIRT [49], which maps chest X-rays to reports. CLIP's mainstream success has resulted in further applications to the medical domain [36]. CheXzero builds upon Radford *et al*. [1] by finetuning the model to the radiology domain, successfully achieving human-like results without explicit labels during training [11]. Similarly, MedCLIP explores CLIP by applying alternative models to Open AI and CheXzero [51]. In particular, they employ Swin transformer [50] and BioClinicalBERT [14] as their respective vision and text encoders. They also perform ablation studies with a ResNet-50 [16] as the vision encoder, which provides the best results in zero-shot classification on COVID-19 and RSNA pneumonia X-ray datasets. More recently, Zhang *et al*. built one of the most extensive medical image-pair datasets and obtained successful image-text/text-image retrieval results [37]. One significant key difference is the increase in context length from 77 to 256. Contrastive learning has also been successfully applied to lung CT images [48, 40]. However, there is currently a lack of extensive literature exploring the application of this technique specifically to CT images and their corresponding reports. Multi-modal learning provides opportunities to provide interesting insights into tasks such as zero-shot learning.

COVID-19. Mohit *et al*. [29] use convolutional neural networks (CNNs) as encoders to build a computer-aided diagnosis system for COVID-19 to assist radiologists in early diagnosis. Building upon this, transformers in supervised diagnosis, as seen in [47], can also be used, as shown in Dong *et al*. [9]. Moreso, this work demonstrates strong results on public datasets by extracting the relevant features from 3D volumes. We see contrastive learning used to build a severity system based on electronic health records (EHRs) [46]. Providing good results, structured data like this is often difficult to obtain, and we primarily focus more on unstructured data such as text.

Transformers. Transformers have emerged as dominant models for natural language processing (NLP) and computer vision, largely due to their effective utilization of attention mechanisms [10, 28]. In NLP settings, self-attention compares word embeddings to capture the relevance and importance of each word in the context [6]. A popular
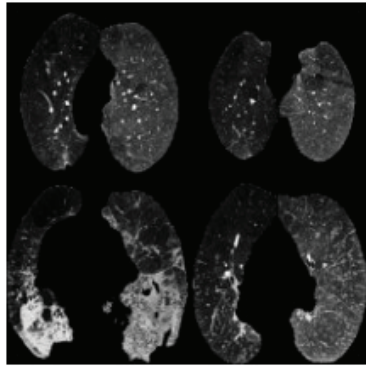
Figure 1. An example of the CT montage, four random slices taken from the CT image. The text is a subsection of the radiology report.

choice, BERT [18], was recently fine-tuned on general and COVID-19-related radiology reports [4, 32]. Influenced by text transformers, vision transformers have been argued to be more efficient in training than traditional CNNs [2]. They divide images into fixed-size patches and utilize self-attention to capture the relationships between them. This study considers three different vision transformers, ViT-B/16, ViT-B/32 [2] and Swin-Transformer [50]. The architectures in these two transformers differ. ViT operates on the entire image as a sequence of flattened patches, whereas Swin Transformer introduces a hierarchical structure of windows to process images. There have now been several implementations of vision transformers applied to medical images [13, 5].

3. Method

3.1. Data pre-processing

CT images are naturally large, given their multi-channel nature. To decrease the memory overhead and filter out unwanted noise [20], we employ a data pre-processing technique influenced by ILD diagnosis research [42]. In particular, the pre-processing steps for each CT scan are as follows:

- We first reduce the size along the axial dimension by 10% on either end. CT scans contain redundant or low information at the beginning and end of the image.

- We then spatially crop the images to ensure there is no additional space and we are focusing on the lung tissue.

- The remaining CT is split into 4 blocks of similar dimensions.

- A single slice is randomly selected from each block and concatenated to form a 4-image slice montage.

- Each 4-slice montage is resized to 224×224.

To speed up our data loading, we perform the aforementioned data pre-processing offline. The pre-processing was performed 10 times for each CT scan. Random slice selection can be considered a form of data augmentation. In Figure 1, we can see an example of the image-text pair. For our text pre-processing, we translate the whole report from German to English using the Google Translate API in Python[1]. Then we slice the radiology report to focus only on the lung parenchyma. Additionally, we filter the resulting text with filters taken from Clinical XLNet [23].

3.2. Encoders selection

Our dataset is considered small compared to previous radiology datasets like CheXpert [19], and MIMIC-CXR [3]. We compensate for the lack of data by building upon previous models specific to our task. We consider established medical-based CLIP methods whilst also experimenting with extracting their vision encoder and finetuning with the RadBERT model. The RadBERT model is pre-trained on 466 million tokens or 4.42M radiology reports [4]. RadBERT was further fine-tuned in a COVID-19 investigation on 19,384 radiology reports this year in Chambon *et al*. [32], making the model weights publicly available. The prior representations of the text data learned in this model suit this study.

The choice of vision encoders is very large, but we are interested in those exposed to relevant data and tasks. There is an observation of previous studies opting to use either vision transformers or standard CNNs such as ResNet-50 [21]. Precisely, we consider the following vision encoders:

- **ResNet-50 [21].** Initialization with pre-trained weights on ImageNet [35] is standard practice in computer vision. We set it up to compare to medical domain-trained models.

- **CheXzero [11].** We initialize the CLIP model [1] with publically available weights trained on the CheXpert and MIMIC-CXR datasets. The results in this paper are very promising in a multi-label task. We are only changing the image modality input. The text data is related to lung conditions specific to our radiology reports.

- **MedCLIP [51].** It considers both a ViT [2] and ResNet-50 [21] trained on their image-text dataset. The datasets picked are relevant to our task. Using

---

[1]The reports are originally in German, given that the hospitals are in the German speaking part of Switzerland.
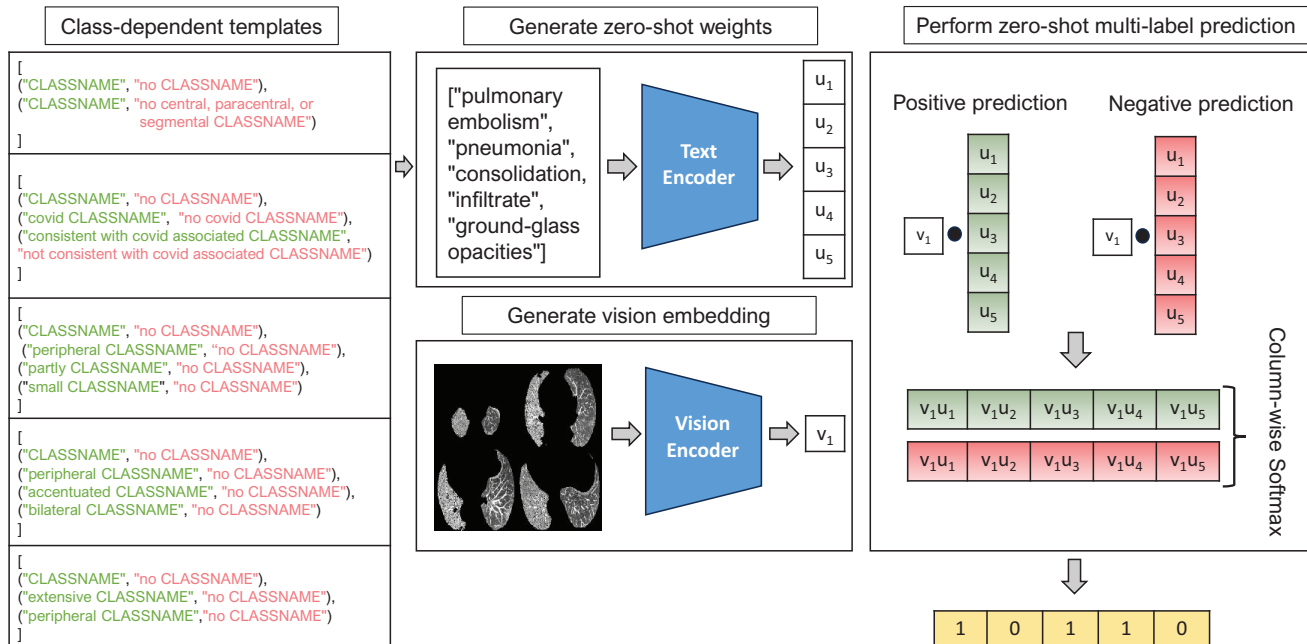
Figure 2. **Zero-shot pipeline.** In the left part, we see the five boxes which represent each class' list of template pairs. For a given class we iterate through its template pairs to generate the postive-negative zero-shot weights. Multiplying the respective CT montage embedding with both positive-negative zero-shot weights gives us a similarity. The last stage is calculating the softmax of these two predictions to result in our final prediction vector.

alternative encoders, we are interested in the results compared to CheXzero.

- **BiomedCLIP [37].** The BioMedCLIP set-up is attractive to us due to the size of the specific dataset it has been trained on. They have also increased their context length when training text encoders.

- **COVID-ViT [47].** A custom ViT transformer used for COVID-19 diagnosis in CT slices. The model has learned to extract relevant features based on the pathological tissue in each CT slice. If correctly aligned, the vision encoder should successfully map the extracted image patterns to the relevant text features.

When considering BiomedCLIP, MedCLIP, and CheXzero, we explore the transfer without any adaptation, i.e., frozen encoders, the training of their vision and text encoders, and the extraction of only their vision encoder and training with the RadBERT text encoder. We also train the ResNet-50 and COVID-ViT with the RadBERT text encoder. When training with the RadBERT, we adapt each vision encoder to map the text output shape.

3.3. Embeddings alignment

Contrastive visual language training requires image and text embeddings to be mapped to the same latent space.

Closely following the forward pass in CLIP [1], we align the output embeddings of the text and vision encoder by calculating the logits of each modality and passing these into separate cross-entropy loss functions. During training, given a batch of B input pairs $(\mathbf{x}_v, \mathbf{x}_u)$, we calculate their respective representation pairs $(\mathbf{v}, \mathbf{u})$ by feeding them into each respective encoder. We use $(\mathbf{v_i}, \mathbf{u_i})$ to denote the i-th pair. The first loss function is an image-to-text contrastive loss for the i-th pair:

$$\ell_i^{(v \to u)} = -\log \frac{\exp\left(\langle \mathbf{v}_i, \mathbf{u}_i \rangle / \tau\right)}{\sum_{k=1}^{B} \exp\left(\langle \mathbf{v}_i, \mathbf{u}_k \rangle / \tau\right)},$$

while similarly, the text-to-image loss:

$$\ell_i^{(u \to v)} = -\log \frac{\exp\left(\langle \mathbf{u}_i, \mathbf{v}_i \rangle / \tau\right)}{\sum_{k=1}^{B} \exp\left(\langle \mathbf{u}_i, \mathbf{v}_k \rangle / \tau\right)}$$

where $\langle \mathbf{v}_i, \mathbf{u}_i \rangle$ represents the cosine similarity and $\tau \in \mathbb{R}^+$ represents a temperature parameter. The temperature parameter controls the range of the logits in the stated losses and the strength of penalties on hard negative samples [15]. To calculate the overall loss, we calculate the average of the losses. Precisely, we add them together and divide them by the number of modalities, i.e., two.

Figure 3. **Word Clouds.** Word cloud generated to assist in class decisions (left). Word cloud with class names removed to assist in template selection (right).

3.4. Zero-shot multi-label classification

Zero-shot methods have gained traction recently, popularised by seminal papers such as Socher *et al.* [33] and Radford *et al.* [1]. The difficulty in evaluating our models lies within the embedding space of our text data. Previous works have utilized shorter, curated text data. This allows the design of simple prompts that match the training distribution, e.g., "A picture of a CAT". However, our report data comes directly from radiologists and reflects a realistic medical setting. To overcome this challenge, we employ prompt-based engineering, as seen in the latest approaches [27, 11, 1]. In collaboration with human experts, we analyze the radiology reports to create sensible classes for the dataset. We first create a word cloud (Figure 3, left) on the text data to visualize the most common words. We settle on the following classes: pulmonary embolism, pneumonia, consolidation, infiltrates and ground glass opacities. Confirming these are sensible and useful classes for radiologists, the human experts label the testing samples for us. Pulmonary embolism is the least common class, diagnosed in approximately one in ten patients. The prevalence of the other classes ranges from 65% to 80%. These classes lay unstructured in the text data we are training on and remain hidden from the vision encoders as in traditional zero-shot learning [44]. For a given class, our prompt is a positive-negative template paired with the word CLASS-NAME, which is replaced by the class. The choice of templates required manual analysis of the reports to estimate what prompts would be a good choice. To justify our template choice, we removed the class names from the text data and generated a word cloud visible in Figure 3 (right). Using this and with the manual reading of the reports, we identify prompts which occur with our class names; for example, the phrase "bilateral infiltrates" creates our prompt "bilateral CLASSNAME" for the class infiltrates.

Following Tiu *et al.* [11], to map the models' prediction to probabilities, we use a softmax layer for each template pair. Instead of using the same template pairs for each class, we propose each class has a list of its template pairs. This

is visualized in Figure 2. More in detail:

- We iterate over each class. In each class, we iterate over the template pairs. For a given template, we substitute the class into the template (e.g., no pulmonary embolism) and pass it into the text encoder. We normalize the embeddings and concatenate the results to get our zero-shot weights. We are left with a 5D vector which is our zero-shot weights for each class.

- We pass the respective CT montage into the vision encoder and normalize the embeddings.

- We calculate the cosine similarity of the resulting embeddings by multiplying the zero-shot weights and the vision embeddings.

- We estimate the class by calculating the softmax of the positive-negative predictions.

We do not scale by the temperature parameter τ in the zero-shot evaluation. We do not include this as we are relying on the representations learned by our encoders. The parameter is no longer needed to control the range of the logits. As a CT montage can have more than one label, we solve a multi-label zero-shot problem by performing binary classification for each class.

Variable	All patients
Age (years $\pm$ SD)	61.4 $\pm$ 14.2
Male / Female	276 / 84
BMI ($\frac{Kg}{m^2} \pm$ SD)	27.9 $\pm$ 8.5
Data split (Training / test)	368 / 92

Table 1. Descripive statistics of the patient cohort.

4. Experimental Settings

4.1. Dataset

After the ethics approval, the study collected data from University Hospital Zurich in Switzerland, resulting in over

Model	Encoders	Fine-tuned	Context length	Macro Avg. F1 ($\uparrow$)	HL ($\downarrow$)	Sub. Acc. ($\uparrow$)
CheXzero [11]	ViT-B/32	GPT2	✗	77	0.43	0.54	0.00
CheXzero [11]	ViT-B/32	GPT2	✓	77	0.71	0.29	0.23
MedCLIP [51]	Swin-T	BCB	✗	77	0.62	0.45	0.04
MedCLIP [51]	Swin-T	BCB	✓	77	0.51	0.48	0.12
BioMedCLIP [37]	ViT-B/16	PMB	✗	256	0.33	0.49	0.06
BioMedCLIP [37]	ViT-B/16	PMB	✓	256	0.61	0.44	0.09

Table 2. **Frozen vs fine-tuned encoders**. Comparison between the frozen and fine-tuned models. HL stands for the Hamming loss, and Sub. Acc. for subset accuracy. PMB, BCB respectively, denote PubMedBERT and BioClinicalBERT.

460 image-text pairs from individual patients taken during the COVID-19 pandemic. A small summary of the patient cohort can be seen in Table 1. CT scans were taken on either Siemens SOMATOM Definition AS, SOMATOM Definition Flash, or SOMATOM Force machines. We only consider thorax CT images taken in a single session. After resampling the CT images and preparing the montages (Section 3.1), we are left with 13,240 montages and 460 reports. We split the dataset 80:20 based on patient ID, so montages from the same CT are not be seen in training and evaluation. We trained on 11,010 montages along with their respective reports. When evaluating our test set, we chose a single montage-report pair for each patient and performed the zero-shot evaluation.

4.2. Training details

In all training scenarios, we set a maximum of 100 epochs, a batch size of 100, and use the AdamW optimizer [17] with betas of (0.9, 0.98). When fine-tuning existing methods (CheXzero, MedCLIP, BioMedCLIP), we adapt the hyperparameters to achieve the best results. Specifically, we set the learning rate to [5e-6, 5e-5, 5e-5] and the weight decay to [1e-4, 1e-4, 1e-3], respectively. For training Rad-BERT and alternative vision encoders, we fix the learning rate at 5e-5 and the weight decay at 1e-3.

We implemented offline image data preparations to accelerate data loading. At the bottom of the image processing pipeline described in Section 3.1, we convert montages to PIL images and tensors. During training, we use a single NVIDIA RTX A6000 GPU to train the encoders.

4.3. Zero-shot evaluation

To assess the zero-shot capabilities, we consider calculating the macro average F1 score, Hamming loss, and the subset accuracy. The F1 score is defined as

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

Where precision and recall are respectively defined as $\frac{TP}{TP+FP}$ and $\frac{TP}{TP+FN}$. The macro average F1 gives us the

average of the F1 over all classes.

Hamming loss is often considered in multi-label evaluation. The Hamming loss measures the fraction of instances where a model's predictions do not equal the true labels. It is obtained by dividing the number of incorrect predictions by the total number of instances and classes.

$$HL = \frac{1}{NL} \sum_{l=1}^{L} \sum_{i=1}^{N} Y_{i,l} \oplus X_{i,l},$$

N is the total number of data samples and L is the total number of classes. $\oplus$ is Exlusive-OR, $X_{i,l}$ ($Y_{i,l}$) stands for boolean that the i-th index (prediction) is equal to the l-th label.

Lastly, the subset accuracy measures how accurately every sample is predicted. For instance, if our target array is [0, 1, 1, 0, 0], it is considered correct only if all elements of the vector are predicted accurately. The prediction [0, 1, 1, 1, 0] is hence considered wrong. This makes the subset accuracy the most challenging metric to fulfill.

Model	Templates	Macro Avg. F1	HL	Sub. Acc.
CheXzero	CI	0.55	0.47	0.05
	CD	0.71	0.29	0.23
RN50	CI	0.46	0.51	0.12
	CD	0.50	0.45	0.21
MedCLIP	CI	0.57	0.44	0.13
	CD	0.58	0.42	0.15

Table 3. **Class-independent (CI) vs. class-dependent (CD) templates**. CheXzero is the ViT-B/32 plus GPT2. The ResNet-50 (RN50) is pre-trained with ImageNet. MedCLIP is the Swin transformer. The RN50 and MedCLIP are coupled with the RadBERT text encoder with a context length of 100 and 200, respectively. All models are fine-tuned on our dataset.

5. Results

In this section, we empirically investigate the best training solution for reliably mapping a relatively small set of

Model	Vision Encoder	Context length	Truncation side	Macro Avg. F1 (↑)	HL (↓)	Sub. Acc. (↑)
CheXzero [11]	ViT-B/32	100 100 200 200	← → ← →	0.53 0.59 0.54 0.57	0.52 0.45 0.48 0.46	0.08 0.14 0.09 0.13
MedCLIP [51]	RN50	100 100 200 200	← → ← →	0.53 0.55 0.46 0.54	0.49 0.48 0.56 0.51	0.12 0.13 0.08 0.09
MedCLIP [51]	Swin-T	100 100 200 200	← → ← →	0.5 0.59 0.58 0.53	0.51 0.46 0.42 0.53	0.09 0.09 0.15 0.09
BioMedCLIP [37]	ViT-B/16	100 100 200 200	← → ← →	0.48 0.42 0.50 0.51	0.53 0.59 0.53 0.51	0.10 0.10 0.05 0.10
CovidViT [47]	ViT	100 100 200 200	← → ← →	0.51 0.55 0.49 0.54	0.49 0.45 0.52 0.49	0.10 0.10 0.05 0.10
–	RN50 IN-1k	100 100 200 200	← → ← →	0.51 0.50 0.53 0.45	0.54 0.49 0.51 0.54	0.06 0.21 0.12 0.08

Table 4. **Combining encoders**. Comparison between models with fixed text encoder (RadBERT) and multiple vision encoders. The left and right arrows, respectively, mean truncation from the end and beginning of reports. HL stands for the Hamming loss, and Sub. Acc. for subset accuracy. RN50 IN-1k refers to the ResNet-50 initialized from ImageNet pre-trained weights. PMB, BCB respectively, denote PubMedBERT and BioClinicalBERT.

CT images to corresponding radiology reports and consequently performing zero-shot predictions. As anticipated, we consider three approaches: 1) applying frozen baseline methods, 2) finetuning baseline methods, and 3) training an alternative text and vision encoder. We also study the impact of class-dependent and class-independent templates. We perform additional ablations considering the truncation side and context length.

5.1. Baselines and finetuning

For a comprehensive comparison, we consider three baselines that provide publically available weights applying contrastive visual language models in the medical domain. CheXzero [11], MedCLIP [51], and BioMedCLIP [37] are all available and are relevant to our task.

Frozen encoders. As baseline methods, we load the pre-trained models and apply straight-away the zero-shot evaluation on our test set without any modifications. All evaluated baseline models feature transformers. The results for the frozen encoders are shown in Table 2. We see the subset accuracies are extremely low, from a minimum of 0% to a maximum of 6%. However, considering none of the models has been exposed to CT montages before, we see the effect of large datasets does not compensate for the shift to our domain, and the models need specific finetuning.

Fine-tuned encoders. Next, we finetune each baseline

on our small dataset and again perform the evaluation. As expected, we observe an increase in the exact match performance (Table 2). MedCLIP and BioMedCLIP respectively gain 8 and 3 percent points. Notably, CheXzero has adapted extremely well to our dataset, increasing the subset accuracy from 0% to 23%. The other metrics are also the best we observe in this study, i.e., Hamming loss of 0.29 and macro average F1 score of 0.71. This finding suggests that previous pre-training on large chest X-ray datasets and their corresponding reports helps achieve better results in this fine-grained task. Unexpectedly, the macro average F1 and Hamming loss for MedCLIP decreases. We suspect that the concurrent decrease of these two metrics and the increase in subset accuracy imply a recognition improvement for a subset of the classes and a deterioration in others. We finally see a big increase in the macro average F1 score for BioMedCLIP but not so much in the other metrics.

Class-independent vs. class-dependent templates. In contrast to previous studies, each zero-shot class has its list of template pairs. This enables us better target vision features to specific prompts for a class. For example, the positive prompt for pneumonia was often "consistent with", whereas this prompt would not make sense for a different class like a pulmonary embolism. As shown in Table 3, CheXzero improves drastically in all three metrics, in particular with a subset accuracy increase of 18%, when apply-

ing class-dependent prompts. All previous results, shown in Table 2, are obtained with class-dependent templates.

5.2. Combining pre-trained models.

Investigating the combination of pre-trained models trained specifically on COVID radiology reports and chest image datasets is required to identify the best setups. All text encoders use the COVID-finetuned RadBERT, which has prior knowledge of radiological terms such as "pulmonary embolism" and "ground glass opacities". We also select the existing pre-trained CLIP models in the medical domain and extract their vision encoders. These vision encoders have all been exposed to chest X-rays, and some have even been exposed to single CT slices. Furthermore, we analyze a vision transformer used in COVID diagnosis and a pre-trained ResNet-50 on ImageNet. The difficulty of the task is acknowledged in the results achieved, receiving the best subset accuracy of 21% (Table 4). Interestingly, the latter is obtained with RN50, which has not been pre-trained on medical datasets.

We run further ablations considering text and multiple vision encoders. We experiment with class-dependent vs. class-independent templates, changing the context length and truncating from different sides.

Class-independent vs. class-dependent templates. Table 3 also shows positive trends for the top-two combinations in Table 4, those trained with RadBERT and an alternative vision encoder. With regards to subset accuracy, the ResNet-50 increases by 9% and MedCLIP marginally improves by 2%. All three metrics improve when applying class-dependent prompts.

Context length and truncation. Due to our uncurated radiology reports being longer than standard image-text pairs, we test varying context lengths equal to 100 or 200. Previous work in contrastive visual language learning deals with shorter text, so we explore increasing this to capture more information. Furthermore, we test whether truncating the tokens from the left or the right affects the performance. At the beginning of the reports, we can see the knowledge describing the lungs. At the end of the report, we see shorter statements that match our classes. We test whether more valuable representations can be learned from the beginning or end.

Our results show the best subset accuracy result with the shorter context length and truncating from the right. The best Hamming loss was recorded with a longer context length, truncated from the left, and still records a good subset accuracy. The average subset accuracy for context length 100 is 9.3%, and for 200, it is 9.41%. Equally, for truncation on the left, it is 7.8%, and for the right, it is 9.6%. We conclude that context length did not alter the results massively, but the truncation side worked better when truncating from the right.

6. Conclusions

In our paper, we investigated the development of a zero-shot tool that assists radiologists in detecting pulmonary embolisms and identifying intricate lung details, including ground glass opacities and consolidations, automatically.

To meet this goal, we collected uncurated COVID-19 CT scans and corresponding reports from a university hospital to ensure our study faces real-world scenarios. Secondly, we run tests based on image-text pretraining and fine-tuning for multi-label zero-shot classification. A clear challenge represented the variability of the text data and the consequent framing of the zero-shot targets. We partially addressed such an issue using a class-dependent zero-shot template scheme and pre-trained vision-text medical models. In parallel, we are in the process of curating data from three additional hospitals. This data will enable us to further develop and apply the described methods to longitudinal data, specifically for the prognosis of long COVID-19. In future work, we would like to see such techniques applied to 3D volumes to solve tasks such as disease prognostication and progression as discussed recently in review articles.[8, 7].

We hope our work inspires future research to meaningfully use data collected throughout the pandemic and improve the automatic identification of fine-grained lung pathological patterns.

7. Acknowledgements

References

[1] Chris Hallacy Aditya Ramesh Gabriel Goh Sandhini Agarwal Girish Sastry Amanda Askell Pamela Mishkin Jack Clark Gretchen Krueger Ilya Sutskever Alec Radford, Jong Wook Kim. Learning transferable visual models from natural language supervision. 2021. 1, 2, 3, 4, 5

[2] Alexander Kolesnikov Dirk Weissenborn Xiaohua Zhai Thomas Unterthiner Mostafa Dehghani Matthias Minderer Georg Heigold Sylvain Gelly Jakob Uszkoreit Neil Houlsby Alexey Dosovitskiy, Lucas Beyer. An image is worth 16x16 words: Transformers for image recognition at scale. 2020. 3

[3] Seth J. Berkowitz Nathaniel R. Greenbaum Matthew P. Lungren Chih-ying Deng Roger G. Mark Steven Horng Alistair E. W. Johnson, Tom J. Pollard. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. 2019. 3

[4] Xing Lu Jiang Du Eric Y. Chang Amilcare Gentili Chun-Nan Hsu An Yan, Julian McAuley. Radbert: Adapting transformer-based language models to radiology. 2022. 3

[5] Rukhsana Zafar Huma Ameer Muhammad Ali Muhammad Moazam Fraz Arshi Parvaiz, Muhammad Anwaar Khalid. Vision transformers in medical computer vision—a contemplative retrospection. 2022. 3

[6] Niki Parmar Jakob Uszkoreit Llion Jones Aidan N. Gomez Lukasz Kaiser Illia Polosukhin Ashish Vaswani, Noam Shazeer. Attention is all you need. 2017. 2

[7] George PM Assayag D Glaspole I Mackintosh JA Corte TJ Glassberg M Johannson KA Calandriello L Felder F Wells A Walsh S. Barnes H, Humphries SM. Machine learning in radiology: the new frontier in interstitial lung diseases. *Lancet Digit Health.*, 2023. 8

[8] Fontanellaz M Brigato L Heverhagen JT Peters AA Huber AT Hoppe H Mougiakakou S Ebner L. Dack E, Christe A. Artificial intelligence and interstitial lung disease: Diagnosis and prognosis. *Invest Radiol.*, 2023. 8

[9] Zhang G Wei Z Zhai Y Lv G. Dong A, Liu J. Momentum contrast transformer for covid-19 diagnosis with knowledge distillation. *Pattern Recognit.*, 2023. 2

[10] Yoshua Bengio Dzmitry Bahdanau, Kyunghyun Cho. Neural machine translation by jointly learning to align and translate. 2014. 2

[11] Pujan Patel Curtis P. Langlotz Andrew Y. Ng Pranav Rajpurkar Ekin Tiu, Ellie Talius. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. 2020. 1, 2, 3, 5, 6, 7

[12] Kimberly Wilber Oisin Mac Aodha Serge Belongie Elijah Cole, Xuan Yang. When does contrastive visual representation learning work? 2022. 2

[13] Conrad Asotie Omonhinmin Emerald U. Henry, Onyeka Emebob. Vision transformers in medical imaging: A review. 2022. 3

[14] Willie Boag Wei-Hung Weng Di Jin Tristan Naumann Matthew B. A. McDermott Emily Alsentzer, John R. Murphy. Publicly available clinical bert embeddings. *arXiv:1904.03323*, 2019. 2

[15] Huaping Liu Feng Wang. Understanding the behaviour of contrastive loss. 2021. 4

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[17] Frank Hutter Ilya Loshchilov. Decoupled weight decay regularization. *arXiv:1711.05101*, 2017. 6

[18] Kenton Lee Kristina Toutanova Jacob Devlin, Ming-Wei Chang. Bert: Pre-training of deep bidirectional transformers for language understanding. 2018. 3

[19] Michael Ko Yifan Yu Silviana Ciurea-Ilcus Chris Chute Henrik Marklund Behzad Haghgoo Robyn Ball Katie Shpanskaya Jayne Seekins David A. Mong Safwan S. Halabi Jesse K. Sandberg Ricky Jones David B. Larson Curtis P. Langlotz Bhavik N. Patel Matthew P. Lungren Andrew Y. Ng Jeremy Irvin, Pranav Rajpurkar. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. 2019. 3

[20] Jeanny Pan Sebastian Rohrich Helmut Prosch-Georg Langs Johannes Hofmanninger, Florian Prayer. Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem. 2020. 3

[21] Shaoqing Ren Jian Sun Kaiming He, Xiangyu Zhang. Deep residual learning for image recognition. 2015. 3

[22] Kaur P. Kaur J. Outbreak covid-19 in medical image processing using deep learning: A state-of-the- art review. 2020. 1

[23] Sitong Chen Edward T. Moseley Chih-ying Deng Naomi George Charlotta Lindvall Kexin Huang, Abhishek Singh. Clinical xlnet: Modeling sequential clinical notes and predicting prolonged mechanical ventilation. 2019. 3

[24] Gary S Collins Laure Wynants, Ben Van Calster. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. 2020. 1

[25] Yann LeCun Yuandong Tian Li Jing, Pascal Vincent. Understanding dimensional collapse in con- trastive self-supervised learning. 2022. 2

[26] Timothy M. Hospedales Linus Ericsson, Henry Gouk. How well do self-supervised models transfer? 2021. 2

[27] Ming Y. Lu, Bowen Chen, Andrew Zhang, Drew F. K. Williamson, Richard J. Chen, Tong Ding, Long Phi Le, Yung-Sung Chuang, and Faisal Mahmood. Visual language pretrained multiple instance zero-shot transfer for histopathology images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19764–19775, June 2023. 5

[28] Christopher D. Manning Minh-Thang Luong, Hieu Pham. Effective approaches to attention-based neural machine translation. 2015. 2

[29] Kumar Mohit, Rajeev Gupta, and Basant Kumar. Self-supervised contrastive learning for covid-19 classification from computed tomography images. In *2022 IEEE 9th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*, pages 1–5, 2022. 2

[30] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2

[31] Li Y. Oord, A. v. d. and O. R Vinyals. Representation learning with contrastive predictive coding. 2018. 2

[32] Curtis P. Langlotz Pierre Chambon, Tessa S. Cook. Improved fine-tuning of in-domain transformer model for inferring covid-19 presence in multi-institutional radiology reports. 2019. 3

[33] Hamsa Sridhar Osbert Bastani Christopher D. Manning Andrew Y. Ng Richard Socher, Milind Ganjoo. Zero-shot learning through cross-modal transfer. 2013. 5

[34] Zhuang Liu Mannat Singh Kalyan Vasudev Alwala Armand Joulin Ishan Misra Rohit Girdhar, Alaaeldin El-Nouby. Imagebind: One embedding space to bind them all. 2023. 2

[35] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 2015. 3

[36] Christoph Meinel Sedigheh Eslami, Gerard de Melo. Does clip benefit visual question answering in the medical domain as much as it does in the general domain? 2021. 2

[37] Naoto Usuyama Jaspreet Bagga Robert Tinn Sam Preston Rajesh Rao Mu Wei Naveen Valluri Cliff Wong Matthew P. Lungren Tristan Naumann Hoifung Poon Sheng Zhang, Yanbo Xu. Large-scale domain-specific pretraining for biomedical vision-language processing. 2023. 2, 4, 6, 7

[38] Malte Jensen Matthew P. Lungren Serena Yeung Akshay S. Chaudhari Shih-Cheng Huang, Anuj Pareek. Self-supervised learning for medical image classification: a systematic review and implementation guidelines. 2023. 1

[39] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. 2016. 2

[40] Kensaku Mori Akinobu Shimizu Yoshito Otake Masahiro Hashimoto Toshiaki Akashi Kazuhiro Hotta Sota Kato, Masahiro Oda. Classification and visual explanation for covid-19 pneumonia from ct images using triple learning. 2022. 2

[41] Mohammad Norouzi Geoffrey Hinton Ting Chen, Simon Kornblith. A simple framework for contrastive learning of visual representations. 2020. 1, 2

[42] Silva M Sverzellati N. Walsh SLF, Calandriello L. Deep learning for classifying fibrotic lung disease on high-resolution computed tomography: a case-cohort study. 2018. 3

[43] Linda Wang, Zhong Qiu Lin, and Alexander Wong. Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. *Scientific reports*, 2020. 2

[44] Wei Wang, Vincent Wenchen Zheng, Han Yu, and Chunyan Miao. A survey of zero-shot learning. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10:1 – 37, 2019. 5

[45] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017. 2

[46] Tingyi Wanyan, Hossein Honarvar, Suraj K. Jaladanki, Chengxi Zang, Nidhi Naik, Sulaiman Somani, Jessica K. De Freitas, Ishan Paranjpe, Akhil Vaid, Jing Zhang, Riccardo Miotto, Zhangyang Wang, Girish N. Nadkarni, Marinka Zitnik, Ariful Azad, Fei Wang, Ying Ding, and Benjamin S. Glicksberg. Contrastive learning improves critical event prediction in covid-19 patients. *Patterns*, 2(12):100389, 2021. 2

[47] Alice Gao Xiaohong Gao, Yu Qian. Covid-vit: Classification of covid-19 from ct chest images based on vision transformer models. 2021. 2, 4, 7

[48] Xu L Li Q Che F Xu Q Cheng X Liu M Yang M Wang X Zhang F Song B Liu M. Xie T, Wei Y. Self-supervised contrastive learning using ct images for pd-1/pd-l1 expression prediction in hepatocellular carcinoma. 2023. 2

[49] Yasuhide Miura Christopher D. Manning Curtis P. Langlotz Yuhao Zhang, Hang Jiang. Contrastive learning of medical visual representations from paired images and text. 2020. 1, 2

[50] Yue Cao Han Hu Yixuan Wei Zheng Zhang Stephen Lin Baining Guo Ze Liu, Yutong Lin. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv:2103.14030*, 2021. 2, 3

[51] Dinesh Agarwal Jimeng Sun Zifeng Wang, Zhenbang Wu. Medclip: Contrastive learning from unpaired medical images and text. 2022. 2, 3, 6, 7