TIER: Text-Image Entropy Regularization for CLIP-style models

Anil Palepu *apalepu@mit.edu*
Harvard-MIT Health Sciences & Technology

Andrew L. Beam *andrew_beam@hms.harvard.edu*
Harvard University

Abstract

In this paper, we study the effect of a novel regularization scheme on contrastive language-image pre-trained (CLIP) models. Our approach is based on the observation that, in many domains, text tokens should only describe a small number of image regions and, likewise, each image region should correspond to only a few text tokens. In CLIP-style models, this implies that text-token embeddings should have high similarity to only a small number of image-patch embeddings for a given image-text pair. We formalize this observation using a novel regularization scheme that penalizes the entropy of the text-token to image-patch similarity scores. We qualitatively and quantitatively demonstrate that the proposed regularization scheme shrinks the text-token and image-patch similarity scores towards zero, thus achieving the desired effect. We demonstrate the promise of our approach in an important medical context where this underlying hypothesis naturally arises. Using our proposed approach, we achieve state of the art (SOTA) zero-shot performance on all tasks from the CheXpert chest x-ray dataset, outperforming an unregularized version of the model and several recently published self-supervised models.

1 Introduction

Self-supervised vision models that leverage paired text data such as the contrastive language-image pre-trained (CLIP) model (Radford et al., 2021; Zhang et al., 2020) have demonstrated very impressive zero-shot classification performance in a variety of domains (Radford et al., 2021; Tiu et al., 2022; Boecking et al., 2022; Palepu & Beam, 2022). Specifically, users can leverage the unified text and image embedding space for zero-shot classification by providing relevant text queries and assessing image embedding similarities (Radford et al., 2021; Tiu et al., 2022; Kumar et al., 2022).

The CLIP architecture consists of a vision encoder, typically a CNN (He et al., 2016) or vision transformer (**?**), and a text encoder, typically a text transformer (Vaswani et al., 2017). Each encoder produces a single embedding in the joint embedding space that aims to summarize all of the relevant information in their respective modality. A recent CLIP-style architecture from Boecking et al. (2022), which was built on chest x-ray (CXR) data, allows for a more fine-grained representation of images by projecting the final ResNet block's output to the joint embedding space prior to doing a global average pooling. As a result, this model produces a set of local or *patch* embeddings which can be indicative of not just *if* a text and image align, but also *roughly where* they align. As an example, in a CXR positive for cardiomegaly (an enlarged heart), the patch embeddings in the heart area should have a higher cosine similarity than unrelated regions to the text embedding of a description of cardiomegaly.

In certain domains, CXRs included, it is clear that important visual features tend to be fairly localized, often being confined to a relatively small portion of the image. Furthermore, distinct concepts in the caption text likely correspond to different regions in the image, especially for more complex captions that may describe multiple image attributes. In this work, we propose a method to encode this observation into any

CLIP-style model that can produce image-patch embeddings and individual text-token embeddings. To do so, we introduce text-image entropy regularization (TIER), which encourages text-token embeddings and image-patch embeddings to be less 'promiscuous' by regularizing the entropy of a softmaxed distribution of similarity scores. This regularization can be modulated by adjusting two hyperparameters, and because it is based on entropy, it is robust to positional shifts in both the text and the images.

We demonstrate qualitatively and quantitatively that this regularization method shrinks the text-token and image-patch similarity scores towards zero. We evaluate the resulting model by comparing it to an equivalent unregularized baseline, a fully-supervised baseline, and several state-of-the-art, CLIP-style CXR benchmarks (Tiu et al., 2022; Wang et al., 2022). We demonstrate that our method results in robust zero-shot accuracy improvement across a wide range of clinical tasks, setting a new state of the art in many instances.

In summary, we make the following contributions:

- A novel regularization scheme applicable to any CLIP-style model that produces local image and text embeddings. The regularization term shrinks the text-token and image-patch similarity scores to encourage sparser image-text similarities.
- We establish a new state of the art (SOTA) zero-shot classification AUC on the CheXpert test set, surpassing recently introduced self-supervised models and several previously published fully supervised ones.
- We also establish a new SOTA for average zero-shot classification AUC on the Padchest dataset, which measures fine-grained classification performance.

2 Methods

2.1 Data

We utilized the MIMIC-CXR-JPG (Johnson et al., 2019) dataset to train our models and the CheXpert (Irvin et al., 2019) and Padchest (Bustos et al., 2020) datasets to evaluate them.

The MIMIC-CXR dataset (Johnson et al., 2019) consists of 377,095 CXR samples from 65,379 different patients. Many patients have multiple radiological studies within the dataset, with a single study often containing both a frontal and lateral CXR view. These CXRs were evaluated by radiologists, who wrote detailed reports on the clinical findings they observed as well as a sentence or two describing their overall impression of the imaging. We extracted these *impression* sections from the radiology reports to use as the paired text for our image input. We dropped any samples that were missing this impression section, leaving us with a total of 319,446 CXR-impression pairs. We split these MIMIC-CXR image-text pairs into training and validation subsets (with approximately 90% training data) and ensured that no patient was represented in both subsets.

For evaluation, we utilized the separate CheXpert (Irvin et al., 2019) dataset with pre-defined validation and test splits which consisted of 234 and 668 CXRs respectively. These subsets of CheXpert have 14 different clinical labels, determined by consensus of 3 and 5 radiologists respectively. We benchmarked our models' thresholded predictions using labels from an additional 3 radiologists available in the CheXpert test set. For the purposes of our evaluation, we only considered the following 5 clinical labels: Cardiomegaly, Edema, Consolidation, Atelectasis, and Pleural Effusion. These labels were the five competition tasks from the CheXpert competition and the most commonly attempted tasks in the literature, making them a natural set for comparison. We also extracted these labels from the MIMIC-CXR dataset, but we only used them when training our fully supervised CNN baseline; our contrastive models did not have any access to these labels.

We additionally evaluated our models with the Padchest dataset (Bustos et al., 2020), of which we only considered the subset of 39,053 CXRs that were labeled by radiologists. There were over a hundred different labels present in these CXRs, but we focused on the set of 57 labels that were present with frequency of at least 50 in our selected subset, as was done by Tiu et al. (2022).

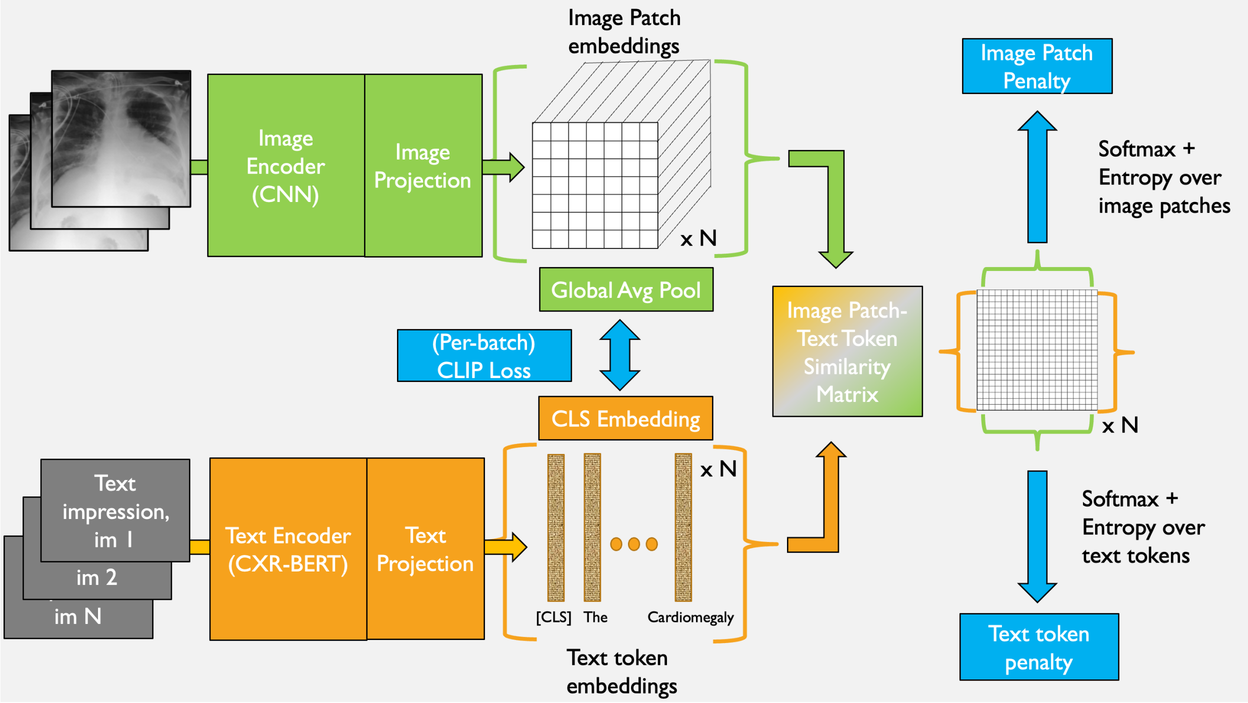

Figure 1: An overview of TIER, our regularized training method.

All images were resized to 224 × 224 pixels with 3 RGB channels. At train time, we performed random data augmentations including random resizing, cropping, affine transformation, and color jitter, while at test time, we simply resized images to 256 × 256 before center cropping to 224 × 224.

2.2 Model Architecture

We based our model on the BioViL architecture (Boecking et al., 2022), which consists of a pre-trained ResNet-50 architecture as the vision encoder and "CXR-BERT-specialized", a transformer, as the text encoder. This model differed from the original CLIP architecture in that it consisted of a radiology-specific text encoder (CXR-BERT-specialized) and was trained with an additional MLM loss, among several other changes (Boecking et al., 2022). This model was also trained using MIMIC-CXR and importantly did not have access to the CheXpert or Padchest datasets, which we used for evaluation.

For our purposes, the most critical feature of the BioViL model is that the final ResNet-50 block provides embeddings that correspond to local, connected regions of the input image (see the green path on the top of Figure 1). Thus, in addition to the single global image embedding, this model also produces a set of 49 embeddings in a 7 × 7 grid, which all share the same joint feature space as the global image embedding. The number of embeddings is a function of the original input size (a larger image input would yield more embeddings) as well as the choice to use the final ResNet block output (using an earlier output would lead to more fine-grained local embeddings). A single multi-layer perceptron with one hidden layer was used to project each local embedding to the joint feature space. We call the output of the ResNet block the *patch* embeddings as they correspond to regional patches of the image.

The text transformer naturally produces a text token embedding for each input token to the model. We use a single multi-layer perceptron with one hidden layer to project these text token embeddings to the joint feature space. The projected embedding from the first text token, [CLS], is contrasted with the global image embedding as is done with typical contrastive language-image pre-trained models (Radford et al., 2021; Palepu & Beam, 2022; Zhang et al., 2020). For a training batch of image-text pairs (x^I, x_T), we use the standard CLIP loss as described in Radford et al. (2021). We add additional penalty terms, described in the following section, to regularize our model beyond this standard CLIP loss. The pseudocode for our method is described in Fig. 2

```
def regularized_loss(Ims, Txts, lambda_patch, lambda_token):
    # image_encoder - ResNet-50, text_encoder - CXR-BERT-specialized
    # Ims[n, h, w, c], Txts[n, l] - minibatch of aligned images & texts
    # W_i[d_i, d_e] - learned projections of image patches to embed
    # W_t[d_t, d_e] - learned projections of text tokens to embed
    # P - number of image patches, T - number of text tokens

    # Setup; compute patch and token representations
    patch_f = image_encoder(Ims) #[n, d_i, P]
    token_f = text_encoder(Txts)  #[n, d_t, T]
    # project to joint embedding space [d_e] and normalize
    patch_e = l2_normalize(dot(patch_f, W_i), axis=1) #[n, d_e, P]
    token_e = l2_normalize(dot(token_f, W_t), axis=1) #[n, d_e, T]

    # CLIP Loss; Compute global embeddings
    image_e = l2_normalize(mean(patch_e, dim=2), axis=1) #[n, d_e]
    text_e = token_e[:, :, 0] #[n, d_e]
    # Compute scaled pairwise cosine similarity matrix
    clip_logits = dot(image_e, text_e.T) * exp(t) #[n, n]
    # Evaluate symmetric CLIP loss function
    labels = np.arange(n)
    loss_i = cross_entropy_loss(clip_logits, labels, axis=0)
    loss_t = cross_entropy_loss(clip_logits, labels, axis=1)
    clip_loss = (loss_i + loss_t)/2

    # Regularization; Compute patch-token similarity matrix
    sim_matrix = batch_multiply(token_e, patch_e) #[n, T, P]
    # Compute patch and token penalties
    patch_entropies = entropy(softmax(sim_matrix, axis = 2)) #[n, T]
    patch_penalty = lambda_patch * mean(patch_entropies)
    token_entropies = entropy(softmax(sim_matrix, axis = 1)) #[n, P]
    token_penalty = lambda_token * mean(token_entropies)

    regularized_loss = clip_loss + patch_penalty + token_penalty
    return regularized_loss
```

Figure 2: Pseudocode for our TIER regularization method

2.3 TIER: Text-Image Entropy Regularization of Image-Patch and Text-Token Similarity Scores

TIER works by first computing a matrix of image-patch and text-token similarities. Specifically, consider an example image-text pair that has $I = \{I_1, \dots, I_P\}$ image-patch embeddings (in our case, $P = 49$), and has $T = \{T_1, \dots, T_T\}$ text-token embeddings (T can vary for each sample as captions can be different lengths). We compute the image patch-text token similarity matrix S by computing a $T \times P$ matrix of cosine similarities between each image-patch embedding and text-token embedding. The embeddings for each input modality are the outputs of an encoder model that is specific to that input, e.g. a CNN or vision transformer for images and a BERT-style transformer for text. Importantly, we select encoders that provide embeddings at the token level, i.e., image-patch embeddings and text-token embeddings. Row i of S indicates the cosine similarity between a text-token T_i and each image patch in I. The columns of S likewise indicate the cosine similarity between a given image patch I_j and each text token in T.

Recall, the goal of our approach is to shrink the elements of S such that each text token is similar to a relatively small number of image patches. To do this, we introduced an entropy-based penalty term that induces shrinkage on the elements of S. First, we perform a row-wise softmax of S and measure the entropy between a text token T_i and all of the image patches in I, shown below:

$$\mathcal{H}(T_i, I) = \sum_{j=1}^{|P|} -p_j * \log(p_j) \tag{1}$$

where p_j is the probability produced by the softmax of the row of S corresponding to T_i. This term will be maximized when each p_j is $\frac{1}{P}$, implying that all of the image patch embeddings have equal similarity to T_i.

Next, we apply the same procedure to the columns of S, applying a column-wise softmax over the text-token similarities to produce probabilities p_1 to p_T for each image patch I_i and calculating the entropy of these probabilities as follows:

$$\mathcal{H}(T, I_j) = \sum_{i=1}^{|T|} -p_i * \log(p_i) \tag{2}$$

We average the $N \times T$ image-patch entropies $\mathcal{H}(T_i, I)$ and the $N \times P$ text-token entropies $\mathcal{H}(T, I_j)$ to produce an *image-patch penalty* and *text-token penalty* for the batch. We control the effects of these penalties on training by weighting them with hyperparameters λ_t and λ_p respectively, adding the weighted penalties to the CLIP loss to compute the total loss as described by Fig. 2.

A random search over the range $(0, 1)$ was used to set the hyperparameters ($\lambda_p = 0.2$, $\lambda_t = 0.1$) for our regularized model. Specifically, we trained our contrastive models for just a single epoch on MIMIC-CXR with 20 pairs of randomly chosen λ_p and λ_t, and chose the pair that maximized zero-shot AUC on the validation set. Both the training procedure and zero-shot classification method are described in later sections.

2.4 Training Details

We begin with the pretrained BioViL architecture and model weights. We use their specialized model, which has already been trained with contrastive learning on the MIMIC-CXR dataset, although in this original training, only frontal images were used. As described below, we retrained an unregularized version of this model to include both frontal and lateral views.

We train two separate CLIP-style models: A regularized model in which $\lambda_p = 0.2$ and $\lambda_t = 0.1$, as well as an unregularized baseline model, in which $\lambda_p = \lambda_t = 0$. All other aspects of model training are identical between these two models. For both models, we freeze the first 8 layers of the BERT encoder, while leaving the rest of the text encoder and vision encoder unfrozen. Each model is trained for 30 epochs using the loss described in the previous section with a learning rate of 0.0001 and batch size of 32.

We also train an additional fully supervised CNN baseline, which utilizes the same vision encoder as the contrastive models but has a multilayer perceptron with one hidden layer and five output nodes. This supervised baseline still uses MIMIC-CXR for training, but instead of text, it is trained with actual labels using binary-cross entropy loss with a learning rate of 0.0001 and batch size of 32.

2.5 Zero-shot classification

We employ a zero-shot classification procedure that leverages our text and image encoders to identify labels of interest in the images. Our method begins with the user selecting K_p positive and K_n negative queries for the label of interest, Q. Positive queries $\{Q_{p1}, ..., Q_{pK_p}\}$ are text descriptions indicative of the presence of that label, while negative queries $\{Q_{n1}, ..., Q_{nK_n}\}$ are text descriptions indicative of the absence of that label; examples which we used for the five CheXpert labels are detailed in Tab. 3. We pass each positive query through the text encoder, project their [CLS] token embeddings to the joint embedding space, and then average these projected embeddings and re-normalize to a unit norm. We do the same for the negative queries so that we have a single positive $\overline{Q_p}$ and negative $\overline{Q_n}$ query embedding associated with each label that we wish to classify:

$$\overline{Q_p} = \frac{\sum_{j=1}^{K_p} Q_{pj}/K_p}{||\sum_{j=1}^{K_p} Q_{pj}/K_p||} \qquad \overline{Q_n} = \frac{\sum_{j=1}^{K_n} Q_{nj}/K_n}{||\sum_{i=1}^{K_n} Q_{nj}/K_n||} \tag{3}$$

For any input image we wish to classify, we use the image encoder to compute its projected global image embedding E_{img} (normalized to unit norm) and take the dot product of this global image embedding with both the positive and negative query embeddings $\overline{Q_p}$ and $\overline{Q_n}$ for every label we wish to predict. We subtract these positive and negative cosine similarity scores to get a zero-shot classification score, Z_Q, for our label of interest.

$$Z_Q = (E_{img} \cdot \overline{Q_p}) - (E_{img} \cdot \overline{Q_n}) \tag{4}$$

Importantly, our zero-shot classification output is not a probability, and its range is actually between $[-2, 2]$. We are primarily interested in assessing discriminative performance of our zero-shot classifiers, so a probability is not necessary; however, if one desired a probability output, they could simply apply a softmax to the positive and negative similarity scores as was done by Tiu et al. (2022) instead of subtracting these scores.

3 Results

3.1 Visualization of the effect of regularization

Qualitatively, our regularization method is able to achieve the desired shrinkage between image patches and text tokens. Fig. 3 and Fig. 4 show patch-level zero-shot classification scores (i.e., the score between each image patch and the global [CLS] text token) overlaid on top of two CXRs, one with cardiomegaly and one without. In these heatmaps, red is indicative of a higher zero-shot score, gray is a neutral score, and blue is a lower (negative) zero-shot score. In both the regularized and unregularized models, we can clearly identify that the cardiomegaly-positive image correctly has a significant amount red on the heatmaps, suggesting a much higher cardiomegaly score. Similarly, both models correctly produce blue heatmaps for the cardiomegaly-negative image, indicating a lower cardiomegaly score.

Important differences between the regularized and unregularized models are apparent when we examine the distribution of blue and red regions of the heatmaps in Fig. 3 and Fig. 4. For the cardiomegaly-positive image (Fig. 3), the regularized model has high similarity primarily on the lower left side of the patient's chest (which corresponds to lower right side of the image), where their heart is located. Likewise, the regularized model shows low similarity (blue) on the lower right side of their chest where one could expect to see some changes in most extreme cases of cardiomegaly. These similarity scores seem rational and are clinically justifiable. On the other hand, while the unregularized model also displays some signal in the clinically relevant regions, it has significantly more extreme similarity scores scattered throughout the image even beyond the heart-adjacent regions.

3.2 Distribution of image-patch similarity scores to global [CLS] text token

To further explore the effect of our regularization method on the image-patch similarities, we utilize a set of 160 positive image-text pairs from MIMIC-CXR. In Fig. 5, we plot the similarity of the projected [CLS] token embeddings to the 49 image-patch embeddings from the corresponding image. In this figure, the image-patch similarities were ranked in descending order before being plotted, with error bars indicating the standard deviation across the 160 samples. We can see that the regularized model on average has significantly lower similarities to the patch embeddings than the unregularized model. To quantify the amount of regularization, we produced Fig. 6, which displays the same information as Fig. 5 except each patch similarity was normalized by dividing the similarity by the sum of all patch similarities in the entire image. In this plot, we can clearly see that the regularized model tends to have a few patches with relatively higher similarities to the [CLS] token embedding, and many with relatively lower similarities; this supports our hypothesis that our regularization scheme shrinks token-level similarity in the model.

3.3 Zero-shot classification

Next, we evaluated our zero-shot classification method for both the regularized and unregularized models on the held-out CheXpert test set. Our primary benchmark for these models is the 'Chexzero' model (Tiu et al., 2022), which recently achieved SOTA zero-shot AUC on this task. For our zero-shot models, we compute the zero-shot score using an ensemble of our five best model checkpoints on the CheXpert validation set (the Chexzero model uses ten checkpoints in its ensemble). We also evaluate another recent self-supervised model, MedCLIP, with the caveat that this model is not strictly zero-shot because the authors utilized clinical labels during their training process. Additionally, we evaluate a fully supervised CNN that uses our vision encoder with an additional classification head. These results can be seen in Tab. 1, which demonstrates that our regularized model achieves SOTA zero-shot AUC on these CheXpert competition tasks; for all five clinical labels, regularization seems to offer a modest but consistent bump in AUC performance.

Label	Regularized TIER	Unregularized Basemodel	Chexzero	MedCLIP	Fully Supervised CNN
Average	**0.90911**	0.90124	0.89801	0.87708	0.88877
Cardiomegaly	**0.91653**	0.90887	0.90327	0.83911	0.86400
Edema	**0.92483**	0.92376	0.92041	0.91242	0.92236
Consolidation	**0.90124**	0.89289	0.89896	0.88653	0.86003
Atelectasis	**0.87033**	0.86168	0.83546	0.79423	0.85869
Pleural Effusion	**0.93276**	0.91899	0.93193	0.95313	0.93876

Table 1: AUCs for various models on CheXpert Test. The highest zero-shot AUC is bolded (the three on the left are performing zero-shot classification, in which they have never previously seen any labels in the training set). Though MedCLIP is trained with contrastive learning, it also utilizes labels during training so we do not consider it to be fully zero-shot.

We also evaluate the zero-shot models against three reference radiologists using the Matthews's correlation coefficient (MCC) and F1 score, which can be seen in Fig. 7. For all three zero-shot models, we selected optimal thresholds for MCC and F1 using their performance on the CheXpert validation set. On the test set, our regularized model outperforms the unregularized model and Chexzero model on average and for a majority of the labels, though it performs worse for consolidation and cardiomegaly F1. All three zero-shot models are somewhat competitive with the radiologists; our regularized model outperforms one of the three radiologists in terms of average MCC and F1 score.

Finally, we use the Padchest dataset to evaluate a broader set of findings, specifically looking at the 57 findings with $n \geq 50$ from the radiologist-labeled subset of Padchest. We constructed positive label queries using the phrase "X is present.", while we constructed negative label queries with the phrase "No X.", replacing X with the label of interest. The notable exception was when we classified "normal" images; in this instance, we used "Abnormal findings." as the negative query. Tab. 2 details the Padchest results for the regularized, unregularized, and Chexzero models. For the majority of Padchest findings, our regularized model outperforms both the unregularized baseline and Chexzero.

3.4 Zero-shot COVID-19 diagnosis

We wished to evaluate our model on a task like COVID-19 detection, since this is a newer diagnosis not present in any of our training data. As a result, our models would not be able to rely on the actual label itself (i.e., the word cardiomegaly in a text query), and therefore the diagnostic capability of our models could be fully attributed to their ability to recognize the descriptive attributes we were querying. Furthermore, discriminating COVID-19 and non-COVID-19 pneumonia from chest imaging is a non-trivial task, with one

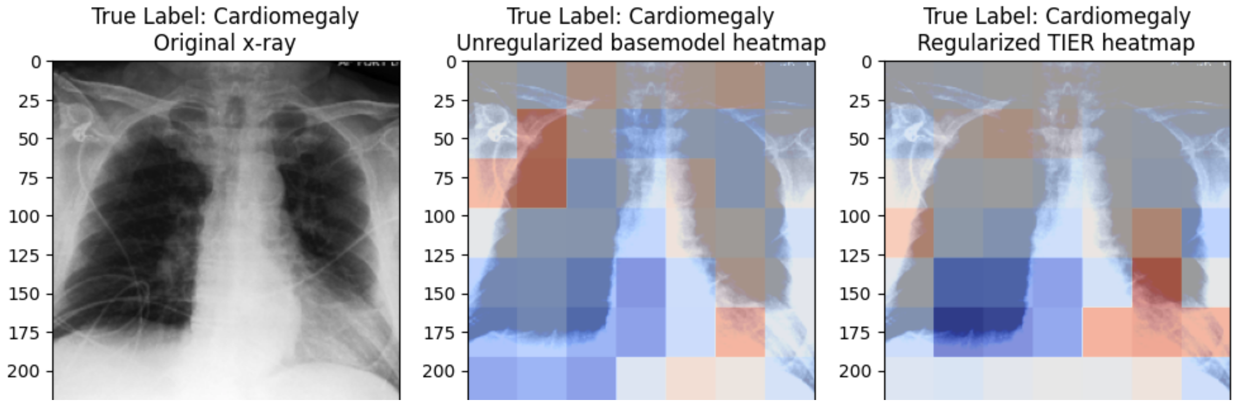

Figure 3: **The penalty term induces shrinkage in the image patch-text token similarity scores**. A CXR positive for cardiomegaly, overlaid with heatmaps displaying zero-shot cardiomegaly score for the unregularized (center) and regularized (right) models. Blue corresponds to a negative zero-shot score, gray is neutral, and red is a positive zero-shot score. As can be seen by comparing the middle and right figures, the regularized model Note: In this instance cardiomegaly is located in the lower-right quadrant of the image.

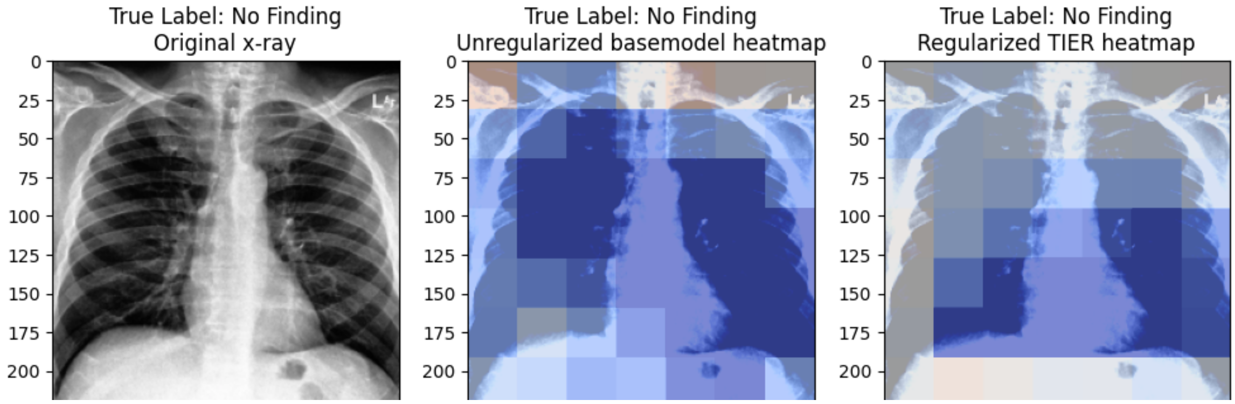

Figure 4: A CXR with no findings, overlayed with heatmaps displaying zero-shot cardiomegaly score for the unregularized (center) and regularized (right) models.

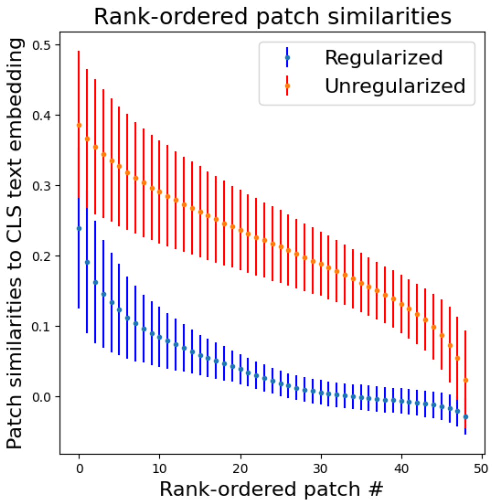

Figure 5: The similarities of each patch to the CLS embedding for a set of 160 MIMIC-CXR images, sorted in descending order.

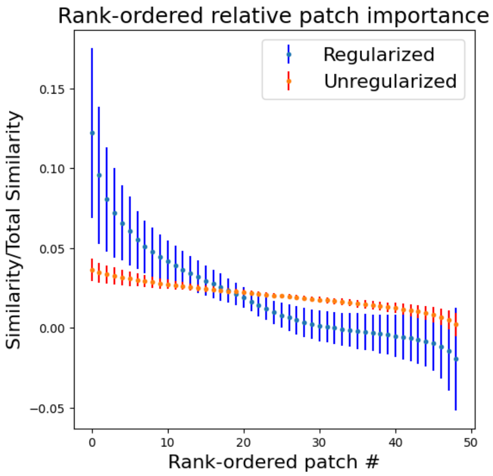

Figure 6: The same plot as Fig. 5 in which each similarity is divided by the sum of the total similarity across all patches for that image.

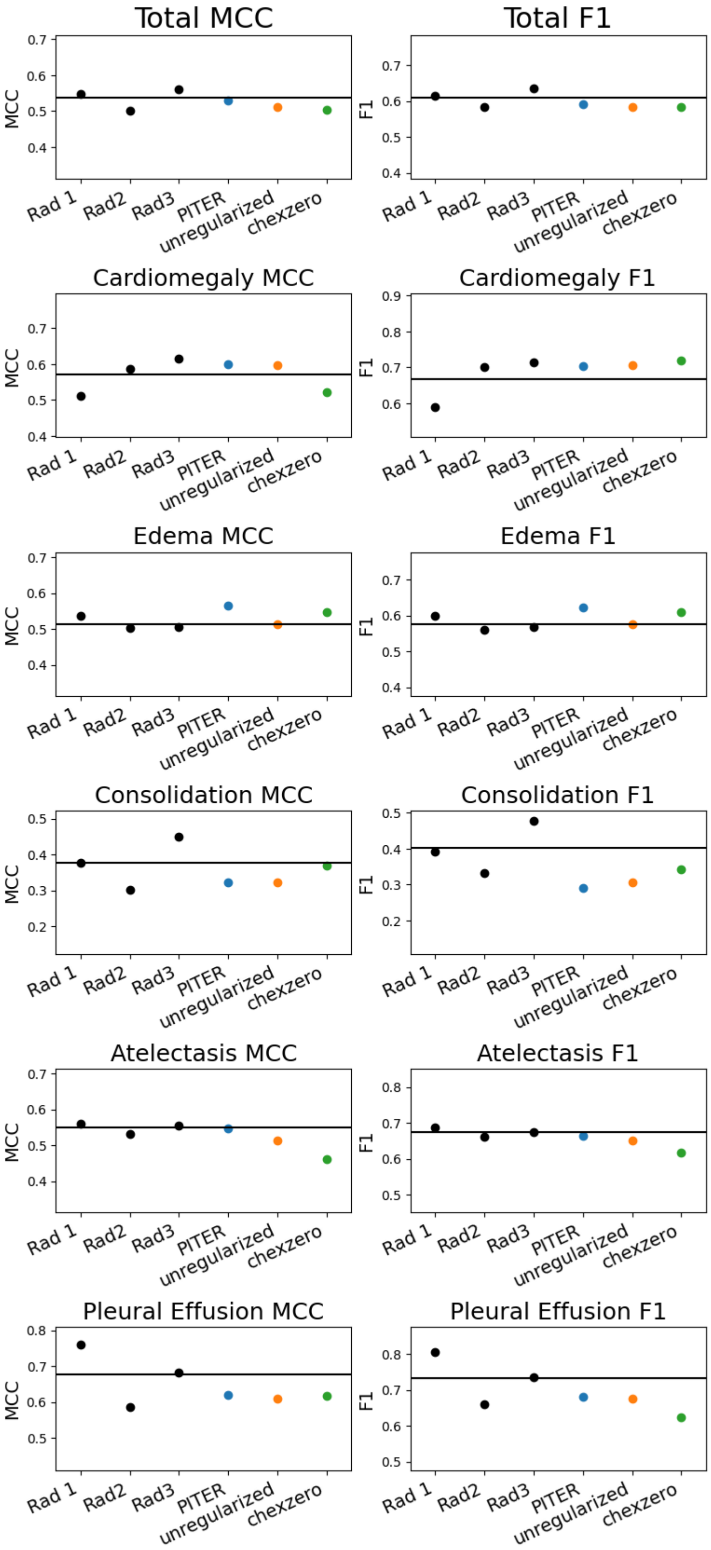

Figure 7: Benchmarking against radiologists. MCC (Matthews correlation coefficient) scores and F1 scores for the three zero-shot models and three radiologists are shown. Our regularized model (blue) has the highest zero-shot MCC and F1 scores on average and for the majority of labels. These zero-shot models are nearly competitive with radiologists, with only slightly lower average performances.

Label	Count	Regularized TIER	Unregularized Basemodel	Chexzero
Number of Evaluations Won (Percent)	57	**29 (50.9%)**	18 (31.6%)	10 (17.5%)
Average AUC	39053	**0.7803**	0.7796	0.7482
Endotracheal Tube	284	0.99172	**0.99322**	0.99203
Pulmonary Edema	87	0.95885	0.95818	**0.95898**
Pulmonary Fibrosis	166	**0.95676**	0.95504	0.92839
Pleural Effusion	1748	0.95234	0.95025	**0.95519**
Heart Insufficiency	546	**0.94052**	0.94034	0.94032
Lung Metastasis	89	**0.90426**	0.90354	0.84494
Vascular Redistribution	129	0.89902	0.89730	**0.90081**
Calcified Pleural Thickening	102	**0.89223**	0.87997	0.85692
Cardiomegaly	3746	0.89021	0.89940	**0.90012**
Pulmonary Mass	247	0.88930	**0.89881**	0.84084
Consolidation	364	**0.88765**	0.87510	0.86340
Alveolar pattern	1353	**0.88490**	0.88007	0.84838
Multiple nodules	102	0.88312	**0.88955**	0.77114
Hypoexpansion	166	**0.87630**	0.87620	0.76866
Cavitation	122	**0.86911**	0.86559	0.84821
Hypoexpansion basal	119	**0.86799**	0.84526	0.74527
Hilar Congestion	601	**0.86572**	0.85993	0.85843
Reticular Interstitial Pattern	72	0.85337	**0.85929**	0.84579
Tuberculosis	59	**0.84425**	0.82670	0.84010
Reticulonodular Interstitial Pattern	51	**0.84258**	0.82305	0.81245
Tuberculosis Sequelae	185	**0.83534**	0.83225	0.69592
Interstitial Pattern	1907	**0.83523**	0.83079	0.81062
Lobar Atelectasis	168	**0.83337**	0.83169	0.79188
Costophrenic Angle Blunting	1683	**0.82273**	0.82268	0.77868
Pneumothorax	98	0.82200	**0.82212**	0.73496
Atelectasis	676	**0.82080**	0.80829	0.80469
Vertebral Fracture	104	0.82073	**0.86166**	0.66401
Pneumonia	1780	**0.81525**	0.79970	0.78387
Emphysema	376	0.80669	0.81655	**0.82435**
Bullas	192	0.80621	**0.85627**	0.58313
Normal	12694	0.79657	**0.80472**	0.76518
Pleural Thickening	213	**0.78842**	0.78460	0.76980
Central Vascular Redistribution	63	**0.78360**	0.76702	0.75158
Infiltrates	1456	**0.77322**	0.76672	0.73997
Humeral Fracture	81	0.76876	**0.79189**	0.72041
Bronchiectasis	667	0.76405	**0.76560**	0.71624
Minor Fissure Thickening	127	**0.75643**	0.75628	0.54595
Hyperinflated Lung	197	**0.75240**	0.72875	0.74708
COPD signs	4823	0.75013	**0.76869**	0.75444
Mediastinal Enlargement	106	**0.74262**	0.73681	0.73973
Vertebral Compression	126	0.73247	**0.75002**	0.70839
Adenopathy	136	0.72426	0.72311	**0.73660**
Hilar Enlargement	447	0.71852	**0.72445**	0.67296
Rib Fracture	140	**0.71615**	0.70734	0.69991
Air Trapping	1952	**0.70145**	0.69475	0.56082
Laminar Atelectasis	1378	0.69865	**0.71266**	0.68271
Ground Glass Pattern	123	**0.69337**	0.68521	0.62526
Calcified Adenopathy	124	**0.64429**	0.62993	0.56879
Vascular Hilar Enlargement	1428	0.62743	**0.63861**	0.63382
Mediastinal Mass	74	0.60398	0.42813	**0.69049**
Unchanged	4036	0.59974	**0.61818**	0.52172
Tracheal Shift	180	0.57974	0.56313	**0.66158**
Nodule	736	0.56563	**0.61066**	0.60075
Clavicle Fracture	74	**0.54489**	0.51931	0.53835
Pseudonodule	795	0.53475	**0.54275**	0.51380
Superior Mediastinal Enlargement	153	0.50927	0.61333	**0.65837**
End On Vessel	63	0.43606	0.44358	**0.52774**

Table 2: **The regularized model outperforms the unregularized and Chexzero models in fine-grained finding prediction.** Zero-shot AUCs for 57 padchest findings. The best AUC for each finding across the three tested models is shown in **bold**

.

study reporting just a 74% average accuracy for three radiologists using chest CT for this task (Bai et al., 2020).

We created positive and negative zero-shot queries to discriminate COVID-19 and non-COVID-19 pneumonia based on differences mentioned in the literature (Bai et al., 2020; Borghesi & Maroldi, 2020). For the positive COVID-19 query, we used the query "Ground glass opacities and consolidation with peripheral distribution with fine reticular opacity and vascular thickening." For the negative COVID-19 query, we used "Pleural effusion present with lymphadenopathy and consolidation with central distribution." which were described by Bai et al. (2020) as findings more specific to non-COVID-19 pneumonia. Using these zero-shot queries, we achieved zero-shot AUCs of 0.759, 0.753, and 0.752 with the regularized, unregularized, and Chexzero models respectively on discriminating COVID-19 from non-COVID pneumonia within the COVID-QU-Ex Dataset (Tahir et al., 2021; 2022).

This performance indicates that we can leverage our model for difficult multi-class classification tasks by simply providing English descriptions of the class-discriminating features. Furthermore, extending this zero-shot procedure to other labels is relatively straightforward, meaning self-supervised vision-language architectures could easily be leveraged in the future to diagnose novel disease based on text descriptions.

4 Discussion & Limitations

In this work, we introduce a regularization method for contrastive language-image pre-trained models which encourages shrinkage of the image-patch and text-token similarities. We demonstrate how our regularization method can benefit zero-shot performance of these models by training a model that achieves SOTA zero-shot classification performance on a broad set of CXR findings. The improvements were robust across a wide range of tasks relative to many strong benchmarks, though in some instances the improvements were modest. Though our work was confined to a medical context, we believe it should be broadly applicable to many other areas where CLIP-style models are used, though these applications were beyond the scope of the present work. We believe our work contributes to a growing literature (Kumar et al., 2022; Mu et al., 2022; Meier et al., 2021) seeking to augment and improve CLIP-style models with inductive biases and domain-specific observations.

References

Harrison X Bai, Ben Hsieh, Zeng Xiong, Kasey Halsey, Ji Whae Choi, Thi My Linh Tran, Ian Pan, Lin-Bo Shi, Dong-Cui Wang, Ji Mei, et al. Performance of radiologists in differentiating covid-19 from non-covid-19 viral pneumonia at chest ct. *Radiology*, 296(2):E46–E54, 2020. 11

Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, et al. Making the most of text semantics to improve biomedical vision–language processing. *arXiv preprint arXiv:2204.09817*, 2022. 1, 3

Andrea Borghesi and Roberto Maroldi. Covid-19 outbreak in italy: experimental chest x-ray scoring system for quantifying and monitoring disease progression. *La radiologia medica*, 125(5):509–513, 2020. 11

Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria de la Iglesia-Vayá. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis*, 66:101797, 2020. 2

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*, pp. 770–778, 2016. 1

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 590–597, 2019. 2

Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019. 2

Bhawesh Kumar, Anil Palepu, Rudraksh Tuwani, and Andrew Beam. Towards reliable zero shot classification in self-supervised models with conformal prediction. *arXiv preprint arXiv:2210.15805*, 2022. 1, 11

Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alex Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. *Advances in Neural Information Processing Systems*, 34:29287–29303, 2021. 11

Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. In *European Conference on Computer Vision*, pp. 529–544. Springer, 2022. 11

Anil Palepu and Andrew L Beam. Self-supervision on images and text reduces reliance on visual shortcut features. *arXiv preprint arXiv:2206.07155*, 2022. 1, 3

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*, pp. 8748–8763. PMLR, 2021. 1, 3

Anas M Tahir, Muhammad EH Chowdhury, Amith Khandakar, Tawsifur Rahman, Yazan Qiblawey, Uzair Khurshid, Serkan Kiranyaz, Nabil Ibtehaz, M Sohel Rahman, Somaya Al-Maadeed, et al. Covid-19 infection localization and severity grading from chest x-ray images. *Computers in biology and medicine*, 139: 105002, 2021. 11

Anas M. Tahir, Muhammad E. H. Chowdhury, Yazan Qiblawey, Amith Khandakar, Tawsifur Rahman, Serkan Kiranyaz, Uzair Khurshid, Nabil Ibtehaz, Sakib Mahmud, and Maymouna Ezeddin. Covid-qu-ex dataset, 2022. URL https://www.kaggle.com/dsv/3122958. 11

Ekin Tiu, Ellie Talius, Pujan Patel, Curtis P Langlotz, Andrew Y Ng, and Pranav Rajpurkar. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nature Biomedical Engineering*, pp. 1–8, 2022. 1, 2, 6, 7

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. 1

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. *arXiv preprint arXiv:2210.10163*, 2022. 2

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. *arXiv preprint arXiv:2010.00747*, 2020. 1, 3

A Appendix

A.1 Chexpert Queries

Class Label	Caption
Cardiomegaly	Cardiomegaly is present. The heart shadow is enlarged. The cardiac silhouette is enlarged.
Pleural Effusion	Blunting of the costophrenic angles represents pleural effusions. The pleural space is filled with fluid. Layering pleural effusions are present.
Edema	Edema is present. Increased fluid in the alveolar wall indicates pulmonary edema.
Consolidation	Consolidation is present. Dense white area of right lung indicative of consolidation.
Atelectasis	Atelectasis is present. Basilar opacity and volume loss is likely due to atelectasis.
No Finding	The lungs are clear. No abnormalities are present. The chest is normal. No clinically significant radiographic abnormalities. No radiographically visible abnormalities in the chest.

Table 3: Query captions used for zero-shot classification. No Finding captions are used as the negative queries for CheXpert classification, while the rest are used as positive queries for their respective labels.

A.2 Code/model availability

Code is available at `https://github.com/apalepu13/TIER_Regularized_CLIP`. For model checkpoints for the regularized/unregularized/fully supervised models, contact the authors. The BioViL model is available at `https://huggingface.co/microsoft/BiomedVLP-CXR-BERT-specialized`.