conVIRT: **Con**trastive Learning of Medical **Vi**sual **R**epresentations from Paired Images and **T**ext

Authors: Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D. Manning, Curtis P. Langlotz
Date of Publication: 2020 (revised 2022)
Summarized by: Suman Dhakal

Abstract: One of the barriers for development of accurate computer vision systems for medical imaging is scarcity of human annotations. Learning visual representation of medical images to support transfer-learning for downstream tasks is crucial for achieving high performance.
Existing work has used three main approaches which includes

- Small scaled expert annotated data with ImageNet transfer learning
- Rule based label extraction from the medical report
- Contrastive self-supervised learning

In this paper, authors have proposed a new unsupervised method for pretraining medical visual representation with paired medical *images* and their *descriptive text* dataset. They transferred this pretrained visual representation for downstream tasks –4 medical image classification tasks, 2 zero-shot image retrieval tasks, and showed that it outperformed strong baseline in most settings.

Link to the original paper: https://arxiv.org/pdf/2010.00747.pdf

Source code (non-official implementation): https://github.com/edreisMD/ConVIRT-pytorch
Dataset: https://github.com/yuhaozhang/convirt

Terminologies

- **Inter-class similarity:**
    - Inter-class similarity is the situation where objects belonging to different classes have visually similar appearance due to minute variations in the morphological features.

- ***ImageNet:***
    - It is a large-scale dataset that is used in computer vision research
    - ImageNet originally consisted of **over 14 million images** in more than **20,000 categories**.
    - The most well-known subset of ImageNet, known as the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), contains around **1.2 million images** across **1,000 categories.**

- ***Pretrain:***
    - This is the term used to denote “training” of ML models in the context of transfer learning.
    - More formally, it refers to the process of training ML models with a large dataset before fine-tuning it on a specific task.

- ***Downstream transfer:***
    - It is a term that is used in the context of transfer learning, it basically means the use of a pre pre-trained model to perform specific tasks.
    - "downstream" refers to the subsequent task or tasks that the pretrained model is applied to.
    - The process of downstream transfer involves
        - **Pretraining**: learn generac features/patterns from large dataset
        - **Downstream task**: fine-tune the mode on task specific dataset related to the specific task.

- ***Self-supervised learning***
    - A Machine Learning paradigm that involves training models to learn from data itself without the need of extensive human label annotations.
    - Example, a word sequence prediction model trained to predict the next word based on the previous input sentence.

- ***Visual Representation***
    - Visual representation refers to the process of converting raw visual data into numerical forms that can be understood and analyzed by computers.
    - It is interchangeable with “feature map”.
    -
- ***Linear projection***
    - A "linear projection" is a simple linear/dense layer without activation function, that is, **a matrix multiplication and a bias vector addition**
    - When doing linear projection, we can project a vector x of dimension n to a vector y of dimension size m by multiplying a projection matrix W of shape n×m.
    - Mathematically,

      $$\text{Output} = \text{Input} \times \text{Weights} + \text{Bias}$$
- ***Non-Linear projection***
    - Dense layer with activation function

- ***In-domain initialization***
    - Pretraining a model with domain (that the downstream task belongs to) specific data.

Introduction

Learning visual representations is basically pretraining of CNN layer(backbone) that could be used for downstream tasks and to achieve remarkable performance which could be hard to achieve with random CNN layer setup.

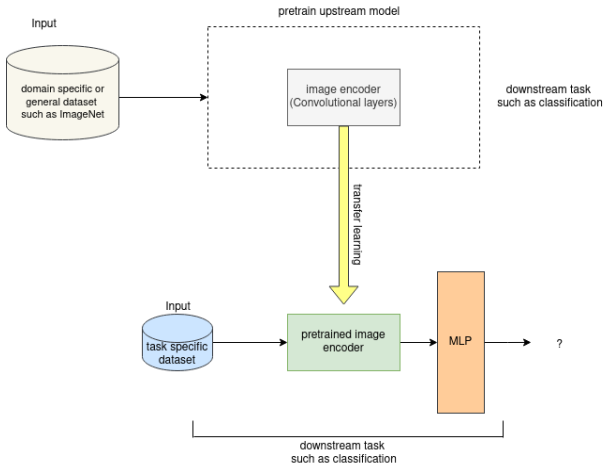
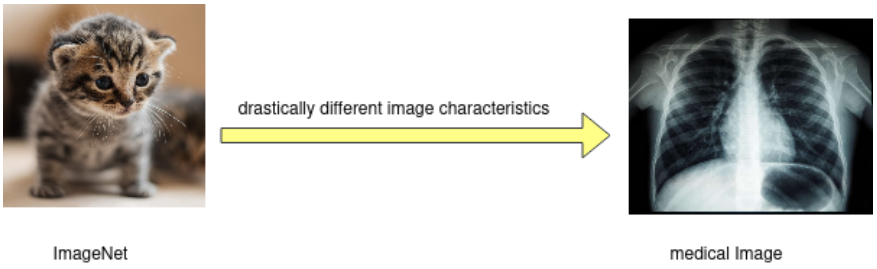


Fig: Generic overview of transfer learning

Existing approaches includes

- **Small scaled expert annotated data + ImageNet**
    - It is a common practice to use pretrained ImageNet weights and fine-tune it with small scaled expert annotated data for specific tasks.
    - **Limitations:**
        - ImageNet is pretrained on general images whereas medical images carry different characteristics than those images, as a result no accurate visual representation for medical images.



- **Rule based label extraction from medical report**
    - Medical reports are textual information and are prepared by lab technicians or domain experts which come along with medical images.
    - Instead of hiring label annotators, people have practiced rule based label extraction to directly extract labels for medical images.
    - **Limitations:**
        - Inaccurate and hard to generate generalized labels

- **Contrastive self-supervised learning**
    - Contrastive self-supervised learning has been able to overcome the barrier due to the scarcity of labeled dataset for learning visual representation. However, it is still unable to perform well in domain specific areas like medical imaging where images have high inter-class similarity.
    - **Limitation:**
        - High inter-class similarity among medical image make it hard for contrastive

- **Proposed ConVIRT**
    - Medical images are naturally produced as a result of medical examinations, observations, tests, and diagnoses.
    - These reports contain information about the patient's medical history, symptoms, test results, diagnoses, and treatment recommendations.
    - The proposed conVIRT method for learning medical visual representation has overcome scarcity of annotated labels by directly relying on medical reports.
    - Similarly, it is a supervised learning that learns by contrasting image-text pairings, thus, does not go through inter-class similarity situations.

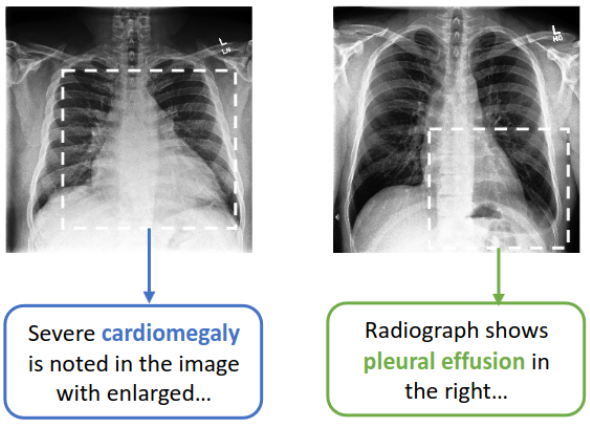


Figure 1: Two example chest X-ray images with different abnormality categories, along with sentences from their paired textual report and example views indicative of their characteristics.

Methods

simCLR framework

Note: It is just to show how simCLR framework is similar to the conVIRT framework. Consider ResNet-18 as an example, since the framework does not specify which encoder to use.

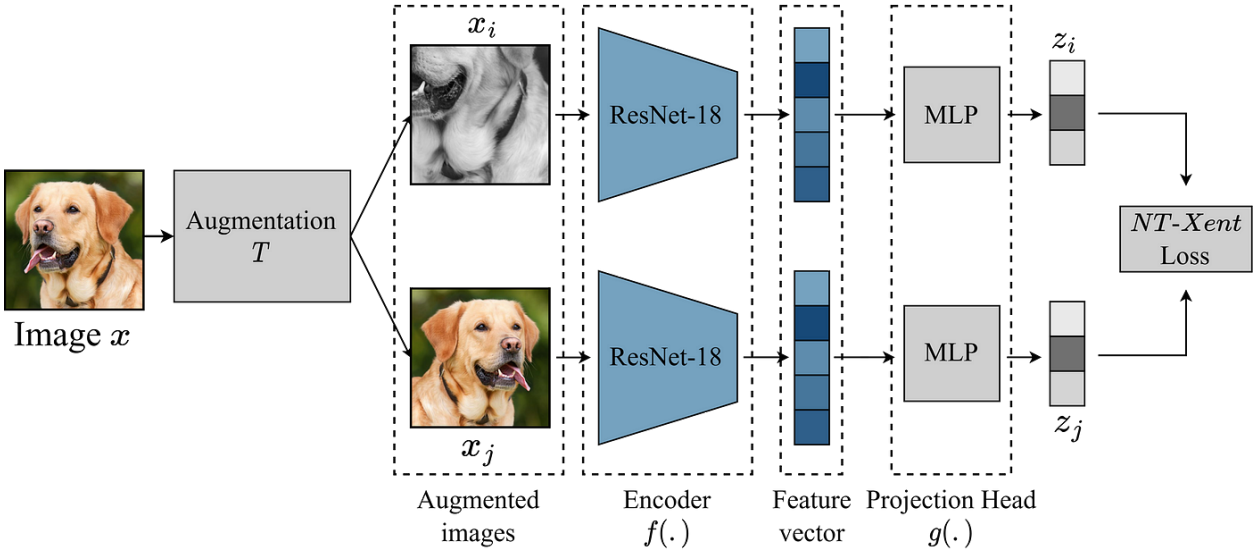


Fig: an architecture of simCLR

ConVIRT framework

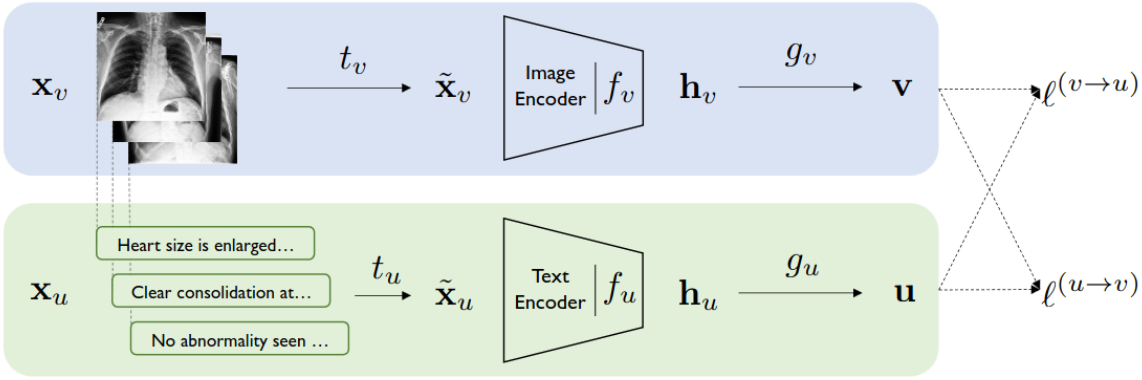


Figure 2: Overview of our ConVIRT framework. The blue and green shades represent the image and text encoding pipelines, respectively. Our method relies on maximizing the agreement between the true image-text representation pairs with bidirectional losses $\ell^{(v\to u)}$ and $\ell^{(u\to v)}$.

Where,

X_v = image input

X_u = text input

t_v = image transformation functions ~ T; transformation function/s sampled from T

t_u = a text transformation functions ~ T; transformation function/s sampled from T

T = a family of transformation function

$\tilde{x}_v$ = image after applying t_v

$\tilde{x}_u$ = text after applying sampling function t_u

f_v = image encoder that transforms $\tilde{x}_v$ into fixed-dimensional vector; for instance, ResNet50

f_u = text encoder that transforms $\tilde{x}_u$ into fixed-dimensional vector; for instance, BERT

h_v = fixed dimensional vector

h_u = fixed dimensional vector

g_v, g_u = non-linear projections for encoded image and text respectively

v, u = fixed dimensional vectors; text and

$$v = g_v(f_v(\tilde{x}_v))$$
$$u = g_u(f_u(\tilde{x}_u))$$

Loss Function

At training time, we sample a minibatch of N input pairs $(\mathbf{x}_v, \mathbf{x}_u)$ from training data, and calculate their representation pairs $(\mathbf{v}, \mathbf{u})$. We use $(\mathbf{v}_i, \mathbf{u}_i)$ to denote the i-th pair. The training objective of ConVIRT involves two loss functions. The first loss function is an image-to-text contrastive loss for the i-th pair:

$$\ell_i^{(v \to u)} = -\log \frac{\exp(\langle \mathbf{v}_i, \mathbf{u}_i \rangle / \tau)}{\sum_{k=1}^{N} \exp(\langle \mathbf{v}_i, \mathbf{u}_k \rangle / \tau)}, \quad (2)$$

Where,

- $\langle v_i, u_i \rangle$ represents the cosine similarity
- $\tau \in$ R+represents a temperature parameter.

Similarly, text-to-image contrastive loss for the i-th pair:

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

Our final training loss is then computed as a weighted combination of the two losses averaged over all positive image-text pairs in each minibatch:

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \left(\lambda \ell_i^{(v \to u)} + (1 - \lambda) \ell_i^{(u \to v)} \right), \tag{4}$$

where $\lambda \in [0, 1]$ is a scalar weight.

Training Details

conVIRT framework does not specify the choice for image and text encoder, transformation and projection function.
Authors have chose:

- g_v *and* g_u as a separate learnable single-hidden-layer neural network
  - $g_v(\cdot) = W^{(2)}\sigma(W^{(1)}(\cdot))$ where σ is a ReLU; same for g_u
- Image encoder f_v = ResNet50 architecture
- text encoder f_u = BERT encoder
  - Followed by max-pool layer over all output vectors
  - Initialized with ClinicalBERT weights pretrained on MIMIC clinical notes
  - Embedding and first 6 transformer layers are freezed and fine tuned the last 6 layers
- Image transformation family T contains
  - *cropping, horizontal flipping, affine transformation, color jittering (brightness and contrast adjustment) and Gaussian blur*
- Image transformation function t_v is sampled from T means it randomly applies above transformation functions to the input image
- text transformation function t_u = perform uniform sampling of a sentence for text documents
  - $\tilde{x}_u$ is a randomly sampled sentence from X_u for each minibatch

Experiments

They evaluate conVIRT by pretraining two different image encoders

- Chest image encoder:
    - **Dataset:** publicly available MIMIC-CXR v2 (*chest radiograph image* + *textual report*)
- Bone image encoder:
    - **Dataset:** Collected from Rhode Island Hospital system (*musculoskeletal (i.e., bone)* + *textual report*)

Baseline models

Baseline models that they conVIRT against with

- **Random Init:**
    - ResNet50 with default random initialization

- **ImageNet Init:**
    - Initialized with weights pretrained on ImageNet

- **Caption-LSTM:**
    - pretrain the ImageNet-initialized CNN weights with an image captioning task using the standard CNN-LSTM with attention mode

- **Caption-Transformer:**
    - a CNN-Transformer-based captioning model for caption-based pretraining

- **Contrastive-Binary-Loss**
    - This baseline differs from ConVIRT by contrasting the paired image and text representations with a binary classification head

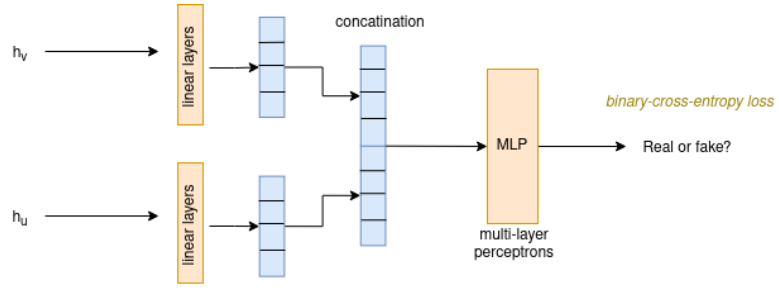


Fig: contrasting the paired image and text representations with a binary classification head

Evaluations Tasks

Image classification Tasks

They compared their pretrained image encoders on four representative medical image classification tasks

- **RSNA** Pneumonia Detection
    - Dataset: https://www.kaggle.com/c/rsna-pneumonia-detection-challenge
- **CheXpert** image classification
    - Dataset: https://www.tensorflow.org/datasets/catalog/chexpert
- **COVIDx** image classification
    - Dataset: https://www.v7labs.com/open-datasets/covidx
- **MURA** bony abnormality detection
    - Dataset: https://stanfordmlgroup.github.io/competitions/mura/

Zero-shot Image-image Retrieval

- This evaluation is similar to the conventional content-based image retrieval setting in which we search for images of a particular category using a representative query image.

Zero-shot Text-image Retrieval

- This setting is similar to the image-image retrieval setting, but instead of using query images, we retrieve images of a particular category with textual queries.

Results

Note: Below results are the average values of multiple settings; conVIRT(our) is the average result of conVIRT pretrained on bone and conVIRT pretrained on chest dataset.

1. Classification Tasks

(a) Linear classification

Method	RSNA (AUC)			CheXpert (AUC)			COVIDx (Accu.)		MURA (AUC)		
	1%	10%	all	1%	10%	all	10%	all	1%	10%	all
General initialization methods											
Random Init.	55.0	67.3	72.3	58.2	63.7	66.2	69.2	73.5	50.9	56.8	62.0
ImageNet Init.	82.8	85.4	86.9	75.7	79.7	81.0	83.7	88.6	63.8	74.1	79.0
In-domain initialization methods											
Caption-Transformer	84.8	87.5	89.5	77.2	82.6	83.9	80.0	89.0	66.5	76.3	81.8
Caption-LSTM	89.8	90.8	91.3	85.2	85.3	86.2	84.5	**91.7**	75.2	81.5	84.1
Contrastive-Binary-Loss	88.9	90.5	90.8	84.5	85.6	85.8	80.5	90.8	76.8	81.7	85.3
ConVIRT (Ours)	**90.7**	**91.7**	**92.1**	**85.9**	**86.8**	**87.3**	**85.9**	**91.7**	**81.2**	**85.1**	**87.6**

(b) Fine-tuning

Method	RSNA (AUC)			CheXpert (AUC)			COVIDx (Accu.)		MURA (AUC)		
	1%	10%	all	1%	10%	all	10%	all	1%	10%	all
General initialization methods											
Random Init.	71.9	82.2	88.5	70.4	81.1	85.8	75.4	87.7	56.8	61.6	79.1
ImageNet Init.	83.1	87.3	90.8	80.1	84.8	87.6	84.4	90.3	72.1	81.8	87.0
In-domain initialization methods											
Caption-Transformer	86.3	89.2	92.1	81.5	86.4	**88.2**	88.3	92.3	75.2	83.2	87.6
Caption-LSTM	87.2	88.0	91.0	83.5	85.8	87.8	83.8	90.8	78.7	83.3	87.8
Contrastive-Binary-Loss	87.7	89.9	91.2	86.2	86.1	87.7	89.5	90.5	80.6	84.0	88.4
ConVIRT (Ours)	**88.8**	**91.5**	**92.7**	**87.0**	**88.1**	88.1	**90.3**	**92.4**	**81.3**	**86.5**	**89.0**

Table 1: Results for the medical image classification tasks: (a) linear classification; (b) fine-tuning setting. All results are averaged over 5 independent models. Best results for each setting are in boldface. COVIDx 1% setting is omitted due to the scarcity of labels in COVIDx.

2. Retrieval Task

Method	Image-Image Retrieval			Text-Image Retrieval		
	Prec@5	Prec@10	Prec@50	Prec@5	Prec@10	Prec@50
Random	12.5	12.5	12.5	12.5	12.5	12.5
ImageNet	14.8	14.4	15.0	–	–	–
In-domain initialization methods						
Caption-Transformer	29.8	28.0	23.0	–	–	–
Caption-LSTM	34.8	32.9	28.1	–	–	–
Contrastive-Binary-Loss	38.8	36.6	29.7	15.5	14.5	13.7
ConVIRT (Ours)	**45.0**	**42.9**	**35.7**	**60.0**	**57.5**	**48.8**
Fine-tuned						
ConVIRT + CheXpert Supervised	56.8	56.3	48.9	–	–	–

Table 2: Zero-shot image-image and text-image retrieval results on the CheXpert 8×200 datasets. *Random* shows results from a random guess; *ConVIRT + CheXpert Supervised* shows results from further fine-tuning the pretrained weights with supervised training data. Text-image retrieval results are not obtained for some methods due to the lack of text encoders.

3. Comparisons to Image-only Contrastive Learning

Method	RSNA (Linear, 1%)	CheXpert (Linear, 1%)	Image-Image (Prec@10)
ImageNet	82.8	75.7	14.4
SimCLR (Chen et al., 2020a)	86.3	77.4	17.6
MoCo v2 (Chen et al., 2020b)	86.6	81.3	20.6
ConVIRT	90.7	85.9	42.9

Table 3: Comparisons of ConVIRT to image-only contrastive learning. For RSNA and CheXpert we show the AUC under linear classification with 1% training data.

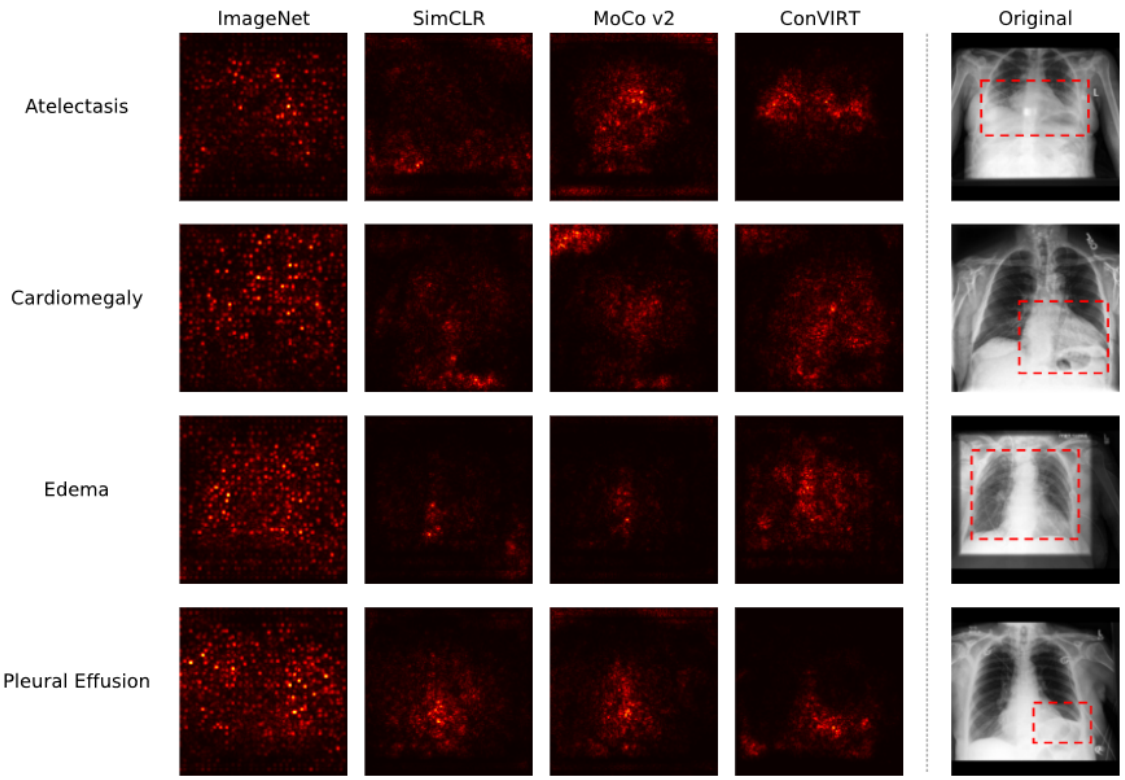


Figure 4: Saliency maps on sampled images for 4 abnormality categories in the CheXpert dataset. For each image we present maps for ImageNet, SimCLR, MoCo v2 and our ConVIRT initializations. Ground truth regions that are indicative of the abnormalities are shown as red boxes in the original images on the right, and are seen to most closely match the regions found by ConVIRT.

Hyper parameter Analysis:

Settings	RSNA Linear (1%, AUC)	Image-Image (Prec@10)	Text-Image (Prec@10)
ConVIRT (default)	90.7	42.9	57.5
$\tau = 0.01$	90.7	40.5	21.0
$\tau = 1$	89.6	25.0	31.0
bs = 16	90.3	40.0	55.8
bs = 128	90.3	39.3	50.3
linear proj.	90.6	40.8	55.8

Table 4: Evaluation results with different hyperparameters, for the RSNA 1% data linear evaluation, image-image retrieval and text-image retrieval tasks. *bs* represents batch size and *linear proj.* represents using linear projection layers for g_v and g_u. Our default model uses $\tau = 0.1$, bs $= 32$ and non-linear projections.

Pretraining Details:

Dataset Preprocessing:

- For both dataset i.e chest and bone
  - resize the image files to have a size of 256 on the larger side
- textual radiology report data
  - Tokenizer : default english tokenizer (in version 4.0.0 of the CoreNLP library)
  - Keep only *findings and Impression section* of the report
- remove all image-text pairings from the dataset where the text section is empty or has less than 3 tokens.
- After preprocessing:
  - chest dataset: 217K
  - Bone dataset: 48K

Image Encoder

- ResNet50

Text Encoder

- Pretrained ClinicalBERT

Hyper parameter for Pretraining:

- Projection layers:
  - Output dimension (d) = 512
  - Temperature ($\boldsymbol{\tau}$) = 0.01
  - A loss weight (λ) = 0.75
- Image transformation family (T):

    - random cropping: [0.6, 1.0]
    - horizontal flipping: p = 0.5
    - affine transformation: [−20, 20]
    - max horizontal and vertical translation fractions = 0.1
    - a scaling factor: [0.95, 1.05]
    - brightness and contrast adjustment ratios: [0.6, 1.4]
    - Gaussian Blur: [0.1, 3.0]
- After transformation t_v all images are resized to 224×224
- Adam Optimizer:
    - initial learning rate = 1e-4
    - weight decay of 1e-6
- Batch size: 32

Note: for more details go for the original paper

Limitations

Their work mainly focuses on comparing **ConVIRT** against conventional ImageNet initialization, image captioning-based initialization, and image-only contrastive learning approaches including **SimCLR** and **MoCo** to demonstrate the data efficiency and effectiveness of image-text pretraining. They did not compare their method against relevant subsequent studies that extended **ConVIRT**, such as **LoVT** or **GloRIA**, mainly because such comparisons are included in these studies.

Conclusion

ConVIRT outperformed other strong in-domain initialization methods, and led to representations with notably higher quality. Compared to ImageNet pretraining, ConVIRT is able to achieve the same level of classification accuracy with an order of magnitude less labeled data