Self-supervised Learning with Radiology Images and Reports

Xu Han xh852@nyu.edu *Courant Institute of Mathematical Sciences*
New York University
New York, NY, United States of America

Carter Perkins csp9835@nyu.edu *Courant Institute of Mathematical Sciences*
New York University
New York, NY, United States of America

Valay Shah vs2393@nyu.edu *Courant Institute of Mathematical Sciences*
New York University
New York, NY, United States of America

1. Introduction

Developing a system that learns meaningful representations of patient radiology data for downstream tasks is one challenge in the healthcare domain that machine learning can address (Ghassemi et al., 2020). This is particularly important in the healthcare domain as manually labeled radiology datasets are not in abundance so leveraging unlabeled information is particularly advantageous for pre-training models. Recent studies have shown that it is better to pre-train exclusively on healthcare datasets rather than using transfer learning with large non-healthcare models to avoid incorporating non-medical features into the representation (Krishnan et al., 2022) — further emphasizing the need for self-supervised learning in healthcare. Moreover, self-supervised pre-training can leverage multimodal data, such as radiology text reports and x-ray images, to further enrich the quality of representation learning. Radiology reports, often used from the MIMIC-CXR dataset (Johnson et al., 2019), are especially useful as they contain information rich sections such as Findings and Impressions, but it is unclear how much contrastive learning frameworks actually learn from these important sections. In this project, we will examine the importance of different radiology sections, and modify an existing contrastive learning framework to improve upon the quality of representations using the radiology text data. Our goal is to investigate the following research questions:

- **RQ1**: What section or combination of sections from a radiology text report result in a better representation?
- **RQ2**: How can a contrastive learning framework be improved to leverage more information from these sections?

We want to see how a multimodal contrastive learning framework utilize the sections of a radiology report and whether an existing framework can be improved if they are not utilizing the full information of text data. Radiology reports contain several sections, such as impressions, findings, indications, and examinations, but we focus on the two main sections which contain the most information about the patient — findings and impressions. We compared the results of models trained with different combinations of the text report sections, and modify an existing architecture to see the extend which we could improve the mutual information and learning.

2. Related Work

ConVIRT (Zhang et al., 2021) In this paper, the authors propose a method for learning visual representations of medical images using unsupervised learning, shown in Figure 1. They do this by pairing images with text data and training a model to encode the images using a bidirectional contrastive objective. The authors then test their method on the downstream task by transferring the pretrained weights to various medical image classification and zero-shot retrieval tasks. They find that their method leads to image representations that outperform strong baselines in most settings. For this research, we will also be working on the same architecture to find out the performance of different sections in a radiology report.

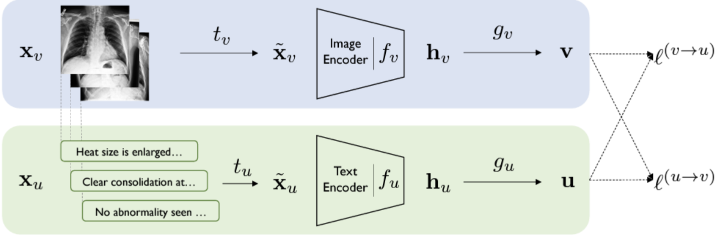


Figure 1: Model architecture introduced in ConVIRT. Image and text are paired and fed into encoders. $l^{(v\to u)}$ and $l^{(u\to v)}$ are bidirectional losses from matching image and report which are defined in Equation (2).

GLORIA (Huang et al., 2021) This paper proposes a framework for learning global and local representations of medical images using attention-based techniques. The authors focus on contrasting image sub-regions with the corresponding text in the associated reports. They generate context aware local representations of images by learning attention weights that emphasize significant image sub-regions for a particular word which in turn would improve performance in various medical image recognition tasks, including image-text retrieval, classification, and segmentation, using limited labeled data. This paper is to develop label-efficient multi-modal medical imaging representations by leveraging radiology reports. In order to learn the attention weights, it is necessary to learn the local representation while the images and the texts are similar, a global representation needs to be learned. The results show that this approach is effective and efficient in these tasks.

PubMedCLIP (Eslami et al., 2021) A new model which is developed on original CLIP model for medical Q&A datasets. The experiments are conducted on two MedVQA benchmark datasets and investigate two MedVQA methods, MEVF (Mixture of Enhanced Visual Features) and QCR (Question answering via Conditional Reasoning).

MoCo-CXR (Sowrirajan et al., 2021) An adaptation of the contrastive learning method Momentum Contrast (MoCo) to produce models with better representations and initializations for the detection of pathologies in chest X-rays.

BioViL (Boecking et al., 2022) BioViL is a machine learning model which uses both text and images to perform the learning in a contrastive way. It pretrains the text encoder on a very large corpus of radiology reports including MIMIC-CXR, MIMIC-III and PubMed and this model is named CXR-BERT. Then they use this learnings in their BioViL model to perform the downstream task of pneumonia segmentation and they outperform all the state of the art model.

3. Methods

3.1. Problem Definition

The first step of our project will be to use the ConVIRT contrastive learning framework to learn a rich representation of radiology images and reports. Machine learning models in healthcare particularly have to deal with a lack of labeled data for training, so pre-training image and text encoders on a large corpus is advantageous for boosting model performance on downstream tasks. MIMIC-CXR particularly emphasizes two sections Findings and Impressions as areas of interest within the reports, and prior work generally involves making use of these sections, but it is not clear if both sections contribute equally towards the representation learning task. To investigate this, we we will compare pre-training with the Findings, Impressions and both Findings & Impressions sections of the text reports to achieve a better understanding of how important these sections are — the model pre-trained with both sections will serve as our baseline as we anticipate this will perform the best as it can leverage data from both report sections. We also modify the ConVIRT architecture (Modified ConVIRT) in an attempt to further improve upon the baseline when training on both sections of the radiology reports to answer **RQ2**. The architecture remains the same as shown in Figure 1. However, there is one change in the text input pre-porcessing where in the ConVIRT paper they pass randomly sampled sentences from the entire report. Instead, we pass all the sentences of the two sections of the report for which we are checking. The aforementioned self-supervised learning frameworks have shown that image and text representations can be improved for downstream tasks through mutual information during the pre-training phase, so we aim to further enforce this phenomenon with our Modified ConVIRT architecture.

3.2. Modified ConVIRT

There are two major changes based on the original ConVIRT model. The first change centered around the text input branch where we separate the text input into two different branches — one for the Findings section and one for the Impressions section. For the two text new branches, each section of the text report is fed separately into the text encoder and then into the same MLP projection network as seen in the original ConVIRT setup as

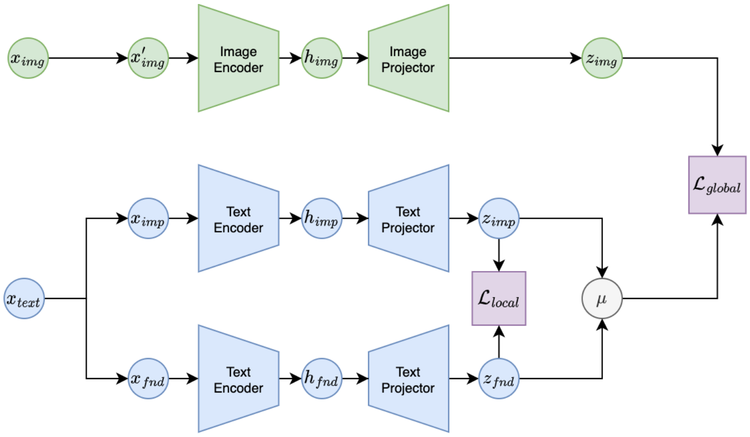


Figure 2: The Modified ConVIRT model

seen in Figure 2. These projected text representations are then aggregated and averaged together to represent the overall text representation of the two different sections, and we use this new text representation in the original loss function.

The second change we made is that we extended the loss function. As we separated the text branch into two for each report section, we want to ensure that the model utilizes both of these sections efficiently and that the report section representations are aligned. We introduce a local loss term to address this:

$$\mathcal{L}_{local} = \frac{1}{N} \sum_{i=1}^{N} \left(\alpha \ell_i^{(fnd \to imp)} + (1 - \alpha) \ell_i^{(imp \to fnd)} \right) \tag{1}$$

where $\alpha \in [0, 1]$ is a scalar weight, and $\ell_i^{(fnd \to imp)}$, $\ell_i^{(imp \to fnd)}$ are the Findings-to-Impressions and Impressions-to-Findings contrastive loss terms respectively. We define the contrastive loss between the i-th pair of two different representations using the ConVIRT formula:

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

where $\langle \boldsymbol{u}_i, \boldsymbol{v}_i \rangle = \boldsymbol{u}_i^{\intercal} \boldsymbol{v}_i / \|\boldsymbol{u}_i\| \|\boldsymbol{v}_i\|$ is the cosine similarity and $\tau \in \mathcal{R}^+$ is the temperature parameter. We denote the loss between the aggregate text representation and the image representation as $\mathcal{L}_{global}$ which follows the Equation (1) formula except with image-to-text and text-to-image contrastive terms. As a final calculation, we use a weighted sum of the local and global loss for our overall loss: $\mathcal{L} = \lambda \cdot \mathcal{L}_{local} + (1 - \lambda) \cdot \mathcal{L}_{global}$ where $\lambda \in [0, 1]$ is a scalar weight.

4. Data and Experiment Setup

4.1. Data

MIMIC-CXR The MIMIC Chest X-ray Database is a collection of chest radiographs in DICOM format with associated text reports. It includes 377,110 images from 227,835 radiographic studies conducted at the Beth Israel Deaconess Medical Center in Boston. In Figures 3 and 4, we showcase an example of image and report sample from the dataset. Since we were interested in how Findings and Impressions contribute differently in representation learning, we first filtered reports with both sections. We then extracted two sections from the reports and integrated Findings and Impressions into two separate datasets for later experiments.

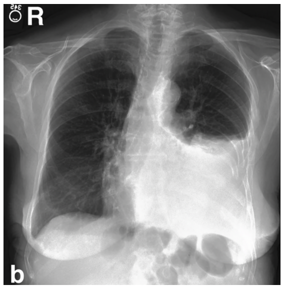
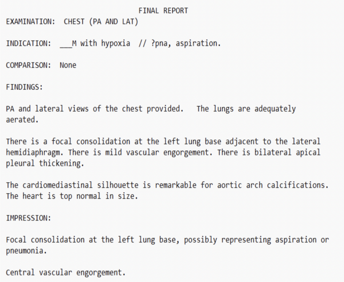


Figure 3: MIMIC-CXR image sample data



```
                                 FINAL REPORT
EXAMINATION:  CHEST (PA AND LAT)

INDICATION:  ___M with hypoxia  // ?pna, aspiration.

COMPARISON:  None

FINDINGS:

PA and lateral views of the chest provided.   The lungs are adequately
aerated.

There is a focal consolidation at the left lung base adjacent to the lateral
hemidiaphragm. There is mild vascular engorgement. There is bilateral apical
pleural thickening.

The cardiomediastinal silhouette is remarkable for aortic arch calcifications.
The heart is top normal in size.

IMPRESSION:

Focal consolidation at the left lung base, possibly representing aspiration or
pneumonia.

Central vascular engorgement.
```

Figure 4: MIMIC-CXR report sample data

MURA The MURA dataset consists of 40,561 X-ray images of bones from 14,863 studies and algorithms are used to analyze these images to determine if there are any abnormalities present. These musculoskeletal conditions, which affect a large portion of the global population, can cause significant pain and disability and often require emergency care. We utilized image data and labels indicating abnormality for evaluation tasks. This dataset did not require data preprocessing and we fed image data and labels directly to the model.

CheXpert A large dataset consists of 224,316 chest radiographs of 65,240 patients. A labeler is designed to automatically detect the presence of 14 observations in radiology reports, capturing uncertainties inherent in radiograph interpretation. If the labeler detects evidence indicating the abnormality, the image has label 1, else -1. If the labeler is unsure about an observation, it will assign label 0 to the image. We only used 8 observations in the dataset and they are *atelectasis*, *cardiomegaly*, *edema*, *fracture*, *pleural effusion*, *pneumonia*, *pneumothorax*, and a special *no finding* category indicating that no obvious abnormality is found in the image.

Data Split The train/test split on MURA is 80/20 and CheXpert already has a train/valid split, where the training set has 223,415 images, and the validation set has 235 images. To create a new valid and test dataset, we used an 80/20 split on the given validation set.

4.2. Experiments

4.2.1. Pre-training

The first experiment to run was training the four different contrastive models on the MIMIC-CXR dataset:

1. ConVIRT trained with both the Findings and Impressions sections included in the text report — this is our **baseline**.

2. ConVIRT trained with only the Findings section in the text report.
3. ConVIRT trained with only the Impressions section in the text report.
4. Modified ConVIRT trained with both the Findings and Impressions sections which are passed separately to the model.

In this pre-training task, we train each of the four approaches on a single GPU for 30 epochs using 10% of the MIMIC-CXR dataset. We are only able to use thir proportion of the data due to limitations with HPC — see section 6 for more details. We utilize a ResNet-18 He et al. (2016) initialized with random weights for our image encoder, and the pre-trained BioClinicalBERT Alsentzer et al. (2019) as our text encoder in each model. Additionally, we used the AdamW optimizer Loshchilov and Hutter (2019) and include our hyperparameters specifications in Table 1.

Hyperparameter	Value
Learning Rate	$1.0 \cdot 10^{-5}$
Image Projection Hidden Dimension	1024
Text Projection Hidden Dimension	1024
Image Projection Output Dimension	512
Text Projection Output Dimension	512
Temperature	0.1
ConVIRT Loss Scalar	0.75
Modified ConVIRT Global Loss Scalar	0.75
Modified ConVIRT Local Loss Scalar	0.75
Modified ConVIRT Total Loss Scalar	0.75

Table 1: Hyperparameter values used during pre-training

4.2.2. Evaluation

Metrics As we use the model to classify each radiology image, we chose F_1 score as our metric, where $F_1 = 2 * \frac{precision*recall}{precision+recall}$. Precision is ratio of the true positive to the sum of true positive and false positive; recall is the ratio of the true positive to the sum of true positive and false negative.

Model Evaluation We evaluated models on two downstream radiology classification tasks: linear classification and fine-tuning inspired from ConVIRT paper with the MURA (Rajpurkar et al., 2018) and CheXpert (Irvin et al., 2019) datasets. For linear classification, we freezed image encoder and only allow fine-tuning of linear classification head. This experiment is designed to assess the quality of learned representations as we only allow changes to the weights of the linear classifier. For the fine-tuning experiment, we allowed changes on both the image encoder and linear classification head to assess real-world practical usage of such models. For both experiments, we train for 10 epochs and allow for early stopping based on the validation F1 score. We also used cross entropy as a loss function and the AdamW optimizer. Finally, we experiment with using 10% or 100% of the MURA and CheXpert datasets for further assessment of our learned representations.

Model / Input	**Dataset**	**Task**	**Data Proportion**	**F1**
Baseline	MURA	Fine-tune	100%	69.13
			10%	61.24
		Linear Classification	100%	53.77
			10%	52.05
Baseline	CheXpert	Fine-tune	100%	41.66
			10%	38.89
		Linear Classification	100%	19.44
			10%	13.89
Imp & Find	MURA	Fine-tune	100%	68.44
			10%	58.34
		Linear Classification	100%	51.21
			10%	51.58
Imp & Find	CheXpert	Fine-tune	100%	38.89
			10%	36.11
		Linear Classification	100%	16.67
			10%	8.33
Modified ConVIRT	MURA	Fine-tune	100%	66.44
			10%	61.43
		Linear Classification	100%	53.52
			10%	52.08
Modified ConVIRT	CheXpert	Fine-tune	100%	38.89
			10%	25.00
		Linear Classification	100%	27.78
			10%	16.66

Table 2: Results

5. Results

We showcase the downstream evaluation experimental results in Table 2. In answering **RQ1**, we noticed the F1 score of the ConVIRT models trained exclusively on either the Findings or Impressions sections resulted in the same performance. Also, the architecture of modified ConVIRT is very similar to the BioVIL architecture but the loss functions are different and we expected the same results as in the baseline model since that is also using both the sections.

6. Discussion and Limitations

We conclude that both sections contribute equally and one possible reason is that Impressions section summarizes the most important ideas from the Findings section such that the learned text representations are similar. Another possible limitation is that we filtered out reports containing only either Finding or Impressions, so the training set decreased from 337k to 225k due to this preprocessing step. We also find that our Modified ConVIRT model does not dramatically improve on the information from the report sections, and more work would need to be done to explore **RQ2**.

A potential issue we have is the labeler used in ConVIRT paper introduced in section 4.1. Even if the labeler has a good performance, the accuracy of labeling an image is not 100%. Without help from professional clinicians and their annotations, we were not able to confirm if the labels assigned by the labeler are the ground truth. We hope, in the future, there is more manually annotated data in order to ensure the accuracy and performance of the model.

We also encountered some other limitations relating to computing resources. Due to issues of parallelizing our training jobs with multiple GPUs, we were able to train only 30 epochs whereas the model was trained for 200 epochs in the paper, so it is reasonable that our models are underfitting the data. Furthermore, we had may interruptions with our workflow due to low utilization of the GPU during pre-training. We observed that our GPU utilization was noisy and sometimes resulted in jobs being killed pre-maturely, and we could not increase the batch size as the GPU memory was maximized.

7. Contributions

An overview of the team's contributions is summarized below and a link to our code is provided[1]:

- **Xu**

  Text report pre-processing

  Hyperparameter tuning

  Initial presentation slide deck

  Modified ConVIRT loss

  Downstream evaluation experiments

  Final report

- **Carter**

  Code workflow

  Modified ConVIRT architecture

  Pre-training experiments

  Downstream evaluation experiments

  Presentation slides

  Final report

- **Valay**

  MIMIC-CXR access and pre-processing

  Pre-training experiments

[1] https://github.com/valay-shah/MLH_Fall22

SLURM management

Text report pre-processing

Presentation slides

Final report

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. Publicly available clinical. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*. Association for Computational Linguistics, 2019. doi: 10.18653/v1/w19-1909. URL https://doi.org/10.18653/v1/w19-1909.

Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, Hoifung Poon, and Ozan Oktay. Making the most of text semantics to improve biomedical vision-language processing, 2022. URL https://arxiv.org/abs/2204.09817.

Sedigheh Eslami, Gerard de Melo, and Christoph Meinel. Does clip benefit visual question answering in the medical domain as much as it does in the general domain? *arXiv preprint arXiv:2112.13906*, 2021.

M. Ghassemi, T. Naumann, P. Schulam, A. L. Beam, I. Y. Chen, and R. Ranganath. A Review of Challenges and Opportunities in Machine Learning for Health. *AMIA Jt Summits Transl Sci Proc*, 2020:191–200, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2016. doi: 10.1109/cvpr.2016.90. URL https://doi.org/10.1109/cvpr.2016.90.

Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, and Serena Yeung. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3942–3951, 2021.

Jeremy Irvin, Pranav Rajpurkar, 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, and Andrew Y. Ng. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:590–597, July 2019. doi: 10.1609/aaai.v33i01.3301590. URL https://doi.org/10.1609/aaai.v33i01.3301590.

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

Scientific Data, 6(1), December 2019. doi: 10.1038/s41597-019-0322-0. URL https://doi.org/10.1038/s41597-019-0322-0.

Rayan Krishnan, Pranav Rajpurkar, and Eric J. Topol. Self-supervised learning in medicine and healthcare. *Nature Biomedical Engineering*, August 2022. doi: 10.1038/s41551-022-00914-1. URL https://doi.org/10.1038/s41551-022-00914-1.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Pranav Rajpurkar, Jeremy Irvin, Aarti Bagul, Daisy Ding, Tony Duan, Hershel Mehta, Brandon Yang, Kaylie Zhu, Dillon Laird, Robyn L. Ball, Curtis Langlotz, Katie Shpanskaya, Matthew P. Lungren, and Andrew Y. Ng. MURA dataset: Towards radiologist-level abnormality detection in musculoskeletal radiographs. In *Medical Imaging with Deep Learning*, 2018. URL https://openreview.net/forum?id=r1Q98pjiG.

Hari Sowrirajan, Jingbo Yang, Andrew Y Ng, and Pranav Rajpurkar. Moco pretraining improves representation and transferability of chest x-ray models. In *Medical Imaging with Deep Learning*, pages 728–744. PMLR, 2021.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis Langlotz. Contrastive learning of medical visual representations from paired images and text, 2021. URL https://openreview.net/forum?id=T4gXBOXoIUr.