Reproduction Project Proposal: Contrastive Learning of Medical Visual Representations from Paired Images and Text

Junyoung Lee (jl298@illinois.edu)

March 30, 2025

1 Introduction

1.1 Problem statement

The paper [ZJM+22] addresses learning effective visual representations for medical images with limited annotated data. Traditional methods relying on expert annotations or ImageNet pretraining are inadequate because medical images require distinct fine-grained features. The paper introduces ConVIRT—a framework leveraging the pairing between medical images and their textual reports through bidirectional contrastive learning. This approach outperforms existing techniques, requiring only 10% of labeled training data compared to ImageNet initialization while achieving comparable or superior performance. Unlike image-only methods like SimCLR[TCH0a] and MoCo v2[XCH0b], ConVIRT efficiently utilizes semantic information from textual data, making it valuable for medical images with high inter-class similarity. This research is significant in healthcare settings where high-quality annotations are expensive. It has influenced subsequent frameworks like CLIP[OAC] and ALIGN[CJ21], demonstrating broader impact on multimodal learning.

2 Methodology

2.1 Specific approach

ConVIRT uses paired medical images and corresponding textual reports to learn in-domain visual representations. The image encoder is ResNet50, while the text encoder is a BERT-based model initialized with ClinicalBERT[EAM19] weights (only last 6 layers fine-tuned). Both encoders connect to projection networks (single-hidden-layer networks with ReLU activations) mapping features into a shared d-dimensional embedding space.

Key techniques include:

- **Contrastive Learning:** Maximizes agreement between true image-text pairs while distinguishing them from mismatched pairs using bidirectional InfoNCE loss. The contrastive loss employs a crucial temperature parameter ($\tau = 0.1$) and an asymmetric weighting ($\lambda = 0.75$) that emphasizes the image-to-text direction.
- **Data Augmentation:** Image augmentations include random cropping, horizontal flipping, affine transformation, brightness/contrast adjustments, and Gaussian blur. Text augmentation uses uniform sentence sampling to select representative spans from reports.
- **Pretraining Strategy:** Pretraining on large-scale paired datasets (MIMIC-CXR[MCJ] chest dataset with 217k pairs and Rhode Island Hospital musculoskeletal dataset with 48k pairs).

Evaluation metrics include:

- **Classification Tasks:** AUC for RSNA Pneumonia Detection[WW20a], CheXpert[JI19], and MURA[PR18]; accuracy for COVIDx[WW20b]. Both linear classification (with frozen encoder) and full fine-tuning settings are evaluated.

- **Retrieval Tasks:** Precision@k (k = 5, 10, 50) for zero-shot image-image and text-image retrieval using the specially created CheXpert[JI19] 8×200 Retrieval Dataset with expert radiologist annotations.

2.2 Novelty/Relevance/Hypotheses to be tested

ConVIRT's novelty lies in cross-modal contrastive learning to align image-text representations, enabling the model to learn clinically meaningful features directly from textual descriptions. The method demonstrates data efficiency by outperforming baselines across multiple tasks using only 10% of labeled data. The authors hypothesize that clinical texts contain semantic cues essential for learning subtle visual patterns, confirmed through expert-validated saliency maps showing ConVIRT's ability to accurately localize clinically relevant image regions compared to image-only methods trained on identical images. The approach reduces annotation burden by utilizing existing clinical documentation, facilitating integration into current workflows without altering standard practices.

2.3 Ablations/extensions planned

Although their hypothesis is legitimate, several improvements inspired by OpenAI's CLIP[OAC] can enhance ConVIRT:

- **Data Scaling:** Increasing the size and diversity of the medical dataset by incorporating additional radiology reports, pathology images, and clinical notes.
- **Zero-shot Transfer:** Evaluating performance on wider medical tasks like segmentation, diagnosis, and cross-domain retrieval.
- **Alternative Architectures:** Testing different text encoders (BioGPT[RL23], BioMed-RoBERTa [SGA20]) and image encoders (deeper CNNs, vision transformers).
- **Hyperparameter Optimization:** Studying effects of temperature settings, batch size, and projection head design.

3 Data Access and Implementation Details

For this reproduction project, the MIMIC-CXR-JPG[MCJ] dataset will be utilized as the primary data source, which requires completing the Physionet credentialing process due to its sensitive patient information content. The dataset provides both chest X-ray images and their associated radiology reports, with repository metadata files from the official github[CoO] guiding the organization of training and evaluation splits.

Since ConVIRT lacks an official public codebase, a new implementation will be developed referencing OpenAI's CLIP codebase[OAC], which shares similar contrastive learning principles. My implementation approach will address the paper's ambiguities regarding:

- Architectural details (projection head dimensions, activation functions, hidden layer dimensions)
- Data transformation specifics (crop size ranges, affine transformation parameters, blur kernel size)
- Training process details (precise learning rate scheduling, validation set composition)
- Implementation details (gradient clipping, regularization techniques, initialization methods)
- Other elements (data preprocessing, tokenization methods, sentence sampling methodology, loss function implementation)

The project will be proceeded with an incremental pipeline expansion design with 2-3 development stages, making design decisions and observing experimental results to ensure reproducibility while resolving unspecified parameters through careful experimentation, prioritizing fidelity to the original methodology.

References

[CJ21] Ye Xia Yi-Ting Chen Zarana Parekh Hieu Pham-Quoc V. Le Yunhsuan Sung Zhen Li Tom Duerig Chao Jia, Yinfei Yang. Scaling up visual and vision-language representation learning with noisy text supervision. *International Conference on Machine Learning 2021*, 2021.

[CoO] https://github.com/yuhaozhang/convirt.

[EAM19] William Boag Wei-HungWeng Di Jindi Tristan Naumann Emily Alsentzer, John Murphy and Matthew McDermott. Publicly available clinical bert embeddings. *In Proceedings of the 2nd Clinical Natural Language Processing Workshop*, 2019.

[JI19] Michael Ko Yifan Yu Silviana Ciurea-Ilcus Chris Chute-Henrik Marklund Behzad Haghgoo Robyn Ball Katie Shpanskaya et al. Jeremy Irvin, Pranav Rajpurkar. A large chest radiograph dataset with uncertainty labels and expert comparison. *In Proceedings of the AAAI Conference on Articial Intelligence*, 2019.

[MCJ] https://www.physionet.org/content/mimic-cxr-jpg/2.1.0/.

[OAC] https://openai.com/index/clip/.

[PR18] Aarti Bagul Daisy Ding Tony Duan-Hershel Mehta Brandon Yang Kaylie Zhu Dillon Laird Robyn L Ball et al. Pranav Rajpurkar, Jeremy Irvin. Mura: Large dataset for abnormality detection in musculoskeletal radiographs. *In 1st Conference on Medical Imaging with Deep Learning (MIDL)*, 2018.

[RL23] Yingce Xia Tao Qin Sheng Zhang-Hoifung Poon Tie-Yan Liu Renqian Luo, Liai Sun. Biogpt: Generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics, 2022*, 2023.

[SGA20] Swabha Swayamdipta Kyle Lo Iz Beltagy-Doug Downey Suchin Gururangan, Ana Marasovic and Noah A. Don't stop pretraining: Adapt language models to domains and tasks. *ArXiv: 2004.10964*, 2020.

[TCH0a] Mohammad Norouzi Ting Chen, Simon Kornblith and Georey Hinton. A simple framework for contrastive learning of visual representations. *International Conference on Machine Learning (ICML)*, 2020a.

[WW20a] Linda Wang and Alexander Wong. Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. *arXiv preprint arXiv:2003.09871*, 2020.

[WW20b] Linda Wang and Alexander Wong. Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. *arXiv preprint arXiv:2003.09871*, 2020.

[XCH0b] Ross Girshick Xinlei Chen, Haoqi Fan and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020b.

[ZJM+22] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D. Manning, and Curtis P. Langlotz. Contrastive learning of medical visual representations from paired images and text. *Machine Learning for Healthcare*, 182:1–24, 2022.