MLHC 2022

# Contrastive learning of medical visual representations from paired images and text

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D. Manning, Curtis P. Langlotz

Presented by Yufeng Zhang

10/30/23

# Outline

- Motivation & Goal
- Introduction
- Methodology
- Experiments & results
- Comments and concurrent works

# Motivation

**Currently limitations**

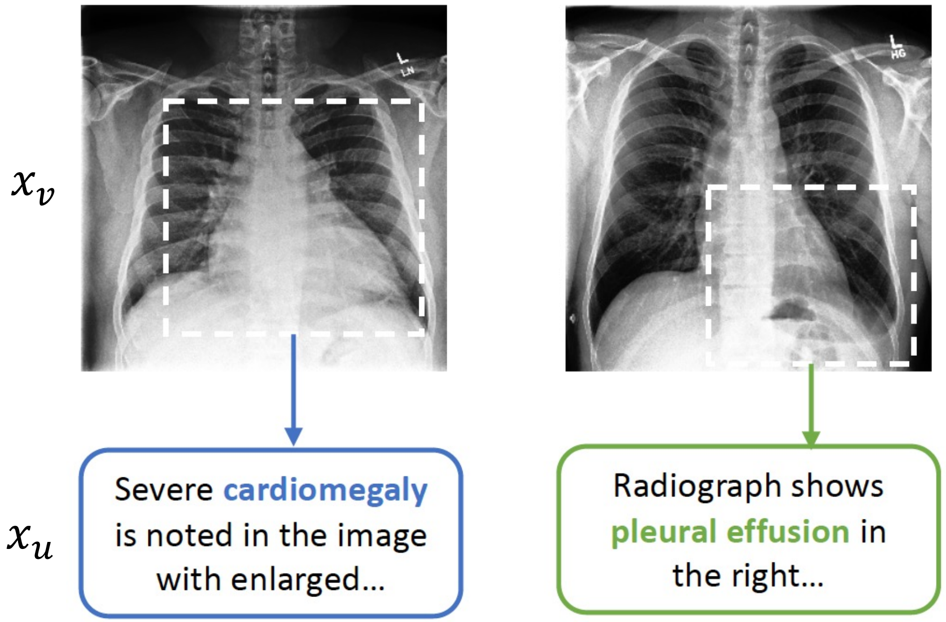
- Scarcity of human annotation
- High inter-class similarity

**ConVIRT**

- learn **better** medical visual representations exploiting naturally occurring **paired descriptive text**
- Bidirectional **contrastive objective** between two modalities

# Goal

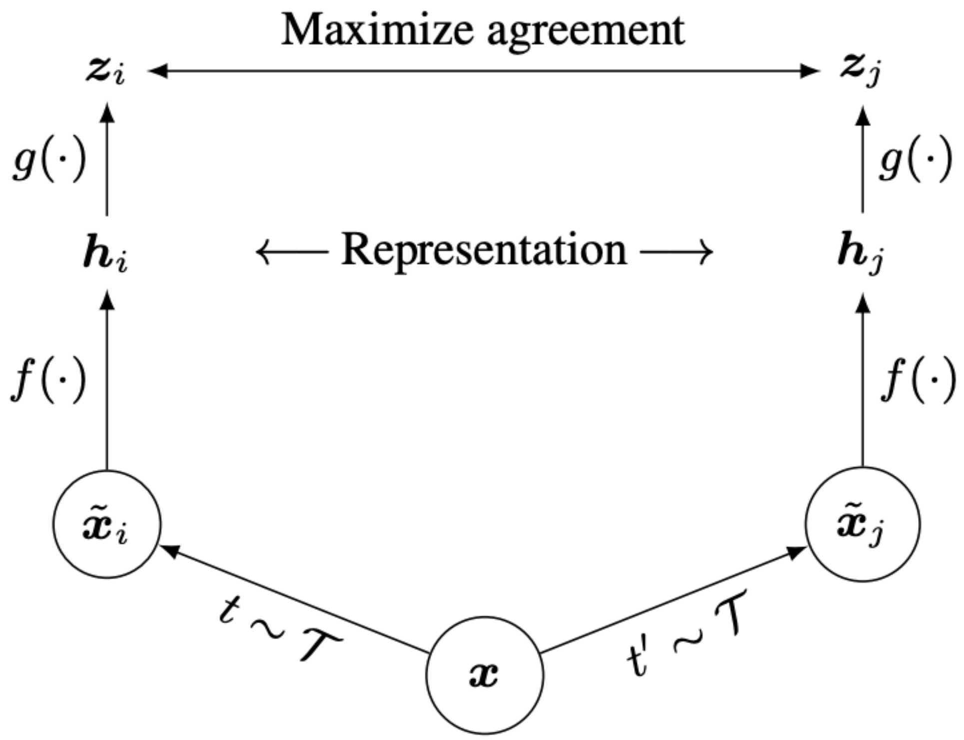
- Learn a parameterized image encoder function $f_v$, which maps an image to a fixed-dimensional vector
- $\mathrm{f_v}$ can be used for <u>classification</u> or <u>image retrieval</u>




Paired medical image-text

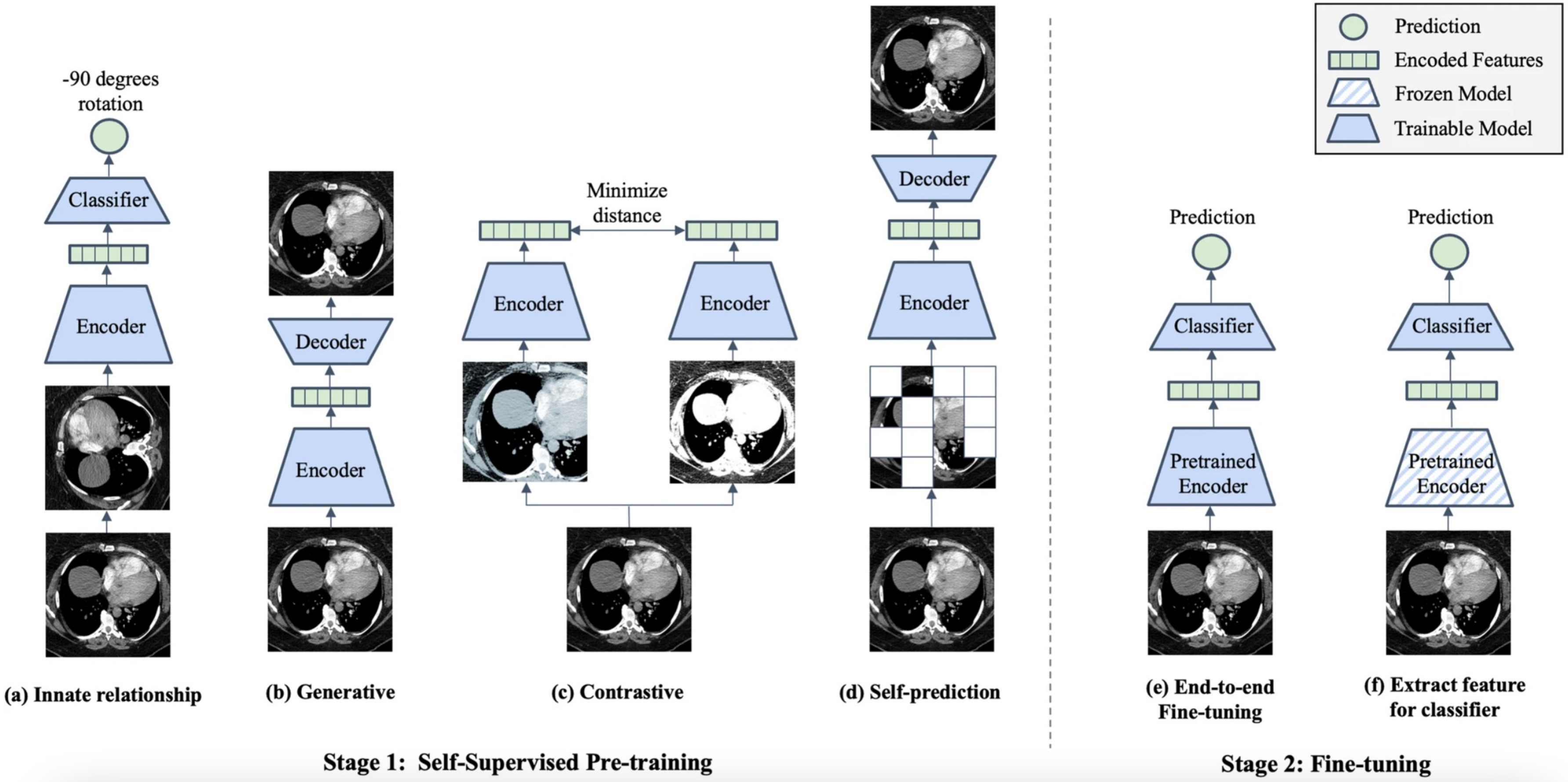
# Contrastive learning recap

- simCLR Framework




Chen, Ting, et al. "A simple framework for contrastive learning of visual representations." *International conference on machine learning*. PMLR, 2020.

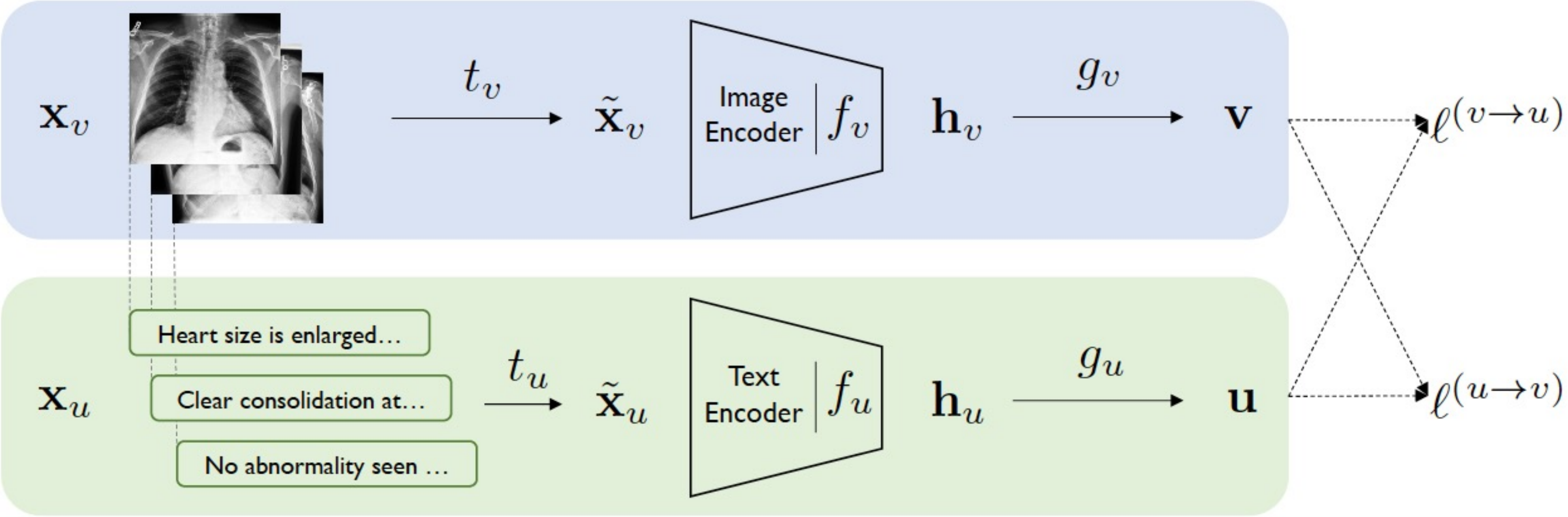
# Application in medical image classification




Huang, Shih-Cheng, et al. "Self-supervised learning for medical image classification: a systematic review and implementation guidelines." *NPJ Digital Medicine* 6.1 (2023): 74.

# Overview of  the proposed ConVIRT framework

standard ResNet50




BERT base encoder initialized with ClinicalBERT pretrained on the MIMIC clinical notes

# Training Loss

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

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

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

# Pretraining Dataset

- Chest X-ray image-text dataset:
  - from MIMIC-III
  - 217k pairs
- Bone image-text dataset:
  - from the Rhode Island Hospital system
  - 48k pairs

# Image classification task

- Dataset
  - RSNA Pneumonia Detection
  - CheXpert
  - *COVIDx*
  - MURA
- Evaluation method
  - Linear probing
  - Fine tuning

(*a*) Linear classification

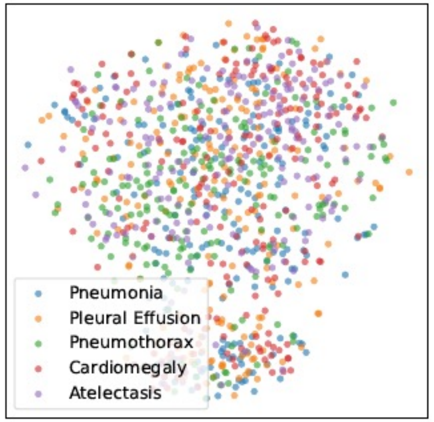
| Method | RSNA (AUC) 1% | 10% | all | CheXpert (AUC) 1% | 10% | all | COVIDx (Accu.) 10% | all | MURA (AUC) 1% | 10% | all |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *General initialization methods* | | | | | | | | | | | |
| Random Init. | 55.0 | 67.3 | 72.3 | 58.2 | 63.7 | 66.2 | 69.2 | 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

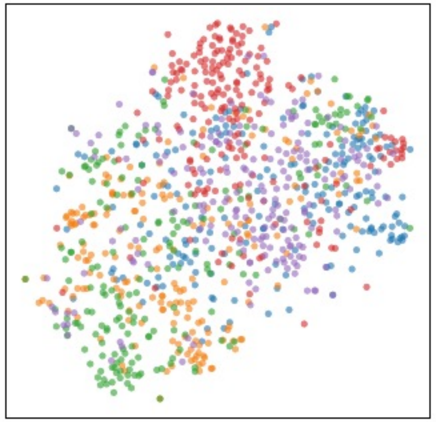
| Method | RSNA (AUC) 1% | 10% | all | CheXpert (AUC) 1% | 10% | all | COVIDx (Accu.) 10% | all | MURA (AUC) 1% | 10% | all |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *General initialization methods* | | | | | | | | | | | |
| Random Init. | 71.9 | 82.2 | 88.5 | 70.4 | 81.1 | 85.8 | 75.4 | 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** |

# Zero-shot image-image/text-image retrieval

| | Image-Image Retrieval | | | Text-Image Retrieval | | |
|---|---|---|---|---|---|---|
| Method | 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 | – | – | – |



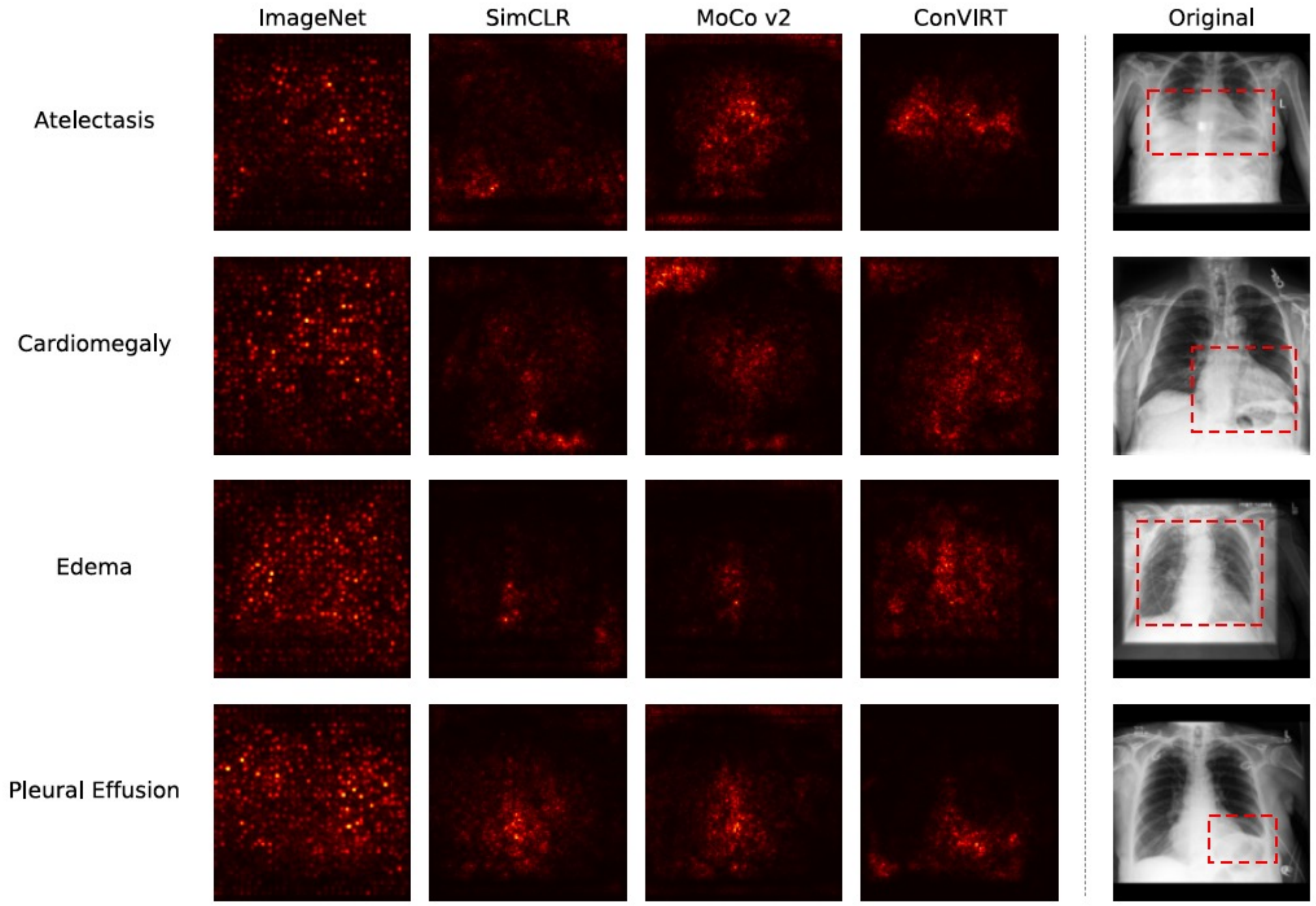

(*a*) ImageNet Pretraining



(*b*) ConVIRT Pretraining

# Comparison 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 |



# Existing Limitations and improvements

# Other visual language models

| CLIP | MedCLIP | ALBEF |
| --- | --- | --- |
| o Simplified framework<br>o Image and text encoder | o Unpaired medical images and text<br>o Semantic Matching Loss | o Momentum Distillation<br>o Triplet loss for pre-training objectives |

Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *International conference on machine learning*. PMLR, 2021.

Wang, Zifeng, et al. "Medclip: Contrastive learning from unpaired medical images and text." *arXiv preprint arXiv:2210.10163* (2022).

Li, Junnan, et al. "Align before fuse: Vision and language representation learning with momentum distillation." *Advances in neural information processing systems* 34 (2021): 9694-9705.

# Criticism

**Final Decision**

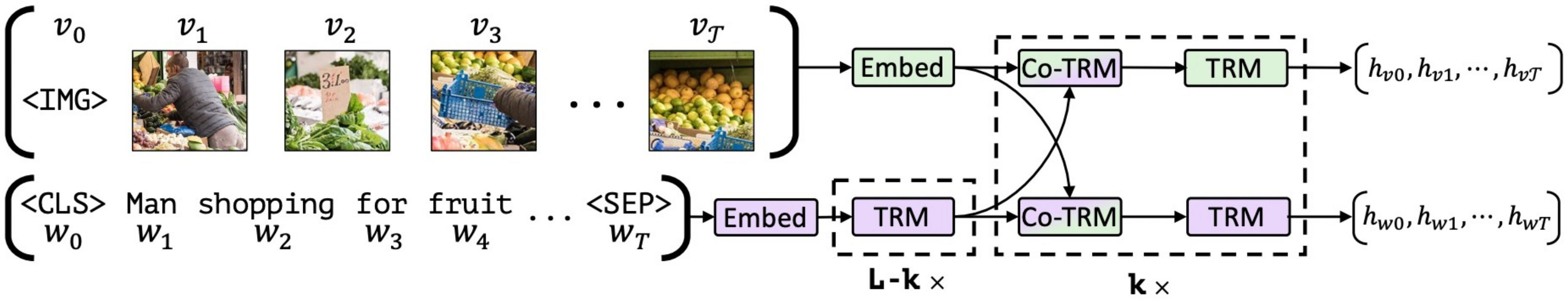
*ICLR 2021 Conference Program Chairs*

07 Jan 2021 (modified: 12 Jan 2021) ICLR 2021 Conference Paper2371 Decision Readers: Everyone

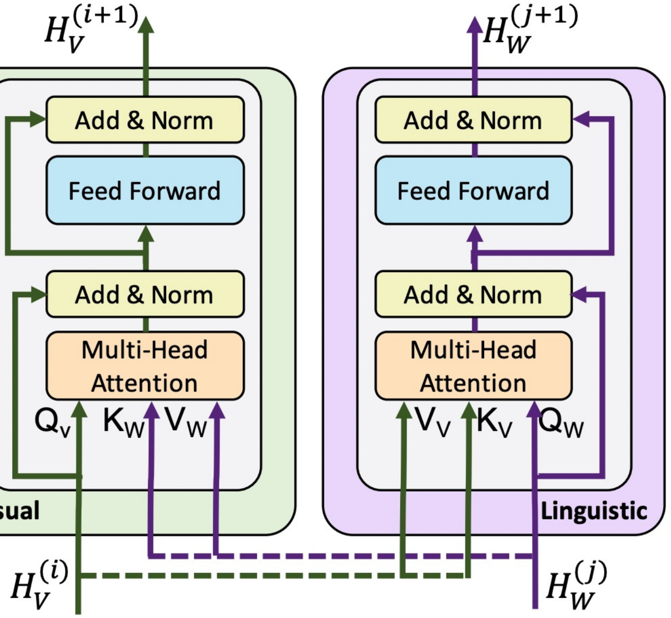
**Decision:** Reject

**Comment:** The proposed ConVIRT learns representations of medical data from paired image and text data. While the paper addresses a relevant problem, the reviewers agree that the method has limited novelty. Two reviewers find and that the experiments are not convincing. One reviewer notes that the paper does not compare to the state-of-the-art methods for the tasks.

# Comparison with ViLBERT




Proposed ViLBERT




Co-attention transformer layer

- Proposed structure
- Loss function

Lu, Jiasen, et al. "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks." *Advances in neural information processing systems* 32 (2019)

Thanks!

# Quiz

- Q1: How does contrastive learning contribute to mitigating the inter-class similarity in medical images?

- Q2: What distinguishes CLIP from ConVIRT in terms of major enhancements?