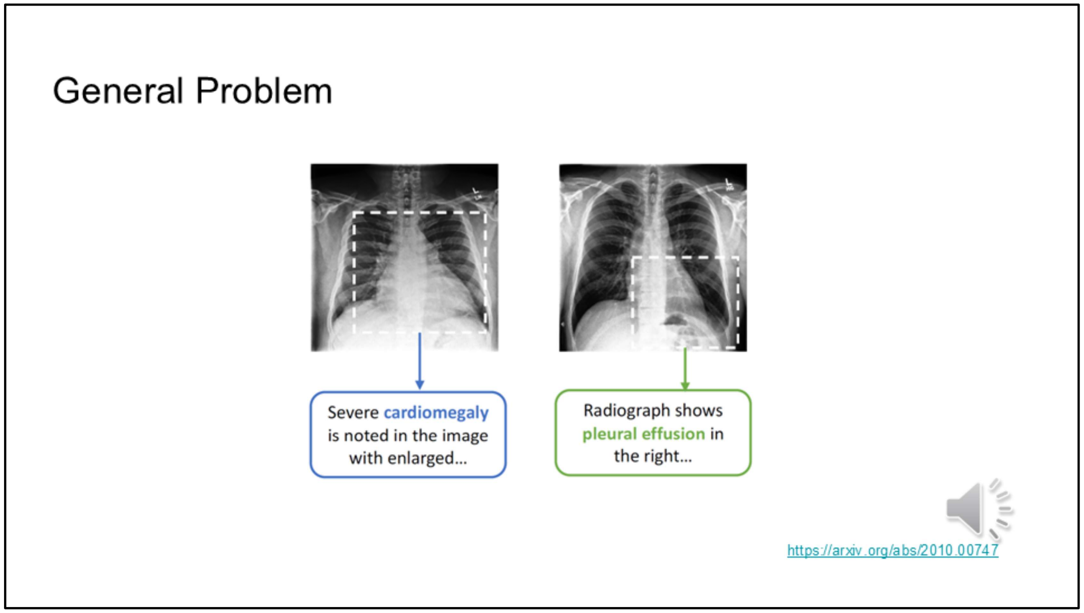
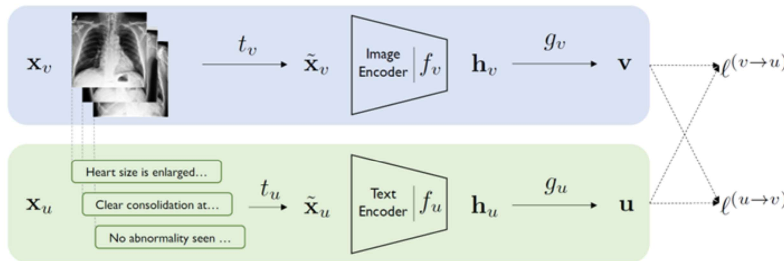
Hi everyone! I'm presenting my reproduction of the ConVIRT framework, Contrastive Learning of Medical Visual Representations from Paired Images and Text



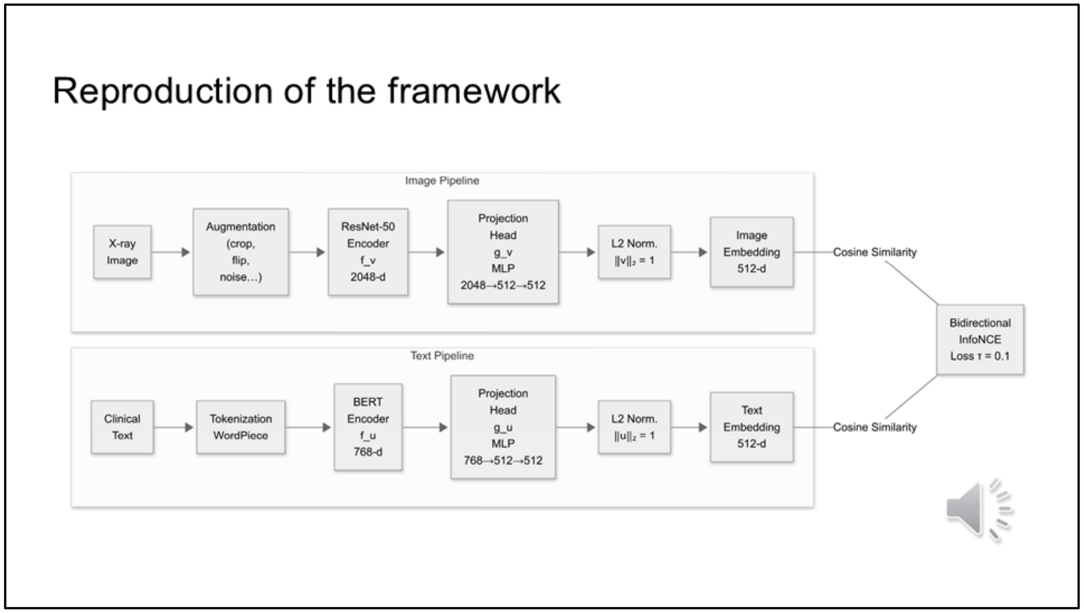
Medical image understanding remains challenging due to limited expert annotations and subtle visual differences between abnormalities. ConVIRT addresses this by leveraging naturally paired image-text data in an unsupervised contrastive learning framework.

The ConVIRT framework



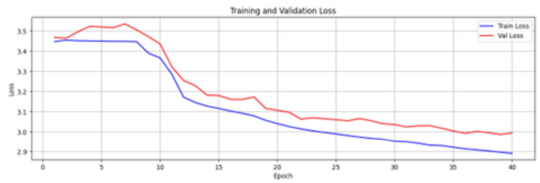
https://arxiv.org/abs/2010.00747

ConVIRT uses a bidirectional contrastive objective with two encoding pipelines: an image pipeline processing X-rays through a ResNet50 encoder, and a text pipeline processing radiology reports through a BERT encoder. Both modalities are projected into a shared space for contrastive learning.



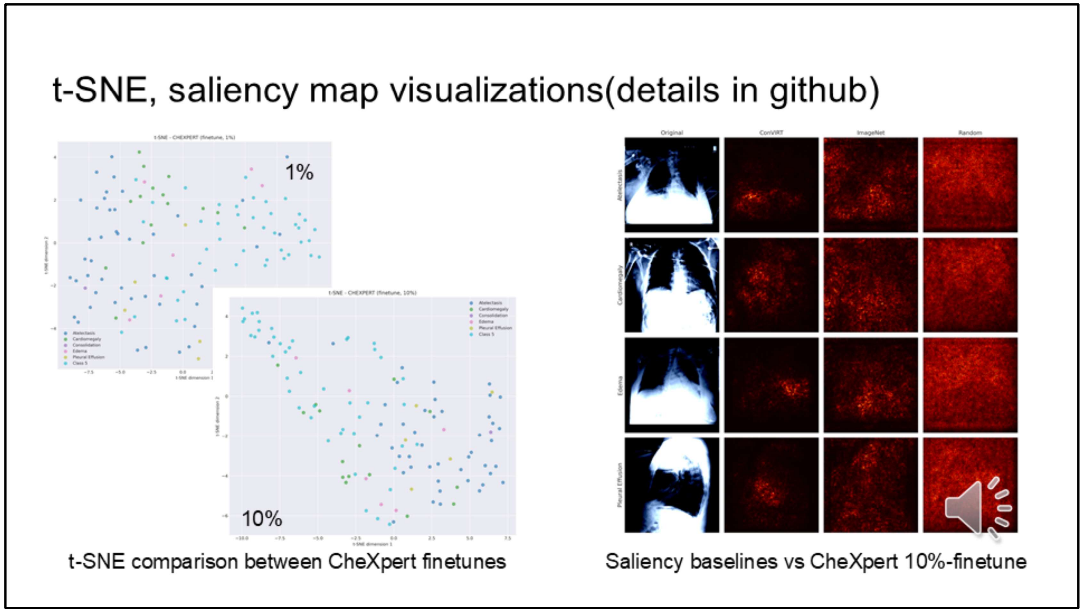
My implementation uses PyHealth with ResNet-50 and ClinicalBERT encoders, followed by projection heads that create 512-dimensional embeddings. The bidirectional InfoNCE loss drives the learning process by maximizing agreement between true pairs while pushing apart mismatched ones.

Training and Results



Dataset	Mode	Training	Metric	Paper	Repro	Diff
RSNA	Linear	1%	AUC	90.7	77.8	-12.9
RSNA	Linear	10%	AUC	91.7	82.5	-9.2
RSNA	Finetune	1%	AUC	88.8	77.8	-11.0
RSNA	Finetune	10%	AUC	91.5	82.7	-8.8
CheXpert	Linear	1%	AUC	85.9	77.5	-8.4
CheXpert	Linear	10%	AUC	86.8	79.8	-7.0
CheXpert	Finetune	1%	AUC	87.0	81.0	-6.0
CheXpert	**Finetune**	**10%**	**AUC**	**88.1**	**87.0**	**-1.1**
COVIDx	Linear	10%	Accuracy	85.9	66.3	-19.6
COVIDx	**Finetune**	**10%**	**Accuracy**	**90.3**	**88.8**	**-1.5**
MURA	Linear	1%	AUC	81.2	46.4	-34.8
MURA	Linear	10%	AUC	85.1	60.0	-25.1
MURA	Finetune	1%	AUC	81.3	52.4	-28.9
MURA	**Finetune**	**10%**	**AUC**	**86.5**	**80.8**	**-5.7**

Due to computational constraints, I trained for 40 epochs rather than the original 200. The closest performance to the original paper came in fine-tuning with 10% training data - just a 1.1% difference for CheXpert where we achieved 87.0% AUC compared to their 88.1%, and a 1.5% difference for COVIDx where we reached 88.8% accuracy versus their 90.3%. The largest discrepancies appeared in MURA evaluations, likely because my implementation didn't use the bone image dataset mentioned in the original paper.



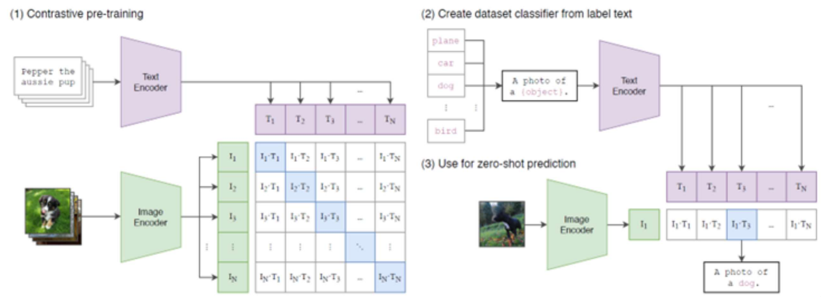
The t-SNE visualizations show better pathology clustering with 10% training data. Similarly, saliency maps reveal that ConVIRT focuses on anatomically relevant regions, like heart borders for cardiomegaly and lung bases for pleural effusion - unlike ImageNet-pretrained models that show scattered attention patterns.

Downstream Zero-shot retrieval evaluation

Method	Image-Image Retrieval			Text-Image Retrieval		
	Prec@5	Prec@10	Prec@50	Prec@5	Prec@10	Prec@50
Random (Paper)	12.5	12.5	12.5	12.5	12.5	12.5
ImageNet (Paper)	14.8	14.4	15.0	-	-	-
ConVIRT (Paper)	45.0	42.9	35.7	60.0	57.5	48.8
ConVIRT (Repro)	26.7	26.0	24.2	29.0	24.5	23.4

For zero-shot retrieval, we achieved 26.7% Precision@5 for image-image retrieval and 29.0% for text-image retrieval. While lower than the original paper's 45.0% and 60.0%, these results still significantly outperform the random baseline of 12.5%, confirming the model's effectiveness despite shorter training.

Next steps based on CLIP



https://arxiv.org/pdf/2103.00020v1

Looking forward, ConVIRT has inspired influential work like CLIP. Future directions include scaling to larger datasets, using more advanced text encoders, and implementing prompt engineering for better image-text alignment. This reproduction confirmed that ConVIRT's approach is effective for medical imaging tasks, even with limited training resources.

Thank you.

Github repository: https://github.com/jl298/repro-convirt

Thank you for your 4minute-attention!