A Reality Check of Vision-Language Pre-training in Radiology: Have We Progressed Using Text?

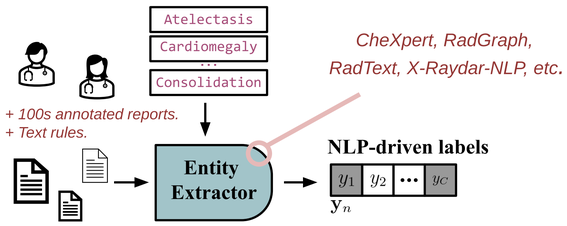
Julio Silva-Rodríguez · Jose Dolz · Ismail Ben Ayed - ÉTS Montréal

IPMI 2025

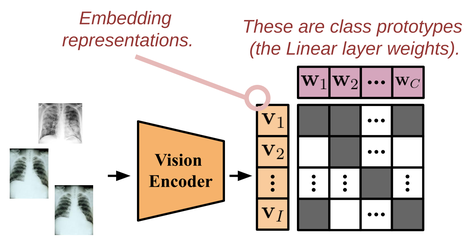
1 *(start here)*

LEARNING TRANSFERABLE ENCODERS FOR CXRs

- **Classical perspective:**

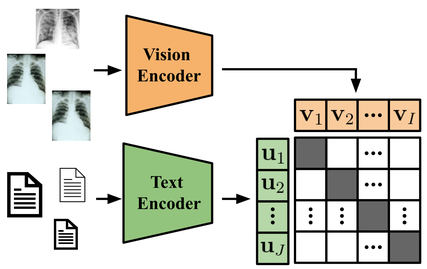


*(a) Disease/findings extraction From clinical **text reports***

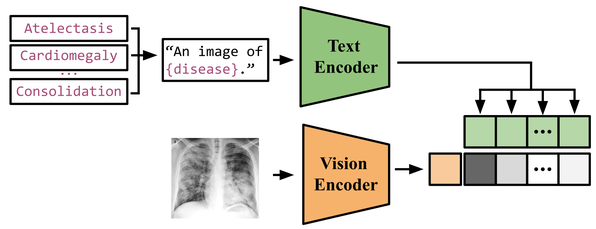


*(b) **Unimodal training** based on (noisy) NLP-driven **labels***

- **Vision-language models (VLMs):**



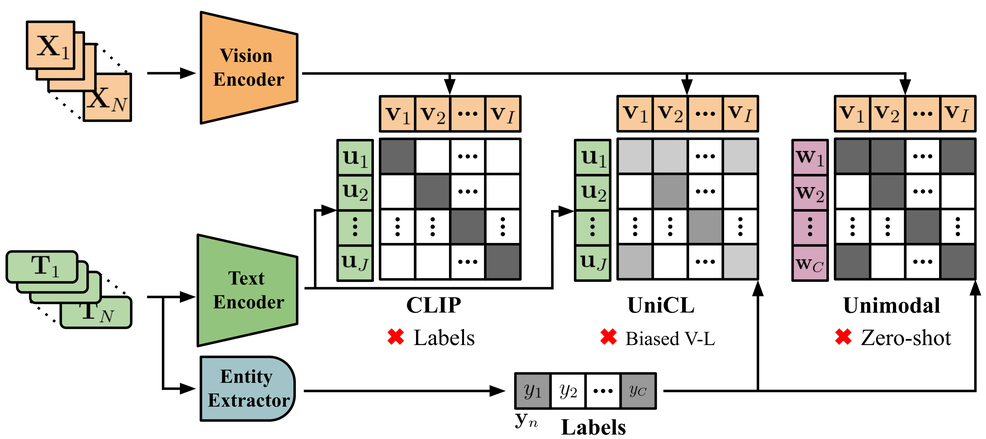
*(a) Large-scale pre-training with (noisy) **text supervision***



(b) Zero-shot / Linear probe transfer

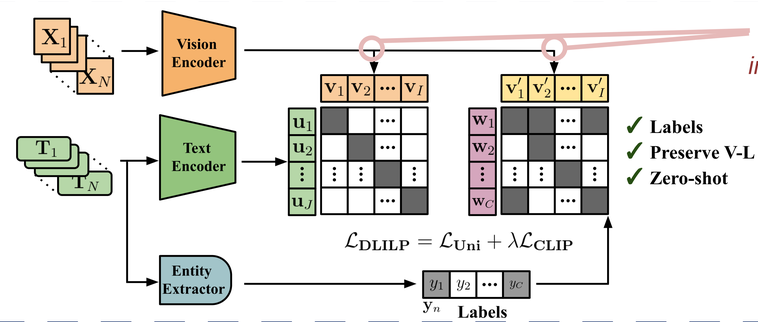
Since 2021: CONVIRT, GloriA, BioVil, MedCLIP, MedKLIP, KED, CXR-CLIP, etc.

BUT…
- **Have we really made progress using VLMs for transfer learning?**
- **How to make better use of both textual and label information?**



- **VLMs do not do magic: their zero-shot transfer capabilities depend on pre-training concept frecuency.**
- **Unimodal pre-training could lead to better models by integrating fine-grained labels – extracted from text reports trough efficient NLP.**

In Nutshell



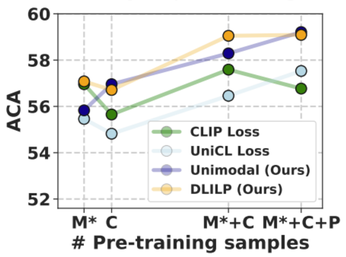
DLILP (*Our solution*)

2

3

1: "*Unimodal leads to more scalable transferability than existing vision-language models.*"
2: "*Label alignment in VLMs might risk biased joint representations.*"
3: "*The zero-shot capabilities of CXR vision-language models have been overestimated.*"

Scalabitily to pre-training data



Known (labeled) *vs. Novel*

Pre-training	$\mathcal{B}$	$\mathcal{N}$	Avg.
(a) Zero-shot generalization			
CLIP Loss [29]	58.61	**30.65**	44.63
UniCL Loss [46]	64.83	25.65	45.24
Unimodal	**64.92**	-	-
DLILP	64.08	30.10	**47.09**
(b) Linear probing ($K = 16$)			
CLIP Loss [29]	64.10	35.35	49.73
UniCL Loss [46]	63.53	32.53	48.03
Unimodal	65.41	35.32	50.37
DLILP	**66.45**	**36.50**	**51.48**

Zero-shot open-set generalization (COVID is not present in pre-training)

Pre-training	2-class Name	2-class Desc.	4-class Name	4-class Desc.
MedCLIP [43]	74.1	78.8	40.5	42.9
MedKLIP [29]	51.8	82.9	20.2	32.5
CLIP Loss [29]	69.6	74.2	32.7	48.8
UniCL Loss [46]	80.5	83.7	45.5	44.8
Unimodal	-	**85.1**	-	**51.6**
DLILP	77.0	81.6	36.6	50.0

COVID: "*the presence of patchy or confluent, band like ground-glass **opacity** or **consolidation***"

Pre-train	#Imgs	Text	#C	Categories
CheXpert (C)[14]	191,026	-	14	[NF, ECard, Card, LLes, LOp, Edem, Cons,
MIMIC (M)[17]	154,595	✓	14	PnMo, Atel, PnTh, PlEff, PlOt, Fract, Dev]
PadChest (P)[4]	96,201	-	84	(see code)

Evaluation	#Train	#Test	#C	Categories
CheXpert$_{5\times200}$	1,000	1,000	5	[Atel, Card, Cons, Edem, PlEff]
MIMIC$_{5\times200}$	1,000	1,000	5	[Atel, Card, Cons, Edem, PlEff]
RSNA [35]	8,400	3,600	2	[NF, PnMo]
SSIM [36]	4800	1200	2	[NF, PnTh]
COVID [5,31]	1,200	4,000	4	[Normal, COVID, N-COVID PnMo, LOp]
NIH-LT[46,11]	920	920	20	[Atel, Card, PlEff, Inf, Mass, Nod, PnMo, PnTh, Cons, Edem, Emph, Fib, PlThi, PnPer, PnMed, SubEm, TAor, CalAor, NF]
VinDr [27]	2,000	2,000	5	[NF, Bro, BrPn, BrLi, PnMo]

Zero-shot in VLMs vs. Unimodal