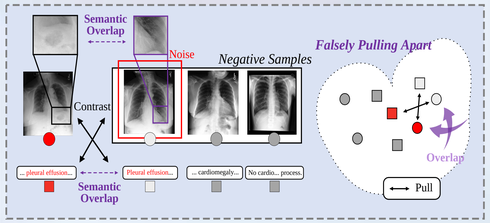
Knowledge Boosting: Rethinking Medical Contrastive Vision-Language Pre-Training

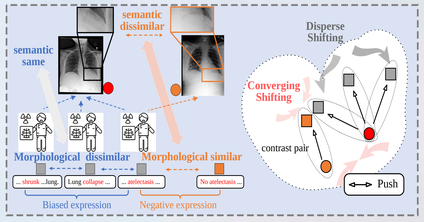
Xiaofei Chen[1], Yuting He[1], Cheng Xue[1], Rongjun Ge[2], Shuo Li[3] and Guanyu Yang[1,4,5*]

[1] Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications (Southeast University), Ministry of Education
[2] Nanjing University of Aeronautics and Astronautics
[3] Dept. of Biomedical Engineering, Case Western Reserve University, OH, USA
[4] Joint International Research Laboratory of Medical Information Processing, Southeast University, Nanjing 210096, China
[5] Centre de Recherche en Information Biomédicale Sino-Français (CRIBs)

MICCAI 2023 Vancouver CANADA

- **Semantic overlap problem in contrastive vision language pre-training.** (Within contrast pair)
- **Semantic shifting problem from biased and correlated expression.** (Between contrast pair)





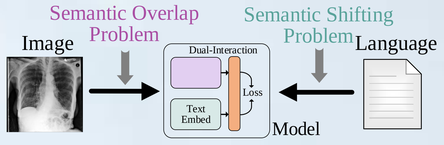
Motivation of Semantic Knowledge Enhancement for *semantic overlap*

- Apply an open-set knowledge representation to estimate negative noise within contrast pair and reduce it.

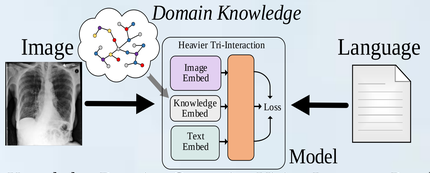
Motivation of Semantic Knowledge Guidance for *semantic shifting*

- Supplement semantic correlation and negative semantics with an comprehensive knowledge representation.

Our KoBo (**Kno**wledge **Boo**sting) vision-language pre-training framework innovates the traditional contrastive pre-training pipeline, inspired by semantic overlap problem and semantic shifting problem which is common in medical scene.

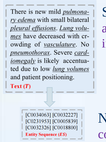


(a) Existing Constative Vision-Language Paradigm



(b) Knowledge-Boosting Constative Vision-Language Paradigm

Comprehensive with excellent graph encoding

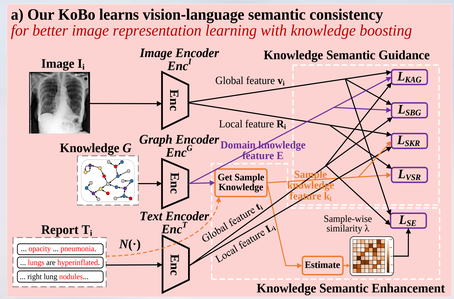


Sample-wise aggregate the knowledge in each report

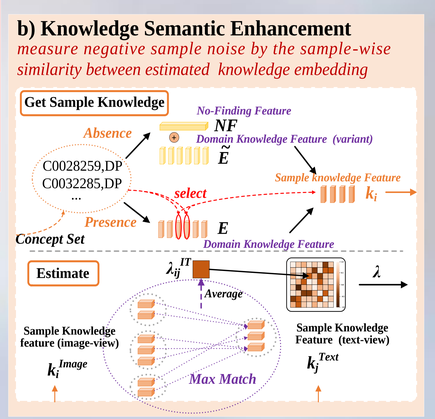
C0028259,DP
C0032285,DP
...

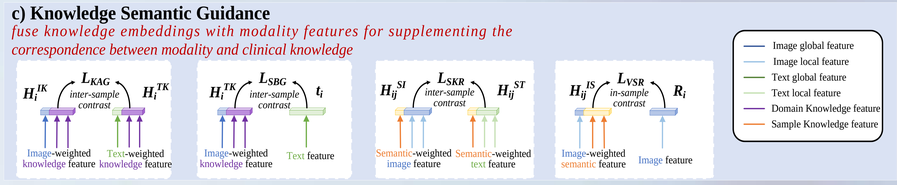
Negative-aware considering negative semantics

Open-set terminologies for knowledge is widely recognized and professional



- **Whole architecture: unified knowledge modeling and boosting of vision-language semantic consistency learning.**
- **KSE Module: utilize knowledge to estimate sample-wise negative noise in contrasive pair.**
- **KSG Module: fully guide the vision-language alignment with the correlated knowledge representation.**

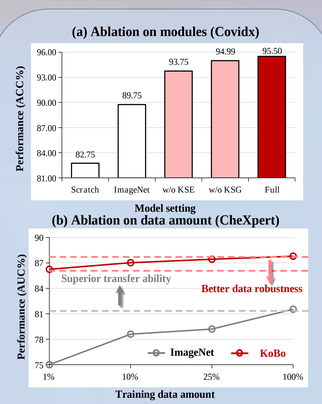




Method	Zero-shot CLS(V+L) CheXpert (Auroc)	RR(V) CheXpert5X200 (mAP)	RR(V+L) MIMIC (mAP)	SR(L) UMNSRS (Pearson)	SR(L) MIMIC (Pearson)	Few-shot-Frozen CLS(V) CheXpert (Auroc)	SEG(V) SIIM (Dice)	CLS(V) Covidx (Acc)
CLIP [18] (*)	0.4702	0.2544	0.7577	0.1985	-0.2879	0.5748	/	0.8975
ConVIRT [26]	0.8252	0.3808	**0.8482**	**0.2506**	0.1429	0.8548	0.4992	0.9475
Gloria [10]	0.8257	0.3875	0.8390	0.2294	0.1100	0.8492	0.5479	0.9250
MGCA [23]	0.8496	0.3906	0.8428	0.1889	0.1809	0.8616	0.5696	0.9375
MedCLIP [25] (*)	0.7805	**0.4298**	0.7258	0.2032	-0.1321	0.8214	0.5619	0.9325
KoBo	**0.8590**	0.3918	**0.8467**	**0.2563**	**0.3712**	**0.8628**	**0.6393**	**0.9550**
KoBo-Vit	**0.8635**	**0.4123**	0.8455	0.1824	**0.4229**	**0.8660**	**0.6554**	**0.9525**

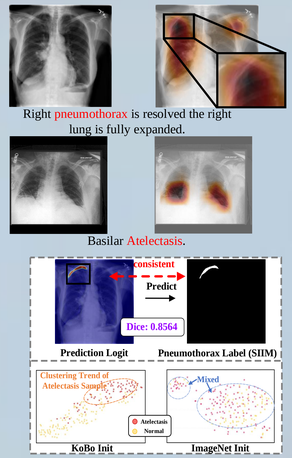
Result1:

Superiority in multiple task with unified pre-training: **state-of-art in classification, segmentation and semantic relatedness, top-2 in retrieval.**



Result2:

1. Effectiveness of module design: **KSE and KSG all contributes.**
2. Data robustness **when training data in find-tuning reduces to 1%: performance rarely decrease**



Result3:

1. Accurate localization of CAM: **vision semantic and language semantic is connected.**
2. Great cluster: **negative semantics is apart.**