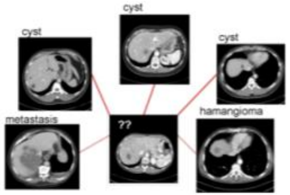
Text-guided visual representation learning for medical image retrieval systems

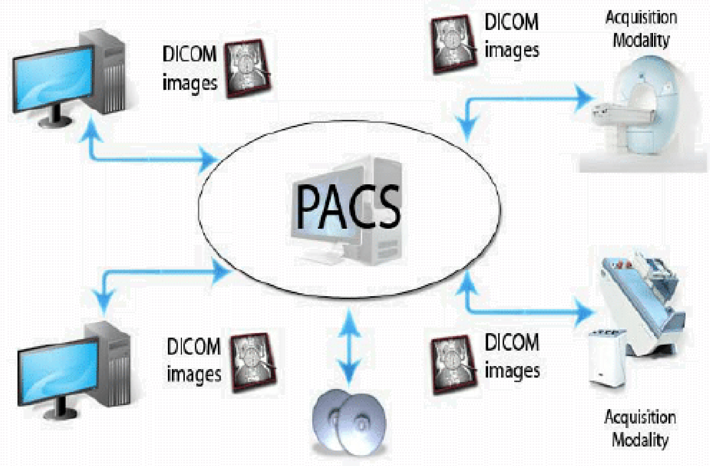
Guillaume Sérieys*, Camille Kurtz*, Laure Fournier**, Florence Cloppet*

*LIPADE, Université Paris Cité, Paris (France)
**Dept of Radiology, Hôpital Européen Georges Pompidou, Paris (France)

firstname.lastname@u-paris.fr

Introduction - Background & Motivation





Clinical routines

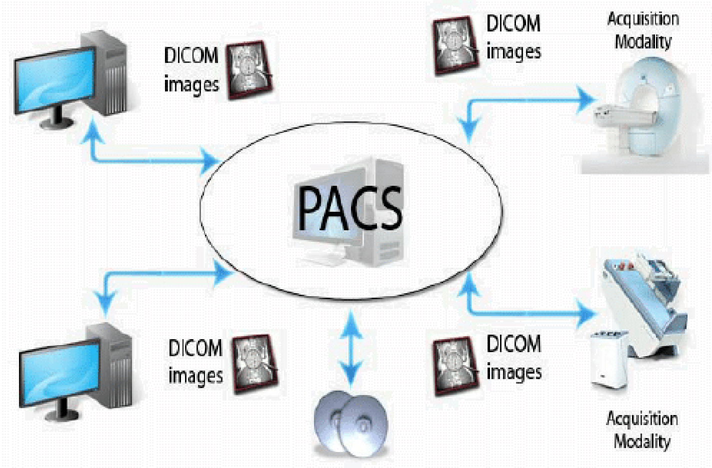
- Clinical and imaging data stored in PACS (Picture **Archiving** and Communication System)
- Physicians VS difficult case

PACS

- Search by keywords

CBIR (Content-Based Image Retrieval) for computer-aided diagnosis

Introduction - Background & Motivation



Clinical routines

- Clinical and imaging data stored in PACS (Picture **Archiving** and Communication System)
- Physicians VS difficult case

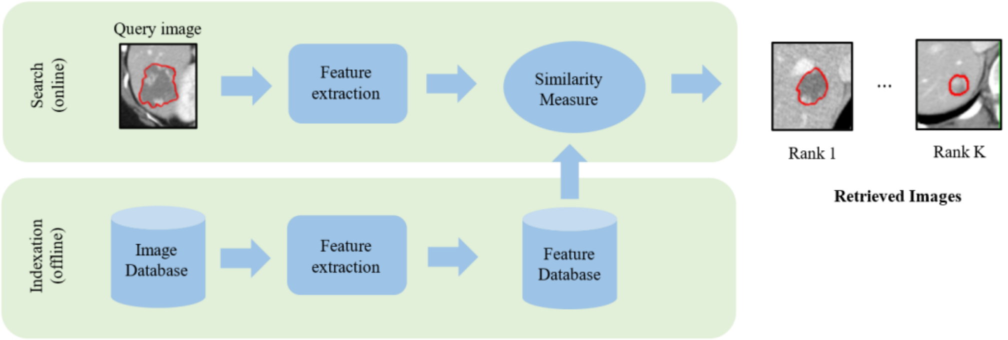
PACS

- Search by keywords

⇨

CBIR (Content-Based Image Retrieval) for computer-aided diagnosis

Introduction - Content-Based Image Retrieval



Approach	Examples	Advantages	Limitations
Hand-crafted descriptor	Color histograms	Describe images as a whole	Inaccurate
	Scale Invariant Feature Transform	Describe key points within images	High dimensionality
Distance metric learning	Contextual constraints	Simplicity	Inadequate for nonlinear data
	Kernel-based	Simplicity	Nonlinear data
Deep learning	Supervised approaches	High performance	High labelling cost
	Unsupervised approaches	Low labelling cost	Lower performances

S. R. Dubey, “A Decade Survey of Content Based Image Retrieval using Deep Learning”, 2021.

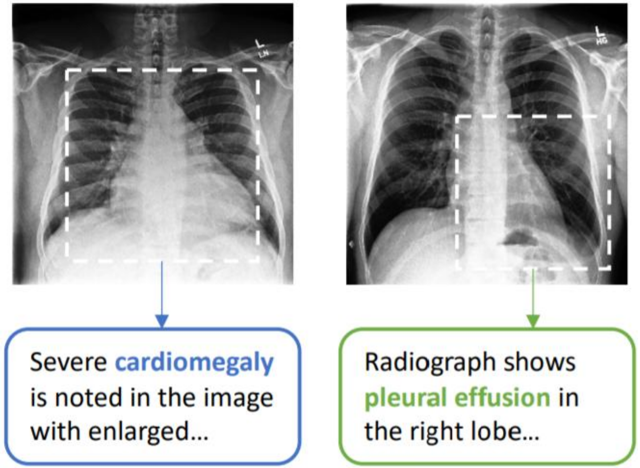
Introduction - Problematic

Feature Extraction

- Deep learning → Automatically learning features
- Large quantity of multimodal and multiparametric data

Medical Imaging specificities

- Large data quantity BUT difficult task: complex nature + scarcity of **labelled** data
- Fine-grained visual features ≠ Natural images
- Transferring model weights from ImageNet pretraining => suboptimal results on medical images



- Additional information in PACS: radiological reports (texts)
  - Automation of label extraction from reports is limited

  => **Expert crafted rules** to extract labels from reports are **inaccurate and domain-specific**

- **Learning visual representation from text supervision**

Y. Zhang et al., "Contrastive Learning of Medical Visual Representations from Paired Images and Text", Oct. 2020.

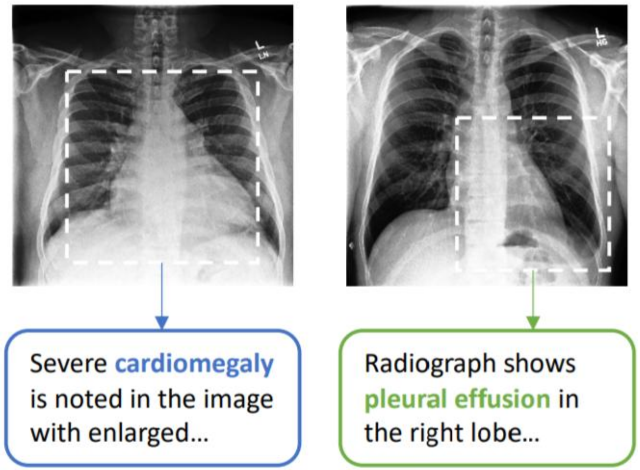
Introduction - Problematic

Feature Extraction

- Deep learning → Automatically learning features
- Large quantity of multimodal and multiparametric data

Medical Imaging specificities

- Large data quantity BUT difficult task: complex nature + scarcity of **labelled** data
- Fine-grained visual features ≠ Natural images
- Transferring model weights from ImageNet pretraining => suboptimal results on medical images



- Additional information in PACS: radiological reports (texts)
  - Automation of label extraction from reports is limited

  => **Expert crafted rules** to extract labels from reports are **inaccurate and domain-specific**

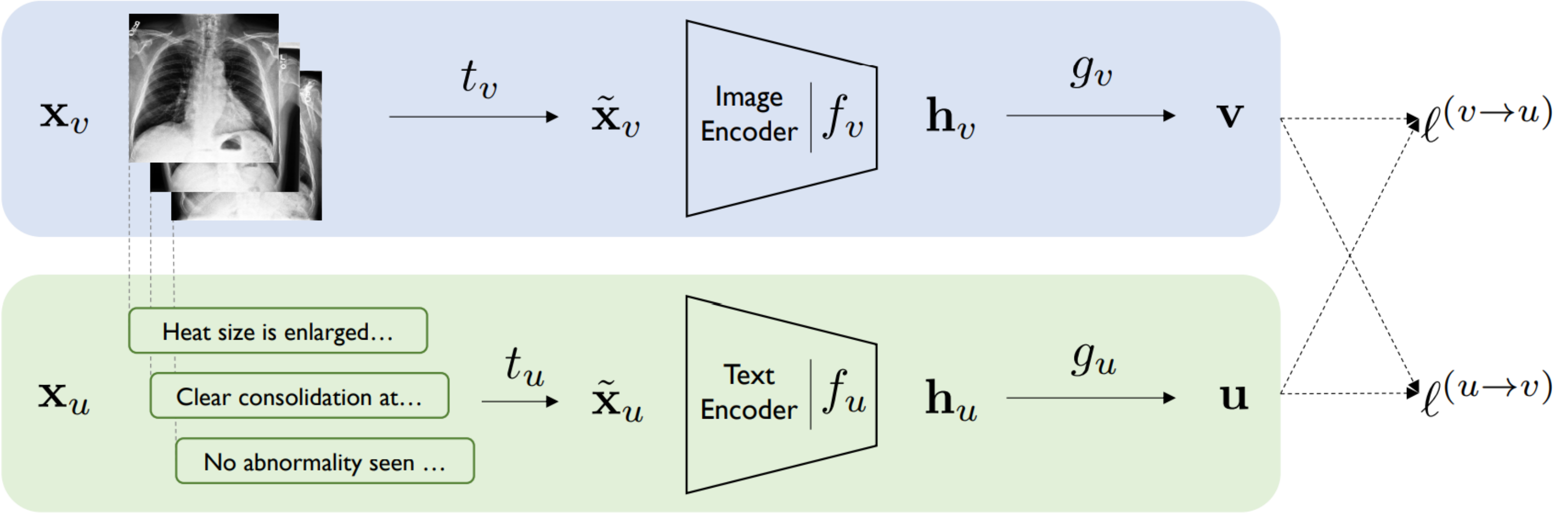
- **Learning visual representation from text supervision**

Y. Zhang et al., "Contrastive Learning of Medical Visual Representations from Paired Images and Text", Oct. 2020.

Methodology - Big Picture

Learning medical visual representation from text supervision

- ConVIRT framework => use positive pairs of image and text in a **contrastive learning fashion**

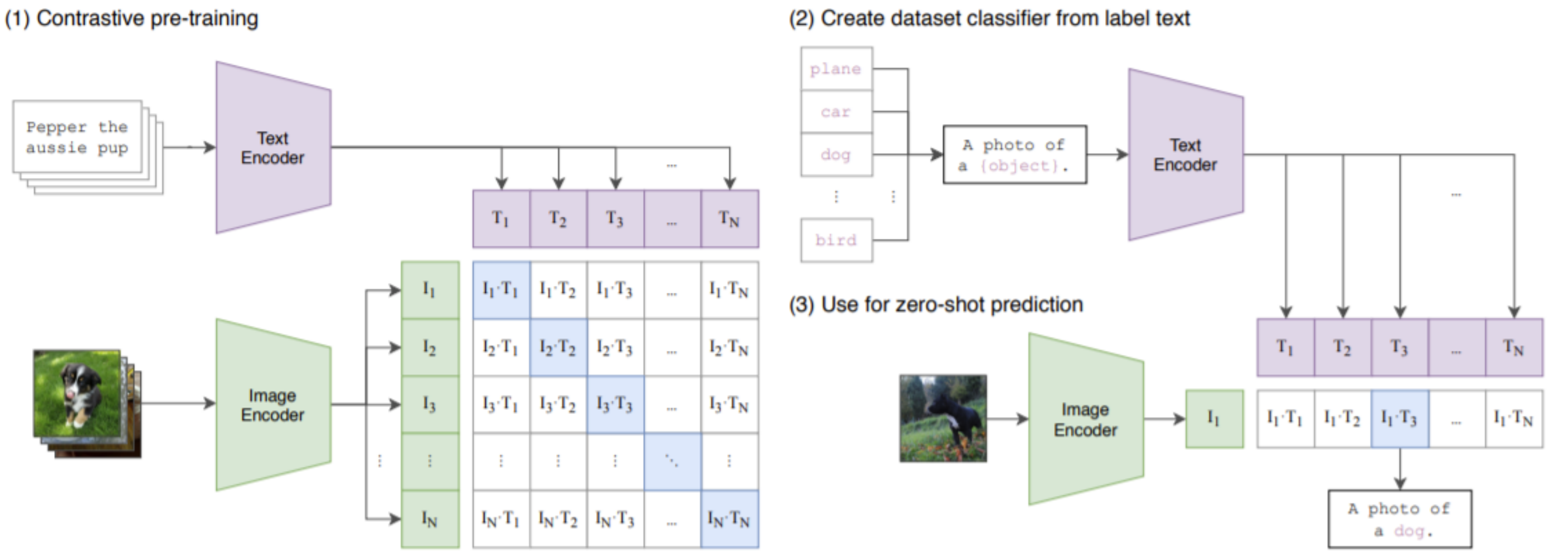


Y. Zhang et al., "Contrastive Learning of Medical Visual Representations from Paired Images and Text", Oct. 2020.

Methodology - Starting point

Learning medical visual representation from text supervision

- CLIP => Clinical CLIP

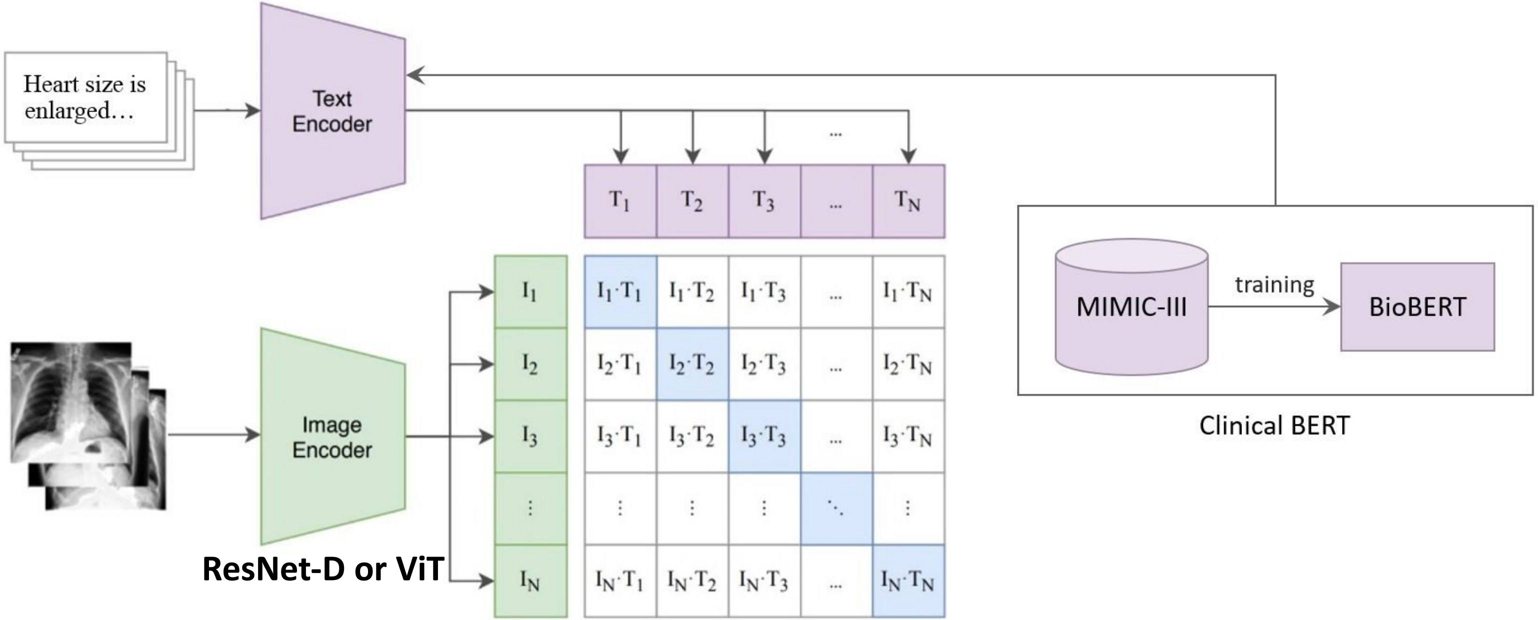


A. Radford et al., “Learning Transferable Visual Models From Natural Language Supervision”, Feb. 2021.

Methodology - Contribution

Learning medical visual representation from text supervision

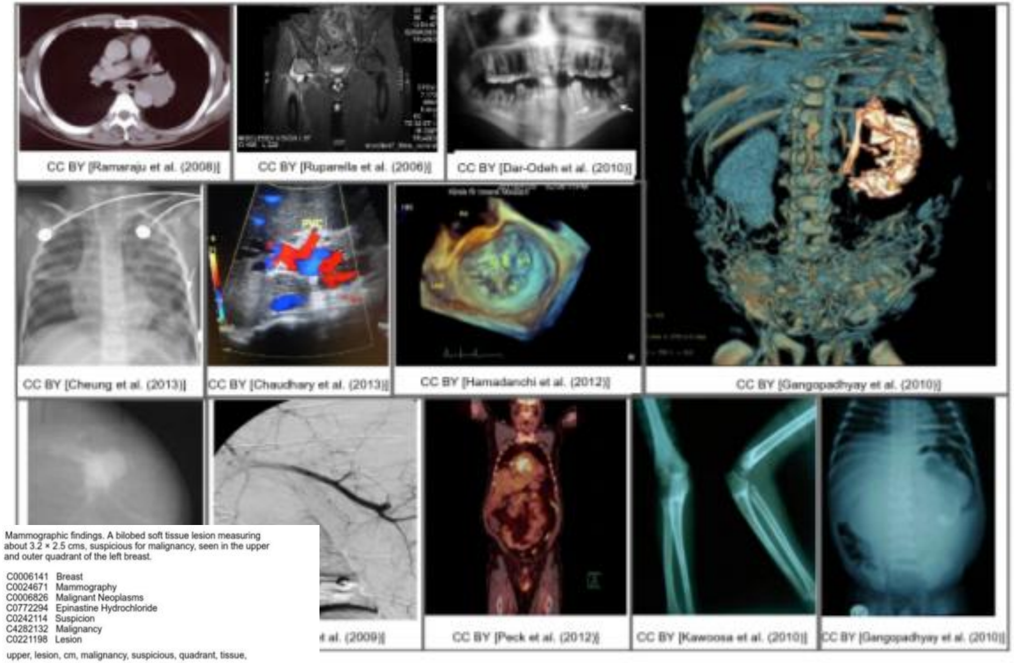
- CLIP => **Towards Clinical CLIP**

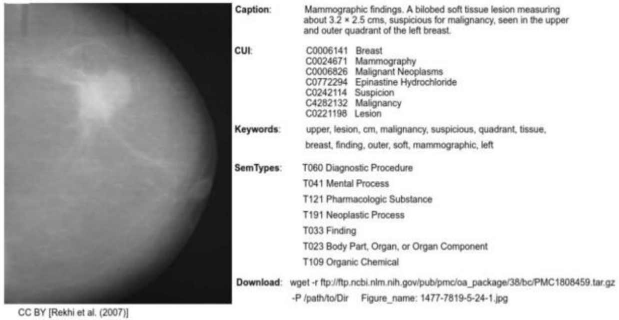


Learning representations from PubMed resources

Pre-Training Data set

- **ROCO (Radiology Objects in Context)**
    - **81,825 radiology images** with corresponding captions, keywords and UMLS (Unified Medical Language System) CUIs (Concept Unique Identifiers) and SemTypes (Semantic Types)
    - **Multimodal** image dataset (CT, X-Ray, PET, MRI, etc.)

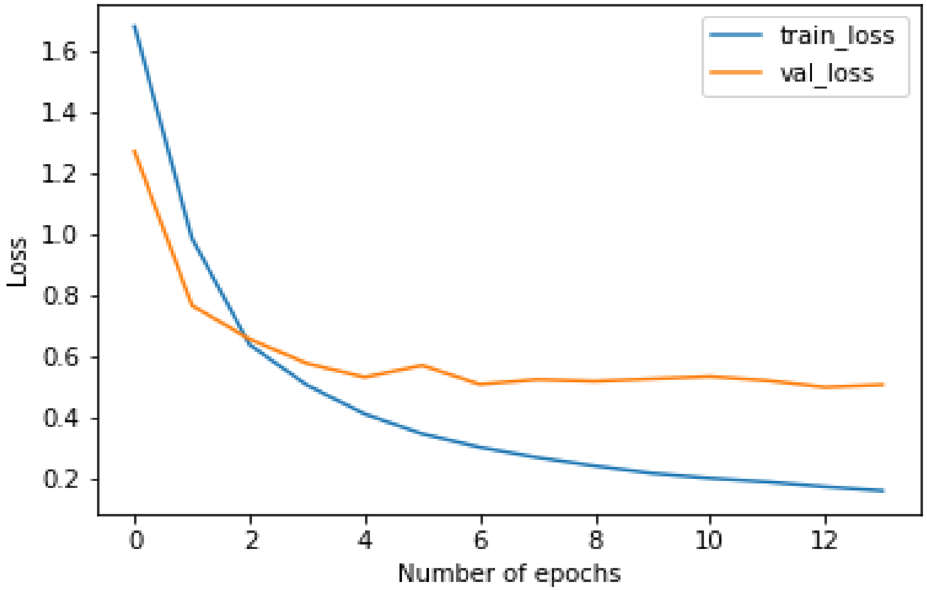




Learning representations from PubMed resources

Pre-Training protocol

- **Number of epochs**: 100
- **Learning rates**: 3e-6 to fine-tune CLIP, 3e-5 otherwise
- **Optimizer**: Adam



Evaluation on the CBIR task

Custom retrieval Data Set

- From ROCO dataset + expert annotations from Hôpital Européen Georges Pompidou
- 8x200 images
- Retrieval Settings
  - Organ
  - Modality
  - Organ + modality

Performance metric

- Precision at K (P@K) with K = number of retrieved images

CheXpert 8x200 retrieval Data set

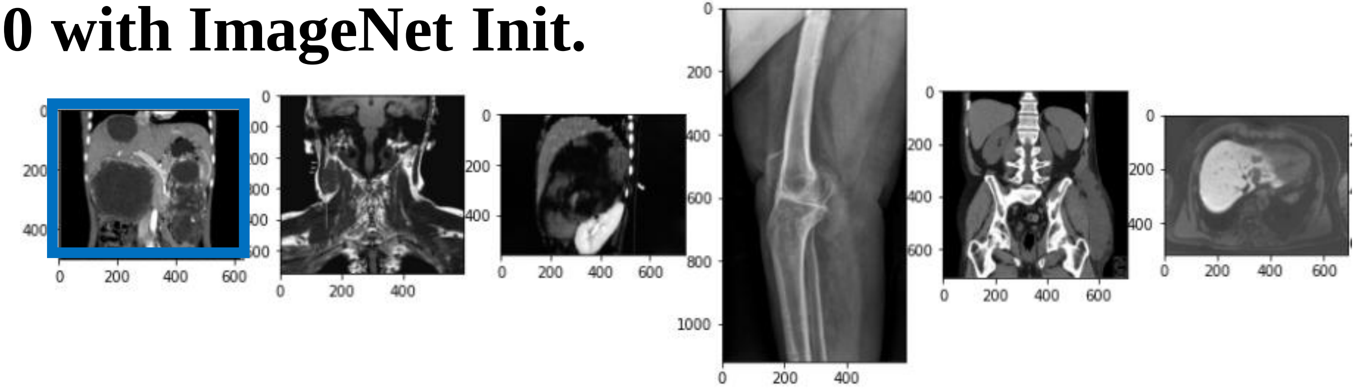
- 224 316 annotated chest radiographs from 65240 patients
- 8 independent categories of diagnosis

Y. Zhang et al., “Contrastive Learning of Medical Visual Representations from Paired Images and Text”, Oct. 2020.

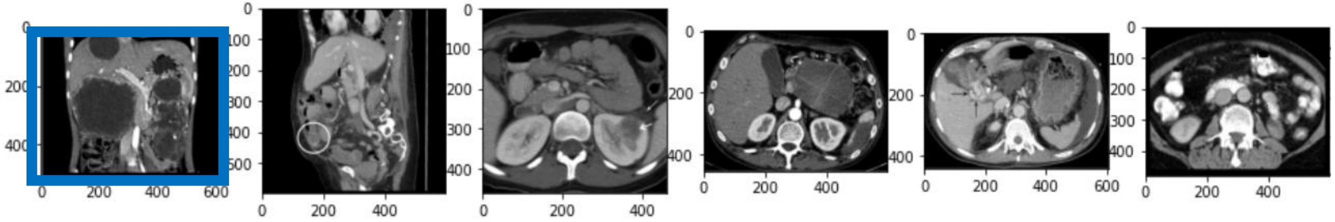
Evaluation - Qualitative results

"Coronal plain computed tomography image showing multiple large tumor masses with edge enhancement inside the abdominal cavity and liver."

ResNet-50 with ImageNet Init.



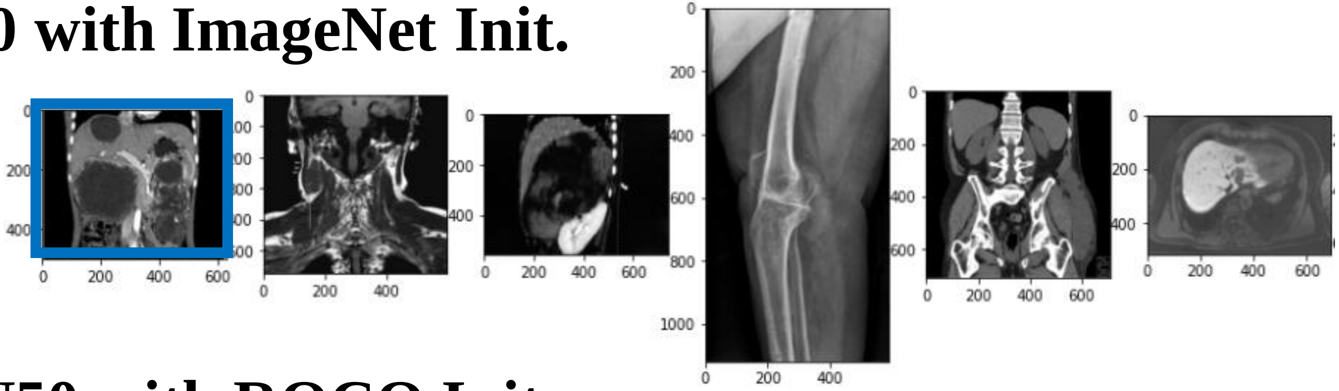
Initial CLIP-RN50



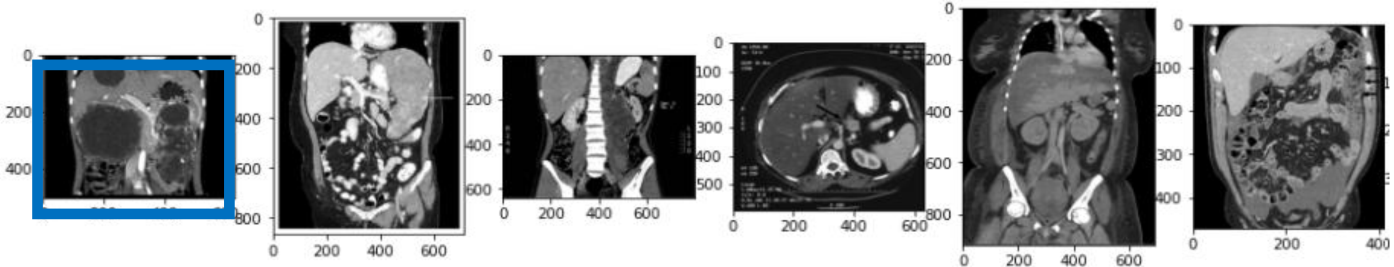
Evaluation - Qualitative results

“Coronal plain computed tomography image showing multiple large tumor masses with edge enhancement inside the abdominal cavity and liver.”

ResNet-50 with ImageNet Init.



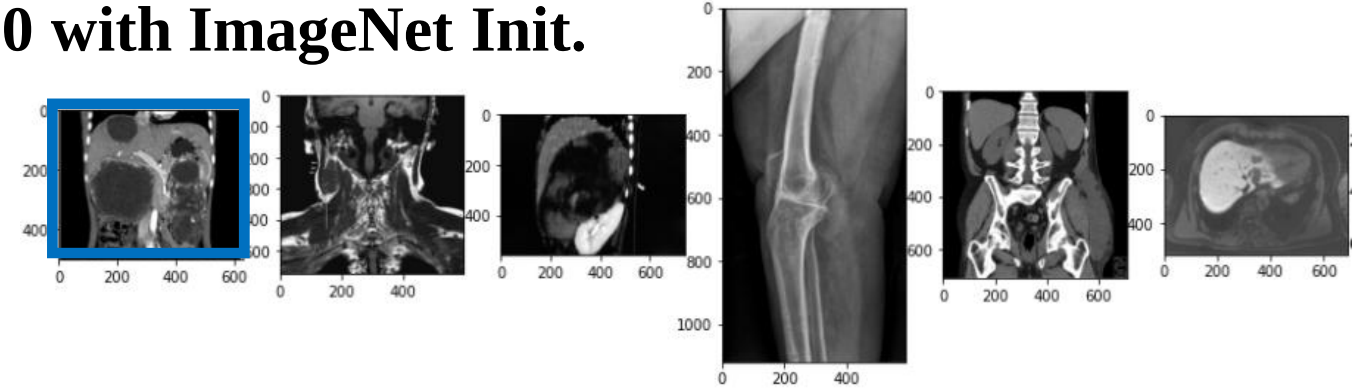
CLIP-RN50 with ROCO Init.



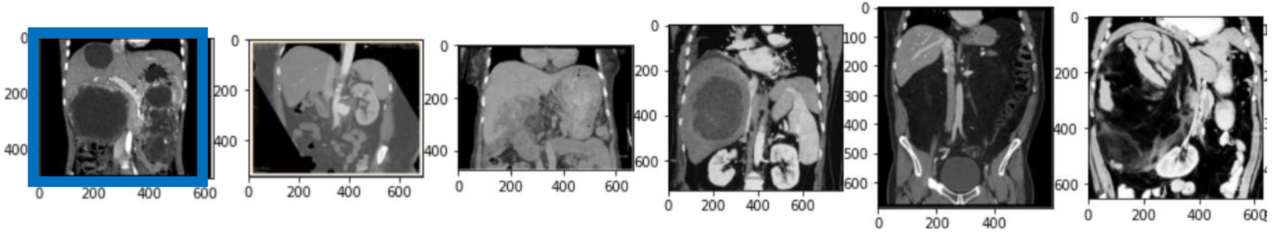
Evaluation - Qualitative results

"Coronal plain computed tomography image showing multiple large tumor masses with edge enhancement inside the abdominal cavity and liver."

ResNet-50 with ImageNet Init.



Clinical CLIP-RN50



Evaluation - Results

Method	ROCO $CUI@K$			Custom retrieval dataset $P@K$									CheXpert 8×200 $P@K$		
	@5	@10	@50	@5	@10	@30	@5	@10	@30	@5	@10	@30	@5	@10	@50
General init. methods				Modality			Organ			Modality/Organ					
Random	1.5	3.0	5.3	20.0	20.0	20.0	12.5	12.3	12.4	11.4	11.4	10.0	12.5	12.5	12.5
ImageNet	10.5	10.6	11.2	46.8	42.6	37.2	30.8	28.3	22.9	44.8	41.4	31.0	16.5	15.5	14.5
CLIP-RN50	14.9	15.1	16.5	84.0	81.2	73.7	50.4	42.5	33.8	60.0	62.9	51.3	17.0	17.0	15.3
CLIP-RN101	15.2	15.5	17.1	85.6	84.6	80.7	47.9	45.2	39.0	65.7	64.8	55.4	11.5	13.1	14.0
CLIP-RN50x4	16.6	17.0	18.5	85.6	85.2	80.5	48.8	45.6	39.4	65.7	65.2	58.4	16.5	16.4	15.7
CLIP-ViT-B/32	15.5	15.8	17.3	86.4	86.4	79.6	53.3	49.2	40.8	70.5	68.1	57.1	16.8	14.4	14.4
In-domain init. methods															
ConVIRT (random init.)	12.3	12.4	13.1	68.4	63.2	52.0	40.8	31.0	24.1	50.5	48.1	37.1	20.3	19.5	15.8
ConVIRT (ImageNet init.)	12.0	12.1	12.9	59.6	54.0	44.8	40.8	35.2	27.0	55.2	48.0	32.9	18.5	17.3	14.5
CLIP-RN50 (ROCO init.)	17.4	17.7	19.3	85.6	86.8	80.9	48.3	46.3	37.4	61.9	63.3	55.6	12.5	15.0	16.0
CLIP-RN101 (ROCO init.)	17.2	17.5	19.1	84.4	85.0	80.6	51.3	49.2	37.5	61.9	65.2	57.0	17.5	18.5	15.3
CLIP-RN50x4 (ROCO init.)	18.0	18.2	20.0	83.2	83.6	78.6	50.8	46.5	36.2	68.6	67.6	57.8	15.0	16.6	15.5
CLIP-ViT-B/32 (ROCO init.)	18.2	18.6	20.4	87.2	86.8	81.3	61.3	58.5	48.3	70.5	69.1	62.9	13.8	15.8	15.0
Clinical CLIP-RN50	18.3	18.9	20.9	**93.2**	**91.6**	**85.4**	65.4	60.2	50.4	69.5	**71.9**	64.6	18.5	17.9	**18.1**
Clinical CLIP-RN101	18.8	19.3	21.3	90.0	89.6	83.9	**65.8**	**62.1**	52.1	71.4	**71.9**	**64.9**	21.0	19.8	**18.1**
Clinical CLIP-RN50x4	**19.1**	**19.5**	**21.4**	**93.2**	**91.6**	83.0	63.8	**62.1**	**52.2**	**72.4**	**71.9**	64.6	**22.5**	**21.0**	17.9
Clinical CLIP-ViT-B/32	18.2	18.7	20.8	90.4	89.2	83.2	62.9	57.9	49.8	69.5	69.5	63.3	17.5	16.3	14.9

Conclusions & Perspectives

Take home message

- It is possible to learn / optimize a visual representation of medical images from weak text supervision => thanks to contrastive learning from pairs of image and text
- “Dormant” data from the medical imaging literature (PubMed) can be re-employed to supervise the learning of neuronal models
- Interest of in-domain text encoders such as Clinical BERT in the model pre-training

Next steps

- To evaluate the performance of our methods in more specific retrieval tasks but by fine-tuning the methods to the specific domain
- To deal with the multimodal aspect of medical imaging: multimodal variational autoencoders (MVAE) to learn a joint representation of multiple modalities
- Go to “real case” applications by considering radiological reports from PACS as originally intended

Y. Zhang et al., “Contrastive Learning of Medical Visual Representations from Paired Images and Text”, Oct. 2020.

Text-guided visual representation learning for medical image retrieval systems

Guillaume Sérieys, Camille Kurtz, Laure Fournier, Florence Cloppet

firstname.lastname@u-paris.fr

This work was supported by diiP, IdEx Université Paris Cité, ANR-18-IDEX-0001 and with access to the HPC resources of IDRIS (GENCI 2021-AD011012656)