MLMI Practical Course

Structured Report Generation

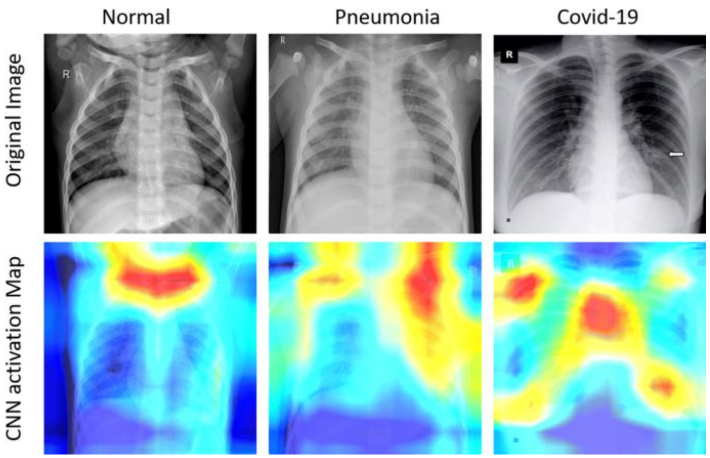
Jingsong Liu, Melis Guelenay,
Priyank Upadhya, Yiheng Xiong

Advisors: Kamilia Zaripova, Matthias Keicher

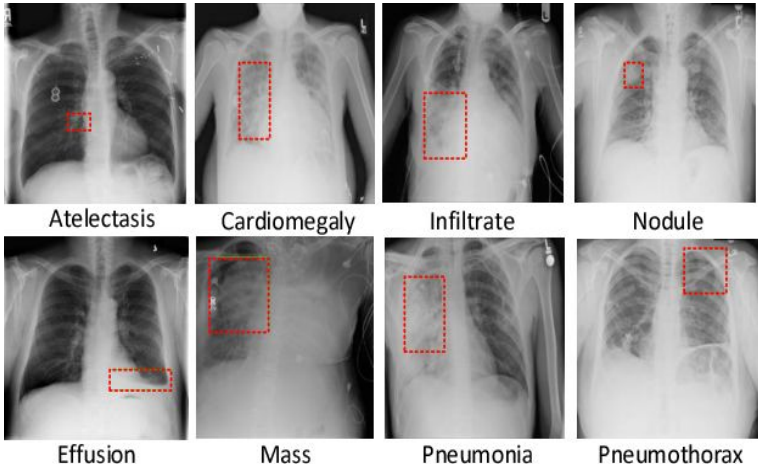
Technische Universität München

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Deep-learning-based Diagnosis



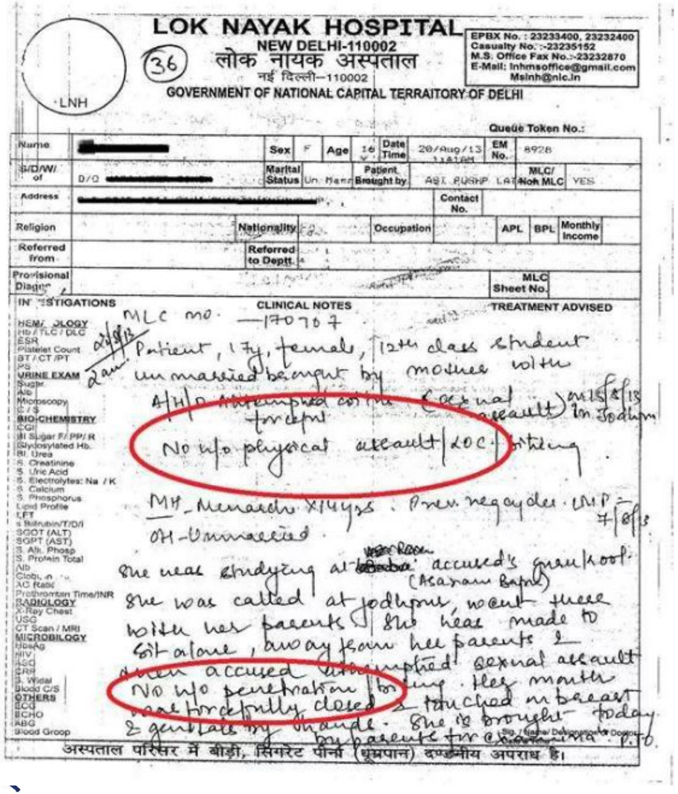
Simple classification



Classification and localization

They are not actually our final goal!

Free-text Medical Report



LOK NAYAK HOSPITAL
NEW DELHI-110002
लोक नायक अस्पताल
नई दिल्ली—110002
GOVERNMENT OF NATIONAL CAPITAL TERRAITORY OF DELHI

EPBX No. : 23233400, 23232400
Casualty No. :-23235152
M.S. Office Fax No.:-23232870
E-Mail: lnhmsoffice@gmail.com
Msinh@nic.in

Queue Token No.:

Name	[redacted]	Sex	F	Age	16 Y	Date Time	20/Aug/13	EM No.	8928
S/D/W/ of	D/O [redacted]	Marital Status	Un Married	Patient Brought by	ASI PUSHP LATA	MLC/ Non MLC	YES		
Address	[redacted]	Contact No.							
Religion		Nationality		Occupation		APL	BPL	Monthly Income	
Referred from		Referred to Deptt.							
Provisional Diagnosis		MLC Sheet No.							

INVESTIGATIONS | CLINICAL NOTES | TREATMENT ADVISED

MLC no. —170707

Patient, 17y, female, 12th class student

unmarried brought by mother with

A/H/O attempted sexual assault on 15/8/13 in Jodhpur

No h/o physical assault / LOC

MH- Menarche X14 yrs. Prev. regular. LMP — 7/8/13

OH - Unmarried

She was studying at ... accused's gurukul (Asaram Bapu)

She was called at Jodhpur, went there with her parents & she was made to sit alone, away from her parents & when accused attempted sexual assault

No h/o penetration ... Her mouth was forcefully closed & touched on breast & genitals by hands. She is brought today by parents for ...

P.T.O.

अस्पताल परिसर में बीड़ी, सिगरेट पीना (धूम्रपान) दण्डनीय अपराध है।

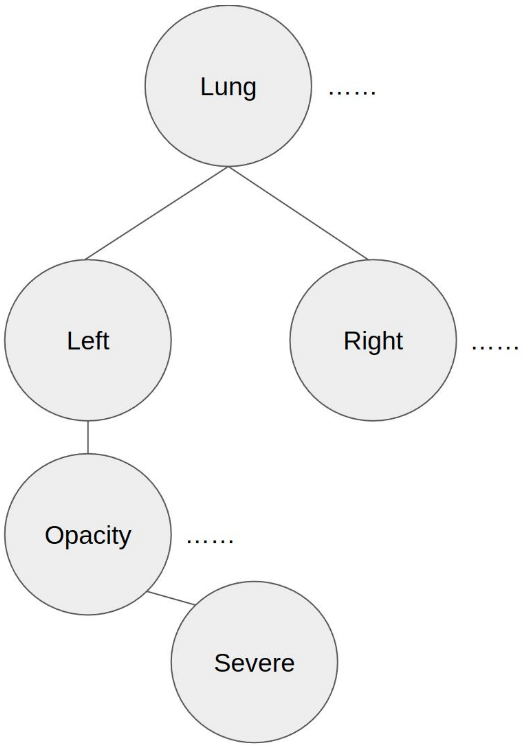
Writing report **BY HAND**?

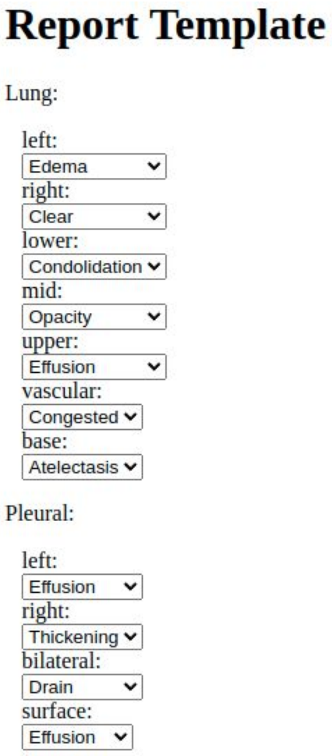
- time-consuming
- error-prone

Automatically generating **Free-text** report?

- largely different between clincians
- hard to evaluate

Structured Medical Report

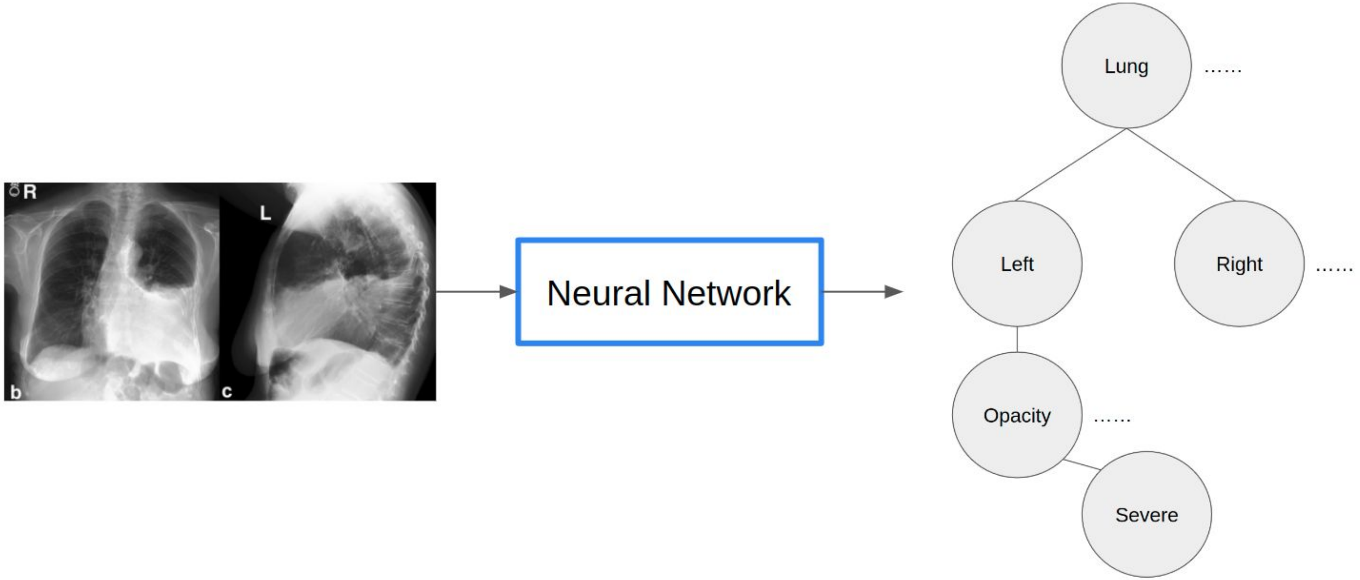




- Well-defined graph-like unified structure
- No ambiguity
- Highly understandable

Generate Structured Report

- Input: Chest X-ray images
- Output: Structured medical report



Related Works

- **<u>Multi-label</u> finding classification[1]**

1. CNN-based detection module
2. Graph CNN to learn label dependency and relationship between the anatomical regions

- **Imaging-captioning-based <u>Free-text</u> report generation[2][3][4]**

1. CNN as feature extractor
2. RNN/LSTM generates sentences

- **Natural-language-prompts-based <u>Structured</u> report generation[5]**

1. Contrastive language-image model
2. Create textual prompts for each structured finding

[1] [Agu et .al MICCAI 21] AnaXNet: Anatomy Aware Multi-label Finding Classification in Chest X-ray
[2] [Jing et .al ACL 18] On the automatic generation of medical imaging reports
[3] [Li et .a al NPIS 18] Hybrid retrieval-generation reinforced agent for medical image report generation
[4] [Wang et .al CVPR 18] Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays
[5] [Keicher et .al ACL 22] Few-shot Structured Radiology Report Generation Using Natural Language Prompts

Dataset

EXAMINATION: CHEST (PA AND LAT)

INDICATION: ___ year old woman with ?pleural effusion // ?pleural effusion

TECHNIQUE: Chest PA and lateral

COMPARISON: ___

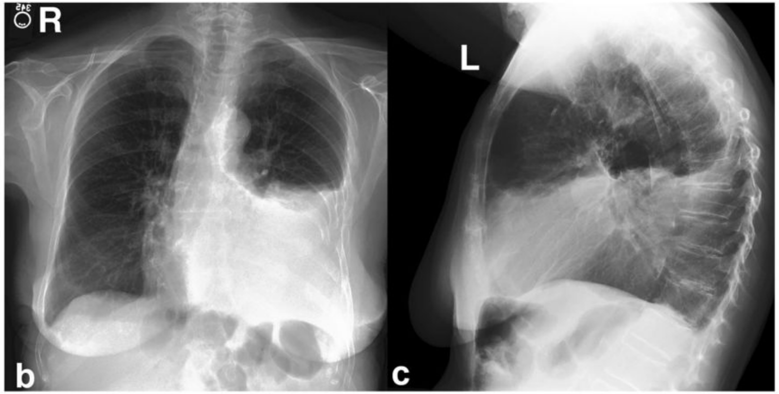
FINDINGS:

Cardiac size cannot be evaluated. Large left pleural effusion is new. Small right effusion is new. The upper lungs are clear. Right lower lobe opacities are better seen in prior CT. There is no pneumothorax. There are mild degenerative changes in the thoracic spine

IMPRESSION:

Large left pleural effusion

a



MIMIC-CXR:

- **large number of imaging study**
- **X-ray images and free-text report**

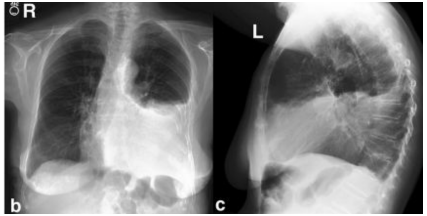
RadGraph:

- derived from MIMIC-CXR via deep learning model
- entities and relations in radiology reports

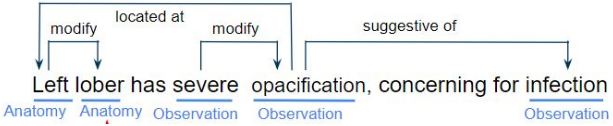
ImaGenome:

- derived from MIMIC-CXR dataset via deep learning model
- scene graph data structure to describe 242k images

RadGraph-based Generated Dataset



Images from MIMIC-CXR



Labeled reports from RadGraph

{"p18/p18004941/s58821758.txt":
[Lober, left,opacification],
[Lober, left, infection],
……[Organ, location, disease]
}

tokenization

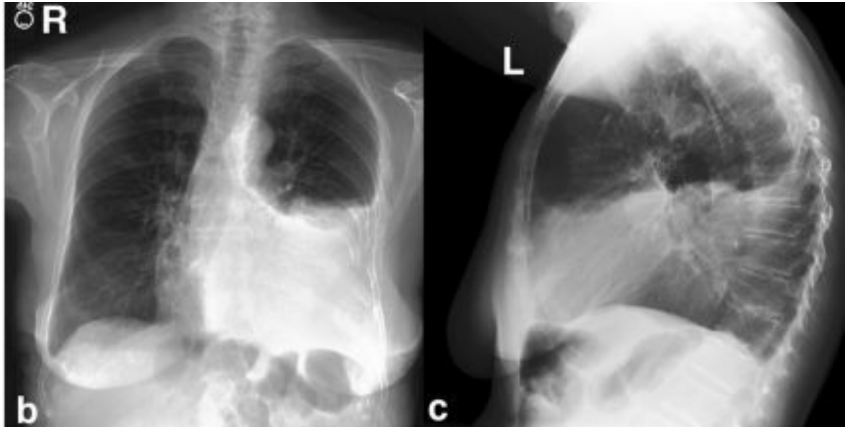
{"p18/p18004941/s58821758.txt":
[1,3,10],
[1,3,6],
……
}

In .json file

- Training set: 425 images & structured reports
- Validation set: 50 images & structured reports
- #diseases: 126, #organs: 47, #locations: 80, #tokens: 597

[Jain et .al NIPS 21l] RadGraph: Extracting Clinical Entities and Relations from Radiology Reports

ImaGeome-based Generated Dataset



Images from MIMIC-CXR


```
"attributes": [
    {
        "right lung": true,
        "bbox_name": "right lung",
        "name": "Right lung",
        "attributes": [
            [
                "nlp|yes|abnormal",
                "technicalassessment|yes|low lung volumes"
            ],
            [
                "anatomicalfinding|yes|lung opacity",
                "anatomicalfinding|yes|pleural effusion",
                "nlp|yes|abnormal"
            ],
            [
                "anatomicalfinding|no|pneumothorax"
            ]
        ],
```


.json file

[Lung Opacity, Lung, Right],
[Pleural Effusion, Lung, Right]
…

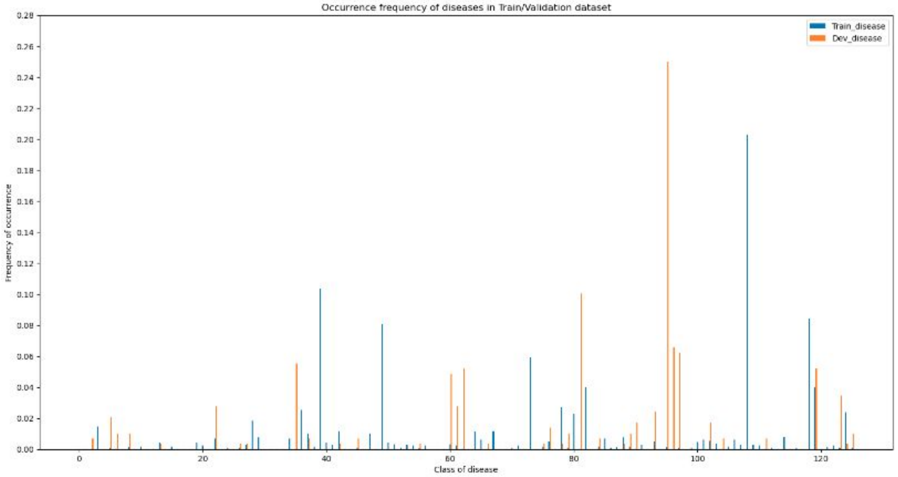
tokenization

PatientId: p18004941
[0, 0, 13],
[3, 0, 13],
…

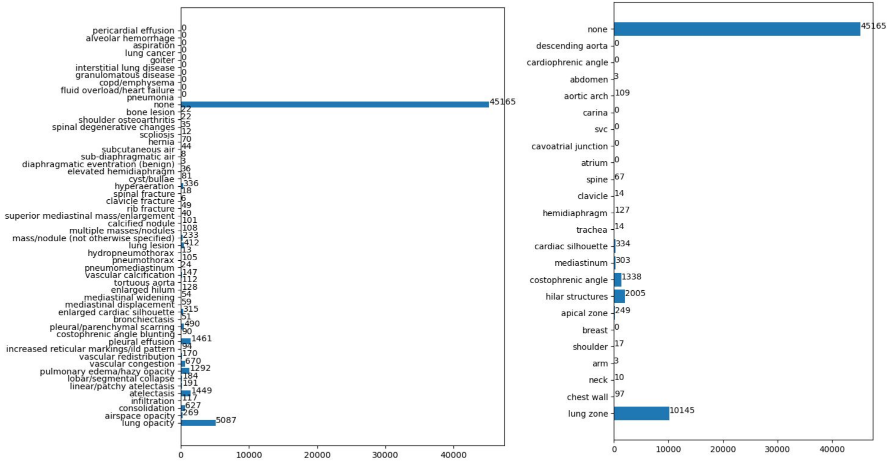
- Training set: 1000 images & structured reports
- Validation set: 100 images & structured reports
- Test set: 200 images & structured reports
- #anatomical findings: 43, #organs: 24, #locations: 14
- #diseases: 10

[Wu et .al l] Chest ImaGenome Dataset for Clinical Reasoning

Highly Inbalanced Dataset



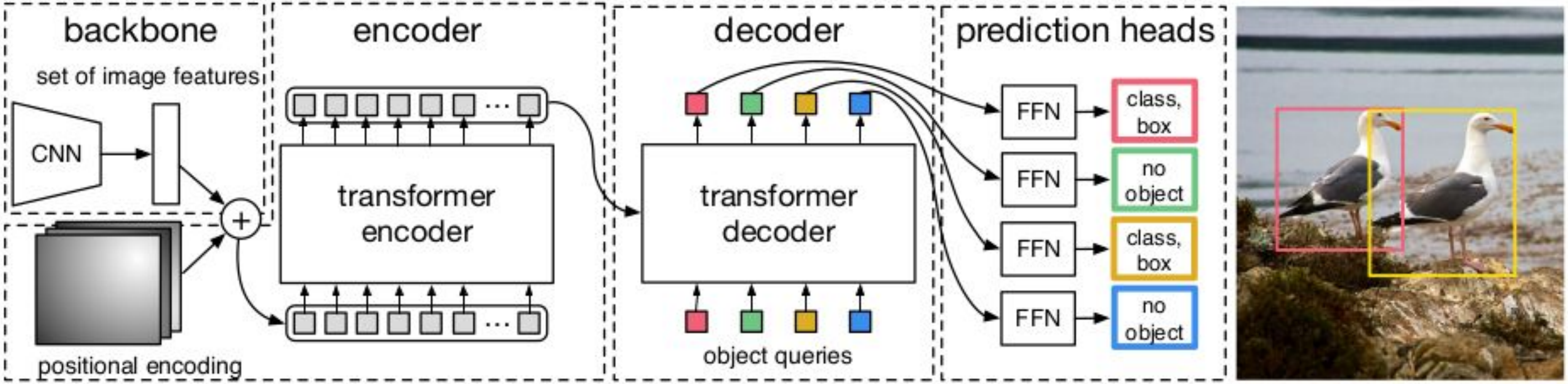
Radgraph



Imagenome

Method - Detection Transformer

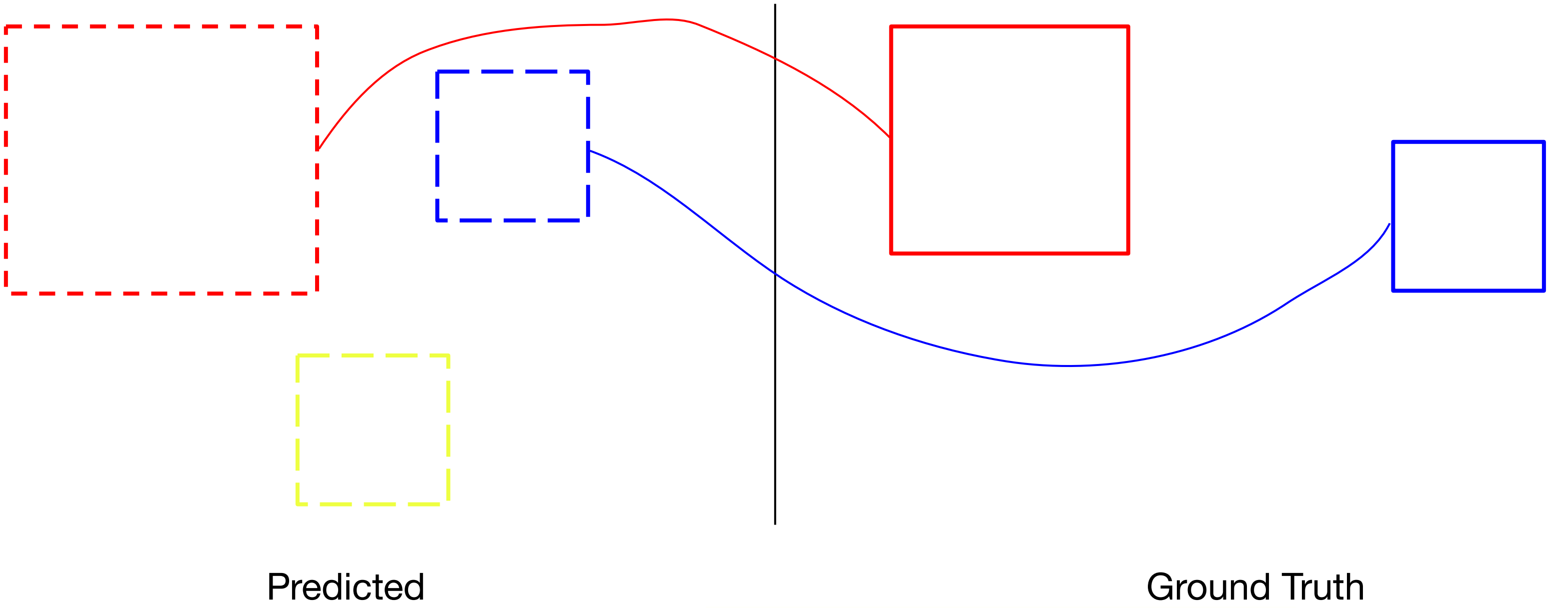
- **Bounding-box-based detection**
  - Object classification
  - Localization: regress *center, height, width*
- View detection as **set prediction**



[Carion et .al ECCV 20] End-to-end Object Detection with Transformers

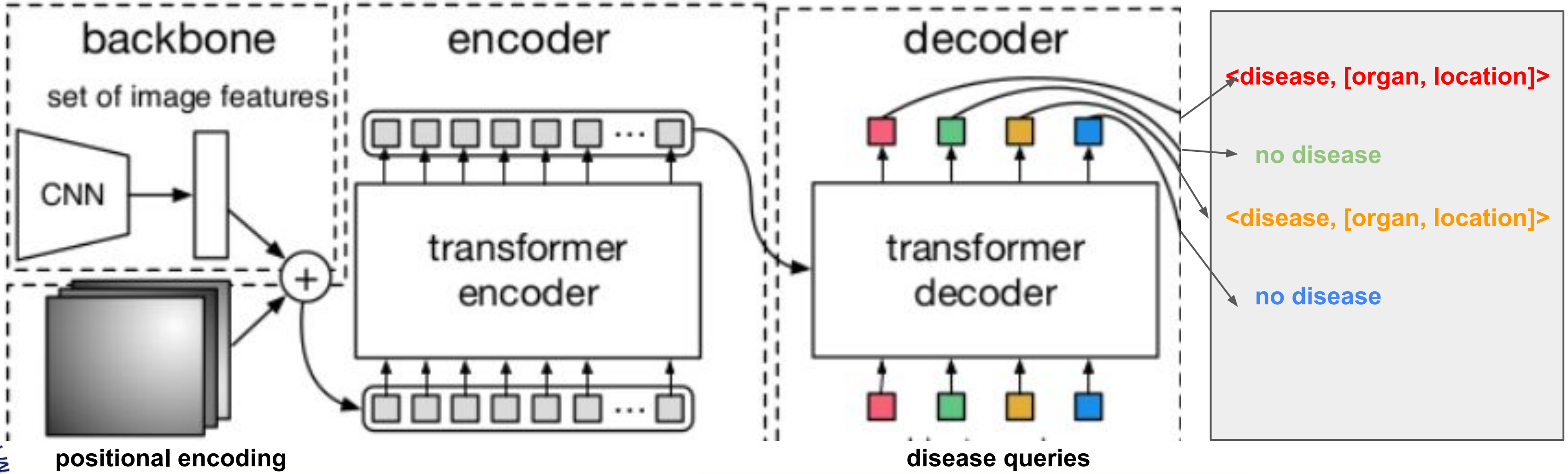
Matching - Bounding-box-based Detection

- Matching cost in detection
  - Object category prediction
  - Similarity of predicted and ground truth boxes



Method - Structure Generation Network (SGN)

- Model structure generation as **natural-language-based detection:**
  - Object classification:
    - Disease classification
  - Localization:
    - Organ classification
    - Location (of organ) classification



Matching - Natural-language-based Detection

- Matching cost in **structured report generation:**
  - **Disease** category prediction
  - Localization prediction
    - **Organ** category prediction
    - **Location** category prediction

- Optimization: minimize matching cost:

$$\hat{\sigma} = \underset{\sigma \in \mathfrak{S}_N}{\arg\min} \sum_{i}^{N} \mathcal{L}_{\text{match}}(y_i, \hat{y}_{\sigma(i)}),$$

- Example:

Predicted probabilities:
[
[0.1, 0.9, 0.1]
[0.2, 0.7, 0.1]
[0.7, 0.1, 0.1]
]
... *N

GT:
[
[0,1,2]
... *N
]

Matching cost = -(0.1+0.7+0.1) + …

Loss Functions

- Cross Entropy Loss

$$\ell(x, y) = L = \{l_1, \dots, l_N\}^\top, \quad l_n = -w_{y_n} \log \frac{\exp(x_{n,y_n})}{\sum_{c=1}^{C} \exp(x_{n,c})}$$

- Focal Loss

$$\mathrm{FL}(p_{\mathrm{t}}) = -\alpha_{\mathrm{t}}(1 - p_{\mathrm{t}})^{\gamma} \log(p_{\mathrm{t}})$$

Experiment Setup

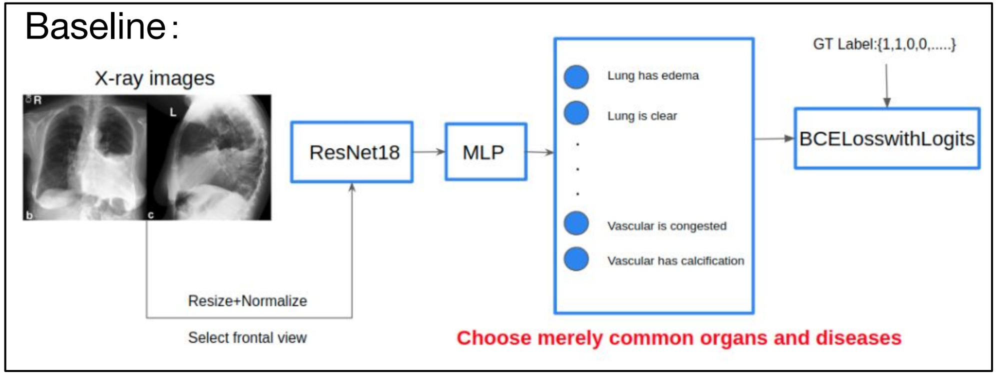
- Training setting:

    - image resized to 224 x 224 and normalized to [0,1]
    - **backbone: ResNet18**
    - data augmentation includes :
        - random crop with at least 75% image size
        - random rotation up to ±15°
        - a color jitter of 10% brightness and 20% contrast and saturation

- Radgraph-based dataset splits: **10-fold** cross validation
- Imagenome-based dataset splits: **random sampling**

Radgraph - Comparison with Baseline

- Multi-label classfication -> precision, recall and f1-score w.r.t. **all possible tuples**



Model	Precision	Recall	F1-score
Baseline	0.12	0.19	0.15
Structure Generation Network	**0.28**	**0.34**	**0.30**

Radgraph - Ablation Study

- Natural-language-based detection problem -> mean average precision w.r.t. 126 **diseases**
  - AP scores **considering organ**:

Model	enlarged	effusion	edema	atelectasis	normal	tortuous	clear	...	mAP
SGN without transformer	**33.33**	0.00	0.00	0.00	18.74	0.00	0.00	...	0.41
Structure Generation Network	31.22	**22.91**	**20.01**	**8.33**	**27.38**	**19.99**	**33.33**	**...**	**2.43**

  - AP scores **considering organ and location**:

Model	enlarged	effusion	edema	atelectasis	normal	tortuous	clear	...	mAP
SGN without transformer	**11.11**	0.00	0.00	0.00	11.24	0.00	0.00	...	0.12
Structure Generation Network	3.12	**6.24**	**19.99**	**8.33**	**24.40**	**6.40**	**33.33**	**...**	**1.55**

Radgraph - Additional Results

- Some triplet ([diesease, organ, location]) token AP scores w.r.t. 90 classes:

Model	[clear, lung, lung]	[normal, mediastinu, contour]	[edema, lung, lung]	[effusion, pleural, left]	[opacity, lung, lower]	[atelectasis, basilar, basilar]	[normal, heart, contour]	…	mAP
Structure Generation Network	34.00	52.51	13.55	8.94	8.13	7.54	11.14	…	17.23

Imagenome - Comparison with SOTA

Method	Lung Opac.	Pleural Ef	Atelectasis	Enl. Card. S	Pulm. Edema	Pneumothor.	Consolidation	**Avg. AUC**
Global view with no localization								
DenseNet169	0.91	0.94	0.86	0.92	0.92	0.93	0.86	0.89
DenseNet169	0.87	0.9	0.79	0.86	0.85	0.83	0.75	0.82
DenseNet121	0.88	0.91	0.81	0.87	0.87	0.87	0.79	0.84
ViT-B16	0.88	0.91	0.8	0.87	0.86	0.85	0.77	0.83
Object detection backbone with high resolution crops								
FasterR-CNN	0.84	0.89	0.77	0.85	0.87	0.77	0.75	0.8
AnaXNet	0.88	0.96	0.92	0.99	0.95	0.8	0.89	0.93
Detector free localization on global view (224 × 224)								
DenseNet121 full	0.83	0.89	0.79	0.87	0.84	0.89	0.83	0.84
FSRG full	0.82	0.89	0.78	0.87	0.84	0.9	0.83	0.84
Structure Generation Network	**0.77**	**0.71**	**0.82**	**1**	**0.67**	**0.5**	**0.93**	**0.64**

Imagenome - Additional Results

- AP scores considering diseases with organs:

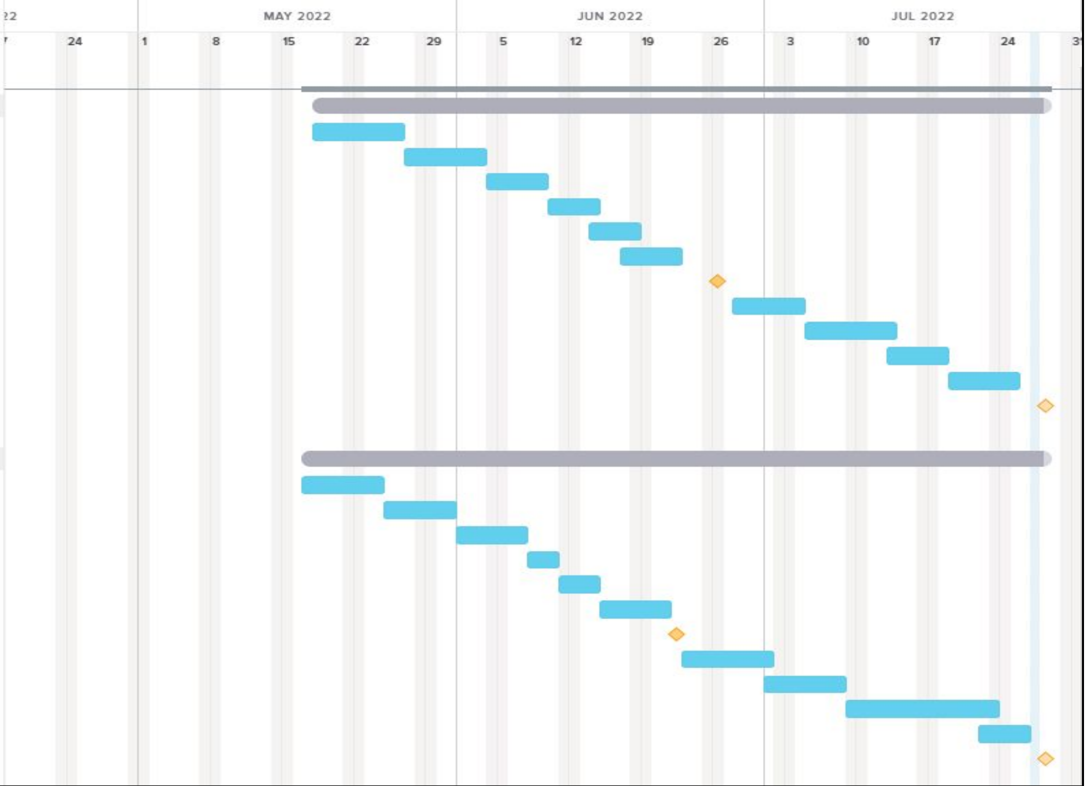
Model	lung opacity	pleural effusion	atelectasis	enlarged cardiac silhouette	pulmonary edema	pneumothorax	consolidation
Structure Generation Network	0.70	0.65	0.74	1.0	0.78	0.03	0.66

Challanges Faced

- Competely new task
- Difficulty to extract labels from given label graphs
- Highly imbalanced dataset
- No existing metrics available to compare results
- Lack of medical knowledge

Milestones and Timelines

	Assigned	Progress
Work Plan		99%
▾ Priyank & Melis		99%
Get familiar with the dataset ImaGenome		100%
Process the dataset ImagGenome		100%
Literature review		100%
Creating a baseline		100%
Define the model and input dataset		100%
Implement the baseline (data, net, eva)		100%
Mid-presentation		☑
Literature about Transformer review & ideas		100%
Implement the Transformer model and test		100%
Check weighted cross entropy loss and focal loss		100%
Add mlp for predicting main diseases		100%
Final-presentation		☐
▾ Jingsong & Yiheng		99%
Get familiar with the dataset RadGraph and MIMIC		100%
Process the dataset RadGraph		100%
Literature review		100%
Define classifications and lables		100%
Generate the new dataset		100%
Implement the baseline (data, net, eva)		100%
Mid-presentation		☑
Improve the complete baseline		100%
Literature about Transformer review & ideas		100%
Implement the new Transformer model		100%
Finish some abslation and comparison experiments		100%
Final-presentation		☐



Summary

- Contributions:
  - Generate two datasets for **structured report generation**
  - Model structured report generation as **multi-classification problem** and **natural-language-based detection problem** and conduct complete experiments
  - Propose **structure generation network** based on detection transformer
  - Demonstrate the feasibility of using **transformer** to generate structured medical reports

- Drawback:
  - Ignore the relations between diseases, organs and locations
  - Mapping in Radgraph is not accurate
  - Overall results are not so impressive
  - Data inconsistency

- Future plans:
  - Try making use of relations between diseases, organs and locations
  - Try using autoregressive model to generate structured reports

Feedback

- Nice advisors and supportive guidance
- Lecture contents are really intuitive
- Project itself is interesting but a bit hard

Take-home Messages

- Dataset can be the most important thing in deep learning (even more important than model itself)
- Background knowledge (here is medical knowledge) is essential (even though we are using some “black box” neural networks)
- Simple architecture generally works better than very complicated architecture

DEMO

Question?