ihealth | Millennium Institute for Intelligent Healthcare Engineering

Medical Imaging Reports: Deep Learning and Explainability

Denis Parra
CS Departament, PUC Chile (iHealth, CENIA, IMFD)
Medical Imaging AI for healthcare: from research to clinical translation
2023

ihealth | Millennium Institute for Intelligent Healthcare Engineering

Acknowledgments

These slides thanks to:

- Pablo Pino, MSc PUC Chile
- Pablo Messina, PhD(c) PUC Chile
- “Speech and Language Processing” (3rd ed.) by Jurafsky and Martin https://web.stanford.edu/~jurafsky/slp3/

ihealth | Millennium Institute for Intelligent Healthcare Engineering

This research in collaboration with:

- Cecilia Besa, PUC School of Medicine
- Jocelyn Dunstan, PUC School of Engineering
- PUC CS Students: Pablo Pino, Pablo Messina
- CENIA engineer: José Cañete
- Upcoming PUC Students: Tamara Quiroga, Francisco Madariaga, Ricardo Schilling
- Undergraduate students: Juan Pablo Barías, Mario Espínola

ihealth | Millennium Institute for Intelligent Healthcare Engineering

TOC

- Motivation
- The MIRG task
- Deep Learning
- Encoder-Decoder Architecture
- Datasets & Training
- Metrics and Evaluation
- SOTA and challenges

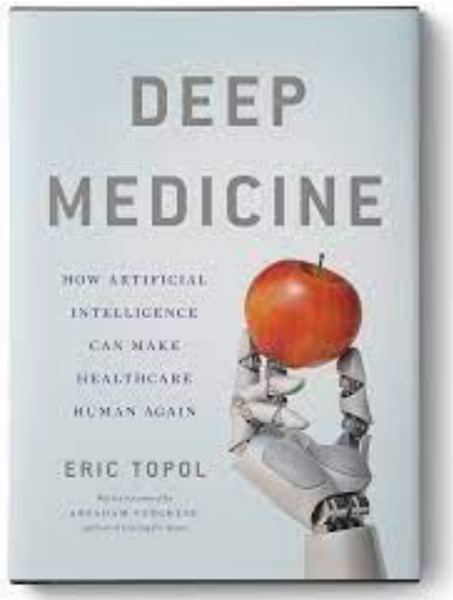
ihealth | Millennium Institute for Intelligent Healthcare Engineering

Motivation

- Eric Topol's "Deep Medicine" (2019) book indicates that in the US, by 2016, there were 800 million medical scans a year, accounting for about 60 billion images. Scaling this up based only on human labor is challenging

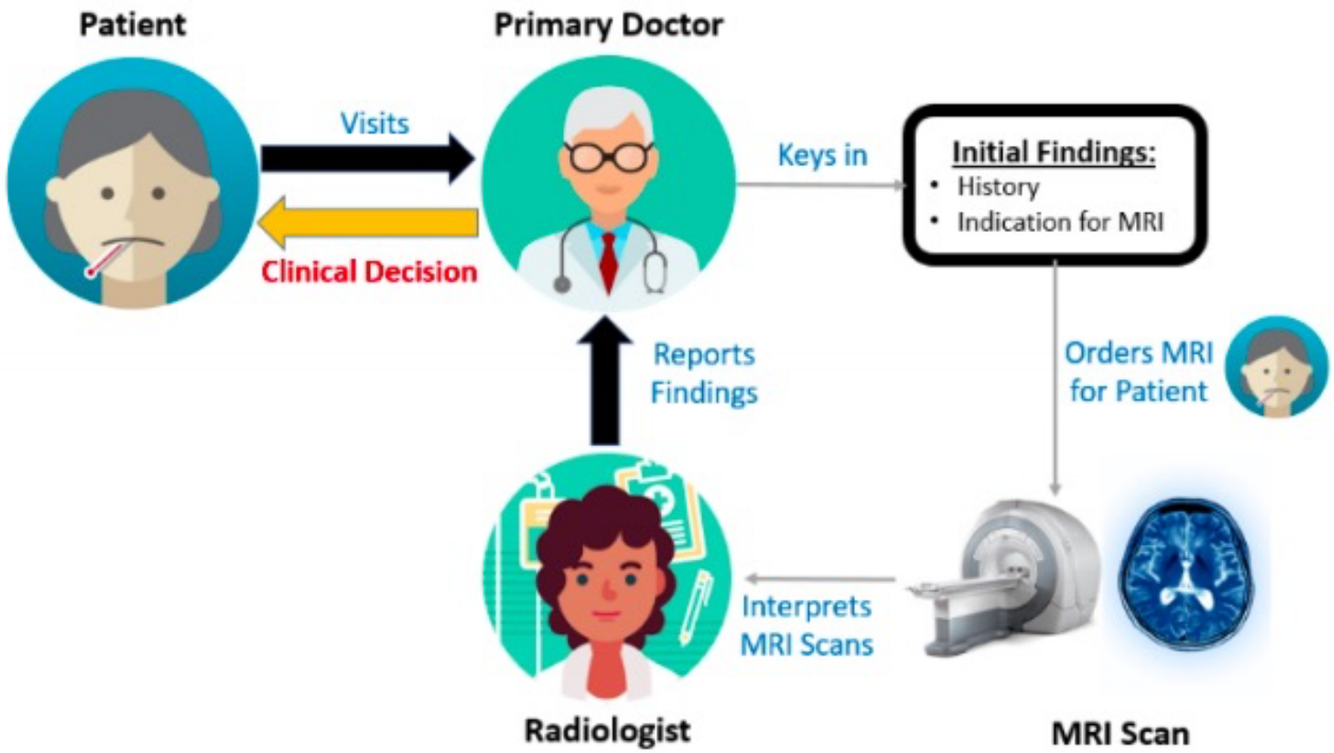
- Using Artificial Intelligence (AI) for medical image report generation (MIRG) could help hospitals deal with this large and growing demand

- MIRG does not mean replacing radiologists, but rather helping them being more efficient and effective



ihealth | Millennium Institute for Intelligent Healthcare Engineering

The patient-physician interaction

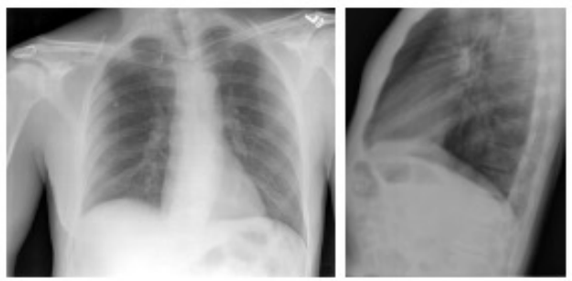


Alwin Yaoxian Zhang, et al.. 2019. Explainable AI: Classification of MRI Brain Scans Orders for Quality Improvement. In Proceedings of (BDCAT '19). DOI:https://doi.org/10.1145/3365109.3368791

ihealth | Millennium Institute for Intelligent Healthcare Engineering

The MIRG task

- Given one or more patient's input image(s), generate a text report of the ***findings*** section of a radiology report



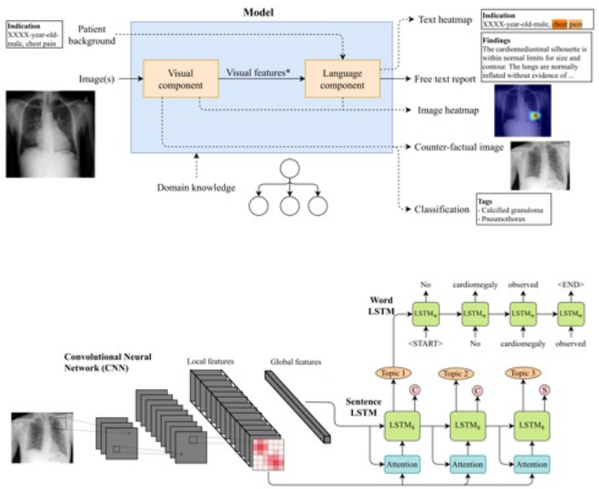
Manual tags: Calcified Granuloma/lung/upper lobe/right
Automatic tags: Calcified granuloma

Comparison: Chest radiographs XXXX.
Indication: XXXX-year-old male, chest pain.
Findings: The cardiomediastinal silhouette is within normal limits for size and contour. The lungs are normally inflated without evidence of focal airspace disease, pleural effusion, or pneumothorax. Stable calcified granuloma within the right upper lung. No acute bone abnormality.
Impression: No acute cardiopulmonary process.

Example from the IU X-ray dataset, frontal and lateral chest x-rays from a patient, alongside the report and the annotated tags. XXXX is used for anonimization.

ihealth | Millennium Institute for Intelligent Healthcare Engineering

iHealth on Report Generation



Messina, Pino et al (2021)
A Survey on Deep Learning and Explainability for Automatic Report Generation from Medical Images.
ACM CSUR

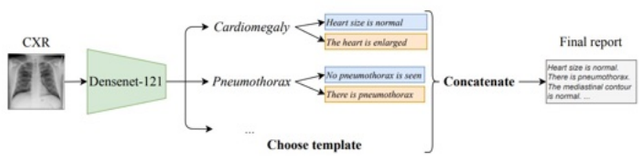
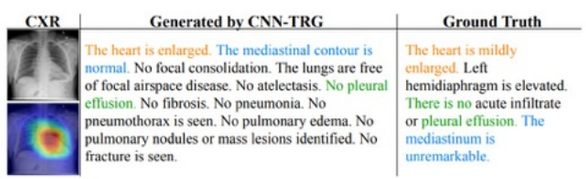
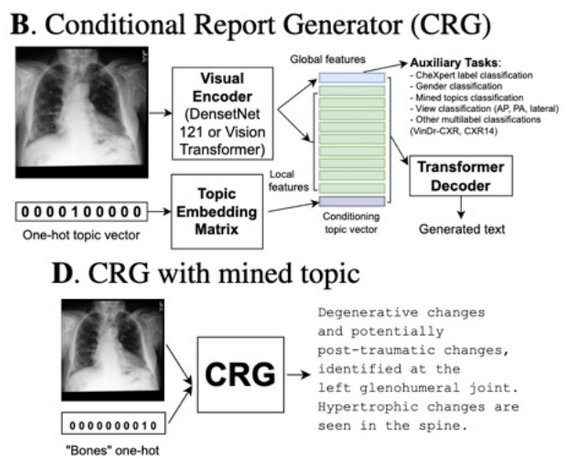


Fig. 2. CNN-TRG model for report-generation



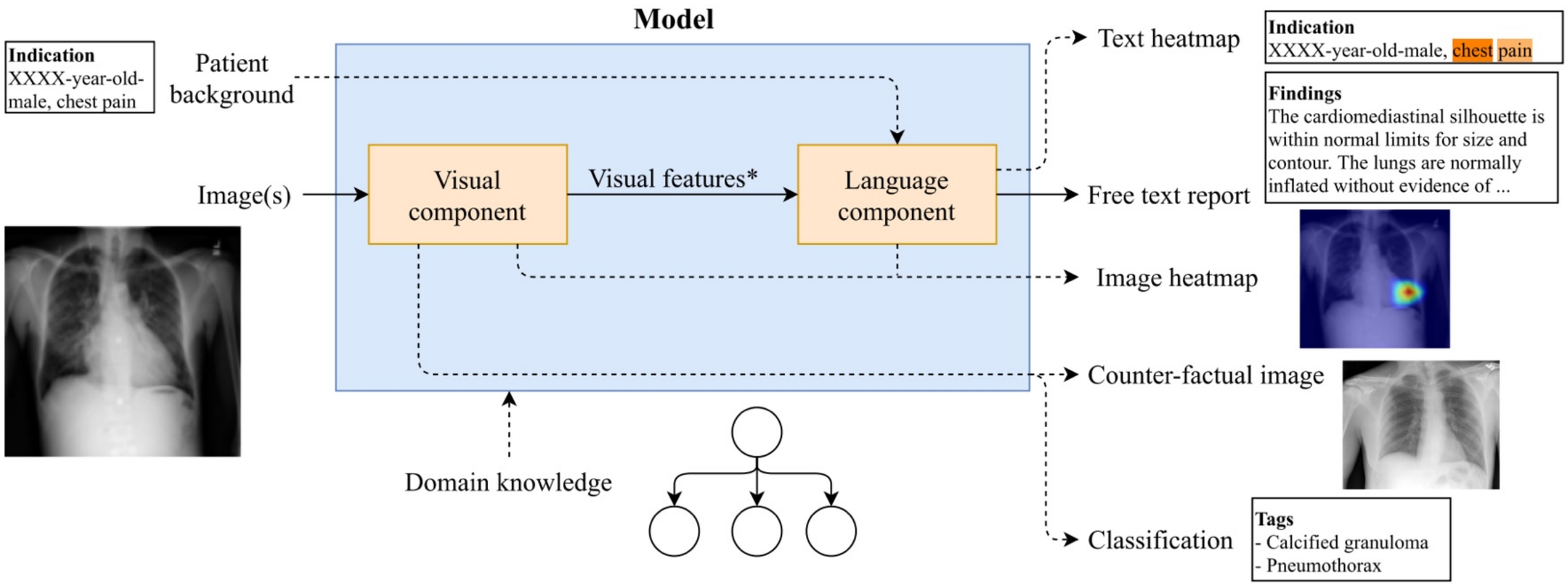
Pino et al (2022) Clinically Correct Report Generation from Chest X-Rays Using Templates,
MLMI 2021



Messina, Cañete et al (2022) Two-stage Conditional Chest X-ray Radiology Report Generation.
MedNeurIPS 2022.

ihealth | Millennium Institute for Intelligent Healthcare Engineering

The MIRG task – A higher level view



Messina, P, Pino, P et al. (2020) “A Survey on Deep Learning and Explainability for Automatic Report Generation from Medical Images”

ihealth | Millennium Institute for Intelligent Healthcare Engineering

Deep Learning

ihealth | Millennium Institute for Intelligent Healthcare Engineering

Deep Learning (2012 - …)

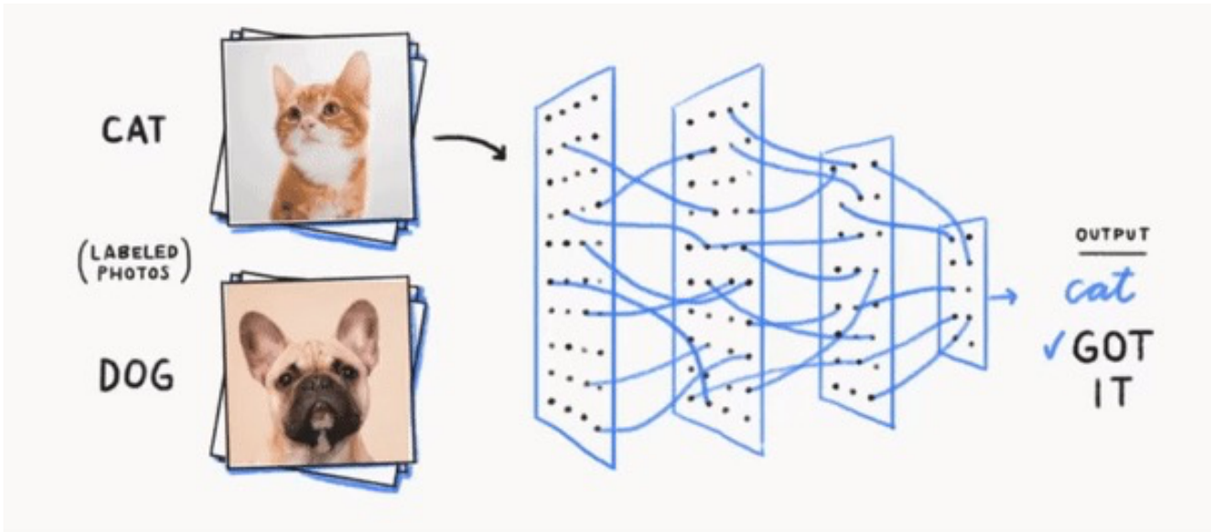


Image from https://mc.ai/aisaturdaylagos-the-torch-panther/

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). **Imagenet classification with deep convolutional neural networks**. NeurIPS 2012.

ihealth | Millennium Institute for Intelligent Healthcare Engineering

Deep Learning (2012 - …)

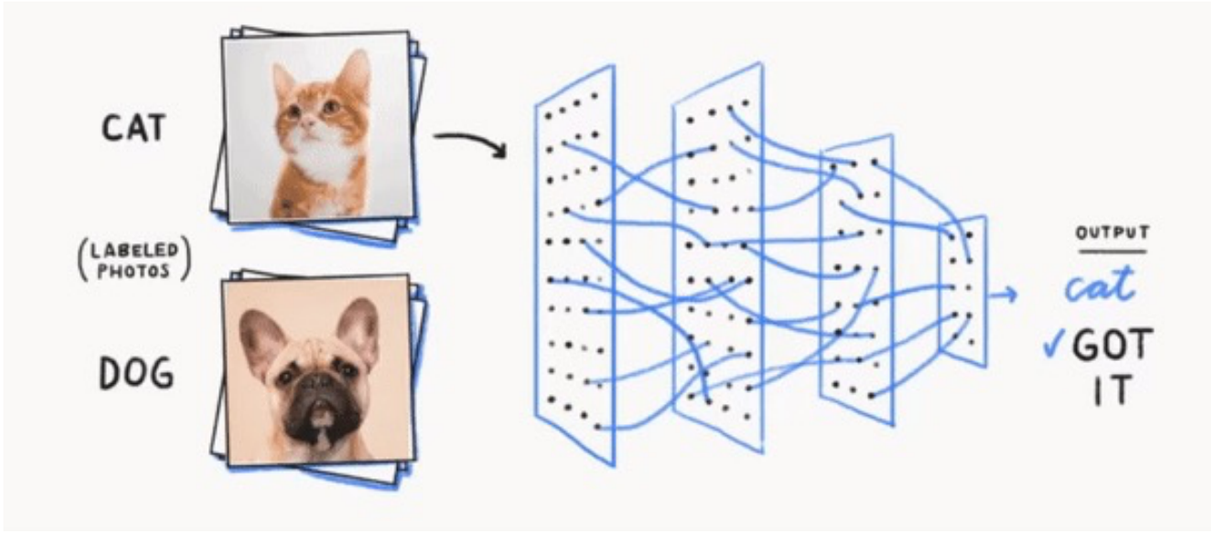
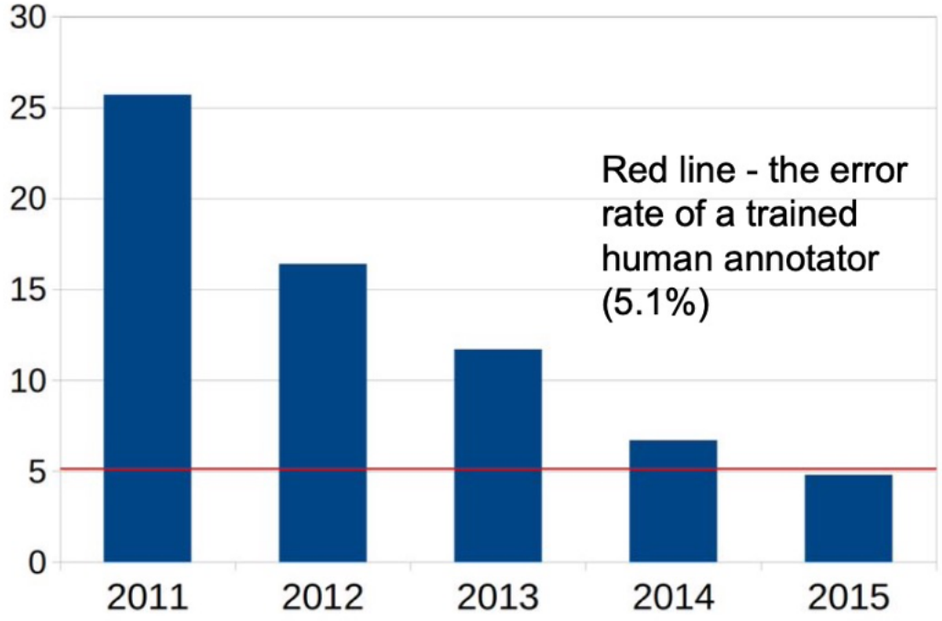


Image from https://mc.ai/aisaturdaylagos-the-torch-panther/

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). **Imagenet classification with deep convolutional neural networks**. NeurIPS 2012.

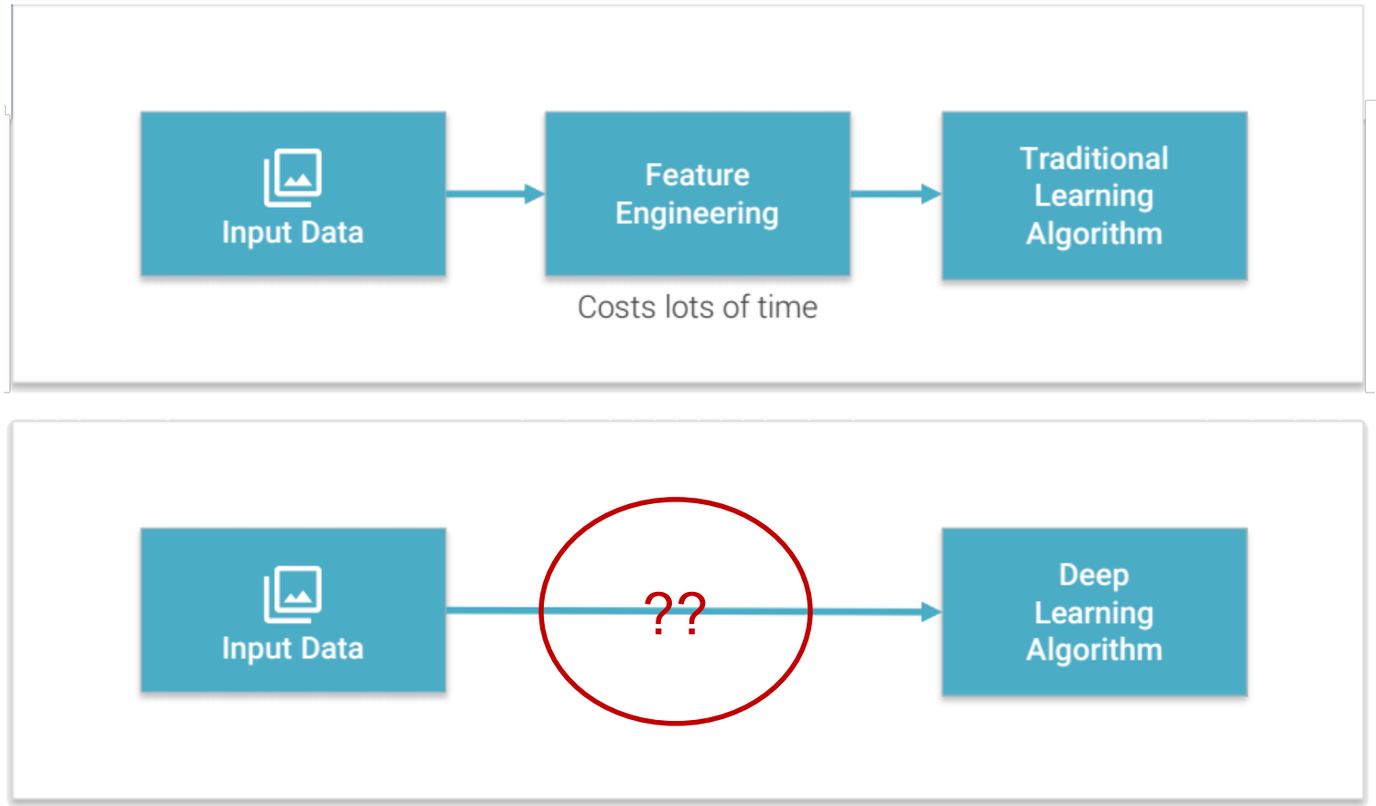


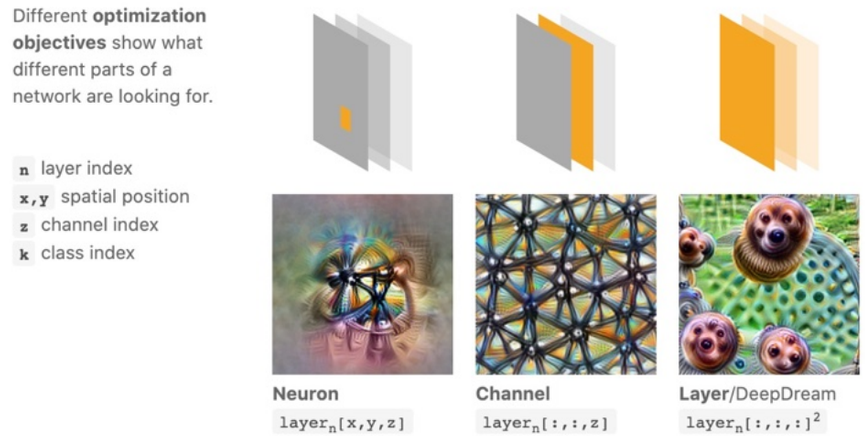
Progress in machine classification of images: the error rate (%) of the ImageNet competition winner by year.

Sandegud, CC0, via Wikimedia Commons

ihealth | Millennium Institute for Intelligent Healthcare Engineering

Deep Learning (2012 - …) – Representation learning





https://distill.pub/2017/feature-visualization/

Olah, et al., "Feature Visualization", Distill, 2017.

ihealth | Millennium Institute for Intelligent Healthcare Engineering

Deep Learning (2012 - …) – Generative models

Transformers: GPT-3

Essay outline
Generation

Generate an outline for a research topic.

Prompt

Create an outline for an essay about Walt Disney and his contributions to animation:

I: Introduction

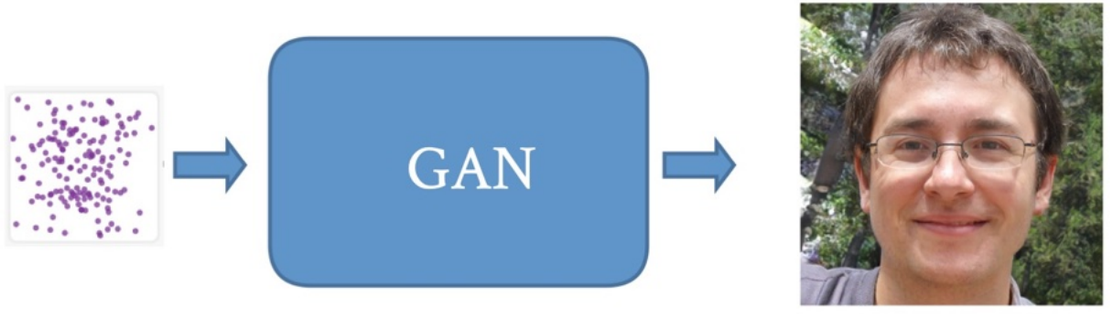
Sample response

: Introduce the topic and your point of view on it.

II: Walt Disney's life: His early life, the years leading up to the release of Snow White and the Seven Dwarfs (1937), and his contributions to animation.

III: Walt Disney's death: The circumstances surrounding his death, his contributions to animation, and the lasting effects of his work.

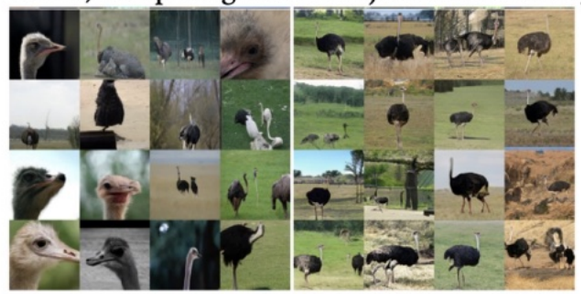
IV: Conclusion: Restate your point of view and address how Walt Disney continues to influence animation today.



https://thispersondoesnotexist.com/

VQVAE-2

- VAEs también se usan para generar, pero con menor calidad que GANs, aunque algunos trabajos muestran mayor diversidad



VQ-VAE (Proposed) BigGAN deep

Razavi, A., Oord, A. V. D., & Vinyals, O. (2019). Generating diverse high-fidelity images with vq-vae-2. *arXiv preprint arXiv:1906.00446*.

Figure 5: Sample diversity comparison for the proposed method and BigGan Deep for Tinca-Tinca (1st ImageNet class) and Ostrich (10th ImageNet class). BigGAN samples were taken with the truncation level 1.0, to yield its maximum diversity. There are several kinds of samples such as top view of the fish or different kinds of poses such as a close up ostrich absent from BigGAN's samples.

TEXT DESCRIPTION

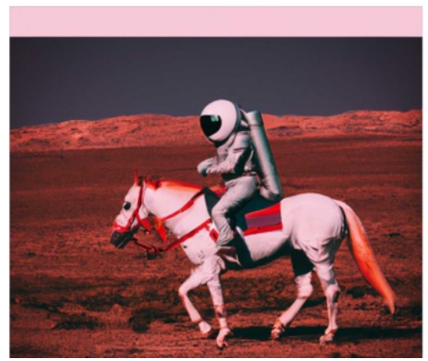
An astronaut Teddy bears A bowl of soup

riding a horse lounging in a tropical resort in space playing basketball with cats in space

in a photorealistic style in the style of Andy Warhol as a pencil drawing

→

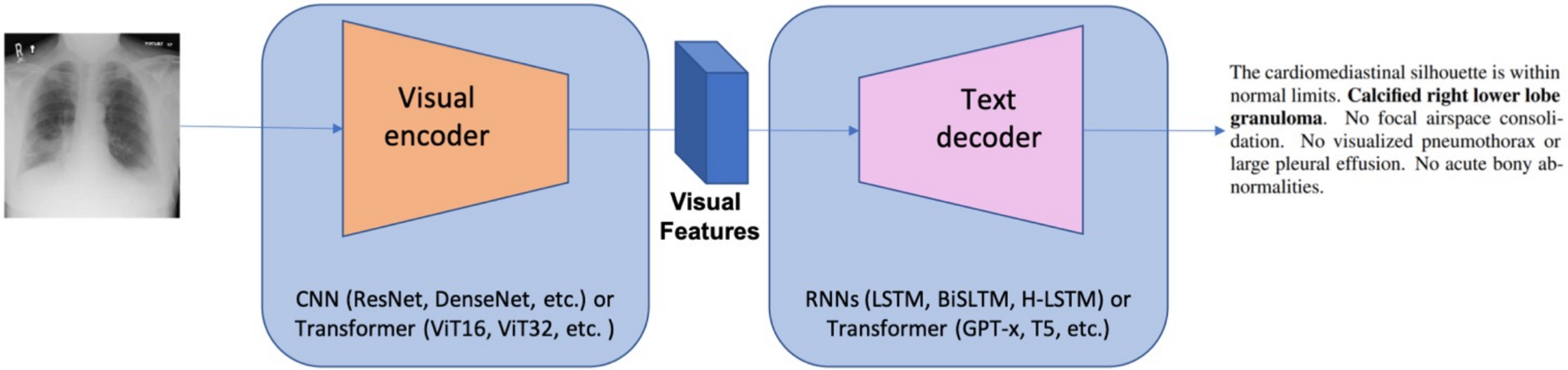
DALL·E 2



ihealth | Millennium Institute for Intelligent Healthcare Engineering

Traditional MIRG solution: Deep Learning

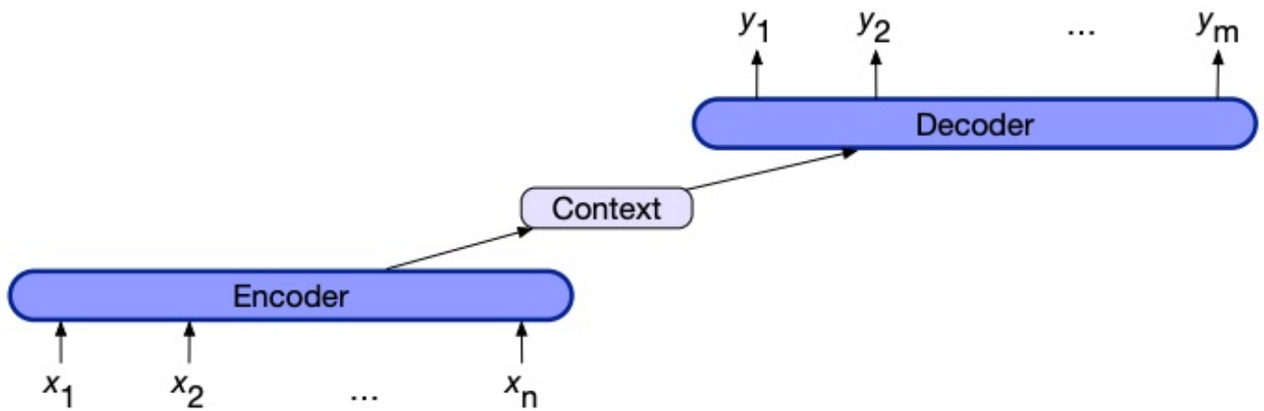
Encoder + Decoder Architecture



Parra, Besa (2022) “Multimodal, multitask and transfer learning for deep radiological report generation”

NLG: The encoder-decoder model for NLP/NLG

Encoder-decoder networks, or **sequence-to-sequence** networks, are models capable of generating contextually appropriate, arbitrary length, output sequences

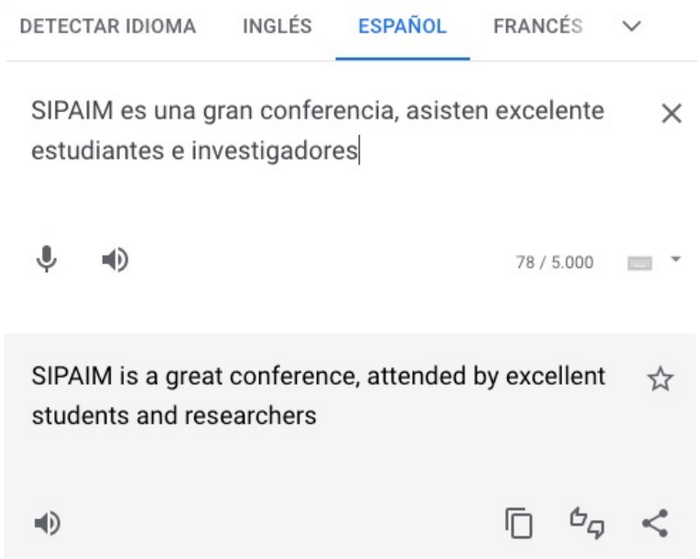


The encoder-decoder architecture. The **context** is a function of the hidden representations of the input, and may be used by the decoder in a variety of ways.

ihealth | Millennium Institute for Intelligent Healthcare Engineering

Applications of E-D model

Machine Translation



Summarization

Figure 4: An excerpt of heat-mapping on Neat-Vision tool with transformed attention values to highlight the importance of sentence with red color.

Question Answering

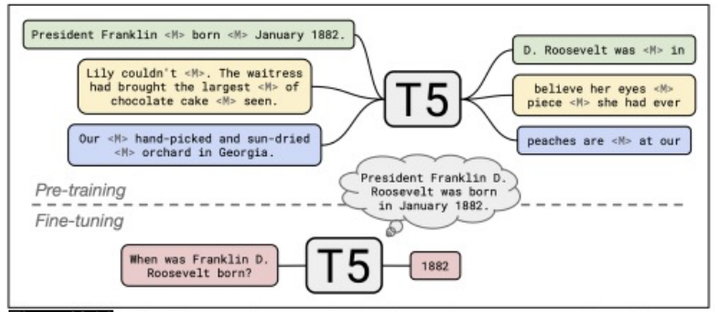
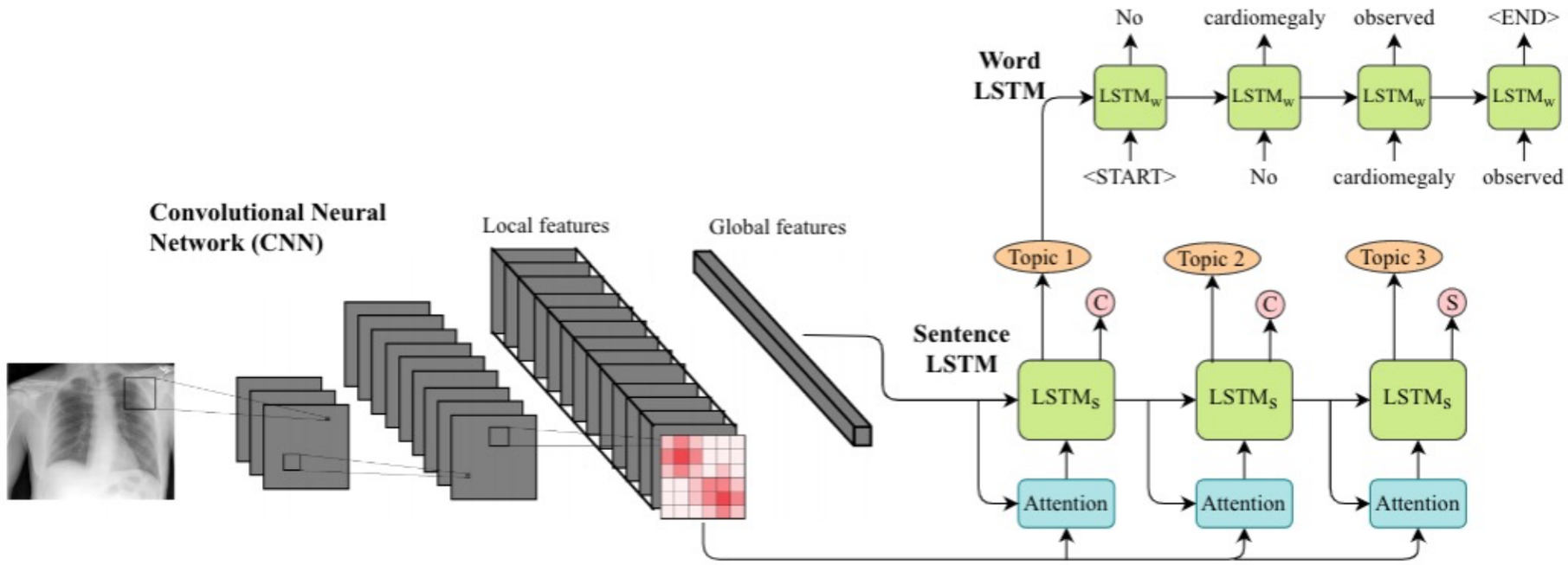


Figure 23.16 The T5 system is an encoder-decoder architecture. In pretraining, it learns to fill in masked spans of task (marked by <M>) by generating the missing spans (separated by <M>) in the decoder. It is then fine-tuned on QA datasets, given the question, without adding any additional context or passages. Figure from Roberts et al. (2020).

ihealth | Millennium Institute for Intelligent Healthcare Engineering

MIRG Architecture example

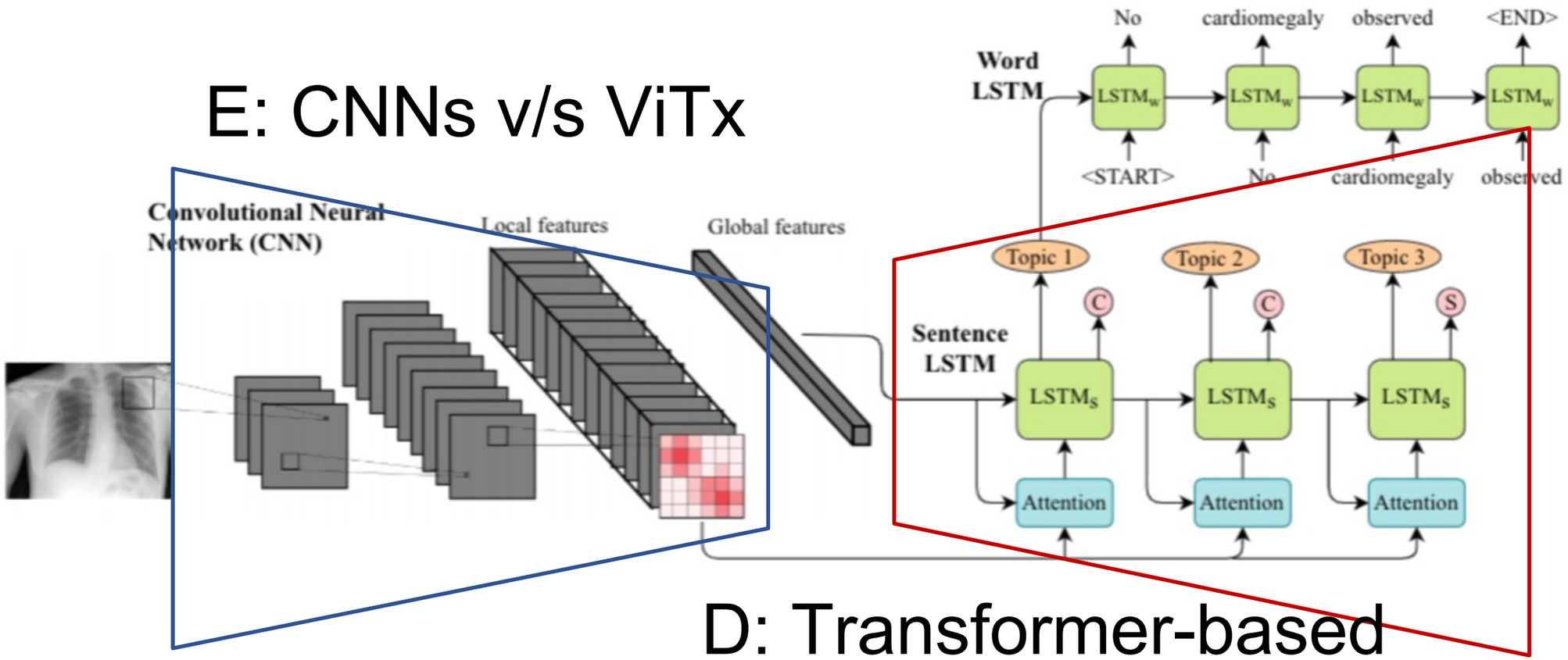
A CNN visual encoder + a language model decoder (e.g. LSTM)



Messina, P, Pino, P et al. (2020) “A Survey on Deep Learning and Explainability for Automatic Report Generation from Medical Images”

ihealth | Millennium Institute for Intelligent Healthcare Engineering

We need to combine architectures and representations!



Messina, P, Pino, P et al. (2020) "A Survey on Deep Learning and Explainability for Automatic Report Generation from Medical Images"

ihealth | Millennium Institute for Intelligent Healthcare Engineering

Datasets & training

ihealth | Millennium Institute for Intelligent Healthcare Engineering

Datasets

- 18 report datasets and 9 classification datasets

Dataset	Year	Image Type	# images	# reports	# patients	Used by papers
Report datasets						
IU X-ray [27]	2015	Chest X-Ray	7,470	3,955	3,955	[15, 35, 39, 47, 60, 65, 66, 83, 84, 87, 90, 114, 117, 126, 136, 142, 143, 145–148, 154]
MIMIC-CXR [67, 68]	2019	Chest X-Ray	377,110	227,827	227,827	[90]
PadChest(sp) [18]	2019	Chest X-Ray	160,868	109,931	67,625	None(5)
ImageCLEF Caption 2017 [34]	2017	Biomedical(2)	184,614	184,614	-	[50]
ImageCLEF Caption 2018 [38]	2018	Biomedical(2)	232,305	232,305	-	None(5)
ROCO [102]	2018	Multiple radiology(3)	81,825	81,825	-	None(5)
PEIR Gross [66]	2017	Gross lesions	7,442	7,442	-	[66]
INBreast(pt) [96]	2012	Mammography X-ray	410	115	115	[85, 122]
STARE [57]	1975	Retinal fundus	400	400	-	None(5)
RDIF(1) [93]	2019	Kidney Biopsy	1,152	144	144	[93]
Classification datasets						
CheXpert [62]	2019	Chest X-Ray	224,316	0	65,240	[148, 154]
ChestX-ray14 [135]	2017	Chest X-Ray	112,120	0	30,805	[15, 65, 84, 87, 136, 143, 145]
LiTS [24]	2017	Liver CT scans	200	0	-	[125]
ACM Biomedia 2019 [54]	2019	Gastrointestinal tract (4)	14,033	0	-	[48]
DIARETDB0 [71]	2006	Retinal fundus	130	0	-	[140]
DIARETDB1 [70]	2007	Retinal fundus	89	0	-	[140]
Messidor [1, 26]	2013	Retinal fundus	1,748	0	874	[140]
DDSM [53]	2001	Mammography X-ray	10,480	0	-	[74]

ihealth | Millennium Institute for Intelligent Healthcare Engineering

Main IMRG Datasets

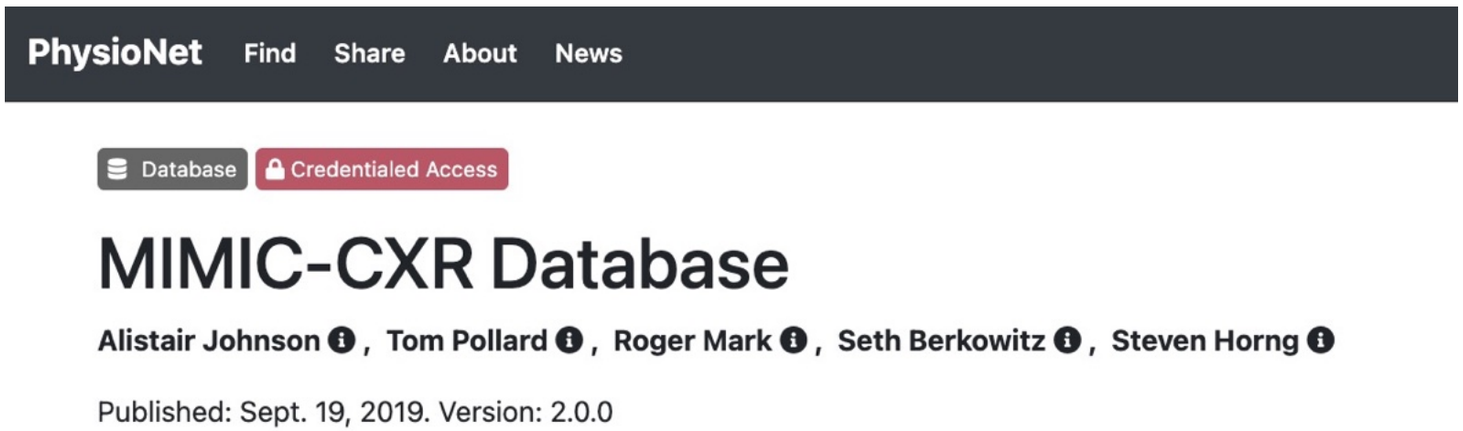
- Open-UI Dataset (2015), MIMIC-CXR (2019).
- Padchest (2019) Is the only one with Spanish reports, but after heavy text processing.

Dataset	Year	Image Type	# images	# reports	# patients	Used by papers
Report datasets						
IU X-ray [27]	2015	Chest X-Ray	7,470	3,955	3,955	[15, 35, 39, 47, 60, 65, 66, 83, 84, 87, 90, 114, 117, 126, 136, 142, 143, 145–148, 154]
MIMIC-CXR [67, 68]	2019	Chest X-Ray	377,110	227,827	227,827	[90]
PadChest$^{(sp)}$[18]	2019	Chest X-Ray	160,868	109,931	67,625	None$^{(5)}$

ihealth | Millennium Institute for Intelligent Healthcare Engineering

MIMIC-CXR How to Use

- Download from physionet
- Complete CITI training and comply with restrictions



Access

Access Policy:
Only credentialed users who sign the DUA can access the files.

License (for files):
PhysioNet Credentialed Health Data License 1.5.0

Data Use Agreement:
PhysioNet Credentialed Health Data Use Agreement 1.5.0

Required training:
CITI Data or Specimens Only Research

ihealth | Millennium Institute for Intelligent Healthcare Engineering

MIMIC-CXR extensions: Chest imagenome

PhysioNet Find Share About News Account Search

Database Credentialed Access

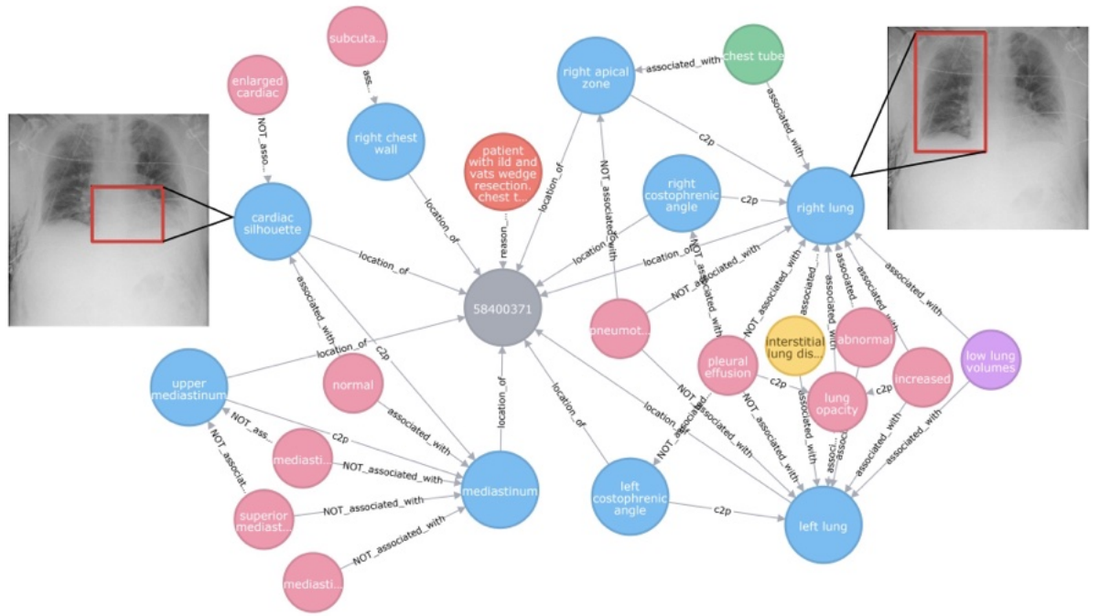
Chest ImaGenome Dataset

Joy Wu, Nkechinyere Agu, Ismini Lourentzou, Arjun Sharma, Joseph Paguio, Jasper Seth Yao, Edward Christopher Dee, William Mitchell, Satyananda Kashyap, Andrea Giovannini, Leo Anthony Celi, Tanveer Syeda-Mahmood, Mehdi Moradi

Published: July 13, 2021. Version: 1.0.0

Table 2 - CXR report knowledge graph evaluation results from 500 reports

Object-Attribute Relations	Sentence-level	Report-level
Number of annotations	21593	16569
Precision	0.932	0.938
Recall	0.945	0.939
F1-score	0.939	0.939



Wu, J. T., Agu, N. N., Lourentzou, I., Sharma, A., Paguio, J. A., Yao, J. S., ... & Moradi, M. (2021). Chest imagenome dataset for clinical reasoning. *arXiv preprint arXiv:2108.00316*.

ihealth | Millennium Institute for Intelligent Healthcare Engineering

Evaluation

ihealth | Millennium Institute for Intelligent Healthcare Engineering

Context

- Quick progress of Deep Learning in Computer Vision and Natural Language Processing can potentially solve the task in a few years

- However, recent research in MIRG shows that:
  - Traditional NLP/NLG metrics (BLEU, ROUGE, CIDEr, etc.) might not measure what is needed for actual clinical use
  - Recent state-of-the-art methods based on sophisticated deep learning architectures achieve disappointing results compared to naïve baselines (clinical correctness or factual accuracy)

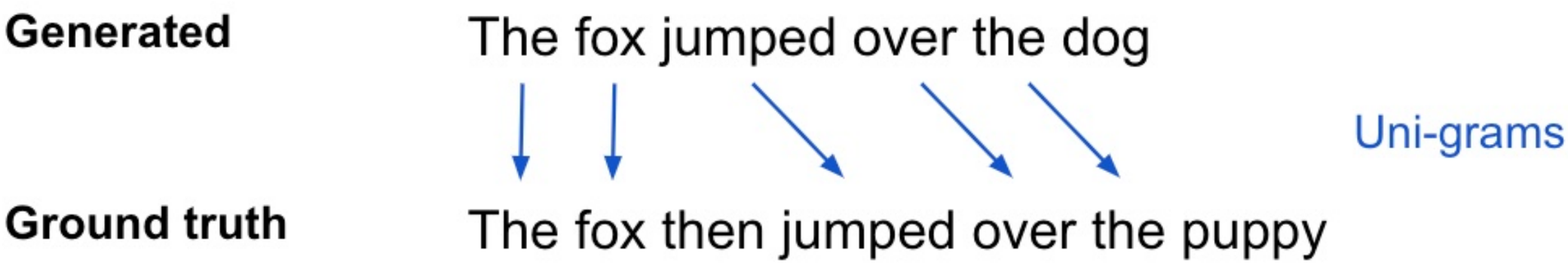
ihealth | Millennium Institute for Intelligent Healthcare Engineering

BLEU (Papineni et al., 2002)

- Counts n-grams matches in the ground truth
- Precision-based
- Brevity penalty
- BLEU-N uses up to N-grams (1-4)

BLEU (Papineni et al., 2002)

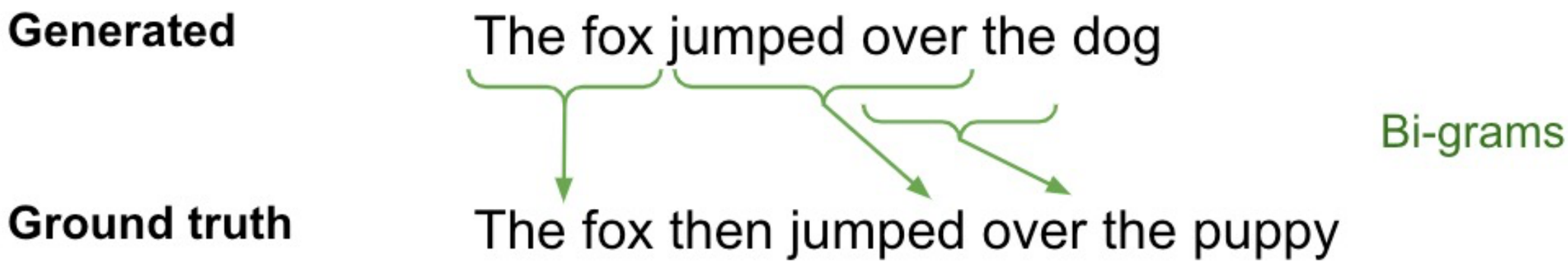
- Counts n-grams matches in the ground truth
- Precision-based
- Brevity penalty
- BLEU-N uses up to N-grams (1-4)



ihealth | Millennium Institute for Intelligent Healthcare Engineering

BLEU (Papineni et al., 2002)

- Counts n-grams matches in the ground truth
- Precision-based
- Brevity penalty
- BLEU-N uses up to N-grams (1-4)



ihealth | Millennium Institute for Intelligent Healthcare Engineering

ROUGE (Lin, 2004)

- Finds the largest common subsequence
- Sentence A is a subsequence of sentence B if:
  - All words of A appear in the same order in B
  - B may have other words in between
- Harmonic average biased toward recall

Generated The fox jumped over the dog

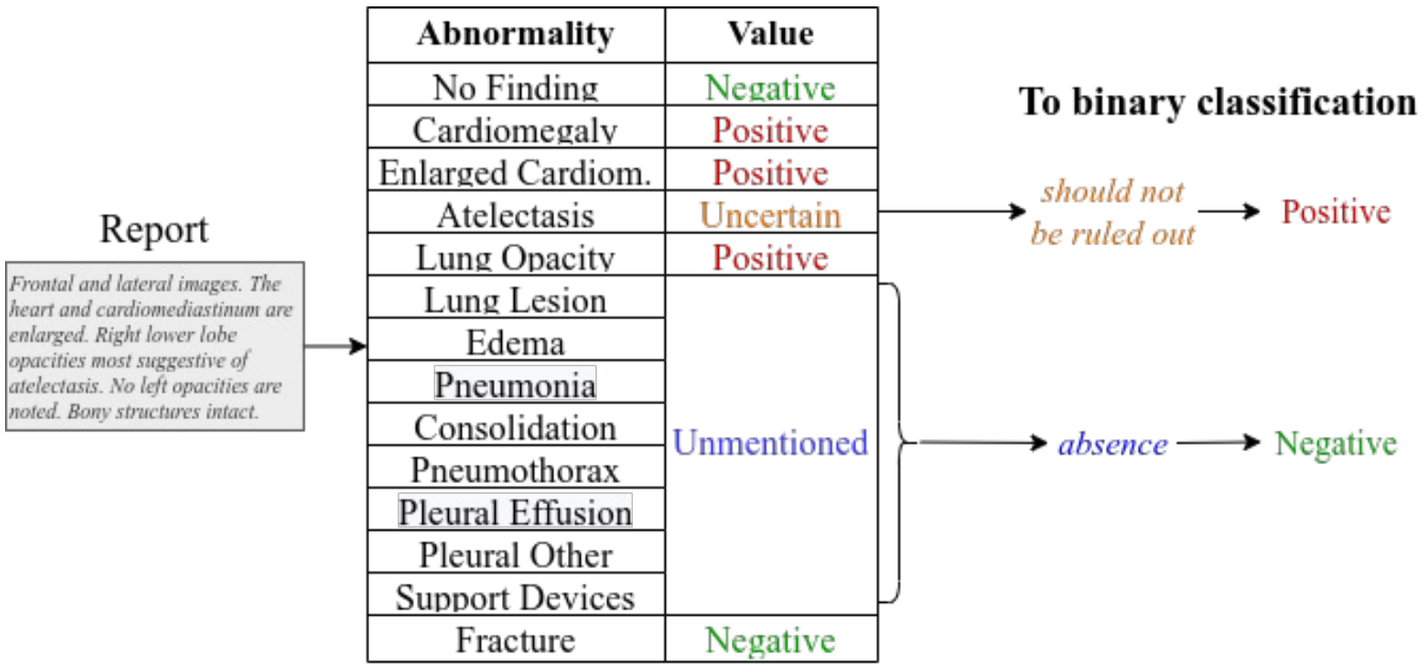
Ground truth The fox then jumped over the puppy

Largest common subsequence

ihealth | Millennium Institute for Intelligent Healthcare Engineering

The Chexpert labeler

- A key part of the current evaluation process is the CheXpert labeler, a tool developed at Stanford based on NegBio which labels the presence of abnormalities (limited to 13 + no finding)



Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., ... & Ng, A. Y. (2019, July). **Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison**. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 33, No. 01, pp. 590-597).

Peng, Y., Wang, X., Lu, L., Bagheri, M., Summers, R., & Lu, Z. (2018). **NegBio: a high-performance tool for negation and uncertainty detection in radiology reports**. *AMIA Summits on Translational Science Proceedings*, *2018*, 188.

NegBio
A tool for negation and uncertainy detection

ihealth | Millennium Institute for Intelligent Healthcare Engineering

Context II: An example

- Findings sections and three generated reports, with BLEU (B), ROUGE-L (R-L) and Chexpert metrics calculated. **Correct sentences are in bold** and and *incorrect sentences are in italics*.

Report	NLP		Chexpert		
	B	R-L	F-1	P	R
Ground-truth: Heart size is mildly enlarged. Small right pneumothorax is seen.	-	-	-	-	-

ihealth | Millennium Institute for Intelligent Healthcare Engineering

Context II: An example

Report	NLP B	NLP R-L	Chexpert F-1	Chexpert P	Chexpert R
Ground-truth: Heart size is mildly enlarged. Small right pneumothorax is seen.	-	-	-	-	-
Heart size is normal. No pneumothorax is seen.	**0.493**	**0.715**	0	0	0

ihealth | Millennium Institute for Intelligent Healthcare Engineering

Context II: An example

Report	NLP		Chexpert		
	B	R-L	F-1	P	R
Ground-truth: Heart size is mildly enlarged. Small right pneumothorax is seen.	-	-	-	-	-
Heart size is normal. No pneumothorax is seen.	**0.493**	**0.715**	0	0	0
The cardiac silhouette is enlarged. *No pneumothorax.*	0.146	0.464	0.5	0.5	0.5

ihealth | Millennium Institute for Intelligent Healthcare Engineering

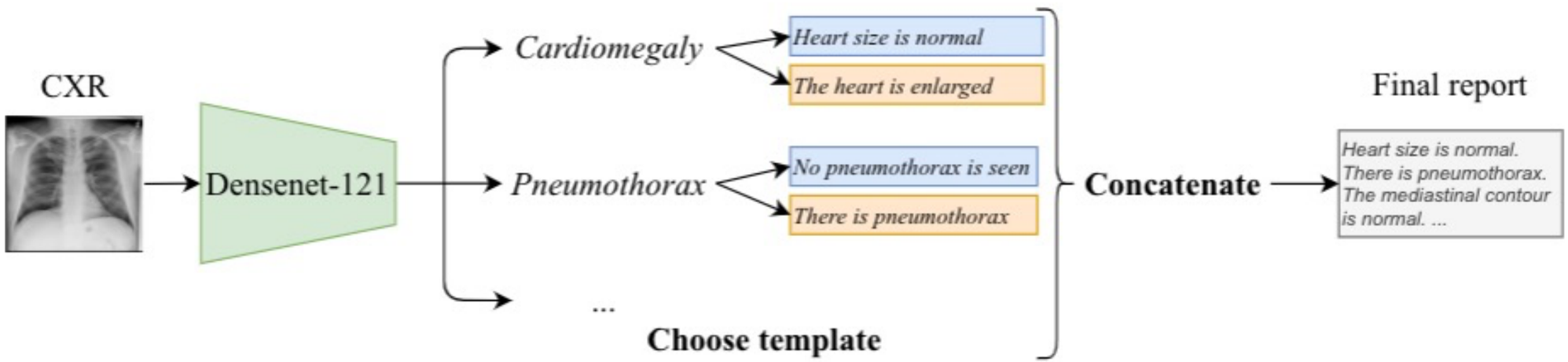
Context II: An example

Report	NLP		Chexpert		
	B	R-L	F-1	P	R
Ground-truth: Heart size is mildly enlarged. Small right pneumothorax is seen.	-	-	-	-	-
Heart size is normal. No pneumothorax is seen.	**0.493**	**0.715**	0	0	0
The cardiac silhouette is enlarged. *No pneumothorax.*	0.146	0.464	0.5	0.5	0.5
Mild cardiomegaly. Pneumothorax on right lung.	0.075	0.289	1	1	1

ihealth | Millennium Institute for Intelligent Healthcare Engineering

Our initial approach: CNN-TRG

- CNN-TRG detects abnormalities in the image using a CNN abnormality classifier and fixed sentences as templates for text generation.



- Sentence generation: ***single*** or ***grouped***-based (e.g. all cardio).

Experiments

- **Task**: Generate the *Findings* section, keeping only frontal chest X-rays
- Using both IU X-ray an MIMIC-CXR:
  - IU X-ray: 7,470 images and 3,955 reports
  - MIMIC-CXR: 377,110 images and 227,827 reports
- Train/validation/test split:
  - IU x-ray: random split 80/10/10, MIMIC-CXR: official train/validation/test split
- Metrics:
  - NLP/NLG: BLEU (B) [0-1], ROUGE-L (R-L) [0-1], CIDEr-D (C-D) [0-10]
  - Clinical correctness: Chexpert-labeler (P, R, F-1) and MIRQI (P, R, F-1)

Experiments II : Baselines

- Naïve models:
  - Fixed constant report
  - Random report
  - 1-NN: copy the report of the most similar image in the dataset
- Deep Encoder-Decoder:
  - Our CNN visual encoder (p.t. as CNN-TRG) + LSTM with attention as decoder (p.t. RadGLove)
- Other models:
  - We present the results reported in the original articles (we only implement CoAtt)

Results in MIMIC-CXR

	Model	NLP B	NLP R-L	NLP C-D	Chexpert F-1	Chexpert P	Chexpert R	MIRQI F-1	MIRQI P	MIRQI R
MIMIC-CXR	Constant	0.137	0.201	0.059	0.021	0.012	0.071	0.163	0.158	0.176
	Random	0.073	0.142	0.078	0.163	0.186	0.151	0.359	0.372	0.362
	1-nn	0.119	0.193	0.151	0.320	0.325	0.319	0.635	0.645	0.641
	CNN-LSTM-att [L]	0.103	0.244	0.479	0.308	0.378	0.297	0.644	0.652	**0.648**
	CoAtt* [10] [L]	0.120	0.252	0.401	0.201	0.356	0.198	0.544	0.551	0.545
	Boag et al. [2] [L]	0.184	-	0.850	0.186	0.304	-	-	-	-
	Liu et al. [16] [L]	0.192	0.306	**1.046**	-	0.309	0.134	-	-	-
	Chen et al. [3] [T]	0.205	0.277	-	0.276	0.333	0.273	-	-	-
	Lovelace et al. [17] [T]	**0.257**	**0.318**	0.316	0.228	0.333	0.217	-	-	-
	CVSE [20] [R,Ab]	-	0.153	-	0.253	0.317	0.224	-	-	-
	RTEX [13] [R]	-	0.205	-	-	0.229	0.284	-	-	-
	CNN-TRG single	0.080	0.151	0.026	**0.428**	**0.381**	**0.531**	**0.668**	**0.749**	0.640
	CNN-TRG grouped	0.094	0.185	0.238	**0.428**	**0.381**	**0.531**	0.666	0.746	0.637

Results in MIMIC-CXR

	Model	NLP B	NLP R-L	NLP C-D	Chexpert F-1	Chexpert P	Chexpert R	MIRQI F-1	MIRQI P	MIRQI R
MIMIC-CXR	Constant	0.137	0.201	0.059	0.021	0.012	0.071	0.163	0.158	0.176
	Random	0.073	0.142	0.078	0.163	0.186	0.151	0.359	0.372	0.362
	1-nn	0.119	0.193	0.151	0.320	0.325	0.319	0.635	0.645	0.641
	CNN-LSTM-att ^{L}	0.103	0.244	0.479	0.308	0.378	0.297	0.644	0.652	**0.648**
	CoAtt*[10]^{L}	0.120	0.252	0.401	0.201	0.356	0.198	0.544	0.551	0.545
	Boag et al. [2]^{L}	0.184	-	0.850	0.186	0.304	-	-	-	-
	Liu et al. [16]^{L}	0.192	0.306	**1.046**	-	0.309	0.134	-	-	-
	Chen et al. [3]^{T}	0.205	0.277	-	0.276	0.333	0.273	-	-	-
	Lovelace et al. [17]^{T}	**0.257**	**0.318**	0.316	0.228	0.333	0.217	-	-	-
	CVSE [20]^{R,Ab}	-	0.153	-	0.253	0.317	0.224	-	-	-
	RTEX [13]^{R}	-	0.205	-	-	0.229	0.284	-	-	-
	CNN-TRG single	0.080	0.151	0.026	**0.428**	**0.381**	**0.531**	**0.668**	**0.749**	0.640
	CNN-TRG grouped	0.094	0.185	0.238	**0.428**	**0.381**	**0.531**	0.666	0.746	0.637

Results in MIMIC-CXR

	Model	NLP B	NLP R-L	NLP C-D	Chexpert F-1	Chexpert P	Chexpert R	MIRQI F-1	MIRQI P	MIRQI R
MIMIC-CXR	Constant	0.137	0.201	0.059	0.021	0.012	0.071	0.163	0.158	0.176
	Random	0.073	0.142	0.078	0.163	0.186	0.151	0.359	0.372	0.362
	1-nn	0.119	0.193	0.151	0.320	0.325	0.319	0.635	0.645	0.641
	CNN-LSTM-att [L]	0.103	0.244	0.479	0.308	0.378	0.297	0.644	0.652	**0.648**
	CoAtt* [10] [L]	0.120	0.252	0.401	0.201	0.356	0.198	0.544	0.551	0.545
	Boag et al. [2] [L]	0.184	-	0.850	0.186	0.304	-	-	-	-
	Liu et al. [16] [L]	0.192	0.306	**1.046**	-	0.309	0.134	-	-	-
	Chen et al. [3] [T]	0.205	0.277	-	0.276	0.333	0.273	-	-	-
	Lovelace et al. [17] [T]	**0.257**	**0.318**	0.316	0.228	0.333	0.217	-	-	-
	CVSE [20] [R,Ab]	-	0.153	-	0.253	0.317	0.224	-	-	-
	RTEX [13] [R]	-	0.205	-	-	0.229	0.284	-	-	-
	CNN-TRG single	0.080	0.151	0.026	**0.428**	**0.381**	**0.531**	**0.668**	**0.749**	0.640
	CNN-TRG grouped	0.094	0.185	0.238	**0.428**	**0.381**	**0.531**	0.666	0.746	0.637

Conclusion

- ***CNN-TRG Clinical Correctness***. Our template-based models outperform all other models (naïve and DL-based) in terms of clinical correctness, both in Chexpert and MIRQI F-1 scores.
- ***NLP vs Clinical Correctness***. Naive models achieve higher NLP performance than CNN-TRG and comparable to some SOTA models, even though they are not clinically useful by design. However, naive models achieve very low performance on Chexpert and MIRQI.

ihealth | Millennium Institute for Intelligent Healthcare Engineering

What to do without templates?

- What do we do if we do not have templates for other abnormalities?

We need an actual generator (NLG)

ihealth | Millennium Institute for Intelligent Healthcare Engineering

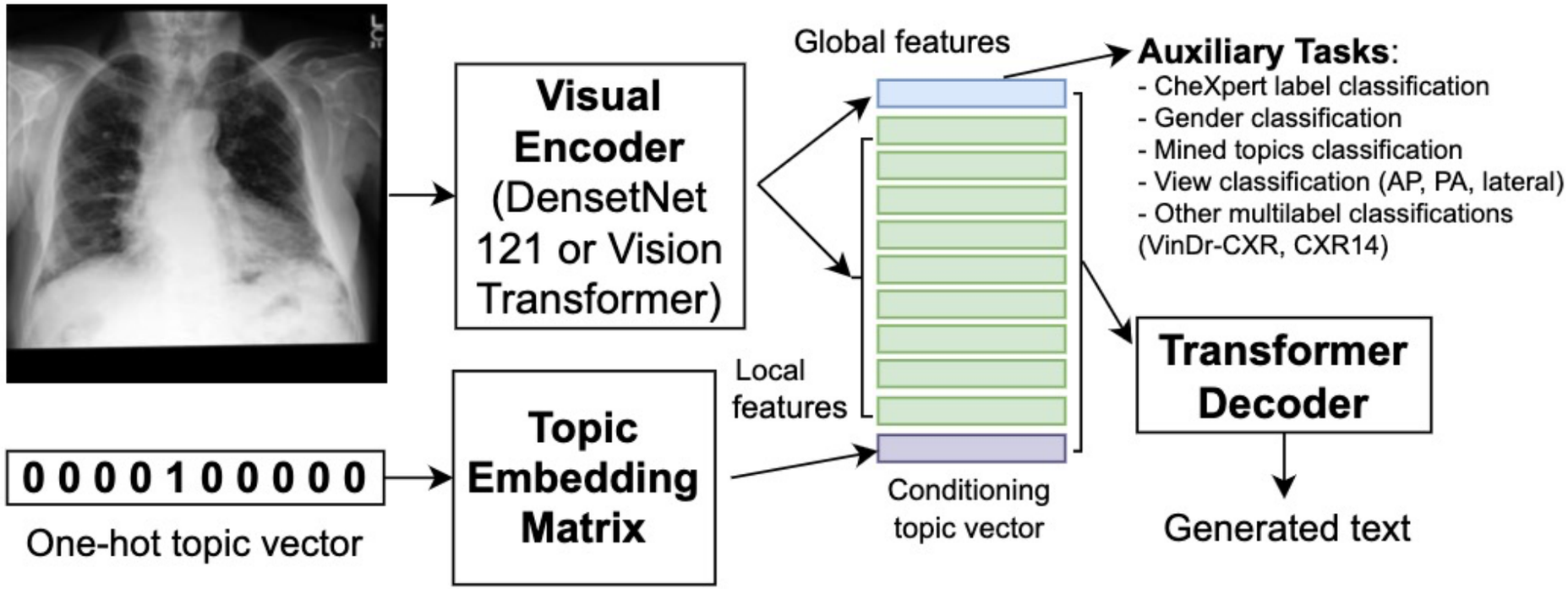
MedNeurIPS 2022: Two-stage conditional report generation

Two-stage Conditional Chest X-ray Radiology Report Generation

Pablo Messina [1,5,6], José Cañete [2,6], Denis Parra [1,5,6], Álvaro Soto [1,6], Cecilia Besa [3,5], and Jocelyn Dunstan [4,5]

ihealth | Millennium Institute for Intelligent Healthcare Engineering

MedNeurIPS 2022: Two-stage conditional report generation



ihealth | Millennium Institute for Intelligent Healthcare Engineering

MedNeurIPS 2022: Two-stage conditional report generation

ID	Model	NLP B	NLP R-L	NLP C-D	Med. Comp.	CheXpert (Macro) F1	CheXpert (Macro) P	CheXpert (Macro) R	CheXpert (Micro) F1	CheXpert (Micro) P	CheXpert (Micro) R
	Other works										
1	Liu et al. [17]	0.192	0.306	**1.046**	-	0.180	0.313	0.126	0.334	**0.634**	0.227
2	Chen et al. 2020 [7]	0.205	0.277	-	-	0.276	0.333	0.273	-	-	-
3	Chen et al. 2021 [6]	0.208	0.283	-	-	0.303	0.352	0.298	-	-	-
4	Lovelace et al. [18]	0.257	0.318	0.316	-	0.228	0.333	0.217	0.441	0.475	0.361
5	Miura et al. [20]	-	-	0.509	-	0.304	0.361	0.360	0.563	0.499	**0.646**
6	Nguyen et al. [22]	**0.339**	**0.390**	-	-	0.412	**0.432**	0.418	**0.576**	0.567	0.585
7	Pino et al. [24]	0.094	0.185	0.238	-	**0.428**	0.381	**0.531**	-	-	-
8	Kong et al. [14]	0.243	0.286	-	-	-	-	-	0.519	0.482	0.563
	Our work										
9	$\text{CRG(DN+TF)}_{\text{chexpert topics}}$: M	0.146	0.196	**0.041**	0.087	0.464	0.377	0.713	0.557	0.428	0.797
10	$\text{CRG(DN+TF)}_{\text{chexpert topics}}$: M+I	0.146	0.196	**0.041**	0.087	0.469	0.388	0.678	0.569	0.448	0.781
11	$\text{CRG(DN+TF)}_{\text{chexpert topics}}$: M+I+Ch	0.146	0.196	**0.041**	0.088	0.463	0.384	0.689	0.568	0.446	0.783
12	$\text{CRG(DN+TF)}_{\text{chexpert topics}}$: M+I+Ch+C14	0.146	0.196	0.040	0.088	0.463	0.386	0.702	0.564	0.440	0.785
13	$\text{CRG(DN+TF)}_{\text{chexpert topics}}$: M+I+Ch+C14+$V_{\text{test}}$	0.145	0.195	0.040	0.088	0.467	0.386	**0.712**	0.569	0.439	**0.811**
14	$\text{CRG(DN+TF)}_{\text{chexpert topics}}$: M+I+Ch+C14+$V_{all}$	0.145	0.195	**0.041**	0.088	0.462	0.383	0.700	0.571	0.444	0.800
15	$\text{CRG(DN+TF)}^{\text{medtok, fve, ft}}_{\text{chexpert topics}}$: M+I	0.146	0.197	0.040	0.086	**0.477**	0.392	0.693	0.575	0.449	0.799
16	$\text{CRG(ViT}_{\text{CLIP}}\text{+TF)}^{\text{medtok, fve, ft}}_{\text{chexpert topics}}$: M+I	**0.150**	**0.199**	0.040	0.087	0.472	0.389	0.653	0.582	0.464	0.779
17	$\text{CRG(DN+TF)}^{\text{medtok, vmf, ft}}_{\text{mined topics predicted by ensemble}}$: M+I	0.102	0.184	0.031	**0.116**	0.448	**0.400**	0.568	**0.588**	**0.487**	0.743

ihealth | Millennium Institute for Intelligent Healthcare Engineering

MedNeurIPS 2022: Is the NLG considering the input image?

Table 2: Visual encoder results on the test split of MIMIC-CXR. For CRG models, CheXpert metrics for the Transformer decoder when conditioned on CheXpert topics are included. Cohen's Kappa measures the agreement between visual encoder and Transformer.

ID	Model	Mined Topics		CheXpert (visual encoder)				CheXpert (transformer)		
		F1 (macro)	F1 (micro)	ROC-AUC (macro)	ROC-AUC (micro)	F1 (macro)	F1 (micro)	F1 (macro)	F1 (micro)	Cohen's Kappa
1	CRG(DN+TF): M	0.213	0.413	0.758	0.823	0.473	0.578	0.465	0.556	0.703
2	CRG(DN+TF): M+I	0.208	0.409	0.750	0.821	0.465	0.579	**0.476**	0.570	0.767
3	CRG(DN+TF): M+I+Ch	0.206	0.392	0.763	0.821	0.472	0.576	0.474	0.569	0.785
4	CRG(DN+TF): M+I+Ch+C14	0.209	0.401	**0.765**	0.823	0.478	0.582	0.470	0.563	0.806
5	CRG(DN+TF): M+I+Ch+C14+V_{test}	0.212	0.405	0.761	0.826	0.483	0.587	**0.476**	0.570	0.785
6	CRG(DN+TF): M+I+Ch+C14+V_{all}	0.210	0.405	0.762	**0.829**	0.481	0.587	0.475	0.572	0.806
7	CRG(DN+TF)^{medtok}: M+I+Ch+C14+V_{test}	0.216	**0.422**	**0.765**	**0.829**	**0.486**	**0.593**	0.473	0.569	0.781
8	CRG(ViT_{CLIP}+TF)$^{medtok,\ fve,\ ft}$: M+I	**0.219**	0.396	0.743	0.823	0.471	0.590	0.474	**0.582**	**0.811**
9	TC(DN+ChEmb+Bilstm)$^{e=191}$: M+I+Ch+C14+V_{all}	0.233	0.443	0.715	0.804	0.450	0.566	-	-	-
10	TC(ChEmb+Bilstm)$^{e=74,ft}$: M+I+Ch	0.230	0.497	**0.744**	0.804	0.463	0.562	-	-	-
11	TC(DN+ChEmb+Bilstm)$^{e=180,ft}$: M+I+Ch	0.234	**0.516**	0.740	0.813	0.465	0.573	-	-	-
12	TC(DN+ChEmb+Bilstm)$^{e=568,ft}$: M+I+Ch+C14+V_{all}	**0.239**	0.507	0.741	**0.817**	**0.466**	**0.579**	-	-	-
13	TC Ensemble	**0.310**	**0.603**	-	-	-	-	-	-	

ihealth | Millennium Institute for Intelligent Healthcare Engineering

Flamingo-Med

Table 2: Visual encoder results on the test split of MIMIC-CXR. For CRG models, CheXpert metrics for the Transformer decoder when conditioned on CheXpert topics are included. Cohen's Kappa measures the agreement between visual encoder and Transformer.

ID	Model	Mined Topics		CheXpert (visual encoder)				CheXpert (transformer)		
		F1 (macro)	F1 (micro)	ROC-AUC (macro)	ROC-AUC (micro)	F1 (macro)	F1 (micro)	F1 (macro)	F1 (micro)	Cohen's Kappa
1	CRG(DN+TF): M	0.213	0.413	0.758	0.823	0.473	0.578	0.465	0.556	0.703
2	CRG(DN+TF): M+I	0.208	0.409	0.750	0.821	0.465	0.579	**0.476**	0.570	0.767
3	CRG(DN+TF): M+I+Ch	0.206	0.392	0.763	0.821	0.472	0.576	0.474	0.569	0.785
4	CRG(DN+TF): M+I+Ch+C14	0.209	0.401	**0.765**	0.823	0.478	0.582	0.470	0.563	0.806
5	CRG(DN+TF): M+I+Ch+C14+V_{test}	0.212	0.405	0.761	0.826	0.483	0.587	**0.476**	0.570	0.785
6	CRG(DN+TF): M+I+Ch+C14+V_{all}	0.210	0.405	0.762	**0.829**	0.481	0.587	0.475	0.572	0.806
7	CRG(DN+TF)^{medtok}: M+I+Ch+C14+V_{test}	0.216	**0.422**	**0.765**	**0.829**	**0.486**	**0.593**	0.473	0.569	0.781
8	CRG(ViT_{CLIP}+TF)$^{medtok,\ fve,\ ft}$: M+I	**0.219**	0.396	0.743	0.823	0.471	0.590	0.474	**0.582**	**0.811**
9	TC(DN+ChEmb+Bilstm)$^{e=191}$: M+I+Ch+C14+V_{all}	0.233	0.443	0.715	0.804	0.450	0.566	-	-	-
10	TC(ChEmb+Bilstm)$^{e=74,ft}$: M+I+Ch	0.230	0.497	**0.744**	0.804	0.463	0.562	-	-	-
11	TC(DN+ChEmb+Bilstm)$^{e=180,ft}$: M+I+Ch	0.234	**0.516**	0.740	0.813	0.465	0.573	-	-	-
12	TC(DN+ChEmb+Bilstm)$^{e=568,ft}$: M+I+Ch+C14+V_{all}	**0.239**	0.507	0.741	**0.817**	**0.466**	**0.579**	-	-	-
13	TC Ensemble	**0.310**	**0.603**	-	-	-	-	-	-	

ihealth | Millennium Institute for Intelligent Healthcare Engineering

LLAVA-Med

Table 2: Visual encoder results on the test split of MIMIC-CXR. For CRG models, CheXpert metrics for the Transformer decoder when conditioned on CheXpert topics are included. Cohen's Kappa measures the agreement between visual encoder and Transformer.

ID	Model	Mined Topics		CheXpert (visual encoder)				CheXpert (transformer)		
		F1 (macro)	F1 (micro)	ROC-AUC (macro)	ROC-AUC (micro)	F1 (macro)	F1 (micro)	F1 (macro)	F1 (micro)	Cohen's Kappa
1	CRG(DN+TF): M	0.213	0.413	0.758	0.823	0.473	0.578	0.465	0.556	0.703
2	CRG(DN+TF): M+I	0.208	0.409	0.750	0.821	0.465	0.579	**0.476**	0.570	0.767
3	CRG(DN+TF): M+I+Ch	0.206	0.392	0.763	0.821	0.472	0.576	0.474	0.569	0.785
4	CRG(DN+TF): M+I+Ch+C14	0.209	0.401	**0.765**	0.823	0.478	0.582	0.470	0.563	0.806
5	CRG(DN+TF): M+I+Ch+C14+V_{test}	0.212	0.405	0.761	0.826	0.483	0.587	**0.476**	0.570	0.785
6	CRG(DN+TF): M+I+Ch+C14+V_{all}	0.210	0.405	0.762	**0.829**	0.481	0.587	0.475	0.572	0.806
7	CRG(DN+TF)^{medtok}: M+I+Ch+C14+V_{test}	0.216	**0.422**	**0.765**	**0.829**	**0.486**	**0.593**	0.473	0.569	0.781
8	CRG(ViT_{CLIP}+TF)$^{medtok,\ fve,\ ft}$: M+I	**0.219**	0.396	0.743	0.823	0.471	0.590	0.474	**0.582**	**0.811**
9	TC(DN+ChEmb+Bilstm)$^{e=191}$: M+I+Ch+C14+V_{all}	0.233	0.443	0.715	0.804	0.450	0.566	-	-	-
10	TC(ChEmb+Bilstm)$^{e=74,ft}$: M+I+Ch	0.230	0.497	**0.744**	0.804	0.463	0.562	-	-	-
11	TC(DN+ChEmb+Bilstm)$^{e=180,ft}$: M+I+Ch	0.234	**0.516**	0.740	0.813	0.465	0.573	-	-	-
12	TC(DN+ChEmb+Bilstm)$^{e=568,ft}$: M+I+Ch+C14+V_{all}	**0.239**	0.507	0.741	**0.817**	**0.466**	**0.579**	-	-	-
13	TC Ensemble	**0.310**	**0.603**	-	-	-	-	-	-	

ihealth | Millennium Institute for Intelligent Healthcare Engineering

SOTA & Challenges

ihealth | Millennium Institute for Intelligent Healthcare Engineering

State-of-the-art (SOTA)

- Very active research area, but validation and generalization has big room ahead

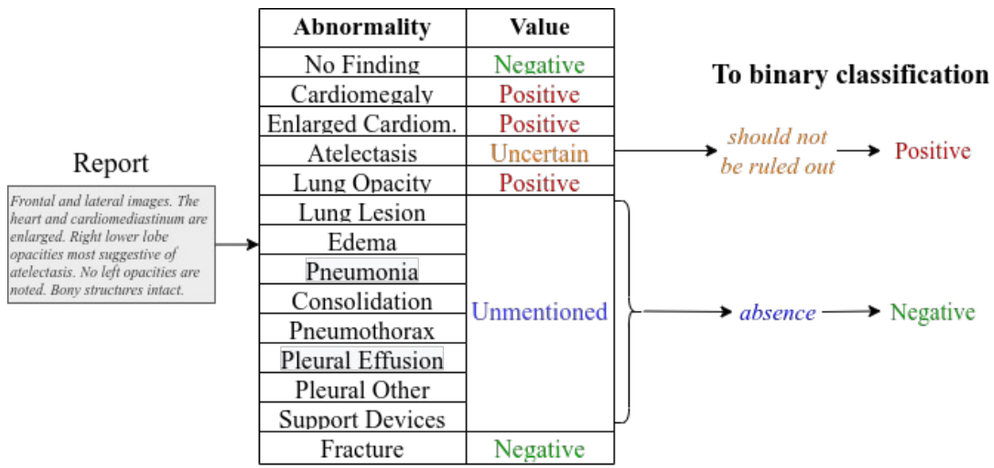
Approach	Method	Year	BLEU1-4	ROUGE-L	F-1	Precision	Recall
Baseline	Random	-	0.073	0.142	0.163	0.186	0.151
Retrieval	Kougia et al. (RTEX)	2021	-	0.205	-	0.229	0.284
LSTM	Boag et al.	2020	0.184		0.186	0.304	-
LSTM	Liu et al.	2019	**0.192**	**0.306**	-	0.309	0.134
Transformer	Nguyen et al.	2021	**0.339**	**0.222**	-	-	-
Transformer	Miura et al. (M2 TRANS)	2021	-	0.114	**0.567**	**0.503**	**0.651**
Template	Pino et al. (CNN-TRG)	2021	0.094	0.185	**0.428**	**0.381**	**0.531**

Table 1. Results of top-performing methods using different approaches for RadRG. The first row (randomly generated reports) provides a comparative baseline. Bold font indicates the top 2 performing methods in each metric. F-1, Precision, and Recall refer to medical correctness metrics computed with the Chexpert labeler.

ihealth | Millennium Institute for Intelligent Healthcare Engineering

The Chexpert labeler – move ahead and faster

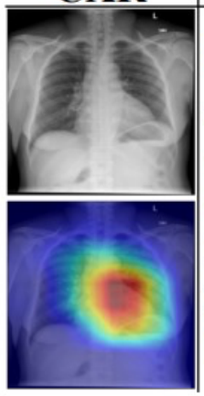
- How to make the Chexpert labeler deal with evaluation on:
  - Other abnormalities (beyond 13 original)
  - Other parts of the body (beyond Chest)
  - Other image modalities



ihealth | Millennium Institute for Intelligent Healthcare Engineering

Explainability

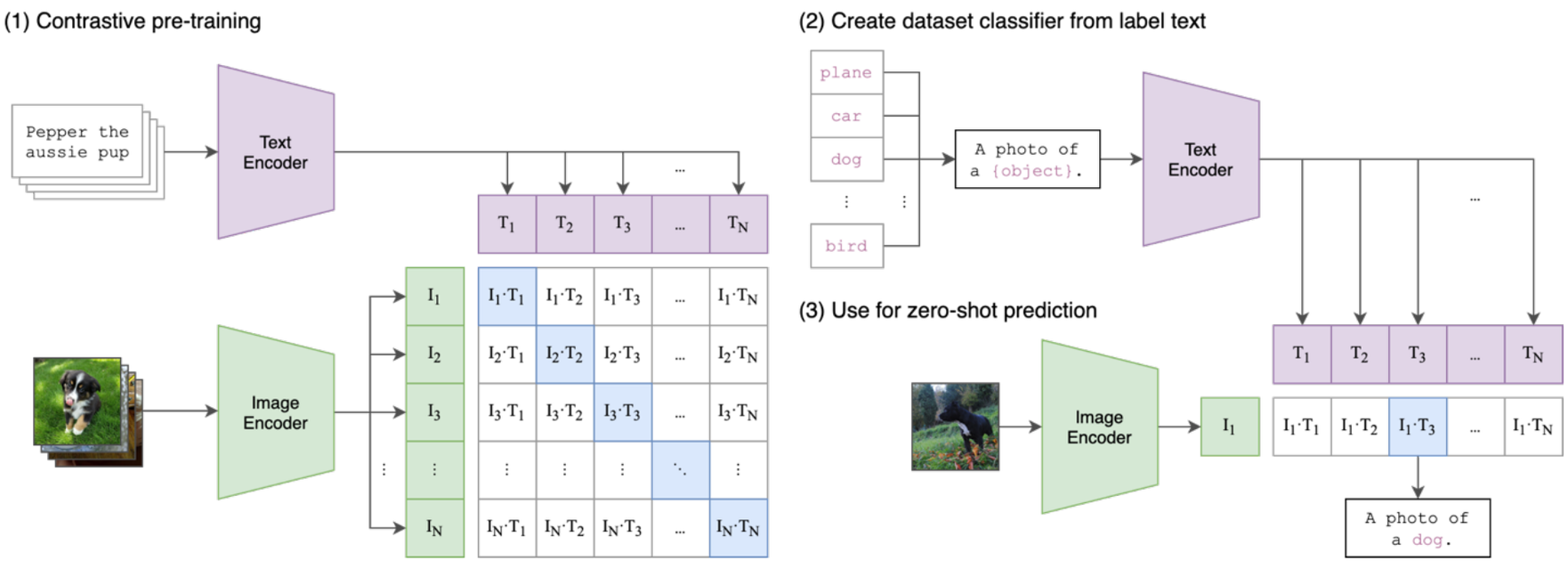
- Methods like CAM, gradCAM, integrated and saliency maps are useful but they are susceptible to adversarial attacks

CXR	Generated by CNN-TRG	Ground Truth
	The heart is enlarged. The mediastinal contour is normal. No focal consolidation. The lungs are free of focal airspace disease. No atelectasis. No pleural effusion. No fibrosis. No pneumonia. No pneumothorax is seen. No pulmonary edema. No pulmonary nodules or mass lesions identified. No fracture is seen.	The heart is mildly enlarged. Left hemidiaphragm is elevated. There is no acute infiltrate or pleural effusion. The mediastinum is unremarkable.

ihealth | Millennium Institute for Intelligent Healthcare Engineering

Contrastive Multimodal learning

- Methods like CLIP seem promising, but radiology reports have long text so traditional methods might work in a straight-forward manner



ihealth | Millennium Institute for Intelligent Healthcare Engineering

Challenges

- Expand to other pathologies & images' types (MRI, CT-Scan, Echotomography, etc.)
- Generate to other languages beyond English
- Deal with multimodal input (text, images, videos, tabular)
- Interpretable AI: explain predictions of results
- Improve generalization (OOD samples): transfer learning, metalearning, etc.
- Evaluation: NLG metrics vs actual clinical diagnostic

ihealth | Millennium Institute for Intelligent Healthcare Engineering

Acknowledgment

This work was funded by ANID - Millennium Science Initiative Program - ICN2021_004