Clinically Correct Report Generation from Chest X-Rays using Templates

Pablo Pino, Denis Parra, Computer Science Dept., PUC Chile & IMFD

Cecilia Besa, Claudio Lagos, School of Medicine, PUC Chile & CardioMR

Machine Learning for Medical Imaging (MLMI) Workshop 2021

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 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

The MIRG task

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

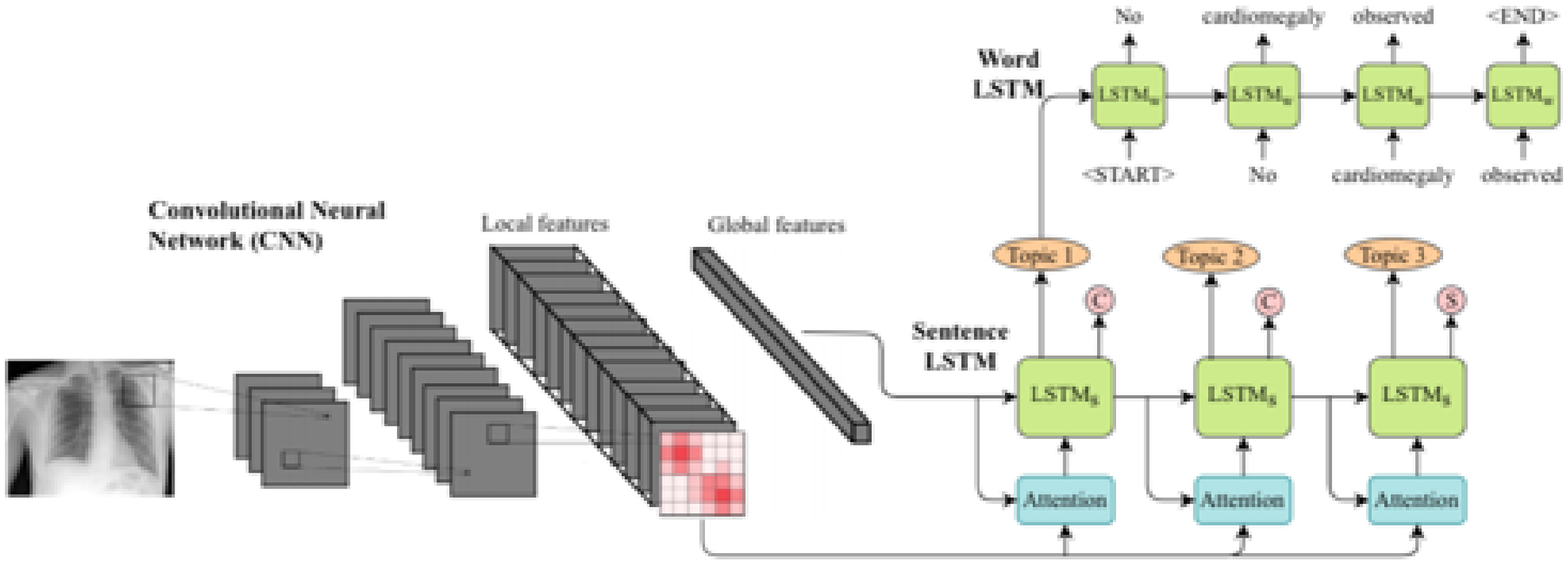
Manual tags: Calcified Granuloma/lung/upper lobe/right
Automatic tags: Calcified granuloma

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.

Traditional architecture: Deep Learning

- 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"

Context

- Quick progress of Deep Learning in Computer Vision and Natural Language Processing can potentially help 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)

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.	-	-	-	-	-

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.	-	-	-	-	-
Heart size is normal. No pneumothorax is seen.	**0.493**	**0.715**	0	0	0

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.	-	-	-	-	-
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

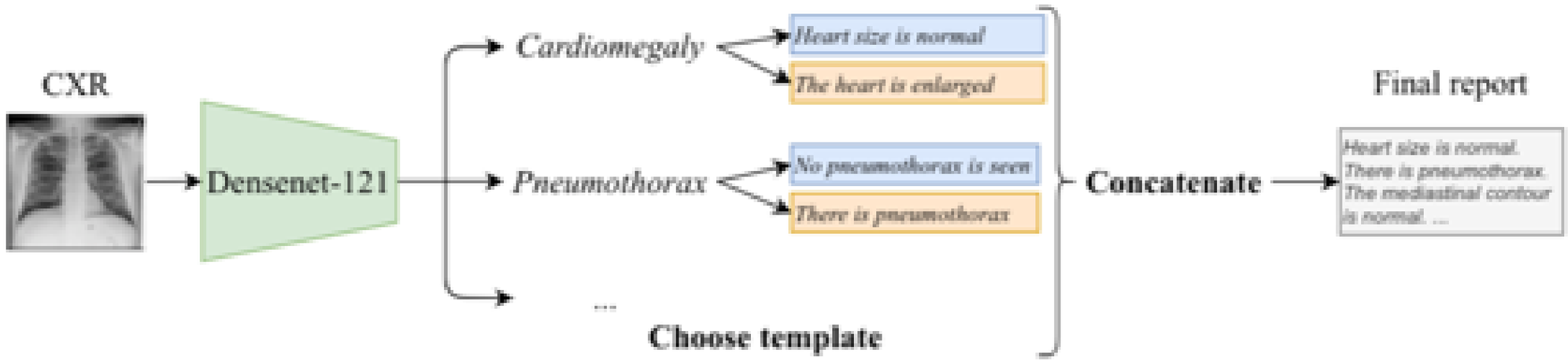
Context II: An example

- 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.	-	-	-	-	-
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**

Our 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 IU X-ray

	Model	NLP B	NLP R-L	NLP C-D	Chexpert F-1	Chexpert P	Chexpert R	MIRQI F-1	MIRQI P	MIRQI R
IU X-ray	Constant	0.297	0.366	0.307	0.038	0.026	0.071	0.469	0.462	0.481
	Random	0.202	0.284	0.145	0.066	0.065	0.068	0.374	0.378	0.384
	1-nn	0.220	0.301	0.245	0.145	0.150	0.144	0.497	0.508	0.500
	CNN-LSTM-att^{L}	0.202	0.319	0.208	0.140	0.159	0.148	0.484	0.492	0.487
	CoAtt*[10]^{L}	0.231	0.316	0.221	0.144	0.162	0.147	0.491	0.503	0.491
	Zhang et al.[31]$^{L,f+i}$	0.271	0.367	0.304	-	-	-	0.478	0.490	0.483
	CLARA [1]^{R}	0.302	-	**0.359**	-	-	-	-	-	-
	KERP [14]^{R}	0.299	0.339	0.280	-	-	-	-	-	-
	RTEX [13]^{R}	-	0.202	-	-	0.193	0.222	-	-	-
	S-M et al.[27]$^{R,f+i}$	**0.515**	**0.580**	-	-	-	-	-	-	-
	CNN-TRG single	0.167	0.282	0.030	**0.239**	**0.225**	**0.357**	**0.529**	0.534	**0.540**
	CNN-TRG grouped	0.273	0.352	0.249	**0.239**	**0.225**	**0.357**	**0.529**	**0.535**	**0.540**

Results in IU X-ray

	Model	NLP B	NLP R-L	NLP C-D	Chexpert F-1	Chexpert P	Chexpert R	MIRQI F-1	MIRQI P	MIRQI R
IU X-ray	Constant	0.297	0.366	0.307	0.038	0.026	0.071	0.469	0.462	0.481
	Random	0.202	0.284	0.145	0.066	0.065	0.068	0.374	0.378	0.384
	1-nn	0.220	0.301	0.245	0.145	0.150	0.144	0.497	0.508	0.500
	CNN-LSTM-att^{L}	0.202	0.319	0.208	0.140	0.159	0.148	0.484	0.492	0.487
	CoAtt*[10]^{L}	0.231	0.316	0.221	0.144	0.162	0.147	0.491	0.503	0.491
	Zhang et al.[31]$^{L,f+i}$	0.271	0.367	0.304	-	-	-	0.478	0.490	0.483
	CLARA [1]^{R}	0.302	-	**0.359**	-	-	-	-	-	-
	KERP [14]^{R}	0.299	0.339	0.280	-	-	-	-	-	-
	RTEX [13]^{R}	-	0.202	-	-	0.193	0.222	-	-	-
	S-M et al.[27]$^{R,f+i}$	**0.515**	**0.580**	-	-	-	-	-	-	-
	CNN-TRG single	0.167	0.282	0.030	**0.239**	**0.225**	**0.357**	**0.529**	0.534	**0.540**
	CNN-TRG grouped	0.273	0.352	0.249	**0.239**	**0.225**	**0.357**	**0.529**	**0.535**	**0.540**

Results in IU X-ray

	Model	NLP B	NLP R-L	NLP C-D	Chexpert F-1	Chexpert P	Chexpert R	MIRQI F-1	MIRQI P	MIRQI R
IU X-ray	Constant	0.297	0.366	0.307	0.038	0.026	0.071	0.469	0.462	0.481
	Random	0.202	0.284	0.145	0.066	0.065	0.068	0.374	0.378	0.384
	1-nn	0.220	0.301	0.245	0.145	0.150	0.144	0.497	0.508	0.500
	CNN-LSTM-att^{L}	0.202	0.319	0.208	0.140	0.159	0.148	0.484	0.492	0.487
	CoAtt*[10]^{L}	0.231	0.316	0.221	0.144	0.162	0.147	0.491	0.503	0.491
	Zhang et al.[31]$^{L,f+i}$	0.271	0.367	0.304	-	-	-	0.478	0.490	0.483
	CLARA [1]^{R}	0.302	-	**0.359**	-	-	-	-	-	-
	KERP [14]^{R}	0.299	0.339	0.280	-	-	-	-	-	-
	RTEX [13]^{R}	-	0.202	-	-	0.193	0.222	-	-	-
	S-M et al.[27]$^{R,f+i}$	**0.515**	**0.580**	-	-	-	-	-	-	-
	CNN-TRG single	0.167	0.282	0.030	**0.239**	**0.225**	**0.357**	**0.529**	0.534	**0.540**
	CNN-TRG grouped	0.273	0.352	0.249	**0.239**	**0.225**	**0.357**	**0.529**	**0.535**	**0.540**

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

Discussion

- ***Template Sets***: The clinical performance is the same for both **single** and **grouped** sets, since their clinical meaning is unchanged, but the grouped set achieves higher NLP performance.
- ***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. On the other hand, naive models achieve very low performance on Chexpert and MIRQI.

Future work

- ***Expand to other pathologies and types of images (MRI, CT-Scan, Ecography, etc.):*** current work is limited to the 13 abnormalities classified by Chexpert and focuses only on X-rays.
- ***Deal with multimodal input***: consider not only the image, but also the background information, specially to generate the *Impression* section of the report.
- ***Explainable AI***: our solution allows to easily integrate visual explainability methods such as CAM o Grad-CAM

Thanks!

Any comments of questions to:

Denis Parra dparra@ing.puc.cl

Pablo Pino pdpino@uc.cl


MICCAI2021
VIRTUAL
24TH INTERNATIONAL CONFERENCE ON MEDICAL IMAGE
COMPUTING & COMPUTER ASSISTED INTERVENTION
September 27 - October 1, 2021 • Strasbourg, FRANCE
MICCAI
University of Strasbourg
ICUBE