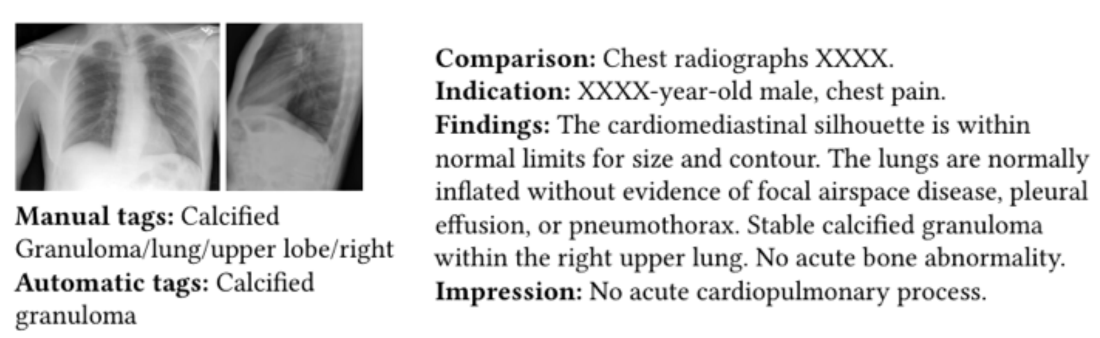
Clinically Correct Report Generation from Chest X-Rays using Templates

Pablo Pino, Denis Parra, Cecilia Besa, Claudio Lagos
Pontificia Universidad Católica de Chile

Introduction

- Using Artificial Intelligence (AI) for medical image report generation (MIRG) could help hospitals deal with a large and growing demand of image-based clinical exams

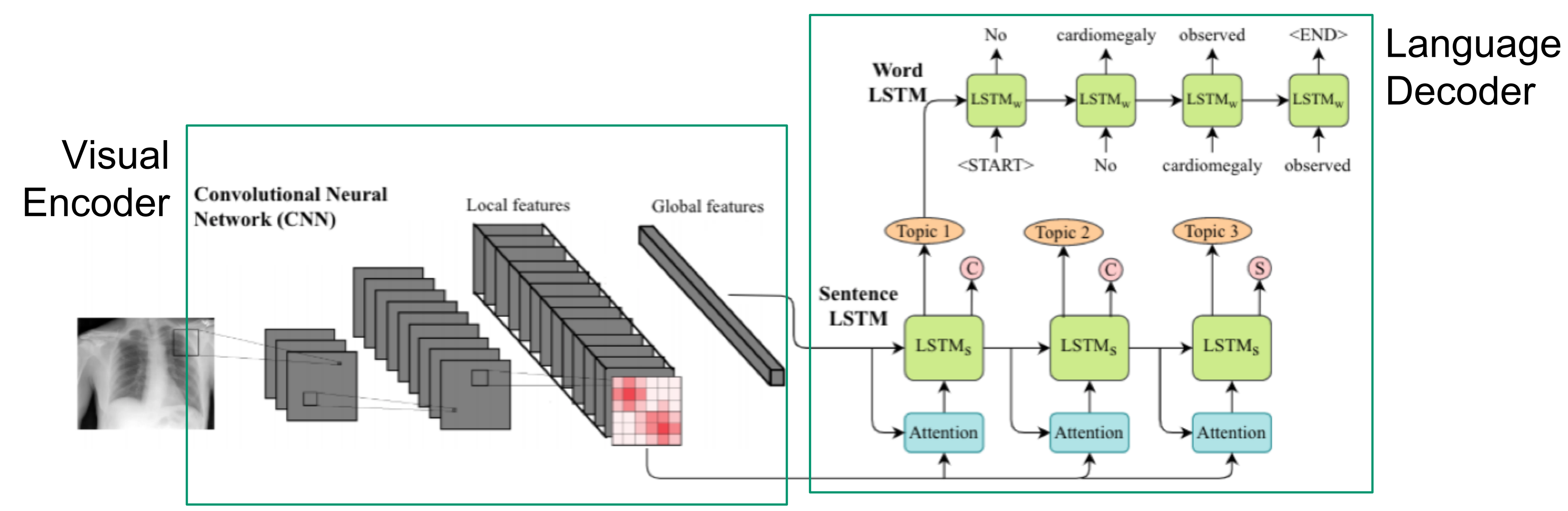


- SOTA models focus too much on NLP metrics (BLEU, ROUGE, etc.), which undermines its performance on clinical correctness (matching the diagnostics)
- In this article we focus on generating the **Findings** section from chest X-rays.

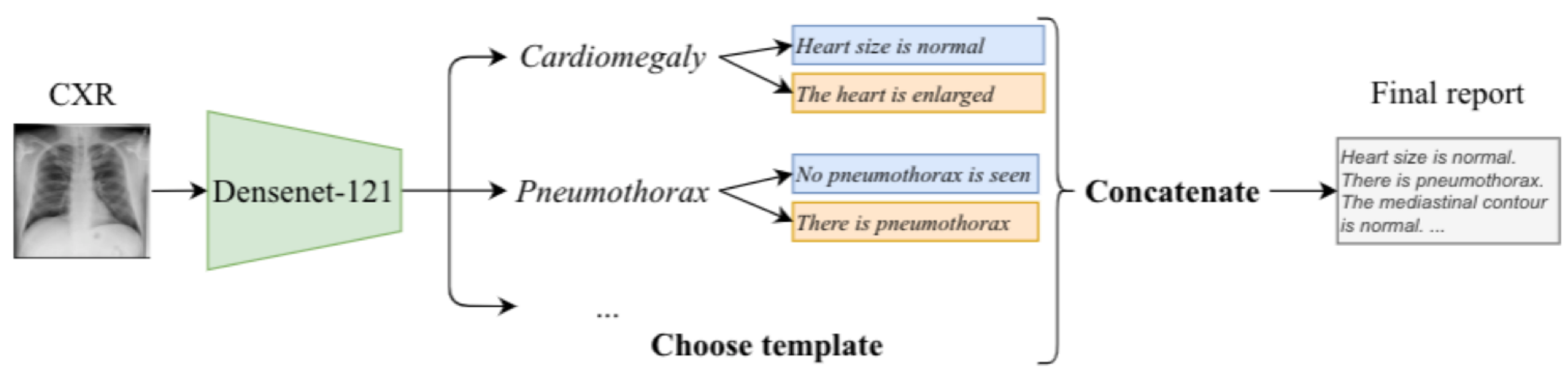
Experiments

- We use both IU X-ray and MIMIC-CXR datasets, keeping only frontal X-rays.
- We compare our proposed solution (CNN-TRG) with naive baselines as well as SOTA methods
- Among the naive baselines, we consider a fixed constant report, random reports, and a memory-based 1-NN report (retrieve the report of the most similar image)
- We compared the results of SOTA models reported in the literature.

Traditional Architecture: Encoder Decoder



CNN-TRG



Results

IU dataset

	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**

MIMIC-CXR dataset

	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

Conclusions

- ***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.

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 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

Acknowledgements

- ANID/FONDECYT 1191791. Instituto Milenio IMFD, convenio IMC-ANID.