Inspecting state of the art performance and NLP metrics in image-based medical report generation

Pablo Pino[1], Denis Parra[1], Pablo Messina[1], Cecilia Besa[2], Sergio Uribe[2]
[1]Department of Computer Science, [2]School of Medicine, Pontificia Universidad Católica de Chile
pdpino@uc.cl

Ground truth:
The cardiac silhouette **is enlarged**. The lungs **are hyperexpanded**. No pneumothorax or pleural effusion.

Generated:
The cardiac silhouette **is normal in size**. The lungs **are clear**. No pneumothorax or pleural effusion.

Fig 1: IU X-ray dataset [4] example (report extract). The generated report is clinically incorrect, but achieves high scores in NLP metrics: BLEU = 0.68, ROUGE-L = 0.82.

Introduction

- We study the task of **automatic medical report generation from images**, where many Deep Learning (DL) models have been proposed in the recent years [1,6-9]
- Authors mostly evaluate with NLP metrics [1], although these metrics **may not be able to capture medical facts** [2,3,6]
- Figure 1 shows an example with 2 out of 3 sentences clinically incorrect, though it achieves high scores:
  BLEU = 0.68 and ROUGE-L = 0.82

Motivation

- Are NLP metrics able to measure medical correctness?
- Are state of the art models achieving clinically useful results?

We **benchmark simple baselines and DL models** against SOTA, using the IU X-ray dataset [4]

Models

Simple baselines:
- *Constant* report
- *Random* report
- *Nearest-neighbor*: report from the most similar image
- *Top words/sentences*: most common words/sentences in a random order

DL models: CNN-LSTM with (and without) an attention mechanism over the image regions

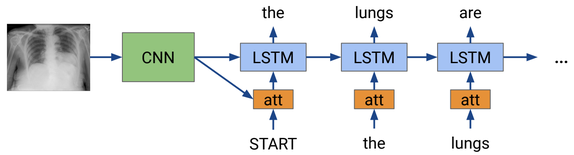

Fig 2: CNN-LSTM with attention model.

Metrics

- NLP metrics: BLEU, ROUGE-L, CIDEr-D
- CheXpert Labeler [5]
  - Attempts to measure clinical correctness
  - Classifies presence of 14 diseases from each report
  - Compute macro average Accuracy and ROC-AUC

Results

Model	B	R-L	C-D	CheXpert Acc	CheXpert AUC
Liu et al. [6]	0.225	**0.359**	**1.490**	**0.916**	-
TieNet [7]	0.182	0.311	**1.335**	0.902	-
KERP [8]	**0.298**	0.339	0.280	-	-
Xue et al. [9]	**0.310**	**0.380**	-	-	-
Constant	**0.289**	**0.357**	**0.293**	**0.915**	0.500
Random	0.188	0.264	0.112	0.894	0.508
Nearest Neighbor	0.211	0.288	0.230	0.903	**0.518**
Top sentences	0.198	0.281	0.166	0.911	0.498
Top words	0.124	0.224	0.075	0.835	**0.509**
CNN-LSTM	0.225	0.338	0.284	0.912	0.505
CNN-LSTM-att	0.211	0.314	0.187	**0.918**	**0.508**

Table 1: Results on the IU X-ray dataset [4]. B: BLEU, R-L: ROUGE-L, C-D: CIDEr-D. All metrics range from 0 to 1, except C-D from 0 to 10. Top-3 results per column are in bold.

Key results:

- *Constant* model achieves high performance
  - Near SOTA performance on BLEU, ROUGE-L and CheXpert-Accuracy
  - Surpasses some literature methods and our DL models in NLP metrics
- CheXpert-AUC is near 0.5 for all implemented methods
  - Thus, generated reports seem to be only slightly better than random, considering the 14 CheXpert diseases
  - Most literature models evaluate only with NLP metrics [1]
- Liu et al. [6] and TieNet [7] surpass baselines in CIDEr-D
- Liu et al. [6] and *Constant* achieve similar Chexpert-Accuracy

Conclusion

Either **SOTA do not perform well in a clinical sense**
and/or **NLP metrics are not able to differentiate** these methods

Future work:
- Further inspection of SOTA performance in medical accuracy
- Do NLP metrics correlate with expert medical judgement?
- How to measure medical correctness in a generated report? Are CheXpert metrics enough?

References

[1] P. Messina et al. “A Survey on Deep Learning and Explainability for Automatic Image-based Medical Report Generation”. In arXiv:2010.10563 (2020).
[2] W. Boag et al. “Baselines for Chest X-Ray Report Generation”. In: Proc of the Machine Learning for Health NeurIPS Workshop. 2020, pp. 126–140.
[3] Y. Zhang et al. “Optimizing the Factual Correctness of a Summary: A Study of Summarizing Radiology Reports”. In: Proc of the 58th Annual Meeting of the ACL. 2020, pp. 5108–5120.
[4] D. Demner-Fushman et al. “Preparing a collection of radiology examinations for distribution and retrieval”. In: Journal of the American Medical Informatics Assoc. (2015), pp. 304–310.
[5] J. Irvin et al. “Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison”. In: Proc of the AAAI Conf. on Artificial Intelligence. 2019, pp. 590–597.
[6] G. Liu et al. “Clinically Accurate Chest X-Ray Report Generation”. In: Machine Learning for Healthcare Conf. 2019, pp. 249–269.
[7] X. Wang et al. “Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays”. In: Proc of the IEEE Conf. on CVPR. 2018.
[8] C. Y. Li et al. “Knowledge-Driven Encode, Retrieve, Paraphrase for Medical Image Report Generation”. In: Proc of the AAAI Conf. on Artificial Intelligence. 2019, pp. 6666–6673.
[9] Y. Xue and X. Huang. “Improved Disease Classification in Chest X-Rays with Transferred Features from Report Generation”. In: IPMI. 2019, pp. 125–138.