1st Serbian International Conference on Applied Artificial Intelligence (SICAAI)
Kragujevac, Serbia, May 19-20, 2022

Baselines for automatic medical image reporting

Franco Alberto Cardillo

Institute for Computational Linguistics
National Research Council, Via G. Moruzzi 1, 56124 Pisa, Italy
Email: francoalberto.cardillo@ilc.cnr.it

Abstract

Despite the high number of deep learning models presented in the last few years for automatically annotating medical images, clear baselines to compare model upon are still missing. Furthermore, though there are only two datasets publicly available for the task, there is neither a shared and commonly adopted procedure for preprocessing the raw data nor an unanimous way in which the intermediate tasks have been defined. The work here presented tries to fill this gap by clearly characterizing the datasets, defining the learning task and providing some baselines that can be especially helpful when trying to replicate the results in languages with less resources than those available in English.
Key words: computer vision, natural language generation, image classification

1 Introduction

The a demand for image-based medical examinations has been rising for the past years and has nowadays become so high as to make it impossible for radiology departments to report on the acquired images in a timely manner[1]. That is a major problem since a short turnaround of written reports from radiologists to clinicians is a key factor from several points of view: it enables an early planning of a correct treatment, increasing the likelihood of healthier clinical courses, it reduces costs and, more in general, it improves the patient experience [2]. For the previous reasons several approaches for automatizing this important step are being proposed.

The task of automatic medical image reporting (AMIR) consists in the generation of a narrative text, expressed in natural language, describing the diagnostic content of one ore more medical images given as input data to a computer program. It is an inherently multi-modal task involving images and texts, whose solution requires a successful combination of computer vision (CV) and Natural Language Processing (NLP) algorithms [1, 2]. Any algorithm tackling the previous task can be roughly described as implemented with a pipeline of two computational steps: the first step processes the input image and maps its visual content onto a feature space accessed in the second step for generating a verbal description of an appropriate surface form, with correct lexicon and grammar, a task normally referred to as Natural Language Generation (NLG). The current state of the art performance is held by deep learning (DL) models [3], usually based on a convolutional neural network (CNN), acting as the visual encoder, and a recurrent neural network, acting as the decoder generating text. Recently, large language models (or surrogate smaller versions) have been introduced improving upon previous results results [4] However, as noted also in other works [2], this initial set of works

[1]Radiology Review, A national review of radiology reporting within the NHS in England, CareQuality Commission, July 2018. https://l.cnr.it/nhseng18

[2]American College of Radiology, Qualified Clinical Data Registry, January 2022. https://l.cnr.it/acrqdr22

is hardly comparable: baselines are not assessed, very complex systems are compared between each other without establishing how well they perform with respect to a baseline measurerement, they use image encoding steps not uniformly defined. It is not even clear whether or not images are needed to seed the generation of text [5]. Even if reviews list up to eight datasets available for the task [6], basically only two datasets are actually usable and, indeed, used. Moreover, The datasets have very different characteristics and their preprocessing is rarely described in full details. This status of the research area makes it very difficult to replicate the results, especially when trying to build an analogous system for languages other than English.

2 Approach

The work will try and provide sound criteria for assessing a baseline and comparing different approach, with a specific attention to choosing methods and pre-processing steps that are language-agnostic and that can be used also when dealing with non-English texts and with low-resource languages. More specifically:

- in all the available datasets the images are associated to multiple labels and are extremely imbalananced. Many works with important contributions either do not specify which labels are selected and how (e.g, [7]) or they seem to arbitrary thresholds that may have a deep impact on the image encoding accuracy. This work will try and select onbjective criteria for selecting labels and distributing in the training and test data partitions.

- In all the datasets the most frequent class is the *"normal"* one, i.e. most of the images do not show any suspicious area. This work will try and establish whether separating normal images from the rest might lead to better image encoding and text generation.

- Many works encode image labels using random vector embeddings or pretrained word-embeddings. However, a detailed study on which is the best choice is missing. This point will be included in the experimentations.

3 Aknowledgment

This work has been partially supported by the EU Horizon 2020 DeepHealth Project (GA No. 825111).

References

[1] R. Bernardi et al. Automatic Description Generation from Images: A Survey of Models, Datasets, and Evaluation Measures (Extended Abstract). In *Proc. of the 26th Intl. Joint Conf. on Artificial Ingelligence*, 2017.

[2] J. Pavlopoulos et al. A survey on biomedical image captioning. In *Proc. of the 2nd Workshop on Shortcomings in Vision and Language*, Minneapolis, Minnesota, 2019. Association for Computational Linguistics.

[3] M. M. A. Monshi et al. Deep learning in generating radiology reports: A survey. *Artificial Intelligence in Medicine*, 106, 2020.

[4] O. Alfarghaly et al. Automated radiology report generation using conditioned transformers. *Informatics in Medicine Unlocked*, 24, 2021.

[5] Zaheer Babar et al. Encoder-decoder models for chest X-ray report generation perform no better than unconditioned baselines. *PLoS ONE*, 16(11), November 2021.

[6] H. Ayesha et al. Automatic medical image interpretation: State of the art and future directions. *Pattern Recognition*, 114, 2021.

[7] Baoyu andothers Jing. On the automatic generation of medical imaging reports. *Proc. of the 56th Annual Meeting of the ACL*, 2017.