A multimodal approach to automated generation of radiology reports using contrastive learning

Giorgio Leonardi, Luigi Portinale and Andrea Santomauro

Computer Science Institute, DiSIT Università del Piemonte Orientale, Alessandria (Italy)

Abstract

In the present paper, we present a preliminary and ongoing work concerning the problem of generating a suitable diagnostic report from radiology (x-ray) images. The task is tackled with a contrastive learning approach, where radiology images and textual reports are embedded in a common space, with the goal of getting similar (close in space) embeddings between a given image and the corresponding report. Once the embeddings for an image and the corresponding textual report are generated (through suitable fine-tuned models), we propose to feed them to a contrastive learning engine in such a way that image and textual embeddings are pushed close in the embedding space if image and text are related, while they are moved away otherwise. Preliminary analysis shows promising results in terms of effectiveness of the contrastive learning approach, but also suggest relevant issues to be investigated such as the importance of context and the role of suitable encoder/decoder modules to properly deal with the textual generation.

Keywords

Multimodal machine learning, contrastive learning, automated radiology report generation

1. Introduction

Healthcare data are intrinsically multimodal [1]; for each patient, clinicians collect different types of information stored and, in the most of the cases, structured in a way we can directly use them in a machine learning task. For instance, the data generated by the triage procedures provide the most general information about patients (e.g. age, gender, blood pressure, etc.). In addition, more detailed information is collected from specific tests and exams (e.g. x-ray images, blood exam, etc.). This information is usually completed with written reports containing the conclusions drawn by the physicians.

These kinds of data are provided through different modalities: structured information such as EHR, time-series data as in ECGs, images in case of x-rays, CATs or PETs, and free text in case of diagnostic reports. Consequently, the interest in jointly exploiting different modalities for machine learning tasks in healthcare applications have been growing in the last years [2, 3], due to both the recognition of the superiority of multimodal learning when different modalities are available [4] and to the advancement of several techniques able to combine, in a common space, the embeddings obtained from the different modalities (e.g. images and text) [5, 6, 7]. In particular, the latest research focuses on two main categories of approaches

HC@AIxIA 2022: 1st AIxIA Workshop on Artificial Intelligence For Healthcare

andrea.santomauro@uniupo.it (A. Santomauro)

^{© 02022} Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

which are currently explored: attention-based approaches, where attention or self-attention mechanisms are exploited to relate images and text fragments [8, 9] and contrastive learning approaches [10, 6, 11].

The attention-based approach is usually adopted in image-text matching problems related to multimodal retrieval [12, 13] while, in the context of a task concerning text generation from images (e.g., image captioning), we claim that a contrastive approach should be more natural. In the present paper, we present a preliminary and ongoing work concerning the problem of generating a suitable diagnostic report from radiology images. The task is tackled with a contrastive learning approach, where radiology images and textual reports are embedded in a common space, with the goal of getting similar (close in space) embeddings between a given image and the corresponding report.

The general idea is to adopt a contrastive loss in such a way that, given a pair <image, text>, the corresponding embeddings are pushed close in the embedding space if image and text are related, while they are pushed away otherwise. Next sections will outlined the approach and report some preliminary results.

2. Contrastive Learning

Contrastive Learning (CL) is a machine learning methodology where unlabeled sample data are compared against each other, in order to teach a model which points are similar and which are different. One of the most prominent application of this paradigm is the siamese network model [14], where a pair of convolutional neural nets (CNN) are trained in order to learn a proper similarity function for pair of images (e.g., in order to recognize a person in a database of pictures of different people). Samples belonging to the same distribution are pushed towards each other in the embedding space. In contrast, those belonging to different distributions are moved away from each other.

In a CL setting, a way to obtain the desired goal is to train the model with the so called *contrastive loss*. Suppose that G_W is a parametric function (with set of parameters W) to be learned. Let X_1, X_2 be two input points, $D_W(X_1, X_2)$ a distance function (to be learned) between input points, and $y \in \{0, 1\}$ a binary label such that y = 0 if X_1, X_2 are deemed similar and y = 1 otherwise. Following LeCunn's original formulation [15], eq 1 shows the formula of the loss where: $(y, X_1, X_2)^i$ is the *i*-th labeled sample pair, P is the number of training pairs, and m > 0 is the so-called margin.

$$\mathcal{L}(W) = \sum_{i=1}^{P} L(W, (y, X_1, X_2)^i)$$

$$L(W, (y, X_1, X_2)^i) = (1 - y)\frac{1}{2}(D_W^i)^2 + y\frac{1}{2}\max(0, m - D_W^i)^2$$
(1)

For similar points (y = 0), the loss is given by their distance; on the contrary when two points are dissimilar (y = 1) they contribute to the loss function only if their distance is within the margin m. The main rational behind formula 1 is that we want to bring closer similar items, and move away, in the embedding space, dissimilar items. The margin hyperparameter avoids pushing away items that are already far enough.



Figure 1: Learning Architecture

In the present work, contrastive loss is used in a multimodal fashion. As mentioned above, we consider the problem of generating suitable (diagnostic) reports associated with a radiology image; in this preliminary work we just consider simple x-ray images. The general idea is the following:

- produce an image embedding using a suitable CNN with the x-ray image as input;
- produce a text embedding using a suitable deep encoder with a free text diagnostic report as input;
- learn a combined embedding model trained by means of contrastive loss, using as similar pairs x-ray images and their corresponding reports, and as dissimilar pairs x-ray images and unrelated reports

Next section will outline the proposed architecture.

3. Proposed Architecture

The learning part of the architecture we are proposing (for the training phase) is summarized in figure 1. Following the general scheme introduced in the previous section, to produce x-ray embeddings we adopted a particular CNN with the goal of extracting the most important hidden features of an x-ray image; in particular we considered a pretrained version of CheXNet, a 121-layer CNN trained on ChestX-ray14, one of the largest publicly available chest X-ray dataset [16]. We then stacked a simple fully-connected neural net (FCNN) to generate the embedding. The structure of such FCNN is reported in figure 2. Concerning text embeddings, we used HuggingFace's Encoder/Decoder model [17] pretrained on a bert2Bert task [18], and in particular the Encoder model for textual embedding generation.

Once all the embeddings are obtained, we standardize them to avoid big differences in value scale, and we generate a specific training set for the contrastive learning task (see next section). During this step, CheXNet's parameters are fine-tuned, while we freeze the EncoderDecoder's



Figure 2: FCNN for x-ray embedding generation





parameters; the aim is to map images' embedding in the textual embedding space, so we can exploit the decoder part to generate images related reports. Figure 3 shows the report generation part of the architecture. For generating the report of a given x-ray image, we submit the latter to the fine-tuned version of the CheXNet+FCNN to generate the suitable embedding that we assume to be close to the text corresponding to the diagnostic report. The embedding is then fed to the decoder and the corresponding report is finally generated.

4. Experiments and preliminary results

To perform an experimental analysis of the approach, we considered a public dataset available through the Open-i service of the National Library of Medicine, National Institutes of Health, (Bethesda, MD, USA), and containing coupled elements of x-ray images and reports [19]. Images in the dataset are grayscale DICOM images; a preprocessing step has been performed in order to assure that every image is of the same size. We also perfomed image normalization. Associated diagnostic reports are stored as XML files which contain different information; in this work we are interested in the *findings* part of the report, a free text information representing the diagnostic findings obtained by the radiologist after the inspection of the image. More than one image can be associated with a given report; for the aim of this work we consider them as independent images with the same diagnostic report. Figure 4 shows the histograms correspoding to the number of reports with a given number of associated images. We can notice



Figure 4: Images associated to reports

that the majority of the reports have two associated images, that are front X-ray and lateral X-ray. Some reports have no images associated, and we discarded them. We also discard images having an empty findings part in the corresponding report. We finally performed further sanity checks and text normalization, finally resulting in 3851 items represented as pairs <x-ray_image, finding_text>. We performed the embedding generation of every image and text in each item using the architecture of figure 1, and in particular the part upstream of the contrastive engine. We get 3851 pairs <x-ray_embedding, finding_embedding >; we split this dataset of embedding pairs into 80% for training and 20% for test. The training set has been used as input of the contrastive engine of figure 1 for the fine tuning of the model. Early stopping has been used to stop training, with a final validation loss of about $2e^{-07}$.

The experimental analysis on the test set has been used to check the following issues:

- Embeddings quality
- Generated text quality

Concerning embedding quality we measured the the distance between the generated report

embedding, and the original ones. In particular, given an x-ray image, we use the architecture of figure 3 to generate the corresponding predicted report; we then compared the predicted report with the original one by measuring the L2 distance between the embeddings. Table 1 reports the average normalized distance obtained on the test set, as well as some confidence intervals on such value The average quality of the embeddings is then quite promising, even if

AVG Distance	Std_dev	95% CI	99% CI
0.21045	0.13991	0.21045 ± 0.00988	0.21045 ± 0.013

Table 1

Embedding distances on test set

not completely satisfactory. We think that more effort should be devoted to the exploitation of multiple images associated to a given report, without considering them as independent.

Concerning the quality of generated reports, we found some problems with particular words that can be generated out of context, because of embedding values close to the ones of more suitable (in terms of context) words. This issue can be addressed in the encoder training which is based on BERT: BERT is trained on a corpus of general purpose text, and not in a specific field (e.g., radiology), thus our guess is that more specific contextual information should be provided. We plan to investigate more on this issue and by trying different approaches such as:

- Inserting prior knowledge
- Using another Encoder/Decoder architecture, specific for the Medical field (e.g., MedBert [20]).

5. Conclusions

We have presented a proposal for a multimodal learning architecture focusing on a contrastive learning approach to the automated generation of radiology reports from x-ray images. The work is very preliminary and the results obtained so far have evidenced some issues to focus on for future works, among which the exploitation of multiple images for the same report, the use of contextual knowledge in the textual encoder/decoder and the room for alternative architectural modules. In particular, we think that a closer comparison with the CLIP architecture [6], a multimodal architecture that shares several aspects with the issues discussed in the present work, will be really beneficial for instantiating these ideas in the specific task of diagnostic report generation from radiology images.

References

- Q. Cai, H. Wang, Z. Li, X. Liu, A survey on multimodal data-driven smart healthcare systems: Approaches and applications, IEEE Access 7 (2019) 133583–133599. doi:10.1109/ ACCESS.2019.2941419.
- [2] T. Syeda-Mahmood, K. Drechsler (Eds.), Multimodal Learning for Clinical Decision Support and Clinical Image-Based Procedures, volume 12445 of *Lecture Notes in Computer Science*, Springer, 2020.

- [3] A. Kline, H. Wang, Y. Li, S. Dennis, M. Hutch, Z. Xu, F. Wang, F. Cheng, Y. Luo, Multimodal machine learning in precision health, 2022. URL: https://arxiv.org/abs/2204.04777.
- [4] Y. Huang, C. Du, Z. Xue, X. Chen, H. Zhao, L. Huang, What makes multi-modal learning better than single (provably), in: M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, J. W. Vaughan (Eds.), Advances in Neural Information Processing Systems, volume 34, Curran Associates, Inc., 2021, pp. 10944–10956.
- [5] D. Francis, B. Huet, B. Merialdo, Embedding images and sentences in a common space with a recurrent capsule network, in: 2018 International Conference on Content-Based Multimedia Indexing (CBMI), 2018, pp. 1–6. doi:10.1109/CBMI.2018.8516480.
- [6] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, 2021. URL: https://arxiv.org/abs/2103.00020.
- [7] K. Li, Y. Zhang, K. Li, Y. Li, Y. Fu, Image-text embedding learning via visual and textual semantic reasoning, IEEE Transactions on Pattern Analysis and Machine Intelligence early access (2022). doi:10.1109/TPAMI.2022.3148470.
- [8] M. Popattia, M. Rafi, R. Qureshi, S. Nawaz, Guiding attention using partial-order relationships for image captioning, 2022. URL: https://arxiv.org/abs/2204.07476.
- [9] Y. Wu, S. Wang, G. Song, Q. Huang, Learning fragment self-attention embeddings for image-text matching, in: Proceedings of the 27th ACM International Conference on Multimedia, MM '19, 2019, p. 2088–2096.
- [10] C. Jia, Y. Yang, Y. Xia, Y. Chen, Z. Parekh, H. Pham, Q. V. Le, Y. Sung, Z. Li, T. Duerig, Scaling up visual and vision-language representation learning with noisy text supervision, 2021. URL: https://arxiv.org/abs/2102.05918.
- [11] J. Yang, C. Li, P. Zhang, B. Xiao, C. Liu, L. Yuan, J. Gao, Unified contrastive learning in image-text-label space, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19163–19173.
- [12] H. Nam, J.-W. Ha, J. Kim, Dual attention networks for multimodal reasoning and matching, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 299–307.
- [13] K.-H. Lee, X. Chen, G. Hua, H. Hu, X. He, Stacked cross attention for image-text matching, in: roceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 201–216.
- [14] Y. Taigman, M. Yang, M. Ranzato, L. Wolf, Deepface: Closing the gap to human-level performance in face verification, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR14, 2014, pp. 1701–1708.
- [15] R. Hadsell, S. Chopra, Y. LeCun, Dimensionality reduction by learning an invariant mapping, in: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), volume 2, 2006, pp. 1735–1742. doi:10.1109/CVPR.2006.100.
- [16] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Y. Ding, A. Bagul, C. P. Langlotz, K. S. Shpanskaya, M. P. Lungren, A. Y. Ng, Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning, 2017. URL: http://arxiv.org/abs/1711.05225.
- [17] T. Wolfe, al., Transformers: State-of-the-art natural language processing, ???? URL: https: //huggingface.co/docs/transformers/model_doc/encoder-decoder, last accessed 22 October 2022.
- [18] C. Chen, Y. Yin, L. Shang, X. Jiang, F. Wang, Z. Wang, X. Chen, Z. Liu, Q. Liu, bert2BERT:

Towards reusable pretrained language models, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 2022, pp. 2134–2148.

- [19] D. Demner-Fushman, M. D. Kohli, M. B. Rosenman, S. E. Shooshan, L. Rodriguez, S. Antani, G. R. Thoma, C. J. McDonald, Preparing a collection of radiology examinations for distribution and retrieval, J Am Med Inform Assoc. Mar;23(2) (2016) 304–10. doi:10.1093/ jamia/ocv080.
- [20] L. Rasmy, Y. Xiang, Z. Xie, C. Tao, D. Zhi, Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction., NPJ Digital Medicine 4 (2021).