Heliyon 10 (2024) e27516

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

XRaySwinGen: Automatic medical reporting for X-ray exams with multimodal model

Gilvan Veras Magalhães Junior [a,*], Roney L. de S. Santos [a], Luis H. S. Vogado [a], Anselmo Cardoso de Paiva [b], Pedro de Alcântara dos Santos Neto [a]

[a] Departamento de Computação, Universidade Federal do Piauí, Teresina, Brazil
[b] Núcleo de Computação Aplicada, Universidade Federal do Maranhão, São Luis, Brazil

ARTICLE INFO

Dataset link: http://openi.nlm.nih.gov/

Dataset link: https://nihcc.app.box.com/v/ChestXray-NIHCC

Keywords:
Multimodal
Transformers
Medical report
Computer vision
Natural language processing

ABSTRACT

The importance of radiology in modern medicine is acknowledged for its non-invasive diagnostic capabilities, yet the manual formulation of unstructured medical reports poses time constraints and error risks. This study addresses the common limitation of Artificial Intelligence applications in medical image captioning, which typically focus on classification problems, lacking detailed information about the patient's condition. Despite advancements in AI-generated medical reports that incorporate descriptive details from X-ray images, which are essential for comprehensive reports, the challenge persists. The proposed solution involves a multimodal model utilizing Computer Vision for image representation and Natural Language Processing for textual report generation. A notable contribution is the innovative use of the Swin Transformer as the image encoder, enabling hierarchical mapping and enhanced model perception without a surge in parameters or computational costs. The model incorporates GPT-2 as the textual decoder, integrating cross-attention layers and bilingual training with datasets in Portuguese PT-BR and English. Promising results are noted in the proposed database with ROUGE-L 0.748, METEOR 0.741, and NIH CHEST X-ray with ROUGE-L 0.404 and METEOR 0.393.

1. Introduction

1.1. Motivation

The application of Artificial Intelligence (AI) in image interpretation tasks has often been limited to the classification problems [1–4], which does not carry further details about the patient's current state. Nonetheless, using AI to do the automatic generation of medical reports [5–10], it is possible to create a descriptive text detailing aspects found in the X-ray such as location, size, magnitude, and other patterns, essential to generating high-quality report [11]. Despite the progress in automated generation of medical reports and the potential to improve patient care and reduce radiologist workload such as [12–16], the problem is still far from being resolved.

* Corresponding author.
E-mail addresses: gilvanveras@ufpi.edu.br (G. Veras Magalhães), roneysantos@ufpi.edu.br (R. L. de S. Santos), lhvogado@ufpi.edu.br (L. H. S. Vogado), paiva@nca.ufma.br (A. Cardoso de Paiva), pasn@ufpi.edu.br (P. de Alcântara dos Santos Neto).

https://doi.org/10.1016/j.heliyon.2024.e27516
Received 30 March 2023; Received in revised form 29 February 2024; Accepted 1 March 2024

Available online 12 March 2024

Radiology is crucial in modern medicine for its non-invasive imaging techniques that enable the diagnosis and monitoring of various medical conditions [17]. However, the formulation of unstructured medical reports necessitates a substantial time investment and is susceptible to errors, thereby introducing a potential hazard of critical medical oversights that could pose a threat to patient safety [18]. The field of Artificial Intelligence has made great progress in activities that depend on specialized knowledge [19–22], including specific tasks such as interpretation of medical images [23,24] and text generation [25,26].

1.2. Objective

The main objective of this work is the proposition of a multimodal model to automatically generate textual medical reports from X-ray images. To achieve this goal, it was applied Computer Vision (CV) techniques, such as normalization, scaling, and attribute extraction to represent X-ray images in a vector. Also, it was applied Natural Language Processing (NLP) techniques to the textual medical reports, such as textual normalization, removal of patient personal information, and unnecessary characters, and the textual tokenization, so that the report is understood in computational language.

1.3. Contribution

In this work, it is placed a significant contribution. The novel modeling of the Swin Transformer as the image encoder of the multimodal model, allows performing a hierarchical mapping of the image and, together with the window displacement approach, it is possible to enhance the perception of the model without increasing the number of parameters, the computational cost and even make it more efficient in terms of memory usage. In addition, the model has GPT-2 as the textual decoder by the addition of cross-attention layers connected to the image encoder that is subjected to training with datasets in two different languages: Portuguese (pt-br) and English.

2. Related works

2.1. Convolutional Neural Networks (CNN)

The automatic generation of medical reports started with [27], who proposed a CNN-RNN architecture to generate texts from images. As studies advanced in the area, the [28]'s attention layer was introduced in several experiments, and models such as [29], started to combine the usual architecture of CNN-RNN with attention mechanism to synthesize multi-view visual features based on a sentence-level in a late fusion fashion.

Convolutional Neural Networks have played an essential role in medical image analysis for many years [30,13,27,14,31–38]. However, despite having an architecture that is easy to understand and implement, CNNs use convolution, a local operation limited to a small neighborhood of an image, while Transformers use a mechanism called self-attention, a global operation [28].

2.2. Transformers

Transformer architecture makes it possible to extract information from the entire image and effectively capture relationships between distant points. In addition, several state-of-the-art results were obtained in different CV and NLP tasks. Therefore, the multimodal model proposed in this work comprises an image encoder and a decoder based on the Transformers architecture.

Initial approaches using transformers for the problem of automatic generation of medical reports for radiological examinations retained the CNN architecture for the image encoder and introduced transformers as a decoder to generate textual medical image reports as can be seen in [5–8].

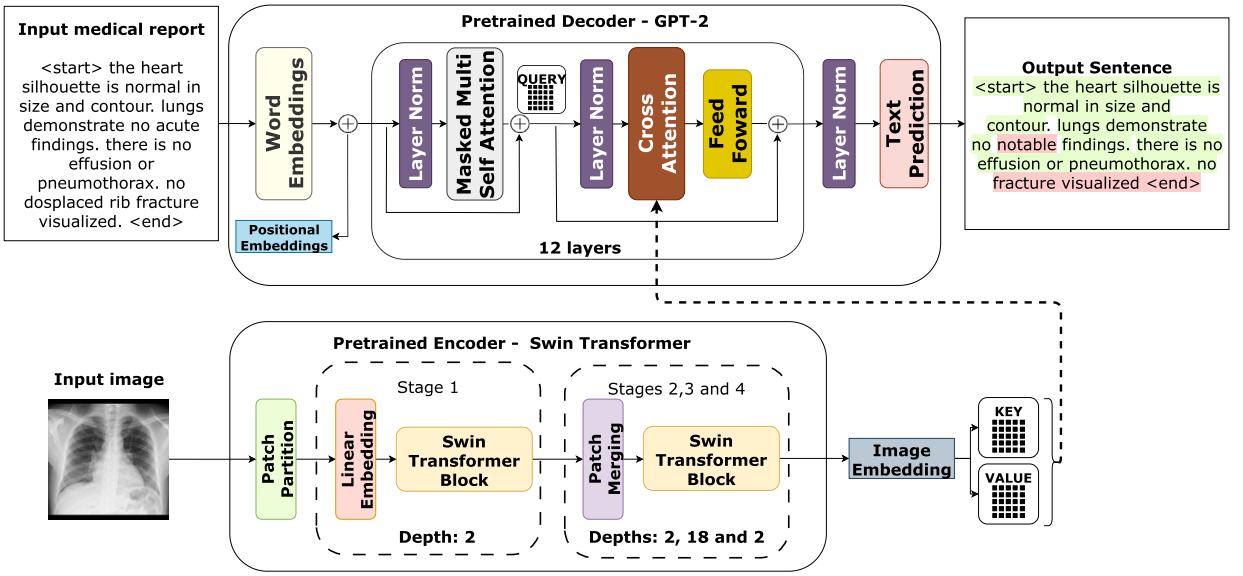
New architectures based on transformers emerged, including vision transformers [39], which enabled the emergence of new works. For example, the work developed by [40] or [9], explored the effect of loss of spatial bias information in the encoder using a pre-trained vision transformer architecture combined with different pre-trained language transformers as the decoder.

3. Methods

3.1. XRaySwinGen

XRaySwinGen is a modified transformer [28] architecture that performs the X-ray image analysis and generates medical reports automatically. In other words, it is an End-to-End Sequence to Sequence embedding task where Image pixels are input sequences and the medical report describing the image is the desired output. This approach was strongly influenced by state-of-the-art [12,41,42] and the Liu et al. paper [43] that had notable advances in adapting transformer from language to vision arising from differences between the two domains, such as large variations in the scale of visual entities and the high resolution of pixels in images compared to words in text.

In this work, the extraction of features from X-ray images is done with the hierarchical mapping of features and the Shifted Window [43] to capture both local and global dependencies effectively, an evolution of the base ViT architecture [44]. The backbone Swin Transformer [43] was pre-trained with the imagenet-1k database [45], in addition, the original patch size 4x4 and the configuration Swin-L were followed. The medical report generation is performed by a trained version of GPT-2 with 124 million parameters



Source: Adapted from [43] and [46]

Fig. 1. XRaySwinGen proposed multimodal model architecture.

Table 1
Summary of used X-ray image databases.

Database	Train	Validation	Test	Total	Reference
Proposed database	17,537	924	3,509	21,970	-
IU X-Ray	3,000	452	342	3,794	[47]
NIH Chest X-ray	82,197	25,596	4327	112,120	[48]

and vocabulary size 50257, which was made available to researchers by OpenAI[1] [46]. When the instances of the pre-trained encoder and decoder models are loaded, their embeddings are tied together using a cross-attention layer. Moreover, the model contains 12 cross-attention layers.

The encoder initially splits an input RGB image into non-overlapping patches, where each patch is treated as a "token" and its feature is defined from the concatenation of the raw values of pixels. The features dimension of each patch is $4 \times 4 \times 3 = 48$, where 3 refers to the amount of image channels. To enhance the perception of the model, the Shift Window operation is performed. In it, there is a displacement in the sequence of patches by a certain number of positions allowing the model to capture global and local features of the image, being able to process images at different levels of granularity, which is not possible with traditional CNNs. The encoder embeddings are used as KEY and VALUE, while on the decoder side, the embeddings are used as QUERY in the cross-attention head.

During training, both the X-ray image and the medical report are the model inputs. By using these inputs, the model is targeted to generate the same caption, leading to an understanding of the correlation between words in medical reports and X-rays in the input images. Throughout each training iteration, the model generates a sequence of words as a prediction. This predicted sequence is then compared to the actual caption, and the resulting difference, known as the loss, is fed back into the model. This process enables the model to learn and improve its ability to generate more accurate reports over time. The architecture described is presented in Fig. 1.

3.2. Training details

The XRaySwinGen model is trained with 3 different databases. The first database was obtained from private Brazilian Hospitals and has a total of 21,970 images, where their resolutions range from 727x692 to 4892x4020 pixels in DICOM format and the medical reports are in Portuguese PT-BR. The second database was proposed by [47] and collected at the Indiana University Hospital in the United States. The third database, which was proposed by [48] has 112,120 frontal chest X-ray images, belonging to 30,805 unique patients and it was called NIH Chest X-ray. The X-ray images from all the databases were converted to 224x224 resolution to fit the model input. In Table 1, the databases used are presented with the number of sample information, as well as the proportion used

[1] https://openai.com/blog/better-language-models/.

Table 2
Results obtained by the XRaySwinGen model.

MODEL	DATA	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR	SPICE
Ours	Proposed Dataset	0.731	0.691	0.672	0.661	0.748	0.741	0.659
Ours	IU X-ray	0.377	0.239	0.168	0.124	0.300	**0.322**	0.202
Ours	NIH Chest X-ray	0.396	0.329	-	-	0.404	0.393	0.170
Chen et al.	IU X-ray	**0.470**	**0.304**	**0.219**	**0.165**	**0.371**	0.187	-

for training, validation, and testing. The models are trained with a batch size of 6 for 25 epochs using the Adam optimizer. Also, the encoder and decoder embedding dimensions are 192 and 768, respectively. The experiments were performed in a Google Colab[2] programming environment with an Intel Xeon CPU @2.20 GHz, 13 GB of RAM, and a Tesla graphics card. T4 16 GB. The FP16 precision was used in training as it is commonly used in modern hardware accelerators, such as graphics processing units (GPUs) and tensor processing units (TPUs), to accelerate calculations in deep learning [49,50].

3.3. Evaluation metrics

To evaluate the quality of the medical reports generated by the models, the Bilingual Evaluation Understudy (BLEU) [51], ROUGE-L [52], METEOR [53] and SPICE [54] scores were calculated. These metrics are widely covered in works within the field [35,11,22], enabling the comparison and tracking of the evolution of works over the years.

BLEU is an evaluation metric commonly used for NLP tasks, particularly machine translation. However, we will use in this work to evaluate the text generation problem. It measures the quality of text predicted by the model compared to the reference text by comparing the n-gram overlaps between the two. The more overlap between the predicted text and the reference text, the higher the BLEU score. BLEU-1 evaluates word by word individually, BLEU-2 evaluates combinations of two by two words, BLEU-3 evaluates grouping by three words, and BLEU-4 evaluates combinations by four words; thus the evaluation is stricter.

This metric is calculated following Equation (1) below:

$$BLEU = BP \cdot exp\left(\sum_{n=1}^{N} w_n \log p_n\right), \tag{1}$$

4. Results

According to the results achieved by our models and listed in Table 2, the metrics are aligned with published work in the research area. Regarding the model trained with the Proposed Dataset, which is a PT-BR dataset, even with a great diversity of natural textual medical reports present in the dataset, the model achieves the highest scores. Standardization in patient care and completion of medical reports certainly enabled these results to be achieved.

About the model trained on IU X-ray data, it is crucial to note that while our model yields lower results than Chen et al. [12] model in BLEU[1-4] and ROUGE-L, it outperforms the METEOR score. In other words, our model demonstrates higher accuracy in placing tokens within sentences. In addition, there is a mismatch in the test dataset size between our work and Chen's, as not all the images for IU X-ray have an associated medical report.

Finally, there are 15 distinct classes in the NIH Chest X-ray database (see Table 3), which were converted into text to compose an adapted medical report. It is worth noting that on this basis, the medical reports are much smaller, mainly composed of one or two words only, so, when checking the metrics it is possible to see that the BLEU[3-4] values were affected, thus assuming lower values. However, the BLEU[1-2], ROUGE-L, and METEOR reached values of 0.396, 0.329, 0.404, and 0.393 respectively, suggesting that the model started the process of generalization and relationship of the image with the texts. As this database has the most significant number of training samples, submitting it to a new training session with a new parameter adjustment and a more significant number of epochs can achieve better results.

5. Discussion

This work proposed a multimodal model to automatically generate textual medical reports from X-ray images. We were the first to include the Swin Transformer architecture as an image encoder in the task of automatically generating medical reports from radiological exams to maximize the capture of image details, bearing in mind that any pixel can compromise the final result. In addition, the work is hindered by challenges such as the scarcity of detailed medical reports in natural language and the prevalent focus on image classification rather than comprehensive textual information in existing databases.

[2] https://colab.research.google.com/.

Table 3
Mapping of NIH Chest X-ray classes.

Numeric class	Text
0	No Finding
1	Atelectasis
2	Cardiomegaly
3	Effusion
4	Infiltration
5	Mass
6	Nodule
7	Pneumonia
8	Pneumothorax
9	Consolidation
10	Edema
11	Emphysema
12	Fibrosis
13	Pleural_Thickening
14	Hernia

To carry out experiments and validate our work, we used three databases, two of which are public and in English (IU X-ray and NIH Chest X-ray) and one was obtained from Brazilian private hospitals and is in Portuguese PT-BR (Proposed Dataset). The results point to some important conclusions: The model can quickly learn to generate text in a new language even if it is not the same as your decoder; The model can generate reports for a wide variety of diagnoses with good accuracy; Compared to the state-of-the-art model by Chen et al. [12] the model appears promising with room for parameter adjustment, data augmentation and fine-tuning.

The proposed dataset is composed of 40% abnormal and 60% normal high-definition images, and the number of words in their medical reports varies between 30 and 120. In general, medical reports are very detailed and objective with observations or descriptions separated by a period. The model trained with this dataset achieved the highest results in all the evaluation metrics. High BLEU[1-4] scores indicate a significant degree of similarity between the generated medical report and the human-generated reference. These BLEU results are corroborated by a high ROUGE-L score, indicating a substantial overlap in long-range content between the model output and the reference. Lastly, the results achieved in METEOR and SPICE scores indicate that the model is performing well in terms of both fluency and semantic similarity with reference, in other words, the model is generating coherent and contextually relevant text.

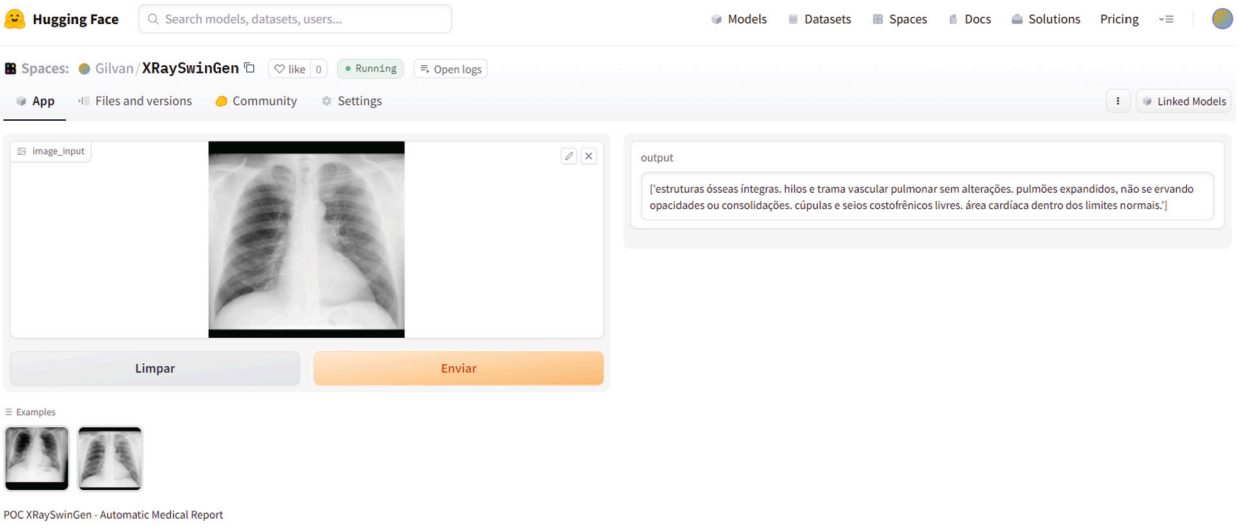
The model trained with the IU X-ray dataset results in lower BLEU[1-4] and ROUGE-L scores than the model of Chen et al., which might be aggravated by the sample distribution. During the processing of the dataset, it was filtered frontal images and some images did not match a report and vice versa which resulted in the numbers presented in Table 1. Despite the reduced dataset, it was possible to achieve close results on the metrics mentioned before and surpass Chen et al. [12] on METEOR. The SPICE score was not presented in their work, but our result of 0.202 indicates that the agreement between the model generation and the human reference is not good in terms of semantic content. A future modification in training this model will be to acquire the updated database and use the frontal and lateral X-ray images.

The model trained with NIH Chest X-ray yields a higher BLEU[1-2] score than the model trained with IU X-ray. BLEU[3-4] was not calculated because the reports consist of a maximum of two words here. Even though it was a classification base, the aim was to use its size to understand how the model would behave with this amount of data. The model successfully learned the predefined classes in the database, as evidenced by the results of the BLEU[1-2], ROUGE-L, and METEOR metrics. However, since it is a generative model with a limited number of classes, the model frequently hallucinates and produces sentences slightly longer than the expected class; therefore, the SPICE score is low.

As a side result of the research, an API/webpage was created to receive an X-ray image, returning the textual information about the patient's diagnosis. The proof of concept API/webpage is currently hosted at HuggingFace.[3] Fig. 2 shows the main window of the running API/webpage.

As part of future work, the intention is to adjust the parameters of the proposed model to enhance performance, achieve improved results, and reduce the utilization of computational resources. For instance, modifications may include training the models with different configurations such as Swin-T, Swin-S, or Swin-B depths, varying patch size, window size, and embedding size. New experiments are planned with both the databases employed in this study and the exploration of additional databases like MIMIC-CXR-JPG [15], ROCO [55], PEIR Gross [13], and ImageCLEF [56]. Another suggested avenue for future exploration involves conducting experiments that involve the combination of all the databases. Additionally, there is a proposal to refine image captioning metrics with a particular emphasis on human evaluation, ensuring nuanced aspects such as contextual relevance and visual semantics are adequately captured, surpassing existing standards like CIDEr or ChexBert. Finally, the plan includes the incorporation of new data learning techniques, such as Contrastive Learning [57], and Data Augmentation [58], among others.

[3] https://huggingface.co/spaces/Gilvan/XRaySwinGen.



Source: author's collection

Fig. 2. API/webpage created to use XRaySwinGen.

CRediT authorship contribution statement

Gilvan Veras Magalhães: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Roney L. de S. Santos:** Writing – review & editing, Visualization, Validation, Supervision, Investigation, Formal analysis. **Luis H. S. Vogado:** Validation, Resources, Data curation. **Anselmo Cardoso de Paiva:** Writing – review & editing, Validation, Supervision, Methodology, Investigation, Formal analysis. **Pedro de Alcântara dos Santos Neto:** Writing – review & editing, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Available at the links below:

- IU X-ray: http://openi.nlm.nih.gov/
- NIH Chest X-ray: https://nihcc.app.box.com/v/ChestXray-NIHCC

Acknowledgements

This work was carried out with the support of the Coordination for the Improvement of Higher Education Personnel - Brazil (CAPES) - Financing Code 001, Maranhão Research Support Foundation (FAPEMA), National Council for Scientific and Technological Development (CNPq) and Brazilian Company of Hospital Services (Ebserh) Brazil (Proc. 409593/2021-4). The results presented in this paper were obtained through research carried out at the "CENTRO DE REFERÊNCIA EM INTELIGÊNCIA ARTIFICIAL - CEREIA", based at the Federal University of Ceará in partnership with the Hapvida NotreDame Intermédica Group and supported by the Grant 2020/09706-7, São Paulo Research Foundation (FAPESP).

References

[1] S. Kim, B. Rim, S. Choi, A. Lee, S. Min, M. Hong, Deep learning in multi-class lung diseases' classification on chest X-ray images, Diagnostics 12 (4) (2022) 915.
[2] S.Z.Y. Zaidi, M.U. Akram, A. Jameel, N.S. Alghamdi, A deep learning approach for the classification of TB from NIH CXR dataset, IET Image Process. 16 (3) (2022) 787–796.
[3] Y. Tian, J. Wang, W. Yang, J. Wang, D. Qian, Deep multi-instance transfer learning for pneumothorax classification in chest X-ray images, Med. Phys. 49 (1) (2022) 231–243.
[4] M. Nawaz, T. Nazir, J. Baili, M.A. Khan, Y.J. Kim, J.-H. Cha, CXray-EffDet: chest disease detection and classification from X-ray images using the efficientdet model, Diagnostics 13 (2) (2023) 248.

[5] Y. Xiong, B. Du, P. Yan, Reinforced transformer for medical image captioning, in: Machine Learning in Medical Imaging: 10th International Workshop, MLMI 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13, 2019, Proceedings, Springer-Verlag, Berlin, Heidelberg, 2019, pp. 673–680.
[6] F. Liu, X. Wu, S. Ge, W. Fan, Y. Zou, Exploring and distilling posterior and prior knowledge for radiology report generation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 13753–13762.
[7] D. You, F. Liu, S. Ge, X. Xie, J. Zhang, X. Wu, Aligntransformer: hierarchical alignment of visual regions and disease tags for medical report generation, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24, 2021, pp. 72–82.
[8] O. Alfarghaly, R. Khaled, A. Elkorany, M. Helal, A. Fahmy, Automated radiology report generation using conditioned transformers, Inform. Med. Unlocked 24 (2021) 100557.
[9] M.M. Mohsan, M.U. Akram, G. Rasool, N.S. Alghamdi, M.A.A. Baqai, M. Abbas, Vision transformer and language model based radiology report generation, IEEE Access 11 (2022) 1814–1824.
[10] F. Yu, M. Endo, R. Krishnan, I. Pan, A. Tsai, E.P. Reis, E.K.U.N. Fonseca, H.M.H. Lee, Z.S.H. Abad, A.Y. Ng, et al., Evaluating progress in automatic chest X-ray radiology report generation, Patterns 4 (9) (2023).
[11] E. Çallı, E. Sogancioglu, B. van Ginneken, K.G. van Leeuwen, K. Murphy, Deep learning for chest X-ray analysis: a survey, Med. Image Anal. 72 (2021) 102125.
[12] Z. Chen, Y. Song, T.-H. Chang, X. Wan, Generating radiology reports via memory-driven transformer, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, 2020, pp. 1439–1449, https://aclanthology.org/2020.emnlp-main.112.
[13] B. Jing, P. Xie, E. Xing, On the automatic generation of medical imaging reports, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 2577–2586.
[14] C.Y. Li, X. Liang, Z. Hu, E.P. Xing, Hybrid retrieval-generation reinforced agent for medical image report generation, in: Proceedings of the 32nd International Conference on Neural Information Processing Systems, Curran Associates Inc., 2018, pp. 1537–1547.
[15] A.E.W. Johnson, T.J. Pollard, S.J. Berkowitz, N.R. Greenbaum, M.P. Lungren, C. Deng, R.G. Mark, S. Horng, MIMIC-CXR: a large publicly available database of labeled chest radiographs, CoRR, arXiv:1901.07042, 2019.
[16] B. Jing, Z. Wang, E. Xing, Show, describe and conclude: on exploiting the structure information of chest X-ray reports, arXiv preprint, arXiv:2004.12274, 2020.
[17] S.K. Zhou, D. Rueckert, G. Fichtinger, Handbook of Medical Image Computing and Computer Assisted Intervention, Academic Press, 2019.
[18] G. Zhao, Z. Zhao, W. Gong, F. Li, Radiology report generation with medical knowledge and multilevel image-report alignment: a new method and its verification, Artif. Intell. Med. 146 (2023) 102714.
[19] R.J. Woodman, A.A. Mangoni, A comprehensive review of machine learning algorithms and their application in geriatric medicine: present and future, Aging Clin. Exp. Res. (2023) 1–35.
[20] M. Rana, M. Bhushan, Machine learning and deep learning approach for medical image analysis: diagnosis to detection, Multimed. Tools Appl. 82 (17) (2023) 26731–26769.
[21] T.W. Cenggoro, B. Pardamean, et al., A systematic literature review of machine learning application in COVID-19 medical image classification, Proc. Comput. Sci. 216 (2023) 749–756.
[22] T. Pang, P. Li, L. Zhao, A survey on automatic generation of medical imaging reports based on deep learning, Biomed. Eng. Online 22 (1) (2023) 1–16.
[23] P. Kaur, S. Harnal, R. Tiwari, F.S. Alharithi, A.H. Almulihi, I.D. Noya, N. Goyal, A hybrid convolutional neural network model for diagnosis of COVID-19 using chest X-ray images, Int. J. Environ. Res. Public Health 18 (22) (2021) 12191.
[24] M. Yang, H. Tanaka, T. Ishida, Performance improvement in multi-label thoracic abnormality classification of chest X-rays with noisy labels, Int. J. Comput. Assisted Radiol. Surg. 18 (1) (2023) 181–189.
[25] N. Fatima, A.S. Imran, Z. Kastrati, S.M. Daudpota, A. Soomro, A systematic literature review on text generation using deep neural network models, IEEE Access 10 (2022) 53490–53503, https://doi.org/10.1109/ACCESS.2022.3174108.
[26] S. Biswas, ChatGPT and the future of medical writing, 2023.
[27] I. Allaouzi, M. Ben Ahmed, B. Benamrou, M. Ouardouz, Automatic caption generation for medical images, in: Proceedings of the 3rd International Conference on Smart City Applications, Association for Computing Machinery, New York, NY, USA, 2018, pp. 1–6.
[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).
[29] J. Yuan, H. Liao, R. Luo, J. Luo, Automatic radiology report generation based on multi-view image fusion and medical concept enrichment, 2019.
[30] G. Litjens, T. Kooi, B.E. Bejnordi, A.A.A. Setio, F. Ciompi, M. Ghafoorian, J.A. van der Laak, B. van Ginneken, C.I. Sánchez, A survey on deep learning in medical image analysis, Med. Image Anal. 42 (2017).
[31] X. Wang, Y. Peng, L. Lu, Z. Lu, R.M. Summers, TieNet: Text-image embedding network for common thorax disease classification and reporting in chest X-rays, 2018.
[32] X. Huang, F. Yan, W. Xu, M. Li, Multi-attention and incorporating background information model for chest X-ray image report generation, IEEE Access 7 (2019) 154808–154817.
[33] B. Jing, Z. Wang, E. Xing, Show, describe and conclude: on exploiting the structure information of chest X-ray reports, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2019, pp. 6570–6580.
[34] G.O. Gajbhiye, A.V. Nandedkar, I. Faye, Automatic report generation for chest X-ray images: a multilevel multi-attention approach, in: N. Nain, S.K. Vipparthi, B. Raman (Eds.), Computer Vision and Image Processing, Springer, Singapore, 2020, pp. 174–182.
[35] M.M.A. Monshi, J. Poon, V. Chung, Deep learning in generating radiology reports: a survey, Artif. Intell. Med. 106 (2020) 101878.
[36] B. Pandey, D. Kumar Pandey, B. Pratap Mishra, W. Rhmann, A comprehensive survey of deep learning in the field of medical imaging and medical natural language processing: challenges and research directions, J. King Saud Univ, Comput. Inf. Sci. 34 (8, Part A) (2022) 5083–5099.
[37] S. Kumar, M.K. Chaube, S.H. Alsamhi, S.K. Gupta, M. Guizani, R. Gravina, G. Fortino, A novel multimodal fusion framework for early diagnosis and accurate classification of COVID-19 patients using X-ray images and speech signal processing techniques, Comput. Methods Programs Biomed. 226 (2022) 107109.
[38] S. Kumar, S.K. Gupta, V. Kumar, M. Kumar, M.K. Chaube, N.S. Naik, Ensemble multimodal deep learning for early diagnosis and accurate classification of COVID-19, Comput. Electr. Eng. 103 (2022) 108396.
[39] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: transformers for image recognition at scale, arXiv preprint, arXiv:2010.11929, 2020.
[40] H. Lee, H. Cho, J. Park, J. Chae, J. Kim, Cross encoder-decoder transformer with global-local visual extractor for medical image captioning, Sensors 22 (4) (2022) 1429.
[41] Y. Miura, Y. Zhang, E. Tsai, C. Langlotz, D. Jurafsky, Improving factual completeness and consistency of image-to-text radiology report generation, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, 2021, pp. 5288–5304, https://aclanthology.org/2021.naacl-main.416.
[42] H. Nguyen, D. Nie, T. Badamdorj, Y. Liu, Y. Zhu, J. Truong, L. Cheng, Automated generation of accurate & fluent medical X-ray reports, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2021, pp. 3552–3569, https://aclanthology.org/2021.emnlp-main.288.
[43] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012–10022.

[44] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, 2020.
[45] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: a large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 248–255.
[46] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are unsupervised multitask learners, OpenAI Blog 1 (8) (2019) 9.
[47] 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. 23 (2) (2016) 304–310.
[48] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, R.M. Summers, ChestX-ray8: hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017, pp. 2097–2106.
[49] N.-M. Ho, W.-F. Wong, Exploiting half precision arithmetic in Nvidia GPUs, in: 2017 IEEE High Performance Extreme Computing Conference (HPEC), IEEE, 2017, pp. 1–7.
[50] S. Markidis, S.W. Der Chien, E. Laure, I.B. Peng, J.S. Vetter, Nvidia tensor core programmability, performance & precision, in: 2018 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW), IEEE, 2018, pp. 522–531.
[51] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, BLEU: a method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 2002, pp. 311–318.
[52] C.-Y. Lin, ROUGE: a package for automatic evaluation of summaries, in: Text Summarization Branches Out, Association for Computational Linguistics, 2004, pp. 74–81, https://aclanthology.org/W04-1013.
[53] A. Lavie, A. Agarwal, Meteor: an automatic metric for MT evaluation with high levels of correlation with human judgments, in: Proceedings of the Second Workshop on Statistical Machine Translation, Association for Computational Linguistics, 2007, pp. 228–231.
[54] P. Anderson, B. Fernando, M. Johnson, S. Gould, Spice: semantic propositional image caption evaluation, in: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, the Netherlands, October 11-14, 2016, Proceedings, Part V 14, Springer, 2016, pp. 382–398.
[55] O. Pelka, S. Koitka, J. Rückert, F. Nensa, C.M. Friedrich, Radiology objects in context (ROCO): a multimodal image dataset, in: Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis: 7th Joint International Workshop, CVII-STENT 2018 and Third International Workshop, LABELS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 3, Springer, 2018, pp. 180–189.
[56] A. García Seco de Herrera, C. Eickhof, V. Andrearczyk, H. Müller, Overview of the ImageCLEF 2018 caption prediction tasks, in: Working Notes of CLEF 2018-Conference and Labs of the Evaluation Forum (CLEF 2018), Avignon, France, September 10-14, 2018, in: CEUR Workshop Proceedings, vol. 2125, 2018.
[57] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, A. Joulin, Unsupervised learning of visual features by contrasting cluster assignments, Adv. Neural Inf. Process. Syst. 33 (2020) 9912–9924.
[58] C. Shorten, T.M. Khoshgoftaar, A survey on image data augmentation for deep learning, J. Big Data 6 (1) (2019) 1–48.