Medical Report Generation from Medical Images Using Vision Transformer and Bart Deep Learning Architectures

Murat Ucan[1[0000-0001-9219-2262]], Buket Kaya[2[0000-0001-9505-181X]], Mehmet Kaya [3[0000-0003-2995-8282]] and Reda Alhajj[4[0000-0001-6657-9738]]

[1] Department of Computer Technologies, Dicle University, Diyarbakır, Turkey
murat.ucan@dicle.edu.tr
[2] Department of Electronics and Automation, Firat University, Elazığ, Turkey
bkaya@firat.edu.tr
[3] Department of Computer Engineering, Firat University, Elazığ, Turkey
kaya@firat.edu.tr
[4] Department of Computer Science, University of Calgary, Calgary, AB, Canada
alhajj@cpsc.ucalgary.ca

Abstract. Generating medical reports from medical images using traditional methods is a time-consuming process that is prone to human error and requires experience. Failure to generate fast reports from medical images delays the treatment of patients, and misdiagnosis can lead to adverse conditions that can cause the death of patients. The main objective of this study is to develop a high-performance deep learning model that can autonomously generate medical reports from medical images. The proposed model consists of a Vision Transformer (ViT) encoder and a Bidirectional Autoregressive Transformer (BART) decoder. Training and testing on the model was conducted using images and reports from the Indiana University Chest X-Ray dataset. The developed model is analyzed with measurable parameters and then compared with its competitors in the literature using the same dataset. The proposed Vi-Ba architecture achieved success scores of 0.150, 0.154, 0.274 in bleu-4, meteor and rouge word matching evaluation metrics, respectively. The Vi-Ba model achieved high reporting performance compared to the studies reviewed in the literature. The results show that the proposed architecture can be used by specialized doctors in hospitals to diagnose diseases faster and more accurately. In this way, misdiagnosis and treatments will be reduced and human life will be protected.

Keywords: Deep Learning, Vision Transformer, ViT, Bidirectional Autoregressive Transformer, BART, Medical Report Generation, Chest X-rays

1 Introduction

Medical images produced using medical imaging technologies are used by specialized doctors for the detection of many diseases. However, the interpretation of medical images is a difficult, time-consuming and error-prone process. Interpretation of medical images can only be done by specialized doctors with many years of experience, so it is difficult. Doctors need to observe and reflect on each image and turn it into a medical text, so it is time-consuming. Doctors have long working hours, human factors and the complexity of the images, so it is an error-prone field of study. In order to solve these problems, the main problem of this study is to develop an autonomous system that can support doctors in making decisions in the diagnosis of diseases.

In underdeveloped countries where there are not enough specialized doctors, or in local hospitals in rural areas, medical imaging equipment is available, but unfortunately it is not always possible to interpret the medical images produced. However, early diagnosis is a very important factor in the cure of diseases. The autonomous reporting architecture proposed in this paper can support first level physicians for preliminary diagnosis and initiation of treatment. In addition, in a hospital where specialized doctors are available, the doctors' use of the decision support mechanism proposed in the study can reduce misdiagnosis and start treatment. In this context, the proposed study is important for faster and more accurate treatment initiation.

Many studies that can support doctors in diagnosing diseases using medical images are being carried out by researchers. Studies that try to produce single-word outputs from images of diseases are classification studies. Basically, determining which disease class an image belongs to is carried out with these studies. Ucan et al. [1] studied an 8-class dataset using gastroenterologic images. The researchers achieved a validation success of 0.935 in their study, which produced single-word classification results. Butun et al. [2] worked on lymph node and classified the images with two classes. The researchers achieved an accuracy of 0.986 as a result of their study. There are many similar classification-based studies on different medical images in the literature.

Another artificial intelligence-based field of study that can support doctors in making decisions in the diagnosis of diseases using medical images is segmentation studies. Segmentation can be defined as painting the diseased areas a different color and separating them from other surfaces for medical image studies. These areas are examined at the time of diagnosis and faster and more accurate decisions can be made. Cinar et al. [3] used brain MRI images to stain brain tumors in different colors. Helaly et al. [4] worked on the detection of Alzheimer's disease from MR images. The researchers designed a system that can guide specialist doctors in the detection of Alzheimer's disease, which is difficult to detect and significantly affects the quality of life.

Classification and segmentation studies are two important fields of study to support doctors in the detection of diseases. However, classification studies produce single-word results, while segmentation studies only involve staining the diseased surface. Although these fields of study are important, they are very lacking in detail. The diseased area needs to be identified with medical reports and doctors need to be given more descriptive information about the treatment [5]. For this reason, the main purpose of this study is to create a medical report consisting of a large number of words and sentences from images.

There are also studies involving the generation of paragraph-level medical texts using medical images. Singh et al. [6] focused on chest X-ray images in their study. They used CNN for feature extraction and LSTM architectures for autonomous medical report writing. The researchers achieved 0.374 success in the blue-1 evaluation metric. Another study involving medical report generation from chest X-ray images was conducted by Wang et al. [7]. They proposed to use an encoder-decoder architecture in their work called TieNet. They used popular CNN-based deep learning architectures in the encoder part and multilayer LSTM architecture in the decoder part. With their proposed reporting architecture, they achieved a bleu-1 evaluation metric score of 0.2860.

Another study using the LSTM architecture as a decoder was conducted by Harzig et al. [8] conducted another study using LSTM architecture as a decoder. They used ResNet-152 architecture for feature extraction from images and dual word LSTM architecture for report writing. The researchers achieved a bleu-1 evaluation metric score of 0.373 with the architecture they named HLSTM+att+Dual in medical report writing from chest X-ray images. Another study using transformer architectures in autonomous medical reporting is Alfarghaly et al. [9] in autonomous medical reporting. They used the DenseNet-based Chexnet [10] architecture for feature extraction from images. They used transformer-based GPT architecture for writing medical reports. The researchers achieved a bleu-1 evaluation metric score of 0.387 in their architecture called CDGPT2.

This study aims to autonomously write medical reports for 14 different diseases that can be detected from chest X-ray images using deep learning architectures. Encoder - decoder architecture is used in the study. Vision transformer architecture is preferred on the encoder side of the autonomous report generation model and bart model is preferred on the decoder side. Experiments on the deep learning architecture were performed using the Indiana University Chest X-Ray dataset collection. The main contributions of the study to the literature are given below.

- A novel and high-performance encoder-decoder architecture called Vi-Ba is proposed for autonomous medical report generation using deep learning methods.
- The developed architecture can be used in hospitals to support doctors in decision making. In this way, wrong diagnoses and treatments can be reduced.
- The autonomous reporting system will be much faster than observational reporting with traditional methods. This will contribute to early treatment initiation and positively affect recovery processes.
- In medical reports, each doctor writes reports with his/her own choice of words, sentences and reporting style. There is no specific standard between reports. The autonomous architecture developed will contribute to the creation of standard reports by creating a standard report template.

2 Methodology

2.1 Dataset and Implementation Details

Indiana University Chest X-Ray dataset was used in the training, validation and testing parts of our deep learning architecture for medical report generation from chest X-ray images [11]. The dataset contains chest X-ray images and reports. The three chest X-ray images were randomly selected from the dataset and the reports associated with the images are given in Figure 1.

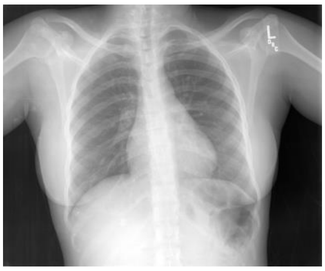
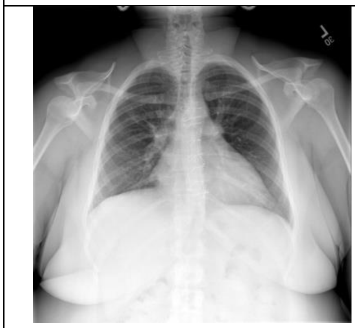
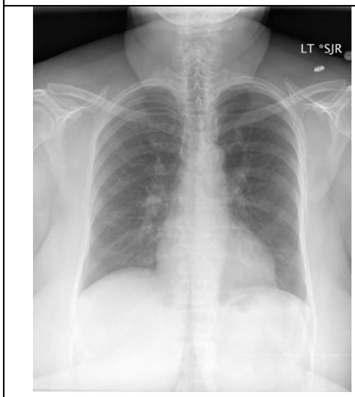
Chest X-ray Image	Report
	The cardiac silhouette and mediastinum size are within normal limits. There is no pulmonary edema. There is no focal consolidation. There are no XXXX of a pleural effusion. There is no evidence of pneumothorax.
	Borderline cardiomegaly. Midline sternotomy XXXX. Enlarged pulmonary arteries. Clear lungs. Inferior XXXX XXXX XXXX.
	Heart size and mediastinal contour are within normal limits. There is no focal airspace consolidation or suspicious pulmonary opacity. No pneumothorax or large pleural effusion. Mild degenerative change of the thoracic spine.

Fig. 1. Randomly selected images and reports from the indiana university chest x-ray collection dataset

Reports were written by doctors who are experts on chest images using traditional observational methods. The dataset contains a total of 6469 image report pairs written in English. The reports are generally observational reviews of disease symptoms and medical findings. Each report in the dataset is paired with one or more images. Multiple images of the same patient, such as frontal and side views, are available in the dataset.

The images in the dataset were pre-processed for resizing to be used in the deep learning architecture. The dataset was used by pairing one image with another image. The image-report pairs in the dataset were then divided into two main subgroups as 90% training-validation and 10% testing. The training-validation group was then divided into two subgroups, 90% training and 10% validation. In the final result, there are 5239, 583 and 647 image-report pairs in the training, validation and test subgroups respectively.

2.2 Vision Transformer (ViT) - Encoder

In order to generate a paragraph-level medical report from chest X-ray images, the first step is to extract the features indicating the diseases from the images. In this stage, Vision Transformer architecture is used in the proposed model. Vision Transformer (ViT) is a transormer structure that has been used in natural language processing applications for many years [12]. As in language applications, the image is divided into parts and then transformed into an array and processed. In many problems, images are divided into small representations called patches of size 16x16 [13]. It has achieved high success in many areas compared to classical CNN-based architectures.

One of the biggest advantages of the ViT architecture is the attention mechanism used in it. The attention mechanism provides a significant benefit in capturing long-range dependencies in medical images [14]. Another important advantage of the ViT architecture is that it has a modular structure. Thanks to its modular structure, it is simple to add new layers and modules to the ViT architecture. This makes it easy to combine the ViT architecture with other architectures and allows for performance optimization in solving problems. The disadvantages of the ViT architecture include the need to work with a large number of images and the computational cost. These disadvantages are minimized by having a sufficient number of images in the dataset and using the largest dataset available in the open access literature. Figure 2 shows an overview of the vision transformer architecture used in the encoder part of the proposed reporting model.

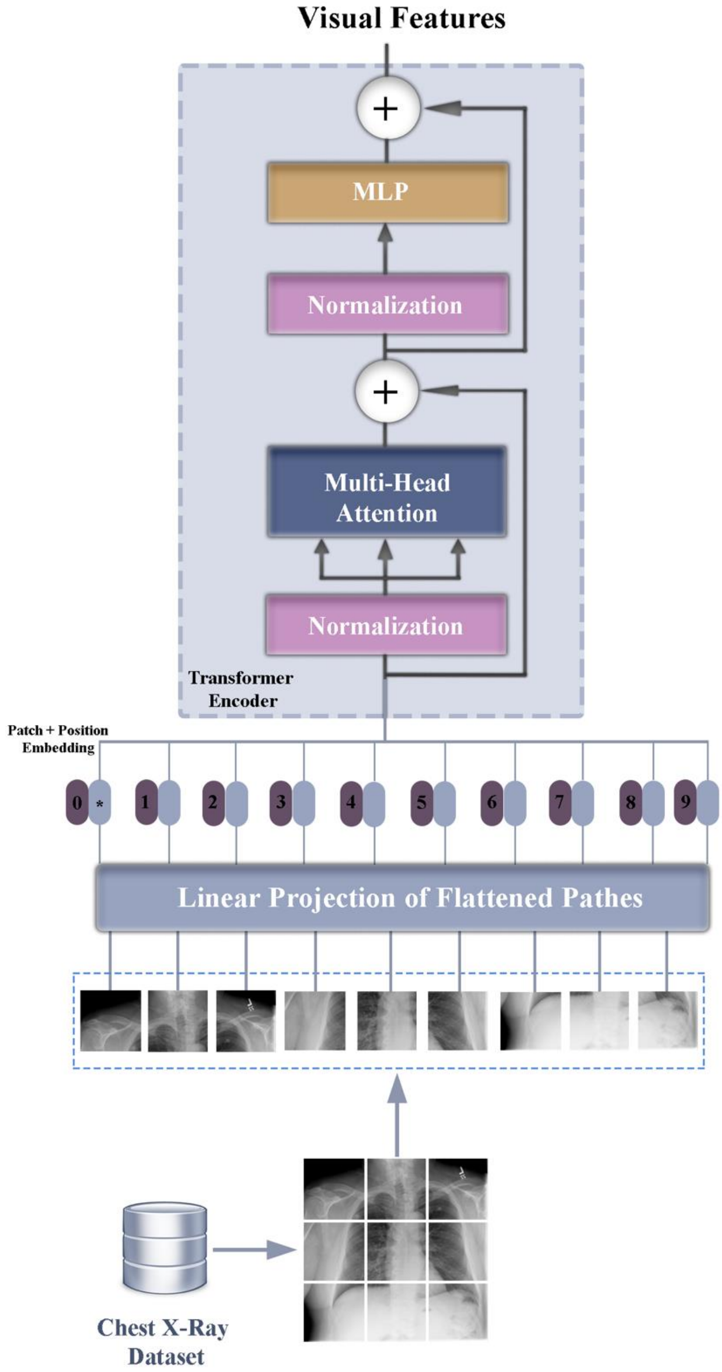


Fig. 2. Vision Transformer (ViT) model diagram for encoder section

2.3 Bidirectional Autoregressive Transformer (BART) - Decoder

The Bidirectional Autoregressive Transformer (BART) model is a variant of the transformer architecture and focuses on the text generation problem. It is a model developed by the Facebook artificial intelligence team that combines the advantages of BERT and GPT architectures [15]. In the problem of report generation from chest X-ray images, a vector representing the X-ray image is given to the decoder. The decoder uses the information in this vector to generate the medical report at the paragraph level. The BART model has the advantages of understanding the context, producing consistent text and learning successfully with less data [16]. The model, which is also used in natural language processing problems, has superior performance thanks to its ability to learn the two-way context of the language.

Medical reports have a complex structure and the texts are context dependent by nature. For this reason, BART architecture was preferred in the decoder part of the developed architecture. The fact that the BART architecture is superior in bidirectional understanding increases the success of the architecture in generating reports. Moreover, since the BART model is pre-trained on a wide range of training data, it can be fine-tuning for domain-specific tasks such as medical reporting. This is an important advantage of the BART model that increases its success in domain-specific tasks.

2.4 Evaluation Metrics

Word overlap evaluation metrics are used to compare autonomously generated reports from medical images with reports written by doctors. Metrics such as accuracy, precision and F1-score used in classification problems are insufficient to evaluate reports consisting of a large number of words and sentences. Bleu [17], Rouge [18] and Meteor [19] have used evaluation metrics. These metrics provide measurable parameters for how similar the autonomously generated text is to the original text in areas such as machine translation and medical reporting.

Bleu is a mathematical model that calculates how similar the texts produced autonomously by machine learning algorithms are to human translations. There are 4 different types of bleu metric, namely Bleu-1, Bleu-2, Bleu-3 and Bleu-4, and the types of bleu metric express n-gram values. Bleu-4 metric expresses the longest links between words and has n-gram values of (0.25, 0.25, 0.25, 0.25, 0.25). For these reasons, the Bleu-4 evaluation metric is the most accurate metric for determining the closeness between autonomously generated texts and reference texts.

The Meteor evaluation metric calculates the similarity between the reference text and the autonomously generated text with a measurable value using parameters such as term similarity, originality and prevalence. Another important word overlap evaluation metric is the rouge metric. This recall-oriented metric calculates how much the prediction text and the original text overlap.

3 Results and Discussions

Indiana University Chest X-Ray dataset was used in the training, validation and testing phases of the proposed medical reporting model. Vision transformer architecture is preferred in the encoder part and bart architecture is preferred in the decoder part of the model developed as encoder - decoder. After simple preprocessing, the images are given to the vision transformer architecture used in the feature extraction phase. In the decoder stage, the features extracted from the images are used to generate texts thanks to the powerful structure of the bart architecture. The encoder - decoder architecture developed in general terms is given in Figure 3 below.

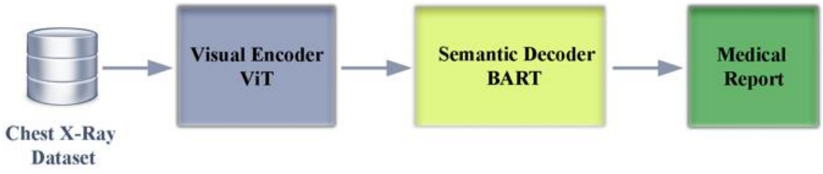


Fig. 3. Proposed Vi-Ba Medical Reporting Encoder - Decoder Model Diagram

Our medical reporting architecture Vi-Ba was developed using Pyhton programming language and pro+ version, a paid subscription of Google colab platform. In the training process for 10 epochs, batch size 8 was used. Learning Rate parameter was used as 5e-5. The hyperparameters used are given in table 1 below.

Table 1. Hyperparameters used in the proposed encoder and decoder architecture

Hyperparameter	Value
Batch size during training	8
Batch size during evaluation	8
Learning rate	5e-5
Weight decay for adamw optimizer	0.01
Train epochs	10

In order to observe the performance of the developed Vi-Ba model during the training periods, training and validation loss graphs were also drawn. The training and validation loss graphs are given in Figure 4. By analyzing the training loss and validation loss graphs, we can get information about the learning performance of the model and whether there is an overfitting problem. When the training and validation loss graphs of our deep learning model are examined, it is understood that a successful training process was carried out and there is no overfitting problem.

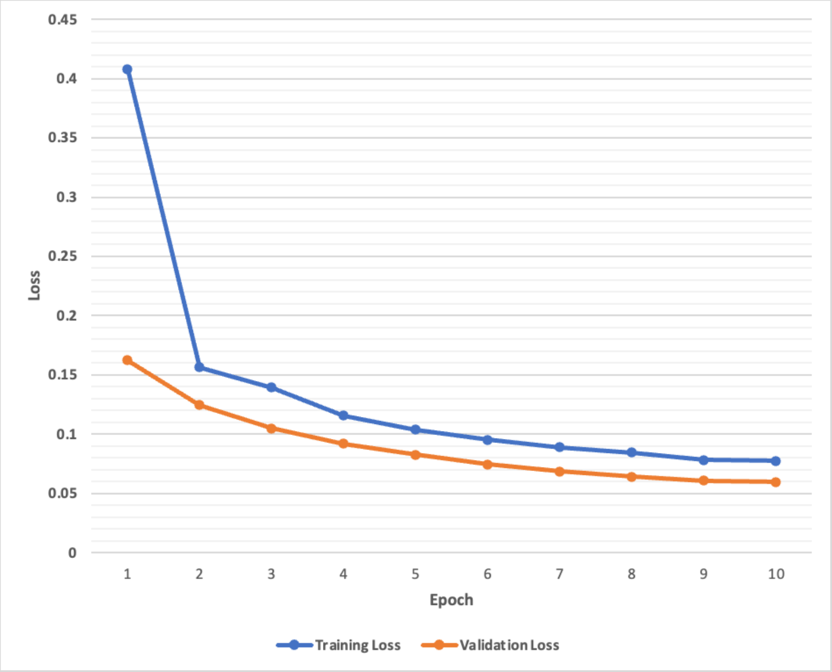


Fig. 4. Training and Validation Loss Graph of Vi-Ba Deep Learning Model Proposed for Medical Reporting

Numerical analyses were also performed in the study to clearly measure the success and competence of the model. The proposed architecture is compared with other popular studies using the same dataset. In this way, the advantages and disadvantages of the model over other models are revealed. Blue-4, Rouge and Meteor evaluation metrics are the most commonly used evaluation metrics in the literature. In Table 1 below, the model successes calculated with the evaluation metrics are compared with other popular studies in the literature.

Table 2. Comparison of medical reports produced by Vi-Ba architecture and other popular architectures in the literature with evaluation metrics

Method	Bleu-4	Meteor	Rouge
CNN-RNN [20]	0.095	0.159	0.267
RTMIC [5]	0.096	-	-
VSGRU [9]	0.116	0.150	0.251
TieNet [7]	0.073	0.107	0.226
(Our) Vi-Ba	0.150	0.154	0.274

The measurable evaluation parameters used in the table are used to measure how similar autonomously generated reports are to reports written by medical professionals who are experts in their field. Thanks to the metrics here, it is examined whether a similar report is produced with doctors. When the results obtained are analyzed, it is observed that the model produces successful results and produces better reports compared to its competitors in the literature. In the RTMIC model, the meteor and rouge metric were not calculated in the publication. Therefore, it is not included in the evaluation table. The proposed Vi-Ba model achieved the highest score in the blue-4 metric. It achieved the second highest score in the meteor metric, but the difference is only three percent. In the Rouge evaluation metric, the proposed Vi-Ba model also achieved the highest score.

4 Conclusion

The main objective of this study is to develop a high-performance deep learning model that can autonomously generate medical reports from medical images. The proposed model consists of a vision transformer encoder and a bart decoder. Training and testing on the model was conducted using images and reports from the Indiana University Chest X-Ray dataset. In the analysis, it is observed that the proposed architecture achieves higher success compared to other studies in the literature. The proposed Vi-Ba architecture achieved success scores of 0.150, 0.154, 0.274 in bleu-4, meteor and rouge word matching evaluation metrics, respectively. The analysis with measurable parameters showed that the autonomously generated reports achieved successful results.

The main contribution of our work to the literature is to present a new and successful model that can be used in chest X-rays and other medical images. This model can be used in practical life to support doctors in making decisions in the diagnosis of diseases. Diagnosis and reporting times from medical images will be shortened and the number of misdiagnoses and treatments will be reduced.

The limitations of our study are that training and testing processes are carried out using only chest X-ray images. In this context, only the images and medical reports in the Indiana University Chest X-Ray dataset were used in the study. Medical images from other fields were not analyzed within the scope of this study. In future studies, more comprehensive results will be obtained by applying the proposed architecture to images from other fields such as MRI, ultrasound and CT. In this way, it is planned to contribute to the autonomation of the processes of writing medical reports for diseases in different parts of the human body.

5 Acknowledgment

This research was supported by the Scientific and Technological Research Council of Turkey (TUBITAK) under Grant No 123E171.

References

1. UCan, M., Kaya, B., Kaya, M.: Multi-Class Gastrointestinal Images Classification Using EfficientNet-B0 CNN Model. In: 2022 International Conference on Data Analytics for Business and Industry (ICDABI). pp. 1–5. IEEE (2022)
2. Bütün, E., Uçan, M., Kaya, M.: Automatic detection of cancer metastasis in lymph node using deep learning. Biomed Signal Process Control. 82, 104564 (2023). https://doi.org/10.1016/j.bspc.2022.104564
3. Cinar, N., Ozcan, A., Kaya, M.: A hybrid DenseNet121-UNet model for brain tumor segmentation from MR Images. Biomed Signal Process Control. 76, 103647 (2022). https://doi.org/10.1016/j.bspc.2022.103647
4. Helaly, H.A., Badawy, M., Haikal, A.Y.: Toward deep MRI segmentation for Alzheimer's disease detection. Neural Comput Appl. 34, 1047–1063 (2022). https://doi.org/10.1007/s00521-021-06430-8
5. Xiong, Y., Du, B., Yan, P.: 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 10. pp. 673–680. Springer (2019)
6. Singh, S., Karimi, S., Ho-Shon, K., Hamey, L.: From Chest X-Rays to Radiology Reports: A Multimodal Machine Learning Approach. In: 2019 Digital Image Computing: Techniques and Applications (DICTA). pp. 1–8. IEEE (2019)
7. Wang, X., Peng, Y., Lu, L., Lu, Z., Summers, R.M.: TieNet: Text-Image Embedding Network for Common Thorax Disease Classification and Reporting in Chest X-rays. (2018). https://doi.org/10.48550/arXiv.1801.04334
8. Harzig, P., Chen, Y.-Y., Chen, F., Lienhart, R.: Addressing Data Bias Problems for Chest X-ray Image Report Generation. (2019). https://doi.org/10.48550/arXiv.1908.02123
9. Alfarghaly, O., Khaled, R., Elkorany, A., Helal, M., Fahmy, A.: Automated radiology report generation using conditioned transformers. Inform Med Unlocked. 24, 100557 (2021). https://doi.org/10.1016/j.imu.2021.100557
10. Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., Lungren, M.P., Ng, A.Y.: CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. (2017)
11. Demner-Fushman, D., Kohli, M.D., Rosenman, M.B., Shooshan, S.E., Rodriguez, L., Antani, S., Thoma, G.R., McDonald, C.J.: Preparing a collection of radiology examinations for distribution and retrieval. Journal of the American Medical Informatics Association. 23, 304–310 (2016). https://doi.org/10.1093/jamia/ocv080
12. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929. (2020)
13. Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y., Yang, Z., Zhang, Y., Tao, D.: A Survey on Vision Transformer. IEEE Trans Pattern Anal Mach Intell. 45, 87–110 (2023). https://doi.org/10.1109/TPAMI.2022.3152247
14. Manzari, O.N., Ahmadabadi, H., Kashiani, H., Shokouhi, S.B., Ayatollahi, A.: MedViT: A robust vision transformer for generalized medical image classification. Comput Biol Med. 157, 106791 (2023). https://doi.org/10.1016/j.compbiomed.2023.106791
15. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461. (2019)

16. Zhou, F., Qin, B., Lan, G., Ye, Z.: News Text Generation Method Integrating Pointer-Generator Network with Bidirectional Auto-Regressive Transformer. In: 2023 2nd International Conference on Artificial Intelligence and Intelligent Information Processing (AIIIP). pp. 114–118. IEEE (2023)
17. Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
18. Lin, C.-Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
19. Banerjee, S., Lavie, A.: METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. pp. 65–72 (2005)
20. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3156–3164 (2015)