ISSN: 2688-8645

A Review of Automated Report Generation Technologies for Ophthalmic Medical Imaging

Yuanshuo Zheng, Pan Su

Department of Computer, North China Electric Power University, Baoding 071003, China

Abstract

The integration of medical and technology has significantly improved diagnostic accuracy, but it has also led to certain challenges. Hospitals and clinics generate, store, and examine large volumes of medical imaging data daily, placing increasing pressure on radiologists. As a result, there is a growing need for technology-assisted report generation. Automated medical report generation technology can reduce physician workload, minimize diagnostic errors, and expedite the clinical diagnostic process. Different types of medical images are generated based on various imaging principles, each with its own advantages and limitations in aiding diagnosis, thus requiring distinct approaches for processing. This paper explores the application prospects of deep learning in the generation of ophthalmic medical reports.

Keywords

Ophthalmic Medical Imaging; Deep Learning; Automated Report Generation.

1. Introduction

With the advancement and development of modern technology, the use of medical machines for auxiliary diagnosis is expected to be a trend in the future. In China, the top ten most common eye diseases include refractive errors (such as myopia, hyperopia, presbyopia, and astigmatism), conjunctivitis, dry eye disease, cataracts, blepharitis, retinal diseases, strabismus, amblyopia, glaucoma, and uveitis. Among these, cataracts are the leading cause of blindness, with an incidence rate of 80%-90% in individuals over 60 years old. In non-blinding conditions, refractive errors are becoming an increasing concern in society, with nearly 70% of individuals under 20 years old being myopic, and this rate continues to rise.

Ophthalmic diseases often require medical imaging for auxiliary diagnosis, leading to an increasing demand for the analysis and processing of medical images. Medical imaging refers to the technology and processes used to obtain internal tissue images of the human body or specific body parts in a non-invasive manner for medical or research purposes. It originated in the 19th century with the discovery of X-rays by Roentgen, marking the beginning of the use of physical imaging in clinical practice. Over the next few decades, various imaging technologies, including X-ray imaging, magnetic resonance imaging (MRI), ultrasound imaging, and nuclear medicine imaging, were developed, forming the medical imaging industry. Currently, the market size for medical imaging equipment accounts for approximately 16% of China's entire medical device industry, making it the largest subsector in the country's medical device market. The workload for the analysis and diagnosis of ophthalmic medical images is also increasing. Therefore, to improve the efficiency of ophthalmic consultations and shorten the time required for generating medical image reports, it is imperative to enhance the automation of this process.

Unlike natural images, medical images are acquired from medical equipment and contain pathological information about internal tissues, enabling doctors to diagnose patients more efficiently. Traditional medical report generation typically requires significant manual effort. To reduce the workload of radiologists while maintaining the quality of healthcare, the

ISSN: 2688-8645

automatic generation of image captions based on deep learning has become an interdisciplinary research challenge that needs further refinement. In the field of image processing, deep learning has demonstrated powerful recognition capabilities, significantly improving experimental results in natural image classification, object detection, segmentation, and image captioning. This has greatly accelerated the automation of workflows, enhancing the quality and standardization of healthcare.

2. Current Research Status and Development Trends

The development of medical imaging technology spans just over a century. From the initial emergence of imaging techniques to the advent of actual medical imaging devices, it took nearly 80 years. It wasn't until the 1980s, with the successful application of magnetic resonance imaging (MRI) for whole-body scanning, that medical imaging truly entered a new era.

Since the 1980s, with the rapid growth of information technology, computer science, applied mathematics, materials science, and manufacturing industries, there has been significant progress, especially with the interdisciplinary application of knowledge. This has further advanced the development of medical imaging technology. Medical imaging has evolved from two-dimensional (2D) to three-dimensional (3D), and then from three-dimensional to four-dimensional (4D) imaging. Accompanying these advancements are developments in qualitative and quantitative analysis techniques for medical imaging, as well as significant progress in molecular, physiological, functional, metabolic, and genetic imaging, along with specific enhancement and AI-powered image recognition technologies. As medical imaging technology becomes increasingly integrated with clinical applications, new imaging technologies and analytical techniques are becoming more widespread.

In recent years, the automatic generation of reports has gained significant attention in the fields of machine learning and automated report generation. The goal is to produce semantically coherent, information-rich reports that describe reference diagnostic images, such as chest X-rays, lung CT scans, or fluorescein angiography. Such technologies hold great clinical potential in alleviating the heavy workload of radiologists and reducing diagnostic errors through improved interpretation.

3. Ophthalmic Imaging Techniques

Medical imaging is a field where deep learning has achieved significant success, and fundus imaging is one of its important branches. Fundus images are used for various tasks in ophthalmic disease diagnosis, such as disease grading, segmentation of lesions, and identification of critical biomarkers. These tasks correspond to several key areas in deep learning, including classification, segmentation, detection, and synthesis. Deep learning techniques have demonstrated strong capabilities in automating these processes, enhancing diagnostic accuracy and efficiency in ophthalmology.

3.1. Fundus Photography

The human pupil is small, and the light flux entering the eye is minimal. Moreover, the pupil automatically constricts in response to bright light, further reducing the light flux reaching the retina. Therefore, in order to capture a larger area of the retina, specialized fundus cameras are required. Fundus images are 2D images of the retina captured by monocular cameras. Standard fundus photography uses visible light to capture color fundus images, which roughly resemble what is observed with the naked eye. In contrast, fluorescein fundus photography utilizes three different wavelengths of laser light to penetrate varying depths of the retina, with the resulting images being superimposed to create a false-color fundus image. Short wavelengths target the vitreoretinal interface and the superficial retinal layers, medium wavelengths target the retinal

ISSN: 2688-8645

vascular layer, and long wavelengths focus on the retinal pigment epithelium and choroidal layers. Fundus photography is simple, non-invasive, cost-effective, and can be easily implemented in primary healthcare settings and non-ophthalmological departments. Following screening, cases with stages 3 or higher can undergo further fluorescein angiography (FFA) for a more detailed examination, effectively improving the cost-effectiveness of screening.

Ultra-widefield fundus imaging is based on a laser confocal scanning system and can capture images of the retina from the equator to the ora serrata in a single scan. The imaging principle can be exemplified by the ultra-widefield laser scanning fundus examination technology. An elliptical mirror with two conjugate points is used, where light reflected from one focal point passes through the other. This principle is utilized in ultra-widefield imaging, where the laser scanning head and the examined eye are positioned at two conjugate focal points. This setup allows low-energy laser beams to enter the pupil. As the laser scanning head rotates precisely and stably around the conjugate focal points, the retina is scanned as though the retina were inside the eye, allowing for a single scan that covers up to 200° of the retinal area (approximately 80% of the retina), including the far peripheral retina beyond the vortex veins. The laser energy reflected from the retina is directed to the elliptical mirror and transmitted through the same scanning system. After frequency conversion by a color probe, the data is transformed into a high-resolution digital image.

3.2. Fluorescein Fundus Angiography (FFA)

Fluorescein Fundus Angiography (FFA) is one of the most common and important diagnostic methods in the evaluation of retinal diseases. It plays a critical role in the differentiation, diagnosis, treatment, and prognosis of various retinal conditions. This technique emerged in the 1960s and involves the rapid intravenous injection of sodium fluorescein through the forearm vein. As sodium fluorescein circulates through the retinal vessels, a specialized fundus camera continuously captures the fluorescent patterns emitted by the blood vessels. This allows for the observation of the dynamic circulation process within the retina, ultimately providing evidence-based medical insights into the pathophysiology, diagnosis, and prognosis of various retinal diseases. FFA is crucial for diagnosing, treating, and assessing the prognosis of these conditions.

FFA has broad indications and is commonly used for the examination of most retinal diseases. It is suitable for conditions such as age-related macular degeneration, diabetic retinopathy, optic nerve disorders (e.g., optic disc edema, optic neuritis, ischemic optic neuropathy), intraocular tumors, central serous chorioretinopathy, acute posterior multifocal placoid pigment epitheliopathy, retinal vitreous membrane cysts, retinal pigmentary degeneration, multiple evanescent white dot syndrome, vascular occlusive retinal diseases, retinal vascular striae, Stargardt's macular dystrophy, and idiopathic polypoidal choroidal vasculopathy, among others.

3.3. Optical Coherence Tomography (OCT)

Optical Coherence Tomography (OCT) is an emerging optical imaging technique that relies on interference between ballistic and serpentine photons, which are scattered back from the medium. When the path length difference of these photons and the reference light is within the coherence length of the light source, interference occurs. In contrast, scattered photons with a path length difference greater than the coherence length do not interfere, allowing the extraction of the ballistic and serpentine photons that carry information from the sample for imaging purposes. This technique enables non-invasive, high-resolution tomography of biological tissues and holds significant potential for various applications.

OCT technology evolved from the optical coherence reflectometer (or low-coherence optical reflectometer). In 1991, David Huang and colleagues at the Massachusetts Institute of

ISSN: 2688-8645

Technology (MIT) first reported Optical Coherence Tomography (OCT) in *Science*. Later, Schmitt et al. applied this technology to measure the optical properties of biological tissues, yielding excellent results. In 1996, Carl Zeiss Meditec Inc. of California developed the OCT system for ophthalmology, which was launched as a clinical medical device.

3.4. Fundus Autofluorescence (AF)

Fundus Autofluorescence (FAF), also known as AF, refers to a method of observing the fluorescence emitted from normal or abnormal tissues in the retina using a specialized fundus camera. FAF imaging captures the unique "fluorescence" pattern emitted by the retina when it is exposed to short-wavelength laser light, typically before any fluorescein injection. The most commonly used wavelength for this purpose is a short 532 nm laser, which provides excellent autofluorescence images. FAF is a form of in vivo imaging technology for the retina, enabling the assessment of the metabolic activity of fluorescent substances in the retina under both normal and pathological conditions.

4. Automated Medical Image Report Generation Method

Recent research in automatic radiology report generation models has made significant strides [1][2][3][4][5]. Most existing studies adopt traditional encoder-decoder architectures, where convolutional neural networks (CNNs) serve as encoders and recurrent (e.g., LSTM/GRU) or non-recurrent networks (e.g., Transformer) function as decoders, following the image-captioning paradigm [6][7][8]. While these methods have achieved notable performance, they still face limitations in fully utilizing the information within both radiological images and reports.

In [9], an effective and simple method for radiology report generation is proposed, enhanced by a cross-modal memory network (CMN) that promotes interaction between modalities (i.e., image and text). The method uses a memory matrix to store cross-modal information and performs memory querying and response on visual and textual features. The corresponding responses are then fed into the encoder-decoder architecture, generating reports enriched by this explicitly learned cross-modal information. In [10], a novel framework is introduced that automatically annotates chest X-rays using image features and text embeddings extracted from relevant reports. A multi-level attention model is integrated into an end-to-end trainable CNN-RNN architecture to highlight meaningful words and image regions, transforming the annotation framework into a chest X-ray reporting system that outputs disease classifications and spontaneously generates preliminary reports. In [11], a new spatial attention model is proposed to extract spatial image features. This is followed by an adaptive attention mechanism, introducing a new LSTM extension that generates an additional "visual sentinel" vector and a sentinel gate that determines how much new information the decoder should derive from the image. The resulting adaptive attention encoder framework automatically decides when to rely on visual signals and when to rely solely on the language model.

Further advancements include the SFNet model in [12], which uses object detection algorithms to automatically separate lesion areas from background regions in medical images, allowing the encoding vector to carry more information related to lesions. In [13], a relational memory (RM) is proposed to store information from previous stages, and a new memory-driven conditional layer normalization (MCLN) is introduced to integrate the relational memory into the Transformer. In [14], a dynamic graph combining specific and general knowledge is used to enhance visual representations learned via contrastive learning, creating a framework called DCL. [15] introduces a hybrid retrieval-generation augmented agent (HRGR Agent), which uses a retrieval strategy module to decide between generating sentences and retrieving specific sentences from a template database based on prior knowledge collected from existing medical reports. This agent is trained using reinforcement learning (RL) with guidance from sentence-

level and word-level rewards. In [16], a self-enhancing framework is proposed to facilitate X-ray report generation, using auxiliary tasks to explicitly learn strongly correlated visual and textual features for image-report pair matching. In [17], a model known as Neural Image Captioning (NIC) is proposed, replacing the encoder RNN with a deep CNN, using the last hidden layer as input to the RNN decoder for sentence generation. In [18], a recognition architecture is presented to address the shortcomings of part- and texture-based models. It uses two cell neural network-based feature extractors whose outputs are combined at each image location using an outer product, merging multiple locations to produce image descriptors. In [19], CLIP is used as a visual feature extractor, training on image-text pairs to generate text-enhanced image features. Additionally, a memory-enhanced sparse attention mechanism is introduced and embedded into the transformer encoder and decoder. To integrate semantic text concepts, a lightweight medical concept generation network is introduced to predict fine-grained semantic concepts and integrate them into the medical report generation process.

5. Conclusion

In summary, the automatic generation of ophthalmic medical image reports has seen rapid development in recent years and is gradually being integrated into clinical practice. By combining various technologies such as image processing, deep learning, and natural language processing, automated report generation not only improves workflow efficiency but also enhances the accuracy and consistency of reports. However, despite significant progress, current technologies still face challenges such as data diversity, model generalization capabilities, and clinical interpretability. In the future, with continued algorithm optimization, the expansion of datasets, and deeper integration of technology with clinical practice, ophthalmic medical image report generation systems are expected to become more widespread in practical applications, providing stronger support for ophthalmologists' decision-making.

Acknowledgments

The Fundamental Research Funds for the Central Universities (2024MS129).

References

[1] Baoyu Jing, Pengtao Xie, and Eric Xing. On the Automatic Generation of Medical Imaging Reports.In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume1: Long Papers), 2018, pages 2577–2586.

[2] Yuan Li, Xiaodan Liang, Zhiting Hu, and Eric P Xing. Hybrid Retrieval-Generation Reinforced Agent for Medical Image Report Generation. In Advances in neural information processing systems, 2018, pages 1530–1540.

[3] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chihying Deng, Roger G Mark, and Steven Horng. MIMIC-CXR: A large publicly available database of labeled chest radiographs. 2019, arXiv preprint arXiv:1901.

[4] Guanxiong Liu, Tzu-Ming Harry Hsu, Matthew McDermott, Willie Boag, Wei-Hung Weng, Peter Szolovits, and Marzyeh Ghassemi. Clinically Accurate Chest X-Ray Report Generation. In Machine Learning for Healthcare Conference, 2019, pages 249–269.

[5] Baoyu Jing, Zeya Wang, and Eric 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, 2019, pages 6570–6580.

[6] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and Tell: A Neural Image Caption Generator. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pages 3156–3164.

[7] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pages 6077–6086.

[8] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. M2: Meshed-Memory Transformer for Image Captioning. 2019, arXiv preprint arXiv:1912.08226.

[9] CHEN Zhihong, et al. Cross-modal memory networks for radiology report generation. 2022, arXiv preprint arXiv:2204.13258.

[10] WANG Xiaosong, et al. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018, p. 9049-9058.

[11] LU Jiasen, et al. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017, p. 375-383.

[12] ZENG Xianhua, et al. Generating diagnostic report for medical image by high-middle-level visual information incorporation on double deep learning models. Computer methods and programs in biomedicine, 2020, 197: 105700.

[13] CHEN Zhihong, et al. Generating radiology reports via memory-driven transformer. 2020, arXiv preprint arXiv:2010.16056.

[14] LI Mingjie, et al. Dynamic Graph Enhanced Contrastive Learning for Chest X-ray Report Generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023, p. 3334-3343.

[15] LI Yuan, et al. Hybrid retrieval-generation reinforced agent for medical image report generation. Advances in neural information processing systems, 2018, 31.

[16] WANG Zhanyu, et al. A self-boosting framework for automated radiographic report generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021, p. 2433-2442.

[17] VINYALS, Oriol, et al. Show and tell: A neural image caption generator. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2015, p. 3156-3164.

[18] LIN Tsung-Yu, ROYCHOWDHURY Aruni, MAJI Subhransu. Bilinear CNN models for fine-grained visual recognition. In: Proceedings of the IEEE international conference on computer vision. 2015, p. 1449-1457.

[19] WANG Zhanyu, et al. A medical semantic-assisted transformer for radiographic report generation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer Nature Switzerland, 2022, p. 655-664.