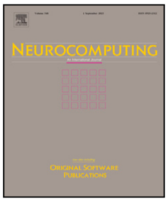
Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



VG-CALF: A vision-guided cross-attention and late-fusion network for radiology images in Medical Visual Question Answering

Aiman Lameesa [a], Chaklam Silpasuwanchai [b], Md. Sakib Bin Alam [c,*]

[a] Department of Computer Science, Faculty of Science and Technology, American International University - Bangladesh (AIUB), Dhaka, Bangladesh
[b] Computer Science and Information Management Program, School of Engineering and Technology, Asian Institute of Technology (AIT), Pathum Thani, Thailand
[c] Department of IT, Department of Computer Science and Engineering, University of Information Technology and Sciences (UITS), Dhaka, Bangladesh

ARTICLE INFO

Communicated by Q. Huang

Keywords:
Vision-guided
Cross-attention
Late-fusion
Medical visual question answering

ABSTRACT

Image and question matching is essential in Medical Visual Question Answering (MVQA) in order to accurately assess the visual-semantic correspondence between an image and a question. However, the recent state-of-the-art methods focus solely on the contrastive learning between an entire image and a question. Though contrastive learning successfully model the global relationship between an image and a question, it is less effective to capture the fine-grained alignments conveyed between image regions and question words. In contrast, large-scale pre-training poses significant drawbacks, including extended training times, handling substantial data volumes, and necessitating high computational power. To address these challenges, we propose the Vision-Guided Cross-Attention based Late Fusion (VG-CALF) network, which integrates image and question features into a unified deep model without relying on pre-training for MVQA tasks. In our proposed approach, we use self-attention to effectively leverage intra-modal relationships within each modality and implement vision-guided cross-attention to emphasize the inter-modal relationships between image regions and question words. By simultaneously considering intra-modal and inter-modal relationships, our proposed method significantly improves the overall performance of MVQA without the need for pre-training on extensive image-question pairs. Experimental results on benchmark datasets, such as, SLAKE and VQA-RAD demonstrate that our proposed approach performs competitively with existing state-of-the-art methods.

1. Introduction

The ability to answer questions from medical images is critical in Medical Visual Question Answering (MVQA) task. MVQA is a specialized type of Visual Question Answering (VQA) task that involves interpreting medical-related visual concepts by blending image and language information. To be more specific, MVQA systems take medical images and clinical questions as inputs and provide textual answers as responses [1]. Patients can benefit from such systems by receiving immediate responses to their questions and making better decisions [2]. It can also assist doctors in getting a second opinion on a diagnosis, potentially reducing the risk of misdiagnosis [3].

Past work mostly focuses on using pipeline approaches to process images and questions, where the vision and language information are extracted separately and then concatenated them for answering [4,5]. These approaches generally comprise four key elements: an image encoder, a question encoder, an attention-driven fusion of visual and textual features, and an answer classifier [6,7]. Previous research used skip-thought vectors and recurrent neural networks as question encoders. In terms of visual encoders, previous MVQA studies [1,8] used Mixture of Enhanced Visual Features (MEVF) [9], which leverages the advantages of Model-Agnostic Meta-Learning (MAML) [10] and Convolutional Denoising Autoencoder (CDAE) [11] to deal with limited labeled data. While MEVF, as a pre-trained vision encoder, effectively addresses specific issues in the VQA-RAD dataset [12], it may not generalize well to other datasets, such as SLAKE.

Recently, Radford et al. [13] introduced Contrastive Language based on Image Pre-training (CLIP), a pre-training approach, which involves predicting the correct association between captions and images, facilitating the learning of image representations from scratch on a large-scale dataset. Along with image-captioning, CLIP can be fine-tuned on other representative tasks, such as, VQA [14]. Eslami et al. [12] developed PubMedCLIP, demonstrating the effectiveness of contrastive learning between vision and language using a fine-tuned CLIP on MVQA. Their approach yielded notable performance gains on both SLAKE and VQA-RAD datasets over the pipeline approaches. However, contrastive learning primarily focuses on capturing the global relationship between the entire medical image and the question, but it is less

* Corresponding author.
E-mail address: sakibsba.cs@gmail.com (Md.S.B. Alam).

https://doi.org/10.1016/j.neucom.2024.128730
Received 8 March 2024; Received in revised form 28 August 2024; Accepted 12 October 2024
Available online 24 October 2024
0925-2312/© 2024 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.