Weight Affinity based Contextual Channel Attention with Exponentially Reduced Compression Factor for Dual Segmentation and CAM guided Classification of Generalized Medical Image

Md. Abtahi Majeed Chowdhury, Dr. Md. Kamrul Hasan

Department of Electrical and Electronic Engineering, Bangladesh University of Engineering and Technology (BUET), Dhaka, 1205, Bangladesh

E-mail: 1806106@eee.buet.ac.bd and khasan@eee.buet.ac.bd

Abstract. *Objective.* Computerized medical image segmentation and classification has played a crucial role in identifying affected tissues and diagnosing diseases precisely. Proliferation of publicly available clinical datasets has paved a way for researched to develop deep-learning based methods. Still a robust, reliable, theoretically and experimentally refined method, which may be effective across diverse medical image modalities remains elusive. *Approach.* This Paper introduces a novel interactive channel-affinity based contextual attention module for simultaneous image segmentation and classification task. The proposed method attained state-of-the-art results in segmentation while maintaining competitive results in classification showcasing inter-depending feature fusion between those task. It also proposes a specialized attention module for dual task establishing inter-channel interaction based on kernel-affinity, preserving contextual bias. Several XAI approaches applied for improving it's reliability. *Experimental Results.* Experimental evaluations conducted on four diverse publicly available medical image dual segmentation and classification datasets: BUSI, BUSI-Corrected, ISIC-2017, Waterloo; Segmentation only datasets: CVC-ClinicDB, DSB-2018, ISIC- 2018, PH2 demonstrate the proposed method's superior performance. The proposed method surpasses state-of-the-art in all the above mentioned datasets. Furthermore, our proposed method demonstrates robust multi-domain segmentation capabilities, exhibiting consistent and reliable performance. *Significance* The proposed method provides robust and reliable segmentation and classification performance on medical images, and thus it has the potential to be used in a clinical setting for the diagnosis of patients

Keywords— CAM, XAI, k-means, clustering, CNN, MBConv, FusedMBConv

1. Introduction

Medial Image Segmentation is a crucial area of interest in digital image processing field which has diversified applications in image analysis, augmented reality, machine vision and various other domains. Moreover, automatic computer-aided image segmentation in medical field turned out to be vital due to the need of precision and robustness while unreliability in manual diagnosis has been seen [4]. Deep learning based methods has enhanced diagnostic accuracy and consistency over diverse applications, along with precise identification of structured procedure. Automation of such labor-intensive tasks has improved time-efficiency and scalability. Support for early diagnosis and intervention with detection of subtle features has accelerated its practical usage. With the advance of medical technology, integration of medical image processing has assisted in tackling cross-modality imaging techniques, hence ensuring its versatility. Ultimately, to build a personalized medical system and accelerating drug development deep learning based lesion segmentation and classification has provided data-driven insights.

AlexNeT [30] is considered to be the pioneer research in this field as it introduced convolutional neural network based classification. InceptionNeT [40] has introduced modules to efficiently capture multi-scale features using parallel convolutional paths. [23] has introduced depthwise separable convolutions to reduce model size. ResNeT [20] has addressed vanishing gradient problem and added residual connections. [37] has introduced inverted residuals with linear bottlenecks to further optimize efficiency. EfficientNeT [41] has produced optimal result with depth, width and resolution scaling. Compound scaling of all the parameters has improved accuracy. Integration of channel attention in [42] has further improved results showing immense potential of attention modules in classification and segmentation pipelines. Invention of vision transformer [13] has paved an alternative way of image processing. Some of the drawbacks of vision transformer has been addressed in [44] involving global and local feature extraction. Inspite of impressive results, these networks lacks reliability as the region of feature extraction cannot be determined.

UNet [36] architectures has revolutionized lesion segmentation in medical images combining low-level features with high level features through skip connections between encoder and decoder. UNet++[54] has replaced skip-connections with a dense convolutional path to bridge semantic gap between encoder and decoder features. It has improved details with better feature alignment. ResUNet++[28] has presented ASPP module to capture multi-scale features with dilated convolutions. MultiResUNet[25] further capture fine details and high-level context simultaneously with residual path and skip-connection. However, more advanced segmentation networks have been introduced to incorporate background information with foreground and address unattended bottleneck layer.

CPFNet[16] provides fusion of find-grained spatial details and high-level contextual understanding with pyramid pooling. TransAttUNet[9] has introduced bottleneck attention layer to incorporate low-level feature effectively and achieved significant improvement. PraNet[15] highlighted the significance of background segmentation aggregating with foreground segmentation to improve reliability and precision. COMA-Net[1] introduced complementary attention guided bipolar refinement modules for background and foreground segmentation. Twin-SegNet[2] proposed dynamic coupling complementary segmentation networks for medical image segmentation. Hi-gMISnet[39] is a GAN based segmentation network using DWT based multilayer fusion and dual mode attention. As these architectures heavily focuses on precise segmentation, preservation of key features to classify or differentiate between lesions cannot be ensured with only them.

Classification and segmentation networks have evolved separately and successfully. Yet, outcomes of concurrent researches in medical imaging field have necessitated architectures to improve reliability and explainability through fusion of those tasks. As a recent advancement in XAI[35], CAM(class activation map)[53] has been introduced to visualize the region of feature of extraction. More advanced visualization methods like Grad-CAM[38],[8],[48],[17] have been introduced to incorporate ROIs with classifier. CAM-QUS[43] has achieved significant stability and accuracy by engaging cam-based and quantitative losses with classifier. Hence, these finding proves that precise aggregation of the lesion from which convolutional network is extracting features is a matter of concern for building classifier to achieved reliability.

Some recent researches [33][33][49][27] attempted the integration of classifiers or class information of the lesions into segmentation networks to preserve contextuality. Yet, very few researches focused on the simultaneous segmentation and classification. Depending on the mode of images, classifiers can extract features from foreground and background, shapes and contrast of a lesion or a particular set of lesions, ultimately, differentiating those region of interest from the whole image. This tendency can further improve the precision of lesion segmentation and achieve attention to details with global feature extraction.

Intensive research has been done on the attention modules. Attention modules help establish inter-channel interaction which previously not possible, thus reducing probability of failure in feature extraction. Spatial and Channel attention modules achieved significant results. Self-attention and multi-head attention modules in [46] has revolutionized with multiplicative operations inside convolutional pipeline. Specifically, transformer based networks have very high accuracy in classification tasks, yet fails to achieve competitive results in segmentation tasks. Some success have been achieved in adaptive vision transformers [31][44] and hybrid ViT-CNN models [10][18] with optimal usage of spatial and channel attention modules.

Squeeze-and-excitation block[24] uses global average pooling to extract the average value over each slice. Pooled values goes through 2 FC layer where the number of neurons are reduced dramatically. CBAM[51] integrates spatial attention and channel attention sequentially to enhance feature representation. Both spatial and channel attention uses GAP and GMP layer along respective axis and multiply their attention masks with input. Due to compression over channel axis, most of the channel information are lost. Cross-channel interaction will be established but the relational basis can neither be explained nor be relied.

ECA-Net [50] proposed an proposed an interesting solution to these. Instead of using fully connected layers, it directly captures cross-channel relation using local 1-D shifted convolution over channel axis. It doesn't compress channel information but it has neighborhood channel bias. As the initialization process of the channel weights are randomized, high correlation between neighbor channels are theoretically undeterminable. As the purpose of channel attention is to boost the output of convolution layers, boosting bias should be determined through highly correlated channel interaction.

Another drawbacks of channel-attentions modules that it completely loses spatial info. In classification phase, the existence of a specific feature at any region over the whole space is the only matter of interest. But it contradicts with the segmentation purpose where separate regions convey separate information. GCNet[6] uses weighed average pooling instead of global average pooling to maintain spatial bias. FCA-Net[34] uses prominent frequency components to provide frequency domain attention along with channel attention.

In model compression techniques [19], [21], [45], [22], [29], they tries to find out the significant layers to extract features while excluding irrelevant feature maps. Model compression with clustering employs algorithms like K-means to partition neural network weights into clusters based on similarity. Each weight is replaced by its corresponding cluster centroid, effectively reducing the precision and diversity of parameters. This results in lower storage requirements as only centroid values and cluster indices are stored. This technique minimizes redundancy in the parameter space and is frequently integrated with quantization or pruning to optimize computational and memory efficiency for deployment on resource-constrained hardware. From these techniques, we can infer that channel weights even after clustering can preserve its' properties and hence assume high result with very low computational complexity.

To evaluate the performance and generalizability of our proposed network, we conducted extensive testing on eight diverse medical image segmentation and classification datasets covering a range of imaging modalities. The experimental results demonstrate that our method consistently outperforms established segmentation approaches across all datasets and excels in extracting high-resolution features from input images. In summary, this paper makes the following significant contributions:

(i) Introduced a novel channel-attention module which groups the channels based on weight-affinity and establish a reliable inter-channel interaction. It exponentially decreases compression factor from traditional channel attention modules while maintaining edge over computational complexity. Meanwhile, to improve segmentation performance, patching is introduced to achieve a localized impact of channel attention on feature maps.

(ii) A supervised class activation map (CAM) based classifier network is incorporated with the encoder to strengthen the segmentation process. We have shown that classifier can dramatically improve segmentation task while achieving explainability and reliability. Additionally, a class activation map at the bottleneck supervises the high-level feature and ensures that encoded features are derived from region of interest.

(iii) Our architecture has been tested intensively over diverse medical image modality. Its' high explainablity feature make it a best fit for practical medical application. Moreover, we have compared our method against several other architecture and obtained state-of-the-art result in all the metrics.

(iv) the weight-affinity based channel attention opens the gate to attain reliability to attention module and rather than treating them as black-box, we are theoretically supervising their mechanism and their effect on application.