Class-consistent Contrastive Learning Driven Cross-dimensional Transformer for 3D Medical Image Classification

Qikui Zhu[1] , **Chuan Fu**[2] , **Shuo Li**[1]
[1]Department of Biomedical Engineering and
Department of Computer and Data Science, Case Western Reserve University, OH, USA
[2]Department of Computer Science, Chongqing university, Chongqing, China
QikuiZhu@163.com, fuchuan2024@163.com, slishuo@gmail.com

Abstract

Transformer emerges as an active research topic in medical image analysis. Yet, three substantial challenges limit the effectiveness of both 2D and 3D Transformers in 3D medical image classification: 1) Challenge in capturing spatial structure correlation due to the unreasonable flattening operation; 2) Challenge in burdening the high computational complexity and memory consumption due to the quadratic growth of computational complexity and memory consumption for 3D medical data; 3) Challenge in discriminative representation learning, due to data-sensitivity. To address the above challenges, a novel Cross-dimensional Transformer (CdTransformer) and a creative Class-consistent Contrastive Learning (CcCL) are proposed. Specifically, CdTransformer consists of two novel modules: 1) Cross-dimensional Attention Module (CAM), which breaks the limitation that Transformer cannot reasonably establish spatial structure correlation when meeting 3D medical data, meanwhile, reduces the computational complexity and memory consumption. 2) Inter-dimensional Feed-forward Network (IdFN), which addresses the challenge of traditional feed-forward networks not being able to learn depth dimension information that is unique to 3D medical data. CcCL innovatively takes full advantage of the inter-class and intra-class features from the slice-distorted samples to boost Transformer in learning feature representation. CdTransformer and CcCL are validated on six 3D medical image classification tasks. Extensive experimental results demonstrate that CdTransformer outperforms state-of-the-art CNNs and Transformers on 3D medical image classification, and CcCL enables significantly improving Transformer in discriminative representation learning.

1 Introduction

Transformers [Vaswani *et al.*, 2017] have achieved performance breakthroughs in capturing long-range token dependencies and global feature extraction. However, it meets three challenges (Figure.1(a)) when applied to 3D medical image data: 1) The flattening operation of Transformer destroys the 3D spatial structure of data, making it difficult to capture spatial structure correlations and brings irrelevant information and ambiguity to feature representation learning for 3D medical image data. 2) Huge computational complexity and memory consumption due to the computational complexity and memory consumption of Transformer is quadratic to the spatial size. 3) Challenge in learning the feature representation with limited positive and negative samples, particularly for discriminative representation learning, due to transformer-based models are only effective when large training datasets are available. Although many Transformer methods [Fu *et al.*, 2024; Guo *et al.*, 2021; Chen *et al.*, 2021b; Zhu *et al.*, 2022; Zhu *et al.*, 2023a] have achieved encouraging performance in 3D data analysis, the aforementioned challenges remain

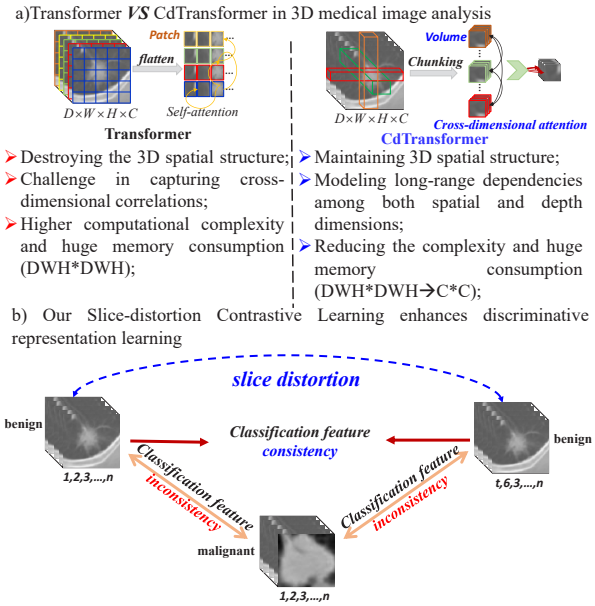

Figure 1: Our CdTransformer and CcCL overcome the disadvantages of Transformers in 3D medical image analysis by maintaining local 3D spatial structure, reducing the complexity and memory consumption, and improving discriminative representation learning.

unresolved and limit the effectiveness of Transformer on 3D medical image analysis.

Existing 3D Transformers are also unable to overcome the aforementioned challenges. For example, the 3D Transformer used in [Zhou *et al.*, 2023] employs Local Volume-based Multi-head Self-attention (LV-MSA) and Global Volume-based Multi-head Self-attention (GV-MSA) to construct feature pyramids for learning representations on 3D volumes. Although LV-MSA reduces computational resources, it can only extract local information. Furthermore, GV-MSA faces the challenge of high computational complexity while learning global feature representations. The Multi-plane and Multi-slice Transformer [Jang and Hwang, 2022] extracts 3D feature representations by constructing attention relationships among multi-plane (axial, coronal, and sagittal) and multi-slice images, but it is unable to avoid the disruption of 3D spatial structure. Additionally, the Multi-plane and Multi-slice Transformer, built upon the Transformer architecture, also encounters the challenge of high computational complexity for 3D input. Despite the various strategies employed by existing methods to integrate Transformers with 3D data, these 3D Transformers still fail to overcome the challenges inherent in 3D medical data. Hence, there is an urgent need for a computationally efficient, 3D structurally aware Transformer in the field of 3D medical image analysis.

We propose a novel Cross-dimensional Transformer (CdTransformer) that consists of 1) a Cross-dimensional Attention Module (CAM) and 2) an Inter-dimensional Feed-forward Network (IdFN) to exploit long-range token dependencies among 3D spatial pixels and extract the global discriminative representations to overcome above challenges and further improve the ability of Transformer on 3D medical image analysis (Figure.1(a)). Specifically, CAM introduces a novel cross-dimensional attention that promotes information sharing and fusion between different dimensions for building connections between 3D spatial pixels, addressing the challenge that Transformer lacking reasonable spatial structure correlation establishes mechanisms. More significantly, CAM converts the quadratic memory consumption of Transformer to linear memory consumption for addressing high complexity and memory consumption problems. IdFN adopts a novel Inter-dimensional attention to activate and fuse the effective features from both spatial and depth views, overcoming the challenge of Feed-forward Network ignoring the depth dimension information.

Apart from architectural novelties, a novel contrastive learning, named Class-consistent Contrastive Learning (CcCL) (Figure.1(b)), is proposed to overcome the limitation that Transformer-based methods are ineffective when meeting limited training datasets [Chen *et al.*, 2020]. Specifically, CcCL innovatively exploits the inter-class feature and intra-class feature from the slice-distorted samples by taking full advantage of Transformer's strengths in long-range dependencies learning. By maximizing the consistent of positive slice-distorted pairs while minimizing the inconsistency of negative slice-distorted pairs, CcCL enables Transformer to learn class-aware discriminative features under limited training datasets. We conduct comprehensive experiments and demonstrate the significance of CcCL in boosting discriminative representation learning with limited positive and negative samples.

Our proposed CcCL and CdTransformer were validated on six 3D medical image classification tasks including five 3D MedMNIST datasets [Yang *et al.*, 2023] and one well-known LIDC-IDRI dataset [Kuan *et al.*, 2017]. Extensive experimental results demonstrate that CdTransformer outperforms state-of-the-art CNNs and Transformers on 3D medical image classification.

Our main contributions include:

- A cross-dimensional Transformer with linear computational complexity and memory consumption has been established for 3D medical image analysis, which addresses three limitations of Transformer in 3D medical image analysis by effectively capturing spatial structure correlations while mitigating the challenges of high computational complexity and memory consumption.
- Our class-consistent contrastive learning innovatively takes full advantage of the inter-class feature and intra-class feature from the slice-distorted samples to boost the effectiveness of the Transformer in learning discriminative representation.
- Our cross-dimensional attention module innovatively establishes the 3D spatial structure correlation in 3D medical image data, where Transformer cannot.
- Our inter-dimensional feed-forward network enables advanced aggregating spatial and depth contexts from both spatial and depth views, which addresses the limitation that the feed-forward Network of Transformer ignores the depth dimension information.
- Experimental results on six 3D medical image datasets demonstrate that our cross-dimensional Transformer serves as a generalized module, exhibiting remarkable capabilities in 3D medical data analysis.

2 Related Work

Many authors try to use Transformer [Dosovitskiy *et al.*, 2020] for 3D medical image analysis [Zhao *et al.*, 2023; Zhu *et al.*, 2023b; Qin *et al.*, 2022]. For example, Xie et al. [Xie *et al.*, 2021] proposed a hybrid model of CNN Transformer, namely CoTr, for 3D medical image segmentation. Inside the model, the deformable Transformer (DeTrans) that employs the deformable self-attention mechanism is introduced to reduce the computational and spatial complexities of modelling the long-range dependency on multi-scale and high-resolution feature maps. Wang et al. [Wang *et al.*, 2022] proposed an fNIRS classification network based on Transformer, named fNIRS-T, for functional near-infrared spectroscopy classification. To explore the spatial-level and channel-level representation of fNIRS signals, two Transformers, fNIRS Spatial-level Transformer (fNIRS-ST) and fNIRS Channel-level Transformer (fNIRS-CT), are employed inside model. fNIRS-ST can extract local brain area features and fNIRS-CT for the hemodynamic response of a single channel. Hatamizadeh et