CSE 4095 Literature Review

Computer Vision with Transformersin Medical Modeling

Jacob Murphy, Rany Kamel, and Andrew Eikleberry
University of Connecticut
Department of Computer Science and Engineering

***Abstract*—Computer vision is an important area of machine learning which provides ample opportunity for exploration, particularly in its various applications and implementations such as with the Transformer architecture. The Transformer architecture is an emerging machine learning model showing immense potential in the fields of natural language processing (NLP) and, more recently, computer vision (CV). This paper provides a detailed study of recent literature on the topic of transformers and their applications in the area of computer vision. In addition, this paper discusses the benefits of utilizing transformers for CV applications, and their benefits over other machine learning models such as convolutional neural networks (CNNs).**

I. Transformers in Computer Vision

A. Jiarui Bi, Zengliang Zhu, Qinglong Meng

The article serves as an in-depth summary of various Transformer implementations for computer vision applications. The authors review 15 articles pertaining to vision-based transformer applications including object detection, tracking, and visual segmentation. Though initially designed for use in Natural Language Processing, there as been a recent explosion of literature on the use of vision-transformers. The article explains that transformer architecture addresses many of the issues seen in other machine learning technologies, such as RNNs and CNNs. Transformers are becoming increasingly popular for computer-vision tasks as they offer a multitude of benefits including less inductive bias and bigger respective fields.

The article then goes into detail, discussing various applications of vision-based transformers. Most notable to the subject of this paper is the discussion on image segmentation. The article discusses three algorithms related to image segmentation. PloyTransform is an image-segmentation algorithm which combines multiple segmentation methods to generate geometry-based masks in object identification. TeTrIS introduces methods to incorporate shape prior data into image segmentation. Finally, VisTR is a video instance segmentation framework based on transformers, consisting of a CNN, encoder and decoder transformer, and instance sequence segmentation and matching modules.

In conclusion, this article covers various recent literature on the topic of transformer architecture and its use in computer vision tasks. It serves to educate and provide examples which are relevant to the topic of this paper.

II. An Image is Worth 16x16 Words

A. Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby

This article introduces a novel transformer architecture designed for computer vision. Called the Visual Transformer (ViT), it was intended to replace Convolutional Neural Networks with a complete transformer solution, requiring significantly less computational resources

III. UNETR: Transformers for 3D Medical Image Segmentation

A. Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R. Roth, Daguang Xu

This article discusses the advantages of transformer architecture as it relates to image segmentation and medical imaging. Fully Convolutional Neural Networks (FCNNs) are used in the majority of medical imaging tasks. However, they are limited in their accuracy due to the inherit locality of their convolutional layers. According to the article, transformers have no such limitation and thus could serve as an improved method of machine-learning-driven 3d medical image segmentation. To this end, the article introduces a novel architecture, UNEt TRansformers (UNETR). UNETR utilizes a transformer as the encoder to learn sequence representations of the input volume and effectively capture the global multi-scale information, while also following the successful "U-shaped" network design for the encoder and decoder. The article also details a test method in which they conclude with the new architecture demonstrating state of the art performance.

With regards to bench-marking, the authors present qualitative multi-organ segmentation comparisons. These comparisons demonstrate UNETR outperforming comparable technologies. Compared to 2D transformer-based models, UNETR shows higher boundary segmentation accuracy and identifies boundaries accurately between kidney and spleen, gallbladder, liver and stomach, and portal vein against liver. The authors also show that UNETR captures much finer

detail compared to other models.

To conclude, the article introduces a new transformer-based architecture for 3D medical image segmentation. The findings therein, as well as the code base and data set provided will serve as the basis of this project.

REFERENCES

[1] J. Bi, Z. Zhu and Q. Meng, "Transformer in Computer Vision," 2021 IEEE International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology (CEI), Fuzhou, China, 2021, pp. 178-188, doi: 10.1109/CEI52496.2021.9574462.

[2] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R. Roth, Daguang Xu, "UNETR: Transformers for 3D Medical Image Segmentation" The Computer Vision Foundation, 2022

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale" ICLR, 2021