CSE 4095 Final Paper

Computer Vision with Transformers in Medical Modeling

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

***Abstract*—Computer vision is a crucial sub-field of machine learning that offers numerous opportunities for research and innovation, particularly with regard to various applications and implementations, such as the Transformer architecture. The Transformer architecture has emerged as a powerful machine learning model, demonstrating remarkable potential in the domains of natural language processing (NLP) and, more recently, computer vision (CV). This project aims to investigate the application of Transformer architecture in computer vision tasks, specifically focusing on medical image segmentation.**

In the medical domain, the accuracy of modeling and medical imaging is of paramount importance for patient care. When it comes to human health, medical professionals demand the highest possible level of precision. While Convolutional Neural Networks (CNNs) are widely employed for medical imaging driven by machine learning, they have certain limitations in terms of accuracy. Building on recent literature, this project seeks to employ Transformer architecture to enhance the accuracy of computer-generated medical images.

I. Project Revisions and Difficulties

A. Project History and Revisions

Initial efforts into the project were somewhat vague in terms of deliverables. At the time of the proposal, the stated goal of the project was to simply "improve upon" existing models for medical imaging segmentation. This was not nearly specific enough to be accomplished in the time allotted for the project. Once feedback was received, the team decided to revise the problem statement to set a more measurable goal. The updated and current problem statement reflects a desired outcome of altering an already established code base to be used in medical image segmentation, with the goal of analysing tumors within the brain.

The revised problem statement reads as follows: Within the medical field, accuracy of modeling and medical imaging is critical to patient care. When human health is involved, medical professionals require as close to perfection as is feasible. CNNs, while commonly used for machine-learning-driven medical imaging, are limited in this degree. This project will expand on recent literature by utilizing transformer architecture in order to detect and model cancerous brain tumors, facilitating medical assessment.

As should be apparent, the new problem statement is much more specific, measurable, and within scope. While the task itself is still difficult, with this new problem statement there is still a clear path towards a deliverable final project.

B. Project Difficulties

Over the course of the project the team encountered and overcame a number of difficulties. The project's code is based on the Medical Open Network for AI (MONAI) code base, which required implementation and modification to suit the specific needs of the project. The code relied heavily on multiple hard-coded environment assumptions. Many of these assumptions proved to be incorrect when applied to the team's environment. This necessitated the exploration of alternative solutions to ensure that the code could be properly executed and adapted for the project's goals.

One such solution was running the code on Google Colab, which proved to be the best environment in terms of compatibility and capability. Ultimately Google Colab proved to be the most optimal environment for the project.

An additional issue was that the data set required for the model was prohibitively large. As the data set consisted of NIfTI (.nii) MRI files instead of more common formats like JPEG or PNG, its size was significantly larger than those typically used in training computer vision applications, and thus manipulating the data to be usable for the project was a tedious task.

Finally, the pre-trained models provided by the repository did not prove to be particularly beneficial for the purposes of the project. Ultimately the team chose to instead train the model themselves.

C. Lessons Learned

Throughout the course of the project, the team learned several valuable lessons that can be applied to future projects. These lessons encompass the areas of code adaptation, environment selection, data set management, and model training.

When utilizing existing codebases, it is important to carefully examine any hard-coded assumptions and verify their applicability to the project at hand. Adapting another's

code to a new application can at save a lot of time, but depending on the implementation, it can also serve to delay progress. This also reinforces a number of previous lessons instilled upon the team in past classes with regards to proper software engineering practises.

Choosing the appropriate environment for running code can significantly impact the success of a project. In this case, Google Colab proved to be the optimal choice due to its compatibility and capabilities. This highlights the importance of testing various environments and selecting the one that best meets the project's needs. This also shows how one can end up in a seemingly endless spiral of dependancies, which can be prevented by proper environment planning. On a related note: large, unconventional data sets can pose significant challenges. It is imprtant to understand and plan for such data sets to ensure one can handle the dataset required for the project.

To address these issues and improve the performance of our proposed model, several steps could be taken in future iterations:

Given more time and resources, the team could train the UNETR-based segmentation model more thoroughly. This proved to be the major crux of the project, so working to this end would result in a great improvement. Additionally, implementing data augmentation techniques to expand the training dataset and enhance the model's ability to learn from diverse input samples could lead to great model accuracy. This can include rotation, scaling, and flipping of MRI images to simulate various tumor appearances and sizes.

II. Model Application and Performance

A. Model

Our proposed model utilizes the UNETR transformer model to segment brain tumors into 3 categories: Whole Tumor, Tumor Core (non-growing section of the tumor), and Enhancing Tumor (rapidly growing section of the tumor). Our intent was to feed the resulting segmentation to a custom CNN model, and train it to classify whether the brain tumors are benign (a non-harmful tumor) and malignant (a cancerous tumor). The rational is that cancerous tumors would have bigger 'Enhancing Tumor' regions, as the tumor is rapidly growing, whereas the benign tumor would have a much smaller or non-existent enhancing layer.

Due to our inability to use the pretrained models provided, due to the reasons mentioned prior, we decided to train a novel model for the segmentation component using the UNETR architecture. Due to the lack of time and resources, however, our resulting model was not trained thoroughly. The model's output was thus very noisy and nowhere near the Ground Truth of any of its inputs. This made it improbable to train the CNN component of the model, as it would essentially be trained on incorrect data.

III. Research and Framework

A. DETR

One article the team looked at when developing the project was "Transformer in Computer Vision" written by Jiarui Bi and published by the IEEE. This article goes over an overview of transformers models in regards to computer vision. One model of computer vision transformer that it went over is the DETR model which was presented by Nicolas Carion which looked at object detection as a set prediction problem. The model is capable of predicting all objects in an image at once, once it is trained end-to-end with a set of loss functions. These functions allow the DETR to perform parallel decoding and better performance on large objects.

The article stresses that the DETR suffers from slow convergence and limited feature resolution. In order to fix the issues prevalent in DETR, the updated model Deformable DETR was proposed. The deformable update could be extended to incorporate multiple scale targets however this upped the complexity of the model compromising its efficiency. This created the need for the development of the UP-DETR which was proposed by Zhigang Dai1 and was inspired by the transformers used in NLP. It randomly crops images and feeds them to queries in the decoder. The implementation of UP-DETR freezes the CNN backbone and proposes a patch feature reconstruction branch that is jointly optimized with patch detection.

B. MONAI and UNETR

The MONAI code base is an open network for AI and offers many open source applications and frameworks for use in the medical field. The framework the team based the project after is a UNETR transformer which is notable in that the UNETR is a pure vision transformer and does not rely on a CNN for extraction. UNETR was introduced in the article "UNETR: Transformers for 3D Medical Image Segmentation", written by Ali Hatamizadeh and fellow staff at NVIDIA in order to remove the use of the FCNN encoder which they believed limited the capability of long ranged spatial dependencies. The UNETR architecture utilizes a stack of transformers as the encoder which works via skip connections. This concept is similar to the ViT whci was the first to propose the use of a purely transformer architecture for 2D images. The UNETR takes advantage of a purely transformer architecture as well however it is unique in that it is designed for use on 3 dimensional data sets which are then segmented and processed as a series of 2D data sets. The article goes on to test the UNETR model to the Ground Truth model, TransBTS model, CoTr model, and the UNet model. The Qualitative results showed that the multi-organ segmentation was better performed by the UNETR than any of the other models which utilized a CNN based transformer architecture. It demonstrates

the increased capacities of the long ranged dependencies which the team was attempting to do by erasing the need for a CNN. Moreover, the results of the segmentation tests over the BTCV model in the article demonstrated a “new state of the art benchmark and validates its effectiveness. Specifically for small anatomies”. The article also compares the decoding capabilities of the UNETR with that of other models such as the Native UPsampling and Progressive UPsampling, outperforming both decoders by 4.3 percent and 7.5 percent respectively. In addition to its performance improvements, the UNETR model has only a moderate model complexity which again outperformed the CNN based models.

The MONAI framework was built to review the BTCV challenge dataset which contained 50 CT scans of abdomens which contained colorectal cancer chemotherapy patients and a retrospective ventral hernia study. The framework was capable of segmenting these scans, as well as differentiating between the spleen, left and right kidneys, gallbladder, Eso, liver, stomach, aorta, IVC, P and S Vein, pancreas and Ad Glands.

C. ViT

In the article “An Image Is Worth 16X16 Words: Transformers for Image Recognition At Scale” written by Alexey Dosovitskiy, an overview of the visual transformer ViT, its applications and its performance are discussed and compared to other architectures with similar designs. Much like the previously discussed architectures, this is a vision transformer. It uses patch embedding techniques to flatten patches of an image into linear projections and puts these embedded pieces through a transformer encoder. This is one of the first architectures to propose not using a CNN for 2D images, and instead using purely a transformer encoder and decoder. In this article, the ViT is tested in comparison to BiT which stands for Big Transfer. The ViT performed superiorly to other transformer models which utilized a CNN and proved that for larger datasets, an architecture that does not take advantage of a CNN performs better while smaller datasets were better processed by the BiT CNN based architecture.

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