A Project Report

On

Optimizing Vision Transformer to deploy on resources constrained edge embedded devices

BY

Anurag Maiti

2022A7PS0205H

Under the supervision of **Dr. Anakhi Hazarika**

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS OF

CS F266: STUDY PROJECT

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE PILANI (RAJASTHAN)

HYDERABAD CAMPUS

(May 2024)

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to Dr. Anakhi Hazarika, EEE Department, for her valuable guidance and support throughout this study. I would also like to thank Swayam Joshi and Pradyum Agarwal for their contributions and meaningful suggestions without whom this study would remain incomplete.

Birla Institute of Technology and Science-Pilani,

Hyderabad Campus

Certificate

This is to certify that the project report entitled "**Optimization of Vision Transformers Model and deploying on edge-devices**" submitted by Mr.Anurag Maiti (ID No. 2022A7PS0205H) in partial fulfillment of the requirements of the course CS F266, Study Project Course, embodies the work done by him under my supervision and guidance.

Date:3-05-2024 **(Dr. Anakhi Hazarika)**

BITS- Pilani, Hyderabad Campus

ABSTRACT

Vision Transformers (ViTs) have recently surfaced as powerful competitors to traditional convolutional neural networks(CNNs), demonstrating remarkable performance in multiple computer vision tasks. This report presents our progress in the understanding and implementation of Vision Transformers with a comprehensive study on the techniques used such as self-attention mechanisms, positional encoding, and tokenization. After extensive research through multiple papers and Pytorch documentation, we have implemented the base model of Vision Transformers as well as identified key areas which might be subject to optimization.

CONTENTS

INTRODUCTION

The project aims to make Vision Transformers (ViTs) work better on small devices like smartphones or IoT gadgets. These ViTs are new ways of understanding images that can be faster and use fewer resources than traditional methods like Convolutional Neural Networks (CNNs). We want to do this because it could make image recognition tasks faster and cheaper, which is important for many applications.

We started by learning about ViTs and how they work. Then, we learned about the basic building blocks of deep learning in PyTorch, the software we use. With this knowledge, we built a simple ViT model to see how it performs.

Our goal is to find ways to make ViTs run faster and use less memory on small devices. We want to do this by changing parts of the ViT model to make it more efficient. For example, we might simplify some calculations or use special techniques to reduce the amount of memory it needs.

However, there are challenges. ViTs can be very complex and need a lot of computing power, which can be hard for small devices to handle. We need to find ways to make them work well on these devices without slowing them down too much or using too much memory. This means we have to be careful about how we design and optimize the model.

BACKGROUND

Transformers are a type of neural network architecture that has revolutionized natural language processing (NLP) tasks. They are designed to handle sequential data by capturing dependencies across long distances, making them particularly effective for tasks like language translation, text generation, and sentiment analysis.

In NLP models, transformers play a crucial role in encoding and decoding text sequences. The transformer architecture includes self-attention mechanisms that allow the model to weigh the importance of different words in a sentence, capturing complex relationships and dependencies within the text.

Recently, it has been realized that self-attention can be applied to images as well to capture long-range dependencies across image patches instead of words which allow transformers to understand the context of an entire image instead of just local features.
This has led to the development of transformer-based models such as Vision Transformers. In computer vision, transformers can process image patches as sequences of tokens, enabling them to understand spatial relationships and extract meaningful features from images.

LITERATURE REVIEW

Vision Transformers were proposed by Dosovitskiy et al. (2021) in AN IMAGE IS WORTH 16x16 WORDS[1] the first paper we started our study with ,understanding the basics of the Transformer model.It gives details on the basic building blocks of flattening, positional embedding, normalization, attention heads, and classification blocks. It then compares the performance of state-of-the-art CNNs with their models of varying sizes when run on datasets of varying sizes. The interpretation of an image as a sequence of patches and its processing through a standard Transformer encoder, commonly used in NLP, has proven to be a remarkably effective and scalable approach.

The following paper we used to familiarize ourselves with the model is ATTENTION IS ALL YOU NEED[2]**.** This focused more on the multi-head self-attention part of the model. It details the formula of Attention(Q, K, V) = softmax(QK^T / $\sqrt{d_k}$)V with explanations. In their work, the authors introduced the Transformer, which stands out as the inaugural sequence transduction model exclusively built on attention mechanisms. This innovation replaces the recurrentlayers typically employed in encoder-decoder architectures with multi-headed self-attention mechanisms.

Our further look out for more optimized techniques brought us to new concept by Xiangxiang Chu et al. (2023) CONDITIONAL POSITIONAL ENCODINGS[3]. Unlike previous fixed or learnable positional encodings that are computed only once and used through out, Conditional Positional Encoding(CPE) is dynamically generated. CPE can also accommodated longer sequences of input during training. This is implemented with PEG i.e Positional Encoding Generator and can be easily incorporated with existing ViT model.

While investigating Conditional Positional Encoding we embarked upon paper by Wenhai Wang et al. (2021) Pyramid Vision Transformers which overcomes certain difficulties of regular Vision Transformers. PVT can be trained on be trained on dense partitions of an image to achieve high resolution outputs unlike the ViT which gives low resolution output with high computational power.It uses a progressive shrinking pyramid like architecture to reduce computation on large feature maps .This boosts performance of PVT in tasks like object detection and semantic segementation.

OBJECTIVES

1) Optimizing the base model of ViT

To optimize the base Vision Transformer (ViT) model for efficient execution on edge devices, we have outlined a series of objectives:-

- Reviewing the foundational papers on ViT, Multi-head Self-Attention (MSA) blocks, and Positional Encoding to gain a comprehensive understanding of the model.
- Transitioning from Python TensorFlow, Scikit, etc., to PyTorch for enhanced flexibility.
- Independently implementing the base code of ViT to deepen our understanding and facilitate future optimizations
- Exploring components of the architecture that can be enhanced, such as modifying positional encoding for colored images and adapting the model for grayscale images to work with colored ones.
- Iteratively implementing, testing, and refining updates against the base model to achieve optimal performance while maintaining efficiency for edge deployment.

2) Implementing on hardware

The next objective which is to be worked upon in the future is to implement the optimized model on hardware

METHODOLOGY/PLAN OF ACTION

We began by immersing ourselves in the base paper and studying the attention mechanism paper to gain a comprehensive understanding of the Vision Transformer (ViT) model and its distinctions from conventional CNN models. This involved extensive research, including studying various resources such as YouTube tutorials and Medium articles, to grasp the fundamental design principles of ViT.

Transitioning to the implementation phase, we found a helpful guide on Medium and translated our theoretical knowledge into code. Alongside this, we meticulously documented our learnings on Google Colab, creating a comprehensive repository of notes for future reference.

During this period, we also made the strategic decision to transition from Python to PyTorch, aligning with the preferred framework for ViT implementation. This allowed us to update our Colab notebook with PyTorch-specific methodologies, ensuring continuity in our approach.

In summary, our approach involved a combination of theoretical understanding, practical implementation, and meticulous documentation, providing a solid foundation for further exploration of the ViT model.

Following this initial phase, we delved into optimization possibilities. We experimented with different positional encoding techniques to improve performance and explored the use of separate query, key, and value tensors, which were then combined using a Sequential layer. Additionally, we started investigating pyramid vision transformers and conditional positional encoding, both of which hold promise for reducing computational resources and improving accuracy, respectively.

Moving forward, we plan to explore more optimization techniques and conduct experiments to compare the performance of our optimized models against the base model on higher-end devices. This will allow us to run for a higher number of epochs and further refine our approach.

.

CONCLUSION

To sum up, our project set out to understand, apply, and refine the Vision Transformer (ViT) model for improved precision and effectiveness, with an emphasis on computer vision tasks. We gained a deep grasp of ViT's architecture by carefully examining its constituent parts, including multi-head self-attention and positional encoding, and by thoroughly reviewing seminal works. We independently built the underlying ViT model before moving on to a realistic implementation using PyTorch. We also noted possible areas for optimization, especially in the positional encoding and multi-head self-attention blocks. Our approach, which combined theoretical knowledge with practical application and documentation, created a solid foundation for the ViT model's future revisions and improvements. The enhanced model has potential for deployment on edge as we proceed.

REFERENCES

Sources used:

1. [2010.11929] An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale
2. [1706.03762] Attention Is All You Need
3. Conditional Positional Encodings for Vision Transformers
4. https://medium.com/mlearning-ai/vision-transformers-from-scratch-pytorch-a-step-by-step-guide-96c3313c2e0c
5. https://keras.io/examples/vision/image_classification_with_vision_transformer/
6. https://d2l.ai/chapter_attention-mechanisms-and-transformers/vision-transformer.html
7. https://www.v7labs.com/blog/vision-transformer-guide
8. https://colab.research.google.com/drive/1bYyJxKJ1uPv8kt0ZAa0nAT-MZKhPPZLW?usp=sharing
9. https://arxiv.org/abs/2102.12122