Illinois Institute of Technology
CS577 - Deep Learning Project Proposal

Sai Manohar Vemuri | A20514848
Aditya Shivakumar | A20513537
Sesha Shai Datta Kolli | A20516330

Object Detection using Vision Transformers (ViT)

Abstract:

Object detection is a fundamental computer vision task with numerous applications, such as autonomous vehicles, surveillance, and image analysis. Traditional object detection methods often rely on CNNs, which have shown remarkable performance. However, recent advancements in the field of deep learning have introduced Vision Transformers (ViTs) as a promising alternative for various computer vision tasks. This project aims to explore and implement object detection using Vision Transformers to leverage their potential advantages in capturing long-range dependencies and handling complex visual data.

Problem Statement:

The problem addressed in this project is to develop an efficient and accurate object detection system using Vision Transformers. The objective is to detect and localize objects of interest within images while achieving competitive performance compared to traditional CNN-based methods. The primary challenges include building the custom ViT architecture for object detection, handling varying object sizes, and optimizing for computational efficiency.

Related Work:

"Beal et al. [1] implemented a model called ViT-FRCNN by modifying the original Vision Transformers model for the object detection task. They removed the final transformer state, which outputs the class, and replaced it with the layer that generates the spatial feature map. This feature map is then fed to the detection network, which consists of a Faster R-CNN model. It predicts the presence or absence of objects in the image using RPN. Later, the extracted features are ROI-pooled and then fed to a detection head, which regresses the bounding box coordinates and also predicts the class label."

Yanghao Li et al. [2] suggest that, since it is computationally expensive to train the model to compute global self-attention, it is advisable to use a pretrained model to perform global self-attention. They explored the use of restricted self-attention, where the feature map is divided into non-overlapping windows, and self-attention is computed within each window."

Liu et al. [6] addressed the data-hungry issue of existing detection transformers by modifying how key-value pairs are constructed in the cross-attention layer. They concluded that multi-scale deformable attention samples sparse features from local regions. They implemented sparse feature sampling using RoIAlign for local feature sampling and used layer-wise bounding box refinement to improve the detection performance. They also performed Label Augmentation to allow the model to receive more guidance from the data, potentially leading to faster and more effective training."

Method:

[1]. Dataset Preparation:
The dataset we intend to use is COCO. We choose the mini version of this dataset consisting of 25,000 images and annotations with bounding boxes and class labels. The following is the link to the dataset: https://github.com/giddyyupp/coco-minitrain

[2]. Data Augmentation Techniques:
Explore advanced data augmentation techniques that can generate synthetic data effectively, reducing the need for a large annotated dataset. Techniques like CutMix, MixUp, etc.; can be beneficial. We might use one of the augmentation techniques.

[3]. Model Selection Based on Resources:
Evaluate the available computational resources, including hardware capabilities and budget constraints.
Decide between two strategies:
a. Building a Custom Model: If sufficient resources are available, design and train a custom Vision Transformer model tailored to the specific task and constraints.
b. Utilizing Pretrained Models: If computational resources are limited, select a pretrained Vision Transformer model as a starting point and fine-tune it for object detection, optimizing for both accuracy and efficiency.

[4]. Training and Fine-tuning:
For the custom model, train it from scratch on the object detection dataset, optimizing architecture and hyperparameters for computational efficiency.
For pretrained models, fine-tune them on the object detection dataset, leveraging transfer learning while optimizing for efficiency.

[5]. Performance Evaluation and Resource Utilization Metrics:
Evaluate the model, considering both standard metrics (IoU, mAP, F1-score) and resource utilization metrics (inference time, memory usage).
Benchmark the selected modeling approach against available resources, assessing its effectiveness and efficiency.

References:

[1]. Beal, Josh, Eric Kim, Eric Tzeng, Dong Huk Park, Andrew Zhai and Dmitry Kislyuk. "Toward Transformer-Based Object Detection." ArXiv abs/2012.09958 (2020)

[2]. Li, Yanghao & Mao, Hanzi & Girshick, Ross & He, Kaiming. (2022). Exploring Plain Vision Transformer Backbones for Object Detection. 10.1007/978-3-031-20077-9_17.

[3]. Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani et al. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929 (2020).

[4]. Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need." Advances in neural information processing systems 30 (2017).

[5]. Carion, Nicolas, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. "End-to-end object detection with transformers." In European conference on computer vision, pp. 213-229. Cham: Springer International Publishing, 2020.

[6]. Liu, Yahui, Enver Sangineto, Wei Bi, Nicu Sebe, Bruno Lepri, and Marco Nadai. "Efficient training of visual transformers with small datasets." Advances in Neural Information Processing Systems 34 (2021): 23818-23830.