Illinois Institute of Technology

CS-577 Intermediate Project Report

Group-31:
Sai Manohar Vemuri: A20514848
Sesha Shai Datta Kolli : A20516330
Aditya Shivakumar : A20513537

Object Detection using Vision Transformers (ViT)

INTRODUCTION:

Object detection, a fundamental task in computer vision, plays a pivotal role in a wide array of applications. The capability of object detection has diverse applications, ranging from surveillance and autonomous vehicles to medical imaging and content-based image retrieval, establishing itself as an important task of modern computer vision.

Traditional approaches to object detection heavily rely on Convolutional Neural Networks (CNNs), which have demonstrated impressive performance over the years. CNNs excel in capturing local features, utilizing convolutional layers to extract hierarchies of patterns and structures, which has not only revolutionized computer vision tasks but also significantly improved the accuracy and efficiency of various vision-related applications. Additionally, their ability to perform feature extraction and classification simultaneously, along with their robustness to variations, has made them an important technique for object detection.

However, the dynamic field of deep learning constantly introduces novel techniques and models. One such breakthrough is the introduction of Vision Transformers (ViTs), which

are seen as a promising alternative for various computer vision tasks. What sets ViTs apart is their unique architectural design, which includes the self-attention mechanism, which allows it to analyze an entire image, taking into account the intricate relationships between all its elements, instead of focusing solely on local features. This approach equips ViTs with the capability to comprehend global context, making them exceptionally versatile for a diverse range of computer vision tasks.

PROBLEM DESCRIPTION:

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 fine tuning the ViT architecture for object detection, handling varying object sizes, and optimizing for computational efficiency.

The Vision Transformer (ViT) is the first attempt to apply a pure transformer model directly to images, without using any convolutional layers. ViT divides an image into patches and treats them as tokens. Then, it applies a series of transformer blocks to encode the patches and produce a final classification output. ViT demonstrates that transformers can achieve competitive results on image classification benchmarks, such as ImageNet, compared to CNNs.

However, for more complex tasks such as object detection or segmentation, maintaining a high input resolution is crucial to ensure that models can properly identify and reflect fine details in their output. This poses a challenge for ViT, as it requires a large amount of computation and memory to process high-resolution images.

DATASET DESCRIPTION:

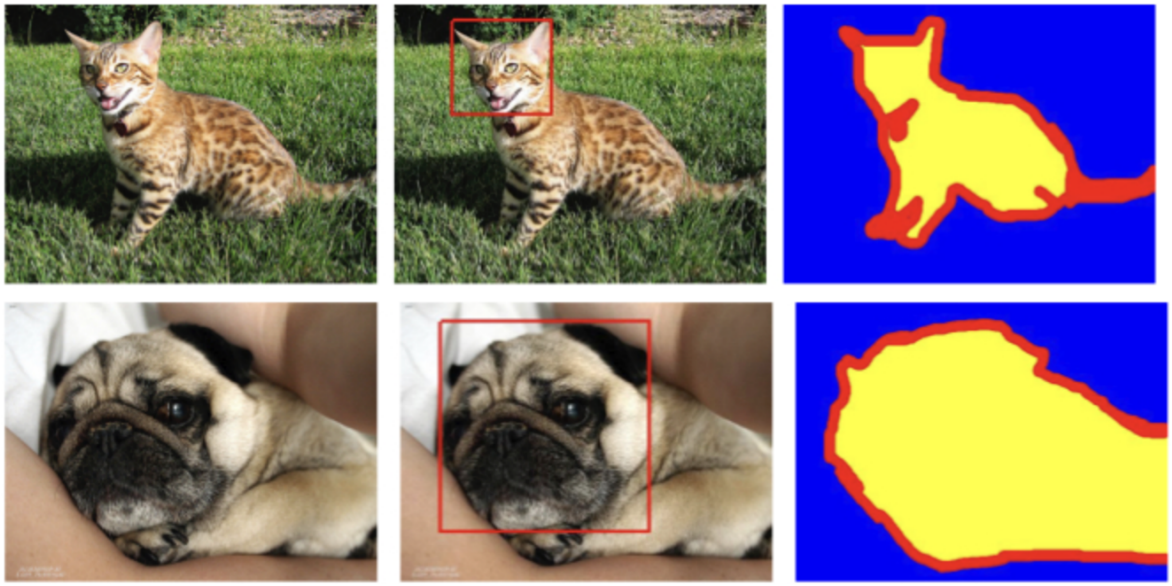
Dataset Name: Oxford-IIIT Pet Dataset

The Oxford-IIIT Pet Dataset is a 37 category pet dataset with roughly 200 images for each class created by the Visual Geometry Group at Oxford. The images have large variations in scale, pose and lighting. All images have an associated ground truth annotation of breed, head ROI, and pixel level trimap segmentation.

Each image in the dataset comes with a set of annotations. These annotations encompass valuable information, including (a) the species and breed name of the animal depicted in the image, (b) a meticulously delineated bounding box, precisely

framing the head of the animal, and (c) a pixel-level segmentation, effectively distinguishing the foreground from the background, achieved through a process known as Trimap.

Examples of annotations:



MILESTONES ACCOMPLISHED:

- Data normalization has been performed to ensure that the input data has a consistent scale.
- A custom dataset class has been created for data loading and processing. The dataset pairs images with their corresponding annotations, including labels and bounding boxes.
- 2-D input Images are flattened and converted into 1D sequences of embedding vectors with positional encoding.
- A multi-head self attention and a multi-layer perceptron blocks are created.
- A transformer encoder block is created, which combines the multi-head self attention and multi-layer perceptron blocks.

MILESTONES TO BE ACCOMPLISHED:

- Fine tuning of the model for better localization of the bounding boxes and Binary classification(cats and dogs).
- Evaluating the performance of the model by using different evaluation metrics.