Implementation of Transformer-Based Visual Segmentation

Nandan Menon
2310110532
nm392@snu.edu.in

Mukhil Charles T
2310110528
mt555@snu.edu.in

Jeevithesh R
2310110489
jr479@snu.edu.in

***Abstract*—This project implements Transformer-based and CNN-based models for visual segmentation tasks including semantic, instance, and panoptic segmentation. We observe the performance of both model types on the COCO dataset and compare their qualitative outputs and standard segmentation metrics. Through this work, we aim to provide insights into the architectural differences and practical implementation aspects of these models.**

***Index Terms*—Visual Segmentation, Transformer, CNN, DeepLabV3, SegFormer, Mask R-CNN, Mask2Former, Panoptic Segmentation, Instance Segmentation**

I. Introduction

Image segmentation is a key task in computer vision, where the goal is to partition an image into meaningful regions. This task plays an important role in applications such as autonomous driving, medical imaging, and robotics. Convolutional Neural Networks (CNNs) have been the dominant architecture for segmentation tasks for several years, owing to their ability to learn spatial hierarchies in images. However, CNNs struggle with capturing long-range dependencies within an image.

In recent years, Transformer-based models, originally designed for natural language processing, have been adapted for vision tasks. Transformers use self-attention mechanisms to model long-range dependencies, making them particularly powerful for segmentation tasks, where global context is crucial.

This project focuses on implementing CNN-based and Transformer-based models for semantic, instance, and panoptic segmentation. We aim to compare the outputs produced by each model and gain insights into their respective strengths and weaknesses.

II. Objectives

The main objectives of this project are as follows:

- To implement state-of-the-art CNN and Transformer models for various segmentation tasks.
- To explore the differences in the performance and design of CNN and Transformer-based models qualitatively.

III. Tasks and Models Implemented

For each segmentation task, we selected the following models:

A. Semantic Segmentation

- **CNN Model: DeepLabV3** DeepLabV3 utilizes atrous convolutions to capture multi-scale contextual information without losing resolution. This architecture is known for its high performance on semantic segmentation tasks.
- **Transformer Model: SegFormer** SegFormer is a lightweight and efficient Transformer model designed specifically for semantic segmentation. It uses a hierarchical encoder to capture multi-scale features and provides competitive performance while maintaining efficiency.

B. Instance Segmentation

- **CNN Model: Mask R-CNN** Mask R-CNN extends Faster R-CNN by adding a segmentation mask prediction branch to the region proposals. It is widely used for instance segmentation tasks.
- **Transformer Model: Mask2Former** Mask2Former is a Transformer model that unifies semantic, instance, and panoptic segmentation tasks with a single attention mechanism. It has shown promising results in instance segmentation due to its ability to handle complex object interactions.

C. Panoptic Segmentation

- **CNN Model: Panoptic FPN** Panoptic FPN combines both semantic and instance segmentation into a unified panoptic output, producing a complete scene understanding.
- **Transformer Model: Mask2Former** Mask2Former is also used for panoptic segmentation, offering a unified architecture that generates both stuff and things categories in a single forward pass.

IV. Dataset

The **COCO (Common Objects in Context)** dataset was used for evaluation in this project. COCO contains over 80 object categories, with annotated pixel-level segmentation masks, bounding boxes, and panoptic annotations. Due to hardware limitations, we focused on the validation set and analyzed a subset of images to visualize model outputs.

V. Evaluation Metrics

The models were evaluated using the following metrics:

- **Mean Intersection over Union (mIoU)** for semantic segmentation, which measures the overlap between predicted and ground truth segmentation masks.
- **Panoptic Quality (PQ)** for panoptic segmentation, which combines segmentation quality and recognition quality into a single metric.

VI. Implementation Details

A. Environment

The project was implemented using Google Colab, leveraging a Tesla T4 GPU for model inference.

B. Workflow

Pretrained models were loaded using the Hugging Face and Detectron2 libraries. Images were preprocessed by resizing and normalizing them before being passed through the models for inference. The models outputted segmentation masks, bounding boxes, and class labels, which were then visualized and analyzed qualitatively.

VII. Results and Analysis

In this section, we present and analyze the qualitative results obtained from our segmentation models.

A. Comparative Analysis

The qualitative results demonstrate that both CNN and Transformer-based models can effectively perform segmentation tasks. However, we observed that:

- Transformer models (SegFormer, Mask2Former) produced more consistent and fine-grained segmentation boundaries, especially for small objects.
- CNN models (DeepLabV3, Panoptic FPN) performed better on large and distinct regions but struggled with fine details.
- Confidence scores in panoptic outputs showed high reliability for clear object classes (like vehicles) but lower scores for cluttered regions.

VIII. Challenges Faced

During the implementation, we encountered several challenges:

- **Hardware Limitations**: Full-scale evaluations on the COCO dataset were computationally expensive. Due to memory and time constraints, we analyzed a smaller subset using pretrained model weights.
- **Model Complexity**: Transformer models like Mask2Former required more memory and computation compared to CNNs, sometimes causing runtime errors.
- **Metric Computation**: Full evaluation across the COCO validation set was infeasible, so only sample outputs were analyzed qualitatively.

IX. Conclusion

Through this project, we successfully implemented and qualitatively evaluated CNN and Transformer-based models for visual segmentation tasks. Our observations suggest that Transformer models have promising capabilities, especially in complex scenes involving multiple overlapping objects. This project provided valuable insights into practical challenges and architectural differences between CNNs and Transformers.

X. References

References

[1] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking Atrous Convolution for Semantic Image Segmentation," arXiv:1706.05587, 2017.

[2] X. Xie, Y. Wang, Z. Zhang, et al., "SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers," NeurIPS 2021.

[3] B. Cheng, I. Misra, A. Girdhar, et al., "Masked-attention Mask Transformer for Universal Image Segmentation," CVPR 2022.

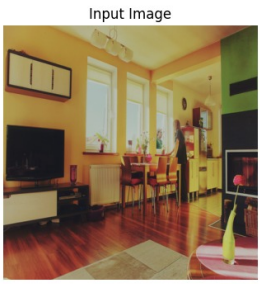
[4] T.-Y. Lin, M. Maire, S. Belongie, et al., "Microsoft COCO: Common Objects in Context," ECCV 2014.

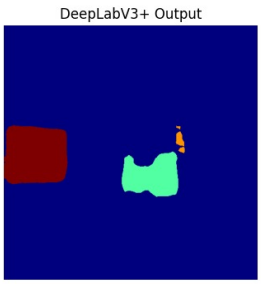
(a) Panoptic Input Image 1

(b) Panoptic Output Image 1

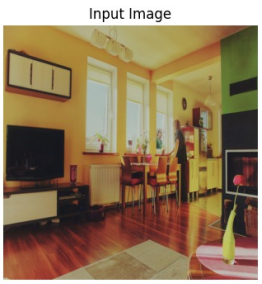
(c) Panoptic Input Image 2

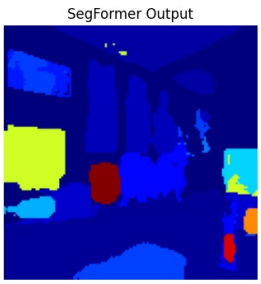
(d) Panoptic Output Image 2





(e) Semantic Output Image 1





(f) Semantic Output Image 2

Fig. 1: Visual results from panoptic and semantic segmentation models. Each row shows paired outputs.