Per-Segment Classification
Leveraging Segments into Contextual Vectors

Amit Levi
Benjamin Cohen

Advisors:
Yaniv Nemcovsky, Chaim Baskin, Gilad Levi

Technion University Computer Science Department

git@github.com:amit1221levi/SATSAT.git

Abstract

In contemporary computer vision research, traditional pixel-based methodologies, although prevalent, often inadequately capture the holistic context and semantic depth necessary for nuanced image understanding. This deficiency has prompted a paradigm shift toward segmentation-based classification techniques that more closely emulate human perceptual processes, which rely on minimal contextual cues for image interpretation. Our proposed framework advances this shift by employing a segment-based approach wherein images are decomposed into vectors of segments. These segments are subsequently tokenized with positional encodings and analyzed through an attention mechanism that assesses inter-segment relational dynamics to assign class memberships effectively. Our methodology draws inspiration from advancements in pixel-level analysis and incorporates BERT's masking strategy, which involves masking a significant proportion of segments to enhance predictive accuracy regarding their classes. This innovative approach not only aligns with the transformative impacts of Transformer architectures in processing visual data but also underscores the critical role of semantic relationships and contextual integrity in computer vision. By leveraging these elements, our framework aims to develop more intuitive, efficient, and nuanced models that transcend traditional analysis paradigms, fostering a deeper comprehension of visual scenes as integrated wholes rather than as mere aggregates of pixels.

1 Introduction

Traditional pixel-based methodologies in computer vision often fall short in capturing the holistic context and semantic depth required for comprehensive image understanding. To address these limitations, we introduce a segment-based classification framework that decomposes images into vectors of segments. These segments are tokenized with positional encodings and analyzed using an attention mechanism that evaluates inter-segment relationships. Our approach draws inspiration from the NLP model's strategy enhancing predictive accuracy by focusing on the contextual dynamics of segments in trams of finding patterns and relationships within sequential input. This method aims to transcend traditional pixel aggregation techniques for several benchmarks , offering a more nuanced and efficient model for segment identification and classification.

Related Work

Advancements in image segmentation technology have revolutionized the approach from traditional techniques to models capable of converting images into vectors of segments without specific task training. Notably, Segment Anything has been at the forefront of this transformation, enabling effective segmentation without the constraints of pre-defined models [1, 2]. For scenarios demanding higher efficiency, lighter models like Fast Segment Anything have been developed to streamline the segmentation process [3]. Incorporating positional encoding into segmentation models has significantly improved their capacity to interpret spatial relationships within images. This enhancement supports the use of sophisticated sorting algorithms which optimize segment arrangement while balancing performance and computational costs [4, 5]. The introduction of Vision Transformer (ViT) models marks a pivotal shift in segmentation, utilizing their self-attention mechanisms to enhance the processing of segment tokens. These models, when fine-tuned, show substantial improvements in handling data processed by advanced tokenizers like SegFormer and BEiT [6, 11]. Furthermore, BEiT models have been effectively applied to train semantic token boosters, aligning segment tokens with their corresponding labels to boost segmentation accuracy and model robustness [7, 8]. Moreover, the application of NLP techniques such as token masking has demonstrated efficacy in enhancing model predictability and robustness. By masking significant portions of input tokens, these models are compelled to predict missing labels, thus reinforcing their predictive capabilities under constrained input conditions [9, 10].

2 Methodology

Our methodology for semantic segmentation transforms images into segment vectors and processes these segments for contextual information, using advanced models and techniques for improving performance in semantic segmentation missions. **Segmentation with SAM:** Initial segmentation masks are generated using SAM. **Positional Encoding:** Location coordinates are added to each segment for spatial context, helping the model understand spatial relationships. **Segment Sorting:** We chose a sorting method based on common borders for efficiency and accuracy. **Contextual Vector Creation:** Segments are transformed into contextual vectors, incorporating positional encoding with shape and size information. Each segment is processed to create masks and labels essential for classification tasks. **Mask Creation:** Binary masks are generated, setting pixels belonging to the segment to one and reducing others by factor R and above we added patches 4 and 8 into the vector. **Labeling:** Segments are assigned labels from a comprehensive semantic segmentation database. **Grid Creation:** Images are divided into labeled patches (4x4 or 8x8), ensuring each patch contains all segment information for accurate classification.

3 Experiments

We trained two SegFormer models, one with our pre-processing tokenization to assess improvements. Finally we integrated the Bidirectional Encoder Representations from Transformers (BIET) model, adapting the BERT architecture for image tasks. We developed a segment token booster to enhance the model's ability to recognize and categorize image segments, using a comprehensive semantic segmentation database for labeling. We compared the Baseline SegFormer, Enhanced SegFormer (with positional encoding and segment-based contextual processing), and Context-Aware SegFormer (with a our pre processing) to assess the improvements brought by each enhancement. **Experimental Setup: Data Collection:** We collected a demo dataset, [sidewalk-oct-22], to ensure fair training and evaluation. **Model Training:** Two pre trained SegFormers Models were trained on 4 GPUs (4GB each). Due to computational limits, we focused on exploring possibilities rather than running benchmarks. **Sort Segments:** Various algorithms, including 3D sorting and K-NN, were explored to order segments. We aimed to position segments with shared boundaries closer. We evaluated using MSE and also simulated transformer sequence-to-sequence by first compressing into small tokens with MAE, where the output in place i is the corresponding label of mask i. Attempts to simulate a GPT model for segment classification were unsuccessful, emphasizing the need to include the global image in the input. Also different positional encodings were tested, with the best results from Fourier transform methods and normalized mask pixels. **Score Function:** Evaluated on 100 images with these definitions: **Mask Label:** Parent object label for each segmented sub-object. - **Label Range:** Sequence of masks from the first to last mini image of the mask label. - **Label Score:** Number of different mask labels within the label range. - **Score:** Sum of label scores for all mask labels. **Evaluation:** The function algorithm evaluation(algorithm, images with labels, masks) assesses algorithms using the images with labels dataset, returning the average score. **Evaluation Metrics:** Metrics such as mean Intersection over Union (IoU) and accuracy were used to assess the model's segmentation and classification performance. **SegFormer Integration:** We integrated SegFormer into our segmentation framework, enhancing it with positional encoding and a segment token booster. 1. **Fine-tuning SegFormer:** The model was fine-tuned on our dataset to establish baseline performance.

4 Results

The results of our experiments indicate that the combination of Segment Anything, positional encoding, Vision Transformers, and SegFormer integration provides significant improvements in segmentation and classification tasks but only on part of the labels . Key findings include: **Enhanced accuracy** in segment identification and classification, Might demonstrating the model's ability to generalize across diverse datasets. **Robust performance** across diverse datasets, highlighting the model's could have capability to handle various segmentation tasks at a high scale.

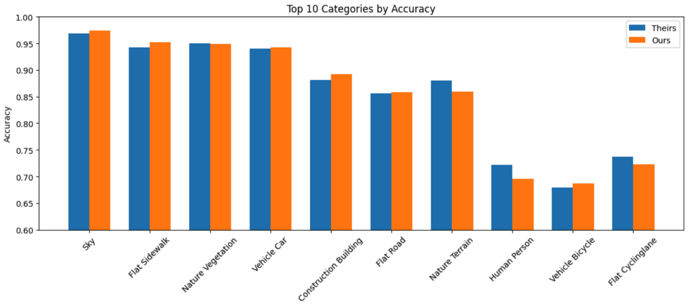

Figure 1: SegFormer semantic segmentation

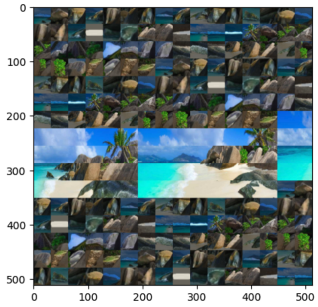

Figure 2: Post segments processing

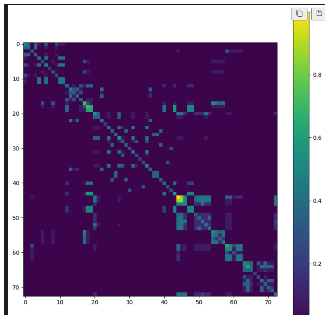

Figure 3: Neighbour masks matrix

Figure 4: Results for post segments processing and neighbour masks matrix

5 Conclusion

Our current work has led to several key insights that will guide our future efforts in improving image segmentation models. One significant finding is the advantage of transforming image segments into patches, specifically in sizes of 8 or 4, to better incorporate the global context of an image. This approach has shown potential in enhancing the model's ability to understand and process complex visual information.

Furthermore, our work have reinforced the importance of including mask information in positional encoding. The findings suggest that the sequential order of segments within the vector does not critically impact performance when the model is exposed to a sufficient variety of examples. This insight simplifies the model architecture by reducing constraints on input ordering, potentially easing the computational load and streamlining the training process.

In the initial testing phases, our model demonstrated improved loss and evaluation metrics over the baseline for the first 10-20 batches. However, this trend did not persist in later tests, where the performance declined, suggesting the possibility of unresolved issues in the implementation . Additional runs are required to ascertain the root cause of this fluctuation and confirm whether it is an artifact of model implementation or indicative of deeper conceptual issues.

6 Summary

Our work demonstrates the effectiveness of integrating advanced segmentation models with positional encoding, Vision Transformers, and SegFormer. Experimental results validate improvements in accuracy and efficiency, highlighting potential applications in computer vision. The integration of segment-based contextual processing and advanced sorting algorithms further enhances model performance for various segmentation and classification tasks.

References

[1] Dosovitskiy, A., et al. (2021). An Image is Worth 16x16 Words:*ICLR*. `https://openreview.net/forum?id=YicbFdNTTy`

[2] Carion, N., et al. (2020). End-to-End Object Detection with Transformers. *ECCV*. `https://arxiv.org/abs/2005.12872`

[3] Chen, Y., et al. (2021). Pre-Trained Image Processing Transformer. *CVPR*. `https://arxiv.org/abs/2012.00364`

[4] Wang, W., et al. (2022). Pyramid Vision Transformer. *CVPR*. `https://arxiv.org/abs/2102.12122`

[5] Liu, Z., et al. (2021). Swin Transformer. *ICCV*. `https://arxiv.org/abs/2103.14030`

[6] Zhu, X., et al. (2021). Deformable DETR Detection. *ICLR*. https://arxiv.org/abs/2010.04159

[7] Sun, K., et al. (2021). Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation. *arXiv:2105.05537*. https://arxiv.org/abs/2105.05537

[8] Huang, Y., et al. (2023). InterFormer *ICCV*, pp. 22301-22311. https://openaccess.thecvf.com/content/ICCV2023/html/Huang_InterFormer_Real-time_Interactive_Image_Segmentation_ICCV_2023_paper.html

[9] Li, X., et al. (2023). ACSeg *CVPR*. https://openaccess.thecvf.com/content/CVPR2023/papers/Li_ACSeg_Adaptive_Conceptualization_for_Unsupervised_Semantic_Segmentation_CVPR_2023_paper.pdf

[10] Peng, B., et al. (2023) *CVPR*. https://openaccess.thecvf.com/content/CVPR2023/papers/Peng_Hierarchical_Dense_Correlation_Distillation_for_Few-Shot_Segmentation_CVPR_2023_paper.pdf

[11] Devlin, J., et al. (2018). BERT: Pre-training of DPT . *arXiv:1810.04805*. https://arxiv.org/abs/1810.04805

7 Future Work(Part B project proposal)

In future iterations of our work will refine and expand our segmentation framework to address more complex challenges in computer vision. A key focus will be the integration of innovative masking techniques, specifically designed to enhance model robustness and learning efficiency. We plan to implement random masking of image patches, which will force the model to learn from a reduced set of data and infer missing information, thereby increasing its predictive capabilities under constrained conditions.

Additionally, we intend to explore the handling of multiple images in a single input batch. This approach will challenge the model to manage and interpret complex images containing numerous elements simultaneously, further testing its capacity to handle real-world variability in visual data. Another might interesting direction involves leveraging state-of-the-art models to process images sequentially. By applying these advanced models to each image individually and then integrating them into a context model arranged in a 16 (4x4) as grid of 16 images, we can calculate losses and compare performance against these cutting-edge benchmarks. This will not only test the efficiency of our segment-based approach but also refine the way segments are incorporated into broader image analysis frameworks, and we suppose out model can use it as stronger base line even then the actual models if the model its self will work properly.

Finally, if we find the points above efficiency will be interesting to advance our methodology by incorporating context information early in the processing stages. The early integration of context will might allow the model to utilize foundational image understanding to enhance further processing, potentially

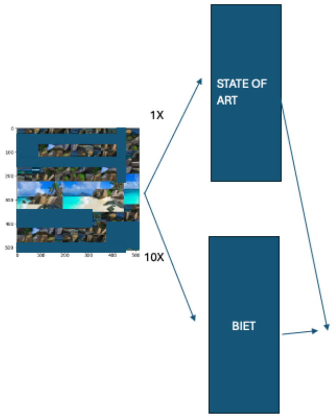

Figure 5: RL

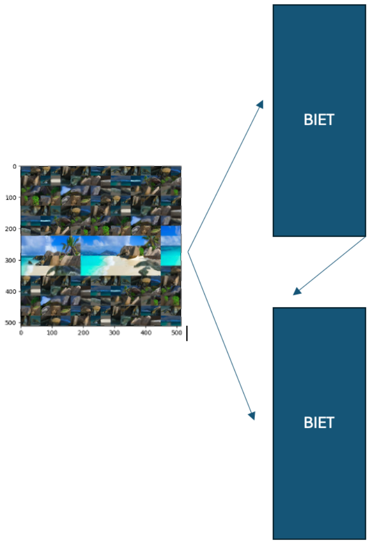

Figure 6: Context feed

leading to more accurate classifications and better handling of complex image structures. These could be implemented by process the image into to BIET one is already context trained the other will get the feed forward and this will could be by taking the baseline against it but the model context its already trained and the two other pre-trained .