SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers

Progress Report

2025 - 08 - 24

CS4681 : Advanced Machine Learning
Index No : 210372D
Manawathilake K.C.K.

1. Table of Contents

2. List of Acronyms

MLP	Multi Layer Perceptron
FCN	Fully Convolutional Network
CNN	Convolutional Neural Network
VGG	Visual Geometry Group (CNN architecture)
NLP	Natural Language Processing
ViT	Vision Transformer
PE	Positional Embeddings
mIoU	Mean Intersection over Union

3. Introduction

There has been a growing interest in image classification, followed by semantic segmentation. Starting from basic image classification models like VGGs and then using the backbone of image classification models for semantic segmentation, the field has gained significant attention. After the success of transformers in NLP, there was also huge interest in applying transformers to vision tasks. ViT for image classification was born from these efforts. Splitting an image into multiple linear embedding patches and feeding them into a standard transformer with positional embeddings led to impressive performance on the ImageNet dataset.

However, even these had limitations. ViT outputs single-scale low-resolution features instead of multi-scale ones, unlike generic CNNs, and it also has high computational costs for large images. Several architectures have been proposed to address these issues, such as Swin Transformers [1] and Twins [2].

SegFormer [3] is a similar architecture introduced with a focus on efficiency, accuracy, and robustness, mainly due to its lightweight all-MLP decoder, which yields impressive results over several renowned datasets, surpassing many state-of-the-art models. They introduced six models, namely SegFormer-B0 (the lightest model with 3.7M parameters) to B5 (with 84.7M parameters), all of which have achieved impressive mIoU compared to other state-of-the-art models.

4. Literature Review

Vision Transformers

A transformer, at its core, is a combination of an encoder and a decoder. The encoder takes a set of tokens and creates contextualized embeddings for them using self-attention and feed-forward layers, and the decoder regenerates another sequence of tokens using the input context. In the domain of vision transformers, this becomes a computationally intensive process, as an image needs to be split into multiple linearly embedded patches and fed with positional embeddings. For an image, this results in a huge amount of computation and was initially considered uncertain in terms of feasibility with the available computational power.

Transformer Backbones

ViT is the first work to prove that a pure transformer can achieve state-of-the-art performance in image classification. Subsequently, methods such as DeiT [4], CPVT [5], TNT, LocalViT, and CrossViT were introduced to enhance this methodology and obtain better-performing image classifiers. PVT [6] was the first to introduce a pyramid structure in transformers, followed by methods such as Swin [1], CvT [7], CoaT, and Twins, which enhanced dense prediction capabilities. These backbones were crucial for various tasks such as tracking, super-resolution,

and colorization. SETR [8] adopts ViT as a backbone to extract features, achieving impressive performance in the domain of semantic segmentation.

SegFormer

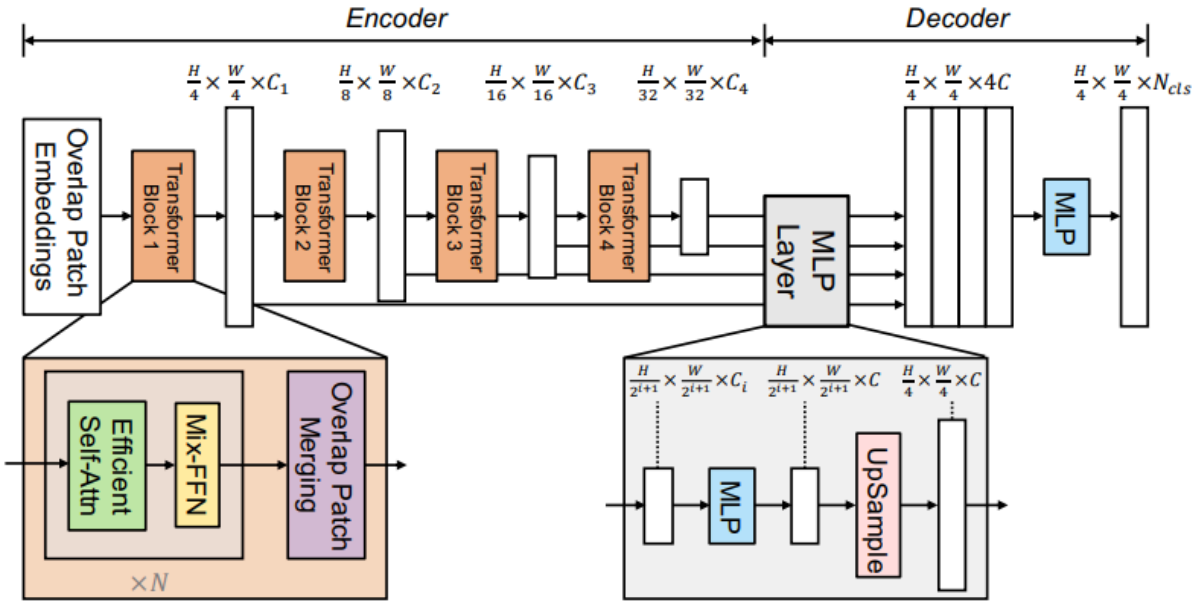

Figure 1: The proposed SegFormer framework.

SegFormer is the proposed architecture for semantic segmentation using ViT backbones. This contains two main modules:

1. A hierarchical transformer encoder to generate high-resolution coarse features and low-resolution fine features.
2. A lightweight all-MLP decoder to fuse these multi-level features to produce the final semantic segmentation mask.

This architecture uses hierarchical transformer encoders to extract features. Starting with divided patches of size 4×4 in each layer, the image size is reduced to half of its length and breadth when forming the features. This allows the model to generate both high-resolution fine features and low-resolution coarse features. Overall, these feature maps are reduced by 2× in each model, from B5 to B0, in order to reduce complexity and improve efficiency.

In the model, in order to improve self-attention, a few layers are used to reshape the Keys in the attention module, reducing the complexity from O(N2) to O(N2/R) (where R=[64,16,4,1] from stage 1 to stage 4).

By introducing a Mix-FFN, this framework eliminates the need for positional encoding (PE). Having a fixed resolution (as in traditional ViT models) decreases accuracy when the testing resolution differs from the training resolution, as the positional codes need to be interpolated.

Unlike other methods used by other models, such as using a 3×3×3 convolution with PE, SegFormer argues that positional encoding is not actually necessary for semantic segmentation. Instead, they introduce Mix-FFN (Mix-Feed-Forward Network), which provides positional information by directly incorporating a 3x3×3 convolution within the feed-forward network (FFN).

Another important architectural implementation proposed by the model is the use of a lightweight all-MLP decoder. It uses a linear layer, followed by upsampling layers, and then a couple of linear layers to finally predict the mask. This allows it to maintain a less computationally demanding method. This efficiency is possible due to the larger effective receptive field of this model compared to other CNN encoders.

Training

- DataSets used
    - ADE20K
    - Cityscapes
    - COCO-Stuff
- Data augmentations
    - Random resizing with ratio 0.5 - 2.0
    - Random horizontal flipping
    - Random cropping to 512x512, 1024x1024
- Optimizer
    - AdamW
- Iterations
    - 160K Iterations for ADE20K, Cityscapes
    - 80K iterations on COCO-Stuff
- Batch size
    - 16 for ADE20K and COCO-stuff
    - 8 for Cityscapes
- Learning Rate
    - Initial value : 0.00006
    - Learning rate schedule : “poly” LR schedule with factor 1.0 (default)

Ablation Studies

1. Influence of the size of the model: Increasing the size of the encoder yielded consistent improvements across all datasets.
2. Influence of C, the MLP decoder channel dimension: Even though performance increased as C increased, this led to larger and less efficient models. Therefore, a value of C=256 was chosen, as it provided competitive performance with reasonable computational cost.
3. Mix-FFN vs. PE: Using Mix-FFN clearly outperformed positional encoding.

4. Effective Receptive Field Evaluation: CNN-based encoders yielded significantly lower accuracy compared to coupling with the proposed transformer encoder. This result is intuitive, as CNNs have a smaller receptive field than transformers.

Robustness

The model was tested on the same dataset with 16 types of algorithmically generated corruptions, including noise, blur, weather, and digital categories. As expected, it was less accurate compared to the original dataset, but its performance was still significantly better than other models such as DeepLabv3+, MobileNetV2, and ResNet.

Other Improvements

SegFormer has been widely adopted as a baseline for semantic segmentation tasks due to its efficiency and lightweight design.

Chen et al. [9] proposed an improved SegFormer approach to achieve enhanced segmentation of photovoltaic (PV) arrays in infrared images, demonstrating superior accuracy compared to the vanilla SegFormer. This was achieved through an inception-enhanced attention mechanism and multi-scale spatial feature extraction to address problems such as segmentation holes and environmental misclassification. Furthermore, the use of a feature pyramid network and bilinear interpolation to enhance the completeness of edge details were notable changes made in the suggested implementation.

Kienzle et al. [10] proposed that the SegFormer architecture incorporates a MixTransformer encoder, a lightweight convolutional decoder, and a token-merging strategy that reduces computational complexity by merging similar tokens instead of pruning. This enhances efficiency without requiring model re-training, thus improving performance on various tasks.

Bai et al. [11] proposed designing a lightweight SegFormer for efficient semantic segmentation. Based on the observation that neurons in SegFormer layers exhibit large variances across different images, they introduced a dynamic gated linear layer, which prunes the most uninformative set of neurons based on the input instance. To improve the dynamically pruned SegFormer, they also introduced two-stage knowledge distillation to transfer knowledge from the original teacher network to the pruned student network.

5. Proposed Methodology

Based on the literature review I have conducted and the availability of resources, there are several possible methods to improve the model's performance. It is important to note that, based on the available computing power and resources, there might be issues with training the larger models. In that case, I will have to focus mainly on the lightweight models, depending on

the resources available during implementation. The following are several possible implementations that can be applied to the given model:

In the paper, the authors specifically mentioned that they have not experimented with widely-used techniques such as OHEM, auxiliary losses, or class-balanced loss. There is a possibility to explore such methods in the implementation, with slightly adjusted hyperparameter values, to improve the performance of the model.

Changes in the decoder MLP layer can also be considered to capture certain domain-specific information, or architectural modifications can be made to the decoder layer to improve performance while reducing the complexity of the model.

6. Project Timeline

Task Id	Task Name	Week (Starting Date and Academic Week No.)					
		2025-08-18	2025-08-25	2025-09-08	2025-09-15	2025-09-22	2025-09-29
		1 (6)	2 (7)	3 (8)	4 (9)	5 (10)	6 (11)
1	Literature Review						
2	Setting up the Environment						
3	Researching on Improvements						
4	Implementation						
5	Finalizing the Repository						
6	Writing the Conference Paper						

7. Progress

So far, I have gone through the provided paper and conducted thorough research on neural network basics, including transformer architectures and vision transformers, and have understood the concepts well. I have also reviewed other similar implementations and extensions of SegFormer, gaining insight into the potential and limitations of current approaches. I have currently created a project schedule and am working on its timeline. As planned, I will be working on setting up the working environment and loading datasets in the upcoming week. In the meantime, I am researching possible modifications that I can implement to improve this model.

8. References

[1] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. arXiv, 2021.

[2] Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haibing Ren, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Twins: Revisiting spatial attention design in vision transformers. arXiv, 2021.

[3] Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., & Luo, P. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. arXiv, 2021.

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-End object detection with transformers. In ECCV, 2020.

[5] Xiangxiang Chu, Zhi Tian, Bo Zhang, Xinlong Wang, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Conditional positional encodings for vision transformers. arXiv, 2021.

[6] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. arXiv, 2021.

[7] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. arXiv, 2021.

[8] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. CVPR, 2021.

[9] W. Chen, S. Jin, Y. Luo and J. Li, "Enhanced Segmentation of PV Arrays in Infrared Images using an Improved SegFormer Approach," 2023 IEEE International Conference on Big Data (BigData), Sorrento, Italy, 2023

[10] D. Kienzle, M. Kantonis, R. Schön and R. Lienhart, "Segformer++: Efficient Token-Merging Strategies for High-Resolution Semantic Segmentation," 2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR), San Jose, CA, USA, 2024

[11] H. Bai, H. Mao and D. Nair, "Dynamically Pruning Segformer for Efficient Semantic Segmentation," ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, Singapore, 2022