SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers

Problem they are trying to solve / Purpose of method

They aim to design a simple and efficient transformer-based model for semantic segmentation, avoiding the complexity and inefficiency of earlier methods.

Traditional segmentation models, especially CNN-based ones, often rely on complex decoder structures or hand-crafted components like dilated convolutions and CRFs to improve performance. While transformer-based models such as SETR and Segmenter capture global context well, they tend to suffer from **high computational cost**, **slow inference**, and **large memory usage**, due to their reliance on **positional embeddings** and **heavy transformer decoders**.

SegFormer addresses these issues with a **faster and more efficient design**. It solves the key challenges of **computational cost**, **scalability**, and **flexibility** in vision transformers.

The goal of SegFormer is to introduce a model that is:

1. **Simple** – no need for complex decoders or positional encodings,
2. **Efficient** – optimized for both computation and memory usage,
3. **Flexible** – performs well across various datasets and image resolutions.

How does it differ from other methods?

SegFormer differs from other segmentation methods in three key ways:

1. **No positional encodings:**
   - Unlike other vision transformers, SegFormer does not use positional encodings. Instead, it relies on overlapping patch embeddings and hierarchical representations, which are sufficient to retain spatial information.
2. **Lightweight MLP decoder:**
   - Instead of using a heavy transformer decoder or complex upsampling modules, SegFormer uses a simple multilayer perceptron (MLP) head to fuse features from different stages of the encoder.
3. **Hierarchical transformer encoder:**
   - SegFormer adopts a hierarchical encoder based on Mix Vision Transformer (MiT), which captures both local and global features efficiently. This is more similar to CNN-like pyramidal processing than flat ViT structures.

These design choices make SegFormer more efficient and scalable while maintaining or exceeding state-of-the-art performance across benchmarks.

How the method works

SegFormer consists of two key parts:

1. A **Mix Vision Transformer (MiT) encoder**,

2. A **lightweight MLP decoder**.

Together, they provide strong hierarchical representations while keeping the model fast and scalable.

Key architectural innovations:

1. **Spatial-Reduction Attention (SRA):**
   Reduces the number of tokens involved in self-attention, lowering computational cost while preserving accuracy.

2. **Mix-Feedforward Network (Mix-FFN):**
   Combines MLP layers with depthwise convolutions. This replaces traditional positional encoding and improves generalization across varying image sizes.

3. **Overlapping Patch Merging:**

   - Instead of splitting the image into non-overlapping patches (like ViT), SegFormer uses overlapping patches to preserve spatial continuity and improve segmentation accuracy.

Decoder:

A simple MLP head takes multi-scale features from the encoder, upsamples them to the same resolution, and fuses them.
Unlike other models, SegFormer avoids any complex decoder structures or heavy upsampling modules.