Note:Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers

Sinkoo

April 3, 2022

1 Abstract

Most recent semantic segmentation methods adopt a FCN with an encoder-decoder architecture, and concentrate on increasing receptive field.And this approach is still limited in learning dependency in distance because of receptive fields. This article provides an alternative perspective by treating semantic segmentation as a sequence-to-sequence prediction task (pure transformer).The paper provided a new segmentation model "SEgmentation TRansformer (SETR)".

2 Method

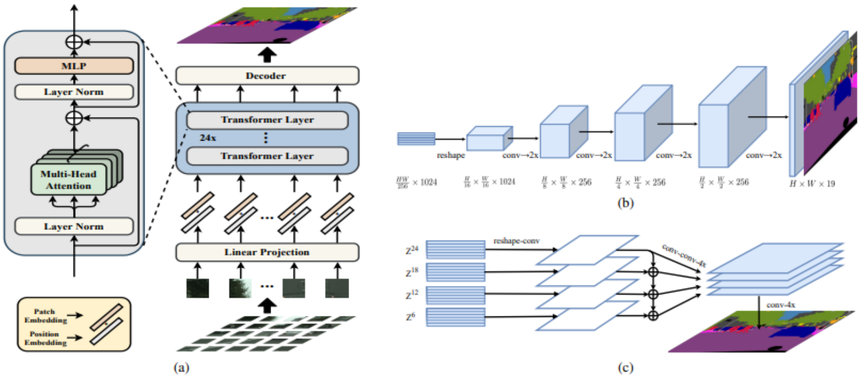


Figure 1. **Schematic illustration of the proposed *SEgmentation TRansformer* (SETR)** (a). We first split an image into fixed-size patches, linearly embed each of them, add position embeddings, and feed the resulting sequence of vectors to a standard Transformer encoder. To perform pixel-wise segmentation, we introduce different decoder designs: (b) progressive upsampling (resulting in a variant called SETR-*PUP*); and (c) multi-level feature aggregation (a variant called SETR-*MLA*).

Image to sequence Transform 2D image to 1D sequence.Purely flatten image (pixel sequence) is out of the question.

SETR divides image($H*W*3$) into a grid of $\frac{H}{16}*\frac{W}{16}$ patches, and then flatten this grid into a sequence. Further mapping each vectorized patch p into a latent C-dimensional embedding space using a linear projection function $f : p \rightarrow e \in R^C$ to get a 1D sequence of patch embeddings for image x. Encoding the patch spacial information, we learn a specific embedding p_i for every location i which is added to e_i to form the final sequence input$E = \{e_1 + p_1, e_2 + p_2, ..., e_L + p_L\}$.

Transformer Use a pure transformer encoder, each transformer layer has a global receptive field.

Decoder designs

Naive upsampling(Naive) Project transformer feature to the dimension of class number. Then bilinearly upsample the output to the full image resolution.

Progressive UPsampling(PUP) Progressive upsampling strategy that alternates conv layers and upsampling operations. Restrict upsampling to 2*;

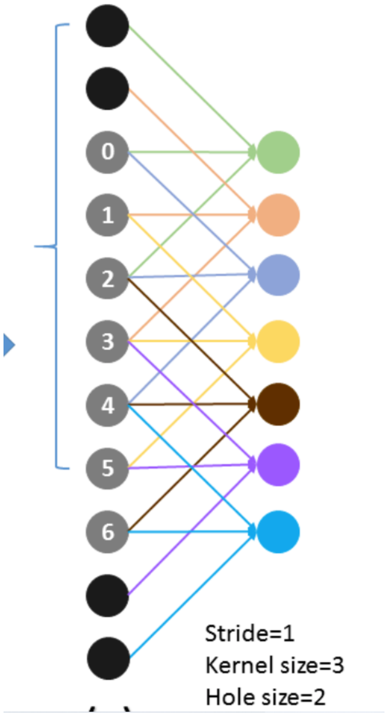
Multi-Level feature Aggregation(MLA) (?)

3 Summary

This paper provides a new way of semantic segmentation with transformers and get rid of the problem with receptive field in traditional FCN, and make to state-of-the-art.
And approach of transform image into sequence and then implement in transform my be helpful in other fields to make use of efficiency of Transformers.

4 Other Knowledge

1, atrous convolution: Based on the idea that neighboring pixels are almost the same, think as redundancy.


0
1
2
3
4
5
6
Stride=1
Kernel size=3
Hole size=2