Transformer based semantic segmentation

6 Apr 2020

Yonk shi

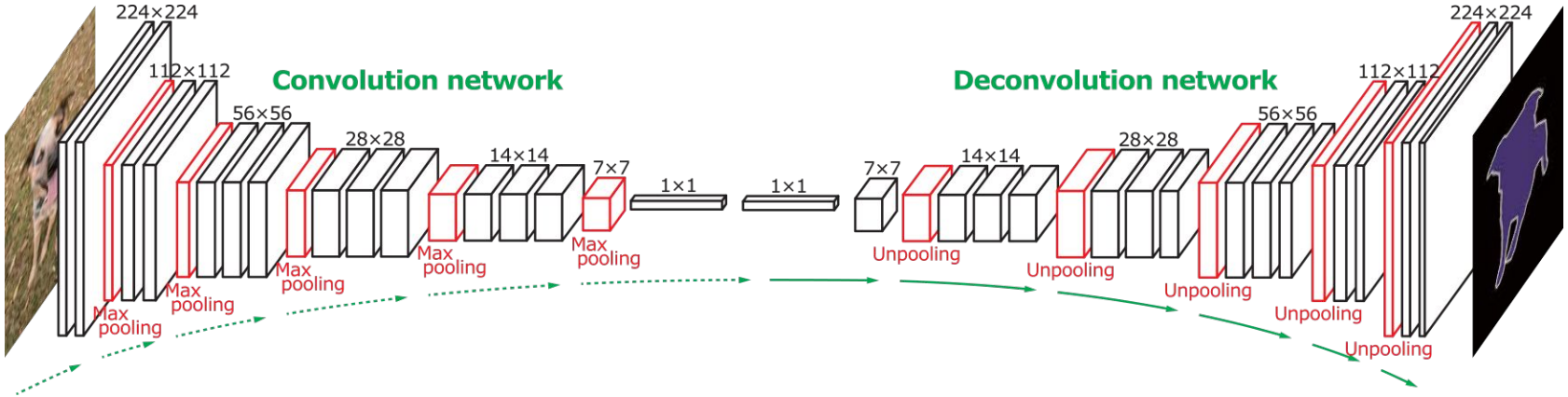
A presentation on: Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers
https://arxiv.org/pdf/2012.15840.pdf

Background: What is semantic segmentation?

Source: http://cvlab.postech.ac.kr/research/deconvnet/

Background

Most current semantic segmentation approaches are based on Fully Convolutional Network (FCN). For example:



Source: http://cvlab.postech.ac.kr/research/deconvnet/

Limitations with FCN based approaches

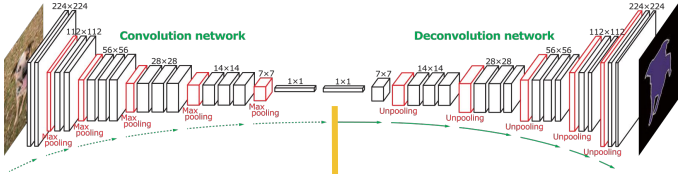
- Unable to capture very long range relationships.
- Maximum scale limited by receptive field

Problems with FCN:

1. Long range relationship

Problems with FCN:

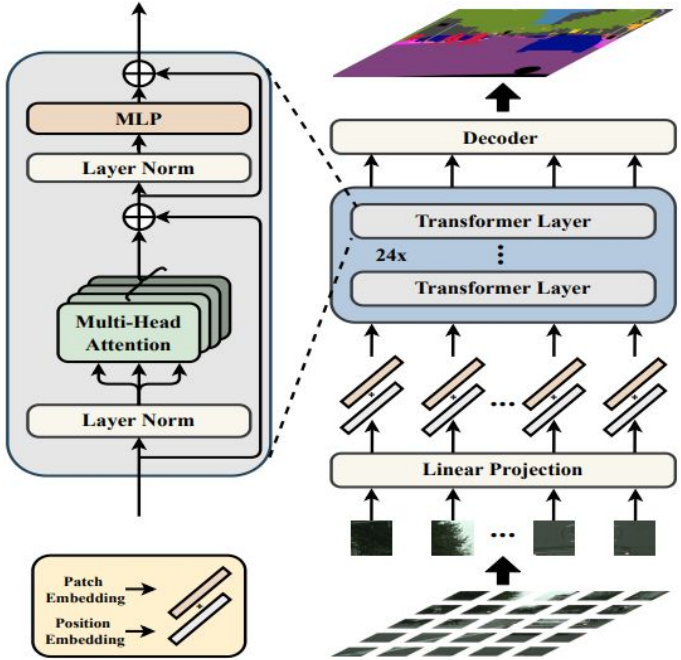
2. Maximum scale limited by receptive field



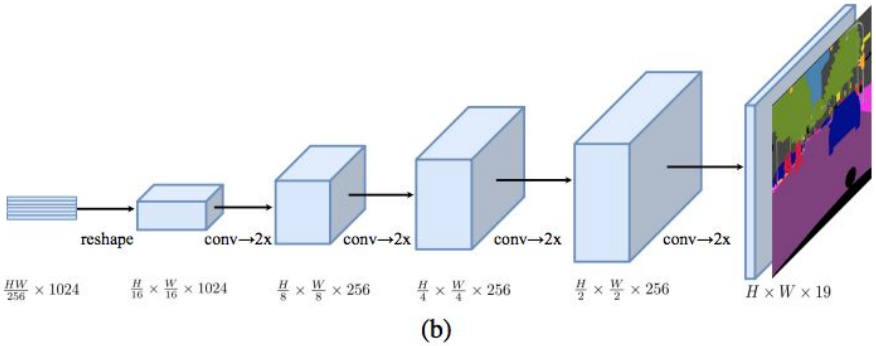
SEgmentation TRansformer (SETR) in a nutshell:

- Uses standard Transformer module
- No downsampling* of images, maintaining dense pixel info through all layers
- Slice image into small patches, flatten into 1-d vector for transformer encoder
- Several flavors of decoders

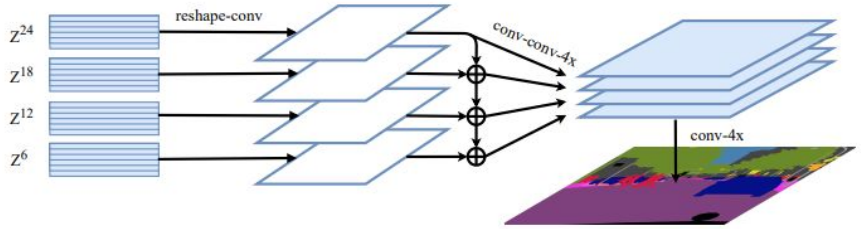
A sequence to sequence to approach to semantic segmentation - SEgmentation TRansformer (SETR)



Decoders


reshape
conv→2x
conv→2x
conv→2x
conv→2x
$\frac{HW}{256} \times 1024$
$\frac{H}{16} \times \frac{W}{16} \times 1024$
$\frac{H}{8} \times \frac{W}{8} \times 256$
$\frac{H}{4} \times \frac{W}{4} \times 256$
$\frac{H}{2} \times \frac{W}{2} \times 256$
$H \times W \times 19$

(b)


z^{24}
z^{18}
z^{12}
z^{6}
reshape-conv
conv-conv-4x
conv-4x

(c)

Results (qualitative)

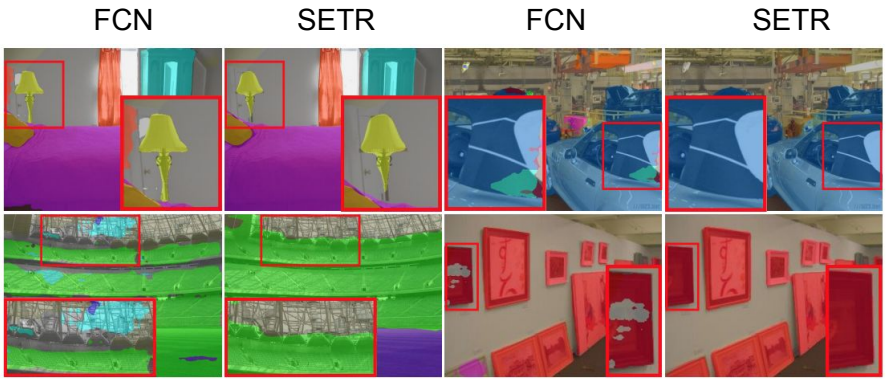


Figure 2. **Qualitative results on ADE20K:** SETR (right column) vs. dilated FCN baseline (left column) in each pair. Best viewed in color and zoom in.

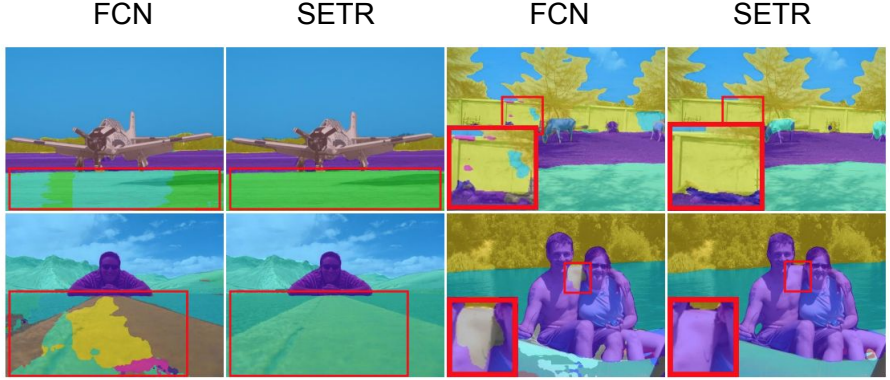


Figure 3. **Qualitative results on Pascal Context:** SETR (right column) vs. dilated FCN baseline (left column) in each pair. Best viewed in color and zoom in.

Results

Method	Backbone	mIoU	Pixel Acc.
FCN (16, 160k, SS) [39]	ResNet-101	39.91	79.52
FCN (16, 160k, MS) [39]	ResNet-101	41.40	80.65
EncNet [54]	ResNet-101	44.65	81.69
PSPNet [59]	ResNet-269	44.94	81.69
DMNet [18]	ResNet-101	45.50	-
CCNet [25]	ResNet-101	45.22	-
Strip pooling [23]	ResNet-101	45.60	82.09
APCNet [19]	ResNet-101	45.38	-
OCNet [53]	ResNet-101	45.45	-
SETR-*Naïve* (16, 160k, SS)	T-Large	48.06	82.40
SETR-*Naïve* (16, 160k, MS)	T-Large	48.80	82.92
SETR-*PUP* (16, 160k, SS)	T-Large	48.58	82.90
SETR-*PUP* (16, 160k, MS)	T-Large	50.09	**83.58**
SETR-*MLA* (16, 160k, SS)	T-Large	48.64	82.64
SETR-*MLA* (16, 160k, MS)	T-Large	**50.28**	83.46

Table 4. **State-of-the-art comparison on the ADE20K dataset.** Performances of different model variants are reported. SS: Single-scale inference. MS: Multi-scale inference.

Method	Backbone	mIoU
FCN (16, 80k, SS) [39]	ResNet-101	44.47
FCN (16, 80k, MS) [39]	ResNet-101	45.74
PSPNet [59]	ResNet-101	47.80
DANet [17]	ResNet-101	52.60
EMANet [31]	ResNet-101	53.10
SVCNet [15]	ResNet-101	53.20
Strip pooling [23]	ResNet-101	54.50
GFFNet [30]	ResNet-101	54.20
APCNet [19]	ResNet-101	54.70
SETR-*Naïve* (16, 80k, SS)	T-Large	52.89
SETR-*Naïve* (16, 80k, MS)	T-Large	53.61
SETR-*PUP* (16, 80k, SS)	T-Large	54.40
SETR-*PUP* (16, 80k, MS)	T-Large	55.27
SETR-*MLA* (16, 80k, SS)	T-Large	54.87
SETR-*MLA* (16, 80k, MS)	T-Large	**55.83**

Table 5. **State-of-the-art comparison on the Pascal Context dataset.** Performances of different model variants are reported. SS: Single-scale inference. MS: Multi-scale inference.

Results (qualitative)

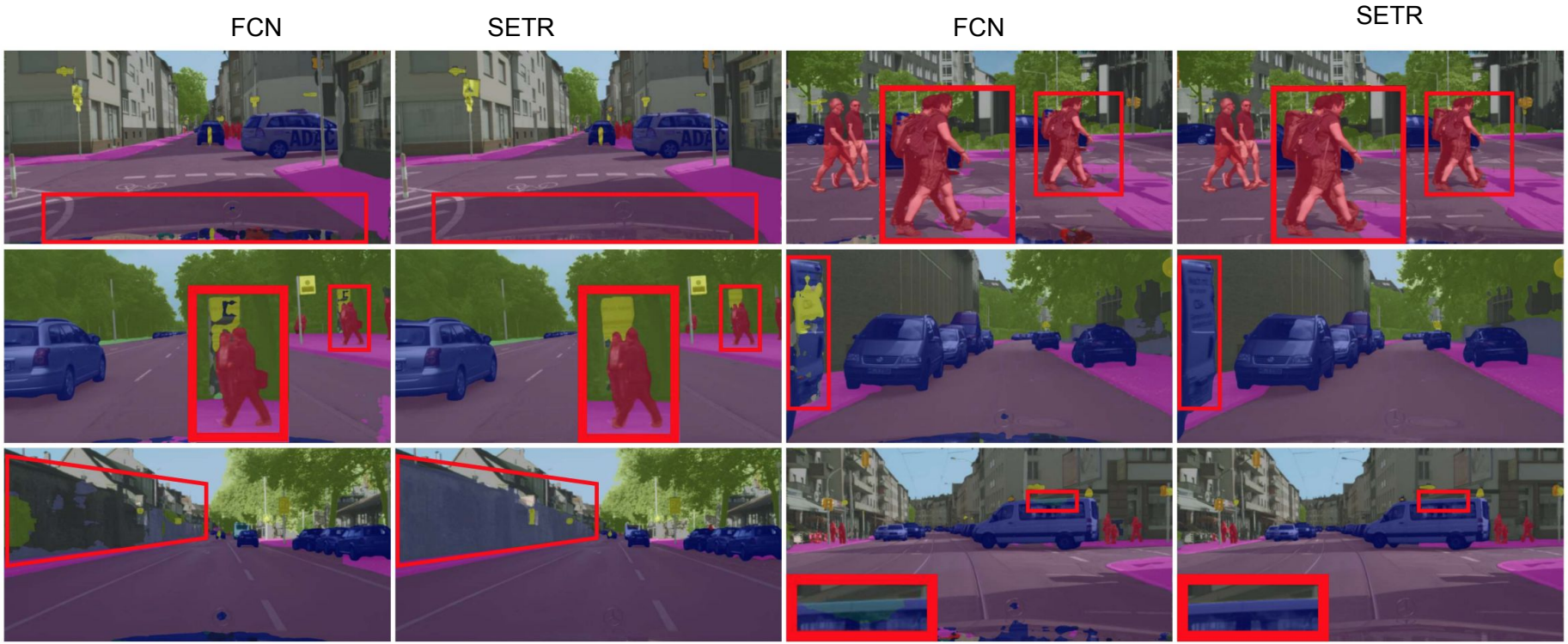


Figure 4. **Qualitative results on Cityscapes:** SETR (right column) vs. dilated FCN baseline (left column) in each pair. Best viewed in color and zoom in.

Results

Method	Backbone	mIoU
FCN (40k, SS) [39]	ResNet-101	73.93
FCN (40k, MS) [39]	ResNet-101	75.14
FCN (80k, SS) [39]	ResNet-101	75.52
FCN (80k, MS) [39]	ResNet-101	76.61
PSPNet [59]	ResNet-101	78.50
DeepLab-v3 [10] (MS)	ResNet-101	79.30
NonLocal [48]	ResNet-101	79.10
CCNet [25]	ResNet-101	80.20
GCNet [4]	ResNet-101	78.10
Axial-DeepLab-XL [47] (MS)	Axial-ResNet-XL	81.10
Axial-DeepLab-L [47] (MS)	Axial-ResNet-L	81.50
SETR-*PUP* (40k, SS)	T-Large	78.39
SETR-*PUP* (40k, MS)	T-Large	81.57
SETR-*PUP* (80k, SS)	T-Large	79.34
SETR-*PUP* (80k, MS)	T-Large	**82.15**

Table 6. **State-of-the-art comparison on the Cityscapes validation set.** Performances of different training schedules (*e.g.*, 40k and 80k) are reported. SS: Single-scale inference. MS: Multi-scale inference.

Caveats & discussion

- This approach has an order of magnitude more parameters than FCN based approaches
- Randomly initialized SETR performs very poorly (hinting at many local minimas?)
- Decoder is depend on image size - less flexible

Model	T-layers	Hidden size	Att head
T-Base	12	768	12
T-Large	24	1024	16

Table 1. Configuration of Transformer backbone variants.

Method	Pre	Backbone	#Params	40k	80k
FCN [39]	1K	R-101	68.59	73.93	75.52
Semantic FPN [39]	1K	R-101	47.51	-	75.80
Hybrid-Base	R	T-Base	112.59	74.48	77.36
Hybrid-Base	21K	T-Base	112.59	76.76	76.57
Hybrid-DeiT	21K	T-Base	112.59	77.42	78.28
SETR-*Naïve*	21K	T-Large	305.67	77.37	77.90
SETR-*MLA*	21K	T-Large	310.57	76.65	77.24
SETR-*PUP*	21K	T-Large	318.31	78.39	79.34
SETR-*PUP*	R	T-Large	318.31	42.27	-
SETR-*Naïve-Base*	21K	T-Base	87.69	75.54	76.25
SETR-*MLA-Base*	21K	T-Base	92.59	75.60	76.87
SETR-*PUP-Base*	21K	T-Base	97.64	76.71	78.02
SETR-*Naïve-DeiT*	1K	T-Base	87.69	77.85	78.66
SETR-*MLA-DeiT*	1K	T-Base	92.59	78.04	78.98
SETR-*PUP-DeiT*	1K	T-Base	97.64	**78.79**	**79.45**


Thank you!