SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation

PR-SMARCLE

김찬영

SMARCLE

Content

- Introduction to SegNet
- Architecture – Encoder
- Architecture – Decoder
- Max-Pooling Indices
- Training
- Results
- Conclusion

SMARCLE

Introduction to SegNet

- Scene Segmentation을 위한 네트워크
- encoder단에 VGG16을 사용
- 메모리 사용량과 연산 속도 측면에서 효율적

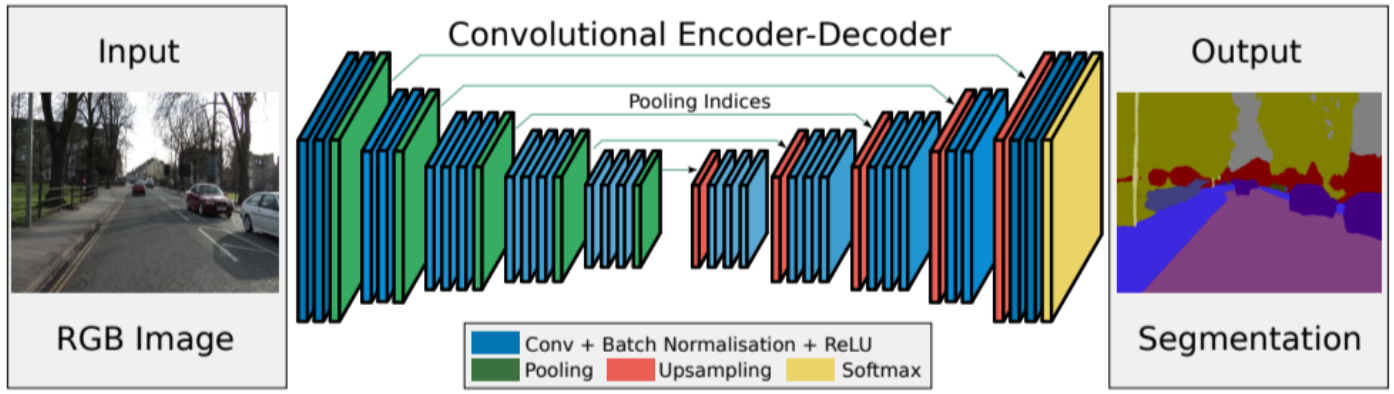


Fig. 2. An illustration of the SegNet architecture. There are no fully connected layers and hence it is only convolutional. A decoder upsamples its input using the transferred pool indices from its encoder to produce a sparse feature map(s). It then performs convolution with a trainable filter bank to densify the feature map. The final decoder output feature maps are fed to a soft-max classifier for pixel-wise classification.

SMARCLE

Badrinarayanan, Vijay, Alex Kendall, and Roberto Cipolla. "Segnet: A deep convolutional encoder-decoder architecture for image segmentation." *IEEE transactions on pattern analysis and machine intelligence* 39.12 (2017): 2481-2495.

Architecture – Encoder

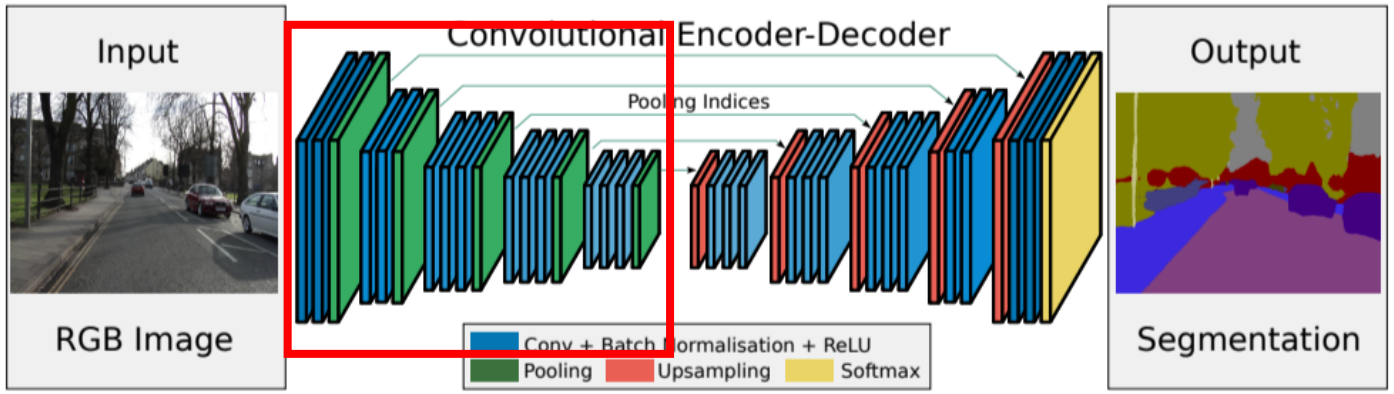


Fig. 2. An illustration of the SegNet architecture. There are no fully connected layers and hence it is only convolutional. A decoder upsamples its input using the transferred pool indices from its encoder to produce a sparse feature map(s). It then performs convolution with a trainable filter bank to densify the feature map. The final decoder output feature maps are fed to a soft-max classifier for pixel-wise classification.

- VGG16에서 FCL를 제외한 구조를 채택
- Max-Pooling Indices Storing 수행

- Conv + BN + ReLU
  - Conv는 ImageNet에서 학습된 VGG의 13개의 Convolutional Layer를 사용

- Max-Pooling
  - Kernel Size : 2x2, stride = 2

SMARCLE

Badrinarayanan, Vijay, Alex Kendall, and Roberto Cipolla. "Segnet: A deep convolutional encoder-decoder architecture for image segmentation." *IEEE transactions on pattern analysis and machine intelligence* 39.12 (2017): 2481-2495.

Architecture – Decoder

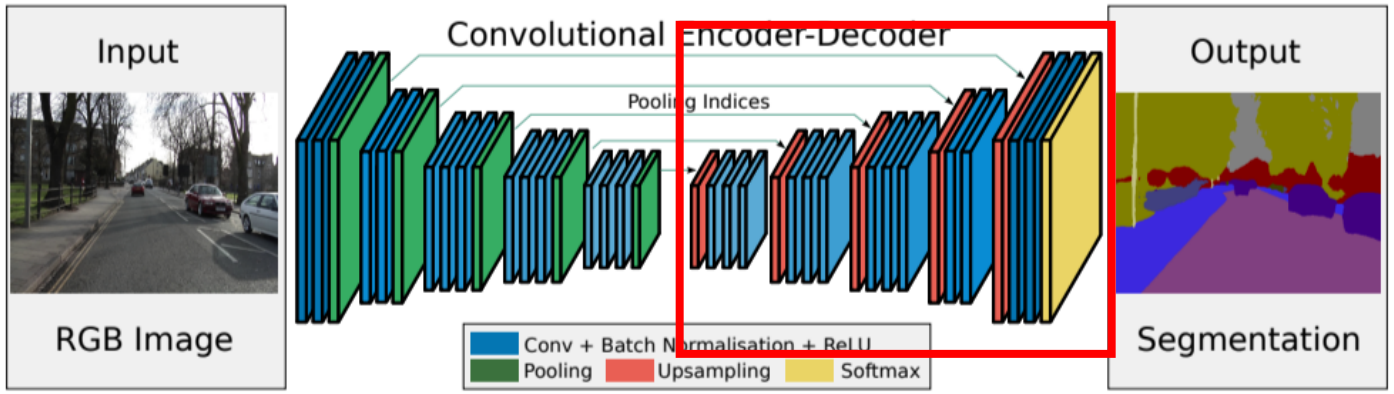


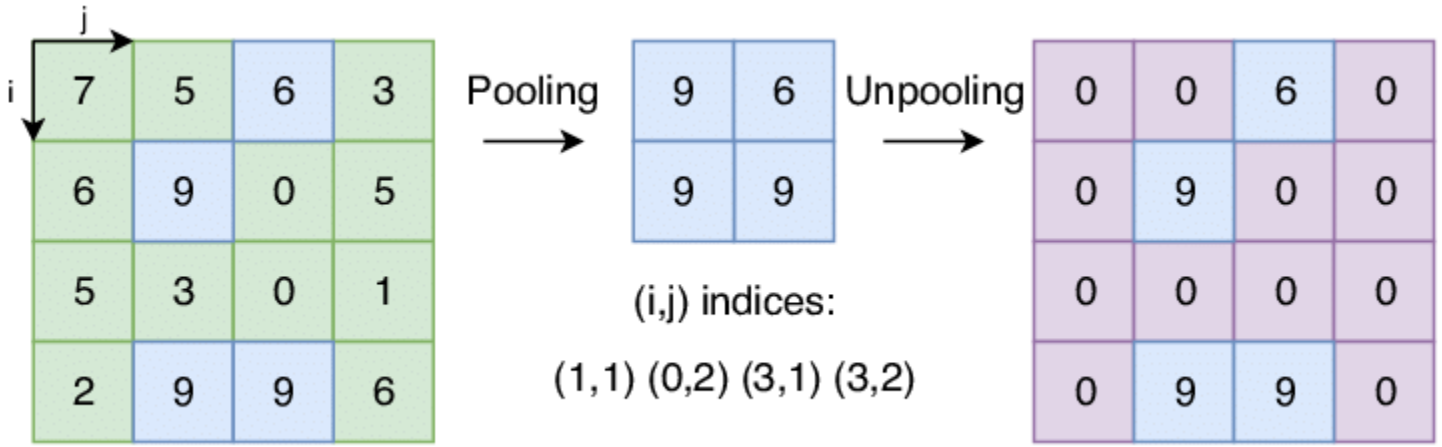
Fig. 2. An illustration of the SegNet architecture. There are no fully connected layers and hence it is only convolutional. A decoder upsamples its input using the transferred pool indices from its encoder to produce a sparse feature map(s). It then performs convolution with a trainable filter bank to densify the feature map. The final decoder output feature maps are fed to a soft-max classifier for pixel-wise classification.

- Encoder에서 저장한 Max-Pooling Indices를 사용해 upsampling을 진행
- 마지막에 픽셀별 K-Class Softmax 분류기

SMARCLE

Badrinarayanan, Vijay, Alex Kendall, and Roberto Cipolla. "Segnet: A deep convolutional encoder-decoder architecture for image segmentation." *IEEE transactions on pattern analysis and machine intelligence* 39.12 (2017): 2481-2495.

Max-Pooling Indices



- Max-Pooling 할 때의 위치 정보만 저장 후 unpooling때 해당 위치 정보에 맞춰 upsampling
- Feature map을 저장하는 것이 아니기 때문에 memory cost가 많이 발생하지 않음
- Boundary에 대한 정확도 향상, parameter 감소를 통한 효율성

SMARCLE

Badrinarayanan, Vijay, Alex Kendall, and Roberto Cipolla. "Segnet: A deep convolutional encoder-decoder architecture for image segmentation." *IEEE transactions on pattern analysis and machine intelligence* 39.12 (2017): 2481-2495.

Training

- Training Dataset : CamVid Road Scenes
- Weight initialization : He initialization (encoder와 decoder 모두)
- Optimizer : SGD with momentum 0.9
- Learning rate : 0.1
- Batch size : 12
- Loss function : cross-entropy
- Class balancing : median frequency balancing
  - 학습 데이터에서 pixel의 대부분을 차지하는 class들에는 1보다 작은 weight를 주고 그렇지 않은 class들에는 1보다 큰 weight를 주어 loss를 계산

SMARCLE

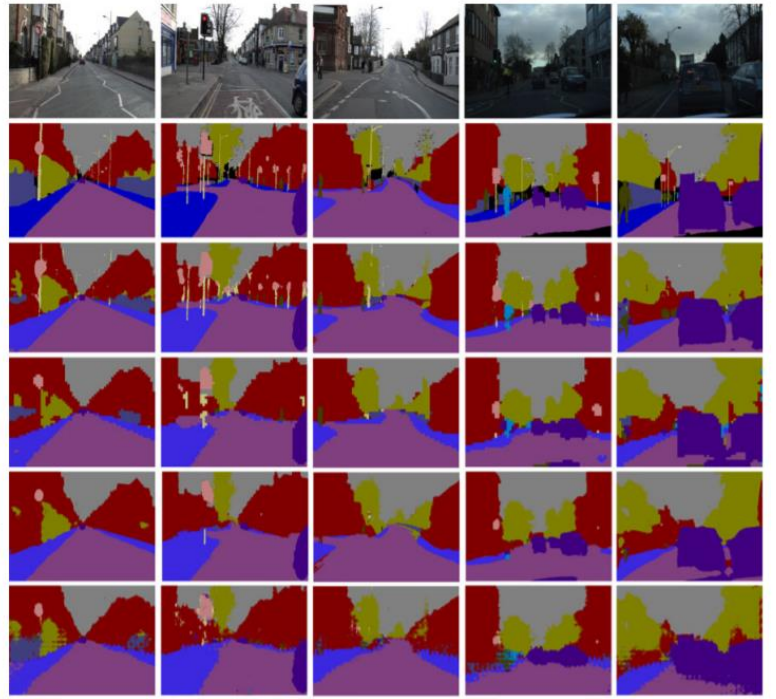
Badrinarayanan, Vijay, Alex Kendall, and Roberto Cipolla. "Segnet: A deep convolutional encoder-decoder architecture for image segmentation." *IEEE transactions on pattern analysis and machine intelligence* 39.12 (2017): 2481-2495.

Results

TABLE 2
Quantitative Comparisons of SegNet with Traditional Methods on the CamVid 11 Road Class Segmentation Problem [22]

Method	Building	Tree	Sky	Car	Sign-Symbol	Road	Pedestrian	Fence	Column-Pole	Side-walk	Bicyclist	Class avg.	Global avg.	mIoU	BF
SfM+Appearance [28]	46.2	61.9	89.7	68.6	42.9	89.5	53.6	46.6	0.7	60.5	22.5	53.0	69.1	n/a*	
Boosting [29]	61.9	67.3	91.1	71.1	58.5	92.9	49.5	37.6	25.8	77.8	24.7	59.8	76.4	n/a*	
Dense Depth Maps [32]	85.3	57.3	95.4	69.2	46.5	**98.5**	23.8	44.3	22.0	38.1	28.7	55.4	82.1	n/a*	
Structured Random Forests [31]	n/a											51.4	72.5	n/a*	
Neural Decision Forests [64]	n/a											56.1	82.1	n/a*	
Local Label Descriptors [65]	80.7	61.5	88.8	16.4	n/a	98.0	1.09	0.05	4.13	12.4	0.07	36.3	73.6	n/a*	
Super Parsing [33]	87.0	67.1	96.9	62.7	30.1	95.9	14.7	17.9	1.7	70.0	19.4	51.2	83.3	n/a*	
SegNet (3.5K dataset training - 140K)	**89.6**	**83.4**	96.1	**87.7**	52.7	96.4	**62.2**	**53.45**	**32.1**	**93.3**	**36.5**	**71.20**	**90.40**	60.10	46.84
CRF based approaches															
Boosting + pairwise CRF [29]	70.7	70.8	94.7	74.4	55.9	94.1	45.7	37.2	13.0	79.3	23.1	59.9	79.8	n/a*	
Boosting+Higher order [29]	84.5	72.6	**97.5**	72.7	34.1	95.3	34.2	45.7	8.1	77.6	28.5	59.2	83.8	n/a*	
Boosting+Detectors+CRF [30]	81.5	76.6	96.2	78.7	40.2	93.9	43.0	47.6	14.3	81.5	33.9	62.5	83.8	n/a*	

*SegNet outperforms all the other methods, including those using depth, video and/or CRF's on the majority of classes. In comparison with the CRF based methods SegNet predictions are more accurate in 8 out of the 11 classes. It also shows a good ≈ 10 percent improvement in class average accuracy when trained on a large dataset of 3.5 K images. Particularly noteworthy are the significant improvements in accuracy for the smaller/thinner classes. * Note that we could not access predictions for older methods for computing the mIoU, BF metrics.*



SMARCLE

Badrinarayanan, Vijay, Alex Kendall, and Roberto Cipolla. "Segnet: A deep convolutional encoder-decoder architecture for image segmentation." *IEEE transactions on pattern analysis and machine intelligence* 39.12 (2017): 2481-2495.

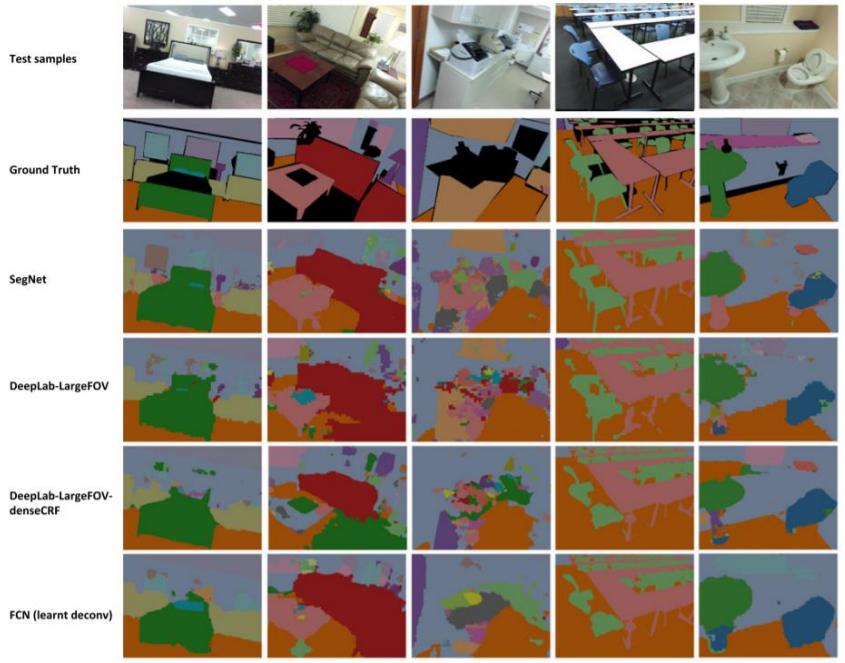
Results

TABLE 4
Quantitative Comparison of Deep Architectures on the SUNRGB-D Dataset When Trained on a Corpus of 5,250 Indoor Scenes

Network/Iterations	80 K				140 K				> 140 K				Max iter
	G	C	mIoU	BF	G	C	mIoU	BF	G	C	mIoU	BF	
SegNet	70.73	30.82	22.52	9.16	71.66	37.60	27.46	11.33	72.63	44.76	31.84	12.66	240 K
DeepLab-LargeFOV [3]	70.70	41.75	30.67	7.28	71.16	42.71	31.29	7.57	71.90	42.21	32.08	8.26	240 K
DeepLab-LargeFOV-denseCRF [3]	not computed								66.96	33.06	24.13	9.41	240 K
FCN (learnt deconv) [2]	67.31	34.32	24.05	7.88	68.04	37.2	26.33	9.0	68.18	38.41	27.39	9.68	200 K
DeconvNet [4]	59.62	12.93	8.35	6.50	63.28	22.53	15.14	7.86	66.13	32.28	22.57	10.47	380 K

Note that only the RGB modality was used in these experiments. In this complex task with 37 classes all the architectures perform poorly, particularly because of the smaller sized classes and skew in the class distribution. DeepLab-Large FOV, the smallest and most efficient model has a slightly higher mIoU but SegNet has a better G,C,BF score. Also note that when SegNet was trained with median frequency class balancing *it obtained 71.75, 44.85, 32.08, 14.06 (180 K) as the metrics.*

G – Global Average
C – Class Average
mIoU – mean Intersection over Union
BF – Boundary F1-measure



SMARCLE

Badrinarayanan, Vijay, Alex Kendall, and Roberto Cipolla. "Segnet: A deep convolutional encoder-decoder architecture for image segmentation." *IEEE transactions on pattern analysis and machine intelligence* 39.12 (2017): 2481-2495.

Results

Network	Forward pass(ms)	Backward pass(ms)	GPU training memory (MB)	GPU inference memory (MB)	Model size (MB)
SegNet	422.50	488.71	6803	**1052**	117
DeepLab-LargeFOV [3]	**110.06**	**160.73**	**5618**	1993	83
FCN (learnt deconv) [2]	317.09	484.11	9735	1806	539
DeconvNet [4]	474.65	602.15	9731	1872	877

TABLE 6

A comparison of computational time and hardware resources required for various deep architectures. The caffe time command was used to compute time requirement averaged over 10 iterations with mini batch size 1 and an image of 360×480 resolution We used nvidia-smi unix command to compute memory consumption. For training memory computation we used a mini-batch of size 4 and for inference memory the batch size was 1. Model size was the size of the caffe models on disk. SegNet is most memory efficient during inference model.

- FCN, DeepLab v1 보다는 느리지만 DeconvNet 보다는 빠름
- FCN, DeepLabv1에 비해 SegNet은 Decoder Architecture가 있어서 느릴 수밖에 없음
- DeconvNet에서 FC Layer를 제거한 구조여서 DeconvNet 보다는 속도가 빠름

SMARCLE

Badrinarayanan, Vijay, Alex Kendall, and Roberto Cipolla. "Segnet: A deep convolutional encoder-decoder architecture for image segmentation." *IEEE transactions on pattern analysis and machine intelligence* 39.12 (2017): 2481-2495.

Conclusion

- Road 및 indoor **scene understanding을 목적**으로 메모리 및 연산 시간에서 **효율적으로 동작**하도록 설계한 **SegNet architecture**를 제안

- SegNet은 **max-pooling indices**를 저장하는 방법을 사용해 **효율적으로 동작하면서도 좋은 성능**을 보인다.

- encoder에서 모든 feature map을 저장하는 아키텍처의 경우 성능은 우수하지만 메모리 사용량이 많아지는 단점이 있다. 그러나 SegNet은 **max-pooling indices** 을 사용하여 메모리 사용량을 줄이면서 성능도 개선하였다.

<SegNet VS U-Net>

- U-Net은 bio 이미지에 사용

- Pooling Indices를 사용하는 대신 전체 feature map이 encoder에서 decoder로 전송된 다음 연결하여 Convolution을 수행

SMARCLE

Badrinarayanan, Vijay, Alex Kendall, and Roberto Cipolla. "Segnet: A deep convolutional encoder-decoder architecture for image segmentation." *IEEE transactions on pattern analysis and machine intelligence* 39.12 (2017): 2481-2495.

My Opinion

- segmentation 모델을 메모리 사용량, 연산 속도 측면을 고려하여 아키텍처를 설계했다는 점에서 의미 있는 시도

- Main contribution으로 주장하는 메모리 사용량과 inference time을 고려했을때 극적인 성능을 개선했다고 느껴지진 않음

- 제안하는 method가 novelty하다고 느껴지진 않음

SMARCLE

Badrinarayanan, Vijay, Alex Kendall, and Roberto Cipolla. "Segnet: A deep convolutional encoder-decoder architecture for image segmentation." *IEEE transactions on pattern analysis and machine intelligence* 39.12 (2017): 2481-2495.

감사합니다

PR-SMARCLE

김찬영

SMARCLE