Applying Convolution Networks to Semantic Segmentation

Abstract

high-level의 computer vision은 이미지에서 의미를 파악하는 task로 이미지 안에 있는 object를 prediction하는 작업, 이미지 내 objet의 위치 정보를 제공하는 작업, 모든 픽셀의 레이블을 예측하는 작업 등이 있다. 이 중 의료분야와 자율주행자동차에 적용되는 Semantic Segmentation에 대해 간단히 이해하고, Deep Convolution Neural Network를 통해 어떻게 적용되었는지 paper [2], [5]를 통해 자세히 살펴본다.

1 Semantic Segmentation

Figure 1: semantic segmentation task

Dense prediction이라고도 하는 Semantic Segmentation은 이미지 내의 object들에 대해 각각의 클래스로 라벨링을하고, 모든 픽셀이 어느 클래스에 속하는지 에측하는 것이다. 예를들어 Figure2는 semantic segmantation task로 입력 이미지에서 사람에 대해 1, 식물과 잔디는 3, 구조물과 건축물에 대해서는 5로 라벨링하여 Segmentation을 통해 각 픽셀이 어떤 class에 속하는지 레이블을 나타낸 Segmentation map을 추출할 수 있다. (다만 이미지 내 사람이 여러 명 있을 경우, 사람들을 각각 분류하지 않고 하나의 class로 생각한다. 각각을 구분짓는 segmentation은 Instance segmentation이라 한다). 결론적으로 Semantic Segmentation은 사진을 보고 하나의 object로 분류하는 것을 넘어 이미지를 완전히 이해하는 컴퓨터 비전이라고 할 수 있을 것이다.

2 Fully Convolution Networks for Semantic Segmentation (FCN)

FCN는 semantic segmentation 벤치마크에서 당시 최고의 성능을 보였고, 다른 dense prediction을 수행하는 모델들의 초석이 되었다. FCN은 모든 layer를 convolution layer로 함으로써, Segmentation을 위해 사용되는 모든 filter들이 학습할 수 있는 end-to-end learning이 가능한 모델로 그 특징에 대해 자세히 알아본다.

2.1 Fully convolutional

1번의 Semantic Segmentation task에서 이미지에 대해 pixel단위로 분류를 하기 때문에 공간, 위치 정보가 매우 중요한 것을 직관적으로 알 수 있다. 하지만 Fully connected layer를 사용하는 딥러닝 모델은 이미지의 위치 정보가 사라지는 한계가 존재하였다. FCN은 이름에서도 알 수 있듯이 이 한계점을 보완하기 위해 모든 fc-layer를 convolutional layer로 대체하여 네트워크를 구성하였다. fc-layer를 conv-layer로 대체함으로써 얻는 또 다른 benefit은 오직 conv-layer로만 이루어져 있기 때문에 임의의 입력 이미지의 크기에 대해서 동일한 크기의 output을 도출해낼 수 있다.
Figure2는 Alexnet의 FC-layer를 1×1 convolution filter로 대체한 것으로, 이처럼 FCN은 Segmentation task를 위해 기존의 model(AlexNet [1], GoogLeNet [4] ,VGGnet [3])에 fine-tuning(convolutionalization)을 함으로써 fully convolutional network로 재구성하였다.

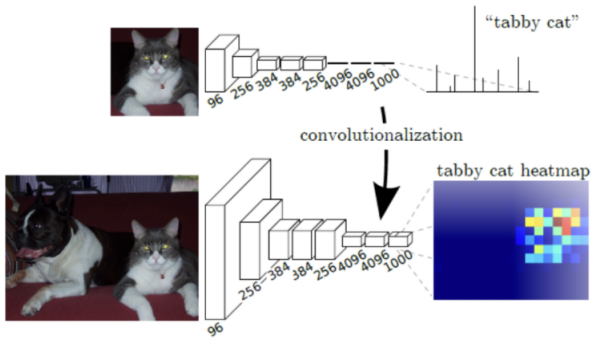

Figure 2: semantic segmentation task

2.2 skip architecture with up-sampling

deep network의 depth에 따라 feature map이 가지고 있는 정보의 성질이 다르다. depth가 얕은 shallow 혹은 fine 층은 appearance정보를 가지고 있다. 반면, 깊이가 점점 깊이가 깊어질수록 즉, deep 혹은 coarse 층은 semantic 정보를 가지고 있다. appearance information은 주로 낮은 수준이라고 하는 직선, 곡선, 색상 등이며 semantic(global) information은 낮은 수준의 정보보다 포괄적인 개체 정보이다.

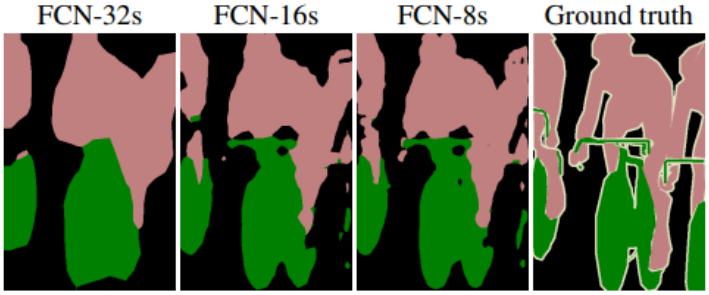

Figure 3: Result according to skip architecture

skip architecture은 깊은 층인 coarse layer의 semantic information를 deconvolution하여 네트워크의 얕은(shallow)층인 fine layer의 appearance information과 combination하는 것이라고 볼 수 있다. Figure2의 모델을 이용한 결과는 Figure3의 FCN-32s로 Ground truth에 비해 성능이 매우 좋지 않다. FCN-32s에서 32s는 최종 output에 stride가 32인 up-sampling 의미하며, 이 때문에 정보손실이 매우커 정교하지 못한 결과를 보여준다. 하지만 FCN-16s, FCN-8s로 갈수록 품질이 개선되는 것을 보여준다. Figure4와 같이 최종 output을 up-sampling을 하는 과정에서 stride를 16, 8로 줄이고, 보다 shallow한 layer의 feature map을 가져와 combination후 Bilinear interpolation을 해줌으로써 Segmentation 품질을 향상시킬 수 있었다.

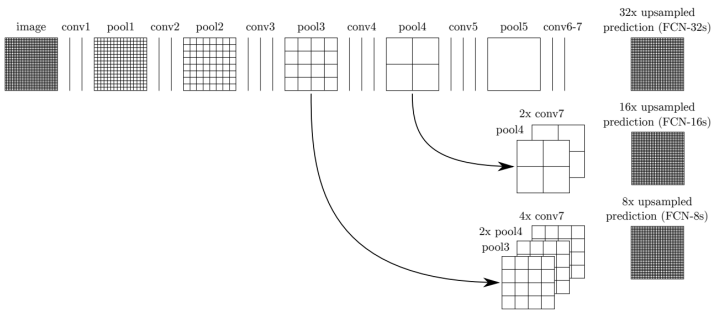

Figure 4: Skip architecture of FCN

3 DilatedNet

FCN은 image classification에서 사용한 model을 성공적으로 repurposing 하여 dense prediction에 적용하여, convolution network를 이용한 semantic segmentation 연구에 큰 motivation을 주었다. 하지만, semantic segmentation은 image classification과 다른 분야이므로 그 architecture를 그대로 따라하는 것이 아니라 구조적인 차이를 고려해야 했다. DilatedNet에서는 dense prediction에 알맞게 네트워크를 수정하는 task를 제시하였으며 state-of-the-art를 달성하였다. dense prediction를 위해 어떻게 네트워크를 수정하였는지 자세히 알아본다. 간단히 요약하면, Dilated Net은 image classification network와 달리 입력으로 rescaled images를 분석하지 않아도 되며 pooling과 subsamping이 없는 모델이다. 또한 parameter를 유지하고 resolution의 손실없이 receptive field를 확장하는 convolution이 도입된다.

3.1 Dilated Convolution

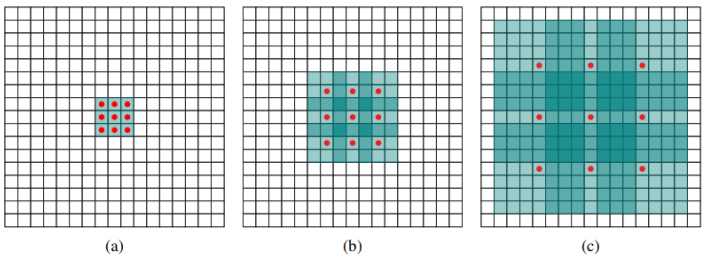

Figure 5: Dilated Convolution 1,2,4-dilated convolution

앞서도 언급하였듯이 semantic segmentation problem에서는 공간, 위치 정보의 보존이 중요하다. 이에 따라 Dilated Net에서는 parameter를 증가시키지 않고 receptive field를 확장할 수 있는 Dilated convolution을 제시한다. 일반적인 convolution filter를 dilation시킴으로써 receptive field를 resolution의 손실 없이 exponential하게 증가시킬 수 있다. dilation factor를 i라고 하였을 때 receptive field의 size는 $F_{i+1} = (2^{i+2}-1)\times(2^{i+2}-1)$이다. Figure5의 (a) 1-dilated convolution으로 F_1 표기되며 9개의 파라미터를 가지고 있는 일반적인 convolution filter와 동일하다. (b)와 (c)는 2-dilated convolution과 4-dilated convolution으로, receptive field가 7×7, 15×15이지만 파라미터의 개수는 동일하게된다. 즉, 연산량을 유지하며 receptive field를 확장하는 것이다. 더불어, dilated convolution을 적용함으로써 기존의 방식인 max-pooling 등을 하지 않아도 receptive field를 확장할 수 있어 resolution 또는 converage의 loss가 없다.

3.2 Network Architecture

Dilated Net의 architecture는 context module과 front-end module로 구성되어 있다. context module은 multi-scalie의 contextual information을 모아 dense prediction model의 성능을 향상시키기 위해 설계되었다. front-end module는 context module 앞에 위치한 입력 RGB이미지를 feature map으로 변형시켜주는 backbone으로 VGG16을 dense prediction에 맞추어 재구성하였다.

3.2.1 context module

Layer	1	2	3	4	5	6	7	8
Convolution	3×3	3×3	3×3	3×3	3×3	3×3	3×3	1×1
Dilation	1	1	2	4	8	16	1	1
Truncation	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
Receptive field	3×3	5×5	9×9	17×17	33×33	65×65	67×67	67×67
Output channels								
Basic	C	C	C	C	C	C	C	C
Large	$2C$	$2C$	$4C$	$8C$	$16C$	$32C$	$32C$	C

Figure 6: Architecture of Context module

comtext 모듈은 입출력의 크기와 채널 수가 동일하여 어느 dense prediction architectures에 사용가능하다. context 모듈의 basic form을 살펴보면 Dilation factor를 다르게하는 것을 볼 수 있다. 7개의 layers의 3×3 filter에 순서대로 Dilation을 1,1,2,4,8,16, 마지막으로 1로 적용한다. 이때, Dilation이 적용된 기준은 64×64의 resolution으로 입력의 resolution에 맞추어 receptive field의 expansion을 멈춘다. 이 dilated convolution 후에는 Truncation으로 ReLU가 수행되며, 최종적으로 $1\times1\times C$의 convolution이 적용되어 module의 output이 추출된다.
기존의 Convolutional network에 적용되는 random distribution을 통한 Initialization은 context module에는 좋지 않은 결과를 보여주어, 초기화 방식을 아래의 식으로 대체하였다.

$$k^b(\mathbf{t},a) = 1_{[\mathbf{t}=0]}1_{[a=b]} \tag{1}$$

(1)은 identity initialization으로, a는 input feature map의 index, b는 output feature map의 index를 의미한다. 일반적으로 이 방법의 초기화는 back-propagation이 잘 안되지만 context network에서는 feature map의 정확도를 높여주어 contextual 정보를 성공적으로 얻을 수 있다. 이 context basic module을 통해 양적으로나 질적으로 dense prediction의 성능에 기여하였다. 특히 매우 큰 receptive field를 적용하여도 3×3 filter의 parameter와 동일한 개수이므로 context module의 총 파라미터 수는 약 $64C^2$이다.

3.2.2 front-end module

front-end module은 VGG-16에서 image-classfication을 위한 요소를 제거하고 dense-prediction에 맞추어 재구성하였다. resolution의 loss를 일으키는 마지막의 두 pooling layer와 전체적으로 striding layer를 제거하였고, 대신 몇 convolutions을 dilated convolutions로 대체하였다. 또한 pooling layer의 제거는 네트워크를 simplifying하는데, 이는 정확도를 더 향상시키는 결과를 내었다. 마지막으로, dense prediction에서는 필요하지 않은 중간 feature maps의 padding을 제거하였다.

3.3 Conclusion

	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mean IoU
Front end	86.3	38.2	76.8	**66.8**	63.2	87.3	78.7	82	33.7	76.7	53.5	73.7	76	76.6	83	**51.9**	77.8	44	79.9	**66.3**	69.8
Front + Basic	86.4	37.6	78.5	66.3	64.1	89.9	79.9	84.9	**36.1**	79.4	**55.8**	77.6	81.6	79	83.1	51.2	81.3	43.7	82.3	65.7	71.3
Front + Large	**87.3**	**39.2**	**80.3**	65.6	**66.4**	**90.2**	**82.6**	**85.8**	34.8	**81.9**	51.7	**79**	**84.1**	**80.9**	**83.2**	51.2	**83.2**	**44.7**	**83.4**	65.6	**72.1**

Figure 7: Semantic Segmentation accuracies

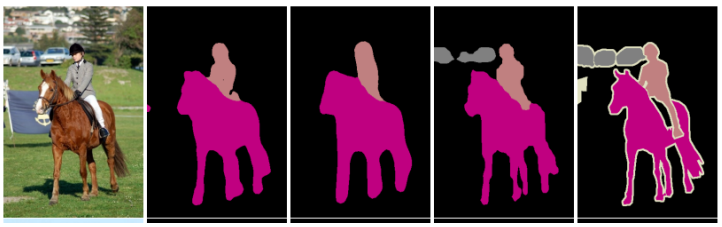

Figure 8: Prediction of Semantic Segmentation

Figure7는 DilatedNet의 각 class를 분석한 accuracy이며, Figure8은 그 결과를 image로 보여준 사진이다. Ground truth와 꽤 유사함을 확인할 수 있다. 이미지를 의미적으로 분석하는 연구가 많아지고, 특히 Semantic segmentation은 자율주행자동차 등에서도 중요한 역할을 하고 있다. 센서, 의료 분야에서도 점점 정교한 분석으로 이상 상황을 감지하거나, 또한 드론에 탑재된 카메라의 이미지를 이용하여 건축물 내부 검사까지 가능할 것 같다. Dilated Net에 다른 model을 plugging하여 정확도를 더 높이고 있고 최근의 연구가 어떻게 진행되고 있을지 기대된다.

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[2] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[3] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[4] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[5] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.