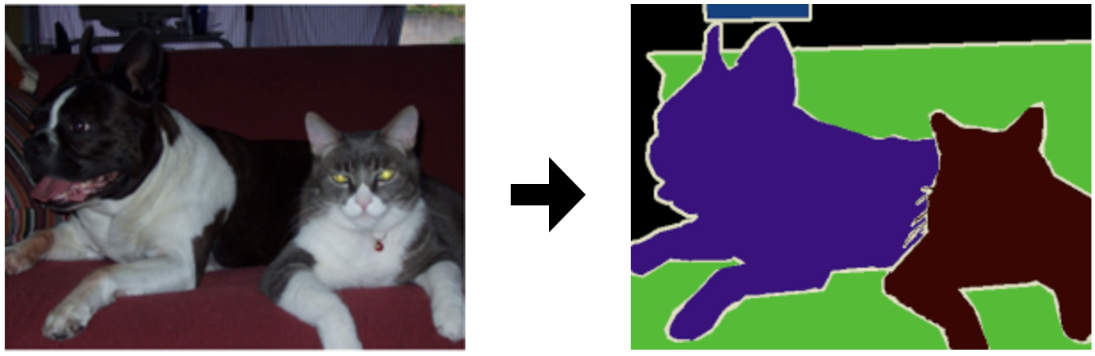
U-Net(FCN)

Fully Convolutional Networks for Semantic Segmentation
(이하 FCN)은 Semantic Segmentation 문제를 위해 제안된 딥러닝 모델이다.



Semantic Segmentation

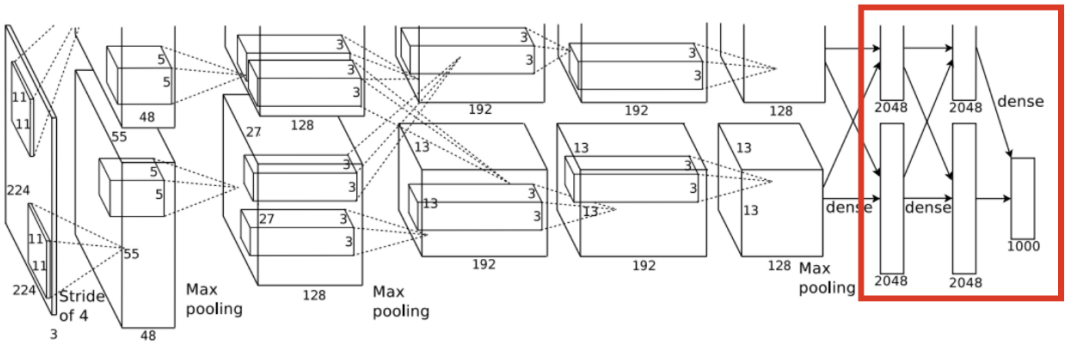
FCN은 Semantic Segmentation 모델을 위해 기존에 이미지 분류에서 우수한 성능을 보인 CNN 기반 모델(AlexNet, VGG16, GoogLeNet)을 목적에 맞춰 변형시킨 것이다.

이러한 [Image classification model] to [Semantic segmentation model]은 크게 다음의 세 과정으로 표현할 수 있다:

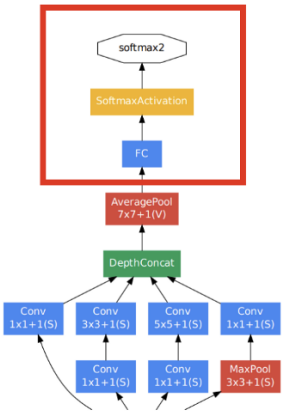
- ***Convolutionalization***
- ***Deconvolution (Upsampling)***
- ***Skip architecture***

Convolutionalization

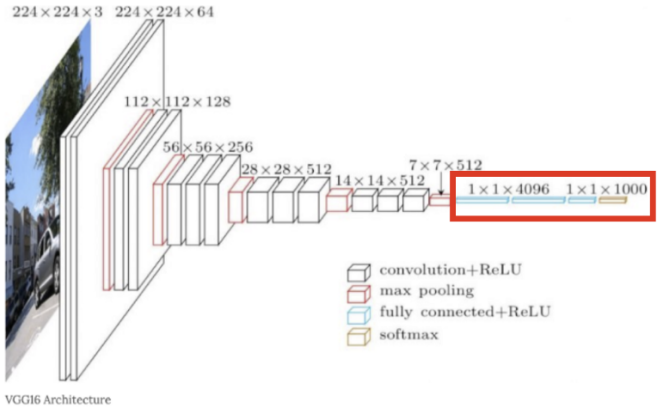
Convolutionalization(컨볼루션화)이라는 표현의 의미를 이해하기 위해서는 기존의 이미지 분류 모델들을 먼저 살펴볼 필요가 있다.



AlexNet



GoogLeNet

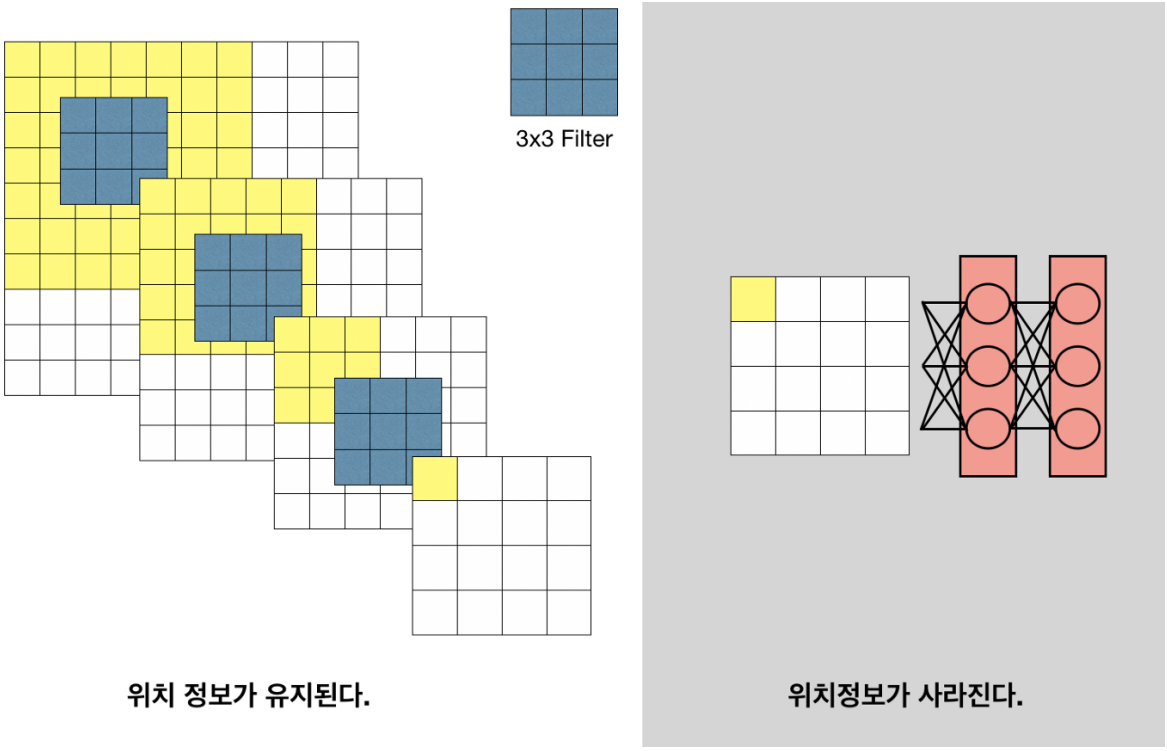


VGG16

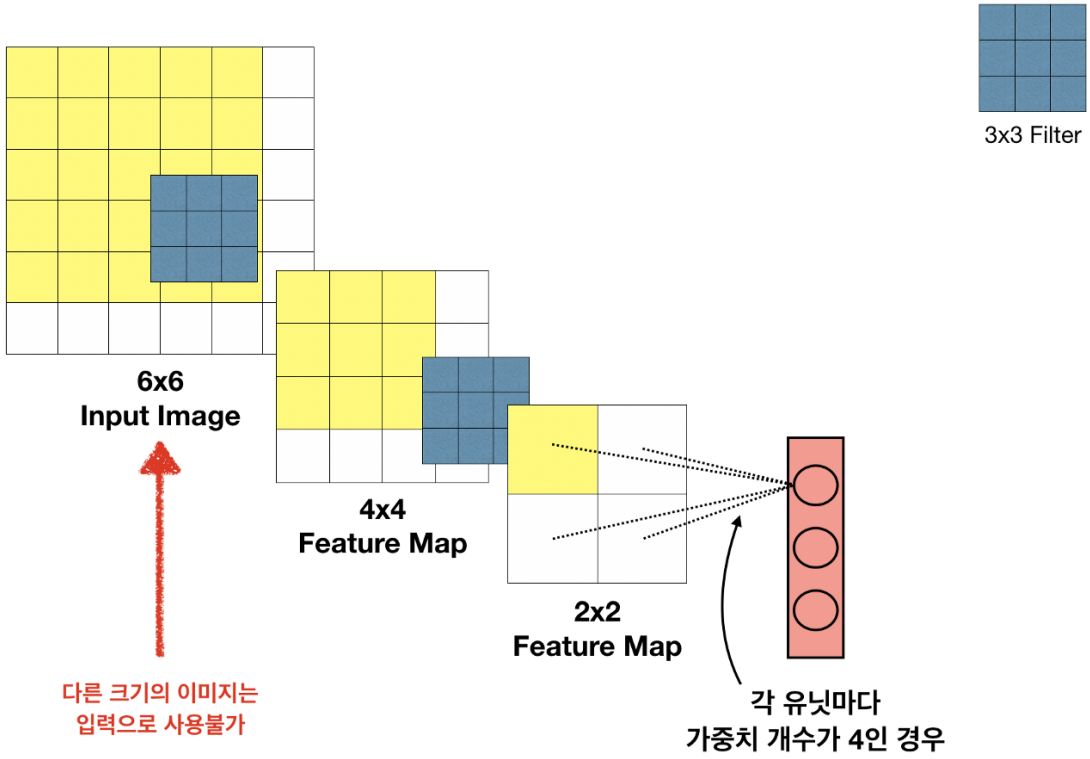
Image classification 모델들은 기본적으로 내부 구조와 관계없이 모델의 근본적인 목표를 위해 출력층이 Fully-connected(이하 fc) layer로 구성되어 있다.

이러한 구성은 네트워크의 입력층에서 중간부분까지 ConvNet을 이용하여 영상의 특징들을 추출하고 해당 특징들을 출력층 부분에서 fc를 이용해 이미지를 분류하기 위함이다.

그런데, Semantic Segmentation 관점에서는 **fc layer가 갖는 한계점**이 있다.
1.이미지의 위치 정보가 사라진다.

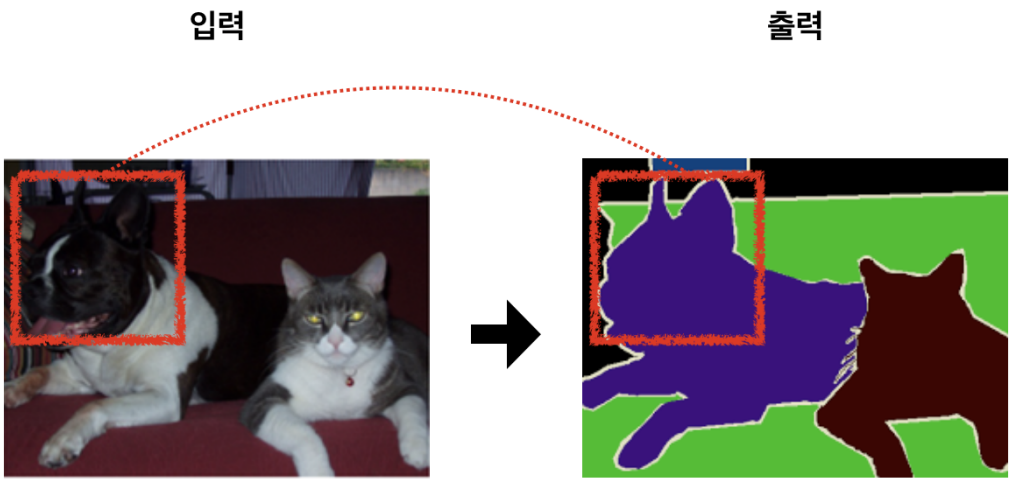


2. 입력 이미지 크기가 고정된다.



Dense layer에 가중치 개수가 고정되어 있기 때문에 바로 앞 레이어의 Feature Map의 크기도 고정되며, 연쇄적으로 각 레이어의 Feature Map 크기와 Input Image 크기 역시 고정된다.

Segmentation의 목적은 **원본 이미지의 각 픽셀에 대해 클래스를 구분하고 인스턴스 및 배경을 분할하는 것**으로 위치 정보가 매우 중요하다.



위치 정보가 매우 중요함

이러한 fc-layer의 한계를 보완하기 위해 모든 fc-layer를 Conv-layer로 대체하는 방법을 택하였다.

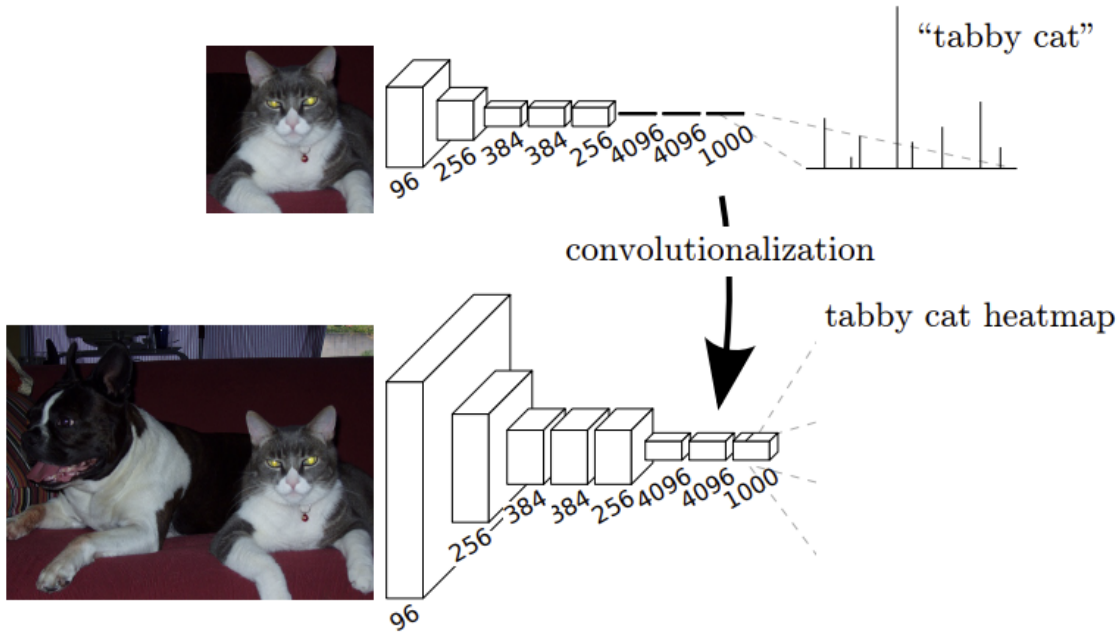
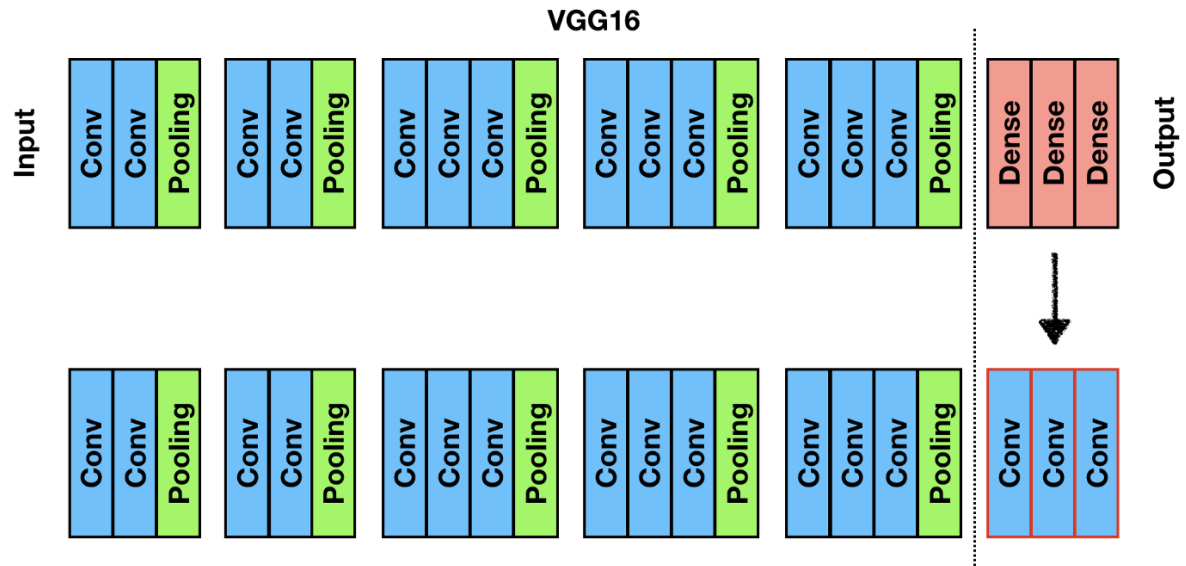


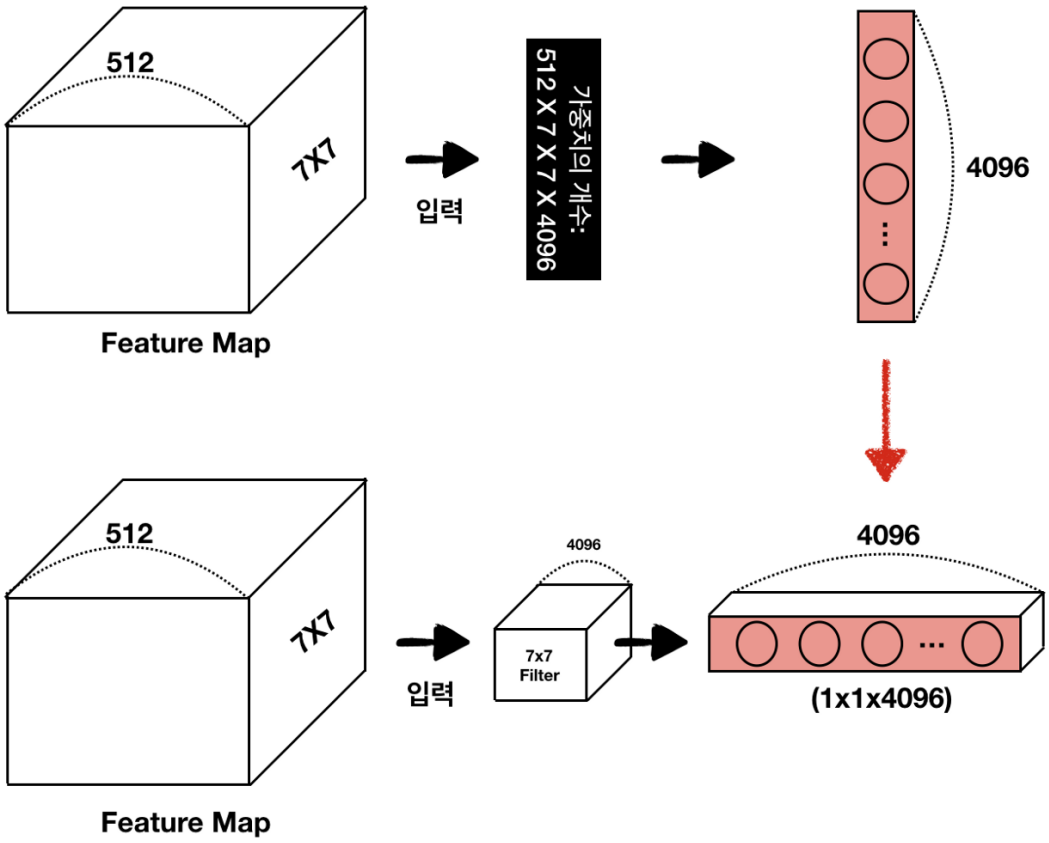
Figure 2. Transforming fully connected layers into convolution layers enables a classification net to output a heatmap. Adding layers and a spatial loss (as in Figure 1) produces an efficient machine for end-to-end dense learning.

Fully-connected layer는 입력의 모든 영역을 receptive field로 보는 필터의 Conv layer로 생각할 수 있다.

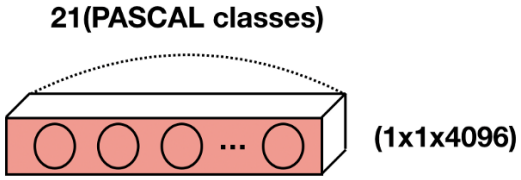
VGG16을 예로 살펴보자. 다음과 같이 출력층 부분의 마지막 3 fc-layers를 모두 Conv-layers로 변경한다.



Dense Layer에서 Conv Layer로 변환하는 방식은 다음과 같다.

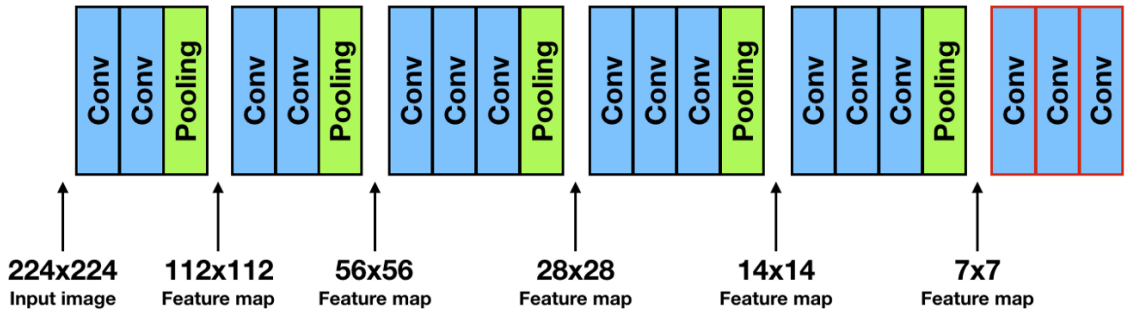


첫 번째 fully-connected layer를 (7x7x512) 4096filter conv로 변경하면 가중치의 수가 유지된다.



마지막 fc-layer의 경우 채널 차원을 클래스 수에 맞춘 1x1 conv로 변환한다.

VGG16에서 다섯 번째 max-pooling(size: 2x2, stride: 2) 연산 후 Feature map의 크기는 7x7이 된다. (입력 이미지의 크기가 224x224 인 경우)



Convolutionalization을 통해 출력 Feature map은 원본 이미지의 위치 정보를 내포할 수 있게 되었다.

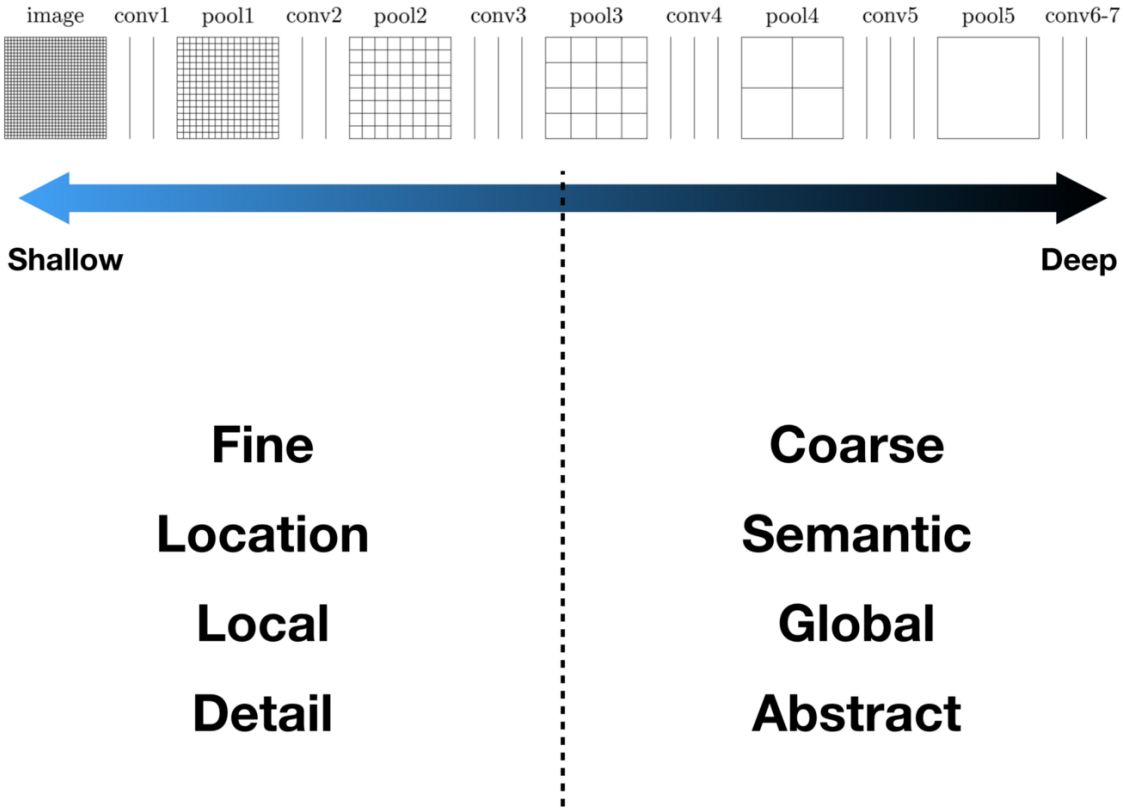
그러나 Semantic segmentation의 최종 목적인 픽셀 단위 예측과 비교했을 때, FCN의 출력 Feature map은 너무 **coarse(거친, 알맹이가 큰)** 하다.

따라서, Coarse map을 원본 이미지 크기에 가까운 Dense map으로 변환해줄 필요가 있다. 적어도 input image size * 1/32 보다는 해상도가 높을 필요가 있다.

Skip Architecture

정확하고 상세한 구분(Segmentation)을 얻기 위해

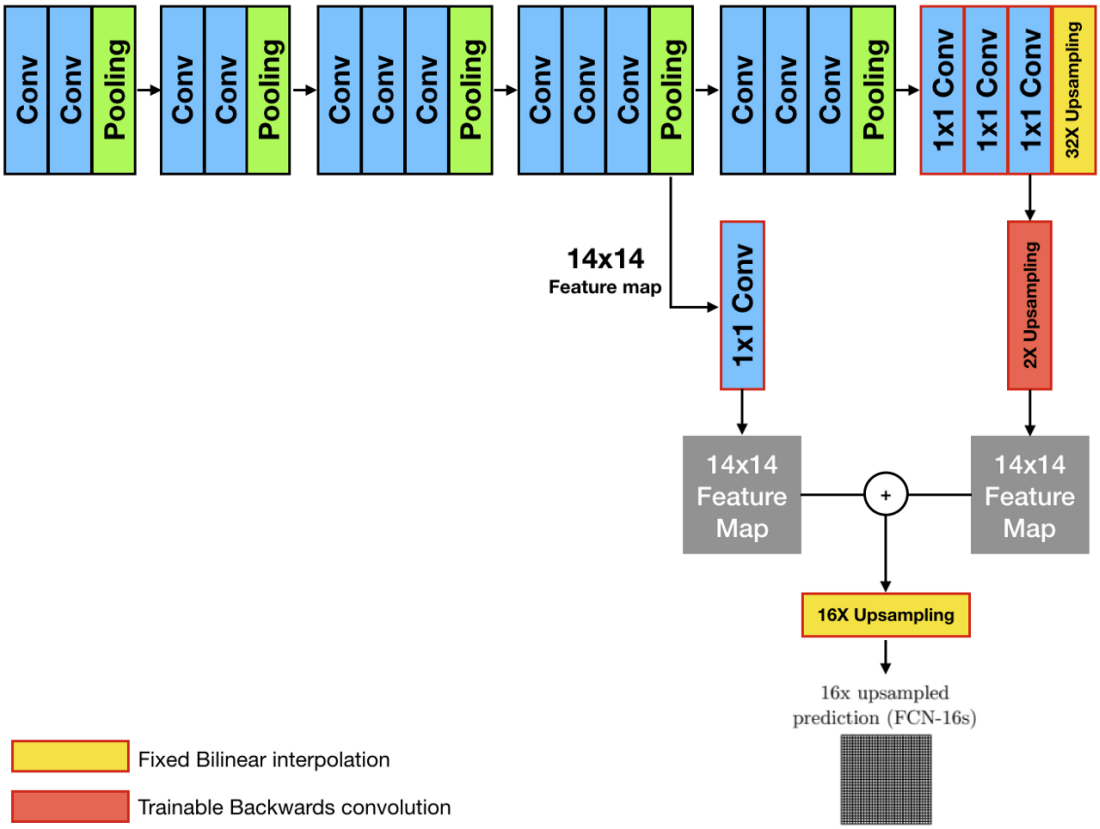
Deep & Coarse(추상적인) 레이어의 의미적(Semantic) 정보와 **Shallow & fine** 층의 외관적(appearance) 정보를 결합한 **Skip architecture**를 정의한다.

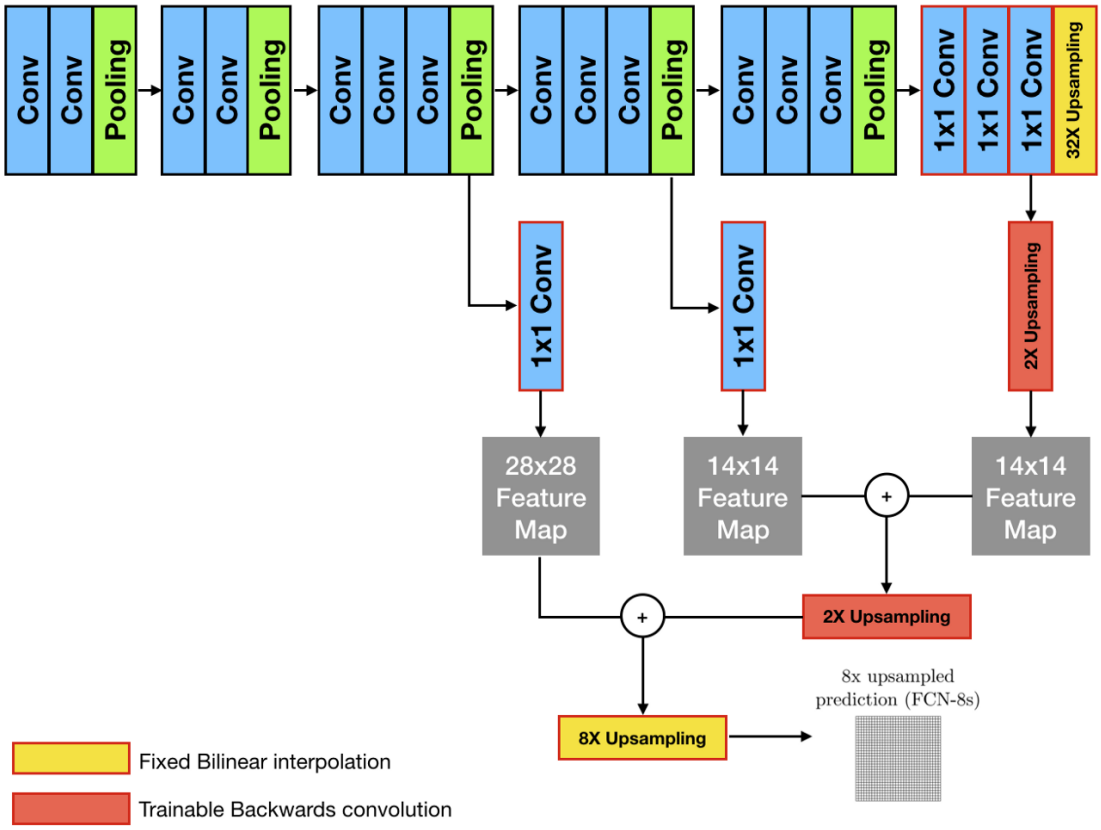


시각화 모델을 통해 입력 이미지에 대해 **얕은 층에서는 주로 직선 및 곡선, 색상 등의 낮은 수준의 특징**에 활성화되고, **깊은 층에서는 보다 복잡하고 포괄적인 개체 정보**에 활성화된다는 것을 확인할 수 있다.

또한 **얕은 층에선 local feature를 깊은 층에선 global feature를 감지**한다고 볼 수 있다.

FCNs 연구팀은 이러한 직관을 기반으로 앞에서 구한 Dense map에 얕은 층의 정보를 결합하는 방식으로 Segmentation의 품질을 개선하였다.


Conv
Conv
Pooling
Conv
Conv
Pooling
Conv
Conv
Conv
Pooling
Conv
Conv
Conv
Pooling
Conv
Conv
Conv
Pooling
1x1 Conv
1x1 Conv
1x1 Conv
32X Upsampling
14x14
Feature map
1x1 Conv
2X Upsampling
14x14
Feature
Map
+
14x14
Feature
Map
16X Upsampling
16x upsampled
prediction (FCN-16s)
Fixed Bilinear interpolation
Trainable Backwards convolution


Conv
Conv
Pooling
Conv
Conv
Pooling
Conv
Conv
Conv
Pooling
Conv
Conv
Conv
Pooling
Conv
Conv
Conv
Pooling
1x1 Conv
1x1 Conv
1x1 Conv
32X Upsampling
1x1 Conv
1x1 Conv
2X Upsampling
28x28
Feature
Map
14x14
Feature
Map
+
14x14
Feature
Map
+
2X Upsampling
8x upsampled
prediction (FCN-8s)
8X Upsampling
Fixed Bilinear interpolation
Trainable Backwards convolution

각 Pooling에 Prediction을 위해 추가된 Conv layer의 필터는 0으로, Trainable Backwards convolution은 Bilinear interpolation으로 초기화한 후 학습을 진행하였다.

이러한 Skip Architecture를 통해 다음과 같이 개선된 Segmentation 결과를 얻을 수 있다.

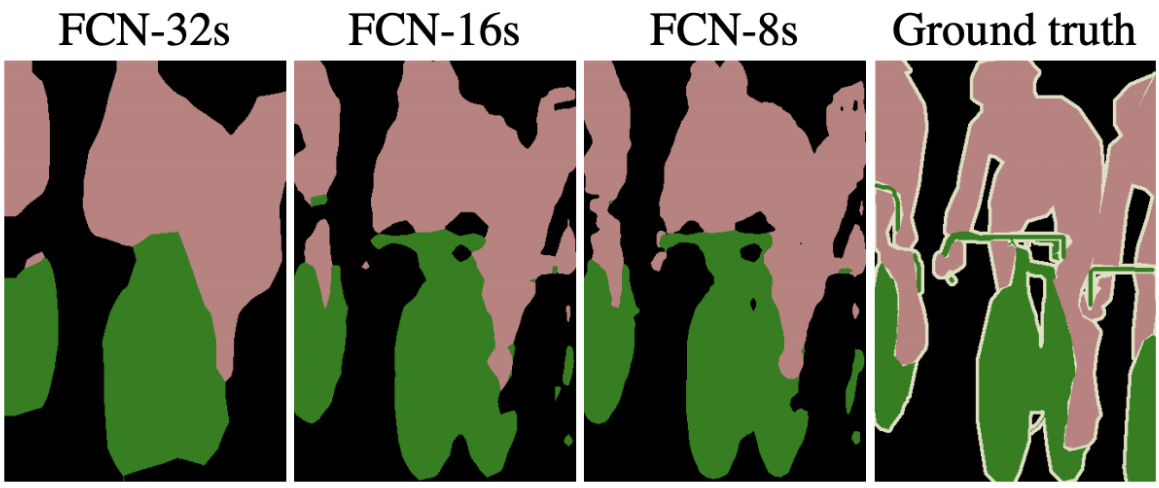


Figure 4. Refining fully convolutional nets by fusing information from layers with different strides improves segmentation detail. The first three images show the output from our 32, 16, and 8 pixel stride nets (see Figure 3).

정리

FCNs은 기존의 딥러닝 기반 이미지 분류를 위해 학습이 완료된 모델의 구조를 **Semantic Segmentation 목적에 맞게 수정**하여 Transfer learning 하였다.

Convolutionalized 모델을 통해 예측된 Coarse map을 원본 이미지 사이즈와 같이 **세밀(Dense)하게 만들기 위해 Up-sampling을 수행**하였다.

또한 Deep Neural Network에서 **얕은 층의 Local 정보**와 **깊은 층의 Semantic 정보**를 결합하는 **Skip architecture**를 통해 보다 **정교한 Segmantation 결과**를 얻을 수 있었다.

FCNs은 End-to-End 방식의 Fully-Convolution Model을 통한 Dense Prediction 혹은 Semantic Segmentation의 초석을 닦은 연구로써 이후 많은 관련 연구들에 영향을 주었다.