고해상도 이미지의
실시간 의미적 분할을 위한 ICNet 활용

ICNet for Real-Time Semantic Segmentation on High-Resolution Images
(Zhao H., Qi X., Shen X., Shi J., and Jia J.; 2018)

※ ICNet : Image Cascade Network

※ 고려대학교 DSBA (Data Science & Business Analytics) 연구실 이윤승 님의 세미나 자료에 일부 내용을 보충한 자료임

목차

목차

1. 되돌아보기 :

   “Computer Vision” “의미적 분할” “Upsampling” “PSPNet vs ICNet”

2. 기반 모델 :

   Pyramid Scene Parsing Network (PSPNet; 2016)

3. 개선 모델 :

오늘은 여기까지!

   Image Cascade Network (ICNet; 2018)

되돌아보기 : Computer Vision

(데이터셋 : CityScapes)

의미적 분할
Semantic Segmentation

- 픽셀 단위(Pixel-wise) 분류
- 같은 클래스 객체가 있더라도, 개별 객체에 대한 고려 안 함
- PSPNet, ICNet

객체 탐지
Object Detection

- 다수 객체의 위치와 클래스 분류
- 같은 클래스 객체도 다른 객체로 간주
- R-CNN, YOLO 계열

객체 분할
Instance Segmentation

- 다수 객체의 위치와 객체별 픽셀 분류
- 같은 클래스 객체도 다른 객체로 간주
- Mask-RCNN

되돌아보기 : 의미적 분할 Semantic Segmentation

인코더-디코더(Encoder-Decoder) : 대표적인 의미적 분할 모델 구조

- 1단계 : 합성곱 연산을 통해 이미지 정보 축약 과정으로 얻은 축소된 특성맵(feature map)
- 2단계 : 축소된 특성맵으로부터 upsampling과 residual connection (= skip 또는 shortcut connection)을 통해 도출된 출력 텐서

  → 예측하고자 하는 클래스 개수만큼의 채널을 가짐

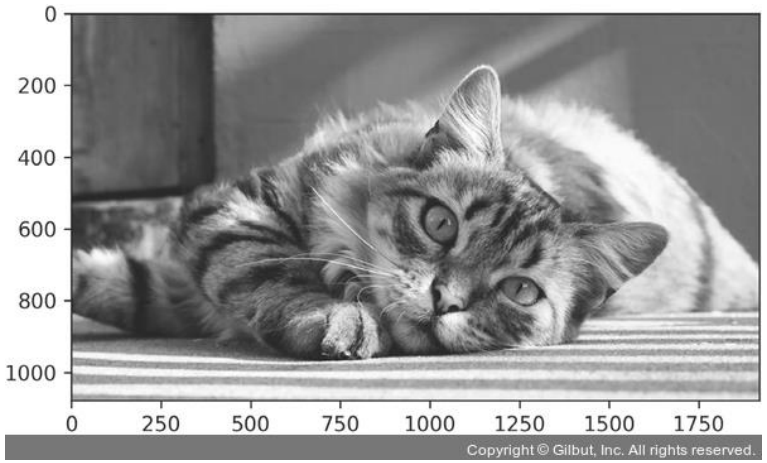
  → Upsampling 기법 : [학습 X] Bilinear 보간법, [학습 O] Transpose-Convolution, Dilated Convolution

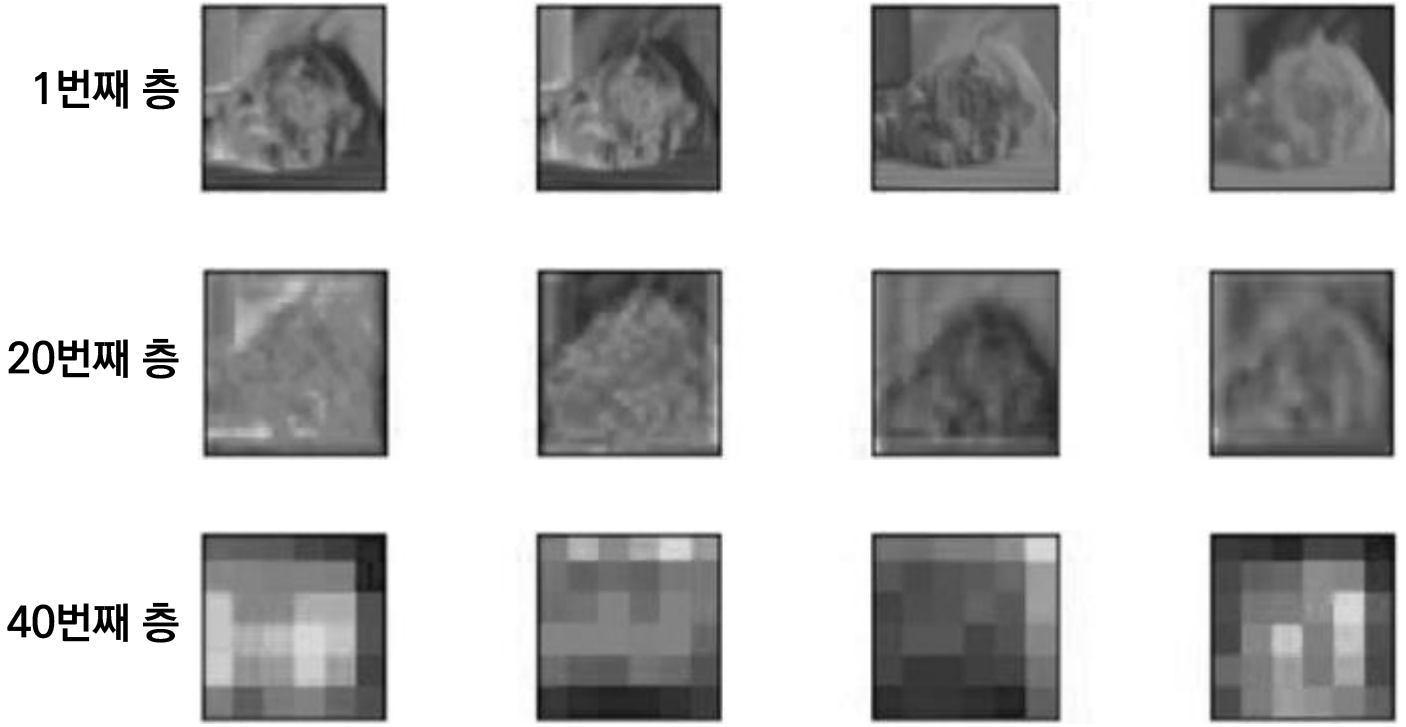
참고 자료 : 특성맵 Feature Map

활성화 맵이라고도 하며, 입력 이미지 또는 다른 특성맵에 필터(ex. 합성곱)를 적용한 결과를 가리킴

- 특정 입력 이미지에 대한 특성맵을 시각화 하면, 특성맵에서 입력 특성을 감지하는 방법을 어느정도 이해할 수 있음
- 입력층과 가까울수록 입력 이미지의 형태가 많이 유지되고, 출력층에 가까울수록 이미지의 특징만 전달 됨

예시) 입력 이미지

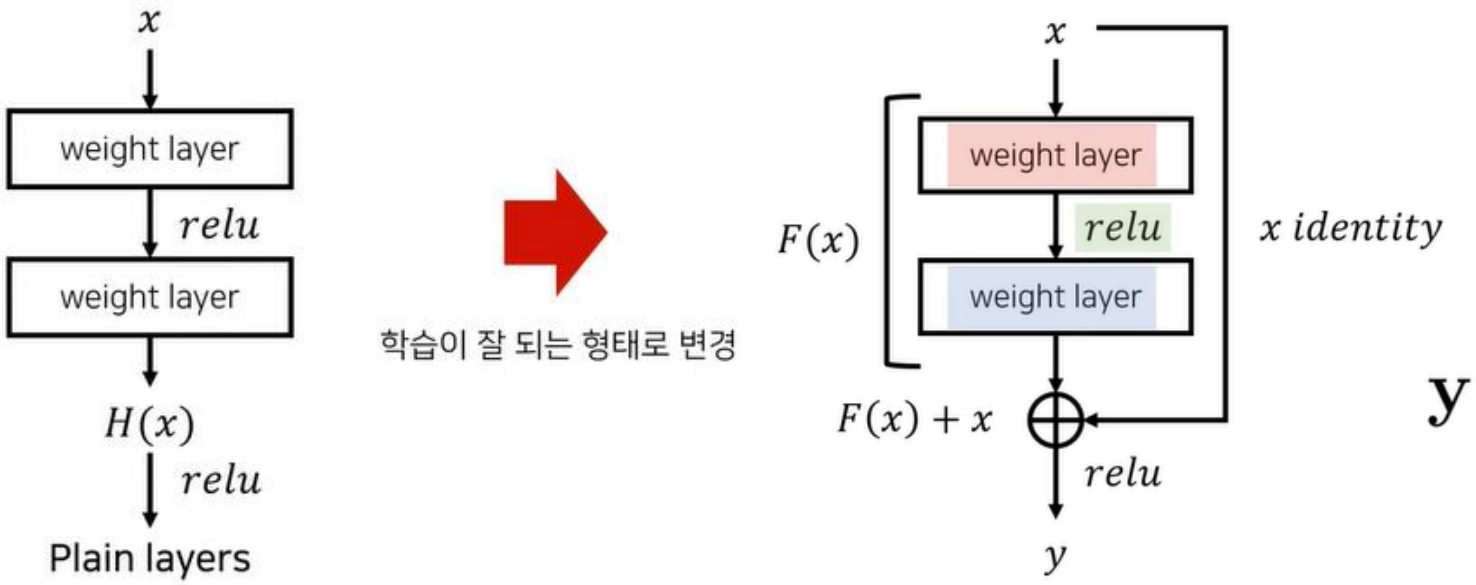




참고 자료 : Residual / Skip / Shortcut Connection

기울기 소실 문제를 개선하기 위해 학습한 함수에 입력(input) 값을 더해주는 방법

- 아래 그림에서 입력 x는 그대로 가져오고, 레이어를 거쳐 도출된 F(x)를 더해주는 형태
- 역전파 시 F(x) 부분만 학습하면 됨

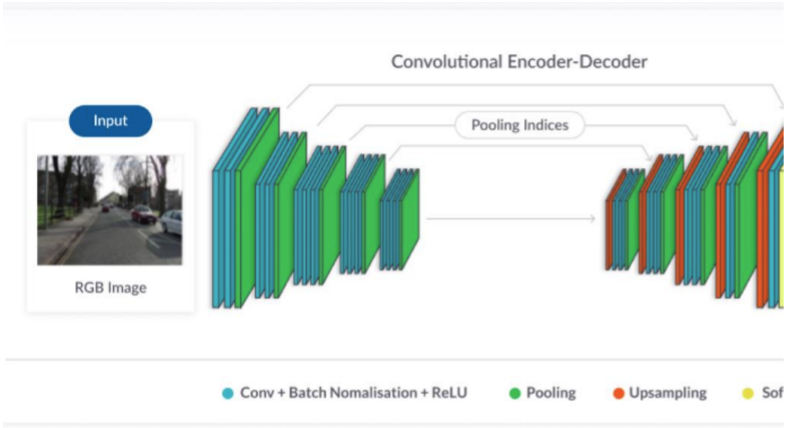


$$\mathcal{F} = W_2\sigma(W_1\mathbf{x})$$

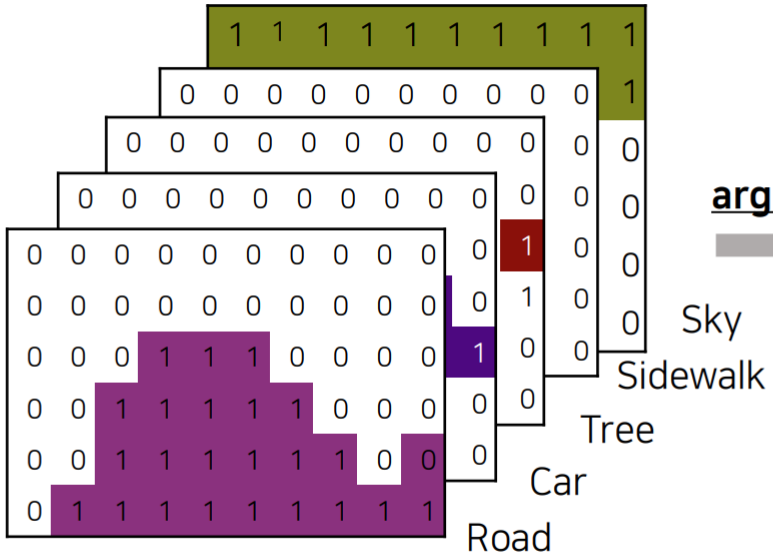
일반적인 형태

$$\mathbf{y} = \underbrace{\mathcal{F}(\mathbf{x}, \{W_i\})}_{\text{multiple convolutional layers}} + \underbrace{W_s\mathbf{x}}_{\text{shortcut}}$$

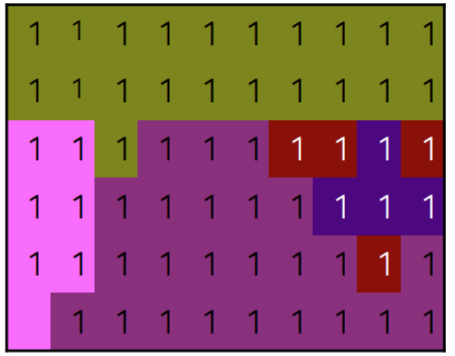
되돌아보기 : 의미적 분할 Semantic Segmentation

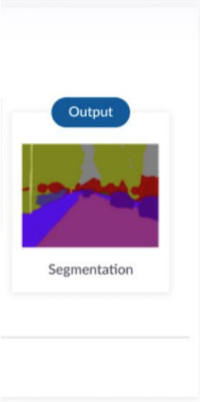


(value의 경우, 실제로는 softmax 값으로 표현됨)



argmax





Input Image
W x H x 3

Output Tensor
W x H x 5

Final Prediction
W x H x 1

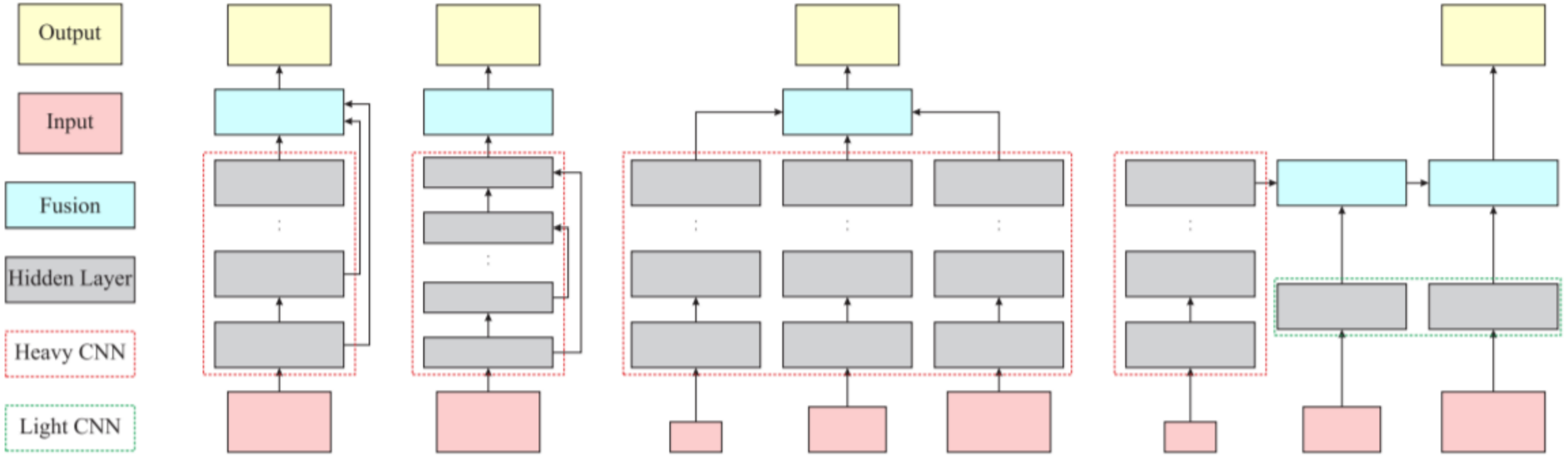
인코더-디코더(Encoder-Decoder) : 대표적인 의미적 분할 모델 구조

- 3단계 : Argmax 연산을 통해 해당 위치의 픽셀에서 softmax 값이 가장 높은 클래스로 예측 수행

  → 손실 함수 : Categorical Cross Entropy

  → Ground Truth 라벨 : 각 픽셀별 클래스 정보가 있어야 함

되돌아보기 : 의미적 분할 Semantic Segmentation



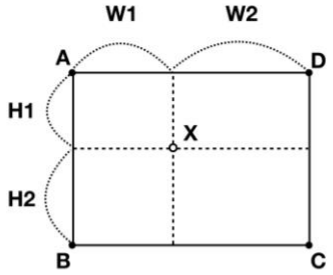
모델	FCN	SegNet, UNet	DeepLab, PSPNet	ICNet
특징	Skip Connection	인코더 디코더	Multi-scale 예측 및 앙상블	낮은 해상도 이미지 → Heavy CNN 높은 해상도 이미지 → Light CNN

되돌아보기 : Upsampling 기법

사용 이유 : 합성곱 신경망을 거쳐 나온 특성맵은 coarse 하므로 픽셀 단위 클래스 분류를 위한 dense 특성맵을 얻기 위함

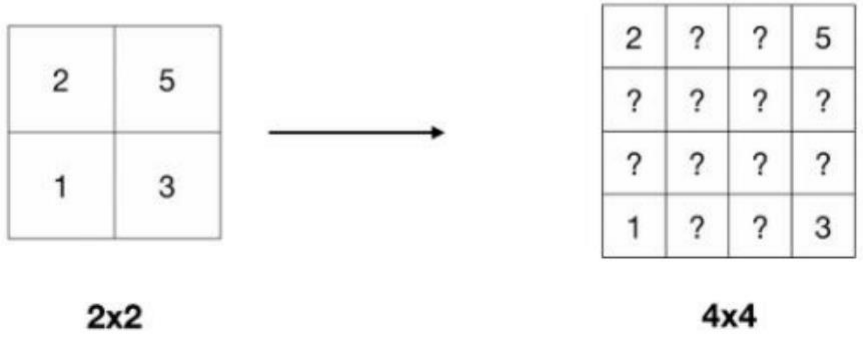
[Bilinear 보간법]

- 기본 원리



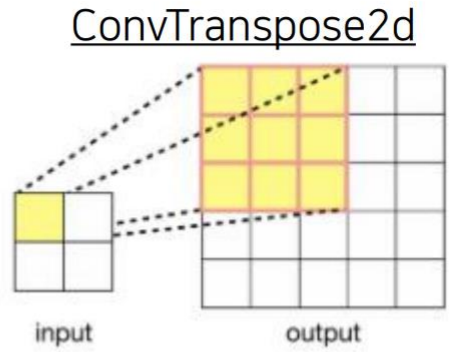
$$X = \left(A\frac{H2}{H1+H2} + B\frac{H1}{H1+H2}\right)\frac{W2}{W1+W2} + \left(D\frac{H2}{H1+H2} + C\frac{H1}{H1+H2}\right)\frac{W1}{W1+W2}$$

- Feature map 적용 (*not learnable*)

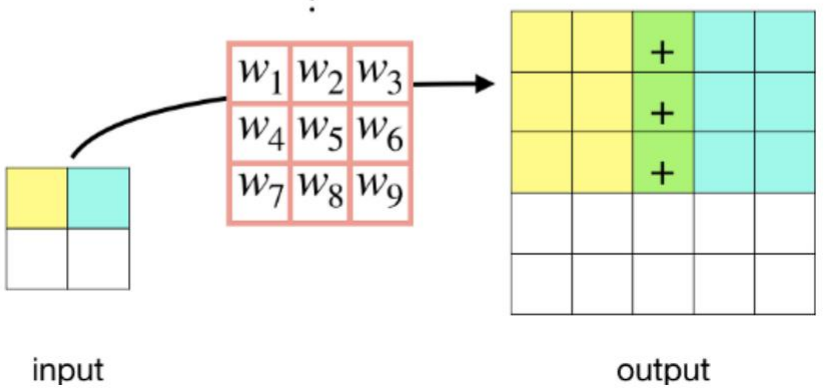


[Transpose Convolution]

- 기본 원리



- Feature map 적용 (*learnable*)



되돌아보기 : Upsampling 기법

사용 이유 : 합성곱 신경망을 거쳐 나온 특성맵은 coarse 하므로 픽셀 단위 클래스 분류를 위한 dense 특성맵을 얻기 위함

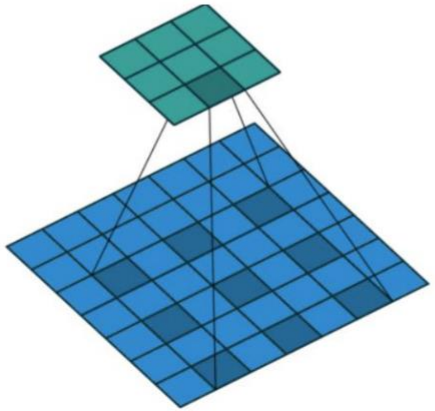
[Dilated Convolution]

- 기본 원리

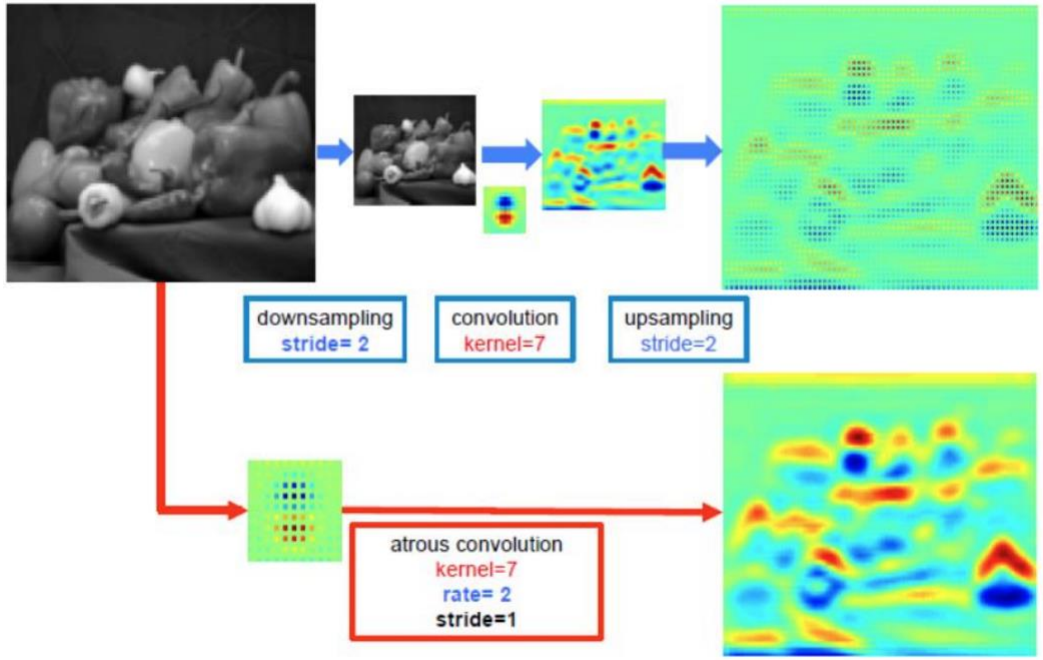
- Conv2d (option: dilation)
- Filter 내부에 zero padding 추가해 receptive field 확장
- 넓은 Receptive field 유지하며 적은 파라미터로 학습 가능

⇒ global context 보존 위해

- Feature map 적용 (*learnable*)



- 예시



되돌아보기 : PSPNet vs. ICNet

[1] PSP-Net (CVPR, 2017)

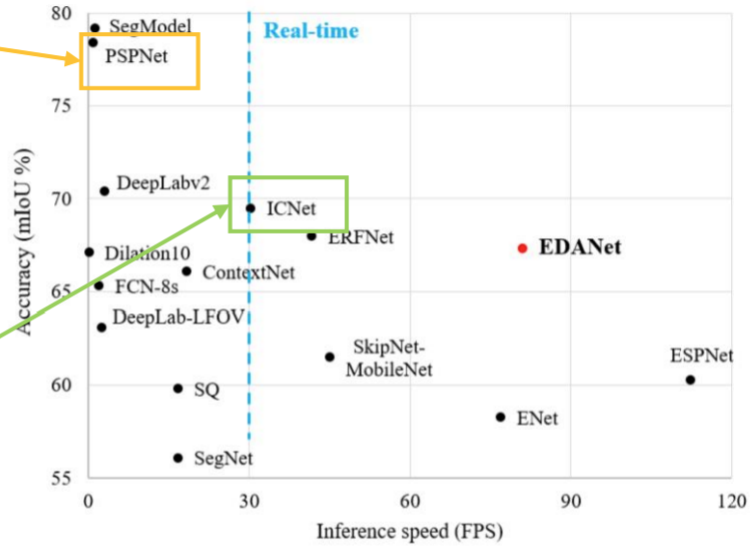
Pyramid Scene Parsing Network

Hengshuang Zhao[1] Jianping Shi[2] Xiaojuan Qi[1] Xiaogang Wang[1] Jiaya Jia[1]
[1]The Chinese University of Hong Kong [2]SenseTime Group Limited
{hszhao, xjqi, leojia}@cse.cuhk.edu.hk, xgwang@ee.cuhk.edu.hk, shijianping@sensetime.com

[2] IC-Net (ECCV, 2018)

ICNet for Real-Time Semantic Segmentation on High-Resolution Images

Hengshuang Zhao[1], Xiaojuan Qi[1], Xiaoyong Shen[2], Jianping Shi[3], Jiaya Jia[1,2]
[1]The Chinese University of Hong Kong, [2] Tencent Youtu Lab, [3]SenseTime Research
{hszhao,xjqi,leojia}@cse.cuhk.edu.hk,
dylanshen@tencent.com, shijianping@sensetime.com



되돌아보기 : PSPNet vs. ICNet

[1] PSP-Net (CVPR, 2017)

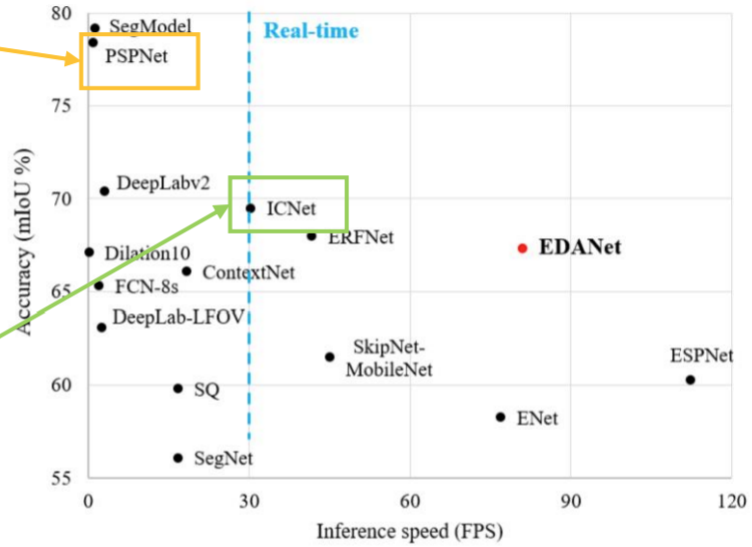
Pyramid Scene Parsing Network

[PSPNet 특징]

- 81.2% mIOU로 정확도가 매우 높음
- 0.78 fps로 매우 느린 모델에 속함

[2] IC-Net (ECCV, 2018)

ICNet for Real-Time Semantic Segmentation on High-Resolution Images



[ICNet 특징]

- PSPNet 기반으로 만들어진 모델
- 70% mIOU로 PSPNet에 비해 낮지만 성능 저하를 최소화 함
- 30 fps로 실시간(real-time) 속도의 모델 임

참고 자료 : mIOU

Mean Intersection Over Union의 줄임 말로, 의미적 분할에서 사용하는 대표적인 성능 측정 방법

- 특히 multi-class 기반의 의미적 분할에 사용됨
- mIOU를 계산하는 과정은 아래와 같음

1. Ground truth와 모델에서 출력된 prediction

0	0	0	0
0	1	1	4
5	5	2	4
5	3	3	4

Ground Truth

0	0	0	0
0	0	1	1
5	5	2	4
5	3	3	4

Prediction

2. 각 클래스별 빈도수 카운트

actual_count =

5	2	1	2	3	3

pred_count =

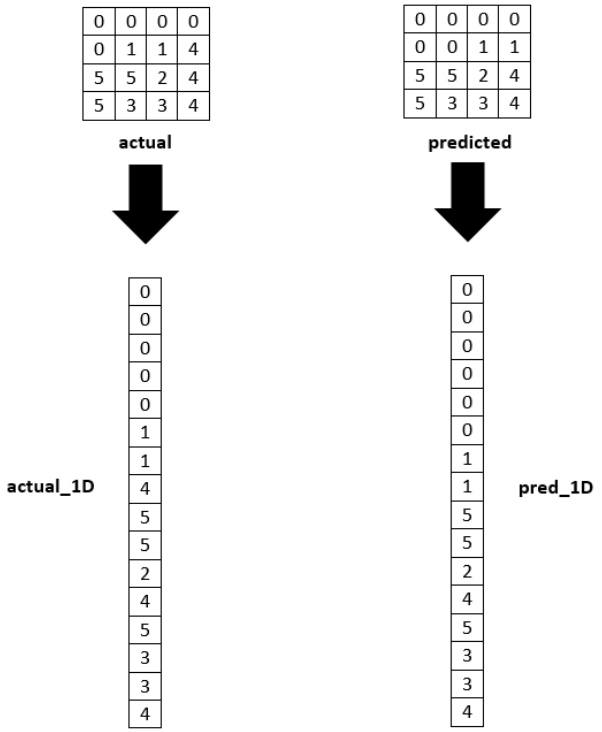
6	2	1	2	2	3

참고 자료 : mIOU

Mean Intersection Over Union의 줄임 말로, 의미적 분할에서 사용하는 대표적인 성능 측정 방법

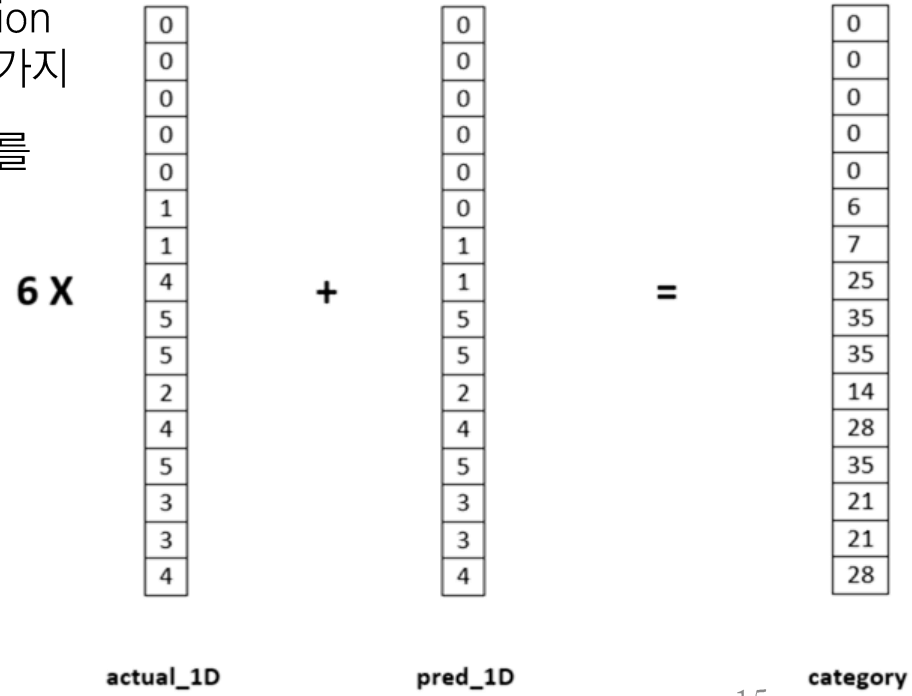
- 특히 multi-class 기반의 의미적 분할에 사용됨
- mIOU를 계산하는 과정은 아래와 같음

3. 행렬을 벡터로 변환



4. 카테고리 행렬 생성

- 클래스가 6개이므로 (GT 클래스, Prediction 클래스) 쌍이 가질 수 있는 카테고리는 36가지
- (0, 0)을 0번, (0, 1)을 1번, 마지막 (5, 5)를 35번 카테고리로 지칭



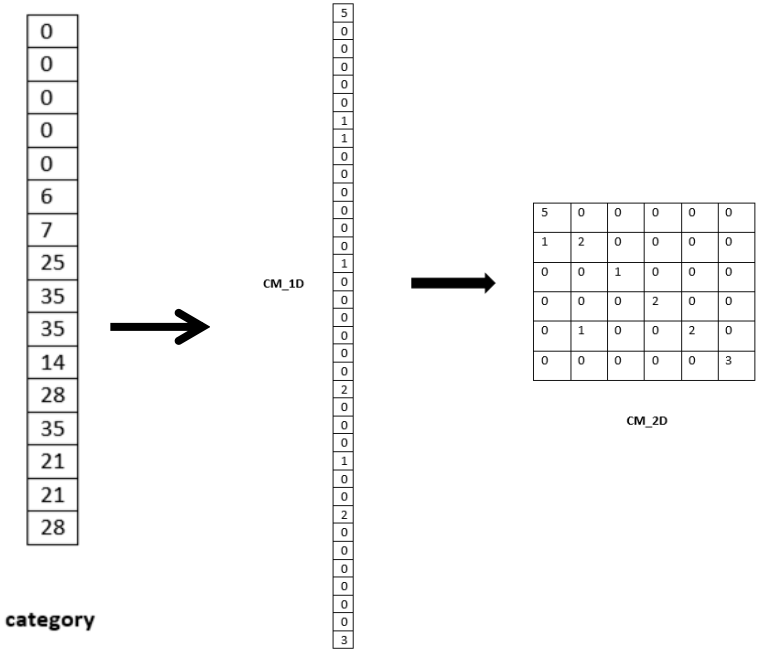
참고 자료 : mIOU

Mean Intersection Over Union의 줄임 말로, 의미적 분할에서 사용하는 대표적인 성능 측정 방법

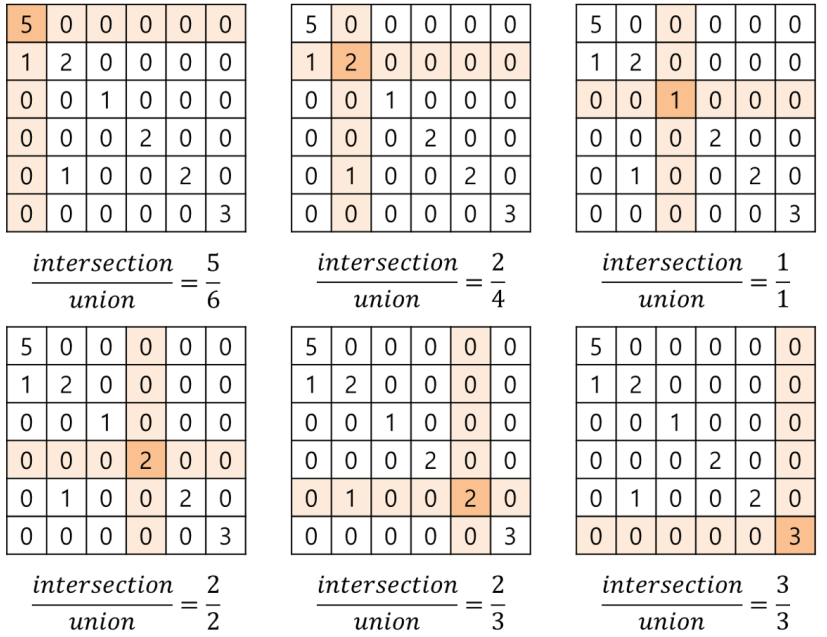
- 특히 multi-class 기반의 의미적 분할에 사용됨
- mIOU를 계산하는 과정은 아래와 같음

5. 혼동 행렬(Confusion Matrix) 생성

- 혼동 행렬의 각 원소 값은 해당 카테고리의 개수



6. 클래스별 IOU 계산



I : Intersection

5	2	1	2	2	3

U : Union

6	4	1	2	3	3

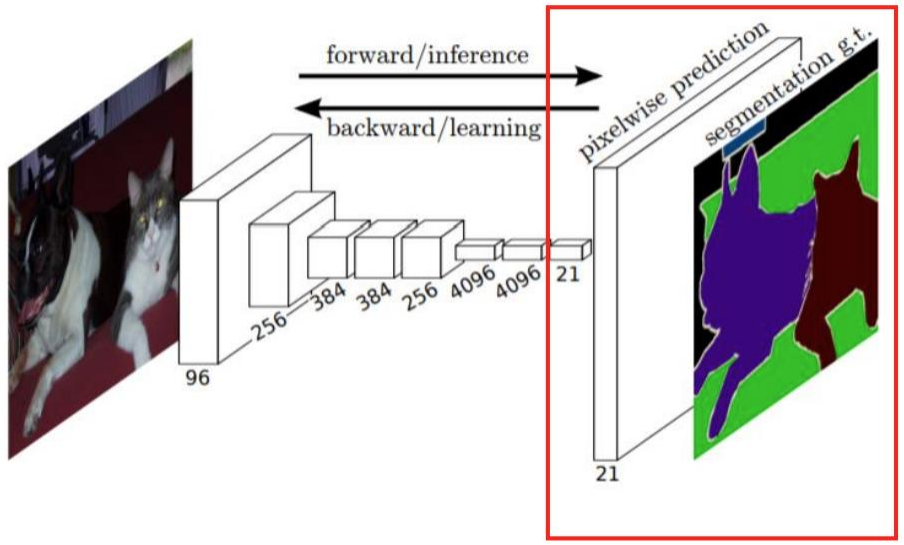
IoU = I / U

5/6	2/4	1/1	2/2	2/3	3/3

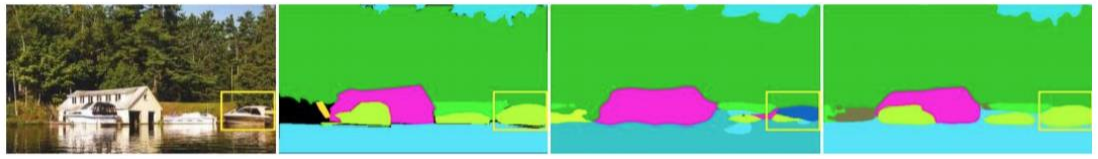
기반 모델 : PSPNet

1) 개요

[Fully Convolutional Networks (FCN)]



Problem



(a) Image (b) Ground Truth (c) FCN (d) PSPNet

context 정보 부족

→ Mismatched relationship

→ Confusion categories

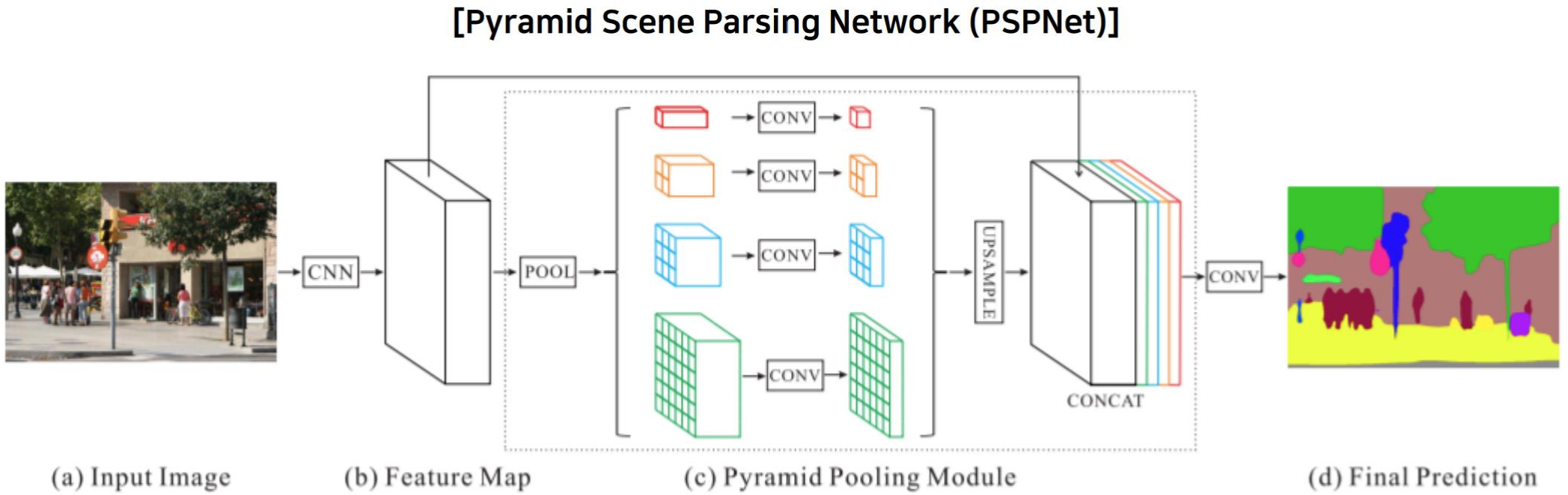
car
boat
water

Fully Convolutional Network (FCN)의 한계점을 해결하고자 제안된 의미적 분할 모델

- FCN 한계점 : Context 정보 부족으로 인한 픽셀 분류 성능 하락

기반 모델 : PSPNet

1) 개요



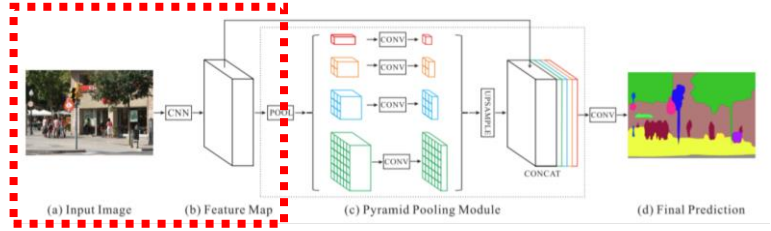
Fully Convolutional Network (FCN)의 한계점을 해결하고자 제안된 의미적 분할 모델

- FCN 한계점 : Context 정보 부족으로 인한 픽셀 분류 성능 하락

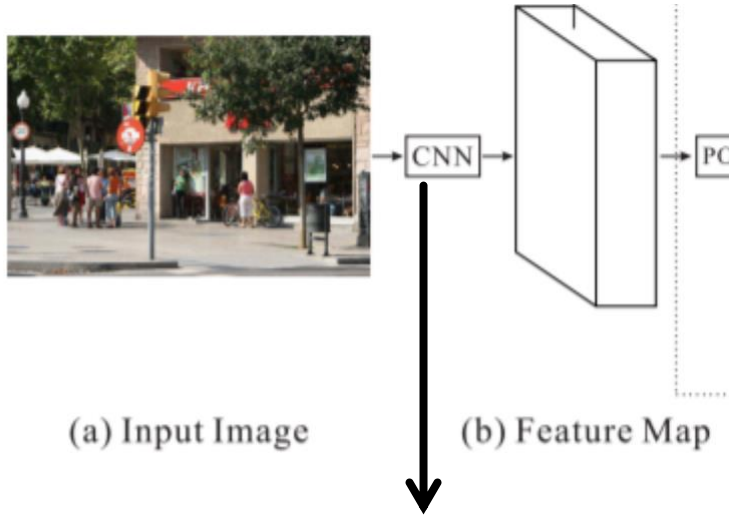
  → 제안 방법론 : Pyramid pooling module 추가 (당시 ImageNet scene parsing, Cityscapes, PASCAL VOC 등 대부분 데이터에서 SOTA 성능 기록

기반 모델 : PSPNet

2) 모델 구조



$W \times H \times 3$ $\quad\quad$ $\frac{W}{8} \times \frac{H}{8} \times N$

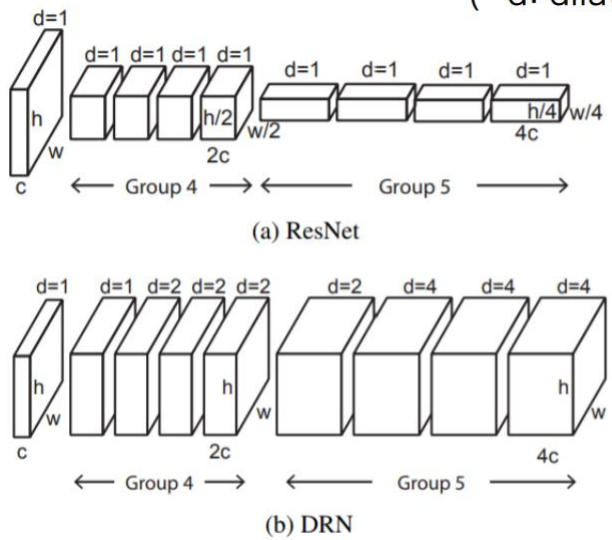


Pretrained CNN (with 분류 문제)

- Dilated Residual Network (= ResNet with dilated convolution)
- Backbone : ResNet-1010이 대표적

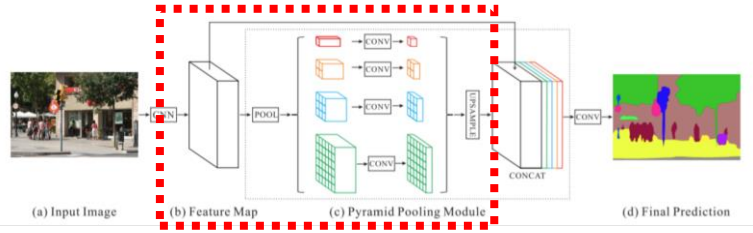
※ Dilated Residual Network

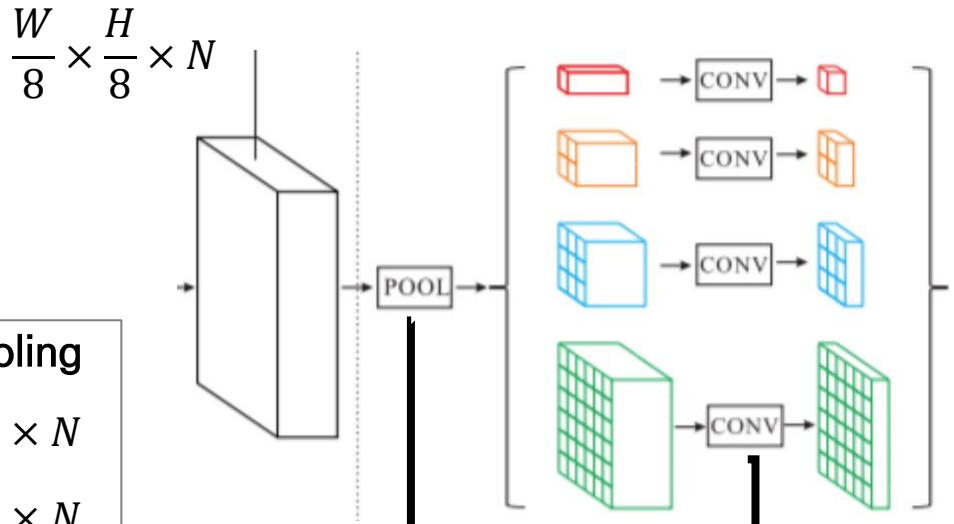
(* d: dilated rate)



기반 모델 : PSPNet

2) 모델 구조





평균 풀링 Average Pooling

- Global $1 \times 1 \times N$
- 2 x 2 $2 \times 2 \times N$
- 3 x 3 $3 \times 3 \times N$
- 6 x 6 $6 \times 6 \times N$

1 x 1 합성곱

$1 \times 1 \times 1$

$2 \times 2 \times 1$

$3 \times 3 \times 1$

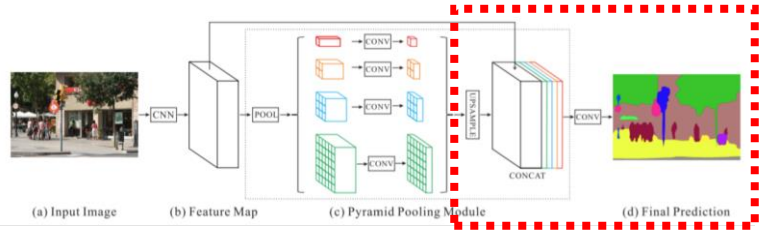
$6 \times 6 \times 1$

Pyramid Pooling Module

- 여러 스케일의 평균 풀링을 통해 rich context를 포함한 특성맵 만듦

기반 모델 : PSPNet

2) 모델 구조



Upsampling

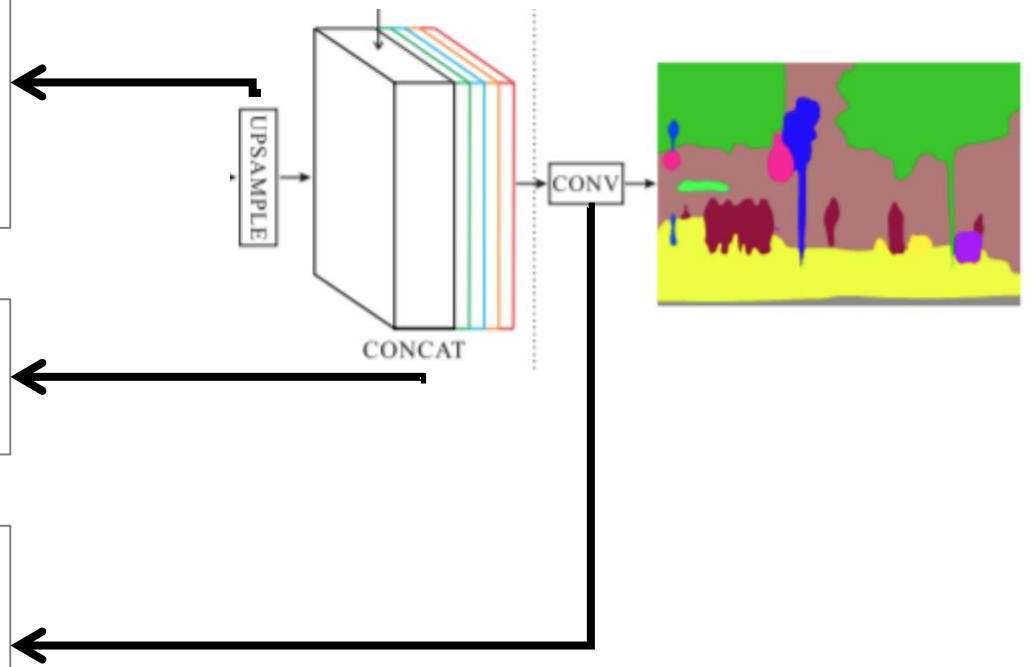
- Bilinear 보간법
- Pyramid pooling module에서 도출된 특성맵이 원래의 특성맵과 동일한 크기를 갖도록 함

특성맵 혼합 Feature Map Fusion

- 여러 스케일의 특성맵들을 하나로 합침(concatenate)

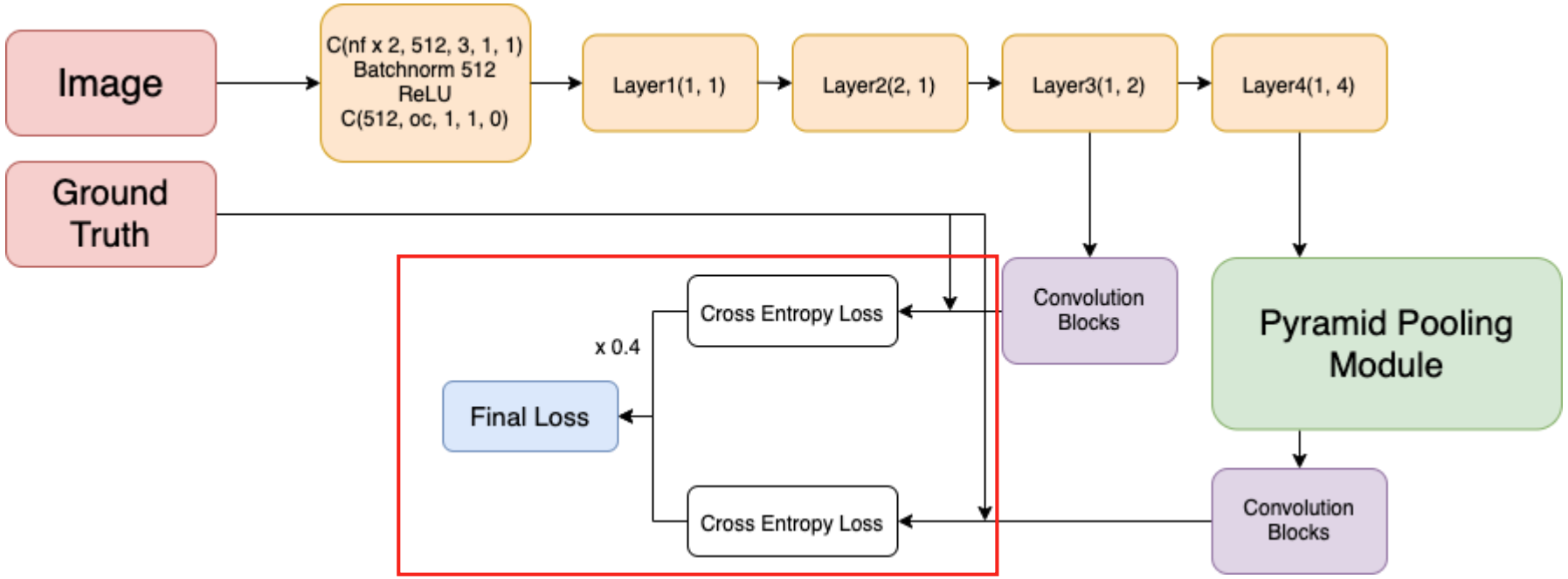
최종 예측 위한 합성곱 층

- 3x3 합성곱과 1x1 합성곱 적용함
- 예측하고자 하는 클래스 개수를 채널로 갖는 텐서 출력



기반 모델 : PSPNet

3) 손실 함수



Cross Entropy

- 픽셀 단위 분류 문제이기 때문임

Auxiliary Loss

- 기울기 소실 문제 해소 + 학습 성능 향상
- 중간 레이어 결과와 ground truth 간 loss 계산을 통해 0.4 배 만큼 최종 loss에 반영함

참고 자료 : Auxiliary Loss

학습(Train)을 잘 하도록 도와주는 보조적인 loss

- GoogLeNet (ILSVRC challenge 2014 우승)에서 처음 도입된 개념
- 신경망이 깊어질수록 기울기 소실 문제로 인해 역전파 시 기울기 전달이 잘 되지 않아, 입력층에 가까운 레이어를 학습시키기 위해 사용됨
- 예를 들어, 기존 신경망 학습은 맨 마지막 레이어에서 계산된 loss만 사용해 학습하는데, auxiliary loss는 레이어 중간중간 손실 함수를 계산하여 레이어 중간에서부터 역전파 할 수 있도록 함

  → 이를 통해 기울기가 잘 전달되지 않는 문제를 개선

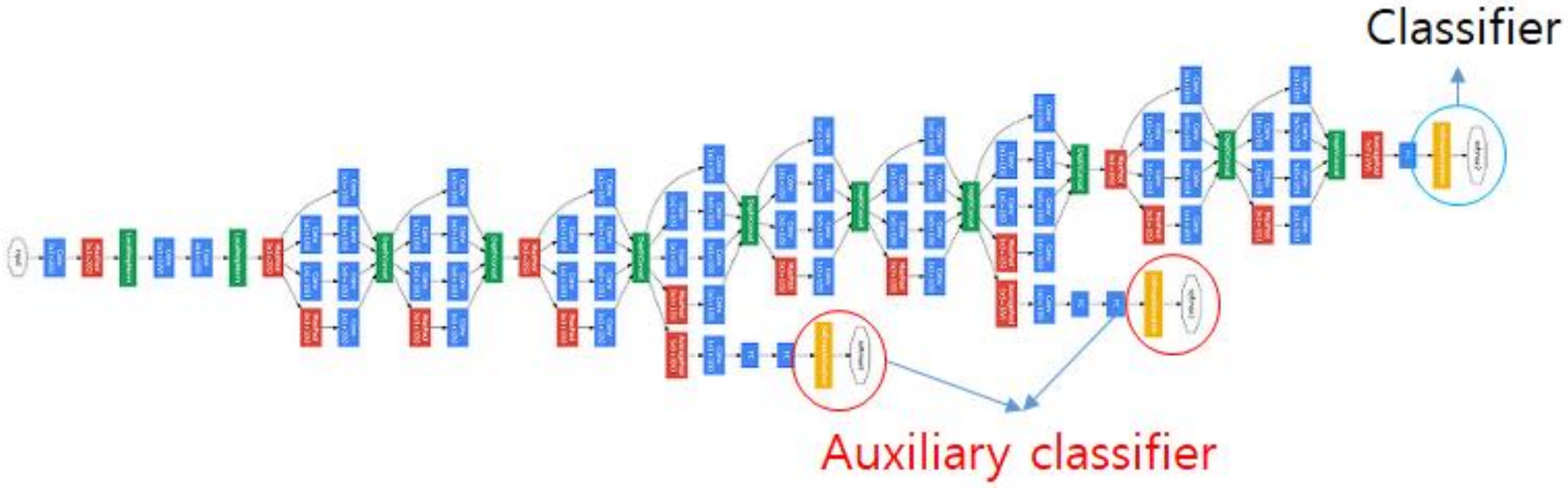
사용 시 주의점!

- 역전파 시, 너무 큰 영향을 주는 것을 방지하기 위해 auxiliary loss에만 일반적으로 0.3 정도를 곱하여 사용함
- 학습 시에만 사용하는 기법이므로, inference 과정에서는 auxiliary 관련 연산을 제외하여야 함

※ 기울기 소실

- 역전파(학습) 시 입력층에 가까운 레이어로 기울기 값이 0에 가깝게 전파되어 학습이 되지 않는 현상

참고 자료 : Auxiliary Loss



[GoogLeNet 구조]

기반 모델 : PSPNet

4) 요약

[1] PSP-Net (CVPR, 2017)

Pyramid Scene Parsing Network

Hengshuang Zhao[1] Jianping Shi[2] Xiaojuan Qi[1] Xiaogang Wang[1] Jiaya Jia[1]
[1]The Chinese University of Hong Kong [2]SenseTime Group Limited
{hszhao, xjqi, leojia}@cse.cuhk.edu.hk, xgwang@ee.cuhk.edu.hk, shijianping@sensetime.com

PSPNet 모델 특징

- Pyramid pooling module을 통한 multi-scale 특성맵 혼합
- 손실 함수에 auxiliary loss 적용

성능 측면

- mIOU, 픽셀 분류 정확도 측면에서 매우 좋은 성능
- Inference 속도는 0.78 fps로 매우 느림

▶ 정확도는 높지만, inference 속도가 매우 느림