Image Segmentation

2024-04-03

이번 차시는 Image Segmentation을 수행하는 FCN과 U-Net을 알아봅시다.

소속 : CV 2조

성명 : 박석우

목차

Computer Vision이란?

"컴퓨터 비전은 이미지나 비디오 데이터를 분석하여 인간의 시각적 인식 능력을 모방하는 기술"

"컴퓨터 비전의 주요 목표는 기계가 시각적 데이터를 '보고', '이해하며', 그 정보를 기반으로 '판단'할 수 있게 하는 것"

GPT-4

세부적인 tasks

- image classification(이미지의 클래스를 결정)
- object detection(객체를 식별하고, 위치를 bb로 표시)
- image segmentation(이미지를 클래스별로 분류)
- pose estimation(인간의 자세를 추정)
- image generation(새로운 이미지 생성)
- ...

오늘 관심있는 주제는 'Image Segmentation'

- Image segmentation은 이미지의 모든 픽셀을 클래스 레이블에 따라 분류하는 것이기 때문에 classification과 본질적으로 같은 task이다.
- semantic segmentation: 같은 클래스의 다른 인스턴스는 구분하지 않음.
- instance segmentation: 같은 클래스의 다른 인스턴스도 구분함.

Computer Vision 이란?

Image Classification

- LeNet (1998): Yann LeCun에 의해 개발된 이 컨볼루션 신경망(CNN)은 손글씨 숫자 인식에 큰 성공을 거두었으며, 딥러닝과 컴퓨터 비전 분야에서 중요한 발전으로 기록되었습니다.
- AlexNet (2012): Alex Krizhevsky, Ilya Sutskever, Geoffrey Hinton에 의해 개발된 이 모델은 2012년 ImageNet 챌린지에서 압도적인 성능을 보여주며 딥러닝 시대의 도래를 알렸습니다.
- VGGNet(2014), ResNet(2015)

---

object detection을 수행하는 R-CNN은 개발된 상태

Image Segmentation

- FCN(2015): FCN은 이미지 세그멘테이션을 위해 특별히 설계된 최초의 딥러닝 모델 중 하나로, 전통적인 컨볼루션 네트워크를 수정하여 어떤 크기의 이미지도 처리할 수 있게 하고, 출력으로 픽셀 단위의 세그멘테이션 맵을 생성합니다. 이는 업샘플링과 스킵 연결을 통해 세부적인 정보를 보존하면서 이루어집니다.
- U-Net(2015): U-Net은 의료 이미지 세그멘테이션을 위해 개발된 모델로, 이름에서 알 수 있듯이 U자 형태의 구조를 가지고 있습니다. 이 구조는 계층적 특징을 활용하여 정확한 세그멘테이션을 생성하도록 설계되었습니다. U-Net은 특히 데이터가 부족한 상황에서도 우수한 성능을 발휘합니다.

+ R-CNN(2014), Fast R-CNN(2015)

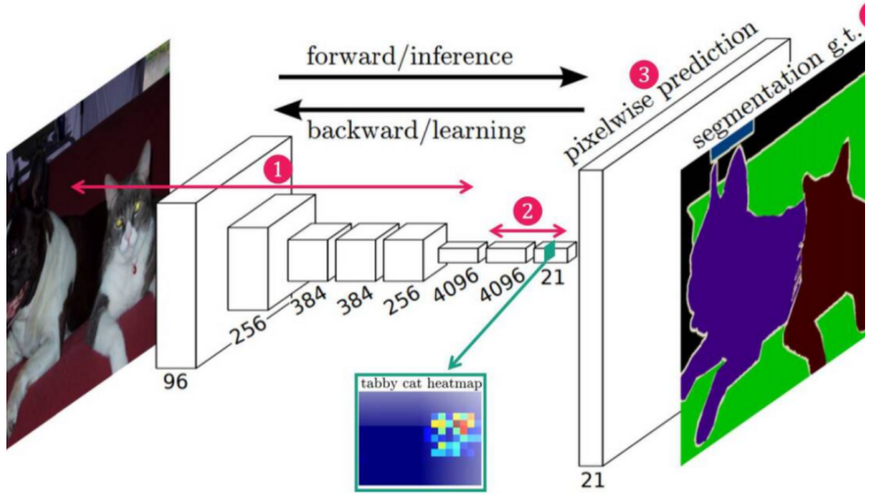
FCN: Fully Convolutional Networks for Semantic Segmenation

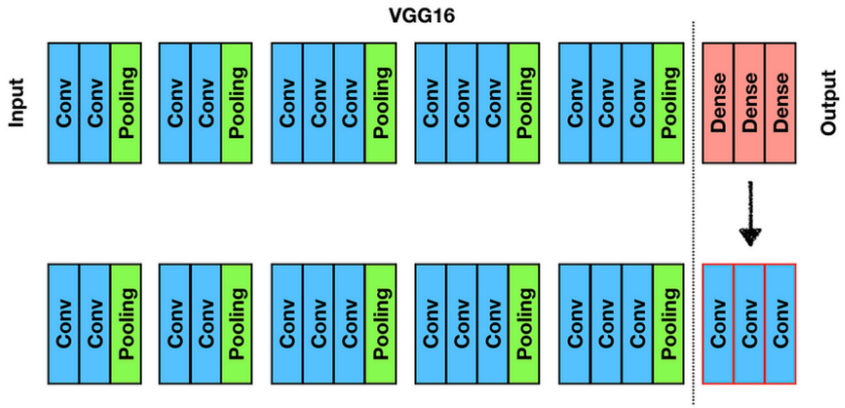
Abstract

Fully Convolutional Networks for Semantic Segmentation

Jonathan Long* Evan Shelhamer* Trevor Darrell
UC Berkeley
{jonlong,shelhamer,trevor}@cs.berkeley.edu

- FCN은 이미지 분류에서 우수한 성능을 보인 CNN 기반 모델(AlexNet, VGGNet, GoogLeNet)을 fully convolutional networks로 변환하고 semantic segmentation에 맞게 fine-tuning하는 transfer learning 방식을 사용한다.
- FCN은 새로운 구조인 skip architecture를 정의해 deep, coarse layer의 semantic 정보와 shallow, fine layer의 appearance information을 결합한다.

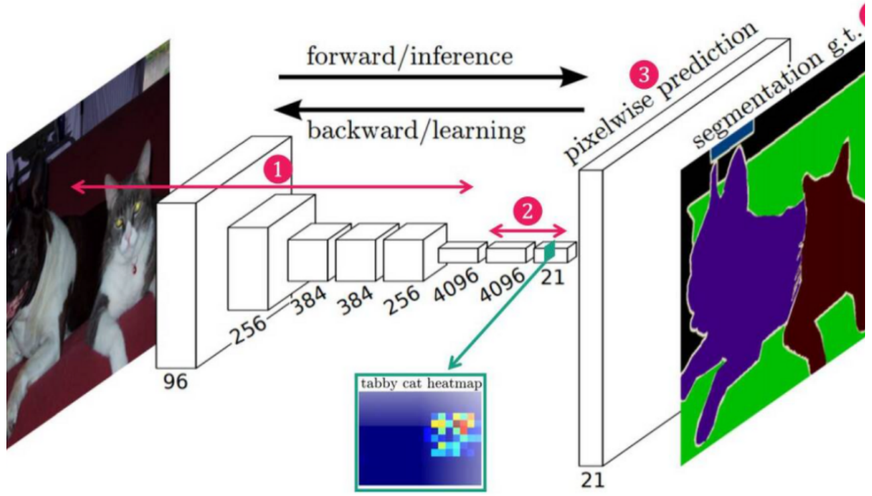


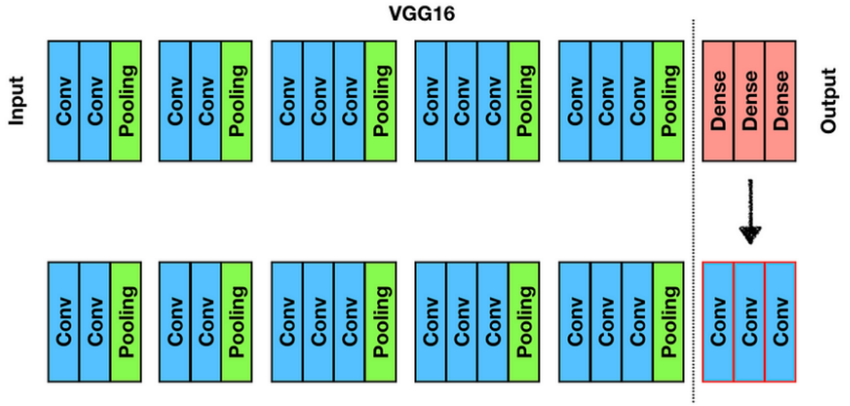


FCN: Fully Convolutional Networks for Semantic Segmenation

1. Introduction

- 기존의 연구들과의 차이점
  - end-to-end 방식
  - pixelwise prediction
  - supervised pre-training
  - input의 크기에 대한 제약이 없어짐
  - 복잡한 pre- and post- processing 사용 x
  - patchwise training x
  - skip architecture 도입





FCN: Fully Convolutional Networks for Semantic Segmenation

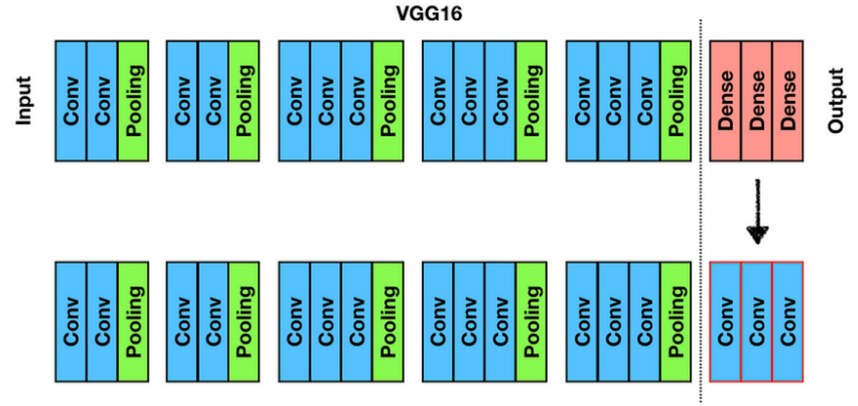
3. Fully convolutional networks

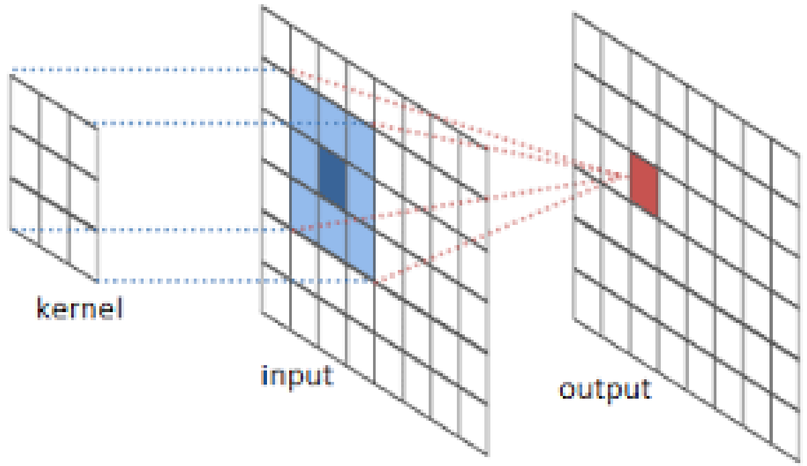
- receptive fields: high layer의 특정 뉴런에 path를 따라 연결되는 이미지의 영역, 즉 관심있는 이미지의 부분
- convnet(합성곱 신경망)은 특정 layer의 (i, j)위치의 data vector를 x라 하고 그 다음 층의 data vector y라 하면 y 다음의 식으로 정의된다.

$$\mathbf{y}_{ij} = f_{ks}\left(\{\mathbf{x}_{si+\delta_i, sj+\delta_j}\}_{0 \le \delta_i, \delta_j \le k-1}\right)$$

- 손실함수가 final layer의 각 지점에서의 loss의 합으로 정의되면 $l(\mathbf{x}; \theta) = \sum_{ij} l'(\mathbf{x}_{ij}; \theta)$

이 되기 때문에 전체 이미지에 계산되는 stochastic gradient descent는 최종 layer의 모든 receptive fields를 하나의 '미니 배치'로 여겨 stochastic gradient descent를 계산하는 것과 같다.

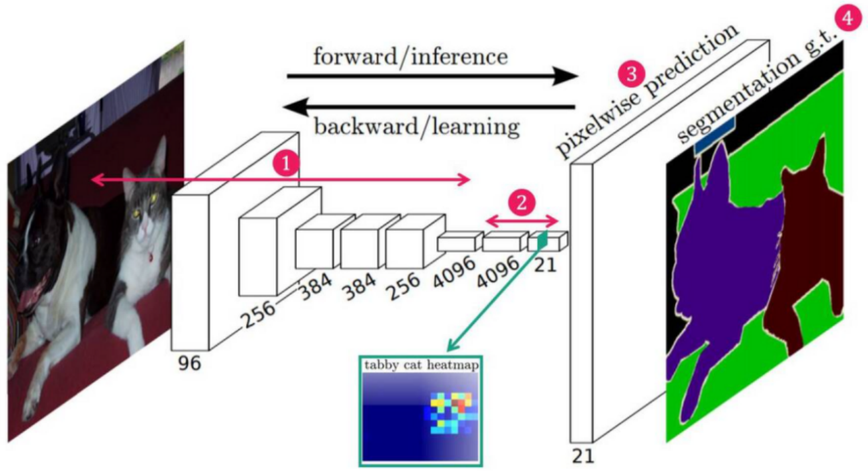


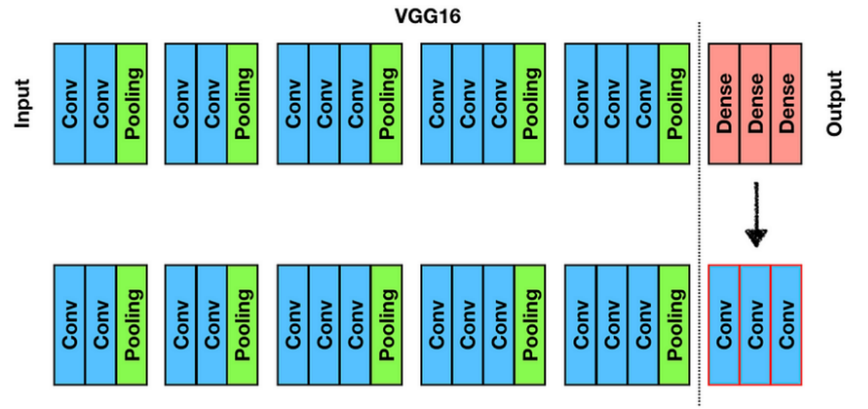


FCN: Fully Convolutional Networks for Semantic Segmenation

3. Fully convolutional networks

- FCN의 구조는 크게 4단계
  - convolution layer를 통한 feature 추출
  - 1*1 convolution layer를 통해 feautre map의 채널 수를 데이터셋 label의 개수와 동일하게 변경
  - upsampling: 낮은 해상도의 heat map을 upsampling하여 입력 이미지와 같은 크기의 map을 형성
    - shift-and-stich(filter rarefaction) v.s.
    - deconvolution
  - 최종 feature map과 label feature map의 차이를 통한 네트워크 학습
    - pathwise training v.s.
    - whole image training

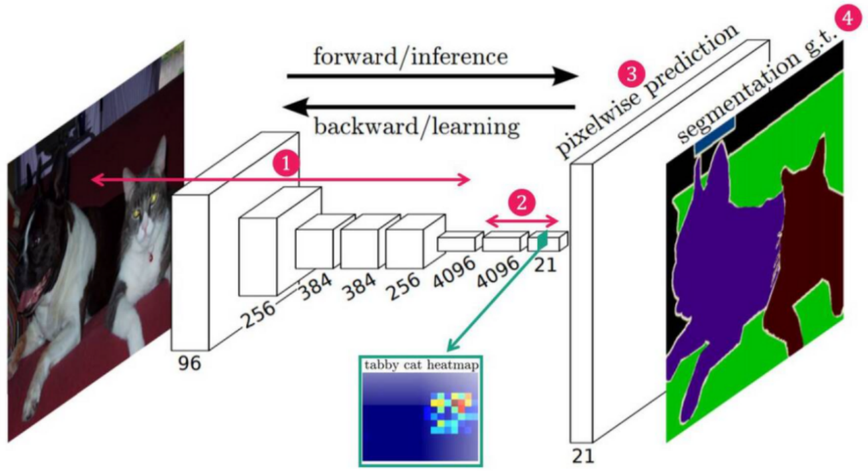


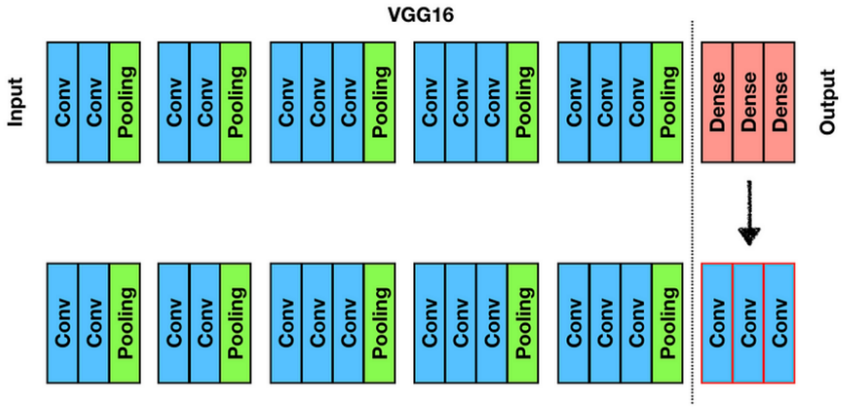


FCN: Fully Convolutional Networks for Semantic Segmenation

3. Fully convolutional networks

- FCN은 입력의 크기가 자유로운 이유?
  - 기존의 convnet은 FC layers가 존재하는데 FC layers는 고정된 수의 뉴런을 가지고 있기 때문. FCN은 이 layers 제거했다!
- 1*1 convolution layer의 역할?
  - 채널 수 조정: feature의 최종 채널 수를 분류하려는 class의 갯수만큼 지정해준다.(background 포함)
  - 피처 학습, 비선형성 추가 → 효율성 up, 표현력 개선
- 왜 하나로 구성하면 될 것을 3개씩 구성할까?
  - 복잡한 관계를 capture하기 위해서
  - 계층적 학습(상위: abstract, 하위: specific)
  - 과적합 방지, better gradient flow, flexibility





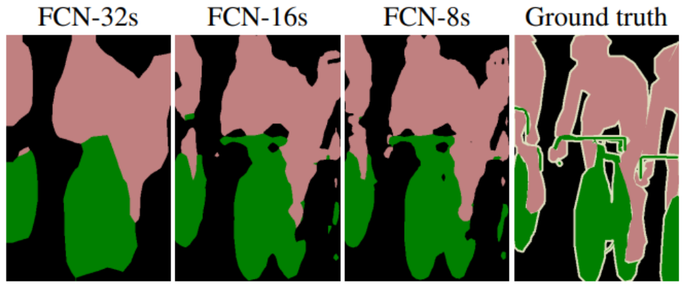
FCN: Fully Convolutional Networks for Semantic Segmenation

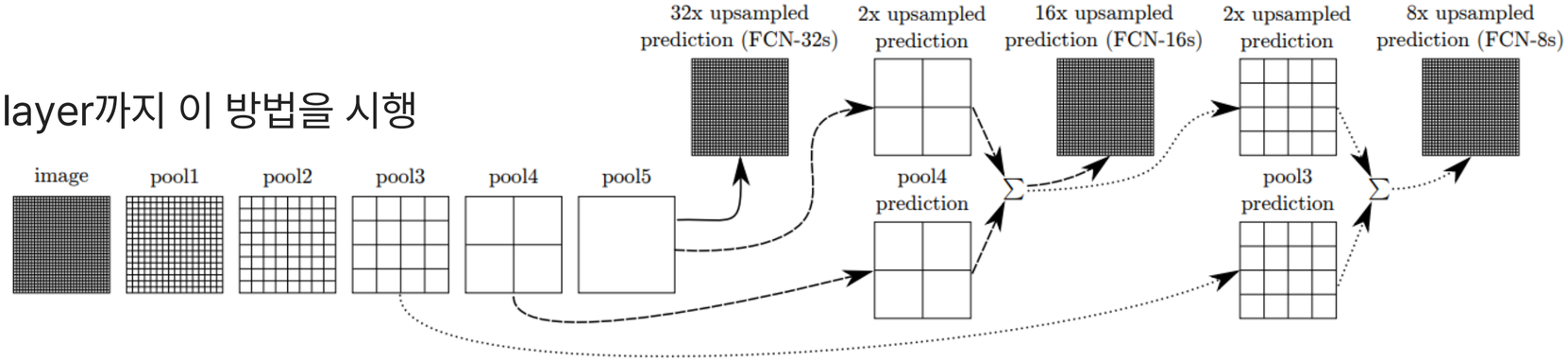
4. Segmentation Architecture

- mean pixel intersection over union(mean IoU, mean IU)
  - 예측된 분할 영역과 실제 영역 간의 겹치는 부분을 전체 합집합으로 나눈 값의 클래스별 평균(IoU = intersection/union), 0~1
- 자체적인 validation 결과 VGG16이 가장 좋은 성능
- skip architecture
  - 초기 up-sampling 결과 detail한 부분을 전혀 살리지 못함
  - fine layer(low layer)와 coarse layer(high layer)를 결합하는 'skip architecture'를 통해 이를 해결
  - 지표적 개선과 시각적 개선이 일어나는 layer까지 이 방법을 시행

Table 1. We adapt and extend three classification convnets to segmentation. We compare performance by mean intersection over union on the validation set of PASCAL VOC 2011 and by inference time (averaged over 20 trials for a 500×500 input on an NVIDIA Tesla K40c). We detail the architecture of the adapted nets as regards dense prediction: number of parameter layers, receptive field size of output units, and the coarsest stride within the net. (These numbers give the best performance obtained at a fixed learning rate, not best performance possible.)

	FCN-AlexNet	FCN-VGG16	FCN-GoogLeNet[4]
mean IU	39.8	**56.0**	42.5
forward time	50 ms	210 ms	59 ms
conv. layers	8	16	22
parameters	57M	134M	6M
rf size	355	404	907
max stride	32	32	32





FCN: Fully Convolutional Networks for Semantic Segmenation

4. Segmentation Architecture

- fine-tuning
  - 전체를 처음부터 학습시키는 것은 시간상 합리적이지 않아 classification 부분의 weights를 가져와 초기 가중치로 설정하고 전체를 fine-tuning하는 방식을 사용함.
- patch sampling
  - 원래 기존의 연구들이 patch sampling을 사용한 이유는
  - patch sampling → higher variance batches → acclerate convergence 인데
  - 본 연구는 최적화된 batch size를 변경하지 않기 위해 p의 비율로 patch sampling하면 한 배치의 이미지를 1/p배 만큼 늘림.
  - 유의미한 속도 개선보다 이미지를 더 추가한 부분이 시간에서 손해를 가져와 whole image training 방법을 채택함.

Table 2. Comparison of skip FCNs on a subset of PASCAL VOC2011 validation[7]. Learning is end-to-end, except for FCN-32s-fixed, where only the last layer is fine-tuned. Note that FCN-32s is FCN-VGG16, renamed to highlight stride.

	pixel acc.	mean acc.	mean IU	f.w. IU
FCN-32s-fixed	83.0	59.7	45.4	72.0
FCN-32s	89.1	73.3	59.4	81.4
FCN-16s	90.0	75.7	62.4	83.0
FCN-8s	**90.3**	**75.9**	**62.7**	**83.2**

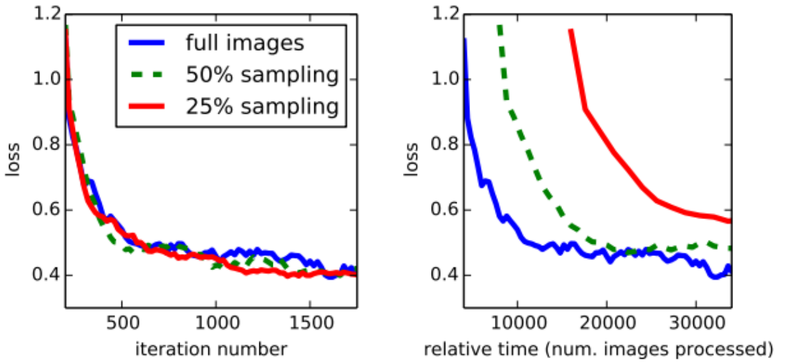


Figure 5. Training on whole images is just as effective as sampling patches, but results in faster (wall time) convergence by making more efficient use of data. Left shows the effect of sampling on convergence rate for a fixed expected batch size, while right plots the same by relative wall time.

FCN: Fully Convolutional Networks for Semantic Segmenation

5. Results

Metrics We report four metrics from common semantic segmentation and scene parsing evaluations that are variations on pixel accuracy and region intersection over union (IU). Let n_{ij} be the number of pixels of class i predicted to belong to class j, where there are n_{cl} different classes, and let $t_i = \sum_j n_{ij}$ be the total number of pixels of class i. We compute:

- pixel accuracy: $\sum_i n_{ii} / \sum_i t_i$
- mean accuracy: $(1/n_{\text{cl}}) \sum_i n_{ii}/t_i$
- mean IU: $(1/n_{\text{cl}}) \sum_i n_{ii} / \left(t_i + \sum_j n_{ji} - n_{ii}\right)$
- frequency weighted IU: $\left(\sum_k t_k\right)^{-1} \sum_i t_i n_{ii} / \left(t_i + \sum_j n_{ji} - n_{ii}\right)$

Table 3. Our fully convolutional net gives a 20% relative improvement over the state-of-the-art on the PASCAL VOC 2011 and 2012 test sets, and reduces inference time.

	mean IU VOC2011 test	mean IU VOC2012 test	inference time
R-CNN [12]	47.9	-	-
SDS [16]	52.6	51.6	$\sim$ 50 s
FCN-8s	**62.7**	**62.2**	$\sim$ **175 ms**

Table 4. Results on NYUDv2. *RGBD* is early-fusion of the RGB and depth channels at the input. *HHA* is the depth embedding of [14] as horizontal disparity, height above ground, and the angle of the local surface normal with the inferred gravity direction. *RGB-HHA* is the jointly trained late fusion model that sums RGB and HHA predictions.

	pixel acc.	mean acc.	mean IU	f.w. IU
Gupta *et al.* [14]	60.3	-	28.6	47.0
FCN-32s RGB	60.0	42.2	29.2	43.9
FCN-32s RGBD	61.5	42.4	30.5	45.5
FCN-32s HHA	57.1	35.2	24.2	40.4
FCN-32s RGB-HHA	64.3	44.9	32.8	48.0
FCN-16s RGB-HHA	**65.4**	**46.1**	**34.0**	**49.5**

FCN: Fully Convolutional Networks for Semantic Segmenation

5. Results

Metrics We report four metrics from common semantic segmentation and scene parsing evaluations that are variations on pixel accuracy and region intersection over union (IU). Let n_{ij} be the number of pixels of class i predicted to belong to class j, where there are n_{cl} different classes, and let $t_i = \sum_j n_{ij}$ be the total number of pixels of class i. We compute:

- pixel accuracy: $\sum_i n_{ii} / \sum_i t_i$
- mean accuracy: $(1/n_{\text{cl}}) \sum_i n_{ii}/t_i$
- mean IU: $(1/n_{\text{cl}}) \sum_i n_{ii} / \left(t_i + \sum_j n_{ji} - n_{ii}\right)$
- frequency weighted IU: $(\sum_k t_k)^{-1} \sum_i t_i n_{ii} / \left(t_i + \sum_j n_{ji} - n_{ii}\right)$

Table 5. Results on SIFT Flow[10] with class segmentation (center) and geometric segmentation (right). Tighe [33] is a non-parametric transfer method. Tighe 1 is an exemplar SVM while 2 is SVM + MRF. Farabet is a multi-scale convnet trained on class-balanced samples (1) or natural frequency samples (2). Pinheiro is a multi-scale, recurrent convnet, denoted RCNN$_3$ ($\circ^3$). The metric for geometry is pixel accuracy.

	pixel acc.	mean acc.	mean IU	f.w. IU	geom. acc.
Liu *et al.* [23]	76.7	-	-	-	-
Tighe *et al.* [33]	-	-	-	-	90.8
Tighe *et al.* [34] 1	75.6	41.1	-	-	-
Tighe *et al.* [34] 2	78.6	39.2	-	-	-
Farabet *et al.* [8] 1	72.3	50.8	-	-	-
Farabet *et al.* [8] 2	78.5	29.6	-	-	-
Pinheiro *et al.* [28]	77.7	29.8	-	-	-
FCN-16s	**85.2**	**51.7**	39.5	76.1	**94.3**

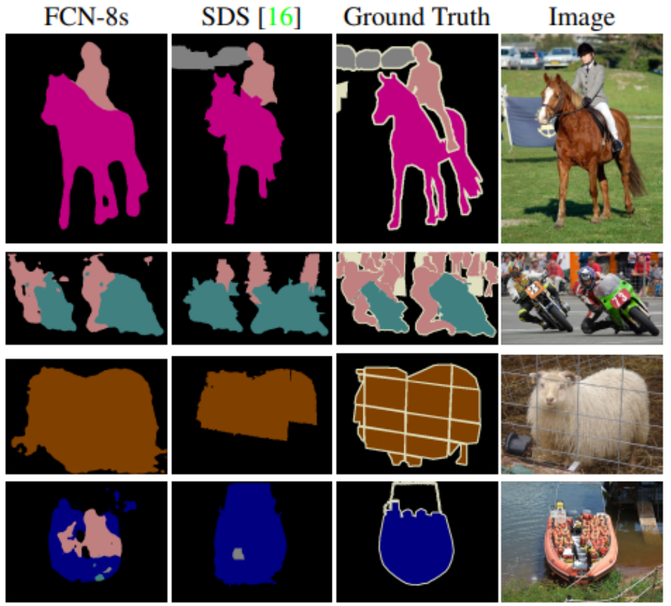


Figure 6. Fully convolutional segmentation nets produce state-of-the-art performance on PASCAL. The left column shows the output of our highest performing net, FCN-8s. The second shows the segmentations produced by the previous state-of-the-art system by Hariharan *et al.* [16]. Notice the fine structures recovered (first row), ability to separate closely interacting objects (second row), and robustness to occluders (third row). The fourth row shows a failure case: the net sees lifejackets in a boat as people.

FCN: Fully Convolutional Networks for Semantic Segmenation

6. Conclusion

- FCN은 multi-resolution layer로 architecture을 개선하여 classification을 segmentation으로 확장한 모델이다.

- 사용한 주요 techniques
  - fully convolutional layers
  - up-sampling by deconvolution
  - skip architecture