

Semantic segmentation report

Junseo Park

School of Computer Science and Engineering, Pusan National University.

1 Introduction

Semantic segmentation(의미론적 분할)은 영상 내 각 픽셀마다 의미 있는 클래스를 부여하는 컴퓨터 비전의 핵심 과제이다. 이는 한 장의 이미지 전체에 하나의 라벨을 예측하는 분류(classification)나 객체의 위치를 경계 상자로 탐지하는 객체 탐지(object detection)와 달리, 보다 세밀하고 밀집된 수준의 장면 이해를 가능하게 한다.

기존의 합성곱 신경망(CNN) 구조(VGG16 등)는 Fully Connected Layer로 인해 입력 크기가 고정되고, 공간적 정보가 손실된다는 한계가 있었다. 이를 극복하기 위해 제안된 Fully Convolutional Network(FCN)은 완전연결층을 1×1 합성곱층으로 대체하여 입력 크기에 대해 픽셀 단위 예측이 가능하도록 하였으며, upsampling과 skip connection을 통해 경계 복원과 다중 해상도 특징 결합을 동시에 수행한다.

본 보고서에서는 VGG16을 기반으로 한 FCN-8s 모델을 직접 구현하고 학습을 통해 성능을 정량적 및 정성적으로 분석한다. Bilinear Upsampling을 적용한 단순한 ConvolutionalVGG 모델과의 비교를 통해 FCN 구조가 갖는 이론적 우수성과 실험적 향상 또한 함께 검증한다.

2 Technical Background

구현한 FCN-8s 모델은 기존의 CNN 분류 네트워크(VGG16)를 기반으로 하며, 완전연결층을 convolutionalization하여 입력 이미지의 공간적 구조를 유지한 채 픽셀 단위 예측이 가능하도록 설계되었다.

2.1 CNN 기반 분류 모델 (VGG16)

VGG16은 3×3 크기의 소형 필터를 깊게 적층하여 이미지의 계층적 특징을 추출하는 대표적인 CNN 구조이다. 그러나 네트워크의 마지막 단계에 존재하는 Fully Connected Layer는 고정된 입력 크기를 요구하고, 공간 정보를 소실시킨다는 단점이 있다. 이러한 한계로 인해 VGG16은 이미지 전체를 하나의 클래스로만 분류할 수 있다.

2.2 Fully Convolutional Network (FCN)

FCN은 기존 CNN의 Fully Connected Layer를 1×1 합성곱층으로 대체하여, 입력 크기에 제약 없이 dense prediction을 수행할 수 있게 한 구조이다. 이를 통해 출력이 Feature Map 형태로 유지되며, 각 위치(pixel)에 대해 독립적인 클래스 확률을 산출할 수 있다. 특히 FCN-8s 모델은 중간 계층의 Feature Map을 결합하는 skip connection 구조와 학습 가능한 업샘플링(Transposed Convolution)을 도입한 의미론적 분할 네트워크이다.

2.3 Upsampling 방법

네트워크의 연속적인 pooling 연산으로 인해 Feature Map의 해상도는 입력 대비 크게 감소한다. 이를 복원하기 위해 FCN은 upsampling 과정을 수행한다. 단순한 Bilinear Interpolation은 고정된 수학적 보간법으로 공간적 구조를 복원하지만 학습되지 않는다. 반면 Transposed Convolution(Deconvolution)은 학습 가능한 필터를 사용하여 입력을 확장함으로써 더 정밀한 경계 복원이 가능하다.

2.4 Skip Connection의 역할

Skip Connection은 저수준의 공간 정보가 풍부한 하위 Feature Map을 고수준의 의미 정보와 결합하는 연결 구조이다. FCN-8s에서는 Pool3, Pool4, Pool5 단계의 Feature Map을 순차적으로 결합하여 세부적인 객체 경계를 보완한다. 이 구조는 단일 해상도 예측의 한계를 극복하고, 작은 객체나 복잡한 경계를 포함한 영역에서도 정확도를 향상시킨다.

3 Model Implementation and Training

3.1 ConvolutionalVGGwithUpsample

본 모델은 기존 VGG16 분류 네트워크를 Semantic Segmentation에 적합하도록 변형한 구조다. 기존의 완전연결층(fc6, fc7, fc8)을 각각 7×7 및 1×1 합성곱층으로 대체(convolutionalization)하여, 입력 이미지의 크기에 제약 없이 픽셀 단위 예측을 수행할 수 있도록 하였다. 마지막 예측 계층은 Conv2d(4096, 21, kernel_size=1)로 정의되어 PASCAL VOC 2012 데이터셋의 21개 클래스(20개 객체 + 배경)에 대한 score map을 생성한다.

모델의 forward() 경로는 입력 이미지를 VGG16의 합성곱 특징 추출부에 통과시킨 후, 변환된 합성곱 계층(fc6, fc7, fc8)을 거쳐 클래스별 예측 맵을 생성하는 방식으로 구성된다. 이후 출력된 score map은 Bilinear Interpolation을 이용해 원본 입력 크기($H \times W$)로 업샘플링된다. 이 과정은 학습되지 않는 보간법을 사용하므로, 세밀한 객체 경계나 작은 구조물의 복원에는 한계가 존재한다.

3.2 FCN8s Architecture

FCN-8s는 VGG16의 합성곱 특징 추출부를 기반으로 하며, 중간 계층의 출력(Pool3, Pool4, Pool5)을 활용하여 다중 해상도 정보를 결합하도록 구성하였다. 각 단계의 feature map에 1×1 합성곱층을 적용해 클래스별 score map을 생성하고, 가장 깊은 feature인 Pool5의 출력을 stride 2의 transposed convolution으로 2배 업샘플링한 뒤 Pool4의 score map과 결합한다. 이때 Pool4의 출력은 0.01의 가중치를 곱해 더해지며, 이는 얇은 계층의 세부 정보가 과도하게 반영되는 것을 방지하고, 고수준 의미(feature)와의 균형을 맞추기 위함이다.

이후 동일한 과정을 한 번 더 반복하여 Pool3의 score map을 추가로 결합한다. Pool3은 0.0001의 스케일로 조정되어 더해지며, 이를 통해 가장 얕은 층의 미세한 위치 정보가 안정적으로 통합된다. 마지막으로 stride 8의 transposed convolution을 통해 전체 출력을 입력 이미지와 동일한 해상도로 복원한다.

모든 deconvolution 계층은 bilinear 커널로 초기화되어 학습 초기에 안정적인 업샘플링을 수행하며, skip connection은 저수준의 위치 정보와 고수준의 의미 정보를 결합해 경계 세부를 보완한다. 최종적으로 각 픽셀에 대해 21개 클래스에 대한 확률 분포가 출력되며, 이를 통해 보다 정밀한 의미론적 분할 결과를 얻을 수 있다.

3.3 Experiment Setup

본 실험에서는 FCN-8s 모델의 학습 안정성과 성능 변화를 관찰하기 위해 세 가지 설정으로 학습을 진행하였다. 모든 실험은 PASCAL VOC 2012 데이터셋을 사용하였으며, 손실 함수는 CrossEntropyLoss를 사용하고, 최적화에는 Adam 옵티마이저를 적용하였다. 실험 조건은 아래 표와 같다.

Table 1: Training configurations for FCN-8s experiments

Experiment ID	Learning Rate	Weight Decay	Iterations
Exp. 1	1×10^{-5}	1×10^{-6}	20,000
Exp. 2	1×10^{-5}	1×10^{-5}	20,000
Exp. 3	1×10^{-5}	1×10^{-6}	40,000

각 iteration은 optimizer.zero_grad() → forward() → loss.backward() → optimizer.step() 순서로 수행되었다. 100 iteration마다 평균 손실을 기록하여 학습 수렴 양상을 관찰하였고, 최종 평가는 검증 데이터셋을 대상으로 한 mIoU(Mean Intersection over Union)로 수행하였다.

4 Experimental Results and Analysis

4.1 Training Loss Analysis

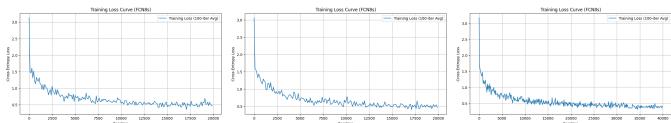


Figure 1: Training loss curves for Exp. 1–3 (100-iteration moving average)

세 실험 모두 초반 손실이 급격히 감소한 후 안정적으로 수렴하는 형태를 보였다. 수치적으로 보면 Exp. 1은 초기 500 iteration 이내에 손실이 3.12에서 1.30 수준으로 급격히 하강한 반면, Exp. 2는 동일 구간에서 3.05에서 1.32로 완화되어 감소 속도가 상대적으로 느렸다. 이는 Exp. 2에서 더 큰 weight decay(1×10^{-5})가 적용되어 정규화 효과가 강해졌기 때문으로, 파라미터 업데이트 폭이 줄어 학습이 완만해진 결과로 해석된다.

Exp. 1과 Exp. 2 모두 평균 손실이 약 0.45 수준까지 감소한 반면에 Exp. 3은 iteration을 40,000으로 늘려 충분한 학습을 수행했으며, 손실이 평균 0.35까지 안정적으로 감소하며 결과적으로 가장 높은 mIoU로 이어졌다.

4.2 Quantitative Evaluation (mIoU)

Table 2: Validation mIoU comparison across experiments

Experiment ID	Learning Rate	Weight Decay	Validation mIoU
Exp. 1	1×10^{-5}	1×10^{-6}	0.5160
Exp. 2	1×10^{-5}	1×10^{-5}	0.5129
Exp. 3	1×10^{-5}	1×10^{-6}	0.5591

세 실험 모두 기준치(0.5)를 상회하는 성능을 기록하였으며, Exp. 3이 가장 높은 mIoU(0.5591)를 달성하였다. 이는 학습 iteration을 40,000으로 늘림으로써 파라미터가 보다 충분히 수렴한 결과로 해석된다. 반면 Exp. 2의 성능이 Exp. 1보다 낮은 이유는, weight decay가 지나치게 커서 파라미터 업데이트 폭이 줄고, 결국 미세한 경계 복원 능력이 저하된 것으로 추정된다.

4.3 Qualitative Evaluation (Visualization)

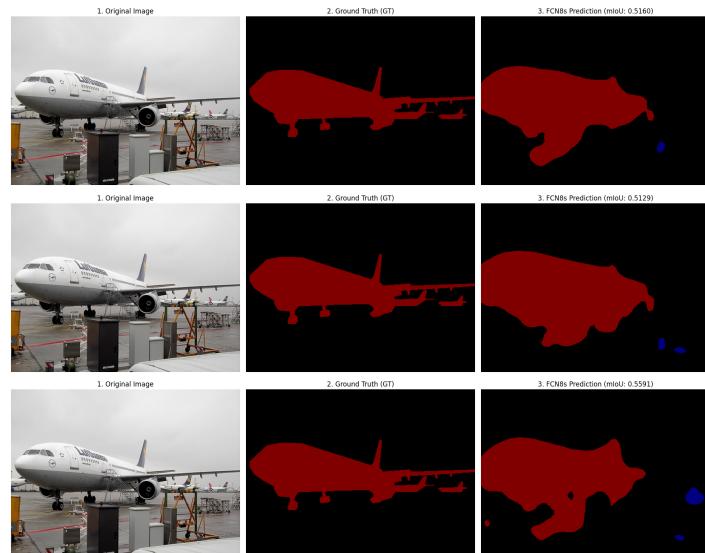


Figure 2: Visualization results of Exp. 1–3 (Original / Ground Truth / FCN8s Prediction)

시각적 결과 또한 정량적 평가와 일관된 경향을 보였다. Exp. 3의 예측 결과는 객체 경계가 보다 매끄럽고 형태 보존이 우수했으며, 배경 영역의 잡음 또한 줄어든 모습을 보였다. Exp. 1은 전반적으로 안정적인 예측을 수행했지만 경계 부근에서 일부 누락된 픽셀이 관찰되었고, Exp. 2는 객체 외곽의 분할 품질이 가장 낮았다. 이는 mIoU 순서(Exp. 3 > Exp. 1 > Exp. 2)와 정확히 일치한다.

4.4 Comparison with ConvolutionalVGGwithUpsample

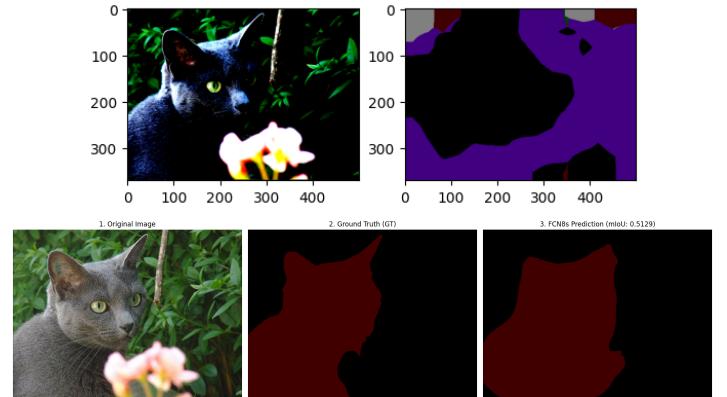


Figure 3: Comparison between ConvolutionalVGGwithUpsample (top) and FCN-8s (bottom)

Bilinear Upsampling 기반의 ConvolutionalVGGwithUpsample 모델은 업샘플링 과정이 학습되지 않기 때문에, 예측 결과가 전반적으로 거칠고 세부 경계가 손실되었다. 고양이 형태를 거의 인식하지 못하거나 일부 영역만 활성화된 모습을 보였다. 반면 FCN-8s는 skip connection과 학습 가능한 transposed convolution을 통해 고양이의 전체 윤곽을 정확히 복원하고, 객체 내부의 일관된 마스크를 생성하였다. 이는 FCN 구조가 공간적 위치 정보와 의미 정보를 효과적으로 결합한다는 점을 시각적으로 입증한다.

5 Conclusion

본 보고서에서는 VGG16을 기반으로 한 Fully Convolutional Network (FCN-8s) 모델을 직접 구현하고, Semantic Segmentation 과제에 적용하여 성능을 분석하였다. FCN-8s는 완전연결층을 1×1 합성곱으로 대체하고, 중간 계층의 Feature Map을 결합하는 skip connection 구조를 통해 공간 정보 손실을 최소화하면서 픽셀 단위의 의미론적 분할을 수행할 수 있음을 확인하였다.

세 가지 학습 설정 중 Exp. 3이 가장 낮은 손실(평균 0.35)과 높은 mIoU(0.559)를 기록하며 가장 안정적인 수렴을 보였으나, 이는 학습 iteration(40,000)을 늘린 결과로 해석된다. 따라서 Exp. 1이 본 과제에서의 표준적 구현으로서 가장 합리적인 학습 효율과 성능 균형을 보여주었다.

또한 Bilinear Upsampling 기반의 단순한 ConvolutionalVGG 모델과 비교했을 때, FCN-8s는 학습 가능한 업샘플링 계층(Transposed Convolution)과 skip connection을 통해 객체의 경계를 보다 정밀하게 복원하고 의미론적 일관성이 향상된 출력을 생성하였다. 이를 통해 FCN 구조가 Semantic Segmentation 문제에서 갖는 구조적 이점을 실험적으로 입증하였다.

결론적으로, 본 실험을 통해 FCN-8s 모델은 학습 안정성, 분할 정밀도, 그리고 구조적 일반화 측면에서 모두 유효함을 확인하였다. 다만, 추가적인 하이퍼파라미터 조정이나 학습 스케줄 최적화 등을 통해 성능을 더욱 향상시킬 여지가 남아 있으며, 이는 향후 연구 또는 후속 실험에서 탐구해볼 가치가 있다.

6 Reference

- Jonathan Long, Evan Shelhamer, and Trevor Darrell. *Fully Convolutional Networks for Semantic Segmentation*. CVPR, 2015.