

# Fully Convolutional Networks for Semantic Segmentation

Jonathan Long, Evan Shelhamer, Trevor Darrell

Summarized by 202355514 강지원

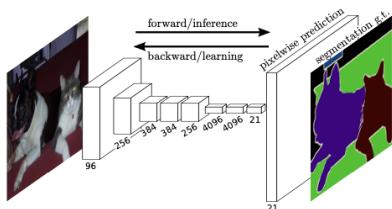


Figure 1: Fully convolutional networks can efficiently learn to make dense predictions for per-pixel tasks like semantic segmentation

이 연구는 딥 분류 네트워크를 FCN으로 변형하여 시맨틱 세그멘테이션을 위한 밀집 예측을 가능하게 하는 새로운 방법을 제시한다. 기존의 방식은 하이브리드 제안 분류기 모델을 사용하여 바운딩 박스와 지역 제안을 샘플링하고 학습하는 방식이지만, 이는 end-to-end 학습을 하지 않는다. 반면, 본 연구는 FCN을 전이 학습을 통해 end-to-end 방식으로 훈련하여 효율적인 밀집 예측을 구현한다. 또한, 패치 단위 훈련과 shift-and-stitch 방식으로 밀집 출력을 FCN 관점에서 효율적으로 처리하며, 후처리 기법 없이 간단하게 예측을 수행한다.

FCN은 변환 불변성(transformation invariance)을 기반으로 하여 로컬 입력 영역에서 작동하며, 각 레이어는 커널 크기와 stride에 따라 비선형 필터를 계산하여 어떤 크기의 입력에도 대응할 수 있는 네트워크를 생성한다. 연구에서는 FCN을 사용해 픽셀 단위 예측을 수행하는 방법을 제시하고, 손실 함수가 마지막 레이어의 공간 차원에 합산될 때, 전체 이미지에 대한 확률적 경사하강법(SGD)을 통해 각 수용 영역을 미니배치로 처리하는 방식과 동일하다는 점을 설명한다. 이를 통해 파라미터를 효율적으로 업데이트하여, feedforward 연산과 역전파의 효율성을 크게 향상시킨다.

$$y_{ij} = f_{ks} \left( \left\{ x_{s(i+\delta_i), s(j+\delta_j)} \right\} \quad 0 \leq \delta_i, \delta_j \leq k \right)$$

위 수식은 각 레이어가 로컬 입력 영역을 기반으로 데이터를 처리하는 방식과 이를 통해 픽셀 단위 예측을 수행하는 방법을 설명한다. 여기서  $y_{ij}$ 는 특정 위치  $(i, j)$ 에서의 출력 데이터를 나타내고,  $x_{ij}$ 는 이전 레이어에서 해당 위치에 대응하는 데이터이다. 함수  $f_{ks}$ 는 컨볼루션에서는 행렬 곱셈, 풀링에서는 평균이나 최댓값 등을 계산하며, 커널 크기  $k$ 와 스트라이드  $s$ 를 사용하여 로컬 영역을 처리한다. 이 방식은 변환 불변성(transformation invariance)을 기반으로 하여, 어떤 크기의 입력에도 대응할 수 있도록 네트워크가 설계된다는 특징을 가진다.

샘플링을 통한 훈련은 클래스 불균형과 공간적 상관성을 해결할 수 있지만, 밀집 예측에서 더 빠르거나 더 좋은 수렴 결과를 얻지는 못했다. 따라서 전체 이미지 훈련이 패치 단위 학습보다 더 효율적임을 보였다. 이 연구에서 제시된 아키텍처는 ILSVRC 분류기를 FCN으로 변형하고, 픽셀 단위 손실을 통해 밀집 예측을 수행하며, fine-tuning을 통해 세그멘테이션을 학습하고 새로운 skip 아키텍처로 예측을 정제한다.

완전 연결층을 컨볼루션으로 바꾸는 방법을 설명하며, 일반적인 인식 네트워크는 고정된 크기의 입력을 받아 비공간적 출력을 생성하지만, 완전 연결층을 입력 영역을 모두 포함하는 커널을 이용한 컨볼루션으로 변환함으로써, 네트워크가 입력 크기와 관계없이 분류 맵을 출력하는 FCN이 될 수 있다. 이를 통해 계산 효율성과 예측 속도를 크게 향상시킬 수 있다.

Overfeat에서 소개된 입력 이동과 출력 교차 기법은 보간법 없이 coarse output에서 밀집 예측을 생성하는 방법으로, 출력이  $f$ 만큼 다운 샘플링된 경우, 0부터  $f-1$  픽셀까지 오른쪽과 아래로 이동하며 생성된  $f^2$ 개의 입력을 convNet에 통과시켜 해당 출력이 수용영역 중심에 맞도록 만든다. 또한, convNet에서 필드와 stride를 변경하여 shift-and-stitch 기법과 동일한 출

력을 생성할 수 있음을 보였다.

앞서 설명한 방식대로 AlexNet, VGGNet, GoogLeNet 이미지 분류 네트워크를 FCN으로 변환하여 시맨틱 세그멘테이션을 수행한다. 각 네트워크에서 최종 분류기 레이어를 제거하고, 완전 연결 레이어를 컨볼루션 레이어로 변환하여 dense prediction을 가능하게 한다. 이후 디컨볼루션을 사용하여 픽셀 단위 예측을 생성한다. 이 과정에서 fine-tuning을 통해 각 네트워크가 시맨틱 세그멘테이션 문제를 해결할 수 있도록 한다. VGG-16은 이미 56.0의 평균 IU로 최첨단 성능을 달성했으나, GoogLeNet은 분류 정확도에서는 비슷한 성능을 보였으나 세그멘테이션 성능에서 미흡함을 보였다.

FCN 분류기는 fine-tuning을 통해 세그멘테이션에 적합하도록 변환할 수 있지만, 그 출력이 너무 거칠어서 세부 사항을 잘 표현하지 못하는 문제가 있다. 이를 해결하기 위해 거친 레이어와 세밀한 레이어를 결합하여 로컬 예측을 더 정밀하게 하고, 전역 구조를 반영할 수 있게 만든다. 이를 통해 deep jet라는 새로운 비선형 로컬 특징 계층을 도입해 세그멘테이션 성능을 향상시킨다.

FCN은 여러 층에서 나온 예측 결과를 결합하여 더 정확한 출력을 만든다. 처음에는 16픽셀 stride에서 예측을 시작하고, 이를 32픽셀 stride 예측과 결합하여 성능을 향상시킨다. FCN-16s로 학습한 후, 추가 결합을 통해 FCN-8s로 성능을 약간 더 개선하지만, 이후 결합은 큰 차이를 만들지 않아 종료하였다. 이 방법으로 mean IU 성능이 62.4에서 62.7로 향상되었다. 따라서 풀링 스트라이드를 줄이거나 shift-and-stitch 기법을 사용해도 VGG-16 네트워크 성능 향상에 제한이 있어 레이어 결합 방법이 더 효과적임을 알 수 있다.

FCN-8s는 PASCAL VOC에서 20% 향상된 성능을 보였으며, 추론 시간이 크게 단축되었다. NYUDv2에서는 RGB-D 입력을 사용한 여러 실험을 진행했지만, 초기 융합 방식에는 큰 이점이 없었고, HHA와 RGB의 후속 융합이 더 효과적이었다. 또한, FCN-16s 모델을 사용해 시맨틱(예: "bridge", "mountain", "sun")과 기하학적(예: "horizontal", "vertical", "sky") 예측을 동시에 수행한 결과, 두 가지 작업에 대해 독립적인 모델을 학습한 것과 동일한 성능을 보였다.

결과적으로 FCN은 모던 분류용 컨볼루션 네트워크의 특수한 경우로, 세그멘테이션을 위한 강력한 모델이다. 분류 네트워크를 세그멘테이션으로 확장하고, 다중 해상도 레이어 조합을 통해 성능을 향상시키고 학습 추론 속도 또한 개선시켰다. 이를 통해 성능과 효율성을 동시에 달성할 수 있었다.

[1] FCN 정의: <https://velog.io/@kwoonho/FCN>

[2] translation invariance 부가 설명: <https://ganghee-lee.tistory.com/43>