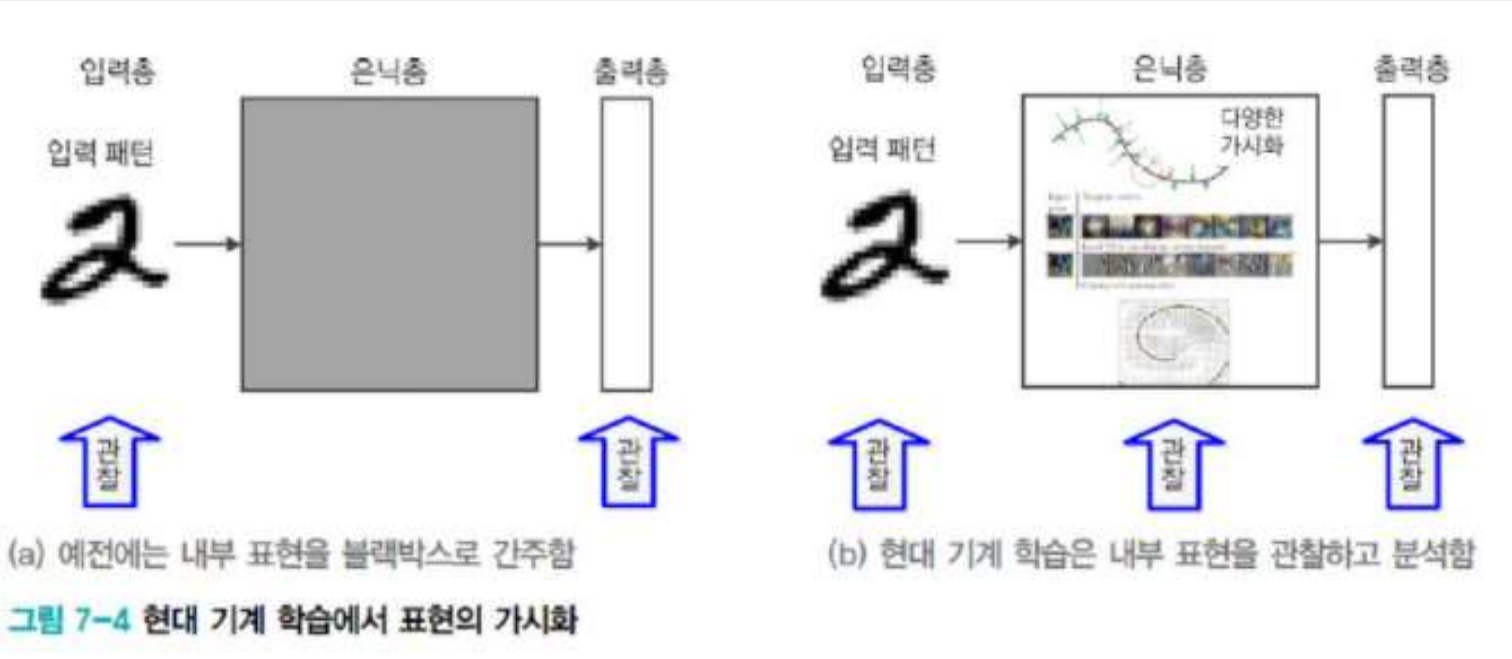
기계 학습

2021210088 허지혜

목차

7.2 내부 표현의 이해



(a) 예전에는 내부 표현을 블랙박스로 간주함

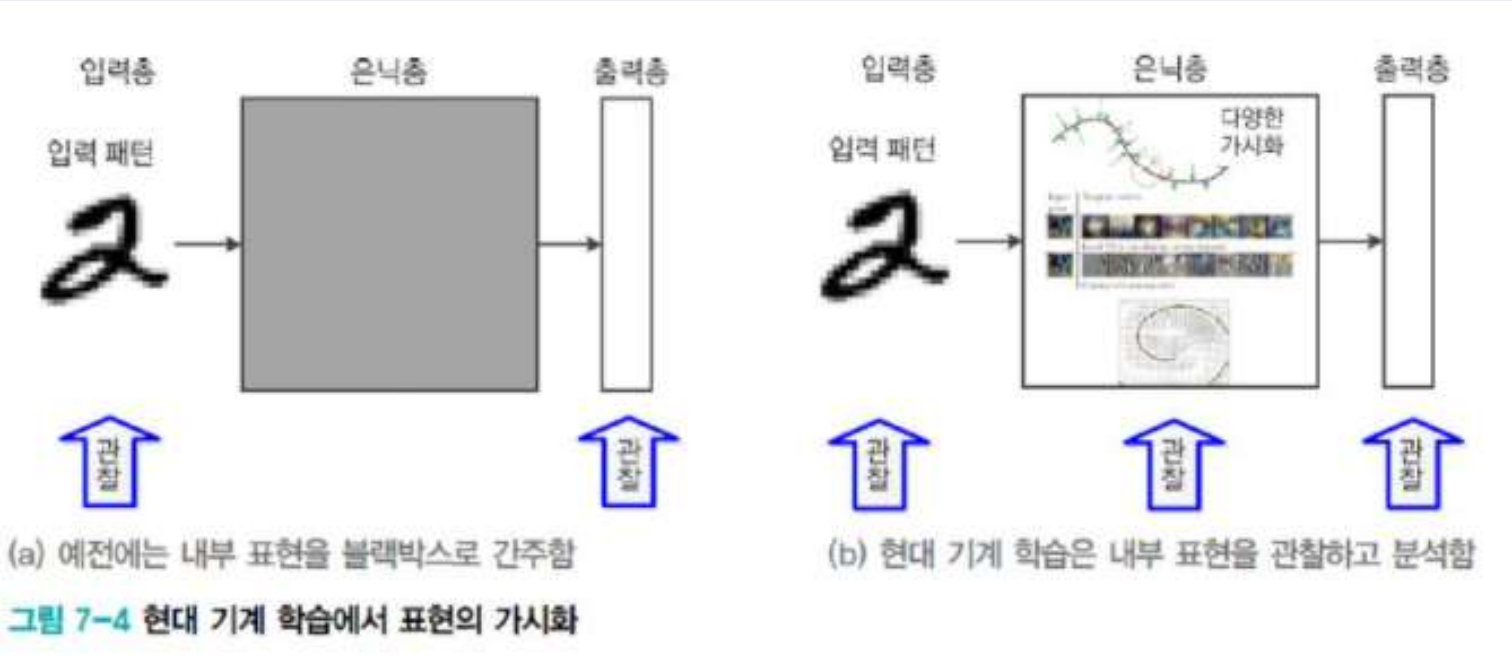
(b) 현대 기계 학습은 내부 표현을 관찰하고 분석함

그림 7-4 현대 기계 학습에서 표현의 가시화

- 그림은 현대 기계 학습에서 나타나는 중요한 경향을 설명한다.

- 옛날과 비교하여 현재는, 학습 결과로 얻은 '내부표현' 즉, 은닉층의 내용을 자세히 들여다보고 분석하려는 노력이 많아졌다.

7.2 내부 표현의 이해



(a) 예전에는 내부 표현을 블랙박스로 간주함 (b) 현대 기계 학습은 내부 표현을 관찰하고 분석함

그림 7-4 현대 기계 학습에서 표현의 가시화

- 내부 표현을 블랙박스로 간주하면 준지도 학습 또는 전이 학습을 제대로 설계할 수 없다.

- 따라서 내부 표현을 가시화(Visualization)하는 기법을 살펴볼 것이다.
현재 가시화 연구가 가장 많이 진행된 기계 학습 모델은 CNN이다.

7.2.1 컨볼루션 필터의 가시화

•학습을 마친 신경망의 내부 표현을 가시화 하는 방법은 다음과 같다.

① 필터 가시화
② 특징 맵 가시화
③ 역투영 가시화

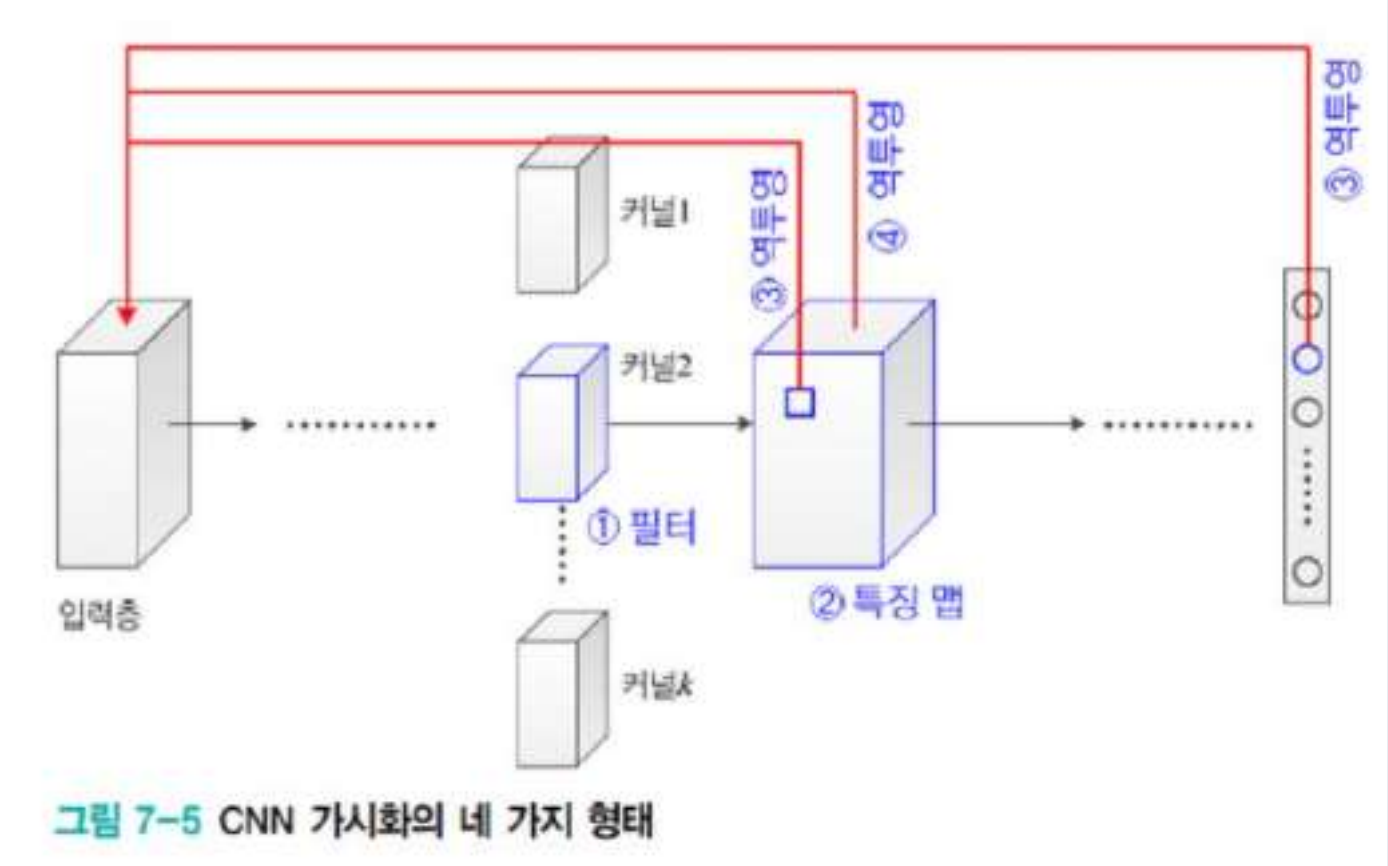


그림 7-5 CNN 가시화의 네 가지 형태

7.2.1 컨볼루션 필터의 가시화

•가장 쉽고 단순한 방법은 '①**필터 가시화**' 이다.
그림은 ImageNet 데이터로 학습한 AlexNet의 첫 번째 컨볼루션층 필터이다.

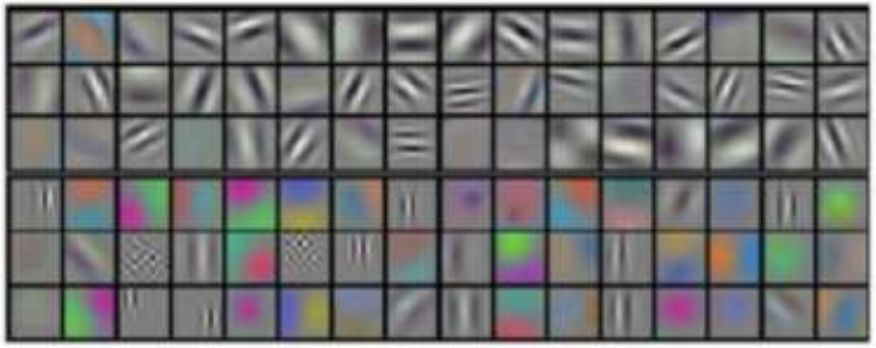


그림 7-6 첫 번째 컨볼루션 층 필터

•**관찰 결과**

첫번째 컨볼루션 층에서는 에지나 블롭이 주로 나타난다.
특정한 데이터 또는 특정한 과업과 무관하게 나타나는 일반적인 현상으로 밝혀졌다.

7.2.2 특징 맵의 가시화

•그 다음은 필터를 적용하여 얻은 '② **특징 맵 가시화**' 하는 방법이다.
특정 맵의 화소는 신경망의 노드에 해당하기 때문에 노드의 활성값을 가시화하는 셈이다.
아래 그림은 가시화 도구에서 획득한 화면 영상이다.

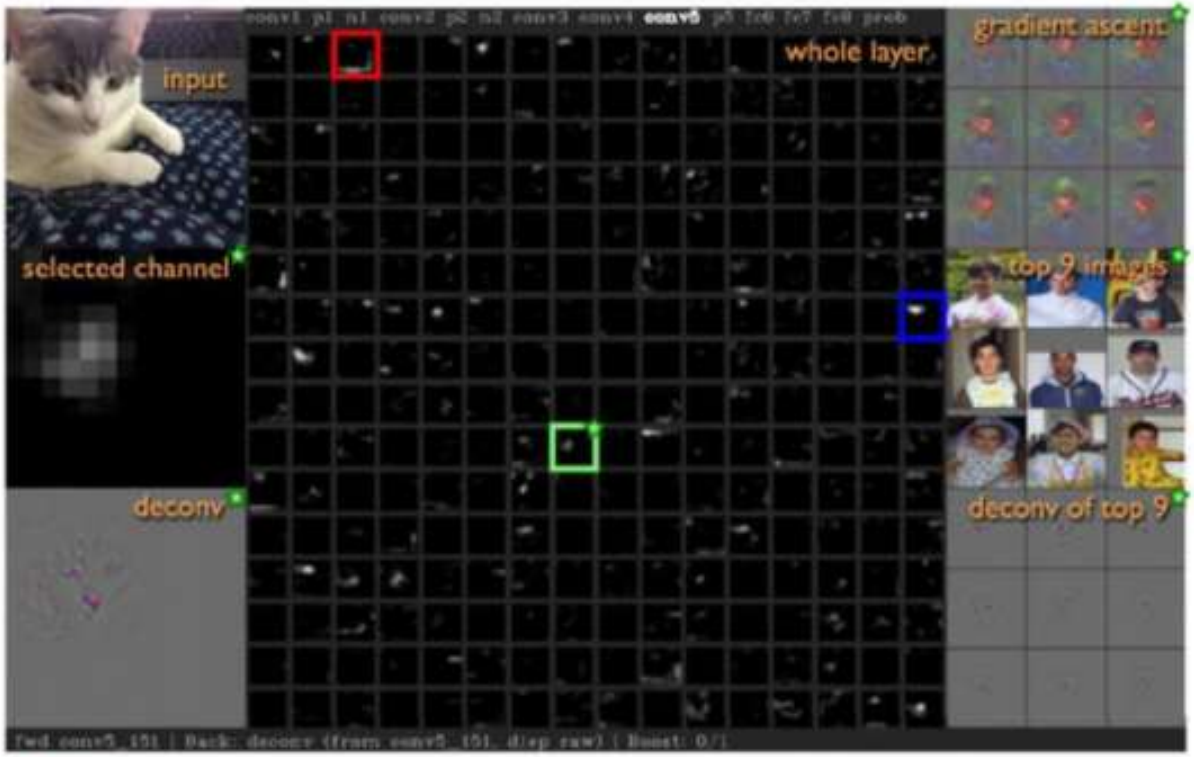


그림 7-7 가시화 도구 화면

7.2.2 특징 맵의 가시화

- 다른 영상을 입력해도 녹색 특징 맵은 얼굴 부위가 활성화됨을 확인할 수 있다.

그림 7-8 얼굴을 검출하는 특징 맵

7.2.2 특징 맵의 가시화

- 이러한 현상은 깊은 신경망에 대한 기존 해석과 다르다.
전 : 모든 층에 걸쳐 정보가 여러 노드로 분산하여 처리된다고 믿음.
후 : 몇 개의 층을 지나면 표현이 지역적으로 된다는 사실을 보임.

- 근데 이 신경망은 1000 부류의 ImageNet으로 학습하였는데 얼굴 부류가 없다.
그런데도 사람이나 동물과 같은 부류를 인식하기 위해 이 물체들을 구성하는 얼굴이라는 구성요소를 알아냈다고 볼 수 있다.
또한, 얼굴의 크기, 방향, 조명, 배경 등이 변하여도 관련 없이 성공적으로 검출한다는 사실을 알 수 있다.

- 이러한 사실은 전이 학습을 설계하는데 유용한 지침이 될 수 있다.

7.2.3 영상 공간으로 역투영

•앞의 두 가시화 기법은 전방 계산 과정에서 발생하는 필터 또는 특징 맵을 보여주는 수준이다.
최근에는 하나의 노드 또는 같은 층에 있는 여러 뉴런의 집합을 활성화하는 입력 신호,
즉, 입력 공간에서의 영상 또는 영역을 찾아 보여 주는 가시화 기법으로 발전하였다.
그런 가시화 기법을 '③ **역투영 가시화**'라고 한다.

• 은닉 노드를 역투영 할 수도 있고 출력 노드를 역투영 할 수도 있다.

7.2.3 영상 공간으로 역투영

- 최적화를 이용한 역투영

관찰하고자 하는 뉴런을 i로 표기하고$a_i(x)$를 영상 x가 입력되었을 때 뉴런 i의 활성값이라고 하면 역투영 문제는 식 (7.1)로 표현된다.

$$\hat{x} = \underset{x}{argmax}(a_i(x)) \qquad (7.1)$$

식 (7.1)의 최적화 문제를 식 (7.2)의 경사 상승법으로 푸는 방법을 제시한다.

$$x_{t+1} = x_i + \eta \frac{\partial a_i(x)}{\partial_x} \qquad (7.2)$$

실제로는 여러 규제 기법을 적용하여 푼다.

$$x_{t+1} = r_\theta(x_i + \eta \frac{\partial a_i(x)}{\partial_x}) \qquad (7.5)$$

7.2.3 영상 공간으로 역투영

- 다컨볼루션을 이용한 역투영

CNN이 사용하는 연산을 역으로 적용하여 시각화하는 다컨볼루션(deconvolution) 기법도 있다. 다컨볼루션은 입력 영상 I를 제공해야 동작하는 기법이다.

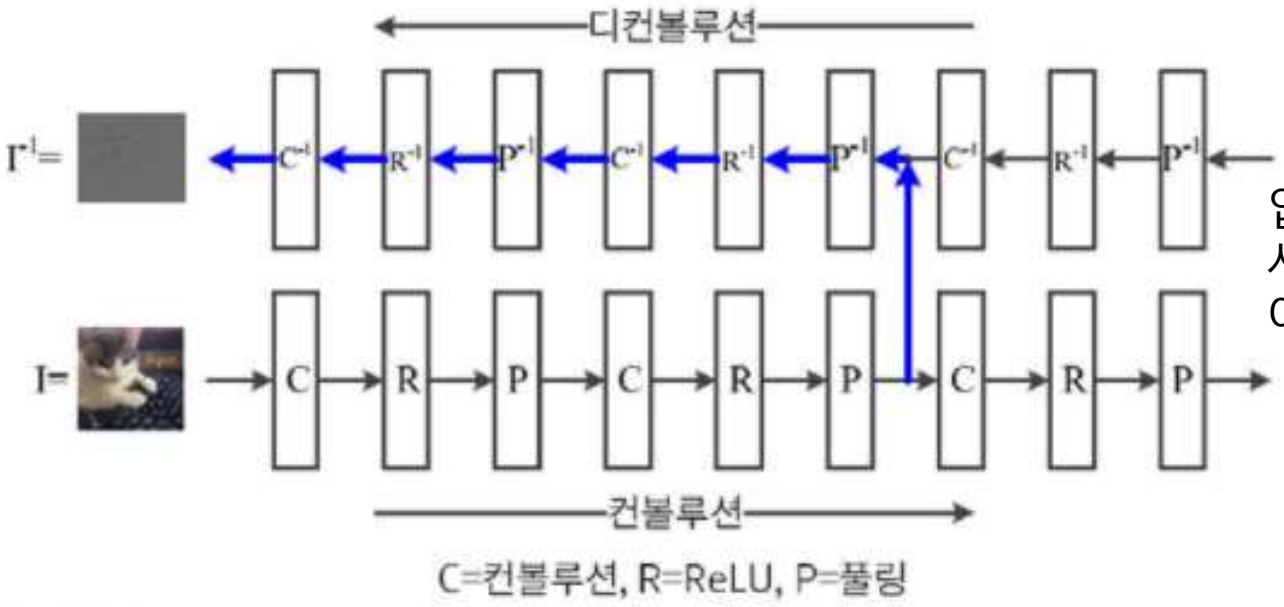


그림 7-9 디컨볼루션의 동작 원리

입력 영상을 가지고 전방 계산을 수행 후 사용자가 지정한 층에서 출발해 역방향으로 여산하면서 입력 공간까지 진행한다.

7.2.3 영상 공간으로 역투영

- 다컨볼루션을 이용한 역투영

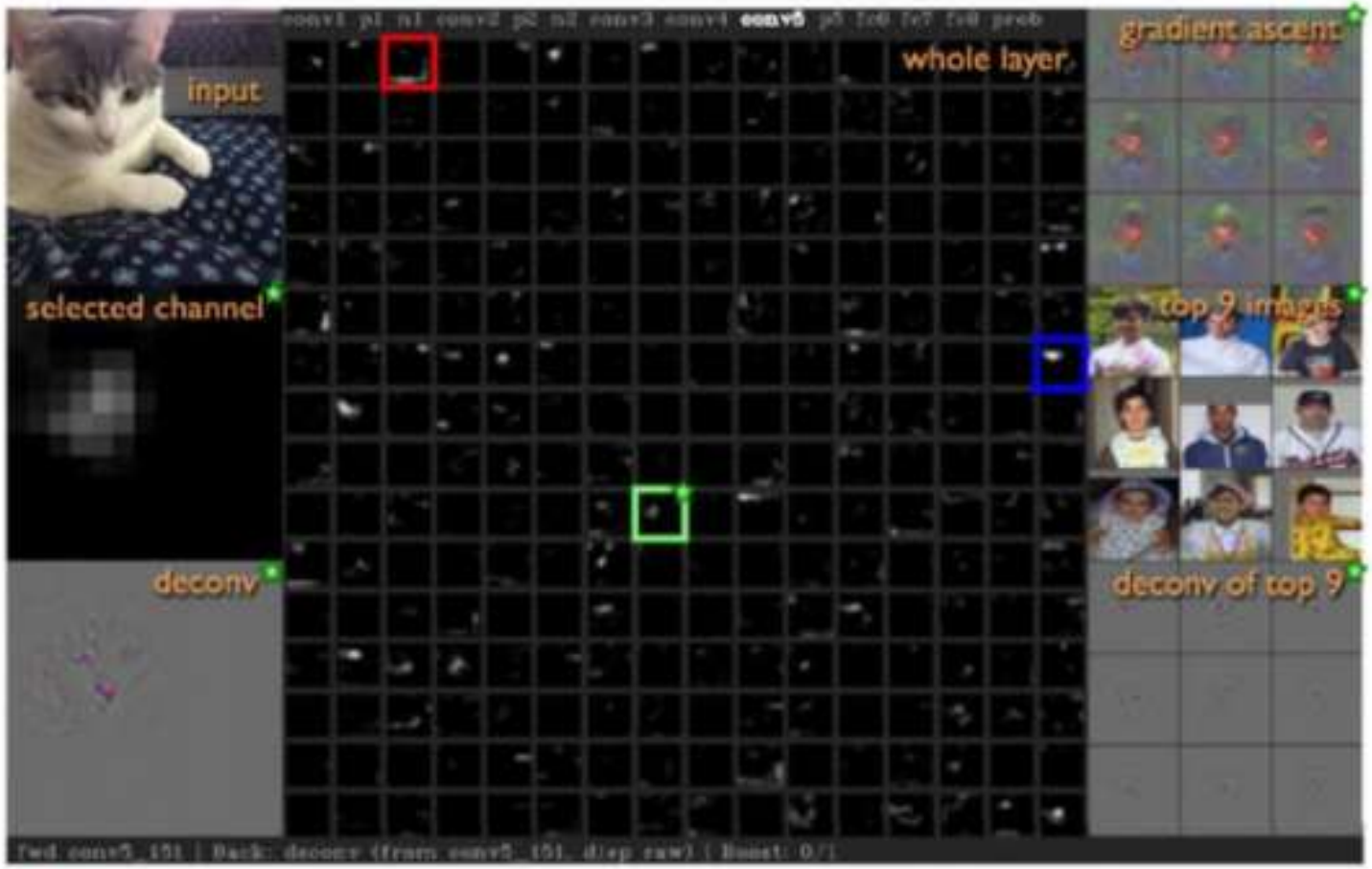


그림 7-7 가시화 도구 화면

gradient ascent이
영상 공간엣서 최적화를 이용한
역투영이다.

deconv와 deconv of top 9
영역이 디컨볼루션을 이용한 영상이다.

끝