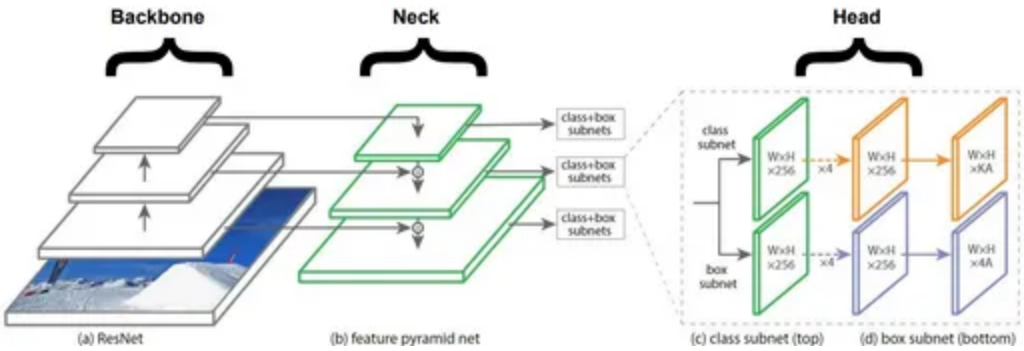
FPN-Feature Pyramid Network for Object Detection

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2117–2125). IEEE. https://doi.org/10.1109/CVPR.2017.631

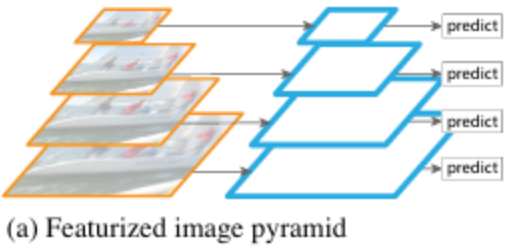
Object Detection 파이프라인



- **백본(Backbone)** : 입력 이미지로부터 주요 특징(feature)을 추출하는 부분
- **넥(Neck)** : 백본에서 추출된 피처 맵을 정제하고 재가공하여 헤드에 전달하는 중간 다리 역할
- **헤드(Head)** : 넥에서 전달받은 최종 피처 맵을 기반으로 실제 **예측(Prediction)을 수행하는 부분**

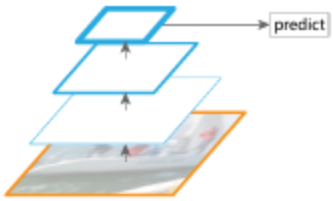
객체 탐지 접근 방식

(이미지 해석 방식 : 피처 맵은 파란색 외곽선으로 표시되며, 더 두꺼운 외곽선은 의미론적으로 더 강한 피처를 의미한다. 즉, 외곽선이 두꺼울수록 '객체를 잘 인식하는 똑똑한 피처'라는 의미이다.)



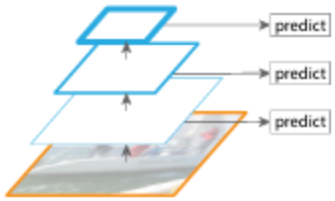
(a) Featurized image pyramid

다양한 크기의 원본 이미지들을 사용하여 피처 피라미드를 구축하는 방식으로, 원본 이미지를 여러 크기로 리사이즈한 뒤, 각각의 이미지를 모두 ConvNet에 통과시켜 피처를 추출한다. 성능은 좋지만, 네트워크를 여러 번 실행해야 하므로 매우 느리다.



(b) Single feature map

Faster R-CNN과 같은 모델들의 기본 방식이다. 속도를 위해 이미지를 한 번만 네트워크에 통과시키고, 가장 마지막에 나오는 최종 피처 맵 하나만을 예측에 사용한다. 빠르지만, 다양한 크기의 객체, 특히 작은 객체를 탐지하는 데 한계가 있다.



(c) Pyramidal feature hierarchy

SSD와 같은 모델의 접근법이다. ConvNet을 한 번 통과하며 나오는 중간 단계의 여러 피처 맵들(C_2, C_3 등)을 모두 예측에 활용한다. 빠르지만, 저수준(high-resolution) 피처 맵들은 의미 정보가 부족하여 성능이 떨어지는 문제가 있다.

(d) Feature Pyramid Network

단일 이미지만 입력으로 받아 (b),(c)처럼 빠르면서도, Top-down 방식과 Lateral connection을 통해 모든 레벨의 피처가 강한 의미 정보를 갖도록 만들어 (a) 방식의 장점인 높은 정확도를 달성한다.

즉, FPN은 ConvNet의 피처 계층을 재활용하되, 고수준의 강한 의미 정보와 저수준의 정밀한 위치 정보를 결합하여 추가 계산 비용을 최소화하면서도 모든 스케일에서 강력한 피처 피라미드를 만드는 것을 목표로 한다.

핵심 아이디어

FPN의 핵심 아이디어는

'의미론적으로 풍부하지만 해상도가 낮은 상위 피처와 의미론적으로는 빈약하지만 해상도가 높은 하위 피처를 융합하는 것' 이다.

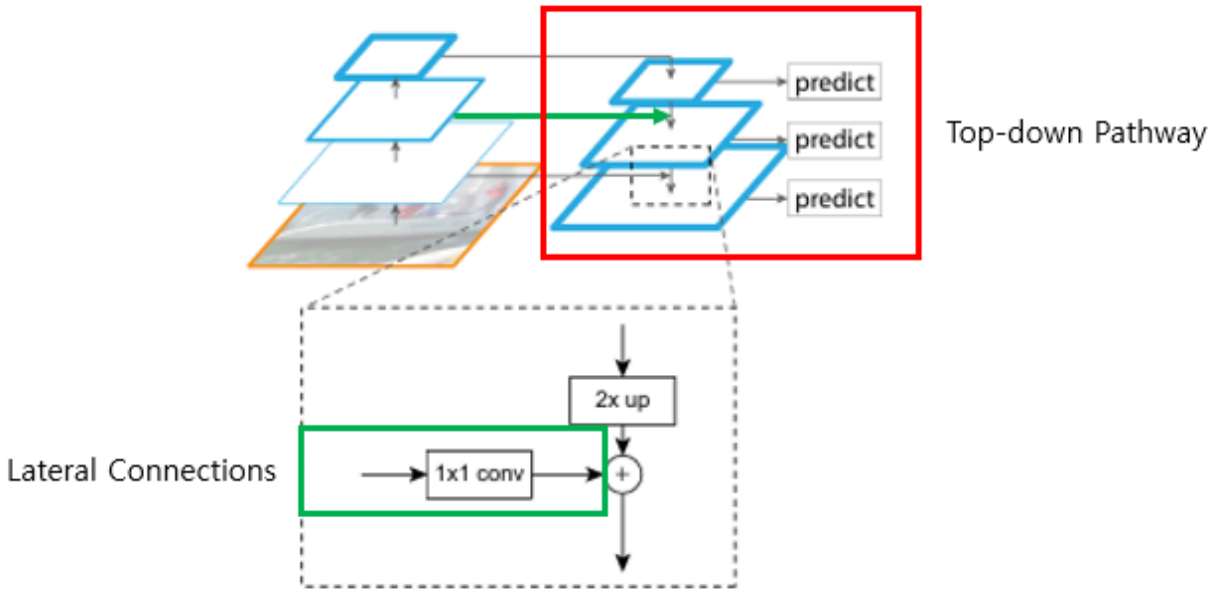
ConvNet의 정방향(bottom-up) 경로를 거치면서 피처 맵은 점차 공간 해상도가 낮아지고(예: 1/4, 1/8, 1/16, 1/32) 채널의 깊이는 깊어진다. 이 과정에서 초기 레이어의 피처 맵(예: C2)은 객체의 윤곽선이나 모서리 같은 저수준 정보를 정밀하게 담고 있지만 '이것이 고양이인지 개인지'와 같은 고수준 의미 정보는 부족하다.

반면, 마지막 레이어의 피처 맵(예: C5)은 '고양이'라는 강한 의미 정보를 담고 있지만, 여러 번의 다운 샘플링으로 인해 객체의 정확한 위치 정보는 희미해진다.

FPN은 이 두 장점을 모두 취하기 위해 두 가지 핵심 장치를 도입한다.

1. Top-down Pathway : 가장 의미 정보가 풍부한 최상위 피처 맵(C5)부터 시작하여, 점차 해상도를 높여가며(upsampling) 하위 레이어로 정보를 전달한다. 이렇게 하면 고수준의 의미 정보가 저해상도에서 고해상도 맵으로 전파된다.

2. Lateral Connections : Top-down 경로에서 업샘플링된 피처 맵을 동일한 공간 크기를 갖는 bottom-up 경로의 피처 맵과 결합한다. 이 '옆길(lateral)'을 통해 하위 피처 맵이 가진 정밀한 위치 정보가 상위 피처의 의미 정보와 합쳐진다.



핵심 수식

FPN을 Fast R-CNN과 같은 RoI(Region-of-Interest) 기반 탐지기에 적용할 때, 다양한 크기의 RoI를 어떤 피라미드 레벨에 할당할지 결정하는 규칙이 필요하다.

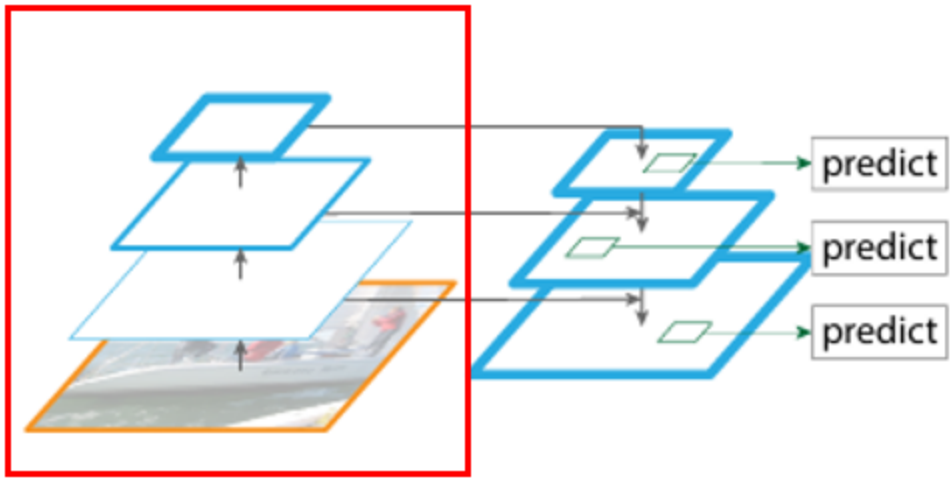
$$k = \lfloor k_0 + log_2(\sqrt{wh}/224 \rfloor$$

- k : RoI가 할당될 피라미드 레벨의 인덱스
- w, k : 입력 이미지 스케일에서의 RoI의 너비와 높이
- 224 : ImageNet 사전 학습에 사용된 표준 이미지 크기
- k_0 : 너비와 높이가 `224x224` 인 RoI를 할당할 기준 레벨

이 수식의 직관적인 의미는, RoI의 크기가 기준(`224x224`)보다 작아지면 더 미세한 해상도의 레벨(k<4)에, 더 커지면 더 거친 해상도의 레벨(k>4)에 할당하는 것이다. 예를 들어, RoI 크기가 절반인 `112x112` 가 되면 $log_2(0.5)$는 -1이므로 `k=3` 이 되어 P_3레벨에 할당된다.

Algorithm/Pipeline(FPN)

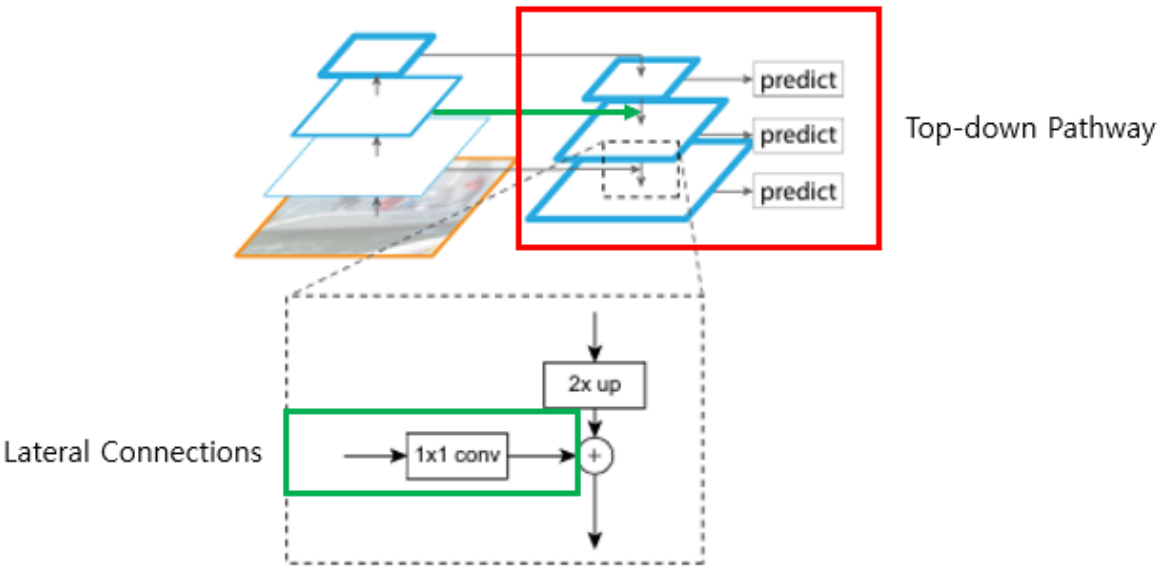
1. **Bottom-up Pathway(정방향 경로)**



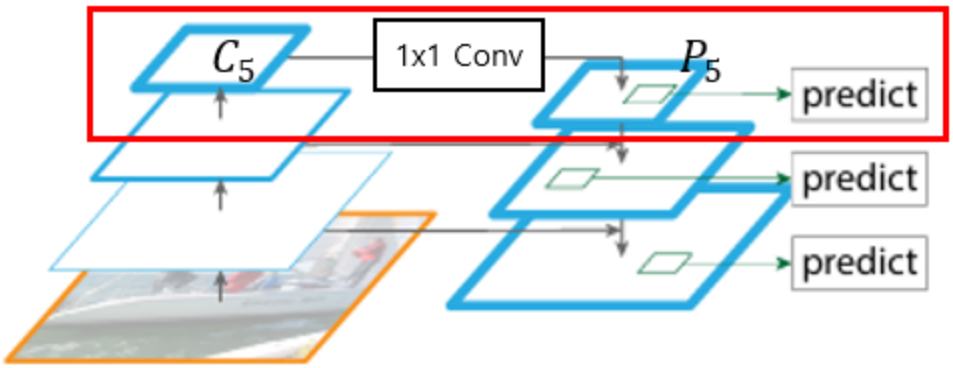
Bottom-up Pathway

- 입력 이미지를 백본 ConvNet(예: ResNet)에 통과시켜 피처 맵 계층을 계산한다.
- 각 스테이지(Stage)의 마지막 레이어에서 출력된 피처 맵을 선택한다. ResNet을 기준으로, 각 스테이지의 Stride가 {4, 8, 16, 32}인 $\{C_2, C_3, C_4, C_5\}$를 얻는다. 이들은 각각 다른 공간 해상도와 의미 수준을 가진다.

2. **Top-down Pathway & Lateral Connections(하향식 경로 및 측면 연결)

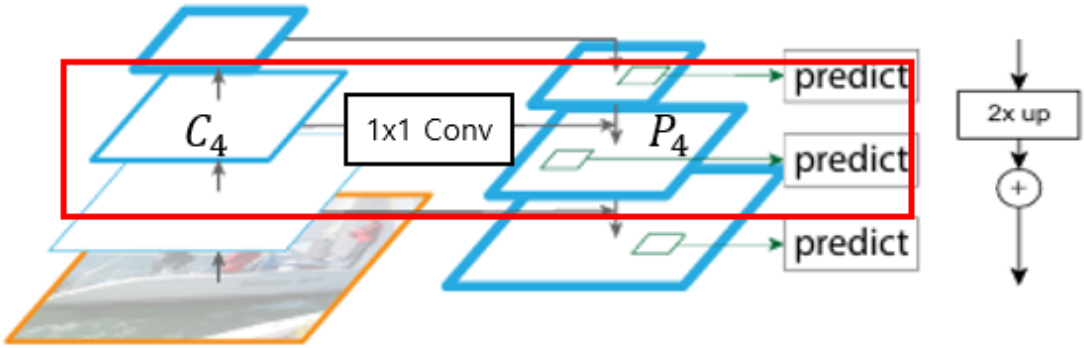


- 가장 상위 레벨인 C_5에 `1x1` 컨볼루션을 적용하여 채널 수를 `d`로 줄여 피라미드의 최상단인 P_5를 생성한다.



추가 설명

- FPN 논문에서 사용하는 백본 네트워크(ResNet-50/101)의 C_5 스테이지 출력은 2048개의 채널을 가진다. FPN 논문에서는 모든 피라미드 레벨의 최종 채널 수를 d=256으로 고정한다. 즉, 1x1 컨볼루션으로 $2048 \times H \times W$ 크기의 피처 맵을 $256 \times H \times W$ 크기로 압축하는 것이다.
- `d=256` 은 모든 피라미드 레벨에 동일한 예측 헤드를 적용하기 위한 통일된 규격이며, 성능을 크게 해치지 않으면서 모델을 가볍고 효율적으로 만들기 위한 실용적인 설계값이다.
- FPN의 가장 큰 장점 중 하나는 각 피라미드 레벨에 **동일한 예측 헤드(Head)를 공유**하여 같이 사용한다는 점이다. 헤드가 동일한 구조와 가중치를 공유하려면, 입력으로 받는 피처 맵의 채널 수가 모두 같아야한다. 또한, 논문 저자들은 2048 채널을 그대로 사용한다면, 예측 헤드의 파라미터 수가 매우 커져 모델이 무거워지고 속도가 느려지므로, FPN의 설계가 여러 선택지에 대해 강건하게(robust) 작동함을 발견했다고 언급하며, 단순함이 FPN의 핵심 철학임을 강조한다.



- C_4는 위와 동일하게 `1x1` 컨볼루션을 적용해 채널 수를 `d` 로 맞춘다.
- P_5를 2배 업 샘플링(간단한 최근접 이웃 보간법 사용)한다.
- 업샘플링된 맵과 채널이 조정된 C_4를 element-wise addition으로 결합하여 P_4의 중간 결과물을 만든다.
- 결과적으로 P_4는 P_5로부터 물려받은 풍부한 의미 정보와 C_4로부터 얻은 정밀한 위치 정보의 장점을 모두 얻게 된다.
- 이 과정을 P_3, P_2가 생성될 때까지 반복한다.

3. **Final Feature Maps(최종 피처 맵 생성)**

- 결합된 각 피처 맵(P_2 ~ P_4)에 3x3 컨볼루션을 적용하여 최종 피라미드 레벨 $\{P_2, P_3, P_4, P_5\}$를 완성한다. 이 과정은 업샘플링으로 인해 발생할 수 있는 에일리어싱(aliasing) 효과를 완화한다.
- 이렇게 생성된 각 피라미드 레벨 P_k 에 RPN(Region Proposal Network)이나 Fast R-CNN의 예측 헤드(Head)를 부착하여 객체 탐지를 수행한다.

성능

- **Faster R-CNN 성능 향상** : ResNet-50 기반의 강력한 Faster R-CNN 기준선 대비, FPN을 적용했을 때 AP는 2.3점, AP@0.5는 3.8점 향상되었다.

Faster R-CNN	proposals	feature	head	lateral?	top-down?	AP@0.5	AP	AP_s	AP_m	AP_l
(*) baseline from He *et al.* [16]†	RPN, C_4	C_4	conv5			47.3	26.3	-	-	-
(a) baseline on conv4	RPN, C_4	C_4	conv5			53.1	31.6	13.2	35.6	**47.1**
(b) baseline on conv5	RPN, C_5	C_5	*2fc*			51.7	28.0	9.6	31.9	43.1
(c) **FPN**	RPN, $\{P_k\}$	$\{P_k\}$	*2fc*	✓	✓	**56.9**	**33.9**	**17.8**	**37.7**	45.8

- **SOTA 달성** : 추가적인 기교 없이, ResNet-101 백본에 FPN을 적용한 단일 모델만으로 COCO 2016 챌린지 우승자를 포함한 기존의 모든 단일 모델을 능가하는 **36.2 AP(test-dev)**를 달성했다.

method	backbone	competition	image pyramid	test-dev					test-std				
				$AP_{@.5}$	AP	AP_s	AP_m	AP_l	$AP_{@.5}$	AP	AP_s	AP_m	AP_l
ours, Faster R-CNN on **FPN**	ResNet-101	-		**59.1**	**36.2**	**18.2**	**39.0**	48.2	**58.5**	**35.8**	**17.5**	**38.7**	47.8
*Competition-winning **single-model** results follow:*													
G-RMI†	Inception-ResNet	2016		-	34.7	-	-	-	-	-	-	-	-
AttractioNet‡ [10]	VGG16 + Wide ResNet§	2016	✓	53.4	35.7	15.6	38.0	**52.7**	52.9	35.3	14.7	37.6	**51.9**
Faster R-CNN +++ [16]	ResNet-101	2015	✓	55.7	34.9	15.6	38.7	50.9	-	-	-	-	-
Multipath [40] (on minival)	VGG-16	2015		49.6	31.5	-	-	-	-	-	-	-	-
ION‡ [2]	VGG-16	2015		53.4	31.2	12.8	32.9	45.2	52.9	30.7	11.8	32.8	44.8

- **속도** : FPN은 추가적인 레이어를 포함하지만, 더 가벼운 예측 헤드를 사용하므로, ResNet-101 기반 Faster R-CNN 기준 모델(0.32)보다 **더 빠른 추론 속도(0.172초)를 보인다.**