객체 인식 설명성 향상을 위한 FPN-Attention Layered 모델의 성능 평가

윤석준 조남익
서울대학교 뉴미디어통신공동연구소
ysj624637@ispl.snu.ac.kr

Performance Evaluation of FPN-Attention Layered Model for Improving Visual Explainability of Object Recognition

Youn, Seok Jun Cho, Nam Ik
Seoul National University Institute of New Media and Communication

요약

DNN을 사용하여 객체 인식 과정에서 객체를 잘 분류하기 위해서는 시각적 설명성이 요구된다. 시각적 설명성은 object class에 대한 예측을 pixel-wise attribution으로 표현해 예측 근거를 해석하기 위해 제안되었다, Scale-invariant한 특징을 제공하도록 설계된 pyramidal features 기반 backbone 구조는 object detection 및 classification 등에서 널리 쓰이고 있으며, 이러한 특징을 갖는 feature pyramid를 trainable attention mechanism에 적용하고자 할 때 계산량 및 메모리의 복잡도가 증가하는 문제가 있다. 본 논문에서는 일반적인 FPN에서 객체 인식 성능과 설명성을 높이기 위한 피라미드-주의집중 계층네트워크 (FPN-Attention Layered Network) 방식을 제안하고, 실험적으로 그 특성을 평가하고자 한다. 기존의 FPN만을 사용하였을 때 객체 인식 과정에서 설명성을 향상시키는 방식이 객체 인식에 미치는 정도를 정량적으로 평가하였다. 제안된 모델의 적용을 통해 낮은 computing 오버헤드 수준에서 multi-level feature를 고려한 시각적 설명성을 개선시켜, 결괴적으로 객체 인식 성능을 향상 시킬 수 있음을 실험적으로 확인할 수 있었다.

1. 서론

현대 사회에서 인공지능이 차지하는 비중은 다방면에서 꾸준히 증가하고 있으며, 심층 신경망(DNN, Deep Neural Network)의 발전은 컴퓨터 비전 분야의 연구를 가속화시키고 있다. 최근 객체 인식(Object Recognition)에서 DNN 모델을 사용하여 객체 인식의 판단 근거를 설명하기 위한 연구가 많이 진행되고 있다. 시각적 설명성(visual explanation)은 객체 인식과정에서 Feature 특성을 잘 포착하는 DNN의 특징을 잘 활용한 CAM(Class Activation Mapping)과 CAM의 구조상 문제로 인해 발생하는 성능의 감소를 gradient signal을 이용해 feature map을 만들어냄으로써 해결한 grad-CAM (gradient-weighted CAM) 등이 있다 [1][2][3]. CAM, Grad-CAM과 같은 post-hoc attention 기술들은 이미 학습이 완료된 모델들에 적용되어 판단 근거를 확인하기 위해 사용된다.

한편 학습 과정에서 네트워크로 하여금 중요한 특징들에 더 집중하고 그렇지 않은 특징에는 덜 집중하도록 능동적으로 학습하게 하는 방법인 trainable attention mechanism을 적용하여 추론 과정에서의 성능을 높이고, 별도의 post-hoc attention 없이 visual explainability를 획득하려는 연구들이 시도되고 있다. Image 내의 다양한 크기의 object를 탐지할 때, 일반적인 CNN 모델의 경우 image 자체의 크기를 조정해가며 object를 찾아냈었다. 이러한 과정은 너무 많은 시간을 소요하였고, 연산량 또한 많아져 메모리 측면에서도 아주 비효율적이었다. FPN(Feature Pyramid Network)은 pyramid feature의 layer 별 설명성 특성을 반영하여 물체의 semantic, boundary 정보를 모두 가지는 visual explanation 조합 방법이다[4]. Low level feature와 high level feature를 동시에 고려하도록 설계된 pyramid 구조를 통해 object recognition에 있어 성능의 개선을 이루어냈다.

최근에는 시각적 설명성을 활용하여 네트워크의 feature를 정교하게 처리하는데 적용되어 정확도뿐만 아니라 설명성을 향상시키는 end-to-end trainable attention network 구조가 등장하였다. Attention 개념은 VQA(Visual Question Answering)와 같이 다양한 feature selection을 요구하는 task들에서 자주 사용되고 있으며, 이는 image detection 및 classification과 같이 computer vision 내의 더 넓은 분야에서도 적용할 수가 있다[1]. 하지만, 최근 object detection 및 segmentation에서 등장하는 pyramidal features 기반 backbone 구조의 경우 attention network가 inter-feature 간 중요도 반영과 계산/메모리 복잡도 측면에서 한계를 가진다.

본 논문에서는 일반적인 FPN에서 객체 인식 성능과 설명성을 높이기 위한 새로운 피라미드 기반 주의집중 네트워크 (FPN-Attention Layered Network)) 방식을 제안하고, 실험적으로 그 특성을 평가하고자 한다. 기존의 FPN만을 사용하였을 때 설명성을 향상시키는 방식이 객체 인식에 미치는 정도를 정량적으로 평가하였다.

2. 제안하는 FPN-ABN Layered 모형

1) Feature Pyramid Network

일반적으로, FPN[4]은 작은 컴퓨팅 overhead로도 다양한 scale의 object들을 탐지할 수 있는 것으로 알려져 있다. FPN은 low-level image feature 및 high-level image feature를 동시에 고려하도록 설계되어 있으며, 크게 bottom-up pathway와 top-down pathway로 나뉘어져 있다. Bottom-up pathway는 image를 convolutional network에 통과시켜 2배씩 작아지는 feature map을 뽑아내는 과정이고, 이 output들을 top-down pathway에서 upsampling을 통해 크기를 맞춰준 후 element-wise summation을 해준다. 기본적인 FPN의 구조는 아래의 그림 1과 같다.

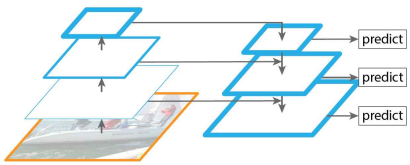


그림 1 Feature Pyramid Network의 구조[4]

2) Attention Branch Network

ABN 모델은 크게 feature extractor, attention branch, 그리고 perception branch로 나눌 수 있으며, 그 중 attention branch 부분을 FPN에 적용하였다. Attention branch는 CAM과 유사한 모델을 기반으로 하여 attention map을 얻어내는데, 획득해낸 map과 기존의 feature map을 attention mechanism을 통해 perception branch로 보낸다는 점에서 CAM과 차이가 있다. 기존의 Attention branch의 구조는 아래의 그림 2와 같다.

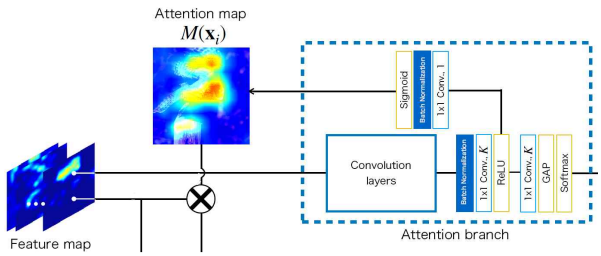


그림 2 Attention Branch의 구조[7]

3) DNN 객체 인식에서의 설명성 평가 척도

Deep Network에서 중요한 특징들에 더 주의 집중을 하는지를 객관적으로 분석할 수 있는 객체 인식에 따른 설명성을 정량적으로 평가할 수 있는 척도를 정의하고 이를 바탕으로 설명성을 평가하기로 한다. 본 논문에서는 설명성 평가를 위하여 GRAD-CAM++에서 사용된 metric인 average drop과 increase in confidence를 적용한다[8][9].

Average drop은 image의 중요한 부분에 occulsion을 적용했을 때의 성능 저하를 의미한다. 심층 신경망 모델은 explanation map을 만드는 과정에서 전체 image space에서 다양한 패턴들을 찾으려 노력하는데, 일반적으로 image 내에 occlusion된 영역이 존재하면 이 과정을 방해할 것이다. 즉, 잘 만들어진 explanation map은 image 내의 중요한 부분들을 잘 찾아내기에 모델의 confidence는 비교적 덜 줄어들 것이다. Average drop[8]은 클래스별 이미지를 시각적 설명성에 따라 masking하였을 때 task network에서 추론한 해당 클래스의 Confidence가 저하되는 정도를 백분율로 나타낸 것으로 식 (1)로 주어진다(N: 데이터셋 내의 image의 수, Y_i^c: image i가 입력으로 주어졌을 때 confidence score, O_i^c: explanation map이 입력으로 주어졌을 때 confidence score).

$$\text{Average drop} = \frac{1}{N}\sum_{i=1}^{N}\frac{\max(0, Y_i^c - O_i^c)}{Y_i^c} \times 100(\%) \quad (1)$$

한편, 경우에 따라서 신경망이 image에서 설명하고자 하는 가장 중요한 패턴이 explanation map의 아주 특이한 곳에 위치 할 수도 있다. 이 경우 increase in confidence는 중요하지 않은 부분에 occlusion을 적용했을 때의 성능 향상을 의미하며, 이렇듯 image 내의 특정 위치에 해당하는 class로 인해 모델의 confidence가 늘어나게 되면 그 외의 영역을 고려하지 않았을 때 성능은 더 좋아진다. Increase in confidence[9]는 아래의 식 (2)로 주어진다.

$$\text{Increase in confidence} = \sum_{i=1}^{N}\frac{Sign(Y_i^c < O_i^c)}{N} \times 100(\%) \quad (2)$$

4) 설명성 향상을 위한 FPN-ABN Layered 모형

설명성을 향상시키기 위하여 FPN의 각 layer에서 구한 feature map을 식(1), (2)에 정의된 두 가지 metric에 근거하여 설명성을 분석하고, 이에 근거하여, metric 값이 상대적으로 낮은 feature pyramid를 선택적으로 attention branch를 추가하여 설명성을 향상시키는 FAL(FPN-ABN Layered Model) 모델을 제안한다. 예를 들어 FPN의 feature pyramid 중 Layer1과 Layer2에 attention branch를 적용시키면 아래의 그림 3과 같이 모델을 표현할 수 있다.

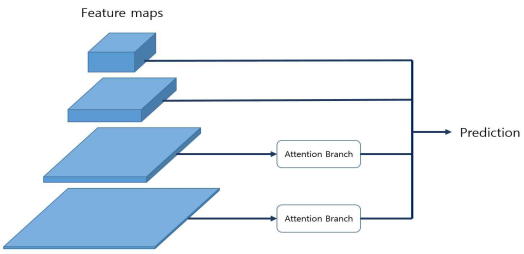


그림 3 Layer 1, 2에 Attention Network를 적용한 FAL Model 예시

3. 실험 및 결과

본 논문에서는 attention branch를 FPN의 Layer1, ..Layer4 중 어느 곳에 적용하는 것이 설명성 향상에 도움이 되는지를 평가하기 위하여 UC Merced Dataset에 대해 FPN 계층별 설명성 평가

를 정량적으로 분석하였다. 실험에 사용한 UC Merced Land Use 데이터셋은 총 21개의 class에 대해 각각 100장의 image로 구성되어 있으며, 모든 image의 크기는 256×256픽셀이다. 다른 모든 class의 image에 대해 training을 시켰으며, airplane image들을 대상으로 validation을 진행하였다.

식(1), (2)에 정의된 두 가지 metric에 근거하여 FPN 모델을 분석해 본 결과 feature pyramid 중 낮은 층, 특히 아래 두 level의 layer에서 상대적으로 % increase in confidence 값이 작게 나오는 것을 확인하였다. 이는 아래의 표 1에서 확인할 수 있다.

표 1 Feature pyramid의 각 계층에 따른 metric 값의 비교
(UC Merced Land Use Dataset class 'airplane')

	FPN Layer 1	FPN Layer 2	FPN Layer 3	FPN Layer 4
Average drop(%)	99.674	99.021	89.995	70.212
Increase in confidence	0	0	10	20

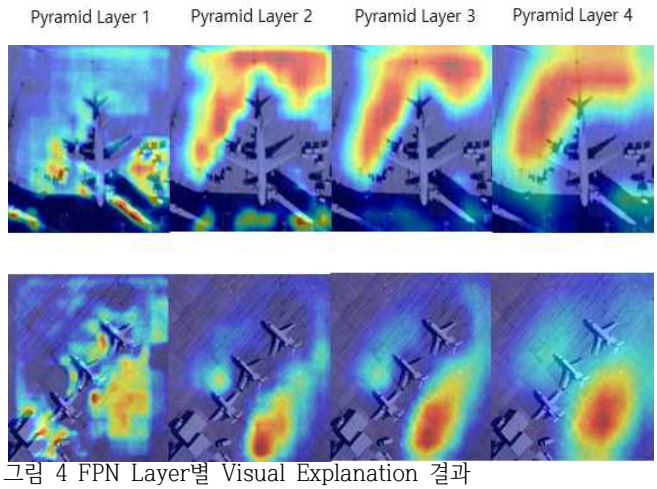


그림 4 FPN Layer별 Visual Explanation 결과

표 1과 그림 4는 FPN 모델의 visual explanation 결과 중 UC merced 데이터셋 내의 'airplane' class에 해당하는 결과를 대표로 뽑아낸 것이다. 그림 4의 heatmap의 경향을 보면 FPN 모델 단독으로는 전반적으로 heatmap이 객체를 제대로 인지하는 데에 어려움이 있는 것을 확인하였다.

한편 FPN Layer 1에서 Layer 4로 Backbone Network가 변화될 때 설명성이 향상된다고 판단할 수 있다. 이러한 Layer 별 객체에 대한 설명성 변화에 착안하여 layer의 feature map에 선택적으로 attention network를 적용해 보았을 때 이것이 성능에 어떠한 변화를 주는지 heat map이 포함된 explanation map을 통해 정성적으로도 성능의 향상 여부를 확인해 보았다.

본 실험에서는 표 1에 나타난 것과 같이 pyramid feature의 하단 부분인 Layer 1, 그리고 Layer 2에서 increase in confidence가 0으로 평가되었고, average drop의 크기도 상대적으로 훨씬 큰 것을 확인하여 이에 따라 FPN Layer 1, 2에 attention branch를 적용해 보았다. 그 결과는 그림 5와 같으며, heatmap이 객체에 더 잘 집중하는 모습을 확인할 수 있다. Layer 1과 Layer 2에서도 성능이 좋아졌지만, 특히 Layer 4에서의 visual explanation이 눈에 띄게 좋아진 것을 확인하였다.

4. 결론

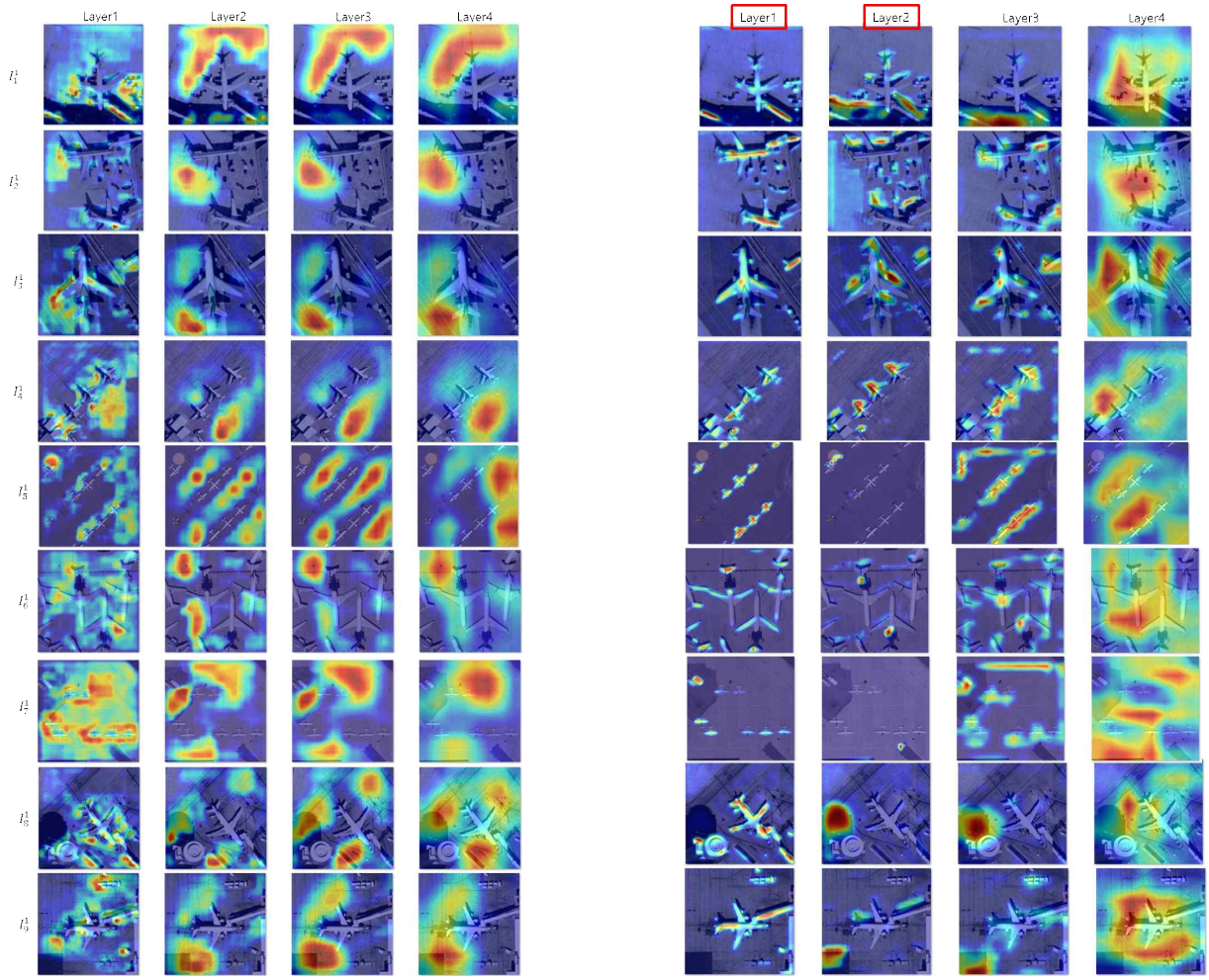
본 논문에서는 일반적인 FPN에서 객체 인식 성능과 설명성을 높이기 위한 피라미드-주의집중 계층네트워크 (FPN -Attention Layered Network) 방식을 제안하고, 실험적으로 그 특성을 평가하였다. 기존의 FPN만을 사용하였을 때 객체 인식 과정에서 설명성을 향상시키는 방식이 객체 인식에 미치는 정도를 average drop과 increase in confidence metric을 사용하여 평가하였다. 실험 결과 제안된 FAL 모델이 multi-level feature 측면에서 시각적 설명성을 개선시켜, 결과적으로 객체 인식 성능을 향상시킬 수 있음을 확인할 수 있었다.

감사의 글

이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(2021R1A2C2007220).
이 논문은 2022년도 BK21 FOUR 정보기술 미래인재 교육연구단에 의하여 지원되었음.

참고문헌

[1] Hendricks, Lisa Anne, et al. "Generating visual explanations."*European conference on computer vision*. Springer, Cham, 2016.
[2] Zhou, Bolei, et al. "Learning deep features for discriminative localization."*Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
[3] Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization."*Proceedings of the IEEE international conference on computer vision*. 2017.
[4] Lin, Tsung-Yi, et al. "Feature pyramid networks for object detection."*Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.
[5] Patro, Badri, and Vinay Namboodiri. "Explanation vs attention: A two-player game to obtain attention for vqa."*Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. No. 07. 2020.
[6] Woo, Sanghyun, et al. "Cbam: Convolutional block attention module."Proceedings of the European conference on computer vision (ECCV). 2018.
[7] Fukui, Hiroshi, et al. "Attention branch network: Learning of attention mechanism for visual explanation."*Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019.
[8] Chattopadhay, Aditya, et al. "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks."*2018 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2018.
[9] Ramaswamy, Harish Guruprasad. "Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization."*Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2020.



(a) (b)

그림 5 FPN과 제안한 FAL 방식의 Visual Explanation 결과 비교