시각적 관계 인식 성능 개선을 위한 객체 검출 기술 연구

최현지°[1], 허유정[2], 장병탁[2]
[1]서울대학교 생명과학부
[2]서울대학교 컴퓨터공학부
app9955@snu.ac.kr, yjheo@bi.snu.ac.kr, btzhang@bi.snu.ac.kr

A Study on Object Detection Technology for an Improved Visual Relationship Detection

Hyunji Choi °[1], Yu-Jung Heo[2], Byung-Tak Zhang[2]
[1] Department of Biological Sciences, Seoul National University
[2] Department of Computer Science and Engineering, Seoul National University

요 약

이미지를 올바르게 이해하기 위해서는 이미지 내 존재하는 객체 뿐 아니라 객체간의 관계에 대한 이해가 필수적이다. 최근 이미지 인식을 위한 딥러닝 기반의 다양한 시각 모델이 제시되었으나, 대부분의 경우 고립된(isolated) 객체에 집중하여 객체의 레이블과 속성을 학습한다. 시각적 관계 인식은 일반적으로 이미지 내의 고립된 객체를 검출한 후 검출된 객체 사이의 관계를 예측하는 순차적 기법으로, 시각적 관계 인식의 성능은 이미지 내 객체 검출에 크게 의존적이다. 따라서, 우리는 이미지 내 객체 검출 기술 연구를 통해 시각적 관계 인식의 성능을 개선하고자 한다. 이 때, 검출된 객체 사이의 관계 예측을 위해서 기존 연구에서 제시된 신경 모티프(Neural Motifs) 모델[1]을 활용한다. 본 논문에서는 Faster R-CNN과 YOLOv3의 객체 검출기를 기반으로 한 신경 모티프 모델의 시각적 관계 인식 성능을 비교하고, 각 객체 검출기의 특성을 실험적으로 확인한다. 여기에 더해, 두 객체 검출기가 상호보완적인 성질을 가지므로, 두 객체 검출기의 결과를 함께 활용한 경우(hybrid) 보다 정확한 객체 검출을 통해 시각적 관계 인식의 성능이 4.1% (recall@100) 개선됨을 보였다.

1. 서 론

이미지 인식, 시각적 장면 이해 분야는 객체 검출 등 이미지 내 독립적인 객체를 판별하는 문제에 초점을 맞추고 발전해왔다[2, 3]. 특히 딥러닝 기술의 발달로 Faster R-CNN[4], YOLOv3[5] 등 빠르고 정확한 객체 검출기가 등장하였다. 그러나 이미지를 완전하게 이해하기 위해서는 독립 객체 뿐 아니라 객체 간 관계를 파악하는 것 또한 매우 중요하다. 시각적 관계 인식에서 출발하여 시각적 질문 답변, 장면 이해와 같은 다양한 고차원적 시각적 인지 문제를 해결할 수 있기 때문이다.

시각적 관계 인식을 위해 초기에는 관계 삼중항(triplet)을 레이블로 정의하여 학습하는 분류기를 구성하였으나, 각각의 관계가 고르게 분포하지 않는 데이터 불균형 문제로 인해 인식 성능의 한계를 보였다[6]. 이후 의미론적(semantic) 접근이나 맥락(context)를 고려하여, 객체와 관계를 각각 분류하는 기법들이 제시되었다. 신경 모티프(Neural Motifs) 모델은 Faster R-CNN으로 객체를 검출한 뒤 주어진 주체와 객체 쌍에 대한 관계의 술부를 예측하는 모델로, 이 때 주어진 주체와 객체 쌍에 대해 특정 술부가 나올 확률이 높다는 맥락 정보를 활용한다[1].

시각적 관계 인식은 일반적으로 이미지 내의 개별 객체 검출 후 검출된 객체 사이의 관계를 예측하는 순차적 기법이기 때문에, 시각적 관계 인식의 성능은 이미지 내 객체 검출에 크게 의존적이다. 따라서, 본 연구에서는 이미지 내 객체 검출 기술 연구를 통해 시각적 관계 인식의 성능을 개선하고자 한다. 보다 구체적으로, 신경 모티프 모델의 성능을 개선하기 위해 Faster R-CNN과 YOLOv3 객체 검출기를 적용하여 시각적 관계 인식의 성능을 비교하고, 최종적으로 두 객체 검출기의 결과를 함께 활용한다(hybrid). 결과적으로, 두 객체 검출기가 상호보완적인 관계를 가지므로, 두 객체 검출기의 결과를 함께 활용한 경우(hybrid) 보다 정확한 객체 검출을 통해 시각적 관계 인식의 성능이 4.1% (recall@100) 개선됨을 보였다.

사사 이 논문은 2019년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원(2015-0-00310-SW.StarLab, 2017-0-01772-VTT, 2018-0-00622-RMI, 2019-0-01367-BabyMind), 한국산업기술평가관리원(10060086-RISF, P0006720-GENKO)의 지원을 받았음.

2. 관련 연구

2.1 객체 검출

객체 검출은 이미지에서 객체가 존재하는 예상 경계 상자를 추출(localization)하고 해당 객체를 분류(classification)하는 것을 의미한다. 현재 객체 검출에 가장 많이 사용되는 모델 중 하나인 Faster R-CNN은 R-CNN에서 출발하였다[4]. R-CNN은 지역 제안(region proposal) 방식을 사용하여 카테고리와 무관하게 선택적 탐색(selective search)을 통해 후보 지역을 먼저 찾는다[2]. 이후 합성곱 신경망(CNN, Convolutional Neural Network)을 사용하여 고정 길이의 특징 벡터를 각 지역에 대해 추출한 뒤 분류하고 경계 상자의 좌표를 얻는다. 이 방법은 높은 정확도를 보이기는 했으나 비교적 느린 단점이 있었다. 이를 개선한 것이 Fast R-CNN이다. Fast R-CNN에서는 이미지 전체에 대해 하나의 합성곱 신경망을 적용하여 특징(feature map)을 추출하고 관심 지역 (RoI, Region of Interest) pooling 층을 이용해 최종 경계 상자와 클래스를 결정한다[7]. 하지만 여전히 선택적 탐색 단계에서 많은 시간이 소요되었다. 따라서 Faster R-CNN에서는 선택적 탐색 대신 지역 제안망을 이용하여 예상 경계 상자를 추출한다[4].

YOLO 계열의 모델은 지역 제안 후 각각에 대해 분류기를 사용하는 두 단계 R-CNN 계열 방식과 달리 한 번의 합성곱 신경망으로 경계 상자 및 라벨을 예측하여 비교적 빠른 속도를 보인다. YOLO는 7*7 격자 셀로 이미지를 나눈 뒤 각 셀에서 경계 상자와 클래스에 대한 점수를 추출하고 비-최댓값 억제(NMS, Non-Maximal Suppresion)를 통해 가장 높은 점수의 상자를 선택한다[3]. 이는 한 개의 셀에서 하나의 클래스만 결정하기 때문에 배경을 잘못 인

식하는 오류는 적으나 가까이 붙어 있는 물체를 잘 인식하지 못하고, 상자 위치가 다소 부정확한 문제가 있었다. 이것을 YOLOv2에서는 배치 정규화, 앵커(anchor) 상자 사용, 다중 비율 학습 등의 사용으로 성능과 속도 모두 개선하였다[8]. 마지막으로 YOLOv3은 특징 추출 신경망 모델의 개선, 다중 비율 상자 추출 등으로 성능 향상을 이끌어냈다[5].

2.2 시각적 관계 인식

시각적 관계 인식은 이미지에서 객체의 위치와 클래스를 결정한 후 객체 간 상호작용을 정의하는 것을 목표로 하며 시각적 질문 답변과 같은 더 고차원적 시각 작업의 기반이 된다[9]. 관계는 {주체-술부-객체}의 삼중항으로 표현 가능하며, 객체 종류가 N개, 술부 종류가 K개일 때 가능한 관계의 수는 O(N*N*k)이기 때문에 개별 물체 인식에 비해 가능한 가짓 수가 훨씬 많다. 따라서 관계 자체, 또는 시각적 어구(visual phrase)를 학습시킨 초기의 시도들은 큰 데이터 셋에서는 효과가 나타나지 않았다[6]. 이후에는 사물과 술부를 분리하여 학습시킨 후 단어 의미를 고려하여 통합된 결과를 도출하거나[10], 사물의 위치, 의미론적 연관성, 외형 등의 신호를 모두 고려한 모델이 발달하였다[9]. 최근에는 관계를 그래프로 표현한 그래프 기반 시각적 관계 인식 모델들이 등장하였다. 그래프 R-CNN은 객체를 검출한 뒤 모든 가능한 관계를 그래프로 표현하고 관계 제안망 (RePN, Relation Proposal Network)을 통해 관계성 지수가 낮은 선(edge)을 지워 나간다[11]. 이후 맥락 정보를 포착하는 주의 그래프 합성곱망 (aGCN, attentional Graph Convolutional Network)을 사용하여 객체와 객체 사이의 관계를 학습한다.

3. 신경 모티프(Neural Motifs) 모델

시각적 관계 인식 모델 중 하나인 신경 모티프 모델은 이미지를 객체 간 관계의 구조적 반복, 또는 모티프의 결합으로 본다[1]. 즉, 주체와 객체 쌍이 정해지면 그 관계를 설명하는 술부는 극히 일부로 한정되며, 이러한 패턴은 조금 더 큰 부분 그래프에서도 나타난다는 가정 하에 쌓인 모티프 네트워크(Stack Motif Network)를 사용한다. 그림 1에서 볼 수 있듯이 전체 검출 과정은 경계 상자, 라벨, 그리고 관계 예측 단계로 분리되고, 각 단계 사이의 양방향 LSTM으로 맥락을 공유한다. Faster R-CNN 객체 검출기가 경계 상자를 예측하면 이 객체 맥락(object context)과 이미 결정된 라벨들을 이용하여 다음 라벨을 결정하고, 다시 이 맥락을 이용하여 관계 맥락(edge context)를 형성한다. 이것이 최종적으로 전체 맥락(global context)를 형성하면서 관계의 라벨을 결정하는데 사용된다. 해당 모델의 보다 자세한 작동 기제는 [1]에서 확인할 수 있다.

4. 실험 설계

4.1 데이터

시각 게놈(Visual Genome) 데이터는 지각적 작업 위주의 기존 데이터를 보완하기 위하여 인지적 작업에 적합한 물체간 관계, 질문-대답 등의 자료를 포함한다[12]. 108,000개가 넘는 이미지와 각 이미지에 평균 35개의 물체, 26개의 특징과 21개의 물체간 관계를 규정해 놓고 있어 컴퓨터 비전 작업을 위한 모델 학습에 적합하다. 관계에는 동작, 위치 정보, 설명 동사, 전치사, 비교 어구 등이 포함되어 있다.

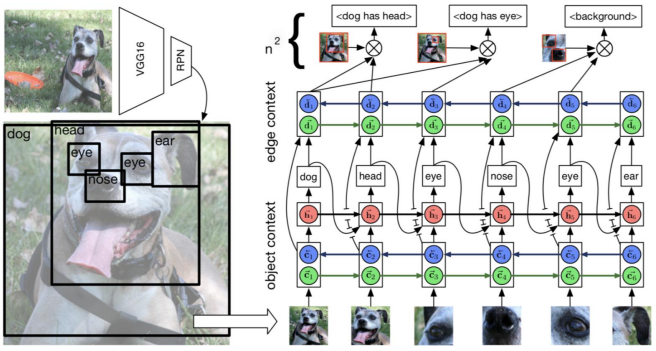

그림 1: 신경 모티프 모델의 구조 [1]

4.2 객체 분류기

YOLOv3 객체 검출기[13]는 시각 게놈 학습 데이터 셋을 이용해 batch 18, learning rate 0.054로 40 epoch 동안 학습시켰고, 시험 데이터 셋으로 mAP 15.9에 도달함을 확인하였다. 해당 모델을 활용하여 26,646개 이미지에서 신뢰값(confidence) 0.0001 이상의 box를 8, 16, 32, 64개 신뢰값 순으로 추출하였다. R-CNN 상자의 경우 기존 신경 모티프 모델에서 사용한 학습된 Faster R-CNN 모델을 이용하였고, 마찬가지로 신뢰값 0.0001 이상의 box를 8, 16, 32, 64개 신뢰값 순으로 추출하여 사용하였다.

두 가지 객체 검출기의 예상 경계 상자를 추출한 결과, R-CNN은 서로 가까이 있는 작은 물체의 경계 상자를 정확하게 잡아내며, 특히 사람의 인식에 특화되어 있음을 확인하였다. 반면 핵심 물체의 인식에 실패하는 경우가 있었는데, 이는 YOLO가 잘 잡아내었다. 따라서 두 가지 모델의 장점을 조합하여 객체를 검출하는 것이 전체 이미지 인식에 도움이 될 것이라고 판단하였다. 두 가지 결과를 함께 활용한 hybrid의 경우, 클래스 별로 비-최댓값 억제를 실시하여 IOU(Intersection over Union)가 0.3 이상인 경우 신뢰값 순으로 8, 16, 32, 64개 추출하였다.

5. 실험 결과

5.1 정량적 수치 평가

장면 검출 평가 결과는 recall@20, 50, 100 메트릭을 사용하였다. 이는 @k개의 기반 진실(ground truth) 중 몇 개를 맞게 예측했는지의 비율을 나타낸다. 기존 신경 모티프 모델을 이용하여 장면 검출(scene detection)한 결과인 recall@100 0.255를 baseline으로 설정하였다. 세 가지 방식으로 추출한 상자를 이용하여 장면 검출을 진행하였을 때의 결과는 표 1와 같았다. 상자 개수가 최대 8개인 경우 R-CNN에 비해 YOLO의 결과가 평균 1.4% 포인트 좋았다. 그러나 32개인 경우 R-CNN의 결과가 평균 1.2% 포인트 좋았다. 이는 YOLO가 더 적지만 정확한 상자를 추출했다고 볼 수 있다.

R-CNN과 YOLO의 상자를 합한 후 비-최댓값 억제를 적용한 Hybrid의 경우는 모든 경우에서 가장 좋은 결과를 보였다. 특히 신뢰값이 높은 32개의 상자를 추출하였을 때 recall@100 값이 최대치를 기록하여 신경 모티프 모델의 baseline 값보다 4.1% 포인트 향상되었다. 이는 두 가지 검출기가 가진 장점이 결합되어 작은 물체와 중앙의 큰 물체를 모두 잘 인식했기 때문이라고 볼 수 있다.

표 1: 검출 평가 결과

객체 검출 방식	평균 상자 개수	R@20	R@50	R@100
baseline[1]		0.138	0.204	0.255
R-CNN	7.9	0.135	0.137	0.137
YOLO	7.9	0.148	0.151	0.151
Hybrid	7.9	0.149	0.151	0.151
R-CNN	15.9	0.182	0.216	0.226
YOLO	15.9	0.185	0.217	0.223
Hybrid	15.9	**0.198**	**0.236**	**0.244**
R-CNN	31.8	0.163	0.237	0.279
YOLO	31.8	0.158	0.227	0.257
Hybrid	31.8	**0.172**	**0.251**	**0.296**
R-CNN	63.5	0.112	0.184	0.243
YOLO	63.5	0.121	0.188	0.241
Hybrid	63.5	0.117	**0.193**	**0.256**

[1] 장면 그래프 검출에는 신경 모티프 모델의 장면 그래프 분류기(scene classification)에 최적화된 모델 체크포인트를 사용하였다.

그림 2에 따르면, R-CNN은 가운데 버스를 인식하지 못한 것에 비해 YOLO를 사용한 경우 창문과 사인까지 정확하게 검출해 내었고, 이를 기반으로 더 정확한 관계까지 검출하였음을 확인할 수 있다. 두 검출기의 상자를 합한 경우 비-최댓값 억제로 인해 신뢰값이 비교적 낮은 가운데 버스의 세부 구성 요소는 누락되었지만, R-CNN의 작은 객체 인식과 YOLO의 핵심 객체 인식이 모두 적용되어 전체적인 장면을 더 잘 설명하는 것을 볼 수 있다.

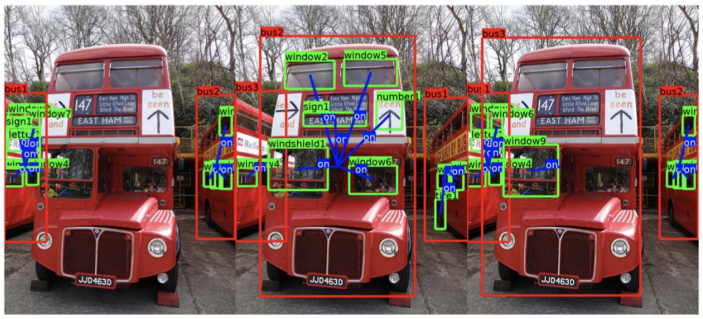

그림 2: 세 가지 방식으로 검출한 객체를 이용한 장면 그래프 검출 결과. 왼쪽부터 RCNN, YOLO, Hybrid 방식의 결과이다.

6. 결론

본 연구는 신경 모티프 모델에 Faster R-CNN 기반 검출기와 YOLOv3 기반 검출기를 적용하였을 때 YOLOv3이 비교적 적은 상자를 추출할 때 RCNN에 비해 정확도가 높고, 많은 상자를 추출할 때 정확도가 떨어진다는 것을 확인하였다. 또한 이 결과에서 착안하여 두 가지 방식의 상호보완적인 성격을 확인하고 각각에 의해 검출된 객체를 비-최댓값 억제 방식으로 결합하여 기존 모델에 비해 4.1% 성능의 향상을 이끌어내었다. 이 결과는 객체 검출기의 개선으로 이미지를 보다 잘 설명할 수 있는 장면 그래프를 이끌어낼 수 있다는 중요한 지표가 되며, 본 연구는 시각적 관계 인식에 기반한 상위 인지 문제 해결에 중요한 기점이 될 것이다.

참고 문헌

[1] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5831–5840, 2018.

[2] Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *CoRR*, abs/1311.2524, 2013.

[3] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016.

[4] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:1137–1149, 2015.

[5] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *CoRR*, abs/1804.02767, 2018.

[6] Mohammad Amin Sadeghi and Ali Farhadi. Recognition using visual phrases. *CVPR 2011*, pages 1745–1752, 2011.

[7] Ross B. Girshick. Fast r-cnn. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448, 2015.

[8] Joseph Redmon and Ali Farhadi. Yolo9000: Better, faster, stronger. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6517–6525, 2017.

[9] Kongming Liang, Yuhong Guo, Hong Chang, and Xilin Chen. Visual relationship detection with deep structural ranking. 2018.

[10] Cewu Lu, Ranjay Krishna, Michael S. Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *ECCV*, 2016.

[11] Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. Graph r-cnn for scene graph generation. In *ECCV*, 2018.

[12] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123:32–73, 2016.

[13] Glenn Jocher, guigarfr, perry0418, Ttayu, Josh Veitch-Michaelis, Gabriel Bianconi, Fatih Baltacı, Daniel Suess, and WannaSeaU. ultralytics/yolov3: Video Inference, Transfer Learning Improvements, April 2019.