심층 신경망을 이용한 영상 기반 물체 속성 및 공간 관계 탐지

이재윤*, 이기호*, 김인철*
*경기대학교 컴퓨터공학부
e-mail : jaeyoon_95@kyonggi.ac.kr,rlgh9250@kyonggi.ac.kr,kic@kyonggi.ac.kr

Detecting Visual Attributes and Spatial Relationships with Deep Neural Networks

Jae-Yun Lee*, Gi-Ho Lee*, In-Cheol Kim*
*Dept. of Computer Science, Kyonggi University

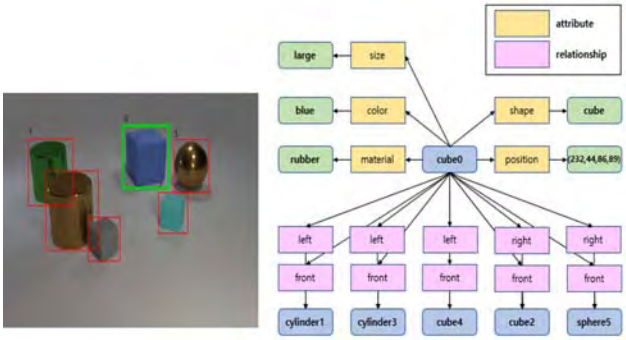
요 약

영상이나 비디오에 담긴 장면을 이해하는 것은 컴퓨터 비전의 궁극적인 목표 중 하나이다. 본 논문에서는 입력 영상으로부터 장면을 구성하는 각 물체들과 그들 간의 공간 관계, 개별 물체들의 다양한 속성들을 탐지해, 지식 그래프를 생성해주는 심층 신경망 기반의 물체 속성 및 공간 관계 탐지 모델을 제안한다. 본 논문에서는 이러한 다양한 복합 시각 인식 작업을 동시에 수행하는 탐지 모델의 구성에 대해 설명하고, 대규모 벤치마크 데이터 집합인 CLEVR을 이용한 탐지 모델의 성능 분석 실험 결과를 소개한다.

1. 서론

영상이나 비디오에 담긴 장면을 이해하는 것은 컴퓨터 비전의 궁극적인 목표 중 하나이다. 영상 기반 장면 이해(visual scene understanding)는 영상 속 물체들을 탐지해내는 작업과, 그들의 속성(attribute)을 인식해내는 작업, 그리고 탐지된 물체들 간의 관계(relationship)를 추정해내는 작업 등 다양한 시각 인식 작업들을 포함한다. 최근 몇 년간 심층 신경망 기술의 발전과 더불어, 영상 분류(image classification), 물체 탐지(object detection), 의미적 분할(semantic segmentation)과 같은 시각 인식 작업들 뿐만 아니라, 영상/비디오 캡션 생성(image/video captioning), 영상 기반 질문 응답(visual question answering) 등과 같이 시각과 언어 이해 능력이 결합된 작업들까지 다양한 지능 기술들이 발전하게 되었다. 이와 같이 영상에 담긴 장면을 이해하는 것뿐만 아니라, 이것을 자연어(natural language)나 정형화된 지식 형태(formal knowledge representation)로 표현하는 기술들도 함께 발전하게 되었다.

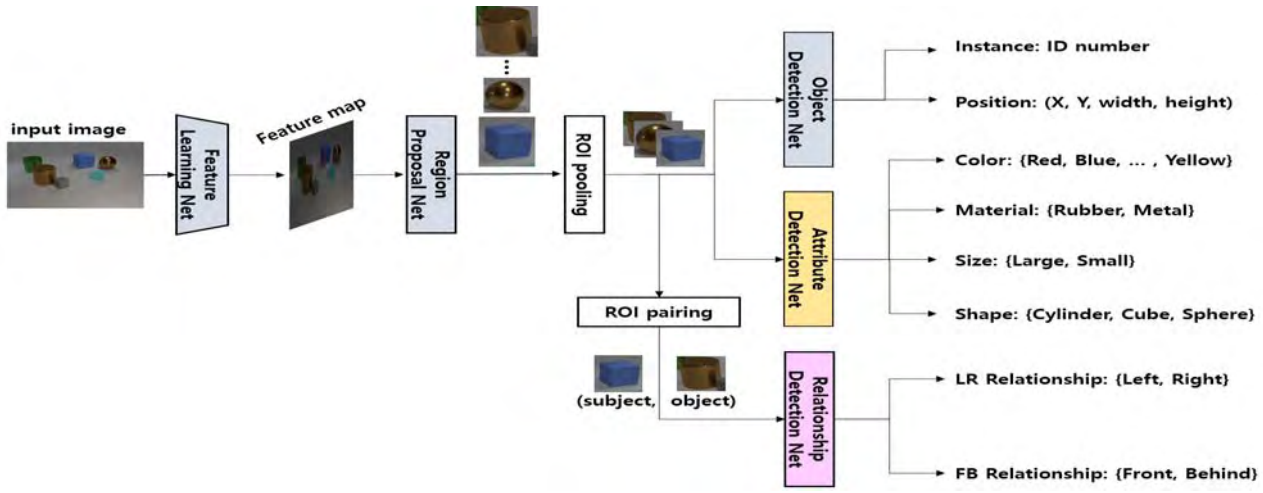
심층 신경망 모델을 이용하여 영상 장면을 이해하고 그것을 정형화된 지식 형태로 출력하고자 했던 대표적인 연구들로는 [1, 2, 3]의 연구들이 있다. 하지만 이 연구들에서는 대부분 물체들 간의 공간적, 의미적 관계들을 탐지해내는 작업들에만 집중하였고, 크기(size), 모양(shape), 색상(color), 재질(material) 등과 같은 각 물체들의 다양한 속성들을 인식해내는 작업들은 다루지 않았다. 일상생활 환경에서 인간과 상호작용하는 많은 지능형 시스템들에게는 서비스 대상 물체들의 다양한 속성들을 인식하는 능력은 필수적이다.



(그림 1) 물체 속성 및 공간 관계 탐지의 예

본 논문에서는 (그림 1)과 같이 좌측의 한 입력 영상으로부터 장면을 구성하는 각 물체들과 그들 간의 공간 관계뿐만 아니라, 개별 물체들의 다양한 속성들도 탐지해, 그림의 우측과 같이, 주어(subject) - 속성(attribute)/관계(relationship) - 목적어(object) 형태의 트리플 지식으로 구성된 지식 그래프(knowledge graph)를 생성해주는 심층 신경망 기반의 물체 속성 및 공간 관계 탐지 모델을 제안한다. (그림 1)은 직육면체(cube) 형태의 0번 물체를 중심으로 영상 내 다른 물체들과의 공간 관계들(left/right, front/behind)과, 0번 물체 자체의 색상, 모양, 크기, 재질 등의 속성들을 탐지해내고, 이들을 지식 그래프로 출력해낸 하나의 사례를 보여준다. 본 논문에서는 이러한 다양한 복합 시각 인식 작업

* 본 연구는 산업통상자원부의 재원으로 기술혁신사업의 지원을 받아 수행한 연구 과제 (No. 10060086, 개인 서비스용 로봇을 위한 지능-지식 집약 ·개방 ·진화형 로봇지능 소프트웨어 프레임워크 기술 개발)입니다



(그림 2) 물체 속성 및 공간 관계 탐지 모델

을 동시에 수행하는 탐지 모델의 구성에 대해 설명하고, 대규모 벤치마크 데이터 집합인 CLEVR v1.0[4]을 이용한 탐지 모델의 성능 분석 실험 결과를 소개한다.

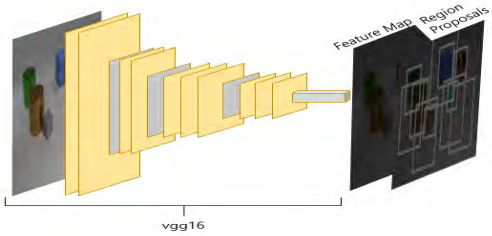
2. 물체 속성 및 공간 관계 탐지 모델

2.1 모델 개요

본 논문에서 제안하는 심층 신경망 기반의 물체 속성 및 공간 관계 탐지 모델은 (그림 2)와 같이 크게 특징 학습(feature learning), 영역 제안(region proposal), 물체 탐지(object detection), 속성 탐지(attribute detection), 공간 관계 탐지(relationship detection) 등의 서브네트워크들로 구성된다. 특징 학습 서브네트워크(Feature Learning Network, FLN)에서는 영상 안에 담긴 물체 정보를 이해하기 위하여 2차원 합성곱 신경망(Convolutional Neural Network, CNN)인 VGG-16을 이용하여 시각 특징(visual feature)을 학습하며, 이 시각 특징은 영역 제안 서브네트워크(Region Proposal Network, RPN)의 입력으로 사용된다. 영역 제안 서브네트워크(RPN)에서는 시각 특징을 토대로 영상 내에서 물체들이 존재할 후보 영역들을 판별하여 제안한다. 물체 탐지 서브네트워크(Object Detection Network, ODN)에서는 제안된 각 후보 영역 안의 물체를 식별해내고 위치 정보를 탐지해내며, 속성 탐지 서브네트워크(Attribute Detection Network, ADN)에서는 각 후보 영역 안의 물체가 가지는 색상, 재질, 크기, 모양 등의 속성들을 탐지해낸다. 마지막으로 공간 관계 탐지 서브네트워크(Relationship Detection Network, RDN)에서는 후보 영역 짝 짓기(ROI pairing)를 거쳐 만들어진 한 쌍의 물체들에 대해 기준 물체를 중심으로 다른 한 물체의 상대적 위치에 따라 좌측과 우측(left/right), 전방과 후방(front/behind) 등의 공간 관계를 판별해낸다.

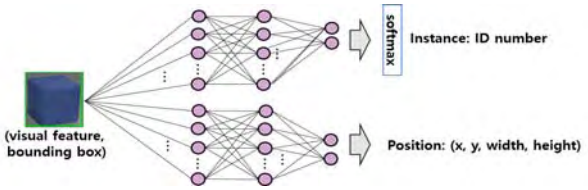
2.2 물체 탐지 네트워크

물체 속성 및 공간 관계 탐지 모델의 전반부는 물체 탐지, 속성 탐지, 공간 관계 탐지에 공통적으로 이용되는 과정으로서, (그림 3)과 같이 입력 영상으로부터 시각 특징 맵(visual feature map)을 추출하기 위한 특징 학습 네트워크(FLN), 영상 안에 물체들이 존재할 후보 영역들을 찾아내는 영역 제안 네트워크(RPN)로 구성된다. 본 모델에서는 ImageNet LSVRC, PASCAL VOC 등 대규모 시각 인식 챌린지들에서 특징 학습을 위해 널리 이용되어온 VGG16 합성 곱 신경망(Convolutional Neural Network, CNN)을 특성 학습 네트워크(FLN)로 채용한다. 그리고 영역 제안 네트워크(RPN)에서는 특성 학습 네트워크(FLN)가 추출한 시각 특징 맵 위에서 슬라이딩 윈도우를 옮기면서 각 기준 위치마다 서로 다른 모양과 크기를 가진 k개의 물체 후보 영역들을 생성한다.



(그림 3) 특징 학습과 영역 제안

이렇게 생성된 서로 다른 모양과 크기의 후보 영역(ROI)들은 ROI 풀링 층(poling layer)을 거쳐 동일한 모양과 크기로 조정된 후, 물체 탐지 네트워크(Object Detection Network, ODN)에 입력된다. 물체 탐지 네트워크(ODN)에서는 각 후보 영역에 대해 그 안에 포함되어 있는 물체를 식별하고 좀 더 정확한 바운딩 박스(bounding box)의 위치와 크기를 추정해낸다. 이를 위해 물체 탐지 네트워크(ODN)은 (그림 4)와 같이 물체 식별과 바운딩 박스 추정을 위해 각각 완전 연결 층(fully connected layer)과 Softmax 층으로 구성하였다.



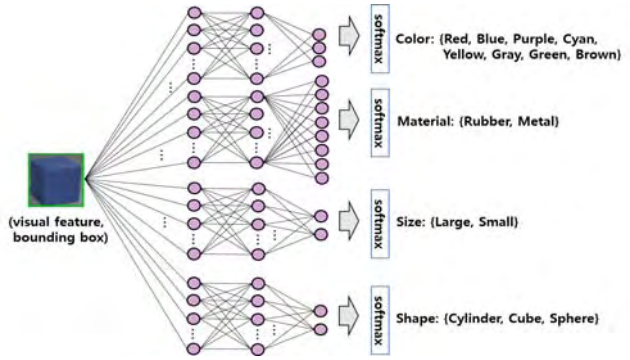
(그림 4) 물체 탐지 네트워크

물체 식별은 일종의 분류(classification) 문제로서

Softmax 층을 두는 반면, 바운딩 박스 위치와 크기인 4차원 실수 벡터 (x, y, width, height)를 추정하는 것은 회귀(regression) 문제이므로 별도로 Softmax 층을 두지 않고 완전 연결 층의 출력을 그대로 이용한다.

2.3 속성 탐지 네트워크

본 탐지 모델의 속성 탐지 네트워크(ADN)에서는 (그림 5)와 같이 ROI 풀링(pooling)을 거친 각 후보 영역에 대해 해당 영역에 포함된 물체가 가지는 다양한 속성 정보를 탐지해낸다. 속성 탐지 네트워크의 입력은 (그림 5)에 표시된 것 같이 각 후보 영역의 시각 특징(visual feature)과 해당 영역의 바운딩 박스(bounding box) 위치와 크기 정보이다.



(그림 5) 속성 탐지 네트워크

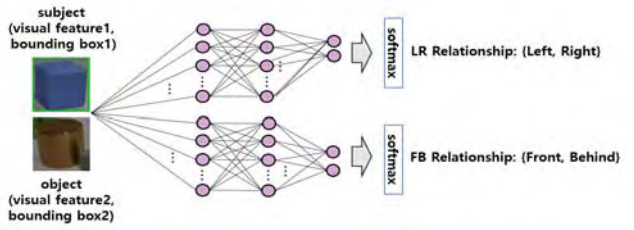
본 연구에서 물체의 모양(shape) 속성은 크게 원기둥(cylinder), 직육면체(cube), 구(sphere) 등 총 3가지 형태로 판별해내며, 물체의 크기(size) 속성은 단순히 상대적인 크기가 큰(large) 것과 작은(small) 것으로만 분류한다. 또한, 물체의 색상(color) 속성도 빨강(red), 파랑(blue), 보라(purple) 등 총 8가지 기본 색상으로만 판별하며, 물체의 제질(material) 속성도 고무(rubber)와 금속(metal)로만 분류하는 것으로 한다. 본 모델에서 물체 속성 탐지 네트워크(ADN)은 (그림 5)와 같이 각 속성별로 2개의 완전 연결 층과 하나의 Softmax 층으로 구성하였으며, 따라서 이 네트워크의 출력은 각 속성 값의 가능성을 평가한 확률 분포가 된다.

2.4 공간 관계 탐지 네트워크

한 영상 장면에 놓인 두 물체 간에는 다양한 공간적, 의미적 관계를 연결 지을 수 있다. 대표적인 공간 관계로는 좌우, 전후, 상하 등 두 물체 간의 상대적인 방향 및 위치 관계가 있다. 이 밖에도 두 물체 간의 포함 관계, 거리 관계 등 다양한 공간 관계들이 존재할 수 있다. 영상으로부터 이러한 물체들 간의 다양한 공간 관계를 자동으로 탐지해내는 기술을 전통적으로 영상 장면 이해의 핵심 기술로 받아들여져 왔다.

[2, 3]과 같은 최근 연구들에서는 개별 물체의 탐지 작업과 공간 관계 탐지 작업 사이의 상호 의존성을 고려하여, 두 작업 간의 반복적인 메시지 교환(iterative message passing)을 통해 보다 최적화된 탐지 결과를 얻어내려는 시도를 보여주었다. 하지만 두 물체 간의 공간 관계가 특별히 그 관계를 만족해야하는 개별 물체들을 특정 지을 수 없거나, 역으로 어떤 물체만이 가질 수 있는 특정 공간 관계가 없을 경우에는 이러한 반복적 메시지 교환을 통한 탐지 작업 최적화는 전체적으로 미미한 탐지 성능 개선 효과에 비해 탐지 지연 시간 증가라는 더 큰 문제를 발생시킬 수 있다.

본 연구에서 실험 대상으로 삼는 CLEVR v1.0[4] 벤치마크 데이터 집합은 작업대 위에 놓인 기본 도형 물체들에 관한 영상 데이터와 그들 간의 좌우, 전후 공간 관계들만을 포함하고 있다. 따라서 이 데이터 집합에는 특별히 특정 물체가 특정 공간 관계를 한정하는 제약도, 특정 공간 관계가 특정 물체들에만 적용되는 제약도 존재하지 않는다. 본 논문에서는 탐지 효율성을 고려하여 이러한 반복적인 탐지 작업 최적화 과정은 생략하고, 물체들 간의 공간 관계 탐지 작업을 최대한 간결하게 처리하도록 네트워크를 설계하였다.



(그림 6) 공간 관계 탐지 네트워크

본 논문에서 제안하는 공간 관계 탐지 네트워크(RDN)는 (그림 6)과 같다. 앞서 설명한 바와 같이 공간 관계 탐지 네트워크에서는 후보 영역 짝 짓기(ROI pairing)를 거쳐 만들어진 한 쌍의 물체들에 대해, 기준 물체(subject)를 중심으로 다른 한 물체(object)의 상대적 위치에 따라 좌우(left/right) 및 전후(front/behind) 공간 관계를 판별해낸다. 입력으로는 두 물체 영역 각각의 시각 특징(visual feature)과 바운딩 박스(bounding box) 정보가 주어진다. 공간 관계 탐지 네트워크(RDN)는 (그림 6)과 같이 좌우 및 전후 공간 관계별로 2개의 완전 연결 층과 하나의 Softmax 층으로 구성하였다.

3. 구현 및 실험

3.1 데이터 집합 및 구현 환경

본 연구에서 사용하는 벤치마크 데이터 집합 CLEVR v1.0[4]는 본래 영상 기반 질문 응답(viasual question answering) 연구를 위해 스텐포드 대학교에서 구축한 데이터 집합이다. 이 데이터 집합에는 영상 장면별로 물체들의 위치와 속성, 그리고 두 물체들 간의 좌우, 전후 공간 관계 정보가 기록되어 있다. CLEVR v1.0 전체 데이터 집합은 70,000개의 훈련용 영상, 15,000개의 검증용 영상, 15,000개의 테스트 영상으로 구성되어 있다. CLEVR v1.0 데이터 집합에서는 각 물체의 3차원 중심점 좌표로 위치 정보를 제공하고 있어, 본 논문의 탐지 모델을 위해 별도로 각 물체의 바운딩 박스 위치와 크기 정보를 추가한 후 실험에 이용하였다.

본 논문에서 제안한 물체 속성 및 공간 관계 탐지 모델의 성능 평가를 위해, Ubuntu 16.04 LTS에서 Python 딥러닝 라이브러리인 Tensorflow와 Keras를 이용하여 탐지 모델을 구현하였고, Geforce GTX

1080 Ti GPU가 탑재된 하드웨어 환경에서 실험을 수행하였다.

3.2 실험

첫 번째 실험에서는 본 논문에서 제안한 모델의 물체 속성 탐지 성능을 평가하기 위한 실험을 실시하였다. 이를 위해 정답 물체(ground truth) 영역과 탐지 물체 영역 간의 겹침 정도 임계값인 *tIoU* (threshold Intersection of Union)를 변경하면서, 물체들의 각 속성별 평균 정확도(average precision)를 측정하였다.

<표 1> tIoU 따른 속성 탐지 정확도

attribute \ tIoU	$\theta > 0.5$	$\theta > 0.6$	$\theta > 0.7$	$\theta > 0.8$	$\theta > 0.9$
Shape	0.36	0.35	0.35	0.36	0.36
Size	0.99	0.99	0.99	0.99	0.99
Color	0.12	0.12	0.12	0.12	0.12
Material	0.99	0.99	0.99	0.99	0.99

$\theta = tIoU(threshold\ Intersection\ of\ Union)$

<표 1>의 실험 결과를 살펴보면, 본 실험에서는 크기(size)와 재질(material) 속성은 모두 99 %의 정확도로 거의 완벽하게 탐지 성능을 보인 반면, 모양(shape)과 색상(color)은 상대적으로 낮은 36%, 12%의 정확도를 보여주었다. 또한, 본 실험에서는 기대하였던 것과는 달리 tIoU의 변화에 속성 탐지 성능 변화는 거의 나타나지 않았다. 각 물체의 모양과 색상 탐지 성능이 특별히 낮은 원인을 분석해보면, 본 실험에서 사용한 CLEVR 데이터 집합은 (그림 7)과 같이 물체들이 서로 가리거나 겹쳐져 있는 영상 장면들이 대부분인 것을 알 수 있었다. 따라서 이러한 경우에는 각 물체의 정확한 모양을 탐지해내기 어려울 뿐만 아니라, 하나의 바운딩 박스 안에 여러 물체들의 색상이 혼재하기 때문에 원천적으로 색상 역시 정확히 한 가지로 판별해내기는 어렵다. 이 실험을 통해 물체들 간에 겹침 정도가 심하지 않은 영상 장면들에 대해서는 본 모델이 물체 속성 탐지에 높은 성능을 보여줄 수 있다는 것을 확인하였다.

(그림 7) 물체들이 서로 가리거나 겹쳐지는 영상 장면들

두 번째 실험에서는 본 논문에서 제안한 모델의 공간 관계 탐지 성능을 분석해보기 위한 실험을 수행하였다. 이를 위해 *tIoU*(θ)가 0.6일 때, 좌우, 전후 공간 관계별 평균 정확도(average precision)를 측정하였다.

<표 2> 공간 관계 탐지 정확도

relationship \ precision	detection
Left/Right	0.50
Front/Behind	0.48

$\theta > 0.6$

<표 2>의 실험 결과를 살펴보면, 좌우(left/right) 관계와 전후(front/behind) 관계의 탐지 정확도가 각각 50%, 48%로 비교적 높게 나타남을 알 수 있다. 이와 같은 실험 결과를 통해, CLEVR 실험 데이터 집합과 같이 물체들 간의 겹침 정도가 심한 경우에도 본 논문에서 제안한 모델은 비교적 높은 공간 관계 탐지 성능을 보여줄 수 있다는 것을 알 수 있다.

세 번째 실험에서는 본 모델의 각 물체별 탐지 성능을 분석해보았다. 이를 위해 Recall@k를 변경하면서, 물체 탐지 정확도(mean average precision, mAP)를 측정하였다. Recall@k는 상위 k개의 바운딩 박스의 재현율(recall)을 의미한다.

<표 3> Recall@k값에 따른 물체 탐지 정확도

data \ Recall	Recall@50	Recall@75	Recall@100
Train data	0.38	0.42	0.44
Validation data	0.38	0.42	0.45

<표 3>의 실험 결과를 살펴보면, k가 증가함에 따라 물체 탐지 정확도도 함께 향상되는 양상을 보이며, Recall@100일 때 훈련 데이터 집합과 검증 데이터 집합의 물체 탐지 정확도는 각각 44%, 45%를 나타냈다. 앞서 설명한대로 물체들 간의 겹침 정도가 심한 CLEVR 실험 데이터 집합의 특성을 감안할 때, 이와 같은 본 모델의 물체 탐지 성능은 비교적 높게 평가할 수 있다

4. 결론

본 논문에서는 입력 영상으로부터 장면을 구성하는 각 물체들과 그들 간의 공간 관계, 개별 물체들의 다양한 속성들을 탐지해, 지식 그래프를 생성해주는 심층 신경망 기반의 물체 속성 및 공간 관계 탐지 모델을 제안하였다. 본 논문에서는 이러한 다양한 복합 시각 인식 작업을 동시에 수행하는 탐지 모델의 내부 구성 네트워크들에 대해 설명하였고, 대규모 벤치마크 데이터 집합인 CLEVR을 이용한 실험을 통해 본 논문에서 제안한 탐지 모델의 성능을 분석하였다. 좌우, 전후 공간 관계이외에 더 다양한 물체들 간의 공간 관계와 의미 관계들을 탐지할 수 있도록 현재의 시스템을 확장하는 연구를 향후 진행할 예정이다.

참고문헌

[1] D. Bo, Y. Zhang, and D. Lin. "Detecting Visual Relationships with Deep Relational Networks," *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[2] D. Xu, Y. Zhu, C. B. Choy, L. Fei-Fei, "Scene Graph Generation by Iterative Message Passing," *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[3] Li, Y., Ouyang, W., Zhou, B., Wang, K., & Wang, X. "Scene Graph Generation from Objects, Phrases and Region Captions," *Proc. of IEEE International Conference on Computer Vision (ICCV)*, 2017.

[4] J. Johnson, B. Hariharan, L. Maaten, L. Fei-Fei, C. L. Zitnick, R. Girshick, "CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning," *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1988-1997, 2017.