KIPS Tr. Software and Data Eng.
Vol.7, No.2 pp.69~76 pISSN: 2287-5905

https://doi.org/10.3745/KTSDE.2018.7.2.69

ORMN: A Deep Neural Network Model for Referring Expression Comprehension

Donghyeop Shin† · Incheol Kim††

ABSTRACT

Referring expressions are natural language constructions used to identify particular objects within a scene. In this paper, we propose a new deep neural network model for referring expression comprehension. The proposed model finds out the region of the referred object in the given image by making use of the rich information about the referred object itself, the context object, and the relationship with the context object mentioned in the referring expression. In the proposed model, the object matching score and the relationship matching score are combined to compute the fitness score of each candidate region according to the structure of the referring expression sentence. Therefore, the proposed model consists of four different sub-networks: Language Representation Network(LRN), Object Matching Network (OMN), Relationship Matching Network(RMN), and Weighted Composition Network(WCN). We demonstrate that our model achieves state-of-the-art results for comprehension on three referring expression datasets.

Keywords : Referring Expression Comprehension, Deep Learning, Contextual Information, Weighted Composition

ORMN: 참조 표현 이해를 위한 심층 신경망 모델

신 동 협† · 김 인 철††

요 약

참조 표현이란 장면 영상 내의 특정 물체를 가리키는 자연어 문장들을 의미한다. 본 논문에서는 참조 표현 이해를 위한 새로운 심층 신경망 모델을 제안한다. 본 논문에서 제안하는 모델은 장면 영상 내 대상 물체의 영역을 찾아내기 위해, 참조 표현에서 언급하는 대상 물체뿐만 아니라 보조 물체, 그리고 대상 물체와 보조 물체 사이의 관계까지 풍부한 정보를 활용한다. 또한 제안 모델에서는 영상 내 각 후보 영역의 적합도 계산을 위해 물체 적합도와 관계 적합도를 참조 표현의 문장 구조에 따라 결합한다. 따라서, 본 모델은 크게 총 네 가지 서브 네트워크들로 구성된다: 언어 표현 네트워크(LRN), 물체 정합 네트워크(OMN), 관계 정합 네트워크(RMN), 그리고 가중 결합 네트워크(WCN). 본 논문에서는 세 가지 서로 다른 참조 표현 데이터집합들을 이용한 실험을 통해, 제안 모델이 현존 최고 수준의 참조 표현 이해 성능을 보인다는 것을 입증하였다.

키워드 : 참조 표현 이해, 딥러닝, 맥락 정보, 가중 결합

1. 서 론

최근 딥러닝을 이용한 영상 및 자연어 처리 분야의 연구가 활발하게 진행되고 있다. 이에 따라 영상 처리와 자연어 처리가 복합된 문제에 관한 연구에도 관심이 높아지고 있다. 참조 표현 이해(referring expression comprehension) 문제 역시 이러한 복합 지능 문제 중 하나이다. 참조 표현(referring expression)이란 영상 내에 특정 물체를 가리키는 자연어 문장을 의미하고, 참조 표현 이해는 Fig. 1과 같이 한 영상에서 참조 표현이 가리키는 대상 물체의 영역(referred region)을 찾아내는 일을 말한다. 참조 표현 이해의 과정은 크게 자연어 참조 표현을 나타내는 의미 특징(semantic feature) - 혹은 언어 특징(linguistic feature)으로도 부름 - 과 영상 내 물체 후보 영역들의 시각 특징(visual feature)과 공간 특징(spatial feature)들을 추출하는 과정, 그리고 참조 표현과 영상에서 각기 추출된 특징들 간의 연관성을 비교하는 과정으로 이루어진다. 일반적으로 자연어 참조 표현의 의미 특징들은 단어의 의미와 순서를 고려하여 순환 신경망(Recurrent Neural Network, RNN)을 통해 추출되는 반면, 영상 내 물체 후보 영역의 시각 특징은 VGGNet, ResNet 등과 같은 합성곱 신경망(Convolutional Neural Network, CNN)을 통해 추출된다. 한편, 영상 내 한 물체 후보 영역의 공간 특징은 영상 내 해당 영역의 위치(location)와 크기(size)로 표현된다.

※ 본 연구는 과학기술정보통신부 및 정보통신기술진흥센터의 대학ICT연구센터육성지원사업의 연구결과로 수행되었음(IITP-2017-0-01642).
※ 이 논문은 2017년도 한국정보처리학회 추계학술발표대회에서 "참조 표현 이해를 위한 물체간의 관계 모델링"의 제목으로 발표된 논문을 확장한 것임.
† 준 회 원 : 경기대학교 컴퓨터과학과 석사과정
†† 종신회원 : 경기대학교 컴퓨터과학과 교수
Manuscript Received : December 11, 2017
Accepted : January 1, 2018
* Corresponding Author : Incheol Kim(kic@kyonggi.ac.kr)

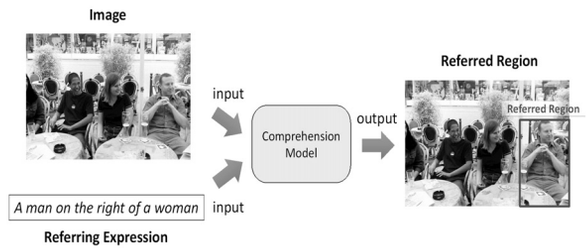


Fig. 1. Referring Expression Comprehension

선행 연구들은 참조 표현으로부터 영상 내 물체 영역들을 평가하기 위해, 참조 표현의 특징과 각 후보 물체 영역의 특징을 일대일(1:1) 비교하는 방법을 취하였다[1-4]. 이러한 방법은 주어진 참조 표현이 갖는 다양한 맥락 정보(contextual information)들을 충분히 이용할 수 없다. 맥락 정보란 참조 표현이 가리키는 대상 물체에 관련된 정보를 의미한다. 예를 들어 Fig. 1에서 참조 표현은 “a man”과 같은 대상 물체 정보(referred object information), “on the right of”와 같은 관계 정보(relationship information), “a woman”과 같은 보조 물체 정보(context object information)를 나타내는 맥락 정보들을 가진다. 이러한 맥락 정보들을 구분 없이 이용하면 그만큼 정확한 후보 영역 평가가 힘들어진다. 이러한 한계점을 보완하기 위해, 참조 표현이 나타내는 정보들을 부분 별로 나누어 영역을 평가하는 연구가 진행되었다[5]. 이와 같이 기존의 선행 연구들에서는 참조 표현이 나타내는 여러 맥락 정보들을 따로 구분하여 이용하려는 노력은 있었으나, 각 맥락 정보들을 어떻게 결합해야 보다 정확한 대상 물체 영역을 찾을 수 있느냐 하는 문제에 대해서는 구체적으로 연구된 바가 없다.

본 논문은 참조 표현 이해를 위한 새로운 심층 신경망 모델과 학습 방법을 제안한다. 본 논문에서 제안하는 모델은 효과적인 참조 표현 이해를 위해, 참조 표현이 가지는 맥락 정보들을 대상 물체 정보, 관계 정보, 보조 물체 정보로 나누어 처리한다. 또한, 본 논문에서 제안하는 모델은 이러한 다양한 맥락 정보들을 참조 표현에 의존적인 방식으로 가중 결합함으로써, 참조 표현에 부합하는 대상 물체 영역을 보다 정확히 탐지해낼 수 있도록 설계하였다. 본 논문에서 제안하는 모델의 우수성을 확인하기 위해, 대규모 참조 표현 데이터 집합[1, 6]들을 이용한 다양한 성능 비교 실험을 수행하고 그 결과를 소개한다.

2. 관련 연구

참조 표현 이해의 연구들은 주어진 참조 표현과 영상 내 존재하는 물체 영역들의 연관성을 비교하여 가장 연관성이 높은 물체 영역을 찾는 방법을 사용한다. 영상 내 존재하는 물체 영역들을 얻기 위해 Faster-RCNN[7], YOLO[8], SSD[9] 등의 물체 탐지 모델을 사용하거나, MS-COCO[10] 같은 영상 데이터 집합에 미리 정의된 물체 영역들을 사용한다. 참조 표현과 물체 영역들의 연관성 비교를 위해, 참조 표현과 물체 영역들의 특징을 추출하여 사용한다. 물체 영역의 특징 추출의 경우, VGGNet, ResNet 등의 합성곱 신경망 모델을 이용하여 시각 특징을 얻고, 영역의 위치와 크기 등을 이용하여 공간 특징을 얻는다. 참조 표현의 특징은 LSTM(Long Short-Term Memory)이나 Bi-LSTM(Bidirectional Long Short-Term Memory)을 이용하여 얻는다.

물체 영역과 참조 영역으로부터 얻은 특징들을 효과적으로 사용하기 위해 많은 연구들이 진행되었다. 초기 연구에서는 각 물체 영역의 평가를 위해, 해당 물체 영역의 특징을 주로 이용하고, 부가적으로 전체 영상의 특징을 이용하였다[1, 2, 4]. 물체 영역 특징으로부터 해당 영역 자체가 나타내는 정보를 얻고, 주어진 참조 표현의 문맥 정보와 같은지 비교하여 해당 물체 영역을 평가한다. 이러한 방법은 대상 물체 후보 영역에만 집중하기 때문에, 참조 표현에서 대상 물체 정보를 제외한 맥락 정보들을 충분히 이용할 수 없었다. 이러한 한계점을 보완하기 위해, 후속 연구들에서는 참조 표현에서 언급하는 대상 물체 영역뿐만 아니라 보조 물체 영역(context region)도 함께 비교하는 방법들이 시도되었다[3, 5]. 이러한 방법은 평가 대상을 한 영역이 아닌, 두 영역으로 구성된 한 영역 쌍(region pair)으로 확장한다. 영역 쌍의 첫 번째 영역은 대상 물체의 후보 영역으로써 사용되고, 두 번째 영역은 보조 물체의 후보 영역으로써 사용된다. 영역 쌍의 각 후보 영역별로 참조 표현의 문맥 정보와 비교하여 영역 쌍을 평가한다. 가장 높은 평가를 받은 영역 쌍에서 대상 물체의 후보 영역이 최종 대상 영역으로 선택된다. 평가 대상을 영역 쌍으로 확장하는 방법 외에도, 평가 하려는 물체 영역과 다른 물체 영역들의 시각적, 공간적 관계를 따로 입력하여 보조 물체 영역들을 고려하는 연구들도 진행되었다[6, 11]. 이러한 방법들은 대상 물체의 특성을 좀 더 구체적으로 분석함으로써 참조 표현 이해에 대한 성능도 높일 수 있었다. 하지만 참조 표현이 나타내는 문맥 정보들을 명시적으로 나누어 분석하지 않기 때문에, 물체 영역에 대한 정확한 평가가 어렵다는 한계점이 존재한다. 이에 따라 주어진 참조 표현이 가지는 문맥 정보들을 분류하고, 각 문맥 정보마다 입력 영역 쌍과의 연관성을 측정하고, 각 측정 결과를 결합하여 영역 쌍을 평가하는 연구가 진행되었다[5]. 이러한 연구는 각 문맥 정보들을 분류하는 동시에, 참조 표현에 나타난 물체간의 관계(relationship)까지 명시적으로 이용한다. 물체간의 관계란 Fig. 1의 예에서 “on the right of”와 같이, 대상 물체와 보조 물체 사이의 공간적 관계나 서술적 관계를 의미한다. 또한, 문맥 정보들을 분류하기 위해 외부 구문 분석기(parser)를 이용하는 방법[12]과, 심층 신경망을 통해 학습되는 구문 분석 모델을 이용하는 방법[5]이 사용되었다.

주어진 참조 표현에서 대상 물체뿐만 아니라 보조 물체 또는 더 나아가 두 물체간의 관계까지 고려하는 연구들은 참조 표현이 갖는 다양한 문맥 정보들을 이용하여 대상 영역을 찾는다. 하지만, 이용되는 문맥 정보들의 종류가 항상 고정되어있다는 점에서 한계점이 나타난다. 일부 문맥 정보만 갖는 간단한 참조 표현의 경우, 고정적으로 여러 문맥

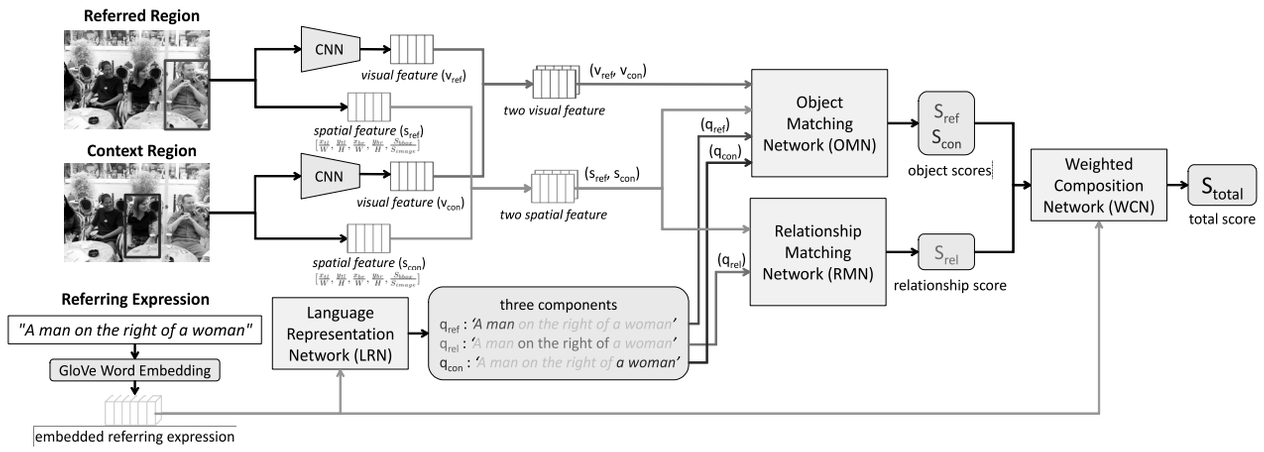


Fig. 2. Object-Relationship Modeling Network(ORMN) Structure

정보로 나누는 방법은 적절하지 않을 수 있다. 본 논문은 이러한 한계점을 해결하기 위해 참조 표현에 의존적인 평가 방식을 이용한다.

3. 참조 표현 이해 모델

3.1 전체 네트워크 구조

본 논문에서는 자연어 참조 표현에 등장하는 대상 물체와 보조 물체, 그리고 이 두 물체들 간의 관계 정보를 효과적으로 이용할 수 있는 물체 관계 모델링 네트워크(Object-Relationship Modeling Network, ORMN)를 제안한다. 본 논문에서 제안하는 ORMN은 Fig. 2와 같이 4 가지 부분 네트워크(sub-network)들로 구성된다. ORMN 모델의 입력은 참조 표현(referring expressions)과 영상(image), 그리고 영상 내 포함된 물체 후보 영역들(object regions)이다. 입력된 물체 후보 영역들은 ORMN 내부에서 Fig. 3과 같이 영역 쌍을 이루게 된다. 만들어진 영역 쌍의 첫 번째 영역은 대상 영역(referred region)의 후보가 되고, 두 번째 영역은 보조 영역(context region)의 후보가 된다. ORMN 모델은 주어진 참조 표현과 가장 연관성이 높은 영역 쌍을 찾기 위해, 각 영역 쌍 마다 참조 표현과의 적합도를 평가한다. 평가를 마친 후, 가장 높은 평가치를 받은 영역 쌍의 대상 영역이 최종 정답 영역으로 선택된다.

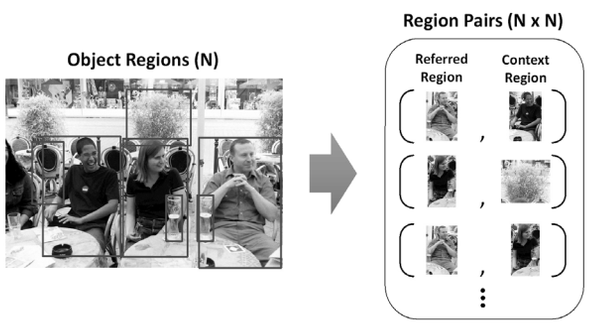


Fig. 3. Region Pair Generation

한편, 입력된 자연어 참조 표현은 언어 표현 네트워크(Language Representation Network, LRN)에 의해 각각 대상 물체 표현(referring object expression), 관계 표현(relationship expression), 보조 물체 표현(context object expression)을 나타내는 세 가지 표현 요소(q_{ref}, q_{rel}, q_{con})들로 나뉜다. 이는 참조 표현이 갖는 문맥 정보들을 각 표현 요소로써 나누는 것을 의미한다. 각 표현 요소는 이어지는 두 네트워크의 입력 데이터로 사용된다. 물체 정합 네트워크(Object Matching Network, OMN)는 영상에서 선택한 대상 물체 후보 영역과 보조 물체 후보 영역의 한 쌍이 자연어 참조 표현에서 추출한 대상 물체 표현 요소(q_{ref})와 보조 물체 표현 요소(q_{con})에 각각 얼마나 잘 부합하는지를 판별한다. 이를 위해 물체 정합 네트워크(OMN)에서는 영상의 두 물체 후보 영역들에서 추출한 시각 특징(visual feature)인(v_{ref}, v_{con})와 공간 특징(spatial feature)인 (s_{ref}, s_{con})을 이용하여 자연어 참조 표현에서 추출한 물체 표현 요소들인 (q_{ref}, q_{con})와의 적합도 (S_{ref}, S_{con})를 평가한다. 관계 정합 네트워크(Relationship Matching Network, RMN)는 영상 내 두 물체 후보 영역들의 공간 특징 (s_{ref}, s_{con})을 이용하여, 이 두 후보 영역들이 자연어 참조 표현에 언급된 두 물체 간의 관계 표현 요소 q_{con}와 얼마나 잘 부합하는 지를 나타내는 적합도 S_{rel}를 계산한다. 가중 결합 네트워크(Weighted Composition Network, WCN)는 세 가지 서로 다른 맥락 정보의 적합도를 나타내는 평가치 (S_{ref}, S_{rel}, S_{con})를 참조 표현 의존적인 방식으로 가중 합산하여 영역 쌍의 최종 평가치 S_{total}를 구한다.

3.2 언어 표현 네트워크

언어 표현 네트워크인 LRN은 Fig. 4와 같이, 주어진 참조 표현을 세 가지 표현 요소로 분할하는 역할을 수행한다. 각 표현 요소는 대상 물체를 나타내는 부분 q_{ref}, 보조 물체를 나타내는 부분 q_{con}, 대상 물체와 보조 물체의 관계를 나타내는 부분 q_{rel}을 의미한다.

참조 표현의 특징을 추출하기 위해, GloVe 임베딩(embedding)

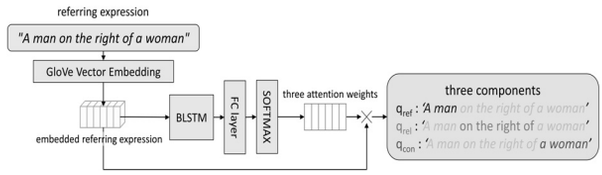


Fig. 4. Language Representation Network(LRN)

[13]을 적용하여 임베딩된 참조 표현(embedded referring expression)을 얻는다. 임베딩된 자연어 참조 표현으로부터 시계열 패턴의 학습에 매우 효과적인 순환 신경망 Bi-LSTM을 적용하여 의미 특징 벡터를 추출한다. 추출된 의미 특징 벡터는 완전 연결 계층과 소프트맥스(softmax) 연산을 거친 후, 대상 물체 표현 요소, 보조 물체 표현 요소, 두 물체간의 관계 표현 요소에 해당하는 세 가지 요소 가중치(three component weights)로 변환된다. 마지막으로 임베딩된 참조 표현과 세 가지 요소 가중치를 곱하고, 세 가지 표현 요소 (q_{ref}, q_{rel}, q_{con})를 얻는다.

3.3 물체 정합 네트워크

영상 내 대상 물체 후보 영역(b_{ref})과 보조 물체 후보 영역(b_{con})이 각각 자연어 물체 참조 요소들(q_{ref}, q_{con})에 얼마나 잘 부합하는지를 평가하는 물체 정합 네트워크인 OMN은 Fig. 5와 같이 구성된다.

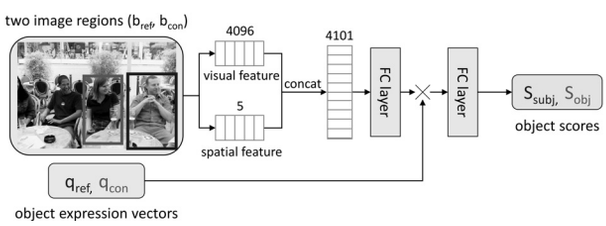


Fig 5. Object Matching Network(OMN)

OMN은 영상 내 각 물체 후보 영역에 대한 특징으로 시각 특징(v_{ref}, v_{con})와 공간 특징(s_{ref}, s_{con})을 이용한다. 각 물체 영역에 해당하는 영상 부분의 시각 특징들은 합성 곱 신경망(CNN)인 VGG-16 모델을 적용하여 추출한다.

$$\left[\frac{x_{tl}}{W}, \frac{y_{tl}}{H}, \frac{x_{br}}{W}, \frac{y_{br}}{H}, \frac{S_{bbox}}{S_{image}}\right] \quad (1)$$

한편, 각 물체 후보 영역의 공간 특징은 Equation (1)과 같이 계산되며, 이들은 전체 영상에 대한 해당 영역의 상대적 위치와 크기 정보를 나타낸다. W와 H는 원 영상의 너비와 높이, S_{image}는 넓이를 나타내고 x, y는 물체 영역의 상대적인 위치 좌표 값, S_{bbox}는 상대적인 크기를 나타낸다. 이러한 시각 특징과 공간 특징을 완전 연결 계층(FC layer)에 입력하여 해당 영역을 나타내는 영역 특징 벡터를 구한다. 그리고 주어진 표현 요소 벡터(q_{ref}, q_{con})와 요소 별 곱셈을 수행하고 완전 연결 계층(FC layer)을 거쳐 물체 적합도 평가치인 (S_{ref}, S_{con})를 계산한다.

3.4 관계 정합 네트워크

영상 내 대상 물체 후보 영역(b_{ref})과 보조 물체 후보 영역(b_{con})의 공간적 관계(spatial relationship)가 자연어 관계체 참조 요소인 q_{rel}에 얼마나 잘 부합하는지를 평가하는 관계 정합 네트워크인 RMN은 Fig. 6과 같이 구성된다.

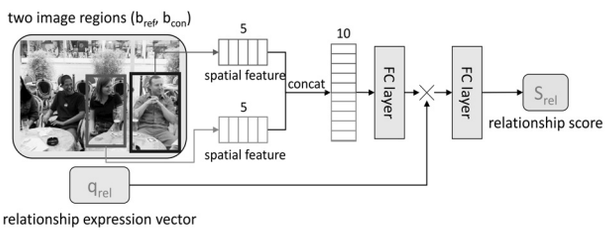


Fig. 6. Relationship Matching Network(RMN)

RMN은 영상 내 두 물체 후보 영역 각각의 공간 특징들(spatial features)로부터 두 물체 영역간의 공간적 관계를 나타내는 관계 특징을 추출하기 위해, 두 물체 영역의 공간 특징들을 단순 결합(concatenation)한 다음, 완전 연결 계층(FC layer)에 입력한다. 두 물체 영역간의 관계 특징과 q_{rel}의 요소 별 곱셈을 수행한 후, 그 결과를 다시 완전 연결 계층(FC layer)에 입력하여 두 물체간의 관계 적합도를 나타내는 평가치 S_{rel}를 계산한다.

3.5 가중 결합 네트워크

물체가 담긴 영상과 자연어 참조 표현으로부터 추출한 다양한 맥락 정보들을 결합함으로써 물체 후보 영역의 최종적인 적합도를 평가하는 가중 결합 네트워크인 WCN은 Fig. 7과 같이 구성된다.

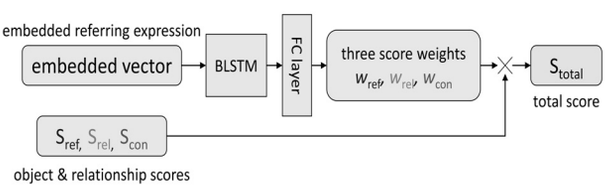


Fig. 7. Weighted Composition Network(WCN)

WCN은 앞서 설명한 OMN과 RMN을 통해 얻어낸 세 가지 평가치 (S_{ref}, S_{rel}, S_{con})를 참조 표현 의존적인 방식으로 가중 결합하여 최종 평가치 S_{total}를 계산한다. 참조 표현을 구성하는 표현 요소별로 차별적인 평가를 위해, LRN의 임베딩된 참조 표현을 순환 신경망 Bi-LSTM 계층에 입력하고, 완전 연결 계층(FC layer)을 거쳐 세 가지 평가 가중치 (w_{ref}, w_{rel}, w_{con})를 계산한다. 그리고 세 가중치(w_{ref}, w_{rel}, w_{con})와 평가치(S_{ref}, S_{rel}, S_{con})를 이용하여 Equation (2)와 같이 가중합을 계산함으로써, 최종 평가치 S_{total}을 구한다.

$$S_{total} = w_{ref}S_{ref} + w_{rel}S_{rel} + w_{con}S_{con} \quad (2)$$

4. 구현 및 실험

4.1 데이터 집합

본 연구에서는 MS-COCO 데이터 집합의 영상을 기반으로 만들어진 대규모 참조 표현 데이터 집합인 Google RefExp[1], RefCOCO[4], RefCOCO+[4]를 실험에 사용한다. Google RefExp는 총 26,711개의 영상 데이터와 각 영상에 대한 104,560개의 참조 표현 데이터로 구성되어 있으며, 주로 길고 복잡한 참조 표현이 포함되어있다. 마찬가지로 RefCOCO는 19,994개의 영상 데이터와 각 영상에 대한 142,209개의 참조 표현 데이터로 구성되어 있으며, 주로 짧은 참조 표현이 포함되어있다. RefCOCO+는 19,992의 영상 데이터와 각 영상에 대한 141,564개의 참조 표현 데이터로 구성되어 있으며, RefCOCO보다 조금 더 길고 복잡한 참조 표현이 포함되어있다. 모델 성능 평가 시, RefCOCO와 RefCOCO+의 경우는 미리 정의된 실험용 데이터 집합(test dataset)을 사용하지만, Google RefExp의 경우는 실험용 데이터 집합이 정의되어 있지 않기 때문에 검증 데이터 집합(validation dataset)을 실험용으로 사용한다. RefCOCO와 RefCOCO+의 실험 데이터 집합은 TestA와 TestB로 나뉜다. TestA는 대상 물체 영역이 사람인 집합이고, TestB는 대상 물체 영역이 물체인 집합이다.

4.2 구현

본 논문에서 제안하는 ORMN 모델을 학습하기 위해서는 입력과 이에 대한 정답 출력의 쌍으로 학습 데이터를 구성해야 한다. ORMN 모델의 경우, 정답 출력은 영상 내의 정답 대상 물체 영역과 정답 보조 물체 영역으로 이루어진 영역 쌍(region pair)이다. 하지만 참조 표현 데이터 집합은 대상 물체 영역과 달리, 보조 물체 영역을 따로 정의하지 않는다. 따라서 학습에 필요한 정답 보조 물체 영역을 결정해주는 방법이 필요하다. 이를 위해 본 연구에서는 대상 물체 영역 b_{ref}의 보조 물체 영역은 b_{ref}를 갖는 영역 쌍 (b_{ref}, b_x)들 중, 가장 높은 평가치를 얻은 영역 쌍의 b_x로 대체한다. 결과적으로 N개의 영역이 주어졌을 때 $N*N$개의 영역 쌍이 만들어지고, 같은 대상 물체 영역 후보를 갖는 영역 쌍 중 가장 높은 평가치의 영역 쌍만 선택한다. 선택된 N개의 영역 쌍으로부터 손실 함수(loss function)인 소프트맥스 크로스엔트로피(softmax cross-entropy)를 구하고, 이를 통해 모델을 학습한다. ORMN 모델의 손실 함수는 Equation (3)과 Equation (4)와 같이 정의한다.

$$S(b_i) = \frac{e^i}{\sum_j^N e^j} \quad (3)$$

$$Loss = -\sum_i^N L_i \log(S(b_i)) \quad (4)$$

$S(b_i)$는 전체 영역 쌍의 평가치에 대한 영역 쌍 b_i의 소프트맥스 연산 결과이고, L_i는 참조 표현이 가리키는 정답 영역을 나타낸 벡터 값을 나타내며 정답 영역은 1, 나머지 영역들은 0의 값을 가진다. 모델의 초기 학습률은 0.01로 설정하고, 일정 수준 학습이 완료되면 학습률을 0.001로 낮추어 학습한다. 손실 함수의 최적화로는 모멘텀 최적화 기법(momentum optimizer)을 사용한다. 이러한 모델 구현을 위해 Ubuntu 16.04 LTS 환경에서 Python 딥러닝 라이브러리인 Tensorflow를 사용하였다. 모델의 학습과 실험은 GeForce GTX TITAN X GPU카드가 설치된 하드웨어 환경에서 수행하였다.

4.3 실험

첫 번째 실험에서는 본 논문의 ORMN 모델에서 제안한 대상 물체와 보조 물체간의 관계(relationship)가 성능 향상에 미치는 효과를 분석하였다. 이를 위해 비교 특징 집합들을 (v_{ref}, s_{ref}, q_{ref})와 같은 대상 물체 특징(REF), (v_{con}, s_{con}, q_{con})와 같은 보조 물체 특징(CON), (v_{rel}, s_{rel}, q_{rel})와 같은 두 물체간의 관계 특징(REL)들로 나누고, 각 특징 집합을 이용했을 때 모델의 성능을 비교 측정하였다. 실험은 Google RefExp 데이터 집합을 이용하여 모델 학습과 평가를 수행하였다. 모델 평가의 지표가 되는 정확도(accuracy)는 모델이 대상 물체 영역을 올바르게 예측한 비율을 나타낸다. 만약 모델이 예측한 대상 물체 영역과 실제 대상 물체 영역(ground-truth)의 IOU(Interaction of Union)가 0.5 이상이면 올바른 예측으로 판단하고, 그렇지 않은 경우는 올바르지 않은 예측으로 판단한다.

Table 1. Performance Comparison among Four Different Feature Sets

Features	Accuracy
REF	68.6 %
REF + CON	68.7 %
REF + REL	69.1 %
REF + REL + CON	**69.5 %**

Table 1의 결과는 대상 물체 특징(REF)만을 이용하는 것보다, 보조 물체 특징(CON)과 관계 특징(REL)을 추가적으로 이용하는 것이 성능에 더 도움을 준다는 것을 보여준다. 특히, 보조 물체 특징보다 관계 특징이 모델의 성능 향상에 더 크게 기여한다는 것을 알 수 있다. 그 이유는 관계에 관한 정보가 찾으려는 대상인 대상 물체 영역과 더 관련성 있기 때문이다. 관계를 모르는 상태에선 보조 물체에 대한 정보는 아무런 도움이 될 수 없다. 이는 Table 1에서도 확인할 수 있다. 결과를 살펴보면, 관계 특징 없이 보조 물체 특징을 추가했을 때의 정확도 상승폭보다 관계 특징을 포함하

여 추가했을 때의 상승폭이 더 높게 나타난다.

두 번째 실험은 본 모델의 WCN에서 채용한 참조 표현 의존적 가중 결합 평가 방법이 모델 성능 향상에 미치는 효과를 분석한다. 이를 위해, 본 실험에서는 세 가지 서로 다른 평가치 결합 방법을 비교한다. 첫 번째는 서로 다른 평가치 (S_{ref}, S_{rel}, S_{con})에 대한 평가 가중치를 모두 1로 설정하는 방법(fixed same weights, Fixed SW), 두 번째는 평가 가중치를 참조 표현에 독립적인 방식으로 학습하는 방법(referring expression-free adjustable weights, RE-free AW). 세 번째는 본 모델에서 제안하는 방식으로 참조 표현의 표현 요소 구성에 따라 적응적으로 평가 가중치를 학습하는 방법(referring expression dependent adjustable weights, RE-dependent AW)이다.

Table 2. Performance Comparison among Three Different Weighting Methods

Weighting	Accuracy
Fixed SW	68.5 %
RE-free AW	69.0 %
RE-dependent AW	**69.5 %**

Table 2의 결과는 본 모델에서 제안하는 참조 표현 의존적인 방식으로 가중 결합하는 RE-dependent AW 방식이 성능 향상에 더 도움이 된다는 것을 보여준다. 이는 참조 표현 마다 집중되어야하는 맥락 정보가 다르다는 것을 의미한다. 이러한 이유는, 관계가 나타나지 않는 참조 표현은 S_{ref}의 비중이 더 커야하고, 관계가 나타나는 참조 표현은 (S_{ref}, S_{rel}, S_{con}) 모두가 고르게 결합되어야 하기 때문이다.

Fig. 8은 RE-dependent AW 방식을 사용했을 때, 참조 표현에 따라 예측된 가중치들을 나타낸다. 왼쪽 그래프는 세 가지 요소 가중치 별로, 참조 표현의 각 단어가 어느 정도의 중요도를 갖는 지 예측한 결과를 나타낸다. 오른쪽 그래프는 각 참조 표현의 단어 구조에 따른 세 가지 평가 가중치를 예측한 결과를 나타낸다. 세 번째 참조 표현의 경우, 관계와 보조 물체에 대한 표현이 존재하지 않아서, 각 요소 가중치에 대한 예측이 잘 이루어지지 않는다. 이러한 단점을 세 가지 요소에 대한 적응적 평가 가중치를 사용하는 RE-dependent AW 방식이 잘 보완할 수 있다.

세 번째 실험은 본 논문에서 제안한 모델의 분리 학습 방법(separate training)의 효과를 분석한다. 이를 위해 본 논문에서 제안한 바와 같이 LRN, OMN, RMN을 미리 학습시키고, 학습된 네트워크에 WCN를 추가하여 학습시키는 분리 학습 방법(separate training)과 전체 네트워크를 한 번에 학습하는 종단 간 학습 방법(end-to-end training)간의 성능을 서로 비교해보았다. Table 3의 결과는 분리 학습 방법의 성능이 종단 간 학습 방법의 성능보다 높다는 것을 보여준다. 그 이유는, 종단 간 학습 방법의 경우 WCN이 이전 네트워크들의 기능을 분담하도록 학습되지만, 분리 학습은 WCN이 이전 네트워크들과 독립적으로 학습되어 제 기능을

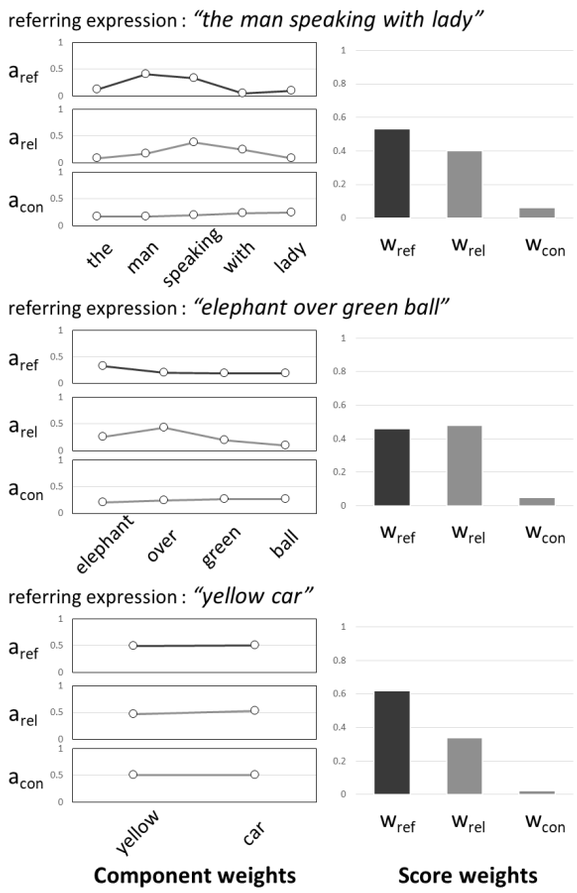


Fig. 8. Weights Comparison among Referring Expressions

수행할 수 있기 때문으로 판단한다.

마지막 실험은 본 논문에서 제안하는 모델인 ORMN의 성능을 최근 다른 선행 연구들과 비교한다. 비교 대상들인 MMI은 Mao의 연구[1]에서, Negative Bag(NB)은 Nagaraja의 연구[2]에서, Comprehension-Guided(CG)는 Luo의 연구[5]에서, CMN은 Hu의 연구[7]에서 각각 제안한 모델들이다. Table 4의 결과를 통해, 본 논문에서 제안한 모델인 ORMN이 기존 모델들보다 더 우수한 성능을 갖는 것을 확인할 수 있었다.

Table 3. Performance Comparison between Two Different Training Methods

Training	Epoch	Accuracy
end-to-end training	140,000	68.9 %
	150,000	69.5 %
	160,000	69.3 %
separate training	150,000 (LRN, OMN, RMN), 20,000 (WCN)	69.8 %
	150,000 (LRN, OMN, RMN), 30,000 (WCN)	**69.9 %**
	150,000 (LRN, OMN, RMN), 40,000 (WCN)	69.8 %

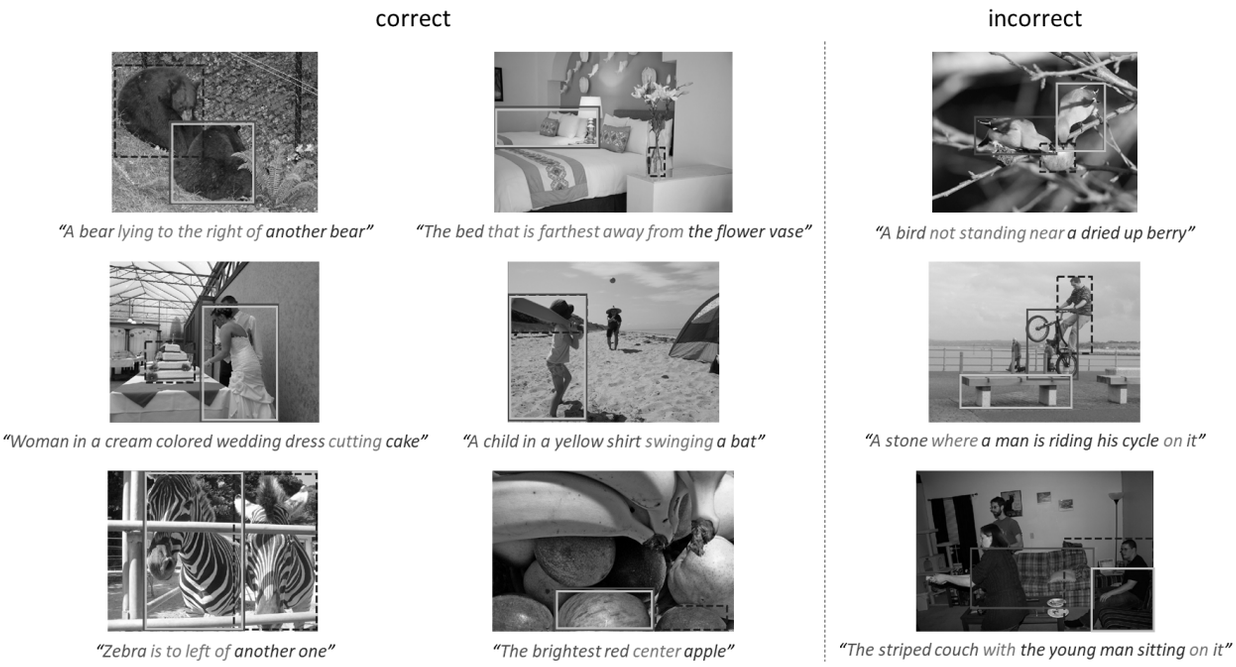


Fig. 9. Referring Expression Comprehension Results using ORMN Model

Table 4. Performance Comparison with Other State-of-the-art Models

Models	RefCOCO		RefCOCO+		Google RefExp
	TestA	TestB	TestA	TestB	val
MMI [1]	71.7 %	71.1 %	52.4 %	47.5 %	62.1 %
NB [2]	75.6 %	78.0 %	-	-	68.4 %
CG [4]	74.0 %	73.4 %	60.3 %	55.0 %	65.3 %
CMN [5]	76.0 %	78.5 %	60.5 %	60.1 %	69.0 %
ORMN	77.4 %	**79.3 %**	60.6 %	**60.2 %**	**69.9 %**

Fig. 9은 ORMN 모델을 이용한 참조 표현 이해를 나타낸다. 각 영상에서 빨간색 테두리 영역과 파란색 테두리 영역은 각각 모델이 예측한 대상 물체 영역과 보조 물체 영역을 나타내고, 주황색 테두리 영역은 해당 참조 표현의 실제 대상 물체 영역(ground-truth)을 나타낸다. 영상의 참조 표현은 대상 물체 표현, 관계 표현, 보조 물체 표현으로 나뉘며, 각각 빨간색, 초록색, 파란색으로 표시한다.

Fig. 9의 'correct' 영역은 ORMN 모델이 대상 물체 영역을 올바르게 예측한 결과를 나타낸다. 결과를 살펴보면 대상 물체 영역뿐만 아니라, 보조 물체 영역까지 정확하게 예측한 것을 확인할 수 있다. 이는 ORMN 모델이 다양한 정보를 갖는 참조 표현을 잘 이해하고, 영상 정보와 잘 부합시킬 수 있다는 것을 의미한다. 'incorrect' 영역은 대상 물체 영역을 올바르게 예측하지 못한 결과를 나타낸다. 먼저 첫 번째 결과는 부정 의미 "not"을 통해 대상 물체를 나타내는 경우이다. 학습에 사용되는 대부분의 참조 표현은 부정 의미를 포함하지 않기 때문에, "not"과 같은 부정 의미 학습이 어렵다. 두 번째와 세 번째 결과는 복잡한 문장 구조를 갖는 참조 표현의 경우이다. 대부분의 참조 표현은 'correct'의 예시들처럼 대상 물체 표현, 관계 표현, 보조 물체 표현 순으로 나열된다. 이에 따라 학습 모델은 일정한 순서로 참조 표현을 나누도록 학습되기 때문에, 두 번째와 세 번째 참조 표현처럼 표현 요소들의 순서가 뒤섞인 경우, 참조 표현에 대한 올바른 판단이 어렵다.

향후 다양한 형태의 참조 표현 이해를 위해, 참조 표현의 정확한 의미 분석이 필요하다. 각 단어들은 참조 표현에서 어떤 의미를 갖는지에 따라 적절한 표현 요소로 분류되어야 하고, 이를 바탕으로 영상 정보와 더 정확하게 비교할 수 있어야 한다.

5. 결 론

본 논문은 참조 표현 이해를 위한 심층 신경망 모델을 제안하였다. 본 논문에서 제안하는 모델(ORMN)은 참조 표현으로부터 대상 물체 정보, 보조 물체 정보, 두 물체 간의 관계 정보를 얻어내고, 얻어낸 각 문맥 정보와 주어진 영상의 시각 및 공간 특징을 비교하여 각 정보 별 평가치를 구한다. 그리고 참조 표현이 실제로 갖는 정보들을 분석하고 의미 있는 평가치를 가중하여 결합한다. 또한 실험을 통하여 참조 표현의 각 정보들을 이용하는 것이 실제로 성능 향상에 기여하는 것을 확인하였고, 각 정보들을 참조 표현의 형태에 따라 가변적으로 이용하는 것이 더 효과적인 방법인 것을 확인하였다. 또한 다양한 데이터 집합을 이용한 비교 실험들을

통해, 본 논문에서 제안한 모델의 우수성을 확인할 수 있었다. 향후 연구로는 참조 표현이 갖는 여러 정보들을 보다 정확하게 분류할 수 있는 방법을 고안하고, 복잡한 관계를 갖는 참조 표현 이해를 위해 모델의 구조를 확장할 계획이다.

References

[1] J. Mao, J. Huang, A. Toshev, O. Camburu, A. L. Yuille, and K. Murphy, "Generation and Comprehension of Unambiguous Object Descriptions," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, pp.11-20, 2016.

[2] R. Luo and G. Shakhnarovich, "Comprehension-Guided Referring Expressions," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2017.

[3] V. K. Nagaraja, V. I. Morariu, and L. S. Davis, "Modeling Context Between Objects for Referring Expression Understanding," *Proceedings of the European Conference on Computer Vision(ECCV)*, 2016.

[4] R. Hu, H. Xu, M. Rohrbach, J. Feng, K. Saenko, and T. Darrell, "Natural Language Object Retrieval," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.4555-4564, 2016.

[5] R. Hu, M. Rohrbach, J. Andreas, T. Darrell, and K. Saenko, "Modeling Relationships in Referential Expressions with Compositional Modular Networks," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.1115-1124, 2017.

[6] L. Yu, P. Porison, S. Yang, A. C. Berg, and T. L. Berg, "Modeling Context in Referring Expressions," *Proceedings of the European Conference on Computer Vision(ECCV)*, pp.69-85, 2016.

[7] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *Proceedings of the Neural Information Processing Systems(NIPS)*, pp.91-99, 2015.

[8] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, pp.779-788, 2016.

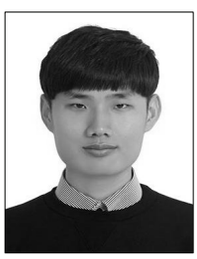
[9] W. Liu, D. Anguelow, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, "SSD: Single Shot MultiBox Detector," *Proceedings of the European Conference on Computer Vision(ECCV)*, pp.21-37, 2016.

[10] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick, "Microsoft COCO: Common Objects in Context," *Proceedings of the European Conference on Computer Vision(ECCV)*, pp.740-755, 2014.

[11] L. Yu, H. Tan, M. Bansal, and T. L. Berg, "A Joint Speaker-Listener-Reinforcer Model for Referring Expressions," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, pp.7282-7290, 2017.

[12] J. Krishnamurthy and T. Kollar, "Jointly Learning to Parse and Perceive: Connecting Natural Language to the Physical World," *Proceedings of the Transactions of the Association for Computational Linguistics(TACL)*, Vol.1, pp.193-206, 2013.

[13] J. Pennington, R. Socher, and C. Manning, "GloVe: Global Vectors for Word Representation," *Proceedings of the Conference on Empirical Methods in Natural Language Processing(EMNLP)*, pp.1532-1543, 2014.



신 동 협

http://orcid.org/0000-0003-4459-8545
e-mail : sdh9446@gmail.com
2018년 경기대학교 컴퓨터과학과(학사)
2018년~현 재 경기대학교 컴퓨터과학과 석사과정
관심분야 : 인공지능, 컴퓨터비전

김 인 철

http://orcid.org/0000-0002-5754-133X
e-mail : kic@kyonggi.ac.kr
1985년 서울대학교 수학과(이학사)
1987년 서울대학교 전산과학과(이학석사)
1995년 서울대학교 전산과학과(이학박사)
1996년~현 재 경기대학교 컴퓨터과학과 교수
관심분야 : 인공지능, 기계학습