Journal of the Korean Society of Marine Engineering, Vol. 41, No. 7 pp. 683~689, 2017
J. Korean Soc. of Marine Engineering (JKOSME)
https://doi.org/10.5916/jkosme.2017.41.7.683

ISSN 2234-7925 (Print)
ISSN 2234-8352 (Online)
Original Paper

객체 검출 기반 이미지 캡션 알고리즘 연구

이준희[1] · 김종찬[2] · 서동환†

(Received July 26, 2017 ; Revised August 4, 2017 ; Accepted August 24, 2017)

A study on image caption algorithm based on object detection

Jun-Hee Lee[1] · Jong-Chan Kim[2] · Dong-Hoan Seo†

요약: 어떠한 이미지에서 검출된 객체간의 상호 연관관계를 이해하는 기술인 이미지 캡션은 향후 자동 감시 시스템에 필수적이기 때문에 컴퓨터 비전 및 자연언어 처리 분야에서 많은 연구가 진행되고 있다. 이미지 전체를 학습하는 기존의 이미지 캡션 방식은 각 이미지 영역의 부분적인 상황 이해가 힘들며 동시에 새로운 이벤트 발생에 취약하다. 본 논문에서는 이를 해결하기 위해 검출된 객체의 경계박스 면적과 객체간의 거리를 이용하여 객체 정보를 생성하는 CNN(Convolutional Neural Network)기반의 객체 검출단과 RNN(Recurrent Neural Network) 기반의 문장 생성단을 결합하여 이미지의 세부적인 문장을 생성하는 모델을 제안한다. 제안한 방식은 CNN 기반의 객체 검출단의 느린 처리속도를 향상시키기 위하여 YOLO(You Only Look Once) 네트워크의 Grid 방식을 적용하였으며, 객체 검출단과 문장 생성단으로 나누어진 두 네트워크를 하나의 인공 신경망 모델로 구성하여 객체 검출 기반 이미지 캡션이 가능하다. 제안한 모델은 캡션 데이터세트인 Flickr 8K, Flickr 30K, MS COCO를 이용하여 학습하였고 BLEU(BiLingual Evaluation Understudy) 스코어 방식을 사용하여 캡션 성능을 검증하였다.
주제어: 객체 검출, 이미지 캡션, Convolutional neural network, Recurrent neural network

Abstract: Image caption, which is a method to understand the interrelationships between objects detected in a certain image, is being actively studied in the fields of computer vision and natural language processing, because it is expected to be an essential part of the future automatic monitoring systems. The conventional image caption method, which learns the whole image, finds it difficult to understand partial situations in different image regions, and is vulnerable in the case of the occurrence of a new event at the same time. In order to solve these problems, this paper presents a new model that generates detailed sentences of images by combining a convolutional neural network (CNN)-based object detection step, which generates object information using the bounding box area of the detected object and the distance between the objects, and a sentence generation step based on recurrent neural networks . The proposed model adopts the grid method of the YOLO (You Only Look Once) network to improve the slow processing speed of the CNN-based object detection step. It is possible by the image caption based on object detection because the two network models involved in the object detection and sentence generation steps are composed of one artificial neural network model. The proposed model learns using the caption data sets Flickr 8K, Flickr 30K, and MS COCO, and the caption performance can be verified using the BLEU (BiLingual Evaluation Understudy) score method.
Keywords: Object detection, Image caption, Convolutional neural network, Recurrent neural network

1. 서 론

최근 복잡해지는 실내외 구조에 따라 동일한 공간에서 요구하는 실시간 보안 시스템의 규모가 점점 증대되고 있다. 특히, CCTV 기반의 영상 감시 장치는 보안시스템의 핵심 기술로서 사람이 실시간으로 감시 영상을 분석하는 방식을 사용하고 있기 때문에 사람의 피로도가 누적됨에 따라 분석의 능률이 떨어지고, 영상 분석에 개인의 주관적인 판단이 들어가기 때문에 감시 영상 분석의 차이가 발생한다. 이러한 문제를 해결하기 위하여 실시간 영상 감시 시스템의 자동화 연구가 활발히 진행되고 있다. 실시간 영상 감시 시스템은 단순히 영상 내의 객체를 파악하는 객체 검출, 이동이나 변화를 파악하는 객체 추적과 객체들의 관계를 통해 문장을 생성해내는 이미지 캡션 기법 등이 대표적인 개발 분야이다. 이들 중, 객체 검출은 컴퓨터 비전분야에서

† Corresponding Author (ORCID: http://orcid.org/0000-0003-3610-0356): Division of Electronics and Electrical Information Engineering, Korea Maritime and Ocean University, 727, Taejong-ro, Yeongdo-gu, Busan 49112, Korea, E-mail: dhseo@kmou.ac.kr, Tel: 051-410-4412
1 Department of Electrical and Electronics Engineering, Korea Maritime and Ocean University, E-mail: ljh9961@kmou.ac.kr, Tel: 051-410-4822
2 Department of Electronics, Kyungbuk College, E-mail: kjc@kbc.ac.kr, Tel: 054-630-5067

매우 우수한 성능을 보이고 있으며 이를 자연언어처리 분야와 결합하여 이미지의 내용을 문장으로 만들어주는 이미지 캡션 기법은 객체 검출 분야의 발전과 더불어 최근 수년간 급격하게 발전하고 있으나 여전히 객체 검출 및 추적 분야에 비해 정확성이 다소 부족하다.

기존 연구들에서는 CNN(Convolutional Neural Network)을 사용하여 이미지 전체를 학습하고 생성된 특징맵을 Recurrent Neural Network(RNN)에 적용하여 이미지 캡션을 생성하는 모델이 대표적이다. 이러한 네트워크의 구조는 CNN의 학습 성능에 따라 객체 검출에 많은 영향을 받고, 이미지 전체를 사용하기 때문에 이미지의 부분적인 상황 이해가 어려우며, CNN의 성능에 따라 이미지 캡션의 성능에 영향을 많이 받는다**[1][2]**.

A. Karpathy *et al.* **[3]**은 이미지와 이를 통해 생성된 캡션을 동일한 벡터 공간에 위치하도록 학습하고, 새로운 이미지에 대한 캡션 생성을 위해 학습된 이미지와 새로운 이미지의 유사도를 이용하여 학습된 이미지의 캡션을 가져오는 방식이다. 이러한 방식은 이미지에서 키워드와 같은 간단한 단어를 생성하기 때문에 문장 수준의 캡션 생성에는 적합하지 않다. 또한 학습이 되어 있지 않은 새로운 이미지가 입력될 경우, 학습된 이미지의 캡션 내에서 단어를 출력하기 때문에 캡션의 정확도가 다소 부정확하다.

C. Liu *et al.* **[4]**은 RNN의 레이어를 인코딩 영역과 디코딩 영역으로 분리하여 인코딩 영역에서는 CNN 기반 객체 검출기의 특징맵을 사용하고 디코딩 영역에서는 단어를 생성하여 결합하는 방식을 적용하였다. 이러한 방식을 통해 이미지에서 객체를 정확하게 검출하고 단어를 결합하기 때문에 정확한 캡션이 가능하지만, 기존의 RNN에 비해 연산이 복잡한 LSTM(Long Short-Term Memory)를 사용하기 때문에 처리 속도가 상대적으로 느리다.

본 논문에서는 부분적인 상황 이해가 어려운 기존의 이미지 캡션 모델의 단점을 극복하기 위해 고속 객체 검출이 가능한 YOLO 네트워크를 통해 객체를 검출하고 검출된 객체의 경계박스의 면적과 객체 간 거리에 따른 가중치를 사용하여 객체 정보를 생성하는 객체 검출단을 구성하고, LSTM의 복잡한 알고리즘으로 인해 발생하는 느린 연산 속도를 해결하기 위해 알고리즘 동작 방식을 최적화 시킨 Gated Recurrent Unit(GRU)**[5]**으로 구성된 문장 생성단을 사용하여 객체 검출 기반 이미지의 캡션이 가능한 모델을 제안한다. 또한 캡션 데이터세트인 Flickr 8K, Flickr 30K, MS COCO를 사용하여 학습을 진행하고 BLEU 스코어 방식을 통해 객관적인 성능을 검증하였다.

2. 관련 이론

2.1 객체 검출

최근 머신러닝 분야의 성장과 함께 CNN은 객체 검출 분야뿐만 아니라 신호의 특징을 추출해야하는 영역에서 매우 우수한 성능을 보여주고 있다. 이미지에서의 객체 검출 기술은 여러 픽셀의 결합을 통해 만들어지는 컴퓨터 비전의 특성을 이용하여 주변 픽셀을 연결하고 학습하는 VGGNeT**[6]**, GoogleNet**[7]**, ResNet**[8]**과 같이 매우 높은 성능을 보이는 CNN 기반의 모델들이 적용되고 있다. 이와 같은 객체 검출 모델은 CNN을 다층 레이어 구조로 설계하여 높은 성능을 보이지만, 이로 인해 연산량이 급격하게 증가하기 때문에 객체 검출의 처리 속도가 느리다. 따라서 객체 검출 모델을 실시간으로 적용하기 위한 다양한 연구가 진행되고 있다.

최근 Faster Region Convolutional Neural Network(Faster R-CNN)**[9]**는 기존의 Fast R-CNN의 이미지 학습 과정에서 이미지를 자르거나 사이즈를 조절하는 전처리작업이 불필요하기 때문에 학습 이미지의 손상이 없으며 CNN 모델과 독립적으로 구성되어 있던 Region Proposal 알고리즘을 하나의 CNN 레이어로 구현함으로서 객체 검출의 처리 속도를 향상시켰다.

YOLO**[10]**는 객체 검출에 사용되는 Proposal 방식의 느린 검출 속도를 해결하기 위해 Grid 방식을 적용하여 검출함으로써 경계박스 예측 시간이 매우 빠르며 검출과 동시에 Class의 분포 확률을 계산하기 때문에 고속 객체 검출이 가능하다.

2.2 이미지 캡션

최근 이미지의 캡션 분야의 연구는 RNN을 사용한 Encoder와 Decoder 방식의 모델**[11]**과 같은 초기 모델을 시작으로 많은 연구가 진행되고 있으며, 단순한 RNN 방식을 통한 문장 생성 시 문장의 길이가 길어질수록 이전 단어에 대한 정보가 소멸되는 Vanishing Gradient Problem을 해결하기 위해 LSTM, GRU와 같은 알고리즘을 많이 사용하고 있다. 또한 문장의 정확도 향상을 위해 문장 생성에 필요한 정보들을 취합하여 문장 생성시 사용하는 Multimodal 레이어, Attention 레이어, Merge 레이어와 같은 방식을 도입하였다.

O. Vinyals *et al.* **[1]**은 CNN과 RNN의 융합을 통해 이미지 캡션에 관한 연구가 활발하게 진행 될 수 있는 가능성을 보였으며, CNN의 특징맵을 사용하여 이미지에서 객체를 검출하고, LSTM을 사용하여 이미지 캡션을 생성함으로써 Vanishing Gradient Problem을 해결하였다.

C. K. Lee **[2]**는 **[6]**의 CNN 특징맵을 사용하여 객체를 검출하고 GRU를 사용하여 이미지 캡션을 진행하였다. 또한 Multimodal Layer를 사용하여 단어 생성의 단계마다 CNN으로부터 이미지 정보를 추출하여 사용하는 특징을 가진다.

J. Mao *et al.* **[12]**는 Multimodal RNN을 사용하여 이미지 캡션을 진행하는 모델을 제안하였다. 이 모델은 단어를 생성하기 위한 Embedding 레이어를 2개의 층으로 구성하고 있으며 Multimodal 레이어와의 연결을 통해 이미지 캡션을 진행하지만 기본적인 RNN을 사용하기 때문에 Vanishing Gradient Problem 문제가 발생하기 쉽다.

3. 제안하는 이미지 캡션 모델

3.1 제안하는 이미지 캡션 모델 구조

본 논문에서 제안하는 모델의 구조는 입력된 이미지로부터 객체를 검출하고 객체 정보를 추출하는 객체 검출단과 추출된 객체의 정보를 통해 단어를 생성해내는 문장 생성단으로 구성된다. 각 이미지의 부분적인 상황 이해를 정확하게 설명하기 위해 검출된 객체 중 메인 객체를 중심으로 각 객체들의 연관성이 많은 영역이 해당 이미지의 부분적인 상황을 설명하는 중요한 핵심 영역이라 볼 수 있다. **Figure 1**은 핵심 영역을 선택하기 위한 객체 정보를 생성하는 과정을 나타낸다.

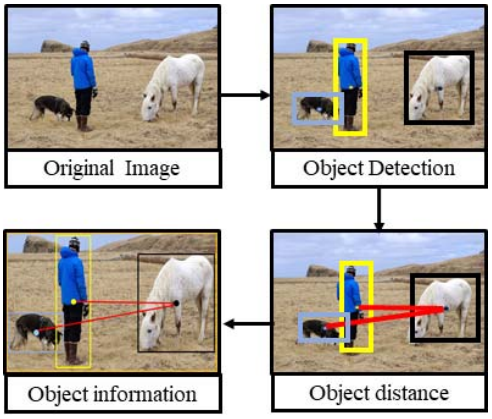


Figure 1: The process of creating object information

핵심 영역을 선택하기 위해 검출된 객체의 경계박스들의 면적을 산출하여 가장 면적이 큰 객체를 메인 객체로 선정하고, 기타 검출 객체들은 서브 객체로 지정한다. 메인 객체와 서브 객체들의 중심 좌표를 구하여 메인 객체의 중심 좌표를 기준으로 각 서브 객체의 중심 좌표와의 거리를 찾아낸다. 각 검출영역의 사이즈와 중심 좌표와의 거리를 통한 가중치를 생성하여 분류하고, 이를 통해 인코딩영역의 GRU에 필요한 이미지 정보의 입력 순서로 사용한다. 디코딩 영역에서는 이러한 이미지 정보가 들어옴과 동시에 문장의 첫 단어를 생성하여 이미지 캡션을 만들어낸다. GRU는 RNN의 특징을 그대로 가지고 있기 때문에 생성한 단어를 토대로 디코딩 영역에서는 시간이 지나감에 따라 이전 시간에 생성한 단어와 현재 시간에 들어오는 이미지의 정보를 토대로 단어를 이어서 생성해 나간다. 또한 가장 높은 확률로 검색이 되는 단어가 문장의 구조로 인해 가장 마지막에 포함되는 경우에도 정확한 문장을 생성할 수 있도록 Merge 레이어를 통해 모든 객체 상태를 고려한 문장을 생성하도록 한다. **Figure 2**는 제안하는 모델의 구조도이다.

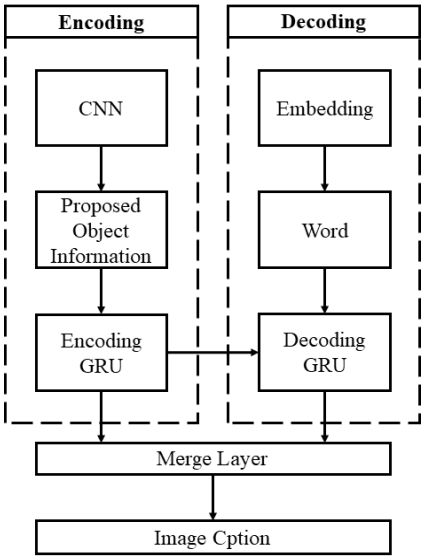


Figure 2: The architectures of proposed model

3.2 메인 객체 추출 및 거리에 따른 가중치 선정

객체 검출은 이미지 캡션에서 문장 생성을 위한 핵심 요소인 주어와 같기 때문에 정확한 이미지 전체 설명의 정확도와 직접적인 연관을 가진다. 정확한 이미지 캡션을 위해서는 생성되는 문장의 주어로 작용하는 객체에 대한 최적화가 필요하다. 본 논문에서는 객체 검출단에서 메인 객체를 선정하는 과정을 사람이 이미지를 보게 되는 경우를 참고하였다. 사람이 이미지를 볼 때 가장 먼저 보게 되는 객체는 이미지에서 가장 큰 객체를 우선적으로 보게 된다. 그러므로 메인 객체 선정을 위해 검출된 객체의 경계박스의 면적을 구하여 메인 객체를 선정한다. **Figure 3**은 검출된 객체의 경계박스를 나타낸다.

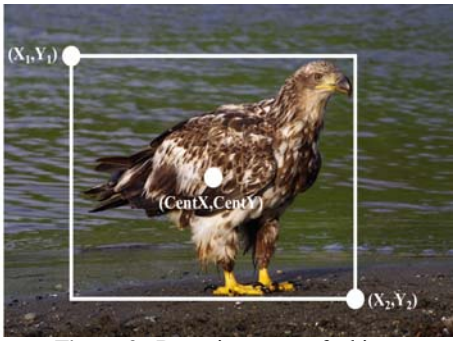


Figure 3: Detection area of object

메인 객체를 선정하기 위한 객체 경계박스의 면적인 S_{area}을 구하는 방법은 **식 (1)**과 같다.

$$S_{area} = X_2 \times Y_1 \tag{1}$$

여기에서 X_1과 Y_1은 경계박스의 좌측 상단의 좌표, X_2와 Y_2는 우측 하단의 좌표이다. 그 후 메인 객체의 중심 좌표를 추출하고, 메인 객체와 각 서브 객체의 중심 좌표간 거리를 구한다. 각 객체의 중심 좌표를 구하는 방법은 **식 (2)**과 같다.

$$centX = \frac{(X_2 - X_1)}{2}$$
$$centY = \frac{(Y_1 - Y_2)}{2} \quad (2)$$

$centX$와 $centY$는 각각 경계박스 중심 좌표의 X와 Y값을 나타낸다. 검출 영역의 중심 좌표를 통해 각 객체간의 거리를 구하는 방법은 유클리드 거리 기법을 이용하며 **식 (3)**과 같다.

$$D_{cent} = \sqrt{(centX_1 - centX_2)^2 + (CentY_1 - CentY_2)^2} \quad (3)$$

D_{cent}는 **식 (3)**을 통해 산출된 메인 객체와 각 객체의 중심 좌표간의 거리를 나타낸다. 각 서브 객체의 입력 순위를 선정하기 위해서는 각 객체에 대한 경계박스의 면적과 메인 객체와의 거리에 따른 가중치를 통해 우선순위를 정해야 한다. 인코딩 영역의 GRU의 입력으로 사용할 서브 객체의 우선순위인 P_{mark}를 정하는 방법은 **식 (4)**로 나타낸다.

$$P_{mark} = \frac{(S_{area} \times \alpha)}{D_{cent}} \quad (4)$$

여기에서 α는 거리에 따른 계수로써 우선순위를 정할 때 동순위가 나오는 것을 방지하기 위한 상수로 작용한다.

3.3 이미지 캡션 모델

이미지 캡션에서 GRU로 구성된 문장 생성단의 핵심은 인코딩 영역을 통해 이미지의 정보가 주어지는 경우 디코딩 영역에서 최적의 단어를 출력하여 이미지에 대한 설명의 정확도를 최대화 시키는 것이다. 객체 검출단에서 선정한 객체의 우선순위를 기준으로 GRU의 인코딩 영역에 객체 정보를 입력한다. GRU는 객체 정보가 들어옴에 따라 디코딩 영역의 GRU를 통해 문장의 시작 단어를 생성한다. **식 (5)**는 객체 정보가 입력되었을 때의 동작을 나타낸다.

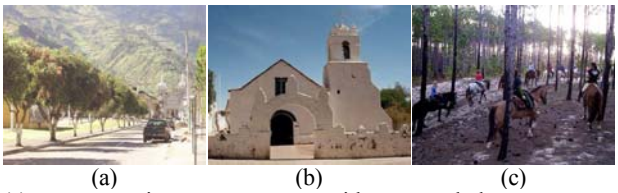
$$\log p(S|I) = \sum_{t=1}^{N} \log p(S_t|I, S_{1:t-1}) \quad (5)$$

I는 이미지의 입력을 나타내며, S_t는 문장 S의 t번째 단어를 의미한다. 여기에서 이미지의 입력은 CNN의 특징 맵과 객체 검출을 통해 추출한 객체 정보를 뜻한다. 또한 각 Decoding 영역의 GRU는 Softmax를 사용하여 어휘를 예측하고 단어를 생성해 낸다. 또한 정확한 이미지 캡션을 위해 Merge 레이어를 구성하여 객체 정보를 임시로 저장한다. 이 영역에서는 앞서 인코딩 영역에 입력으로 들어온 객체 정보를 모두 고려하여 좀 더 정확한 문장 생성이 가능하도록 한다. 문장의 주어에 사용되는 메인 객체가 먼저 입력으로 들어왔지만 문장의 마지막 구성에 배치가 되어도 그 문맥적 의미를 잃지 않게 만들어주는 역할을 한다.

4. 알고리즘 실험 및 고찰

4.1 이미지 캡션 데이터세트

본 논문에서는 이미지 캡션 생성 및 정확도 비교를 위하여 Flickr 8K**[11]**과 Flickr 30K**[12]**, MS COCO**[13][14]**와 같은 이미지 캡션 분야에서 많이 사용되는 검증된 데이터세트를 사용하였다. Flickr 8K, 30K 데이터세트는 각각 Flickr에서 추출한 8,000장, 31,783장의 이미지로 구성되어 있으며 이미지를 설명하는 5개의 캡션을 포함하고 있다. MS COCO 데이터세트는 82,783장의 학습용 이미지, 40,504장의 검증용 이미지로 구성되어 있으며, 일부 이미지에만 각 5개 이상의 캡션 문장이 제공된다. 본 논문에서는 학습을 위한 실험 환경을 구성하기 위해 **[2]**의 실험 환경과 비슷하게 구성하였으며 Flickr 80K 데이터세트의 6,500장의 이미지를 학습에, 1,000장의 이미지를 검증에 사용하였고, 500장의 이미지를 평가에 사용하였다. 또한, Flickr 30K 데이터세트에서는 29,500장의 이미지를 학습에 사용하였고, 1,014장의 이미지를 검증에, 500장의 이미지를 평가에 사용하였다. MS COCO데이터세트는 82,783장의 학습용 이미지를 사용하여 학습을 진행하였고, 검증 및 평가에 각 5,000장씩의 이미지를 사용하였다.



(a) (b) (c)

(a) two cars in a grey street with many dark green trees with red blooms on the left

(b) a small white church with a bell tower and a cross on it

(c) many people on brown horses in light brown sand in the foreground

Figure 4: Example of dataset image and caption

Figure 4는 이미지 캡션에 사용된 데이터세트의 이미지와 해당 이미지에 주어지는 문장들에 대한 예시를 나타낸다. 이미지 캡션 분야는 사람이 이미지를 보고 만든 캡션을 통해 모델에 대한 학습을 진행하며, 모델이 생성한 캡션과 데이터세트 이미지에 주어진 캡션을 비교하여 모델의 성능을 검증한다. 이미지 캡션의 평가 지표는 기존의 여러 이미지 캡션 논문들에서 사용하는 BLEU 스코어 방식인 BLEU-1, BLEU-2, BLEU-3, BLEU-4를 사용하였다**[15]**. BLEU 스코어 방식은

기계번역 시스템의 객관적인 성능평가지표로써 이미지 캡션에서도 사용되며 데이터세트에 제공되는 캡션 문장을 기준으로 하여 모델에서 생성하는 캡션 문장과의 비교를 통해 점수를 산출한다. 식 **(6)**과 **식 (7)**은 BLUE 점수의 산출 방법을 나타낸다.

$$BP=\begin{cases}1 & \text{if } c>r \\ e^{(1-\frac{r}{c})} & \text{if } c\le r\end{cases} \tag{6}$$

$$BLEU=BP\times\exp\left(\frac{1}{N}\sum_{n=1}^{N}\log P_N\right) \tag{7}$$

식 (6)의 BP는 Brevity Penalty로써 문장의 길이가 짧을수록 높은 점수를 받는 경우를 제거하기 위해 사용되며, r은 데이터세트에 제공되는 캡션 문장의 길이, c는 모델에서 생성된 캡션 문장의 길이를 뜻한다. **식 (7)**은 최종적인 BLEU 스코어를 산출하는 식으로 N은 평가에 적용하는 Gram의 수를 나타내며 P_N은 해당 Gram에 대한 정확도를 나타낸다.

Figure 5는 BLEU 점수를 산출하기 위한 간단한 예시 문장을 나타내었다. BLEU 점수 산출에는 일반적으로 1~4까지의 Gram을 이용하며 1-Gram일 때는 Reference 문장과 모델이 생성한 문장에서 한 단어씩 비교하며 4-Gram일 때는 4개의 연속적인 단어를 비교함으로써 문장의 구성에 대한 비교를 의미한다.

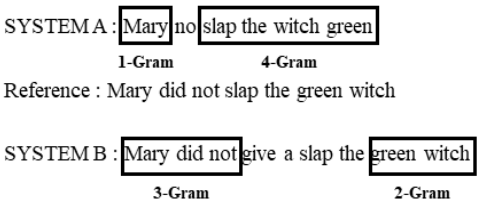


Figure 5: Example of BLEU

Table 1은 **Figure 5**의 문장을 사용하여 각 Gram에서의 정확도를 계산하고 BLEU-4에 대한 점수를 나타내었다. SYSTEM B의 경우 BLEU-4의 경우 훨씬 좋은 성능을 보임을 알 수 있으며, **Figure 5**의 예시를 통해 보았을 때 Reference와의 4-Gram의 정확도가 높기 때문에 SYSTEM A에 비해 문장의 표현력이 사람의 표현과 비슷해져 성능이 좋은 것을 알 수 있다.

Table 1: Results of **Figure 4** BLEU Score

	SYSTEM A	**SYSTEM B**
1-Gram Precision	5/6	7/9
2-Gram Precision	1/5	5/8
3-Gram Precision	0/4	3/7
4-Gram Precision	0/3	1/6
Brevity Penalty	6/7	10/7
BLEU-4	**0%**	**46%**

Figure 2에서 설명한 모델은 Keras를 사용하여 구현하였으며, CNN 모델은 정확한 객체 인식을 위해 ImageNet 데이터세트를 학습한 YOLO를 사용하였고, RNN은 GRU와 LSTM을 구현하여 이미지 캡션을 진행 및 비교 하였다.

Table 2: Experiment results of Flickr 8K dataset

Model	**B-1**	**B-2**	**B-3**	**B-4**
NIC**[1]**	63	41	27	-
GRU-DO4**[2]**	63.5	45.1	30.7	20.4
DeepVS**[3]**	57.9	38.3	24.5	16.0
m-RNN**[14]**	56.5	38.6	25.6	17.0
Proposed (GRU)	**65.2**	**46.6**	**31.3**	**20.7**
Proposed (LSTM)	64.5	43.7	28.4	17.1

Table 3: Experiment results of Flickr 30K dataset

Model	**B-1**	**B-2**	**B-3**	**B-4**
NIC**[1]**	66.3	42.3	27.7	18.3
GRU-DO2**[2]**	63.6	44.5	30.6	20.7
DeepVS**[3]**	57.3	36.9	24.0	15.7
m-RNN**[14]**	60.0	41.2	27.8	18.7
Proposed (GRU)	57.9	**45.4**	**31.9**	**21.1**
Proposed (LSTM)	58.2	44.1	30.4	20.9

Table 4: Experiment results of MS COCO dataset

Model	**B-1**	**B-2**	**B-3**	**B-4**
NIC**[1]**	66.6	46.1	32.9	24.6
GRU-DO1**[2]**	67.4	**50.7**	37.6	27.9
DeepVS**[3]**	62.5	45.0	32.1	23.0
m-RNN**[14]**	67	49	35	25
Proposed (GRU)	**68.1**	45.2	**38.0**	**28.7**
Proposed (LSTM)	67.5	44.6	35.4	28.1

4.2 실험 결과

Table 2는 Flickr 8K 데이터세트를 사용하여 이미지 캡션을 진행한 결과이다. 본 논문에서 제안한 모델이 기존 논문에 비해 높은 성능을 나타내는 것을 보아 캡션 생성이 잘 되며 이미지의 의미를 잘 전달한다는 것을 알 수 있다. 또한 **Table 3**은 Flickr 30K 데이터세트를 사용한 결과이며 이미지의 전체를 통해 캡션을 구성하는 구글의 NIC**[1]**에 비해 제안하는 모델은 객체 검출단에서 메인 객체를 선정하고 서브 객체와의 거리를 통해 객체 정보를 넘겨주기 때문에 검출된 객체가 작을 경우 캡션 구성에서 배제되어 BLEU-1에서는 성능이 다소 떨어지나 BLEU-2와 BLEU-3, BLEU-4의 평가지표는 제안하는 모델이 객체의 우선순위를 통해 단어를 생성하고 연결하며 Merge 레이어에서 입력된 객체 정보를 고려하여 최종적으로 캡션을 생성하기 때문에 단어 간 연결이 매끄러워 문장의 표현 성

능이 좋다는 것을 확인할 수 있다. **Table 4**의 MS COCO 데이터세트를 사용한 경우에는 제안한 모델이 BLEU-2를 제외한 나머지 성능이 좋다는 것을 확인할 수 있다. MS COCO 데이터세트의 경우 객체검출을 위한 데이터세트로도 많이 사용되어 객체의 정보 검출을 위한 객체 검출단의 사전 학습에 사용되어 BLEU-1의 점수가 제일 높았으나, BLEU-2의 점수는 GRU-DO1**[2]**에 비해 낮음을 알 수 있다. 또한 GRU와 LSTM의 성능을 비교하기 위해 추가적으로 한 실험 결과에서 전체적인 성능이 GRU가 좋으나 실질적인 성능차이는 1~3%차이로 미비한 것을 알 수 있다. 제안하는 모델의 성능이 좀 더 강인한 것은 GRU를 사용함으로써 Vanishing Gradient Problem으로 인해 문장이 길어질 경우 성능이 하락 하는 문제점을 보완 할 수 있기 때문이다. 객체 검출단을 통해 객체의 정보를 GRU에 추가적으로 입력함으로써 메인 객체 중심으로 이미지 영역을 지정하여 캡션을 구성하고 포함된 서브 객체의 우선순위에 맞춰 정확한 문장을 구성할 수 있음을 확인할 수 있다. 또한 제안한 모델의 경우 BLEU-4 지표의 성능이 전반적으로 좋기 때문에 생성된 문장의 표현력이 좋은 것을 알 수 있다.

(a) (b) (c)

(a) a big banana with the man.

(b) a white sea gull at a sandy beach with the sea and grey and brown rocks in the background

(c) Brown rocks and sea in background

Figure 6: Example of failed image caption generation

Figure 6은 이미지 캡션의 실패 예시로 **Figure 6 (a)**는 남자가 바나나를 들고 바나나조각상의 아래에 서있는 사진이지만 객체 정보를 넘겨줄 때 메인 객체로 선정된 큰 바나나 조각상의 모형에 초점이 맞춰져 캡션을 진행하기 때문에 남자의 손에 있는 작은 바나나가 경계박스 안에 동일한 바나나로 취급되어 검출기에서 인식하지 못하였다. 또한 **Figure 6 (b)**는 갈매기와 바다는 정확하게 인식하였으나 객체 검출기의 학습데이터에 섬이라는 개념이 없어 섬을 갈색돌로 인식하여 문장을 구성하였다. **Figure 6 (c)**는 눈이 덮인 벌판을 바다로 인식하여 문장을 구성하는 문제가 발생하였다. 이와 같은 문제점을 통해 객체 검출 기반 이미지 캡션에 있어 검출기의 학습데이터와 정확도는 이미지 캡션 정확도에 많은 영향을 주는 것을 알 수 있으며 메인 객체의 선정에 있어서도 적절한 메인 객체를 선정하지 못하면 문장의 주요 부분이 겹쳐 세밀한 부분을 제대로 인식하지 못하는 문제가 발생하는 것을 알 수 있다. 따라서 이미지 캡션 오류를 최소화하기 위해서는 객체 검출의 성능 향상을 위한 다양한 학습데이터가 필요하고, 적절한 메인 객체 선정 과정이 필요하다.

5. 결 론

본 논문에서는 고속 객체 검출이 가능한 YOLO 네트워크를 사용한 객체 검출단을 구성하여 추출된 객체 정보를 객체 영역의 크기와 거리를 가중치로 분류하여 GRU로 구성된 문장 학습단을 통해 캡션 생성이 가능한 네트워크를 제안한다. 제안한 모델은 Flickr 8K, Flickr 30K, MS COCO 데이터 세트를 통해 학습을 진행하였고 BLEU 평가지수를 통해 모델의 성능에 대한 객관적인 평가를 진행하였다. 아울러 기존의 이미지 캡션 모델에 비해 제안한 모델이 BLEU-4의 경우 높은 성능을 보이는 것을 알 수 있었으며 이를 통해 이미지의 의미를 잘 전달하는 것을 알 수 있다. 이 연구를 통하여 다양한 학습데이터를 통해 객체 검출기를 학습시키고 적절한 메인 객체 선정이 병행된다면 객체 검출 기반 이미지 캡션의 정확도가 높아질 것으로 기대된다.

후 기

이 논문은 2016년도 정부(교육부)의 재원으로 한국연구재단 기본연구지원사업의 지원을 받아 수행된 기본연구임 (No.2016R1D1A1B03934812)

References

[1] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and tell: A neural image caption generator,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3156-3164, 2015.

[2] C. K. Lee, “Image caption generation using recurrent neural network,” Journal of Korea Information Science Society, vol. 43, no. 8, pp. 878-882, 2016 (in Korean).

[3] A. Karpathy and L. Fei-Fei, “Deep visual-semantic alignments for generating image descriptions,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3128-3137, 2015.

[4] C. Liu, F. Sun, C. Wang, F. Wang, and A. Yuille, “MAT: A multimodal attentive translator for image captioning,” arXiv preprint arXiv:1702.05658, 2017.

[5] K. Cho, B. Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, “Learning phrase representations using RNN encoder-decoder for statistical machine translation,” arXive preprint arXive:1406.1078, 2014.

[6] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,”

arXiv preprint arXiv:1409.1556, 2014.

[7] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, and A. Rabinovich, “Going deeper with convolutions,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1-9, 2015.

[8] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770-778, 2016.

[9] S. H. Song, H. B. Hyeon, and H. Lee, “A Pedestrian Detection Method using Deep Neural Network,” Journal of Korea Information Science Society, vol. 44, no. 1, pp. 44-50, 2017 (in Korean).

[10] 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, pp. 779-788, 2016.

[11] M. Hodosh, P. Young, and J. Hockenmaier, “Framing image description as a ranking task: Data, models and evaluation metrics,” Journal of Artificial Intelligence Research, vol. 47, pp. 853-899, 2013.

[12] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille, “Deep captioning with multimodal recurrent neural networks (m-rnn),” arXiv preprint arXiv:141236632, 2014.

[13] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, and Y. Bengio, “Show, attend and tell: Neural image caption generation with visual attention,” International Conference on Machine Learning, pp. 2048-2057, 2015.

[14] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, and C. L. Zitnick, “Microsoft coco: Common objects in context,” arXiv preprint arXiv:1405.0312v3, 2015.

[15] K. Papineni, S. Roukos, T. Ward, and W. Zhu, “BLEU: a method for automatic evaluation of machine translation,” Proceedings of the 40th annual meeting on association for computational linguistics, pp. 311-318, 2002.