ISSN 2383-630X(Print) / ISSN 2383-6296(Online)
Journal of KIISE, Vol. 46, No. 4, pp. 369-375, 2019. 4
https://doi.org/10.5626/JOK.2019.46.4.369

객체 Attention을 이용한 이미지 캡션 생성
(Image Caption Generation using Object Attention Mechanism)

박 다 솔 †　　차 정 원 ††
(Da-Sol Park)　　(Jeong-Won Cha)

요 약 이미지 데이터가 폭발적으로 증가함에 따라 이미지를 자연어로 표현하기 위한 이미지 캡션 생성 기술에 대한 연구도 활발하게 이루어지고 있다. 기존 한국어 이미지 캡션 생성 기술에서는 영어권 데이터를 번역하여 사용함으로 인해 동시 발생 객체들에 의한 오류가 있다. 본 논문에서는 입력 이미지에 대한 캡션을 생성하여 추출한 명사와 이미지의 정답 캡션에서 추출한 명사를 이용하는 attention 함수를 새로운 손실 함수로 사용하는 이미지 캡션 모델을 제안한다. 공개된 실험 데이터를 사용한 실험에서 BLEU1 0.686, BLEU2 0.557, BLEU3 0.456, BLEU4 0.372를 보였다. 이를 이용하여 제안된 모델이 고빈도 동시 발생 객체 오류 해결에 효과적임을 입증하고 기존 연구보다 높은 성능을 얻음을 보이며 중복된 출력 문장을 줄임으로써 이미지 캡션의 다양한 표현들이 생성에 효과적임을 보였다. 또한 본 논문에서 제안하는 방법을 이용하여 이미지 캡션 모델을 학습하기 위한 코퍼스를 생성할 수 있다.
키워드: 어텐션 매커니즘, 심층 학습, 이미지 캡셔닝, 자연어 처리, 기계 학습

Abstract Explosive increases in image data have led studies investigating the role of image caption generation in image expression of natural language. The current technologies for generating Korean image captions contain errors associated with object concurrence attributed to dataset translation from English datasets. In this paper, we propose a model of image caption generation employing attention as a new loss function using the extracted nouns of image references. The proposed method displayed BLEU1 0.686, BLEU2 0.557, BLEU3 0.456, BLEU4 0.372, which proves that the proposed model facilitates the resolution of high-frequency word-pair errors. We also showed that it enhances the performance compared with previous studies and reduces redundancies in the sentences. As a result, the proposed method can be used to generate a caption corpus effectively.
Keywords: attention mechanism, deep learning, image captioning, natural language processing, machine learning

· 이 논문은 2017~2018년도 창원대학교 자율연구과제 연구비 지원으로 수행된 연구결과임

† 학생회원 : 창원대학교 친환경해양플랜트FEED공학과
dasol_p@changwon.ac.kr
†† 종신회원 : 창원대학교 컴퓨터공학과 교수(Changwon Nat'l Univ.)
jcha@changwon.ac.kr
(Corresponding author임)

논문접수 : 2019년 2월 25일
(Received 25 February 2019)
논문수정 : 2019년 4월 5일
(Revised 5 April 2019)
심사완료 : 2019년 4월 8일
(Accepted 8 April 2019)

1. 서 론

스마트폰과 각종 센서들의 상용화로 인해 이미지 데이터의 양이 증가함에 따라 이미지 데이터의 활용성이 증가하고 있다. 이미지 데이터의 활용 기술 중 이미지 캡션 생성 기술은 이미지를 설명하는 텍스트를 생성하는 기술이다. 이미지 캡션을 수행하기 위해서는 이미지를 이해하기 위한 이미지 처리 모델과 문장 생성을 위한 자연어 처리 모델이 함께 사용되어야 한다. 이미지 캡션을 이용한 분야가 다양하게 존재하며 그 중 Microsoft에서는 시각 장애인들을 위해 주변 환경을 설명해주는 Seeing AI 앱[1]을 출시하는 등 상용화를 앞두고 있다.

최근 딥러닝을 이미지 처리 및 자연어 처리 분야에 적용하여 높은 성능을 보이고 있으며 이에 따라 이미지 캡션 생성 연구 또한 많은 발전이 있었다. 그러나 한국어 이미지 캡션 생성 연구의 경우 기계 학습이나 딥러닝을 적용할 수 있는 데이터가 부족하여 많은 연구가 이루어지지 못하고 있다. 이를 극복하기 위해 공개된 영어 이미지 캡션 데이터인 MicroSoft COCO를 번역하고 번역으로 인해 발생한 오류는 수기로 수정하여 한국어 데이터를 생성하였다.

본 논문에서는 이미지 캡션 데이터의 문제점들 중 동시 발생 객체들에 대한 오류를 해결하고자 한다. 동시 발생 객체는 하나의 객체가 인식이 되면 이미지에 나타나지 않음에도 불구하고 캡션 결과에 고빈도의 동시 발생 객체가 출력되는 문제이다. 그림 1은 동시 발생 객체 오류의 예를 보여준다.

그림 1의 이미지의 정답 캡션 문장은 '테이블 위에 콜라와 빵이 놓여 있다.'이지만 생성된 캡션 문장은 '테이블에 감자 튀김과 콜라가 있습니다.'이다. 이것은 '감자 튀김'과 '콜라'가 함께 나타나는 빈도가 높기 때문에 '감자 튀김'이 존재하지 않아도 '감자 튀김'을 출력하는 오류의 경우이다. 우리는 이 문제를 '고빈도 동시 발생 객체 오류'라고 한다.

Generated Sentence	There are fries and cola on the table.

그림 1 고빈도 동시 발생 객체 오류의 예

Fig. 1 Example of highly frequent and concurrent object errors

동시 발생 객체 오류를 해결하기 위해 본 연구에서는 입력 이미지를 이용하여 캡션을 생성하여 추출한 명사와 이미지의 정답 캡션에서 추출한 명사를 객체라고 가정한다. 생성한 한국어 데이터를 이용하여 생성한 캡션 내 객체와 정답 캡션 내 객체를 이용하여 attention을 계산하고 이를 새로운 손실 함수로 사용하는 방법으로 동시 발생 오류를 저감하는 결과를 보인다.

2. 이전 연구

이미지 캡션 생성에 대해서는 Recurrent Neural Network(RNN)을 사용한 모델의 연구를 시작으로 하여 많은 연구가 진행되고 있으며, 단순한 RNN 방식을 통한 문장 생성을 진행할 때 문장의 길이가 길어질수록 이전 단어에 대한 정보가 소멸되는 Vanishing Gradient Problem을 해결하기 위해 Long-Short Term Memory(LSTM), Gated Recurrent Unit(GRU)와 같은 알고리즘을 많이 사용하고 있다.

영어권에서 이미지 캡셔닝의 이전 연구는 아래와 같다. [2]는 Multimodal RNN을 이용하여 이미지 캡션 생성을 진행하였다. 이 모델의 언어 모델은 단어를 생성하기 위한 임베딩 레이어를 2개의 층으로 구성했다. Multimodal 레이어와의 연결을 통해 이미지 캡션을 진행한다. 기본적인 RNN을 사용하기 때문에 Vanishing Gradient Problem 문제점이 존재한다. [3]은 Convolution Neural Network(CNN)과 LSTM을 이용해 이미지 캡션을 생성하였는데 CNN은 Inception V3를 이용하였다. LSTM을 사용하여 이미지 캡션을 생성함으로써 Vanishing Gradient Problem을 해결하였다. 어떤 단어가 이미지의 어느 부분에 크게 반응하는지 학습하는 주의집중 메커니즘(Attention mechanism)을 반영하여 개선된 성능을 보였다.

한국어권에서 이미지 캡셔닝의 이전 연구는 아래와 같다. 이미지에 따라 번역 작업을 진행했으며, 크게 Flickr 8K[5]와 Miscrosoft COCO로 구분할 수 있다. Flickr 8K를 사용한 논문인 [4]는 CNN과 LSTM의 변형인 GRU와 Residual Network를 이용하는 방법을 제안하였으며 Flickr 8K 데이터를 대상으로 영어 이미지 캡션을 번역자가 번역해 사용하였다. Residual Network는 네트워크의 깊이가 깊어질수록 학습 및 평가 에러가 줄지 않는 현상을 해결한 기술로, 몇 단계 이전 히든 레이어의 입력을 가중치 없이 현재의 입력에 더해주는 기술이다.이를 이용함으로써 GRU를 사용하는 모델[6]에 비해 성능의 개선이 있었다. [7]은 입력 이미지 정보를 소스로 보고 Context-Gate를 이미지 캡션에 적용하였

다. 디코딩 시 소스 정보와 타겟 정보 비율을 적절히 분배하는 Context-Gate를 적용하여 [4]의 이미지 캡셔닝 모델의 성능보다 개선된 성능을 보였다.

Microsoft COCO를 사용한 논문인 [8]은 이미지 캡션 모델을 구성하기 위해 CNN과 LSTM을 모델을 사용하였다. 이를 어절 단위, 형태소 단위, 의미 형태소 단위로 학습 데이터를 구분하여 모델의 입력으로 사용하였다. 어절 단위보다 형태소 단위로 학습하였을 때 더 높은 성능을 보였다. 캡션 결과를 분석했을 때 크게 저빈도 객체에 대한 오류, 고빈도 동시 발생 객체에 대한 오류가 발생했다.

3. 제안 방법

한국어 이미지 캡션 데이터를 생성하기 위해서 Microsoft COCO 영어 캡션 데이터셋을 한국어로 번역하고 번역으로 인해 발생한 오류는 수기로 수정하였다. 영어권에서는 단어 단위로 캡션을 생성한다. 하지만 한국어 모델에서 어절 단위로 입력을 사용하게 되는 경우 조사에 따라 같은 단어도 다른 단어로 인식하는 문제가 존재한다. 따라서 기존 영어 데이터에 비해 학습 데이터의 양이 훨씬 많아야 하는 문제점이 존재한다. 본 논문에서는 [8]의 실험 결과를 바탕으로 형태소 단위로 캡션을 생성하였다.

본 논문에서는 [3]에서 제안한 구조를 기본으로 한다. RNN의 문제점을 해결할 수 있는 LSTM 모델과 Attention Mechanism을 적용하여 고빈도 동시 발생 객체에 대한 오류를 개선하고자 한다. [3]에서는 입력 이미지를 CNN을 사용하여 512차원으로 인코딩한 후 이를 LSTM의 입력으로 사용해 문장을 생성한다. 그림 2는 본 논문에서 제안하는 구조이다. 본 논문에서는 위에 기술된 오류를 해결하기 위해 새로운 손실 함수를 설계하였고 객체 추출(Object Detection) 모델의 결과물은 명사만을 취급한다라고 가정한다.

본 연구의 제안 모델은 다음과 같이 정의할 수 있다.

$$\Theta^{*} = \arg\max_{\Theta} \sum_{I,S} \log p(S|I,X;\Theta) \quad (1)$$

여기서 Θ는 LSTM 모델의 전체 파라미터이고, I는 이미지, S는 정답 문장이다. 또한 X는 이미지에서 추출된 객체들이다. 여기서 가변 길이 S를 처리하기 위해 다음과 같이 체인 규칙(chain rule)을 적용한다.

$$\log p(S|I;\Theta) = \sum_{t=0}^{N} \log p(S_t|I,S_0,S_1,\cdots,S_{t-1};\Theta) \quad (2)$$

학습 시에 (S,I) 쌍을 이용하여 식 (2)의 식을 최적화

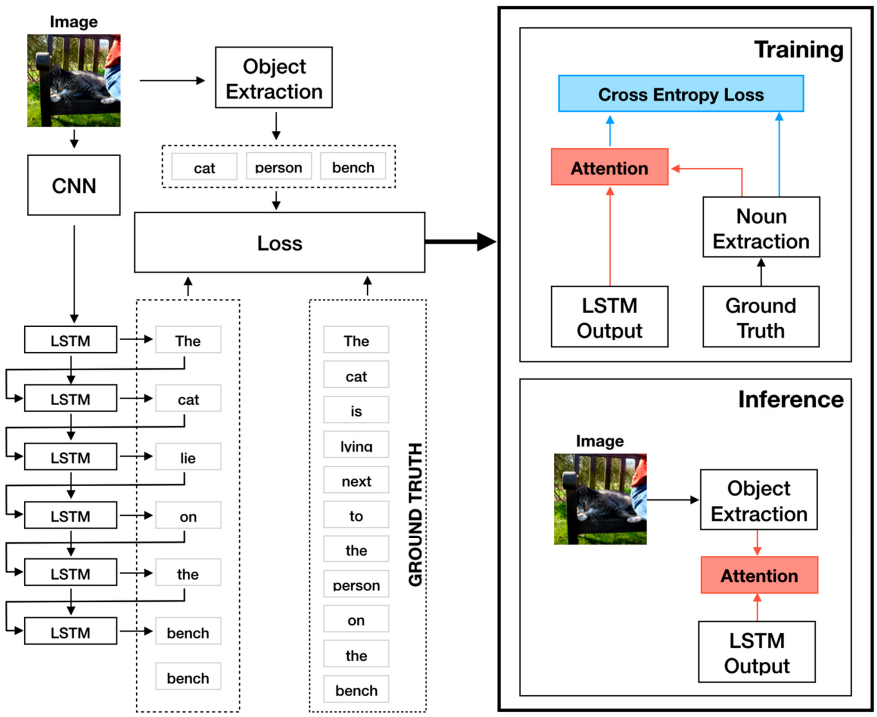

그림 2 제안 모델 구조

Fig. 2 Proposed model architecture

한다. 가변 길이를 처리하기 위해서 LSTM을 사용한다. 또한 이미지를 표현하기 위해 CNN을 사용한다. 이 모델은 현재 이미지 처리 문제와 객체 인식 문제에서 가장 널리 사용하고 있다. 객체 인식 및 추출을 위해서 YOLO 9000[9]를 사용한다.

3.1 학습

LSTM 모델은 $p(S_t|I,S_0,\cdots,S_{t-1};\Theta)$에 의해서 생성된 단어와 이미지에 의해서 생성되는 각 단어들을 이용하여 학습되고 $t-1$에서의 LSTM의 결과물인 m_{t-1}은 t의 LSTM에 입력으로 사용된다. 좀 더 자세하게 기술하면 다음과 같다.

$$z_{-1} = CNN(I)$$
$$z_t = W_e S_t^{'}, \mathrm{t}\in\{0\cdots \mathrm{N}-1\} \quad (3)$$
$$p_{t+1} = LSTM(z_t), \mathrm{t}\in\{0\cdots \mathrm{N}-1\}$$

여기서 각 단어($S_t^{'}$)는 원핫(one-hot) 벡터로 표현한다. $S_0^{'}$는 시작을 표시하는 특별한 문자이고 $S_N^{'}$는 문장의 마지막을 표시하는 특별한 문자이다. 이 식에서 CNN에 의해서 생성된 이미지 정보와 단어 임베딩(W_e)에 의해서 표현된 단어들이 같은 공간에 매핑된다. 이미지는 $t=-1$에 한번 입력된다.

이렇게 만들어진 문장($S^{'}$)은 객체 추출에 의해서 생성된 단어들과 attention을 수행한다.

$$S = S^{'} \cdot X \quad (4)$$

학습 시에는 객체 추출에 의해서 생성되는 단어를 다른 방법으로 생성한다. 우리는 정답 문장에서 명사들을 추출하여 이들을 객체 추출의 결과물이라고 가정한다. 따라서 이 명사들과 LSTM의 출력 문장 내 명사를 추출하여 attention을 적용한다. 객체 추출 명사와 LSTM 출력 문장 내 명사를 이용하여 멀티핫(multi-hot) 벡터를 생성한다. 즉, 벡터의 길이는 사전크기와 같으며 해당 명사가 존재하면 1로 표기하고 존재하지 않으면 0으로 표기한다. 객체 추출 명사의 멀티핫 벡터와 LSTM 출력 문장 내 명사의 멀티핫 벡터를 이용하여 Cosine similarity를 구하여 손실 함수로 적용하였다. 멀티핫 벡터를 Cosine simliarity를 통해 계산하는 것은 Attention을 하는 것과 동일한 효과를 가진다. 생성된 문장 내 객체들과 정답 캡션 내 객체들 중 동시에 나타나는 객체에 가중치를 더 둔다는 것을 의미하고 이로 인해 고빈도 동시 발생되는 객체에 대한 가중치를 낮추는 역할을 한다.

우리의 손실 함수는 각 단계에서 정확한 단어들의 음의 로그 우도(likelihood)의 합을 나타낸다.

$$L(I,S,X) = -\sum_{t=1}^{N} \log p_t(S_t^{'}) \quad (5)$$

식 (5)의 손실 함수는 CNN을 이용한 이미지 정보와 단어 임베딩 정보, 그리고 객체 추출된 단어정보들을 입력으로 하여 LSTM의 모든 파라미터에 대해서 최소화된다.

3.2 추론

하나의 이미지에 캡션을 생성하는 방법은 여러 가지가 있다. 본 논문에서는 BeamSearch 방법을 사용한다. 시간 t까지 k개의 문장을 생성하고 시간 $t+1$에 후보 문장을 검토하여 다시 k개의 문장을 남긴다. 따라서 최종적으로 $S^{*}=\mathrm{argmax}_S p(S|I,X)$를 이용하여 최적 문장을 선택한다.

4. 실 험

본 논문에서는 제안 방법이 한국어뿐만 아니라 다양한 언어에도 효과적으로 동작함을 증명하기 위해 영어와 한국어에 대해 이미지 캡션 생성을 진행하였다. 본 논문에서는 한국어 이미지 캡션 생성을 위해 Microsoft COCO 번역 데이터셋을 사용하였다. Microsoft COCO 데이터셋은 123,287개의 이미지와 하나의 이미지당 5문장의 캡션으로 총 616,435문장으로 구성된다. 이 중 117,211개의 이미지를 학습에 사용하였고 2,025개의 이미지를 검증에 사용하였으며 4,051개의 이미지를 테스트에 사용하였다.

한국어 데이터셋은 [8]의 연구결과에 따라 형태소 단위로 구축하였다. Attention을 위해 문장에서 명사를 추출하는 작업을 진행해야만 한다. 한국어의 경우 POS 태깅을 통해 명사를 추출하였고 영어의 경우 nltk[10]에서 지원하는 stopword를 제외한 나머지를 명사로 간주하여 추출을 진행하였다. 우리는 [8]에서 제안한 모델을 '기본 모델'이라고 칭한다.

기본 모델의 결과는 표 1과 같다. 형태소 단위의 한국어(Korean)와 영어(English)로 적용한 모델은 CNN과 LSTM을 이용한 이미지 캡션 생성 모델이다. 표 2는 제안 모델의 이미지 캡션 생성 성능을 나타낸 것이다.

표 1 기본 모델 실험 결과
Table 1 Experimental results of baseline model

Language	BLEU 1	BLEU 2	BLEU 3	BLEU 4
Korean[7] (Morpheme unit)	0.630	0.445	0.333	0.260
English[7]	0.611	0.441	0.323	0.250

표 2 제안 모델 실험 결과
Table 2 Experimental results of proposed model

Language	BLEU 1	BLEU 2	BLEU 3	BLEU 4
English	0.632	0.463	0.342	0.263
Korean	0.686	0.557	0.456	0.372

제안 모델에서는 영어와 한국어로 실험을 진행하였다. 성

능은 BLEU 1, BLEU 2, BLEU 3, BLEU 4로 측정하였다. 실험 결과 전반적으로 성능이 향상됨을 볼 수 있다. 특히 한국어의 경우 BLEU 4의 성능이 크게 향상되었다.

기본 모델의 고빈도 동시 발생 객체의 오류를 해결함으로써 전체 성능 향상에 도움이 된다는 것을 증명하기 위해 전체 오류 중 고빈도 동시 발생 객체의 오류가 차지하는 비율을 확인하였다. 모델의 전체 오류와 고빈도 동시 발생 객체 오류를 사람이 모두 확인하는 것은 시간과 비용이 많이 드는 작업이다. 그렇기 때문에 오류에 대해 제약사항을 둔다. 전체 오류의 경우 이미지의 정답 캡션 문장과 생성된 캡션 문장을 이용하여 BLEU 4를 계산하여 BLEU 4를 기준으로 정렬한 후 하위 5%, 10%, 15%, 20%를 나누어 해당 범위에 나타나는 오류들을 전체 오류라고 간주한다.

고빈도 동시 발생 객체 오류는 모델의 결과에는 고빈도 동시 발생 객체가 존재하지만 정답 문장에는 고빈도 동시 발생 객체가 존재하지 않는 경우를 고빈도 동시 발생 객체 오류가 발생하였다고 간주하였다. 고빈도 동시 발생 객체는 학습 데이터에서 나타난 객체 쌍의 빈도를 측정하여 상위 1%에 포함되는 객체 쌍을 의미하며 고빈도 동시 발생 객체 오류로 정의하였다. 표 3은 기본 모델에서 한국어와 영어에 대해 전체 오류 중 고빈도 동시 발생 객체 오류의 비율을 나타낸 것이다. 고빈도 동시 발생 객체 오류가 차지하는 비율은 한국어의 경우 15~23%이며, 영어의 경우 72~80%로 고빈도 오류가 전체 성능에 영향을 미치고 있다고 말할 수 있다.

영어의 경우 고빈도 동시 발생 객체 오류의 비율이 매우 높은 것을 확인할 수 있었다. 이는 활용 형태로 사용되는 경우와 동의어의 사용이 많기 때문에 고빈도 동시 발생 객체 오류라고 잘못 인식되어 추출되는 경우가 많기 때문이라고 판단된다. 예를 들어 고빈도 동시 발생 객체가 "테니스(tennis)"와 "라켓(racket)"이고 정답 문장에서는 "테니스(tennis)"만 나타났으나 모델의 결과가 "테니스(tennis)"와 "라켓(racket)"을 생성했다고 하는 오류가 있을 때 정답 문장에는 "라켓(racket)"이 존재하지 않지만 "라켓(racquet)" 과 같이 동의어를 사용하는 경우가 있다. 또 고빈도 단어쌍이 "앉기(sitting)", "테이블(table)"이고 정답 문장에서는 "테이블(table)"만 나타났으나 모델의 결과가 "앉기(sitting)"과 "테이블(table)"을 생성했다고 하는 오류가 있을 때 정답 문장에는 "앉기(sitting)"이 없으나 "앉다(sit)"이 존재하는 경우가 있다. 제안 모델에서 고빈도 동시 발생 객체 오류가 해결된 개수를 확인하기 위해 표 4는 기본 모델에서 발생한 고빈도 동시 발생 객체 오류를 제안 모델이 해결한 오류에 대한 정보이다.

표 3 고빈도 동시 발생 객체 오류가 전체 오류에서 차지하는 비율

Table 3 Proportion of highly frequent and concurrent object errors among the total errors

Language	Group	Number of high frequent co-occurrent objects errors	Number of whole errors	Ratio
Korean	5%	47	202	23%
	10%	75	405	19%
	15%	96	607	16%
	20%	125	810	15%
English	5%	146	202	72%
	10%	311	405	77%
	15%	473	607	78%
	20%	644	810	80%

한국어의 경우 전체 평가 데이터의 결과에서 397개의 고빈도 동시 발생 객체 오류 중 169개를 해결해 42.6%의 많은 오류를 해결하였지만 영어의 경우 총 3,039개의 고빈도 동시 발생 객체 오류 중 424개를 해결해 14.0%의 비교적 적은 오류를 해결하였다.

표 5는 기본 모델과 제안 모델에서 각각 발생하는 고빈도 동시 발생 객체의 오류에 대한 예시를 보여준다. 예를 들어 발생한 고빈도 동시 발생 객체는 [비디오/NNG, 무리/NNG]이며 정답 캡션에 나타난 객체는 '비디오/NNG'이다. '비디오/NNG'와 '무리/NNG'가 함께 나오는 오류 발생 수는 기본 모델에서는 총 10번이 발생하였고, 제안 모델에서는 4번 발생하였다. 이러한 결과로부터 고빈도 동시 발생 객체에 대한 오류는 개선되었다고 할 수 있다.

또 다른 관점에서 낮은 성능에 대해 분석을 진행해보았다. 높은 비율의 중복된 캡션 문장을 출력하는 것이 성능 저하의 원인이라고 판단하였다. 실제 학습 이미지에서 유사한 이미지가 굉장히 많이 나타나는 경우이다. 서로 다른 서핑보드의 이미지는 굉장이 유사한 이미지로 구성되어 있음을 알 수 있고 해당 이미지들은 모두 "한 남자가 서핑보드를 타고 있다"라는 캡션을 생성하

표 4 해결된 오류 개수에 대한 정보

Table 4 Number of resolved errors

	Korean	English
Number of high frequent co-occurrent objects whole errors	397	3,039
Number of high frequent co-occurrent object errors resolved (Number of errors resolved by the proposed model)	169	424
Resolution ratio	42.6%	14.0%

표 5 개선된 고빈도 동시 발생 객체 오류의 예제

Table 5 Examples of improvement highly frequent and concurrent object errors

high frequent co-occurrent objects	The object shown in the correct caption	Number of errors in the baseline model	Number of errors in the proposed model
[비디오/NNG, 무리/NNG]	비디오/NNG	10	4
[시내/NNG, 길/NNG]	시내/NNG	8	4
[데스크/NNG, 책상/NNG]	데스크/NNG	6	4
[침대/NNG, 탁자/NNG]	침대/NNG	6	4
[벤치/NNG, 무리/NNG]	벤치/NNG	6	4

였다. 위와 같이 이미지가 굉장히 유사하게 나타나기 때문에 유사한 이미지 벡터가 생성되었으며 정답이 다양하게 생성되는 이미지 캡션 셋이 생성된다. 따라서 모든 정답 캡션에서 공통적으로 나타나는 단어를 추출하는 경우를 결과로 나타냈다고 볼 수 있다. 이를 해결하기 위해 제안된 모델로 실험을 진행하였고 중복된 캡션 문장의 수가 감소되는 것을 확인할 수 있었다.

학습 데이터의 전체 이미지는 119,236개이며 Best-1을 선택했을 때 중복 제거 후 문장은 18,664개로 100,572(84.35%)개의 중복 문장이 있다. Best-1이 아닌 전체의 경우 하나의 이미지당 3개의 문장이 있기 때문에 357,708개의 문장이며 중복 제거 후 문장은 46,509개로 311,199(87.00%)개의 중복 문장이 있다.

이러한 중복 문장을 출력하는 현상이 개선됨을 확인하기 위해 학습 데이터셋에 대해 캡션 문장을 생성한 후 동일하게 중복이 없는 문장 수를 측정하는 작업을 진행하였다. 기본 모델과 제안 모델을 그대로 사용하였을 때 위와 같은 문제가 있는지 확인하였다. 표 6은 기

표 6 기존 학습 데이터와 수정된 학습 데이터 비교표
Table 6 Comparison of original and modified in training data

Classification	original training data	modification training data
Number of whole image	119,236	
Number of whole Best 1 sentence	119,236	
Number of Best 1 deduplicate sentence	14,142	18,663 (+4,521)
Number of whole Best 3 sentence	357,708	
Number of Best 3 deduplicate sentence	36,550	46,509 (+9,859)

표 7 기존 평가 데이터와 수정된 평가 데이터 비교표
Table 7 Comparison of original and modificated in test data

Classification	original test data	modification test data
Number of whole image	4,051	
Number of whole Best 1 sentence	4,051	
Number of Best 1 deduplicate sentence	1,768	2,006 (+238)
Number of whole Best 3 sentence	12,153	
Number of Best 3 deduplicate sentence	4,674	5,123 (+449)

존 학습 데이터와 수정된 학습 데이터의 비교표이며, 표 7은 기존 평가 데이터와 수정된 평가 데이터의 비교표이다. 수정된 학습 데이터와 수정된 평가 데이터는 제안 모델을 적용하여 생성된 캡션 결과이다. 표 6과 표 7에 명시된 괄호의 숫자는 기존 학습 및 평가 데이터보다 수정된 학습 및 평가 데이터에서 중복 제거 후 문장 수가 증가하였고 Best 1 문장을 기준으로 하였을 때 24.22%의 비율로 증가하였다. 이는 기존의 학습 및 평가 데이터에 비해 제안된 모델에 의해 생성된 문장에서 중복 없는 문장 수가 증가하였음을 보여주는 실험이고 캡션의 다양한 표현들이 생성되는 것에 효과적임을 보인다.

5. 결론 및 고찰

기존 한국어 이미지 캡션 생성 기술에서는 영어권 데이터를 번역하여 사용함으로 인해 동시 발생 객체들에 의한 오류가 있다. 본 논문에서는 입력 이미지를 이용하여 생성된 캡션에서 추출한 명사와 이미지의 정답 캡션의 추출한 명사를 이용하는 attention 함수를 새로운 손실 함수로 사용하는 객체 attention을 적용한 이미지 캡션 모델을 제안한다.

제안 방법이 동시 발생 객체 쌍 오류 해결에 효과적임을 보였으며, 동시 발생 객체 쌍 오류를 해결함으로써 중복이 없는 문장을 생성하는 비율도 증가하였다. 실험 결과에 의해 기본 모델보다 제안 모델의 결과가 문장 내 명사를 정확하게 생성하고 이미지 캡션의 다양한 표현들이 캡션 생성에 있어 효과적임을 보였다. 또한 제안 모델을 영어, 한국어에 적용하여 하나의 언어가 아닌 여러 언어에 적용할 수 있음을 보였다.

향후 연구로는 이미지 캡션의 문제점인 저빈도로 나타나는 객체를 다른 객체로 오인식하는 오류를 해결할 수 있는 방안에 대해 연구를 진행해 볼 예정이다.

References

[1] [Online]. Available: https://www.microsoft.com/en-us/seeing-ai/

[2] Mao, Junhua, et al., "Deep captioning with multimodal recurrent neural networks (m-rnn)," *arXiv preprint arXiv:1412.6632*, 2014.

[3] VINYALS, Oriol, et al., "Show and tell : Lessons learned from the 2015 mscoco image captioning challenge," *IEEE transactions on pattern analysis and machine intelligence*, pp. 652-663, 2017.

[4] Jangseong Bae, Changki Lee, "Korean Image Caption Generation using Deep Learning," *Proc. of the KIISE Korea Computer Congress 2016*, pp. 488-490, Dec, 2016. (in Korean)

[5] Hodosh, Micah, Peter Young, and Julia Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics," *Journal of Artificial Intelligence Research*, 47, pp. 853-899, 2013.

[6] Changki Lee, "Image Caption Generation using Recurrent Neural Network," *KIISE*, Vol. 43, No. 8, pp. 878-882, 2016. (in Korean)

[7] Jangseong Bae, Changki Lee, "Image Caption Generation with Context-Gate," *Proc. of the KIISE Korea Computer Congress 2018*, pp. 488-490, Jun. 2018. (in Korean)

[8] Seong-Jae Park, Jeong-Won Cha, "Generate Korean image captions using LSTM," *Annual Conference on Human and Language Technology*, pp. 082-084, 2017. (in Korean)

[9] Redmon, Joseph, and Ali Farhadi, "YOLO9000: better, faster, stronger," *Proc. of the IEEE conference on computer vision and pattern recognition*, pp. 7263-7271, 2017.

[10] Loper, Edward, and Steven Bird., "NLTK: the natural language toolkit," *arXiv preprint cs/0205028*, 2002.

박 다 솔

2014년 창원대학교 학사. 2017년 창원대학교 석사. 2017년~현재 창원대학교 친환경해양플랜트 FEED공학과(정보통신·컴퓨터전공) 박사. 관심분야는 자연어처리, 딥러닝, 기계학습

차 정 원

숭실대학교(학사). 포항공과대학교(석사,박사). USC/ISI(박사후연수). 2004년~현재 창원대학교 컴퓨터공학과 교수. 관심분야는 자연어처리, 기계학습, 정보검색